概率论与数理统计

夏　强　刘金山　主编

人 民 邮 电 出 版 社
北 京

内 容 提 要

本书是基于普通高等教育“十二五”国家规划教材，针对教学改革而出版的少学时的公共课教材．全书共十一章，前五章为概率论部分，第六章至第十章为数理统计部分，第十一章为 R 软件的介绍．

本书是根据非数学类专业概率论与数理统计教学的基本要求，结合作者多年来教学实践中的经验和体会，在对已有教材进行认真改进的基础上编写而成的．本书论述严谨、通俗易懂、注重应用，力求深入浅出，便于学生学习掌握概率论与数理统计的基本内容和方法，并了解和掌握一些现代统计方法及应用．

本书适合作为普通高等学校非数学、非统计学类各专业概率论与数理统计课程的教材或学习参考书，特别适合工科、理科、经济、管理和农林类各专业的学生使用，也可作为各类科技人员和管理人员的参考书．

♦ 主　　辑　夏　强　刘金山
责任编辑　张　斌
责任印制　沈　蓉

♦ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京　　印刷

♦ 开本：787×1092　1/16
印张：　　2018 年　月第 1 版
字数：　　千字　　2018 年　月北京第 1 次印刷

定价：　　元

读者服务热线：(010) 81055256　印装质量热线：(010) 81055316
反盗版热线：(010) 81055315
广告经营许可证：京东工商广字第 8052 号

前　言

概率论与数理统计是研究随机现象统计规律的现代数学分支之一，有着非常广泛的应用背景，在工业、农业、商业、军事、科学研究、工程技术、经济管理等几乎所有领域都有重要应用. 随着现代科学技术的迅猛发展，特别是计算机和信息技术的发展，近年来概率统计方法在经济、金融、保险、生物、农林、医学和管理等许多领域中得到了广泛应用和深入发展. 正是这种广泛的应用性，使得概率论与数理统计课程成为高等学校大部分专业开设的一门重要的必修或选修课程. 通过本课程的学习，学生可以掌握处理随机性观察数据的基本理论和方法，为各专业知识的深入学习或应用打下良好的基础.

关于概率论与数理统计的教材已非常多，这类教材主要以经典概率统计的理论为基础，讲述其理论、方法与例题分析，目的是为了帮助读者理解和掌握基本的概率统计概念和方法. 随着教学改革的要求，概率论与数理统计的学时在逐渐减少，这就对概率统计的学科发展和教学内容的改革提出了挑战. 为此，本书力求在以下几个方面做一些尝试：

（1）以初等概率统计知识为起点，与大学概率统计知识有机地联系在一起；

（2）以概率论为基础，以数理统计为主线，立足于概率统计的基本理论和方法；

（3）尽量与现代科学技术特别是信息技术发展相适应，强调应用性、实效性；

（4）有一定的可塑性，能广泛适用于普通大学理科、工科以及经济类、管理类、农林类等专业；根据专业的不同，教师可根据其特点和需要选择教学内容和习题；

（5）深入浅出，易教易学，突出重点，强调案例式教学方法，并有针对性地配备了 R 软件程序和示例；

（6）根据课程教学改革的少学时要求，有针对性地进行了整合.

当然，上述想法只是作者编写此书的希望和初衷，本书实际上还远没有达到这样的目标.

本书共十一章，内容包括：随机事件及其概率，随机变量及其分布，变量的数字特征，极限定理，抽样分布理论，参数估计，假设检验，方差分析，回归模型和 R 软件介绍. 文中加 * 号的部分视学生的情况作为选讲内容，各章包含大量例题和习题，有些内容还提供了用 R 软件进行统计分析的程序和示例，书末配有习题参考答案.

本书初稿由夏强副教授执笔，刘金山教授负责全书的统稿和定稿.

由于编者水平有限，书中难免有缺点和错误之处，敬请读者批评指正.

编　者

2018 年 2 月

目　录

第一章　随机事件及其概率

在自然界和人类社会活动中，人们所观察的现象大致上可分为两类. 一类是事先可以预知结果的现象，即在一定条件下，某一确定的现象必然会出现，或根据它过去的状态，完全可以预知它将来的发展状态. 我们称这一类现象为**确定性现象**或必然现象. 例如，在一个标准大气压下，水加热到 100℃ 时必然沸腾. 另一类是事先不能预知结果的现象，即在相同条件下重复进行试验或观测时，每次出现的结果未必相同，或者即使知道它过去的状态，也不能完全确定它将来的发展状态. 我们称这一类现象为**随机现象**或偶然现象. 例如，多次抛掷一枚骰子，朝上一面出现的点数可能是 1，2，…，6 中的任何一个，但在每次抛掷前不能预知出现的点数到底是几. 这类现象的共同特点是：在相同条件下重复进行试验或观测，其结果不止一个，在每次试验之前不能预知该次试验的确切结果.

对于随机现象，人们通过大量的实践发现，在相同的条件下，虽然试验结果在一次试验或观察中到底出现哪个是不确定的，但在大量重复试验中却能呈现出某种规律性，这种规律性称为统计规律性. 例如，多次抛掷一枚均匀的硬币时，带国徽的一面朝上的次数约占总抛掷次数的一半.

概率论与数理统计就是以数量化方法研究随机现象统计规律的学科. 概率论是研究随机现象的模型，即概率分布；数理统计是研究随机现象的数据分析与处理方法. 概率论与数理统计不仅研究能大量重复的随机现象，而且也研究不能重复的随机现象，例如某些经济现象(如经济增长速度、金融产品收益率、股票价格等).

1.1　基本概念

1.1.1　随机试验与事件

在概率论与数理统计中，“试验”是一个广泛的术语. 我们把在一定条件下对某种现象的一次观察、测量或进行一次科学实验，统称为一个试验. 一般称满足下面两个条件的试验为**随机试验**：

(1) 在相同条件下可以重复进行；

(2) 每次试验结果事先不可预知，但所有可能的试验结果事先知道.

一般用字母 E 表示随机试验. 下面是一些随机试验的例子.

E_1：抛掷一颗骰子，观察出现的点数；

E_2：将一枚硬币连续抛掷两次，观察其正反面出现的情况；

E_3：将一枚硬币连续抛掷两次，观察其正面出现的次数；

E_4：观察一天内进入某个超市的顾客人数；

E_5：观察某型号电视机的使用寿命 t；

E_6：记录某地区一昼夜的最低气温 x 和最高气温 y.

对一个随机试验，我们把所有可能的试验结果组成的集合称为该试验的**样本空间**，记为 Ω. 样本空间中的每个元素称为**样本点**. 在上述 6 个试验中，若以 Ω_i 表示试验 E_i 的样本空间，则

$\Omega_1 = \{1, 2, 3, 4, 5, 6\}$；

$\Omega_2 = \{HH, HT, TH, TT\}$，其中 H 表示正面，T 表示反面；

$\Omega_3 = \{0, 1, 2\}$；

$\Omega_4 = \{0, 1, 2, \cdots\}$；

$\Omega_5 = \{t \mid 0 \leqslant t < \infty\}$；

$\Omega_6 = \{(x, y) \mid T_0 \leqslant x \leqslant y \leqslant T_1\}$，其中 T_0 和 T_1 分别表示这一地区的最低气温和最高气温.

对于样本空间应注意下面几点：

(1) 样本空间是一个集合，它由样本点组成，可以用列举法或描述法来表示；

(2) 在样本空间中，样本点可以是一维的，也可以是多维的，样本点个数可以是有限个，也可以是无限的；

(3) 对于一个随机试验而言，试验的目的不同，样本空间往往也不同. 例如，E_2 和 E_3 虽然都是将一枚硬币抛掷两次，但由于试验目的不同，因此样本空间不同，E_2 的样本空间为 $\Omega_2 = \{HH, HT, TH, TT\}$，$E_3$ 的样本空间为 $\Omega_3 = \{0, 1, 2\}$.

我们把样本空间的任一个子集称为一个**随机事件**，简称为事件，常用大写字母 A，B，C，… 表示. 因此，随机事件就是随机试验的某些结果(样本点) 组成的集合. 特别地，由一个样本点组成的单点集合称为**基本事件**. 在一个试验中，事件 A 发生当且仅当 A 中某个样本点出现，这就是事件 A 发生的含义.

例 1.1.1　在抛掷一颗骰子的试验中，若用 A 表示“出现偶数点”，B 表示“出现奇数点”，C 表示“出现 3 点或 3 点以上”. 假设试验的目的是观察出现的点数，试写出样本空间，并用样本点表示事件 A，B，C.

解　该试验有 6 个可能的结果，样本空间为 $\Omega = \{1, 2, 3, 4, 5, 6\}$，事件 A，B，C 分别表示为 $A = \{2, 4, 6\}$，$B = \{1, 3, 5\}$，$C = \{3, 4, 5, 6\}$.

例 1.1.2　从一批计算机中任取一台，观察其无故障运行的时间 T(单位：小时). A 为事件“恰好运行 240 小时”，B 为事件“运行 240.2 小时以上”，C 为事件“运行时间大于 270.5 小时，小于等于 480.7 小时”. 试写出样本空间，并用样本点子集表示事件 A，B，C.

解　样本空间为 $\Omega = \{T \mid T \geqslant 0\}$，事件 A，B，C 分别表示为 $A = \{T = 240\}$，$B = \{T \mid T > 240.2\}$，$C = \{T \mid 270.5 < T \leqslant 480.7\}$.

样本空间 Ω 是其自身的一个子集，因此它也是一个事件，由于它包含所有样本点，所以每次试验它必然发生，因此 Ω 表示必然事件. 空集 $\varnothing$ 不包含任何样本点，每次试验它都不会发生，故 $\varnothing$ 表示不可能事件. 虽然它们不是真正的随机事件，但为了研究问题的方便，我们把它们视为特殊的随机事件.

1.1.2　事件的关系与运算

因为事件是集合，即样本空间的子集，所以事件之间的关系和运算可以按照集合之间

的关系和运算来处理. 根据“事件发生”的含义，我们不难给出事件的关系与运算的定义和规则.

设 Ω 是样本空间，A，B，C 及 A_1，A_2，… 都是事件，即 Ω 的子集，它们有以下关系.

1. 包含关系

若 A 的发生必然导致 B 的发生，则称 B 包含 A 或 A 是 B 的子事件，记为 $B \supset A$ 或者 $A \subset B$，即 A 的元素全属于 B，如图 1.1 所示.

2. 相等关系

若 $A \subset B$ 且 $B \subset A$，则称 A 与 B 相等，记为 $A = B$.

3. 事件的和

对两个事件 A 和 B，定义事件

$$C = \{A \text{ 发生或 } B \text{ 发生}\},$$

称其为 A 与 B 的和事件，记为 $C = A \cup B$. 事件 $A \cup B$ 发生，即 A 发生或 B 发生，意味着 A 与 B 至少有一个发生，如图 1.2 所示.

和事件可以推广到多个事件的情形. 设有 n 个事件 A_1，A_2，…，A_n，定义它们的和事件为

$$C = \bigcup_{k=1}^{n} A_k = \{A_1,\ A_2,\ \cdots,\ A_n \text{ 中至少有一个发生}\}.$$

对无穷多个事件 A_1，A_2，…，A_n，…，可以类似地定义它们的和事件为

$$C = \bigcup_{k=1}^{\infty} A_k = \{A_1,\ A_2,\ \cdots,\ A_n,\ \cdots \text{ 中至少有一个发生}\}.$$

4. 事件的积

对两个事件 A 和 B，定义事件

$$C = \{A \text{ 发生且 } B \text{ 发生}\},$$

称其为 A 与 B 的积事件，记为 $C = A \cap B$（或 $C = AB$）. 事件 AB 发生意味着 A 发生且 B 发生，即 A 与 B 同时发生，如图 1.3 所示.

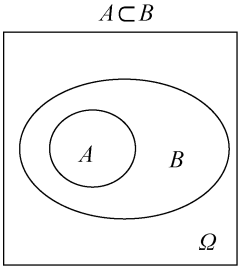


图 1.1 A 包含于 B

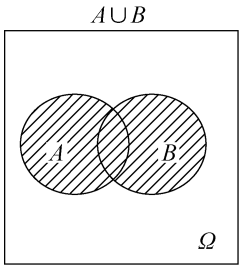


图 1.2 和事件

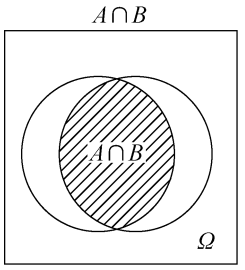


图 1.3 积事件

类似地，可以定义多个事件 A_1，A_2，…，A_n，… 的积事件. 根据事件的个数为有限和无限情况分别有下列积事件：

$$C = \bigcap_{k=1}^{n} A_k = \{A_1,\ A_2,\ \cdots,\ A_n \text{ 同时发生}\},$$

$$C = \bigcap_{k=1}^{\infty} A_k = \{A_1,\ A_2,\ \cdots,\ A_n,\ \cdots \text{ 同时发生}\}.$$

5. 事件的差

对两个事件 A 和 B，定义事件

$$C = \{A \text{ 发生且 } B \text{ 不发生}\},$$

称其为 A 与 B 的差事件，记为 $C = A-B$（或 $C = A \setminus B$），即 A 发生但 B 不发生的事件，如图 1.4 所示. 容易知道 $A-B = A-AB$.

6. 互斥事件

若两个事件 A 与 B 不能同时发生，即 $AB = \varnothing$，则称 A 与 B 是互斥事件，或称它们互不相容，如图 1.5 所示. 若事件 A_1，A_2，…，A_n 中任意两个都互斥，则称事件组 A_1，A_2，…，A_n 两两互斥.

当事件 A 与 B 互斥时，可记它们的和事件 $A \cup B$ 为 $A+B$；对于两两互斥的多个事件的和事件有类似的记法.

7. 对立事件

“A 不发生”的事件称为 A 的对立事件，记为 $\overline{A}$，即 $\overline{A} = \Omega - A$，如图 1.6 所示，并称 A 与 $\overline{A}$ 为互逆事件，它们是互为对立的事件，满足 $A \cup \overline{A} = \Omega$，$A\overline{A} = \varnothing$，$\overline{\overline{A}} = A$.

例如，在抛掷硬币的试验中，设 A 为“出现正面”，B 为“出现反面”，则 A 与 B 互斥且 A 与 B 互为对立；在掷骰子的试验中，设 A 为“出现 1 点”，B 为“出现 3 点以上”，则 A 与 B 互斥，但 A 与 B 不是对立事件.

利用事件对立关系容易知道，对于任意事件 A 和 B，$A-B = A\overline{B}$.

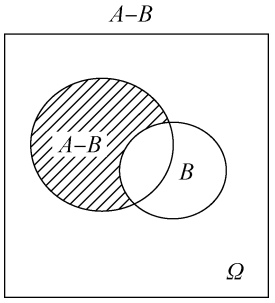


图 1.4　差事件

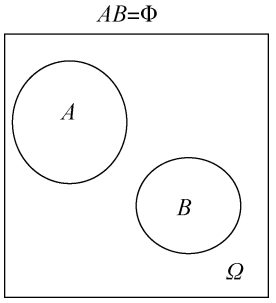


图 1.5　互斥事件

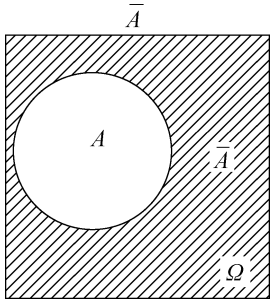


图 1.6　对立事件

设 A，B，C 为事件，根据集合的运算规则，有以下事件运算规则.

1. 交换律：$A \cup B = B \cup A$；$AB = BA$.

2. 结合律：$(A \cup B) \cup C = A \cup (B \cup C)$；$(AB)C = A(BC)$.

3. 分配律：$A(B \cup C) = (AB) \cup (AC)$；$A \cup (BC) = (A \cup B)(A \cup C)$.

4. 对偶律：$\overline{A \cup B} = \overline{A}\,\overline{B}$；$\overline{AB} = \overline{A} \cup \overline{B}$.

对于多个事件情况，上述运算规则仍然成立. 例如：

$A(A_1 \cup A_2 \cup \cdots \cup A_n) = (AA_1) \cup (AA_2) \cup \cdots \cup (AA_n)$；

$A \cup (A_1A_2\cdots A_n) = (A \cup A_1)(A \cup A_2)\cdots(A \cup A_n)$；

$\overline{A_1 \cup A_2 \cup \cdots \cup A_n} = \overline{A}_1\overline{A}_2\cdots\overline{A}_n$；

$\overline{A_1A_2\cdots A_n} = \overline{A}_1 \cup \overline{A}_2 \cup \cdots \cup \overline{A}_n$.

上述运算规则也可以推广到无穷多个事件的情况.

例 1.1.3 向指定目标连续射击三次，观察击中目标的情况. 分别用 A_1，A_2，A_3 表示事件“第一、二、三次射击时击中目标”，试用 A_1，A_2，A_3 表示以下各事件.

（1）只第一次击中；

（2）只击中一次；

（3）三次都未击中；

（4）至少击中一次.

解 （1）事件“只第一次击中”，意味着第二次和第三次都不中. 所以，该事件可表示为 $A_1\overline{A}_2\overline{A}_3$.

（2）事件“只击中一次”，并不指定哪一次击中，意味着三个事件“只第一次击中”、“只第二次击中”和“只第三次击中”至少有一个发生，即它们的和事件发生. 由于上述三个事件两两互斥，所以，该事件可表示为 $A_1\overline{A}_2\overline{A}_3+\overline{A}_1A_2\overline{A}_3+\overline{A}_1\overline{A}_2A_3$.

（3）事件“三次都未击中”，就是事件“第一、二、三次都未击中”，该事件可表示为 $\overline{A}_1\overline{A}_2\overline{A}_3$ 或 $\overline{A_1 \cup A_2 \cup A_3}$.

（4）事件“至少击中一次”，就是事件“第一、二、三次射击中至少有一次击中”，所以，该事件可表示为 $A_1 \cup A_2 \cup A_3$.

1.2 事件的概率

除必然事件和不可能事件外，任何随机事件在一次试验中可能发生，也可能不发生. 我们常常希望知道某事件在一次试验中发生的可能性大小. 例如，知道了某批种子的发芽率就可以科学合理地安排播种；知道了某种流行疾病的传播规律，就可以提前进行预防和控制；知道了某食品在某段时间内变质的可能性大小，就可以合理地制定该食品的保质期，等等. 为了合理地刻画事件在一次试验中发生的可能性大小，我们首先引入频率的概念，然后根据频率的性质定义事件发生的概率，并讨论概率的基本性质.

1.2.1 频率及概率的统计定义

1. 事件的频率

定义 1.2.1 在相同条件下，重复进行 n 次试验，事件 A 发生的次数 n_A 称为事件 A 发生的频数，比值 $\dfrac{n_A}{n}$ 称为事件 A 发生的频率，记为 $f_n(A)$，即 $f_n(A)=\dfrac{n_A}{n}$.

由定义 1.2.1 不难发现，频率满足下列三条性质：

（1）非负性：对任意事件 A，$f_n(A)\geqslant 0$；

（2）规范性：$f_n(\Omega)=1$；

（3）有限可加性：若 A_1，A_2，…，A_k 为两两互斥事件，则

$$f_n\left(\bigcup_{i=1}^{k}A_i\right)=\sum_{i=1}^{k}f_n(A_i).$$

例 1.2.1 考虑某种子发芽率试验. 从一大批种子中抽取 7 批种子做试验，其结果如

表 1.1 所示.

本例中，将观察一粒种子是否发芽视为一次试验，若种子发芽，则事件 A 发生. 从表 1.1 中不难发现，事件 A 在 n 次试验中发生的频率 $f_n(A)$ 具有随机波动性. 当 n 较小时，随机波动的幅度较大；当 n 较大时，随机波动的幅度较小. 随着 n 的逐渐增大，$f_n(A)$ 逐渐稳定于固定值 0.9 附近.

表 1.1　种子发芽率试验数据

种子粒数	10	70	310	700	1500	2000	3000
发芽粒数	10	60	282	639	1339	1806	2706
发芽率	1.0	0.857	0.910	0.913	0.893	0.903	0.902

例 1.2.2　在英文中某些字母出现的频率远远高于其他字母. 表 1.2 所示是一份英文字母出现频率的稳定值统计表.

表 1.2　英文字母出现频率数据

字母	E	T	A	O	I	N	S	R	T
频率	0.1268	0.0978	0.0788	0.0776	0.0707	0.0706	0.0634	0.0594	0.0573
字母	L	D	U	C	F	M	W	Y	G
频率	0.0394	0.0389	0.0280	0.0268	0.0256	0.0244	0.0214	0.0202	0.0187
字母	P	B	V	K	X	J	Q	Z	
频率	0.0186	0.0156	0.0102	0.0060	0.0016	0.0010	0.0009	0.0006	

字母使用频率的研究对于打字机键盘的设计(方便的地方安排使用频率较高的字母键)、信息编码(常用字母使用较短的编码表示)和密码的破译等方面有重要意义.

人们在长期的实践中观察到，随机事件 A 出现的频率 $f_n(A)$ 有如下特点：当试验次数 n 较小时，频率 $f_n(A)$ 在 0 到 1 之间波动较大；当试验次数 n 增大时，频率 $f_n(A)$ 逐渐接近于某一个常数. 这种特性称为频率的稳定性，也就是通常所说的统计规律性. 因此，用频率的稳定值来刻画事件 A 发生的可能性大小是合适的. 实践中，人们常常让试验重复大量次数，计算频率 $f_n(A)$，用它来表征事件 A 发生的概率. 这个概率就是统计定义下的概率.

2. 概率的统计定义

定义 1.2.2　在相同条件下，重复进行 n 次试验，当试验次数 n 增大时，如果某事件 A 发生的频率 $f_n(A)$ 在区间 $[0, 1]$ 上的某一稳定值 p 附近摆动，且随着试验次数 n 的增大，摆动的幅度越来越小，则称数值 p 为事件 A 发生的概率，记为 $P(A) = p$.

概率的统计定义是描述性的，它一方面肯定了随机事件的概率是存在的，另一方面它提供了概率的一个具体估计值，即当试验重复次数 n 充分大时，可以用频率的稳定值作为概率的估计值，这一点是频率方法最有价值的地方. 但其不足之处是需要进行大量的重复试验. 在实际中，要把一个试验无限次地重复下去，往往是不经济的或不现实的. 因此，概率的统计定义有一定的局限性.

1.2.2 概率的定义和性质

在历史上，为了更好地刻画随机事件的概率，人们提出了多种概率定义，其中著名数学家柯尔莫哥洛夫于1933年提出的概率公理化定义最为成熟. 这个定义概括了历史上几种概率定义的共有特性，即不管何种随机现象，只要某函数(以事件A为自变量)满足定义中的三条公理，就称它为概率. 这个定义给予了概率论严格的数学基础，并使得概率论的研究方法和结果能用于其他科学领域. 这一公理化体系迅速得到举世公认，是概率论发展史上的一个里程碑. 有了这个公理化定义后，概率论得到了跨越性的发展.

定义 1.2.3 设随机试验E的样本空间为Ω，其事件域为F，对每个事件$A \subset F$，定义一个实数$P(A)$与之对应. 称集合函数$P(A)$为事件A的概率，如果它满足下列三条公理，

公理 1(非负性)：对任意事件A，$P(A) \geqslant 0$；

公理 2(规范性)：对必然事件Ω，$P(\Omega) = 1$；

公理 3(可数可加性)：对任意可数个两两互斥的事件A_1，A_2，$\cdots$，A_n，$\cdots$，$P(\bigcup\limits_{i=1}^{\infty} A_i) = \sum\limits_{i=1}^{\infty} P(A_i)$.

由概率的上述定义，可以推得概率的一些重要性质.

性质 1 $P(\varnothing) = 0$.

证明 因为$\varnothing = \varnothing \cup \varnothing \cup \cdots \cup \varnothing \cup \cdots$，且不可能事件之间两两互斥，则由公理3有

$$P(\varnothing) = P(\varnothing) + \cdots + P(\varnothing) + \cdots,$$

又由公理1有$P(\varnothing) \geqslant 0$，因此有$P(\varnothing) \geqslant 2P(\varnothing)$，$P(\varnothing) \leqslant 0$，故必有$P(\varnothing) = 0$.

性质 2 若A_1，A_2，$\cdots$，A_n两两互斥，则$P(\bigcup\limits_{i=1}^{n} A_i) = \sum\limits_{i=1}^{n} P(A_i)$.

证明 因为$\bigcup\limits_{i=1}^{n} A_i = \bigcup\limits_{i=1}^{n} A_i \cup \varnothing \cup \varnothing \cup \cdots$，利用公理3和性质1有

$$P(\bigcup_{i=1}^{n} A_i) = P(\bigcup_{i=1}^{n} A_i \cup \varnothing \cup \cdots) = P(A_1) + \cdots + P(A_n) + P(\varnothing) + \cdots = \sum_{i=1}^{n} P(A_i).$$

性质 3 对任意事件A，有$P(A) = 1 - P(\overline{A})$.

证明 因为$A \cup \overline{A} = \Omega$，$A\overline{A} = \varnothing$，所以$1 = P(\Omega) = P(A \cup \overline{A}) = P(A) + P(\overline{A})$，移项即得$P(A) = 1 - P(\overline{A})$.

有些事件直接计算其概率较为困难，但可能其对立事件的概率相对比较容易计算. 对此类事件就可以利用性质3计算其概率.

性质 4 对任意事件A，B，有$P(A-B) = P(A) - P(AB)$. 特别地，若$B \subset A$，则$P(A-B) = P(A) - P(B)$，且$P(B) \leqslant P(A)$.

证明 因为$A = (A-B) \cup AB$且$(A-B) \cap AB = \varnothing$，所以由性质2有

$$P(A) = P((A-B) \cup AB) = P(A-B) + P(AB),$$

移项即得$P(A-B) = P(A) - P(AB)$. 特别地，若$B \subset A$，则$P(AB) = P(B)$，即有$P(A-B) = P(A) - P(B)$，又由于$P(A-B) \geqslant 0$，所以$P(B) \leqslant P(A)$.

性质 5 对任意事件A，$P(A) \leqslant 1$.

证明　因 $A \subset \Omega$，由性质 4 即得 $P(A) \leqslant P(\Omega) = 1$.

性质 6　对任意事件 A，B，有 $P(A \cup B) = P(A) + P(B) - P(AB)$.

证明　因为 $A \cup B = (A - B) \cup B$，且 $(A - B)B = \varnothing$，由性质 2 和性质 4 有

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

性质 6 称为概率的广义加法公式，该性质可以推广到多个事件. 设 A_1，A_2，$\cdots$，A_n 是任意 n 个事件，则有

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \leqslant i < j \leqslant n} P(A_i A_j) + \sum_{1 \leqslant i < j < k \leqslant n} P(A_i A_j A_k) + \cdots + (-1)^{n+1} P(A_1 A_2 \cdots A_n).$$

特别地，对任意事件 A，B，C，有

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC).$$

例 1.2.3　小王参加“智力大冲浪”游戏，他能答出甲、乙两类问题的概率分别为 0.7 和 0.5，两类问题都能答出的概率为 0.3. 求

(1) 答出甲类而答不出乙类问题的概率；

(2) 至少有一类问题能答出的概率；

(3) 两类问题都答不出的概率；

(4) 至少有一类问题答不出的概率.

解　设事件 A，B 分别表示他“能答出甲类问题”和“能答出乙类问题”，则 $P(A) = 0.7$，$P(B) = 0.5$，$P(AB) = 0.3$. 所求概率分别为

(1) $P(A\overline{B}) = P(A - AB) = P(A) - P(AB) = 0.7 - 0.3 = 0.4$；

(2) $P(A \cup B) = P(A) + P(B) - P(AB) = 0.7 + 0.5 - 0.3 = 0.9$；

(3) $P(\overline{A}\,\overline{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0.9 = 0.1$；

(4) $P(\overline{A} \cup \overline{B}) = P(\overline{AB}) = 1 - P(AB) = 1 - 0.3 = 0.7$.

需要指出的是，上例中 $P(AB) = 0.3$，$P(AB) \neq P(A)P(B) = 0.7 \times 0.5 = 0.35$. 一般来说 $P(AB) = P(A)P(B)$ 不一定成立，只有当 A，B 相互独立(见 1.5 节) 时该等式才成立.

例 1.2.4　已知 $P(A) = P(B) = P(C) = \dfrac{1}{4}$，$P(AB) = 0$，$P(AC) = P(BC) = \dfrac{1}{9}$. 求事件 A，B，C 全不发生的概率.

解　由于 $ABC \subset AB$，可知 $P(ABC) \leqslant P(AB) = 0$，因此 $P(ABC) = 0$. 所求概率为

$$\begin{aligned} P(\overline{A}\,\overline{B}\,\overline{C}) &= P(\overline{A \cup B \cup C}) = 1 - P(A \cup B \cup C) \\ &= 1 - [P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)] \\ &= 1 - \left(\frac{1}{4} + \frac{1}{4} + \frac{1}{4} - 0 - \frac{1}{9} - \frac{1}{9} + 0\right) = \frac{17}{36}. \end{aligned}$$

1.3　古典概率模型

古典概率模型是概率论早期研究的主要对象，比较直观. 古典概率的计算是概率论中

最重要的内容之一. 实际上，在应用中有大量的问题需要用古典概率计算方法来解决，而在理论物理等学科研究中也需要用到古典概率方法.

1.3.1 古典概型

若一个随机试验满足

(1) 样本空间中只有有限个样本点(有限性)，

(2) 每个样本点出现的可能性相等(等可能性)，

则称该随机试验为古典型随机试验，称该概率模型为古典概型或等可能概型.

设古典概型的样本空间为 $\Omega=\{\omega_1,\omega_2,\cdots,\omega_n\}$，由于每个样本点出现的可能性相等，可知有

$$P(\{\omega_1\})=P(\{\omega_2\})=\cdots=P(\{\omega_n\}).$$

由于基本事件两两互斥，且 $\{\omega_1\}\cup\{\omega_2\}\cup\cdots\cup\{\omega_n\}=\Omega$，可得

$$\begin{aligned}1&=P(\Omega)=P(\{\omega_1\}\cup\{\omega_2\}\cup\cdots\cup\{\omega_n\})\\&=P(\{\omega_1\})+P(\{\omega_2\})+\cdots+P(\{\omega_n\})\\&=nP(\{\omega_1\}),\end{aligned}$$

于是

$$P(\{\omega_1\})=P(\{\omega_2\})=\cdots=P(\{\omega_n\})=\frac{1}{n}.$$

若事件 A 中含有 $k(k\leqslant n)$ 个基本事件，记为 $A=\{\omega_{i_1}\}\cup\{\omega_{i_2}\}\cup\cdots\cup\{\omega_{i_k}\}$，则由概率性质 2 可得

$$P(A)=\sum_{j=1}^{k}P(\{\omega_{i_j}\})=\frac{k}{n}=\frac{\text{事件 }A\text{ 包含的基本事件数}}{\Omega\text{ 包含的基本事件总数}}\tag{1.3.1}$$

该式是计算古典概型下事件概率的基本公式.

例 1.3.1 从标号为 1，2，…，10 的 10 个同样大小的球中任取一个，分别求事件 A，B，C 的概率，这里 $A=\{$取到 2 号$\}$，$B=\{$取到奇数号$\}$，$C=\{$取到的号数不小于 7$\}$.

解 显然，样本空间为 $\Omega=\{1,2,3,\cdots,10\}$，基本事件总数为 10，事件 A，B，C 包含的基本事件数分别为 1、5、4 个，它们的概率为

$$P(A)=\frac{1}{10},\quad P(B)=\frac{5}{10}=\frac{1}{2},\quad P(C)=\frac{4}{10}=\frac{2}{5}.$$

例 1.3.2 从 6 双不同的鞋子中任取 4 只，求(1) 其中恰有 2 只成双的概率；(2) 至少有 2 只成双的概率.

解 (1) 设 A 表示事件“恰有 2 只成双”. 该事件可以按如下方式完成：先从 6 双鞋子中抽取 1 双，2 只全取，再从剩下的 5 双中任取 2 双，每双中各取 1 只. 因此事件 A 所含的样本点个数为 $C_6^1C_2^2C_5^2C_2^1C_2^1$. 所以

$$P(A)=\frac{C_6^1C_2^2C_5^2C_2^1C_2^1}{C_{12}^4}=\frac{16}{33}.$$

(2) 该问题采用求对立事件概率的方法比较简单. 设 B 为事件“至少有 2 只成双”，则 $\overline{B}$ 为事件“任 2 只鞋子都不能配对”，于是有

$$P(B)=1-P(\overline{B})=1-\frac{C_6^4C_2^1C_2^1C_2^1C_2^1}{C_{12}^4}=\frac{17}{33}.$$

在例1.3.2中，不能把事件B所含的样本点数计为$C_6^1C_2^2C_{10}^2$，即先从6双鞋子中抽取1双，2只全取，再从剩下的鞋子中任取2只. 这是因为，若设每双鞋子标有号码1，2，…，6，则当先取到第i双鞋子的2只时，后取的2只可能恰好为第j双鞋子的2只，即恰好取到第i双和第j双，而同时当先取到第j双鞋子的2只时，后取的2只可能恰好为第i双鞋子的2只，也是取到第i双和第j双，这与前者重复. 我们也可以利用求互斥事件的和事件概率的方法直接求事件B的概率，即

$$P(B)=P(A)+P(C)=\frac{C_6^2C_2^2C_5^2C_2^1C_2^1}{C_{12}^4}+\frac{C_6^2C_2^2C_2^2}{C_{12}^4}=\frac{17}{33},$$

其中C表示事件“恰好取到2双”.

例1.3.3(盒子模型)　设有n个球，每个球都等可能地落入N个盒子中的一个，假设$n\leqslant N$. 求下列事件的概率.

A：某指定的n个盒子中各落入一球；

B：恰有n个盒子各落入一球；

C：某个指定的盒子中落入m个球；

D：恰好$n-1$个盒子里有球.

解　由于每个球都等可能地落入N个盒子中的一个，按照乘法原理，n个球共有N^n种落法. 把每种落法作为一个基本事件，这是一个古典概型问题，基本事件总数为N^n.

按照乘法原理，事件A包含的基本事件数是$n!$，故

$$P(A)=\frac{n!}{N^n}.$$

对事件B，从N个盒子中任选n个，有C_N^n种选法；选定n个盒子后，每个盒子各落入一球的方法为$n!$种. 因此事件B包含的基本事件数是$C_N^n n!=P_N^n$，故

$$P(B)=\frac{P_N^n}{N^n}=\frac{N(N-1)\cdots(N-n+1)}{N^n}.$$

事件B的另一种分析方法是：n个球落入n个盒子中，每个盒子恰好落入一个球，则第1个球有N种落法，第2个球有$N-1$种落法，……，第n个球有$N-n+1$种落法. 根据乘法原理，共有$N(N-1)\cdots(N-n+1)$种落法，由此即得B的上述概率.

对事件C，m个球可以在n个球中任选，共有C_n^m种选法. 其余$n-m$个球可以任意落入另外的$N-1$个盒子中，共有$(N-1)^{n-m}$种落法. 所以，事件C包含的基本事件个数是$C_n^m\cdot(N-1)^{n-m}$. 故

$$P(C)=\frac{C_n^m\cdot(N-1)^{n-m}}{N^n}=C_n^m\left(\frac{1}{N}\right)^m\left(1-\frac{1}{N}\right)^{n-m}.$$

对事件D，注意到“$n-1$个盒子里有球”，意味着其中一个盒子中恰有2个球，其余的$n-2$个盒子中各有一个球. 可先任取落入2个球的一个盒子，有N种取法，再任取$n-2$个盒子，有C_{N-1}^{n-2}种取法，然后将球落进去，落法有$C_n^2\cdot(n-2)!=\frac{n!}{2!}$种，故

$$P(D)=\frac{NC_{N-1}^{n-2}\dfrac{n!}{2!}}{N^n}=\frac{n!\ C_{N-1}^{n-2}}{2N^{n-1}}.$$

下面我们用盒子模型讨论“生日问题”，即 p_n 个人中至少有两个人生日相同的概率是多少？

若把 n 个人看成是 n 个球，将一年365天看成是 $N=365$ 个盒子，则“n 个人的生日全不相同”就相当于有 $n(n\leqslant N)$ 个盒子各有一球，所以由例1.3.3中事件 B 的概率可知，p_n 个人中至少有两个人生日相同的概率是

$$p_n=1-\frac{365(365-1)\cdots(365-n+1)}{365^n}.$$

经计算，可得表1.3中的概率.

表1.3　n 个人中至少有两个人生日相同的概率

n	10	20	30	40	50	60
p_n	0.1169482	0.4114384	0.7063162	0.8912318	0.9703736	0.9941227

这个数值结果是相当惊人的，说明30个以上人中至少有两个人同一天生的概率在0.7以上，60个人中至少有两个人同一天生的概率几乎等于1.

例1.3.4　某公司生产的15件产品中，有12件正品，3件次品. 现将它们随机地分装在3个箱中，每箱5件. 记 A 为事件“每箱中恰有1件次品”，B 为事件“3件次品都在同一箱中”. 试求概率 $P(A)$ 和 $P(B)$.

解　由排列组合公式，将15件产品装入3个箱中，每箱5件，共有 $\dfrac{15!}{5!\ 5!\ 5!}$ 种装法，把每种装法作为一个基本事件，这是一个古典概型问题，基本事件总数为 $\dfrac{15!}{5!\ 5!\ 5!}$.

把3件次品分别装入3个箱中，共有3！种装法. 对于每一种这样的装法，把其余12件正品平均装入3个箱中，共有 $\dfrac{12!}{4!\ 4!\ 4!}$ 种装法. 因此事件 A 的概率为

$$P(A)=3!\ \frac{12!}{4!\ 4!\ 4!}\Big/\frac{15!}{5!\ 5!\ 5!}=\frac{25}{91}.$$

把3件次品装入1个箱中，共有3种装法. 对于每一种这样的装法，把其余12件正品装入3个箱中，其中1箱装2件，其余2箱各装5件，共有 $\dfrac{12!}{2!\ 5!\ 5!}$ 种装法. 因此事件 B 的概率为

$$P(B)=3\,\frac{12!}{2!\ 5!\ 5!}\Big/\frac{15!}{5!\ 5!\ 5!}=\frac{6}{91}.$$

例1.3.5　设有一箱产品共有100件，其中有4件次品，其余均为正品. 求

(1) 从箱中任取3件，取到的全是正品的概率.

(2) 从箱中任取3件，取到恰有2件正品的概率.

解　设 A 为事件“3件全为正品”；B 为事件“恰有2件正品”. 则

（1）由于抽样是从箱中任取 3 件，故此时样本空间 Ω 的样本点数为 C_{100}^3，而 A 包含的样本点数为 C_{96}^3，故

$$P(A)=\frac{C_{96}^3}{C_{100}^3}=0.883\ 611\ 6$$

（2）该样本空间 Ω 的样本点数仍为 C_{100}^3，B 包含的样本点数为 $C_{96}^2C_4^1$，故

$$P(B)=\frac{C_{96}^2C_4^1}{C_{100}^3}=0.112\ 801\ 5$$

注：从例 1.3.5(2) 的解法中我们可以归纳出更一般的抽样模型. 设一箱子中共有 N 只球，其中有 M 只红球，有 $N-M$ 只白球，若从该箱子中任取 n 只球，则其中恰有 m 只红球的概率(设 A_m 表示恰有 m 只红球的事件) 为

$$P(A_m)=\frac{C_M^mC_{N-M}^{n-m}}{C_N^n}\quad m=0,\ 1,\ 2,\ \cdots,\ r,\ r=\min(n,\ M)\tag{1.3.2}$$

上述概率称为**超几何概率**. 在“不放回抽样”问题中经常需要用到该概率公式.

上述超几何概率还可以进一步推广. 设一箱子中共有 N 只球，其中有 N_1 只红球、N_2 只黄球、N_3 只白球($N=N_1+N_2+N_3$)，若从该箱子中任取 n 只球，则 n 只球中恰有 n_1 只红球、n_2 只黄球、n_3 只白球($n=n_1+n_2+n_3$) 的概率为

$$P(A)=\frac{C_{N_1}^{n_1}C_{N_2}^{n_2}C_{N_3}^{n_3}}{C_N^n}.\tag{1.3.3}$$

例 1.3.6(彩票问题)　一种福利彩票称为幸运 35 选 7，即购买时从 01 ~ 35 这些数字中任选 7 个号码，开奖时从 01，02，…，35 中不重复地取 7 个基本号码和 1 个特殊号码. 中各等奖的规则如下.

若 7 个基本号码全中，则得一等奖；若中 6 个基本号码和特殊号码，则中二等奖；若中 6 个基本号码，则中三等奖；若中 5 个基本号码和特殊号码，则中四等奖；若中 5 个基本号码，则中五等奖；若中 4 个基本号码和特殊号码，则中六等奖；若中 4 个基本号码，或中 3 个基本号码和特殊号码，则中七等奖.

下面求各等奖的中奖概率. 因为不重复地选取号码是一种不放回抽样，按照几何概率计算方法，样本空间 Ω 中含有 C_{35}^7 个样本点. 我们把各等奖的抽取看成是从三类号码中抽取：第一类为 7 个基本号码，第二类为 1 个特殊号码，第三类为其余 27 个无用号码. 记 p_i 为第 i 等奖的中奖概率，则由式(1.3.3) 可计算出各等奖的中奖概率如下.

$$p_1=\frac{C_7^7C_1^0C_{27}^0}{C_{35}^7}=0.149\times10^{-6},\ p_2=\frac{C_7^6C_1^1C_{27}^0}{C_{35}^7}=1.04\times10^{-6},$$

$$p_3=\frac{C_7^6C_1^0C_{27}^1}{C_{35}^7}=28.106\times10^{-6},\ p_4=\frac{C_7^5C_1^1C_{27}^1}{C_{35}^7}=84.318\times10^{-6},$$

$$p_5=\frac{C_7^5C_1^0C_{27}^2}{C_{35}^7}=1.096\times10^{-3},\ p_6=\frac{C_7^4C_1^1C_{27}^2}{C_{35}^7}=1.827\times10^{-3},$$

$$p_7=\frac{C_7^4C_1^0C_{27}^3}{C_{35}^7}+\frac{C_7^3C_1^1C_{27}^3}{C_{35}^7}=30.448\times10^{-3}.$$

若记 A 为事件“中奖”，$\bar{A}$ 为事件“不中奖”，由上述概率可以得到

$$P(A)=p_1+p_2+\cdots+p_7=0.033\ 485,\quad P(\bar{A})=1-P(A)=0.966\ 515.$$

这说明，一百个人中约有 3 人中奖，而中头等奖的概率只有 0.149×10^{-6}，即两千万个人中约有 3 个人中头等奖. 未中奖者占绝大多数，即大约有 96.65% 的购彩票者不中奖.

1.3.2　几何概型

古典概型考虑的是有限等可能结果的随机试验的概率模型. 现在我们考虑样本空间为一线段、平面区域或空间立体的等可能随机试验的概率模型，称为几何概型.

如果一个试验具有以下两个特点：

(1) 样本空间 Ω 是一个可以度量的几何区域(如线段、平面、立体)，其度量(长度、面积、体积) 记为 $\mu(\Omega)$.

(2) 向区域 Ω 上随机投掷一点，这里“随机投掷一点” 的含义是指该点落入 Ω 内任一部分区域 A 的可能性只与区域 A 的度量 $\mu(A)$ 成比例， 而与区域 A 的位置和形状无关.

那么，事件 A 的概率由下式计算，即

$$P(A)=\frac{\mu(A)}{\mu(\Omega)} \tag{1.3.4}$$

例 1.3.7　在一个均匀陀螺的圆周上均匀地刻上 $(0,\ 4]$ 上的所有实数，在一桌面上旋转该陀螺，求陀螺停下后，圆周与桌面的接触点位于区间 $[0.5,\ 1]$ 的概率.

解　由于陀螺及刻度的均匀性，它停下来时其圆周上的各点与桌面接触的可能性相等. 根据题意，这是一个几何概型问题，故

$$P(A)=\frac{\text{区间}[0.5,\ 1]\text{ 的长度}}{\text{区间}(0,\ 4]\text{ 的长度}}=\frac{1/2}{4}=\frac{1}{8}.$$

例 1.3.8　(会面问题) 甲、乙两人相约在 7 点到 8 点之间在某地会面，先到者等候另一人 20 分钟，过时就离开. 如果每个人在指定的一小时内任意时刻到达，试计算二人能够会面的概率.

解　记 7 点为计算时刻的 0 时，以分钟为单位，x，y 分别为甲、乙到达指定地点的时刻，则样本空间为 $\Omega=\{(x,\ y)\mid 0\leqslant x\leqslant 60,\ 0\leqslant y\leqslant 60\}$. 设 A 为事件“两人能会面”，则显然有 $A=\{(x,\ y)\mid (x,\ y)\in\Omega,\ |x-y|\leqslant 20\}$，即图 1.7 中阴影部分区域. 根据题意，这是一个几何概型问题，二人能会面的概率为

$$P(A)=\frac{\mu(A)}{\mu(\Omega)}=\frac{60^2-40^2}{60^2}=\frac{5}{9}.$$

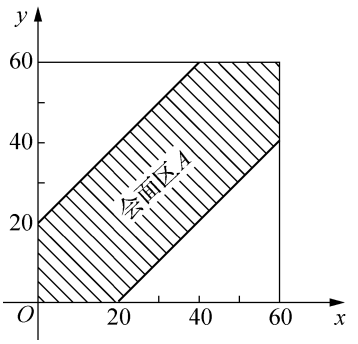


图 1.7　会面问题

几何概型的一个著名例子是蒲丰(Buffon) 投针试验. 通过这个试验还可以求圆周率 π 的近似值.

例 1.3.9(蒲丰投针试验)　在平面上画有平行线束，两条相邻的平行线的距离均为 $2a$，向平面随机投掷一枚长度为 $2l$ 的针，假定 $0<l<a$. 求针与平行线相交的概率 p.

解　设 M 为针的中点，Y 为 M 与最近平行线的距离，θ 为针与平行线的交角(见图

1.8)，则点(Y, θ)“均匀地”散布在矩形$\Omega = \{(y, \theta) \mid 0 \leqslant y \leqslant a, 0 \leqslant \theta \leqslant \pi\}$上. 不难知道针与平行线相交的充要条件是$Y \leqslant l\sin\theta$，即$(Y, \theta)$落在图1.9中的阴影区域上，故针与平行线相交的概率p为阴影区域面积与矩形面积之比，即

$$p = \frac{1}{\pi a}\int_0^{\pi} l\sin\theta \mathrm{d}\theta = \frac{2l}{\pi a}.$$

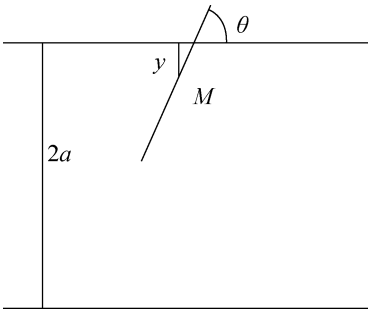


图1.8　蒲丰投针

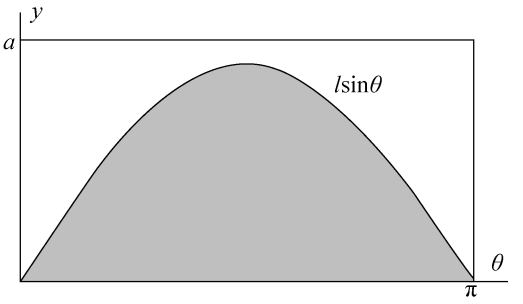


图1.9　投影区域

1.4　条件概率

1.4.1　条件概率定义

在实际问题中，除了要考虑某事件A发生的概率$P(A)$外，有时还要考虑事件B发生条件下事件A发生的概率. 一般情况下，后者的概率与前者不同. 为了区别起见，我们把后者的概率称为条件概率，记为$P(A \mid B)$，读作事件B发生条件下事件A的条件概率. 条件概率是概率论中的一个重要概念，由它可产生三个非常有用的公式，即乘法公式、全概率公式和贝叶斯公式.

为了引进条件概率概念，我们先看一个例子.

例1.4.1　考虑有两个孩子的家庭. 样本空间$\Omega = \{$(男、男)，(男、女)，(女、男)，(女、女)$\}$. 设A为事件“家庭有女孩”，B为事件“家庭有男孩”. 求已知家庭有男孩条件下家庭有女孩的条件概率.

解　显然，问题是求事件B发生的条件下事件A发生的概率. 此时，$A = \{$(男、女)，(女、男)，(女、女)$\}$，$B = \{$(男、男)，(男、女)，(女、男)$\}$. 已知事件B已经发生了，有了这个信息，就知道有两个女孩的家庭在此种情况下不可能出现. 因此，在B发生条件下样本空间可视为$B = \{$(男、男)，(男、女)，(女、男)$\}$，而事件A中属于这个样本空间的点有2个，于是由古典概率方法可得A的条件概率

$$P(A \mid B) = \frac{2}{3}.$$

另外，易见

$$P(A) = \frac{3}{4},\ P(B) = \frac{3}{4},\ P(AB) = \frac{2}{4},\ P(A \mid B) = \frac{2}{3} = \frac{2/4}{3/4},$$

所以有

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

上式启发我们，可以用 $P(AB)$ 与 $P(B)$ 的比值作为条件概率 $P(A \mid B)$ 的定义.

定义 1.4.1 设 A，B 是两个事件，且 $P(B) > 0$，则已知事件 B 发生条件下事件 A 发生的条件概率为

$$P(A \mid B) = \frac{P(AB)}{P(B)} \tag{1.4.1}$$

关于条件概率，应注意如下两点。

1. 条件概率 $P(A \mid B)$ 也是一个概率，因为它满足概率定义中的三条公理：

(1) $P(A \mid B) \geqslant 0$；

(2) $P(\Omega \mid B) = 1$；

(3) 若 A_1，A_2，… 是可数个两两互斥事件，则 $P(\bigcup_{i=1}^{\infty} A_i \mid B) = \sum_{i=1}^{\infty} P(A_i \mid B)$.

因而条件概率应具有概率的所有性质，例如 $P(\overline{A} \mid B) = 1 - P(A \mid B)$.

2. 计算条件概率可选择如下两种方法之一. (1) 在原样本空间 Ω 中，先计算 $P(AB)$，$P(B)$，再按公式 $P(A \mid B) = \dfrac{P(AB)}{P(B)}$ 计算条件概率；(2) 由于事件 B 已经出现，可以将其视为新的样本空间，并在该样本空间下计算事件 A 发生的概率 $P(A \mid B)$.

例 1.4.2 某疾病 D 的医学检验结果可能为阳性(用 A 表示) 或阴性(用 $\overline{A}$ 表示)，有关概率由下表给出.

表 1.4 事件的概率结果

	D	$\overline{D}$
A	0.009	0.099
$\overline{A}$	0.001	0.891

由条件概率的定义可得

$$P(A \mid D) = \frac{P(AD)}{P(D)} = \frac{0.009}{0.009 + 0.001} = 0.9,$$

$$P(\overline{A} \mid \overline{D}) = \frac{P(\overline{A}\,\overline{D})}{P(\overline{D})} = \frac{0.891}{0.099 + 0.891} = 0.9.$$

显然，该检验是相当精确的，对患者的检验结果有 90% 呈阳性，而对健康者的检验结果有 90% 呈阴性. 假定某人的检查的结果是呈阳性，那么这个人患疾病 D 的概率会有多大呢？凭直觉很容易认为这个概率会很大，但正确的结果是

$$P(D \mid A) = \frac{P(AD)}{P(A)} = \frac{0.009}{0.009 + 0.099} = \frac{1}{12} \approx 0.083.$$

本例结果表明，虽然 $P(A \mid D) = 0.9$，$P(\overline{A} \mid \overline{D}) = 0.9$，这两个概率都很高. 但若将检验结果呈阳性用于判断某人患有疾病 D，其正确性只有约 8%.

例 1.4.3 设某种动物从出生起活到 20 岁以上的概率为 0.8，活到 25 岁以上的概率

为 0.5. 求

（1）如果现在有一个 20 岁的这种动物，它能活到 25 岁以上的概率；

（2）如果现在有一个 20 岁的这种动物，它活不到 25 岁的概率.

解　（1）设 A 为“能活到20岁以上”，B 为“能活到25岁以上”. 依题意，$P(A)=0.8$，$P(B)=0.5$. 由于 $B\subset A$，因此 $P(AB)=P(B)=0.5$. 由条件概率定义得

$$P(B\mid A)=\frac{P(AB)}{P(A)}=\frac{0.5}{0.8}=0.625.$$

（2）一个 20 岁的这种动物，它活不到 25 岁的概率为

$$P(\overline{B}\mid A)=1-P(B\mid A)=1-0.625=0.375.$$

1.4.2　乘法公式

由条件概率定义容易推得概率的乘法公式：

$$P(AB)=P(A)P(B\mid A)=P(B)P(A\mid B) \tag{1.4.2}$$

该乘法公式可以推广到多个事件的情形，若 $n\geqslant 2$ 且 $P(A_1A_2\cdots A_{n-1})>0$，则

$$\begin{aligned}P(A_1A_2\cdots A_n)&=P(A_1)\cdot\frac{P(A_1A_2)}{P(A_1)}\cdot\frac{P(A_1A_2A_3)}{P(A_1A_2)}\cdots\frac{P(A_1A_2\cdots A_n)}{P(A_1A_2\cdots A_{n-1})}.\\&=P(A_1)P(A_2\mid A_1)P(A_3\mid A_1A_2)\cdots P(A_n\mid A_1\cdots A_{n-1})\end{aligned}$$

利用概率的乘法公式容易计算若干个事件的积事件概率.

例 1.4.4　在一批由 90 件正品，3 件次品组成的产品中，不放回地连续抽取两件产品，求第一件为正品，第二件为次品的概率.

解　设 A 为“第一件为正品”，B 为“第二件为次品”，求概率 $P(AB)$. 依题意有 $P(A)=\frac{90}{93}$，$P(B\mid A)=\frac{3}{92}$. 由乘法公式得

$$P(AB)=P(A)P(B\mid A)=\frac{90}{93}\times\frac{3}{92}=\frac{45}{1426}\approx 0.0315.$$

例 1.4.5　设袋中有 a 只红球，b 只白球，随机抽出一只，观察其颜色后放回，并加进同样颜色的球 c 只，一共抽取了 $m+n$ 次球. 试求前 m 次取到红球，后 n 次取到白球的概率.

解　设 A_i 为第 i 次取到红球的事件，$i=1,2,\cdots,m+n$，则前 m 次取到红球，后 n 次取到白球的事件为 $A_1A_2\cdots A_m\overline{A}_{m+1}\overline{A}_{m+2}\cdots\overline{A}_{m+n}$. 依题意有

$$P(A_1)=\frac{a}{a+b},$$

$$P(A_2\mid A_1)=\frac{a+c}{a+b+c},$$

$$P(A_3\mid A_1A_2)=\frac{a+2c}{a+b+2c},$$

$$\cdots$$

$$P(A_m\mid A_1A_2\cdots A_{m-1})=\frac{a+(m-1)c}{a+b+(m-1)c},$$

$$P(\overline{A}_{m+1} \mid A_1A_2\cdots A_m) = \frac{b}{a+b+mc},$$

$$P(\overline{A}_{m+2} \mid A_1A_2\cdots A_m\overline{A}_{m+1}) = \frac{b+c}{a+b+(m+1)c},$$

$$P(\overline{A}_{m+n} \mid A_1A_2\cdots A_m\overline{A}_{m+1}\cdots\overline{A}_{m+n-1}) = \frac{b+(n-1)c}{a+b+(m+n-1)c},$$

因此

$$\begin{aligned}
&P(A_1A_2\cdots A_m\overline{A}_{m+1}\overline{A}_{m+2}\cdots\overline{A}_{m+n})\\
&= P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1A_2)\cdots P(A_m \mid A_1A_2\cdots A_{m-1})\\
&\quad\times P(\overline{A}_{m+1} \mid A_1\cdots A_m)P(\overline{A}_{m+2} \mid A_1\cdots A_m\overline{A}_{m+1})\cdots P(\overline{A}_{m+n} \mid A_1\cdots A_m\overline{A}_{m+1}\cdots\overline{A}_{m+n-1})\\
&= \frac{a}{a+b}\,\frac{a+c}{a+b+c}\,\frac{a+2c}{a+b+2c}\cdots\frac{a+(m-1)c}{a+b+(m-1)c}\\
&\quad\times\frac{b}{a+b+mc}\,\frac{b+c}{a+b+(m+1)c}\cdots\frac{b+(n-1)c}{a+b+(m+n-1)c}.
\end{aligned}$$

上述问题所求的概率只与红球、白球出现的次数有关，而与它们出现的次序无关. 特别地，当 $c=0$ 时是有放回抽样的摸球问题，当 $c=-1$ 时是无放回抽样的摸球问题. 历史上玻利亚(Ploya) 曾经用上述模型讨论传染病传播的规律.

1.4.3 全概率公式

为了计算复杂事件的概率，人们经常把一个复杂事件分解为若干个互斥的简单事件的和，通过分别计算这些简单事件的概率，来求复杂事件的概率. 在这种做法中全概率公式起着非常重要的作用.

定义 1.4.2 设 Ω 是某随机试验的样本空间，A_1，A_2，…，A_n 是一组事件. 若 A_1，A_2，…，A_n 满足下列条件：

(1) 两两互斥，即 $A_iA_j=\varnothing$，$i\neq j$，i，$j=1$，2，…，n；

(2) $A_1\cup A_2\cup\cdots\cup A_n=\Omega$.

则称 A_1，A_2，…，A_n 为样本空间 Ω 的一个划分，并称 A_1，A_2，…，A_n 构成一个**完备事件组**.

容易知道，A_1，A_2，…，A_n 是样本空间 Ω 的一个划分，当且仅当事件 A_1，A_2，…，A_n 在每次试验中必有一个发生，且恰有一个发生.

在许多场合，若事件 B 的概率不易直接求出，此时可利用样本空间的划分将事件 B 表示为

$$B = B\Omega = B\cap\left(\bigcup_{i=1}^{n}A_i\right) = \bigcup_{i=1}^{n}A_iB,$$

由于 A_1，A_2，…，A_n 两两互斥，易知 A_1B，A_2B，…，A_nB 也是两两互斥的. 因此，若 $P(A_i)>0(i=1,2,\cdots,n)$，由概率的性质 2 和乘法公式可得

$$P(B) = \sum_{i=1}^{n}P(A_iB) = \sum_{i=1}^{n}P(A_i)P(B \mid A_i) \tag{1.4.3}$$

这个公式称为**全概率公式**. 特别地，当 $n=2$ 时，记 $A_1=A$，$A_2=\bar{A}$，则有最简单的全概率公式

$$P(B)=P(A)P(B\mid A)+P(\bar{A})P(B\mid \bar{A}).$$

例 1.4.6　若 10 张彩票中有 2 张有奖，10 个顾客各抽一张，求第二个顾客中奖的概率.

解　设 A，B 分别表示第一个、第二个顾客中奖，则

$$P(A)=\frac{2}{10},\ P(\bar{A})=\frac{8}{10},\ P(B\mid A)=\frac{1}{9},\ P(B\mid \bar{A})=\frac{2}{9}.$$

由全概率公式

$$P(B)=P(A)P(B\mid A)+P(\bar{A})P(B\mid \bar{A})=\frac{2}{10}\times\frac{1}{9}+\frac{8}{10}\times\frac{2}{9}=\frac{2}{10}=0.2.$$

从这个例子我们看到，第一个顾客和第二个顾客中奖的概率都是 0.2. 事实上，每个人中奖的概率都一样，这就是“抽签公平原理”.

例 1.4.7　设某仓库有一批产品，已知其中有 50%、30%、20% 的产品依次是甲、乙、丙厂生产的，且甲、乙、丙厂生产的次品率分别为$\frac{1}{10}$，$\frac{1}{15}$，$\frac{1}{20}$. 求

(1) 从这批产品中任取一件产品，取到次品的概率；

(2) 若从这批产品中取出一件产品，发现是次品，它是由甲厂生产的概率.

解　(1) 以 A_1，A_2，A_3 分别表示事件“取到的产品是由甲、乙、丙厂生产的”，以 B 表示事件“取到的产品为次品”，则

$$P(A_1)=\frac{5}{10},\ P(A_2)=\frac{3}{10},\ P(A_3)=\frac{2}{10},$$

$$P(B\mid A_1)=\frac{1}{10},\ P(B\mid A_2)=\frac{1}{15},\ P(B\mid A_3)=\frac{1}{20}.$$

由全概率公式

$$\begin{aligned}P(B)&=P(A_1)P(B\mid A_1)+P(A_2)P(B\mid A_2)+P(A_3)P(B\mid A_3)\\&=\frac{5}{10}\cdot\frac{1}{10}+\frac{3}{10}\cdot\frac{1}{15}+\frac{2}{10}\cdot\frac{1}{20}=0.08.\end{aligned}$$

(2) $P(A_1\mid B)=\dfrac{P(A_1B)}{P(B)}=\dfrac{P(A_1)P(B\mid A_1)}{P(B)}=\dfrac{0.5\times0.01}{0.08}=0.625.$

1.4.4　贝叶斯公式

将例 1.4.7(2) 的计算方法推广到一般，即得贝叶斯公式.

设 A_1，A_2，…，A_n 为某试验 E 的样本空间 Ω 的一个划分，且 $P(A_i)>0(i=1,2,\cdots,n)$，B 为一个事件，且 $P(B)>0$，则有

$$P(A_i\mid B)=\frac{P(A_iB)}{P(B)}=\frac{P(A_i)P(B\mid A_i)}{\sum\limits_{j=1}^{n}P(A_j)P(B\mid A_j)}\qquad(1.4.4)$$

这个公式称为**贝叶斯公式**，也称为后验概率公式.

在贝叶斯公式(1.4.4)中，概率 $P(A_i)$ 可视为某种试验之前就已经获得的信息，称其为 A_i 的先验概率. 如果试验产生了结果 B，则概率 $P(A_i \mid B)$ 是事件 B 发生条件下对原有概率 $P(A_i)$ 的修正，通常称 $P(A_i \mid B)$ 为 A_i 的后验概率.

例 1.4.8 对以往数据的分析结果表明，当机器状态良好时，产品的合格率为 98%，而当机器发生某种故障时，产品的合格率为 55%. 每天早上机器开动时，其状态良好的概率为 95%. 试求已知某日早上第一件产品是合格品时，机器状态为良好的概率.

解 设 A 为事件“产品合格”，B 为事件“机器状态良好”. 已知 $P(A \mid B) = 0.98$，$P(A \mid \overline{B}) = 0.55$，$P(B) = 0.95$，$P(\overline{B}) = 1 - P(B) = 0.05$. 由贝叶斯公式，所求概率为

$$P(B \mid A) = \frac{P(AB)}{P(A)} = \frac{P(B)P(A \mid B)}{P(B)P(A \mid B) + P(\overline{B})P(A \mid \overline{B})}$$

$$= \frac{0.95 \times 0.98}{0.95 \times 0.98 + 0.05 \times 0.55} = 0.97.$$

这里概率 $P(B) = 0.95$ 是由以往的数据分析得到的，是先验概率. 而在得到信息(即生产出的第一件产品是合格品)之后再重新加以修正的概率为 $P(B \mid A) = 0.97$，这个概率就是后验概率. 有了后验概率我们就能对机器的状态有新的了解.

贝叶斯公式在通信技术中有大量的应用. 在数字通信过程中，信号通常用高、低电平表示，通信中由于噪声干扰及能量衰减的影响，收到的信号可能不是原来的信号，因此，接收方要通过概率的计算来作出判断，保证通信的质量.

例 1.4.9 假设发报台分别以概率 0.6 和 0.4 发出信号“. ”和“-”，由于通信系统受到干扰，当发出信号“. ”时，收报台未必收到信号“. ”，而是分别以 0.8 和 0.2 的概率收到“. ”和“-”；同样，发出“-”时分别以 0.9 和 0.1 的概率收到“-”和“. ”. 如果收报台收到“. ”，求它收到的是正确信号的概率.

解 设 A 为发报台发出信号“. ”，B 为收报台收到信号“. ”. 则 $\overline{A}$ 为发报台发出信号“-”，$\overline{B}$ 为收报台收到信号“-”. 于是，$P(A) = 0.6$，$P(\overline{A}) = 0.4$，$P(B \mid A) = 0.8$，$P(\overline{B} \mid A) = 0.2$，$P(\overline{B} \mid \overline{A}) = 0.9$，$P(B \mid \overline{A}) = 0.1$.

由贝叶斯公式，所求概率为

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(\overline{A})P(B \mid \overline{A})}$$

$$= \frac{0.6 \times 0.8}{0.6 \times 0.8 + 0.4 \times 0.1} = \frac{12}{13} \approx 0.923.$$

1.5 事件的独立性

设 A，B 是两个事件，若 $P(B) > 0$，则可以定义条件概率 $P(A \mid B)$. 一般而言 $P(A \mid B) \neq P(A)$，这表明事件 B 的发生对事件 A 发生的概率有影响. 只有当这种影响不存在时，才会有 $P(A \mid B) = P(A)$，此时称事件 A，B 相互独立.

定义 1.5.1 若两个事件 A，B 满足 $P(A) = P(A \mid B)$，则称 A 与 B 独立，或称 A，B 相

互独立.

若 A，B 相互独立，则由乘法公式有

$$P(AB)=P(A\mid B)P(B)=P(A)P(B);$$

反之，若 $P(AB)=P(A)P(B)$ 成立，由条件概率有

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

这表明 A 与 B 独立与等式 $P(AB)=P(A)P(B)$ 等价，从而得到两个事件独立的另一种表述方式.

定义 1.5.2　若两个事件 A，B 满足 $P(AB)=P(A)P(B)$，则称 A 与 B 独立，或称 A，B 相互独立.

定义 1.5.1 从互不影响的角度出发表述事件的独立性概念，易于直观理解，但不便于应用. 定义 1.5.2 以严密的数学形式刻画了独立性概念，不仅应用方便，且可将两个事件的独立性概念推广到多个事件的独立性.

定义 1.5.3　若事件 A，B，C 满足

$$\begin{cases}P(AB)=P(A)P(B),\\ P(BC)=P(B)P(C),\\ P(AC)=P(A)P(C),\end{cases}$$

则称事件 A，B，C 两两独立.

定义 1.5.4　若事件 A，B，C 满足

$$\begin{cases}P(AB)=P(A)P(B),\\ P(BC)=P(B)P(C),\\ P(AC)=P(A)P(C),\\ P(ABC)=P(A)P(B)P(C),\end{cases}$$

则称事件 A，B，C 相互独立.

一般地，有下列独立性定义.

定义 1.5.5　若事件 A_1，A_2，…，A_n 满足

$$\begin{cases}P(A_iA_j)=P(A_i)P(A_j),\ \forall\, i\neq j,\\ P(A_iA_jA_k)=P(A_i)P(A_j)P(A_k),\ \forall\, i\neq j\neq k,\\ \cdots\\ P(A_1A_2\cdots A_n)=P(A_1)P(A_2)\cdots P(A_n),\end{cases}$$

则称这 n 个事件 A_1，A_2，…，A_n 相互独立.

关于上述独立性概念，应注意下面几点.

（1）容易证明，必然事件 Ω 和不可能事件 $\varnothing$ 与任何事件都独立. 这一事实并不意外，因为必然事件和不可能事件是确定性事件，它们不受任何事件的影响，也不影响任何事件的发生.

（2）事件的独立性与事件的互斥是两个不同的概念，它们之间没有必然联系. 两个事件互斥表示它们不能同时发生，两个事件独立表示它们彼此互不影响. 当 $P(A)>0$，$P(B)>0$ 时，若 A，B 相互独立，则 $P(AB)=P(A)P(B)>0$，若 A，B 互斥，则 $P(AB)=P(\varnothing)=0$，此时 A，B 相互独立和 A，B 互斥不会同时成立.

(3) 多个事件相互独立一定是两两独立的，但两两独立未必相互独立.

(4) 两个事件独立与两个事件对立也是不同的概念. 两个事件对立是指它们互为逆事件，但它们不一定独立；反之，两个事件独立它们不一定对立.

关于事件的独立性，还有如下结论.

定理 1.5.1 若四对事件 A 与 B，$\overline{A}$ 与 B，A 与 $\overline{B}$，$\overline{A}$ 与 $\overline{B}$ 中有一对相互独立，则另外三对也相互独立.

证明 以下只证明，若 A，B 相互独立，则 A 与 $\overline{B}$ 也独立. 其他可类似证明.

因为 A，B 相互独立，所以 $P(AB)=P(A)P(B)$，于是

$$\begin{aligned} P(A\overline{B}) &= P(A-B) = P(A)-P(AB) \\ &= P(A)-P(A)P(B) = P(A)[1-P(B)] \\ &= P(A)P(\overline{B}). \end{aligned}$$

所以 A 与 $\overline{B}$ 独立，结论成立.

由定义 1.5.5 和定理 1.5.1 可知，以下两个结论成立.

(1) 若事件 A_1，A_2，…，A_n 相互独立，则其中任意 $k(2\leqslant k\leqslant n)$ 个事件也相互独立.

(2) 若事件 A_1，A_2，…，A_n 相互独立，则将 A_1，A_2，…，A_n 中任意多个事件换成它们各自的对立事件，所得的 n 个事件仍相互独立.

在实际问题中，我们一般不是用定义来判断事件之间是否相互独立，而是根据具体问题去判断它们的独立性. 如果认为是独立的，就可以利用独立性结论来简化事件概率的计算.

例 1.5.1 两门高射炮各自独立地射击一架敌机，设甲炮击中敌机的概率为 0.9，乙炮击中敌机的概率为 0.8，求敌机被击中的概率.

解 设 A 为“甲炮击中敌机”，B 为“乙炮击中敌机”，则 $A\cup B$ 为“敌机被击中”. 因为 A 与 B 独立，所以有

$$\begin{aligned} P(A\cup B) &= P(A)+P(B)-P(AB) \\ &= P(A)+P(B)-P(A)P(B) \\ &= 0.9+0.8-0.9\times 0.8 = 0.98. \end{aligned}$$

此题还有另一种解法. 由定理 1.5.1 知，$\overline{A}$ 与 $\overline{B}$ 相互独立，故

$$\begin{aligned} P(A\cup B) &= 1-P(\overline{A\cup B}) = 1-P(\overline{A}\,\overline{B}) \\ &= 1-P(\overline{A})P(\overline{B}) = 1-(1-0.9)(1-0.8) = 0.98. \end{aligned}$$

一般来说，相互独立事件的和事件的概率计算，可转化为其对立事件的积事件的概率计算，从而简化运算过程，即当 A_1，A_2，…，A_n 相互独立时

$$\begin{aligned} P\left(\bigcup_{i=1}^{n}A_i\right) &= 1-P\left(\overline{\bigcup_{i=1}^{n}A_i}\right) \\ &= 1-P\left(\bigcap_{i=1}^{n}\overline{A_i}\right) = 1-P(\overline{A_1})P(\overline{A_2})\cdots P(\overline{A_n}). \end{aligned}$$

例 1.5.2 设某地区的人群中每人血液中含有某种病毒的概率为 0.001，将 2 000 人的血液进行混合，求混合后的血液中含有该病毒的概率.

解 设 $A_i(1\leqslant i\leqslant 2\,000)$ 为第 i 个人的血液中含有病毒的事件，混合后的血液中含有

病毒的事件为 $\bigcup_{i=1}^{2\,000} A_i$，其概率为

$$P\left(\bigcup_{i=1}^{2\,000} A_i\right) = 1-P\left(\overline{\bigcup_{i=1}^{2\,000} A_i}\right) = 1-P\left(\bigcap_{i=1}^{2\,000} \overline{A_i}\right)$$
$$= 1-P(\overline{A_1})P(\overline{A_2})\cdots P(\overline{A_{2\,000}})$$
$$= 1-(1-0.001)^{2\,000} = 1-0.999^{2\,000} \approx 0.864\,8.$$

从该例可以看出，虽然每个人携带病毒的概率很小，但混合后的血液中含有病毒的概率却很大. 在实际中，这类效应值得引起注意. 比如，在购买 35 选 7 的福利彩票时，中特等奖的概率为 $\frac{1}{C_{35}^7} = \frac{1}{6\,724\,520}$，非常小，但我们在报纸上却经常看到有人中了特等奖；一辆汽车在一天中发生交通事故的概率是非常小的，但在一座大城市里交通事故时有发生等. 这些都启示我们不要忽视小概率事件.

习题一

1. 写出下列随机试验的样本空间 Ω.

（1）记录一个班一次数学考试的平均分数(设以百分制记分)；

（2）生产某种产品直到有 10 件正品为止，记录此过程中生产该种产品的总件数；

（3）对某工厂出厂的产品进行检查，合格的记为“正品”，不合格的记为“次品”，若连续查出了 2 件次品就停止检查，或检查了 4 件产品就停止检查，记录检查的结果；

（4）在单位圆内任意取一点，记录它的坐标.

2. 设 A，B，C 为三个事件，用 A，B，C 及其运算关系表示下列事件.

（1）A 发生而 B 与 C 不发生；

（2）A，B，C 中恰好有一个发生；

（3）A，B，C 中至少有一个发生；

（4）A，B，C 中恰好有两个发生；

（5）A，B，C 中至少有两个发生；

（6）A，B，C 中有不多于一个发生.

3. 设样本空间 $\Omega = \{x \mid 0 \leqslant x \leqslant 2\}$，事件 $A = \{x \mid 0.5 \leqslant x \leqslant 1\}$，$B = \{x \mid 0.8 < x \leqslant 1.6\}$，具体写出下列事件.

（1）AB；（2）$A-B$；（3）$\overline{A-B}$；（4）$\overline{A \cup B}$.

4. 一个样本空间有三个样本点，其对应的概率分别为 $2p$，p^2，$4p-1$，求 p 的值.

5. 已知 $P(A) = 0.3$，$P(B) = 0.5$，$P(A \cup B) = 0.8$. 求(1) $P(AB)$；(2) $P(A-B)$；(3) $P(\overline{A}\,\overline{B})$.

6. 设 $P(AB) = P(\overline{A}\,\overline{B})$，且 $P(A) = p$，求 $P(B)$.

7. 对于事件 A，B，C，设 $P(A) = 0.4$，$P(B) = 0.5$，$P(C) = 0.6$，$P(AC) = 0.2$，$P(BC) = 0.4$ 且 $AB = \Phi$，求 $P(A \cup B \cup C)$.

8. 将 3 个球随机地放入 4 个杯子中去，求杯子中球的最大个数分别为 1、2、3 的概率.

9. 在整数 0 ~ 9 中任取 4 个，它们能排成一个四位偶数的概率是多少？

10. 一部五卷的文集，按任意次序放到书架上去，试求下列事件的概率.（1）第一卷出现在旁边；（2）第一卷及第五卷出现在旁边；（3）第一卷或第五卷出现在旁边；（4）第一卷及第五卷都不出现在旁边；（5）第三卷正好在正中.

11. 把 2，3，4，5 四个数字各写在一张小纸片上，任取其中三个按自左向右的次序排成一个三位数，求所得数是偶数的概率.

12. 一幢 10 层楼中一架电梯在底层登上 7 位乘客，电梯在每一层都停，乘客从第二层起离开电梯，假设每位乘客在任一层离开电梯是等可能的，求没有两位及两位以上乘客在同一层离开的概率.

13. 某人午觉醒来发觉表停了，他打开收音机想收听电台报时，设电台每正点报时一次，求他(她）等待时间短于 10 分钟的概率.

14. 甲乙两人相约 8 ~ 12 点在预定地点会面. 先到的人等候另一人 30 分钟后离去，求甲乙两人能会面的概率.

15. 现有两种报警系统 A 和 B，每种系统单独使用时，系统 A 有效的概率为 0.92，系统 B 的有效概率为 0.93，而两种系统一起使用时，在 A 失灵的条件下 B 有效的概率为 0.85，求两种系统一起使用时，

（1）这两个系统至少有一个有效的概率；

（2）在 B 失灵条件下，A 有效的概率.

16. 已知事件 A 发生的概率 $P(A) = 0.5$，B 发生的概率 $P(B) = 0.6$，以及条件概率 $P(B \mid A) = 0.8$，求 A，B 中至少有一个发生的概率.

17. 一批零件共 100 个，其中有次品 10 个. 每次从该批零件中任取 1 个，取出后不放回，连取 3 次. 求第 3 次才取得合格品的概率.

18. 有两个袋子，每个袋子都装有 a 只黑球，b 只白球，从第一个袋中任取一球放入第二个袋中，然后从第二个袋中取出一球，求取得的是黑球的概率.

19. 一个机床有 $\frac{1}{3}$ 的时间加工零件 A，其余时间加工零件 B. 加工零件 A 时，停机的概率是 0.3；加工零件 B 时，停机的概率是 0.4，求这个机床停机的概率.

20. 10 个考签中有 4 个难签，3 个人参加抽签考试，不重复地抽取，每人抽一次，甲先，乙次，丙最后. 证明 3 人抽到难签的概率相同.

21. 两部机器制造大量的同一种零件，根据长期资料统计，甲、乙机器制造出的零件废品率分别是 0.01 和 0.02. 现有同一机器制造的一批零件，估计这一批零件是乙机器制造的可能性比甲机器制造的可能性大一倍，现从这批零件中任意抽取一件，经检查是废品. 试由此结果计算这批零件是由甲机器制造的概率.

22. 有朋友来自远方，他乘火车、轮船、汽车、飞机来的概率分别是 0.3、0.2、0.1、0.4. 如果他乘火车、轮船、汽车来的话，迟到的概率分别是 $\frac{1}{4}$、$\frac{1}{3}$、$\frac{1}{12}$，而乘飞机则不会迟到. 结果他迟到了，试求他是乘火车来的概率.

23. 加工一个产品要经过三道工序，第一、二、三道工序不出现废品的概率分别是 0.9、0.95、0.8. 假定各工序是否出废品相互独立，求经过三道工序而不出现废品的

概率.

24. 三个人独立地破译一个密码，他们能译出的概率分别是 0.2、1/3、0.25. 求密码被破译的概率.

25. 对同一目标，3 名射手独立射击的命中率分别是 0.4、0.5 和 0.7，求三人同时向目标各射一发子弹而没有一发中靶的概率.

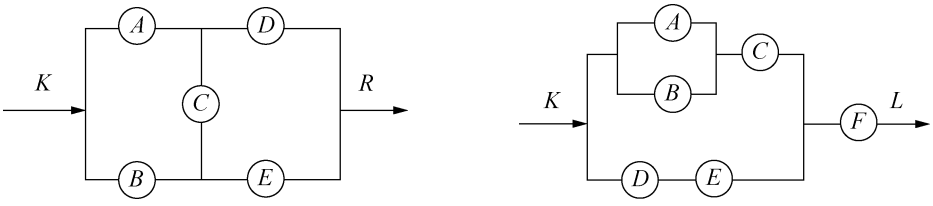
26. 甲、乙、丙三人同时对飞机进行射击，三人击中的概率分别为 0.4、0.5、0.7. 飞机被一人击中而击落的概率为 0.2，被两人击中而击落的概率为 0.6，若三人都击中，飞机必定被击落，求飞机被击落的概率.

27. 证明：若三个事件 A，B，C 相互独立，则 $A \cup B$、AB 及 $A-B$ 都与 C 独立.

28. 15 个乒乓球中有 9 个新球，6 个旧球，第一次比赛取出了 3 个，用完后放回去，第二次比赛又取出 3 个，求第二次取出的 3 个球全是新球的概率.

29. 要验收一批 100 件的物品，从中随机地取出 3 件来测试，设 3 件物品的测试是相互独立的，如果 3 件中有一件不合格，就拒绝接收该批物品. 设一件不合格的物品经测试被查出的概率为 0.95，而一件合格品经测试被误认为不合格的概率为 0.01，如果这 100 件物品中有 4 件是不合格的，求这批物品被接收的概率.

30. 设下图的两个系统 KL 和 KR 中各元件通达与否相互独立，且每个元件通达的概率均为 p，分别求系统 KL 和 KR 通达的概率.



第二章　一维随机变量及其分布

2.1　随机变量的定义

在第一章中，我们把随机试验的所有可能结果组成的集合称为样本空间，并用样本空间的子集来表示随机事件. 为了方便地研究随机试验的各种结果及其发生的概率，本章引入随机变量的概念，把随机试验的结果与实数对应起来，即把随机试验的结果数量化.

定义 2.1.1　设 E 是随机试验，Ω 是其样本空间. 如果对每个样本点 $\omega \in \Omega$，总有一个实数 $X = X(\omega)$ 与之对应，则称 X 为该随机试验的随机变量.

从上述定义我们知道，随机变量 X 是定义在样本空间上 Ω 的实值函数，它的自变量是随机试验的结果. 由于随机试验结果的出现具有随机性，所以随机变量的取值也具有随机性，这是随机变量与一般函数的不同之处.

引进随机变量，就相当于对样本空间的数量化，此时每个随机事件(即样本空间的子集) 都可以用随机变量来描述.

例 2.1.1　抛掷一枚均匀硬币，样本空间为 $\Omega = \{\text{正面，反面}\}$，设

$$X = X(\omega) = \begin{cases} 1, & \omega = \text{正面}, \\ 0, & \omega = \text{反面}. \end{cases}$$

则 X 是定义在样本空间 Ω 上的随机变量，它将原样本空间数量化为 $\Omega = \{0, 1\}$.

例 2.1.2　从某学校学生中任选一人 ω，记其身高为 $X = X(\omega)$，它随 ω 而变，故 X 是定义在集合 $\Omega = \{\omega: \omega \text{ 为该学校学生}\}$ 上的随机变量.

例 2.1.3　观察一部电梯一年内出现故障的次数. 记 ω_i 为“电梯一年内发生 i 次故障”，$i = 0, 1, 2, \cdots$，样本空间为

$$\Omega = \{\omega_i, \ i = 0, 1, 2, \cdots\}.$$

引入随机变量

$$X(\omega_i) = i, \ i = 0, 1, 2, \cdots,$$

它将原样本空间数量化为非负整数集合 $\Omega = \{0, 1, 2, \cdots\}$.

例 2.1.4　考虑测试灯泡寿命的试验，记 ω 为灯泡的使用寿命，样本空间为

$$\Omega = \{\omega \mid 0 \leqslant \omega < \infty\},$$

它是非负实数集合. 对于每个 $\omega \in \Omega$，引入随机变量

$$X(\omega) = \omega, \ 0 \leqslant \omega < \infty.$$

对于任意实数集合 L，事件 $\{\omega \mid X(\omega) \in L\}$ 可以简记为 $\{X \in L\}$，即样本点落在集合 L 的事件. 例如，在例 2.1.1 中，用 $\{X = 1\}$ 表示事件“出现正面”，用 $\{X = 0\}$ 表示事件“出现反面”；在例 2.1.3 中，用 $\{X \leqslant 5\}$ 表示事件“电梯在一年内出现故障的次数不超过 5”；在例 2.1.4 中，若寿命以小时计，则 $\{X \geqslant 1\,000.5\}$ 表示事件“灯泡的使用寿命不小于 1 000 小时 30 分钟”. 这样，我们就可以把对随机事件的研究转化为对随机变量的研究. 由于随机变量的值是实数，因此我们可以利用函数和微积分等数学方法一般地研究随机事件的概率.

随机变量主要分为两种类型. 一类是离散型随机变量，另一类是连续型随机变量. 例 2.1.1 和例 2.1.3 中的随机变量 X 是离散型的，而例 2.1.2 中的学生身高和例 2.1.4 中的灯泡使用寿命属于连续型随机变量.

2.2　随机变量的分布函数

对于随机变量 X，它的随机取值规律称为概率分布. 通常用分布函数，分布律或分布密度来刻画随机变量的概率分布.

定义 2.2.1　设 X 为随机变量，x 是任意实数，称事件 $\{X \leqslant x\}$ 的概率

$$F(x) = P\{X \leqslant x\}, \quad -\infty < x < +\infty \tag{2.2.1}$$

为 X 的分布函数.

由分布函数的定义可知，$F(x)$ 是随机变量 X 落在区间 $(-\infty, x]$ 内的概率，由于该概率随实数 x 的变化而变化，因此它是 x 的函数. 由概率的取值范围可知，分布函数 $F(x)$ 的值域是区间 $[0, 1]$.

此外，对于任意实数 x_1，$x_2 (x_1 < x_2)$，由于 $\{x_1 < X \leqslant x_2\} = \{X \leqslant x_2\} - \{X \leqslant x_1\}$ 且 $\{X \leqslant x_1\} \subseteq \{X \leqslant x_2\}$，则由概率性质可知有

$$P\{x_1 < X \leqslant x_2\} = P\{X \leqslant x_2\} - P\{X \leqslant x_1\} = F(x_2) - F(x_1) \tag{2.2.2}$$

因此，若已知随机变量 X 的分布函数，就可以知道 X 落在任一区间 $(x_1, x_2]$ 的概率，从这个意义上说，分布函数刻画了随机变量 X 的统计规律.

例 2.2.1　某工厂生产的显像管的寿命 X(单位：万小时) 是一随机变量，其分布函数为

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - \mathrm{e}^{-x/2}, & x \geqslant 0. \end{cases}$$

求显像管的寿命超过 2 万小时的概率及寿命超过 2 万小时但不超过 4 万小时的概率.

解　显像管寿命超过 2 万小时的概率为

$$P(X > 2) = 1 - P(X \leqslant 2) = 1 - F(2) = 1 - (1 - \mathrm{e}^{-2/2}) = \mathrm{e}^{-1} = 0.3679.$$

超过 2 万小时但不超过 4 万小时的概率是

$$\begin{aligned} P(2 < X \leqslant 4) &= P(X \leqslant 4) - P(X \leqslant 2) = F(4) - F(2) \\ &= (1 - \mathrm{e}^{-4/2}) - (1 - \mathrm{e}^{-2/2}) = \mathrm{e}^{-1} - \mathrm{e}^{-2} \\ &= 0.368 - 0.135 = 0.233. \end{aligned}$$

分布函数 $F(x)$ 具有如下性质.

性质 1　单调不减性：若 $x_1 < x_2$，则 $F(x_1) \leqslant F(x_2)$.

性质 2　右连续性：$\lim\limits_{x \to x_0^+} F(x) = F(x_0)$，或记为 $F(x_0 + 0) = F(x_0)$.

性质 3　$0 \leqslant F(x) \leqslant 1$，且 $F(-\infty) = \lim\limits_{x \to -\infty} F(x) = 0$，$F(+\infty) = \lim\limits_{x \to +\infty} F(x) = 1$.

可以证明，若某函数 $F(x)$ 满足上述性质，则它一定是某随机变量的分布函数. 例如下列函数

$$F(x) = \begin{cases} 1 - \mathrm{e}^{-x}, & x \geqslant 0, \\ 0, & x < 0, \end{cases}$$

是一个分布函数.

2.3　离散型随机变量

2.3.1　离散型随机变量的分布律

若随机变量的所有可能取值是有限的或可数无限的，则称其为**离散型随机变量**. 对于离散型随机变量，样本点个数是有限或可数无限的，因此只要知道了每个样本点的概率，就可以求出任何随机事件的概率.

设离散型随机变量 X 的所有可能取值为 $x_k(k=1,2,\cdots)$，而 X 取值 x_k 的概率为 p_k，即

$$P\{X=x_k\}=p_k,\quad k=1,2,\cdots. \tag{2.3.1}$$

称(2.3.1)式为离散型随机变量 X 的概率分布或**分布律**，常用表格的形式表示为

x_k	x_1	x_2	$\cdots$	x_k	$\cdots$
$P\{X=x_k\}$	p_1	p_2	$\cdots$	p_k	$\cdots$

以横坐标表示随机变量的可能取值，纵坐标表示随机变量取这些值的概率，并用折线把这些点连接起来，就得到概率分布图，如图 2.1 所示.

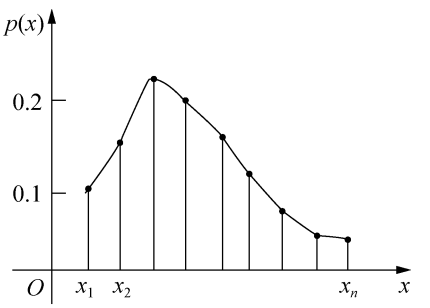


图 2.1　概率分布图

若 X 的分布律为式(2.3.1)，则 X 的分布函数为

$$F_X(x)=P\{X\leqslant x\}=\sum_{x_k\leqslant x}P\{X=x_k\}=\sum_{x_k\leqslant x}p_k,$$
$$-\infty<x<\infty.$$

容易证明，式(2.3.1)中的分布律 p_k，$k=1,2,\cdots$ 满足下列性质.

性质 1　$0\leqslant p_k\leqslant 1$，$k=1,2,\cdots$；

性质 2　$\sum\limits_{k=1}^{\infty}p_k=1$.

例 2.3.1　设袋中有标号为 -1，1，1，2，2，2 的六个球，从中任取一个球，求所取球的标号数 X 的分布律和分布函数 $F(x)$，并画出 $F(x)$ 的图像.

解　X 是一维离散型随机变量，它的可能取值是 -1，1，2，X 的分布律为

X	-1	1	2
p_k	$p_1=\frac{1}{6}$	$p_2=\frac{2}{6}$	$p_3=\frac{3}{6}$

因为 $F(x)=P\{X\leqslant x\}$，所以，当 $x<-1$ 时，$F(x)=P\{X\leqslant x\}=P(\varnothing)=0$；

当 $-1\leqslant x<1$ 时，$F(x)=P\{X=-1\}=\frac{1}{6}$；当 $1\leqslant x<2$ 时，

$$F(x)=P\{X=-1\}+P\{X=1\}=\frac{1}{6}+\frac{2}{6}=\frac{1}{2};$$

当 $x \geqslant 2$ 时，

$$F(x) = P\{X = -1\} + P\{X = 1\} + P\{X = 2\} = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} = 1.$$

综上得

$$F(x) = \begin{cases} 0, & x < -1, \\ \dfrac{1}{6}, & -1 \leqslant x < 1, \\ \dfrac{1}{2}, & 1 \leqslant x < 2, \\ 1, & x \geqslant 2. \end{cases}$$

$F(x)$ 的图形如图 2.2 所示.

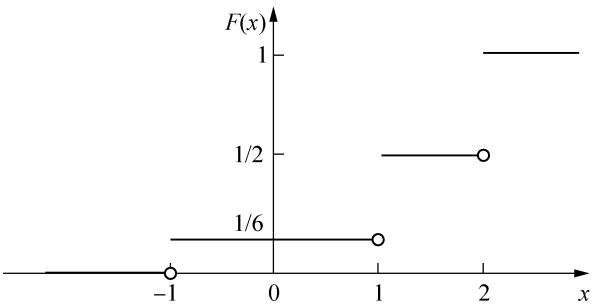


图 2.2　分布函数图

2.3.2　常见的离散型随机变量

1. 两点分布

如果随机变量 X 只取 0 和 1，其分布律为

$$P\{X = 1\} = p, \quad P\{X = 0\} = q \tag{2.3.2}$$

其中 $0 < p < 1$，$q = 1 - p$，则称 X 服从参数为 p 的**两点分布**或**伯努利分布**，记为 $X \sim B(1, p)$.

例 2.3.2　设射击一次命中率为 0.4. 若用 $X = 1$ 表示“中”，用 $X = 0$ 来表示“不中”，则

$$P\{X = 1\} = 0.4, \quad P\{X = 0\} = 0.6.$$

即 X 服从参数为 0.4 的两点分布.

2. 二项分布

在 n 次随机试验中，若每次试验结果的出现不依赖于其他各次试验的结果，则称这 n 次试验相互独立.

设试验 E 只有两个结果 A 和 $\bar{A}$，则称其为伯努利(Bernoulli) 试验. 记 $p = P(A)$，则 $P(\bar{A}) = 1 - p$，即事件 A 发生的概率为 p，不发生的概率为 $q = 1 - p$. 将试验 E 独立地重复进行 n 次，称这 n 次独立重复试验为 n 重伯努利试验.

在 n 重伯努利试验中，事件 A 恰好发生 $k(0 \leqslant k \leqslant n)$ 次的概率记为 $P\{X = k\}$. 因为各次试验是相互独立的，所以事件 A 在指定的 k 次试验中发生，在其余 $n - k$ 次试验中不发生的概率为 $p^k q^{n-k}$，由于这种指定的方式共有 C_n^k 种，故在 n 次伯努利试验中事件 A 恰好发生

k 次的概率为

$$p_k = P\{X = k\} = C_n^k p^k q^{n-k},\ k = 0,\ 1,\ 2,\ \cdots,\ n \tag{2.3.3}$$

其中 $q = 1-p$. 显然 $p_k \geqslant 0$，$k = 0,\ 1,\ 2,\ \cdots,\ n$，且

$$\sum_{k=0}^{n} C_n^k p^k q^{n-k} = (p+q)^n = 1.$$

注意到 $C_n^k p^k q^{n-k}$ 恰好是二项式 $(p+q)^n$ 的展开式中的一般项，所以称满足(2.3.3)式的随机变量 X 服从参数为 n，p 的**二项分布**，记为 $X \sim B(n,\ p)$.

特别地，当 $n = 1$ 时，二项分布就化为两点分布 $B(1,\ p)$，其分布律式(2.3.2)可表示为

$$P\{X = k\} = p^k q^{1-k},\ k = 0,\ 1 \tag{2.3.4}$$

例 2.3.3 某出租汽车公司共有出租车 400 辆，设每天每辆出租车出现故障的概率为 0.02，试求

(1) 一天内有不超过 2 辆出租车出现故障的概率；

(2) 一天内没有出租车出现故障的概率.

解 将观察一辆出租车一天内是否出现故障看成一次试验. 因为每辆出租车是否出现故障与其他出租车是否出现故障无关，于是观察 400 辆出租车是否出现故障就是做 400 次伯努利试验. 设 X 是每天内出现故障的出租车数，则 $X \sim B(400,\ 0.02)$.

(1) 一天内有不超过 2 辆出租车出现故障的概率为

$$\begin{aligned} P\{X \leqslant 2\} &= P\{X = 0\} + P\{X = 1\} + P\{X = 2\} \\ &= 0.98^{400} + 400 \times 0.02 \times 0.98^{399} + C_{400}^2 \times 0.02^2 \times 0.98^{398} \approx 0.013\ 1. \end{aligned}$$

(2) 一天内没有出租车出现故障的概率为

$$P\{X = 0\} = 0.98^{400} \approx 0.000\ 309.$$

例 2.3.4 波兰数学家巴拿赫随身带着两盒火柴，分别放在他的左、右两个衣袋里，每盒有 n 根火柴，他需要火柴时，便随机地从其中一盒中取出一根. 试求他发现其中一盒已空而另一盒中剩下的火柴根数 X 的分布律.

解 设 A 为事件“取左衣袋中的一盒”，显然有 $P(A) = P(\bar{A}) = \dfrac{1}{2}$. 把每取一次火柴看成一次伯努里试验. 当发现左边一盒空而右边一盒剩 k 根时，共做了 $(n+1)+(n-k) = 2n-k+1$ 次伯努利试验，其中 A 发生了 $n+1$ 次，$\bar{A}$ 发生了 $n-k$ 次，其概率为

$$C_{2n-k+1}^{n+1}\,[P(A)]^{n+1}\,[P(\bar{A})]^{n-k} = C_{2n-k+1}^{n+1}\left(\frac{1}{2}\right)^{n+1}\left(\frac{1}{2}\right)^{n-k}.$$

由对称性，发现右边一盒空而左边一盒剩 k 根的概率也是 $C_{2n-k+1}^{n+1}\left(\dfrac{1}{2}\right)^{n+1}\left(\dfrac{1}{2}\right)^{n-k}$，故 X 的分布律为

$$P(X = k) = 2C_{2n-k+1}^{n+1}\left(\frac{1}{2}\right)^{n+1}\left(\frac{1}{2}\right)^{n-k} = C_{2n-k+1}^{n+1}\left(\frac{1}{2}\right)^{2n-k},\ k = 0,\ 1,\ \cdots,\ n.$$

3. 泊松分布

如果随机变量 X 的分布律为

$$P\{X = k\} = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda},\ k = 0,\ 1,\ 2,\ \cdots \tag{2.3.5}$$

其中 $\lambda>0$ 为常数，则称 X 服从参数为 λ 的**泊松分布**，记为 $X\sim P(\lambda)$.

易见，$P\{X=k\}>0$，$k=0,1,2,\cdots$，且有

$$\sum_{k=0}^{\infty}P\{X=k\}=\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}=e^{-\lambda}e^{\lambda}=1.$$

在许多实际问题中，我们所关心的量近似服从泊松分布. 例如，某医院每天前来就诊的病人数；某地区一段时间间隔内发生火灾的次数，或发生交通事故的次数；牧草种子中的杂草种子数；某地区一年内发生暴雨的次数等，都可以用泊松分布来描述.

例 2.3.5　某商店出售某种贵重商品，根据以往经验，每月销售量 X 服从参数为 $\lambda=3$ 的泊松分布. 问在月初进货时，要库存多少件此种商品，才能以 99% 的概率满足顾客的需要？

解　X 的分布律为

$$P\{X=i\}=\frac{3^i}{i!}e^{-3},\ i=0,1,2,\cdots.$$

设月初库存 k 件，则由题意知

$$P\{X\leqslant k\}=\sum_{i=0}^{k}\frac{3^i}{i!}e^{-3}\geqslant 0.99.$$

查本书后面的泊松分布表得 $k=8$，即月初进货时，库存 8 件这种商品，才能以 99% 的概率满足顾客的需要.

在应用中，利用泊松分布近似计算二项分布的概率常常会很方便. 尤其当 n 充分大，p 又很小时，二项分布 $B(n,p)$ 可以用泊松分布 $P(\lambda)$ 来近似，即有近似公式

$$C_n^k p^k(1-p)^{n-k}\approx\frac{\lambda^k}{k!}e^{-\lambda},\ k=0,1,2,\cdots,n \tag{2.3.6}$$

其中 $\lambda=np$.

例 2.3.6　设每次射击时击中目标的概率为 0.001，如果射击 5 000 次，试求至少两次击中目标的概率.

解　设击中目标的次数为 X，则 $X\sim B(5\,000,0.001)$，所求概率为

$$\begin{aligned}P\{X\geqslant 2\}&=1-P\{X=0\}-P\{X=1\}\\&=1-(1-0.001)^{5\,000}-5\,000\times 0.001\times(1-0.001)^{4\,999}\\&=0.959\,640.\end{aligned}$$

下面用近似公式(2.3.6)求上述概率的近似值，这时 $\lambda=5\,000\times 0.001=5$，所以

$$P\{X=0\}\approx e^{-5},\ P\{X=1\}\approx 5e^{-5}.$$

从而

$$P\{X\geqslant 2\}\approx 1-6e^{-5}=0.959\,576.$$

例 2.3.7(人寿保险问题)　有 2 000 个同一年龄的人购买了某保险公司的人寿保险. 每个投保人在 1 月 1 日付保费 800 元，如果投保人在当年死亡，则保险公司必须向投保人的家属支付 200 000 元的赔费. 设在投保的当年每个投保人死亡的概率是 0.002. 求到年底时保险公司亏本的概率 p_1 和保险公司获利不少于 400 000 元的概率 p_2.

解　设每年的死亡人数为 X，则 $X\sim B(2\,000,0.002)$. 又保险公司一年的总收入为 $2\,000\times 800=1\,600\,000$(元)，总支出为 $200\,000X$(元). 故

$$p_1 = P\{200\ 000X > 1\ 600\ 000\} = P\{X > 8\} = 1 - P\{X \leqslant 8\}$$

$$= 1 - \sum_{k=0}^{8} C_{2\ 000}^{k} \times 0.002^k \times 0.998^{2\ 000-k} = 0.021\ 24,$$

$$p_2 = P\{1\ 600\ 000 - 200\ 000X \geqslant 400\ 000\} = P\{X \leqslant 6\}$$

$$= \sum_{k=0}^{6} C_{2\ 500}^{k} \times 0.002^k \times 0.998^{2\ 500-k} = 0.889\ 33.$$

下面用近似公式(2.3.6)计算，这时 $\lambda = np = 2\ 000 \times 0.002 = 4$，

$$p_1 \approx 1 - \sum_{k=0}^{8} \frac{\mathrm{e}^{-4}}{k!} 4^k = 0.021\ 36,$$

$$p_2 \approx \sum_{k=0}^{6} \frac{\mathrm{e}^{-4}}{k!} 4^k = 0.889\ 33.$$

4. 几何分布

如果随机变量 X 只取正整数值 1，2，…，且其分布律为

$$P\{X = k\} = q^{k-1}p,\ k = 1,\ 2,\ \cdots \tag{2.3.7}$$

其中 $0 < p < 1$，$q = 1 - p$，则称 X 服从参数为 p 的**几何分布**，记为 $X \sim Ge(p)$.

例 2.3.8 设有独立重复试验序列，事件 A 在每次试验中发生的概率为 p. 设 X 为 A 首次发生时的试验次数，即 $\{X=k\}$ 为事件“A 在第 k 次试验中发生，而在前面的 $k-1$ 次试验中均不发生”. 则

$$P(X = k) = q^{k-1}p,\ k = 1,\ 2,\ \cdots,$$

即 $X \sim Ge(p)$.

***5. 负二项分布**

负二项分布是几何分布的一种延伸，亦称为**巴斯卡分布**.

在独立重复地伯努利试验中，记每次试验事件 A 出现的概率为 p，如果 X 为事件 A 第 r 次出现时的试验次数，则 X 的所有可能取值为 r，$r+1$，…，称 X 服从负二项分布或巴斯卡分布，其分布律为

$$P(X = k) = C_{k-1}^{r-1} p^r (1-p)^{k-r},\ k = r,\ r+1,\ \cdots \tag{2.3.8}$$

记为 $X \sim Nb(r,\ p)$. 当 $r = 1$ 时，即为几何分布.

这是因为在 k 次伯努利试验中，最后一次一定是 A，而前 $k-1$ 次中 A 出现 $r-1$ 次，由二项分布知其概率为 $C_{k-1}^{r-1} p^{r-1} (1-p)^{k-r}$，再乘以最后一次出现 A 的概率 p 即得(2.3.8)式.

如果将第一个 A 出现的试验次数记为 X_1，第二个 A 出现的试验次数(从第一个 A 出现之后算起)记为 X_2，…，第 r 个 A 出现的试验次数(从第 $r-1$ 个 A 出现之后算起)记为 X_r 则 X_i 独立同分布，且 $X_i \sim Ge(p)$. 此时有 $X = X_1 + X_2 + \cdots + X_r \sim Nb(r,\ p)$，即负二项分布的随机变量可以表达为 r 个独立同时服从几何分布的随机变量之和.

***6. 超几何分布**

第一章中由式(1.3.2)给出的超几何概率称为超几何分布.

设有 N 件产品，其中有 M 件不合格品，有 $N-M$ 件合格品，若从中不放回地随机抽取 n 件，则其中含有的不合格品件数 X 服从超几何分布，记为 $X \sim h(n,\ N,\ M)$. 由式(1.3.2)可知，超几何分布的分布律为

$$P(X=k)=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n},\ k=1,2,\cdots,r \tag{2.3.9}$$

其中 $r=\min\{M,n\}$，且 $M\leqslant N$，$n\leqslant N$，n，M，N 均为正整数.

若要验证式(2.3.9)给出的确实为一个分布律，只需利用1.3.1节中的常用排列组合公式(1.3.3)，其中取 $m=n$，$n_1=M$，$n_2=N-M$，即可得到

$$\sum_{k=1}^{r} C_M^k C_{N-M}^{n-k}=C_N^n$$

由此可知式(2.3.9)中的概率之和为1.

2.4　连续型随机变量

2.4.1　密度函数

定义2.4.1　设 $F(x)=P\{X\leqslant x\}$ 是随机变量 X 的分布函数，若存在非负函数 $f(x)$，使 $F(x)$ 表示为下列变上限积分

$$F(x)=\int_{-\infty}^{x} f(t)\,\mathrm{d}t \tag{2.4.1}$$

则称 X 为连续型随机变量，并称 $f(x)$ 为 X 的概率密度函数，简称为**密度函数**.

若 $f(x)$ 是随机变量 X 的密度函数，则对任意固定的 x 及任意的 $\Delta x>0$，有

$$\frac{P\{x<X\leqslant x+\Delta x\}}{\Delta x}=\frac{F(x+\Delta x)-F(x)}{\Delta x}=\frac{1}{\Delta x}\int_{x}^{x+\Delta x} f(t)\,\mathrm{d}t \tag{2.4.2}$$

上式左端为随机变量 X 落在区间 $(x,\ x+\Delta x]$ 上的平均概率，如果 $f(x)$ 在 x 处连续，则

$$\lim_{\Delta x\to 0}\frac{P\{x<X\leqslant x+\Delta x\}}{\Delta x}=F'(x)=f(x).$$

从这里我们看到，密度函数的定义与物理学中线密度的定义极其类似. 这就是我们将 $f(x)$ 称为密度函数的原因. 若不计高阶无穷小，则由上式可得

$$P\{x<X\leqslant x+\Delta x\}\approx f(x)\Delta x,$$

它表明随机变量 X 落在区间 $(x,\ x+\Delta x]$ 上的概率近似等于 $f(x)\Delta x$.

密度函数 $f(x)$ 有下列性质：

性质1　$f(x)\geqslant 0$；

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

性质3　$P\{a<X\leqslant b\}=\int_{a}^{b} f(t)\,\mathrm{d}t$；

性质4　当 x 是 $f(x)$ 的连续点时，有 $F'(x)=f(x)$.

任意函数 $f(x)$，若它满足上述性质1和性质2，则它一定是某连续型随机变量的密度函数.

可以证明，若 X 是连续型随机变量，则对于任意实数 a，总有

$$P\{X=a\}=0 \tag{2.4.3}$$

这个结果说明，连续型随机变量取任意一点的概率为零. 同时也说明，由 $P(A)=0$，并不能推出 A 是不可能事件. 因为虽然 $P\{X=a\}=0$，但事件 $\{X=a\}$ 并非是不可能事件.

由(2.4.3)式可知，连续型随机变量 X 落在区间 (a,b)，$[a,b)$，$(a,b]$，$[a,b]$ 上的概率都相等，即有

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

它们等于由曲线 $y=f(x)$ 和直线 $x=a$，$x=b$ 及 $y=0$ 所围成的曲边梯形的面积，如图 2.3 所示.

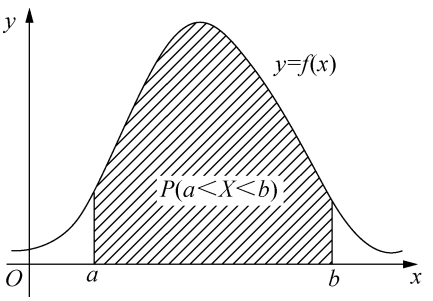


图 2.3　概率的几何意义

例 2.4.1　设 $f(x)=\begin{cases}\dfrac{x}{a}\mathrm{e}^{-\frac{x^2}{2a}}, & \\ 0, & \end{cases}$ 其中 $a>0$ 为已知实数.

(1) 证明 $f(x)$ 是某随机变量 X 的密度函数，

(2) 求 X 的分布函数 $F(x)$，

(3) 求概率 $P\{0\leqslant X\leqslant 1\}$.

解　(1) 显然 $f(x)\geqslant 0$，它满足密度函数的性质 1. 又

$$\int_{-\infty}^{+\infty}f(x)\,\mathrm{d}x=\int_{-\infty}^{0}0\mathrm{d}x+\int_0^{+\infty}\frac{x}{a}\mathrm{e}^{-\frac{x^2}{2a}}\mathrm{d}x=\int_0^{+\infty}\mathrm{e}^{-\frac{x^2}{2a}}\mathrm{d}\left(\frac{x^2}{2a}\right)=-\mathrm{e}^{-\frac{x^2}{2a}}\Big|_0^{+\infty}=1,$$

即 $f(x)$ 满足性质 2，因此 $f(x)$ 是某连续型随机变量 X 的密度函数.

(2) 当 $x<0$ 时，$f(x)=0$，$F(x)=\int_{-\infty}^{x}f(t)\,\mathrm{d}t=\int_{-\infty}^{x}0\mathrm{d}t=0$；当 $x\geqslant 0$ 时，$f(x)=\dfrac{x}{a}\mathrm{e}^{-\frac{x^2}{2a}}$，

$$\begin{aligned}F(x)&=\int_{-\infty}^{x}f(t)\,\mathrm{d}t=\int_{-\infty}^{0}f(t)\,\mathrm{d}t+\int_0^x f(t)\,\mathrm{d}t\\&=\int_{-\infty}^{0}0\mathrm{d}t+\int_0^x\frac{t}{a}\mathrm{e}^{-\frac{t^2}{2a}}\mathrm{d}t=-\mathrm{e}^{-\frac{t^2}{2a}}\Big|_0^x=1-\mathrm{e}^{-\frac{x^2}{2a}}.\end{aligned}$$

综上得

$$F(x)=\begin{cases}0, & x<0,\\ 1-\mathrm{e}^{-\frac{x^2}{2a}}, & x\geqslant 0.\end{cases}$$

(3) $P\{0\leqslant X\leqslant 1\}=F(1)-F(0)=1-\mathrm{e}^{-\frac{1}{2a}}$. 该结果也可以通过对密度函数的积分直接得到，即

$$P\{0\leqslant X\leqslant 1\}=\int_0^1 f(x)\,\mathrm{d}x=\int_0^1\frac{x}{a}\mathrm{e}^{-\frac{x^2}{2a}}\mathrm{d}t=-\mathrm{e}^{-\frac{x^2}{2a}}\Big|_0^1=1-\mathrm{e}^{-\frac{1}{2a}}.$$

2.4.2　常见的连续型随机变量

1. 均匀分布

如果随机变量 X 的密度函数为

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\leqslant x\leqslant b,\\ 0, & \text{其他.}\end{cases}\tag{2.4.4}$$

则称 X 服从$[a,\ b]$ 区间上的**均匀分布**，记作 $X\sim U[a,\ b]$，其中，a，$b(a<b)$ 为常数.

均匀分布的分布函数为

$$F(x)=P\{X\leqslant x\}=\int_{-\infty}^{x}f(t)\,\mathrm{d}t=\begin{cases}0, & x<a,\\ \dfrac{x-a}{b-a}, & a\leqslant x<b,\\ 1, & x\geqslant b.\end{cases}\tag{2.4.5}$$

如果 $X\sim U[a,\ b]$，则对任意满足 $a\leqslant c<d\leqslant b$ 的 c，d，总有

$$P\{c\leqslant X\leqslant d\}=\int_{c}^{d}f(x)\,\mathrm{d}x=\frac{d-c}{b-a}.$$

这表明，X 落在$[a,\ b]$ 的子区间$[c,\ d]$ 上的概率，只与子区间的长度$(d-c)$ 有关(成正比)，而与子区间$[c,\ d]$ 在区间$[a,\ b]$ 中的具体位置无关.

均匀分布是理论和应用中常用的一种分布. 当我们对取值在某一区间$[a,\ b]$ 上的随机变量 X 的分布情况不清楚时，一般可以假定它服从均匀分布 $U[a,\ b]$.

例 2.4.2　设随机变量 X 在区间$[2,\ 5]$ 上服从均匀分布，现对 X 进行三次独立观测，试求至少有两次观测值大于 3 的概率.

解　X 的密度函数为

$$f(x)=\begin{cases}\dfrac{1}{3}, & 2\leqslant x\leqslant 5,\\ 0, & \text{其他.}\end{cases}$$

设 A 为事件“X 的观察值大于 3”，则

$$P(A)=P\{X>3\}=\int_{3}^{5}\frac{1}{3}\mathrm{d}x=\frac{2}{3}.$$

设 Y 为三次独立观测中观测值大于 3 的次数，即事件 A 发生的次数，则 Y 服从二项分布 $B(3,\ 2/3)$，于是

$$P\{Y\geqslant 2\}=C_3^2\left(\frac{2}{3}\right)^2\left(1-\frac{2}{3}\right)^{3-2}+C_3^3\left(\frac{2}{3}\right)^3\left(1-\frac{2}{3}\right)^{3-3}=\frac{20}{27}.$$

2. 指数分布

如果随机变量 X 的密度函数为

$$f(x)=\begin{cases}\lambda\mathrm{e}^{-\lambda x}, & x>0,\\ 0, & x\leqslant 0,\end{cases}\tag{2.4.6}$$

其中 $\lambda>0$，则称 X 服从参数为 λ 的**指数分布**，记为 $X\sim E(\lambda)$.

下面求指数分布的分布函数 $F(x)$.

若 $x<0$，则 $F(x)=\int_{-\infty}^{x}f(t)\,\mathrm{d}t=\int_{-\infty}^{x}0\mathrm{d}t=0$；若 $x\geqslant 0$，则

$$F(x)=\int_{-\infty}^{x}f(t)\,\mathrm{d}t=\int_{-\infty}^{0}f(t)\,\mathrm{d}t+\int_{0}^{x}f(t)\,\mathrm{d}t$$

$$=\int_{-\infty}^{0}0\mathrm{d}t+\int_{0}^{x}\lambda\mathrm{e}^{-\lambda t}\mathrm{d}t=1-\mathrm{e}^{-\lambda x}.$$

综上即得

$$F(x)=\begin{cases}1-\mathrm{e}^{-\lambda x}, & x\geqslant 0,\\ 0, & x<0.\end{cases}\tag{2.4.7}$$

例 2.4.3 设某型号电子管的使用寿命 X 服从参数为 $\lambda=0.001$ 的指数分布，试计算概率 $P\{1000<X\leqslant 1200\}$.

解 X 的密度函数为

$$f(x)=\begin{cases}0.001\mathrm{e}^{-0.001x}, & x>0,\\ 0, & x\leqslant 0.\end{cases}$$

所求概率为

$$P\{1000<X\leqslant 1200\}=\int_{1\,000}^{1\,200}f(x)\,\mathrm{d}x=\int_{1\,000}^{1\,200}0.001\mathrm{e}^{-0.001x}\mathrm{d}x$$

$$=\mathrm{e}^{-1}-\mathrm{e}^{-1.2}=0.066\,7.$$

在实际应用中，指数分布常用来描述各种“寿命”的近似分布. 例如，无线电元件的寿命，动物的寿命，电话系统中的通话时间，随机服务系统中的服务时间等都常用指数分布来近似.

指数分布具有“无记忆性”. 设 $X\sim E(\lambda)$，则对任意的 $s>0$，$t>0$，有

$$P\{X>s+t\mid X>s\}=\frac{P\{X>s+t\}}{P\{X>s\}}=\frac{1-P\{X\leqslant s+t\}}{1-P\{X\leqslant s\}}$$

$$=\frac{1-F(s+t)}{1-F(s)}=\frac{\mathrm{e}^{-\lambda(s+t)}}{\mathrm{e}^{-\lambda s}}=\mathrm{e}^{-\lambda t}=F(X>t).$$

若把 X 解释为寿命，则上式表明，如果已知寿命大于 s 年，则寿命大于 $s+t$ 年的概率只与时间 t 有关，而与目前年龄 s 无关，此即“无记忆性”.

3. 正态分布

正态分布是最重要的一种分布，大量随机变量都服从或近似地服从正态分布. 例如，某零件长度或直径的测量误差，炮弹的弹着点距目标的距离，某族群人体的身高、体重，飞机材料的疲劳应力等，都服从或近似服从正态分布. 可以说，正态分布是自然界和社会现象中最常见的一种分布. 一般来说，如果一个随机变量是由大量微小的、独立的随机因素叠加而成的，则它近似地服从正态分布. 因此，正态分布在理论和实际应用中有着极其重要的作用.

设随机变量 X 的密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad -\infty<x<\infty\tag{2.4.8}$$

其中 σ 和 μ 都是参数，$\sigma>0$，$\mu\in(-\infty,+\infty)$，则称 X 服从**正态分布**，记为 $X\sim N(\mu,\sigma^2)$.

正态分布的密度函数 $f(x)$ 如图 2.4 所示.

从式(2.4.8)和图2.4可知，正态密度函数 $f(x)$ 的图形呈钟形，且有如下特征.

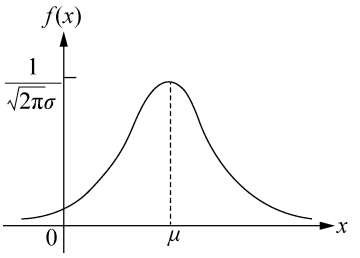


图 2.4　正态密度函数

(1) 关于直线 $x=\mu$ 对称；

(2) 在 $x=\mu$ 处取得最大值 $\dfrac{1}{\sqrt{2\pi}\sigma}$；

(3) 在 $x=\mu\pm\sigma$ 处有拐点；

(4) 当 $|x|\to\infty$ 时，曲线以 x 轴为渐近线；

(5) 如果固定 σ，改变 μ 的值，则图形沿着 x 平移，而图形的形状不变，如图 2.5 所示. 如果固定 μ，改变 σ 值，则最大值 $f(\mu)=\dfrac{1}{\sqrt{2\pi}\sigma}$ 随 σ 的增大而减小，因此随着 σ 的增大，图形变得越来越扁，但图形的对称轴没有改变，如图 2.6 所示.

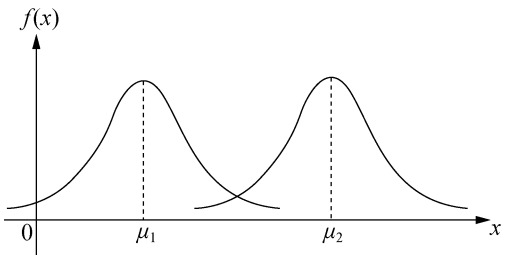


图 2.5　位置参数

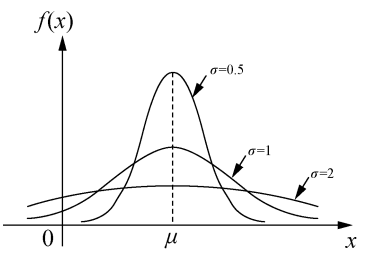


图 2.6　尺度参数

此外，由于正态密度函数 $f(x)$ 关于直线 $x=\mu$ 对称，则对任意 $h>0$，有 $P\{\mu-h<X\leqslant\mu\}=P\{\mu<X\leqslant\mu+h\}$，即两块曲边梯形面积相等，如图 2.7 所示.

正态分布的分布函数为

$$F(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt \tag{2.4.9}$$

它的图形如图 2.8 所示.

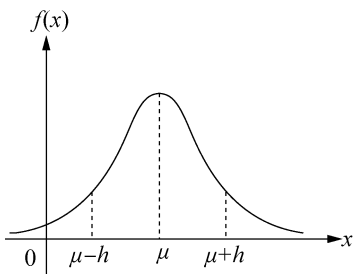


图 2.7　正态分布的对称性

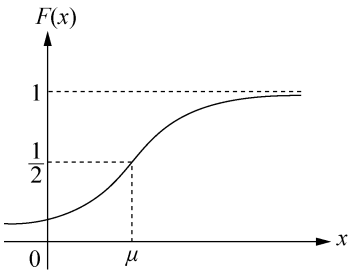


图 2.8　正态分布函数

特别地，当 $\mu=0$，$\sigma=1$ 时，称正态分布 $N(0,1)$ 为**标准正态分布**. 对于标准正态分布 $N(0,1)$，其密度函数通常用 $\varphi(x)$ 来表示，即

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}},\quad -\infty<x<\infty \tag{2.4.10}$$

它的图形如图 2.9 所示的曲线 $\varphi(x)$. 可以证明 $\varphi(x)$ 满足密度函数的两条性质，即非负性 $\varphi(x)\geqslant 0$ 和

$$\int_{-\infty}^{+\infty}\varphi(x)\mathrm{d}x=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x=1.$$

标准正态分布 $N(0,1)$ 的分布函数为

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t,\quad -\infty<x<\infty \tag{2.4.11}$$

本书后面的附表 2 中列出了 $\Phi(x)$ 的部分值，可供查用.

由密度函数 $\varphi(x)$ 的对称性可知，分布函数 $\Phi(x)$ 满足如下重要性质

$$\Phi(-x)=1-\Phi(x). \tag{2.4.12}$$

图 2.9 直观地显示了 $\Phi(x)$ 的这个性质，其中左边阴影部分面积为 $\Phi(-x)$，右边阴影部分是 $1-\Phi(x)$，上曲边为密度函数 $\varphi(x)$，且由式(2.4.12) 和图 2.9 不难知道 $\Phi(0)=0.5$，因为它正好等于整个曲边梯形面积的一半.

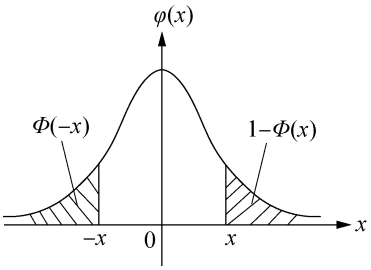


图 2.9　标准正态分布的密度函数

例 2.4.4　设 $X\sim N(0,1)$，求 $P\{-1<X<2\}$.

解
$$\begin{aligned}P\{-1<X<2\}&=P\{X<2\}-P\{X\leqslant -1\}\\&=\Phi(2)-\Phi(-1)\\&=\Phi(2)-[1-\Phi(1)]\\&=\Phi(2)+\Phi(1)-1\\&=0.977\ 3+0.841\ 3-1\\&=0.818\ 6.\end{aligned}$$

在一般正态分布下，随机变量落在一个区间内的概率的计算一般通过标准正态分布函数值来计算. 下面定理给出了一般正态分布函数与标准正态分布函数之间的关系.

定理 2.4.1　设 $X\sim N(\mu,\sigma^2)$，X 的分布函数为 $F(x)$，标准正态分布 $N(0,1)$ 的分布函数为 $\Phi(x)$，则

$$F(x)=\Phi\left(\frac{x-\mu}{\sigma}\right) \tag{2.4.13}$$

证
$$F(x)=P\{X\leqslant x\}=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}\mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}}\mathrm{d}t$$

$$\xrightarrow{\text{令}\frac{t-\mu}{\sigma}=u}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x-\mu}{\sigma}}\mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u=\Phi\left(\frac{x-\mu}{\sigma}\right).$$

由定理 2.4.1，若 $X\sim N(\mu,\sigma^2)$，则事件 $\{a<X<b\}$ 的概率为

$$P\{a<X<b\}=F(b)-F(a)=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right) \tag{2.4.14}$$

例 2.4.5　设 $X\sim N(1.5,4)$，试计算概率 $P\{|X|<3\}$.

解　所求概率为

$$P\{|X|<3\}=P\{-3<X<3\}=F(3)-F(-3)$$
$$=\Phi\left(\frac{3-1.5}{2}\right)-\Phi\left(\frac{-3-1.5}{2}\right)=\Phi(0.75)-\Phi(-2.25)$$
$$=\Phi(0.75)-[1-\Phi(2.25)]=\Phi(0.75)+\Phi(2.25)-1$$
$$=0.773\ 4+0.987\ 8-1=0.761\ 2.$$

例 2.4.6　已知某台机器生产的螺栓长度 X(单位：厘米) 服从参数为 $\mu=10.05$，$\sigma=0.06$ 的正态分布. 规定螺栓长度在 10.05 ± 0.12 内为合格品，试求螺栓为合格品的概率.

解　已知螺栓长度 $X\sim N(10.05,\ 0.06^2)$，记 $a=10.05-0.12$，$b=10.05+0.12$，则 $\{a\leqslant X\leqslant b\}$ 表示螺栓为合格品，其概率为

$$P\{a\leqslant X\leqslant b\}=\Phi\left(\frac{0.12}{0.06}\right)-\Phi\left(\frac{-0.12}{0.06}\right)=\Phi(2)-\Phi(-2)$$
$$=\Phi(2)-[1-\Phi(2)]=2\Phi(2)-1$$
$$=2\times0.977\ 2-1=0.9544.$$

若 $X\sim N(\mu,\ \sigma^2)$，则可从标准正态分布的分布函数表中查出

$$P\{|X-\mu|<\sigma\}=2\Phi(1)-1\approx68.27\%,$$
$$P\{|X-\mu|<2\sigma\}=2\Phi(2)-1\approx95.45\%,$$
$$P\{|X-\mu|<3\sigma\}=2\Phi(3)-1\approx99.73\%.$$

可见在一次试验中，服从正态分布的随机变量 X 基本上在区间 $(\mu-2\sigma,\ \mu+2\sigma)$ 内取值，而且几乎总是落在区间 $(\mu-3\sigma,\ \mu+3\sigma)$ 内. 这个性质在标准制定和质量管理等方面有着广泛的应用，通常称为“3σ 原则”.

注意到 $\Phi(-\infty)=0$ 和 $\Phi(\infty)=1$，可知当 (2.4.14) 式中的 a 和 b 之一为无穷时，定理 2.4.1 仍成立. 这时(2.4.14) 式分别变成

$$P\{X\leqslant b\}=P\{-\infty<X\leqslant b\}=\Phi\left(\frac{b-\mu}{\sigma}\right),$$
$$P\{X>a\}=P\{a<X<\infty\}=1-\Phi\left(\frac{a-\mu}{\sigma}\right).$$

例 2.4.7　假设某地区成年男性的身高(单位：厘米) $X\sim N(170,\ 7.69^2)$，求该地区成年男性的身高超过 175 厘米的概率.

解　由题意，$\{X>175\}$ 表示该地区成年男性的身高超过 175 厘米这一事件，其概率为

$$P\{X>175\}=1-P\{175\leqslant X\}=1-\Phi\left(\frac{175-170}{7.69}\right)$$
$$=1-\Phi(0.65)=1-0.742\ 2=0.257\ 8.$$

2.5　一维随机变量函数的分布

在许多实际问题中，常常需要考虑随机变量的函数的分布. 例如，我们能直接测量到圆轴正截面的直径 D，而所关心的却是该截面的面积 $A=\pi D^2/4$，它是随机变量 D 的函数.

在本节中，我们讨论如何由已知的随机变量 X 的分布，求 X 的函数 $Y=g(X)$ 的分布.

2.5.1　离散型随机变量函数的分布

设离散型随机变量 X 的分布律为 $p_k=P\{X=x_k\}$，$k=1,2,\cdots$，$g(x)$ 是一个单值函数. 令 $Y=g(X)$，则 Y 也是一个离散型随机变量，它的分布律容易由 X 的分布律得到.

例 2.5.1　设 X 有分布律

X	-1	0	1	2	3
p_k	0.2	0.2	0.3	0.2	0.1

求 $Y=(X-1)^2$ 的分布律.

解　Y 的所有可能取值为 0，1，4，且

$$P\{Y=0\}=P\{X=1\}=0.3,$$
$$P\{Y=1\}=P\{X=0\}+P\{X=2\}=0.2+0.2=0.4,$$
$$P\{Y=4\}=P\{X=-1\}+P\{x=3\}=0.2+0.1=0.3,$$

故 Y 的分布律为

Y	0	1	4
q_i	0.3	0.4	0.3

一般地，若 X 是离散型随机变量，其分布律为

X	x_1	x_2	x_3	$\cdots$	x_k	$\cdots$
p_k	p_1	p_2	p_3	$\cdots$	p_k	$\cdots$

则 $Y=g(X)$ 也是一个离散型随机变量. 设 $y_i=g(x_i)$，则 Y 的概率分布为

Y	y_1	y_2	y_3	$\cdots$	y_j	$\cdots$
q_i	q_1	q_2	q_3	$\cdots$	q_j	$\cdots$

其中

$$q_j=\sum_{g(x_i)=y_j}p_i.$$

例 2.5.2　设某城市一个月内发生火灾的次数 $X\sim P(5)$，试求随机变量 $Y=|X-5|$ 的分布律.

解　X 的所有可能取值的集合为 $\{0,1,2,\cdots\}$，其分布律为

$$P\{X=k\}=\frac{5^k}{k!}\mathrm{e}^{-5},\quad k=0,1,2,\cdots$$

由 $Y=|X-5|$ 可知，Y 的所有可能取值的集合为 $\{0,1,2,\cdots\}$. 且对每个 $i=1,2,\cdots$，当 $0<i\leqslant 5$ 时，有 $k=5+i$ 和 $k=5-i$ 两个值使得 $|k-5|=i$；当 $i=0$ 或 $i\geqslant 6$ 时，只有一个 $k=5+i$ 使得 $|k-5|=i$. 于是，随机变量 Y 取值为 i 的概率为

$$q_i = P\{Y=i\} = \begin{cases} \left[\dfrac{5^{5-i}}{(5-i)!}+\dfrac{5^{5+i}}{(5+i)!}\right]\mathrm{e}^{-5}, & i=1,2,3,4,5, \\ \dfrac{5^{5+i}}{(5+i)!}\mathrm{e}^{-5}, & i=0,6,7,\cdots, \end{cases}$$

此即 $Y=|X-5|$ 的分布律.

2.5.2　连续型随机变量函数的分布

对于连续型随机变量 X，求 $Y=g(X)$ 的密度函数 $f_Y(y)$ 的基本方法是，根据分布函数的定义先求 $Y=g(X)$ 的分布函数 $F_Y(y)=P\{Y\leqslant y\}$，即

$$F_Y(y)=P\{Y\leqslant y\}=P\{g(X)\leqslant y\}.$$

将上式化为 X 的分布函数 $F_X(x)$ 的复合函数表达式后，再关于 y 求导数，就得到 Y 的密度函数 $f_Y(y)=F'_Y(y)$. 下面我们通过具体例题来说明如何求 Y 的密度函数.

例 2.5.3　设 $X\sim N(\mu,\sigma^2)$，求 $Y=\dfrac{X-\mu}{\sigma}$ 的密度函数.

解　设 $F_Y(y)$ 和 $f_Y(y)$ 分别为 Y 的分布函数和密度函数，则

$$F_Y(y)=P\{Y\leqslant y\}=P\left\{\frac{X-\mu}{\sigma}\leqslant y\right\}=P\{X\leqslant \mu+\sigma y\}$$

$$=F_X(\mu+\sigma y)=\int_{-\infty}^{\mu+\sigma y}\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x.$$

再由 $f_Y(y)=F'_Y(y)$ 得

$$f_Y(y)=f_X(\mu+\sigma y)\frac{\mathrm{d}(\mu+\sigma y)}{\mathrm{d}y}$$

$$=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{[(\sigma y+\mu)-\mu]^2}{2\sigma^2}}\cdot\sigma=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}},$$

它是标准正态分布的密度函数，即 $Y\sim N(0,1)$.

例 2.5.3 说明，若 $X\sim N(\mu,\sigma^2)$，则 $Y=\dfrac{X-\mu}{\sigma}\sim N(0,1)$. 人们通常把这种变换称为 X 的标准化.

例 2.5.4　设 X 有密度函数

$$f_X(x)=\begin{cases}|x|, & -1<x<1,\\ 0, & \text{其他}.\end{cases}$$

求 $Y=2X+1$ 的密度函数.

解　设 $F_X(x)=P(X\leqslant x)$ 为 X 的分布函数，$F_Y(y)$ 和 $f_Y(y)$ 分别为 Y 的分布函数和密度函数，则

$$F_Y(y)=P\{Y\leqslant y\}=P\{2X+1\leqslant y\}=P\left\{X\leqslant\frac{y-1}{2}\right\}$$

$$=F_X\left(\frac{y-1}{2}\right)=\int_{-\infty}^{\frac{y-1}{2}}f_X(x)\,\mathrm{d}x.$$

再由 $f_Y(y)=F_Y'(y)$，得

$$f_Y(y)=f_X\left(\frac{y-1}{2}\right)\left(\frac{y-1}{2}\right)'=\frac{1}{2}f_X\left(\frac{y-1}{2}\right)$$

$$=\frac{1}{2}\begin{cases}\left|\dfrac{y-1}{2}\right|, & -1<\dfrac{y-1}{2}<1,\\ 0, & \text{其他},\end{cases}=\begin{cases}\dfrac{|y-1|}{4}, & -1<y<3,\\ 0, & \text{其他}.\end{cases}$$

注： 在求 $F_Y(y)$ 关于 y 的导数时，可采用复合函数求导数公式

$$\frac{\mathrm{d}F_X[h(y)]}{\mathrm{d}y}=F_X'[h(y)]h'(y)=f_X[h(y)]h'(y)\tag{2.5.1}$$

或变上、下限积分的导数公式

$$\frac{\mathrm{d}}{\mathrm{d}y}\left[\int_{a(y)}^{b(y)}f_X(t)\,\mathrm{d}t\right]=f_X[b(y)]b'(y)-f_X[a(y)]a'(y)\tag{2.5.2}$$

当函数 $g(x)$ 满足一定条件时，也可以利用下面的定理直接求 $f_Y(y)$.

定理 2.5.1　若随机变量 X 有密度函数 $f_X(x)$，$x\in(-\infty,+\infty)$，$y=g(x)$ 为严格单调函数，且 $g'(x)$ 对一切 x 都存在，记 (a,b) 为 $g(x)$ 的值域，则 $Y=g(X)$ 的密度函数为

$$f_Y(y)=\begin{cases}f_X[h(y)]\,|h'(y)|, & a<y<b,\\ 0, & \text{其他}.\end{cases}$$

这里 $x=h(y)$ 是函数 $y=g(x)$ 的反函数.

注： 如果 X 的密度函数在一个有限区间 $[\alpha,\beta]$ 之外取值为零，则定理 2.5.1 中函数 $g(x)$ 只须满足在 (α,β) 内可导，且在该区间严格单调即可. 当 $g(x)$ 为单调增函数时，$a=g(\alpha)$，$b=g(\beta)$；当 $g(x)$ 为单调减函数时，$a=g(\beta)$，$b=g(\alpha)$.

例 2.5.5　设 $X\sim U[0,1]$，其密度函数为

$$f_X(x)=\begin{cases}1, & 0\leqslant x\leqslant 1,\\ 0, & \text{其他}.\end{cases}$$

求 $Y=\mathrm{e}^X$ 的密度函数.

解　由于 $f_X(x)$ 在 $[0,1]$ 之外取值为零，且函数 $y=g(x)=\mathrm{e}^x$ 在 $(0,1)$ 内可导且严格增加，其反函数为 $x=h(y)=\ln y$. 这里，$\alpha=0$，$\beta=1$，$a=g(\alpha)$，$b=g(\beta)$ 且 $|h'(y)|=1/y$. 由定理 2.5.1 得 $Y=\mathrm{e}^X$ 的密度函数

$$f_Y(y)=\begin{cases}\dfrac{1}{y}, & 1\leqslant y\leqslant \mathrm{e},\\ 0, & \text{其他}.\end{cases}$$

当 $g(x)$ 不是严格单调函数时，不能使用定理 2.5.1，但此时例 2.5.3 和例 2.5.4 中的方法仍然适用.

需要指出，连续型随机变量的函数 $Y=g(X)$ 不一定是连续型的. 如果它是离散型的，可计算其分布律.

例 2.5.6　设加工零件的尺寸误差 $X\sim N(0,\sigma^2)$. 有时正误差和负误差所产生的后果不同. 若用 Y 表示由误差所引起的损失，并设

$$Y=\begin{cases}a, & \text{若 } X\geqslant 0,\\ b, & \text{若 } X<0,\end{cases}$$

其中 $a \neq b$. 则 Y 是 X 的函数，它服从一个两点分布，其分布律为

$$P\{Y = a\} = P\{X \geqslant 0\} = 0.5,$$
$$P\{Y = b\} = P\{X < 0\} = 0.5.$$

习题二

1. 将一颗骰子抛掷两次，以 X 表示两次出现点数的最小值，试求 X 的分布律.

2. 设离散型随机变量 X 的分布律为

$$P\{X = k\} = \frac{1}{2^k},\ k = 1,\ 2,\ 3,\ \cdots,$$

求概率(1)$P\{X = 2,\ 4,\ 6,\ \cdots\}$，(2)$P\{X \geqslant 3\}$.

3. 设在 15 只同类型的零件中有 2 只是次品，在其中取 3 次，每次任取 1 只，作不放回抽样，以 X 表示取出次品的只数，求 X 的分布律和分布函数.

4. 设离散型随机变量 X 的分布律为

$$P\{X = k\} = a\mathrm{e}^{-k},\ k = 1,\ 2,\ \cdots,$$

试确定常数 a.

5. 一大楼内装有 5 个同类型的供水设备，调查表明在任一时刻 t 每个设备被使用的概率为 0.1，试求：

（1）恰有 2 个设备同时被使用的概率；

（2）至少有 3 个设备同时被使用的概率；

（3）至多有 3 个设备同时被使用的概率；

（4）至少有 1 个设备同时被使用的概率.

6. 甲、乙两人投篮，投中的概率分别为 0.6 和 0.7. 今二人各投 3 次，求：

（1）两人投中次数相等的概率；

（2）甲比乙投中次数多的概率.

7. 有甲、乙两种味道和颜色都极为相似的名酒各 4 杯. 如果从中挑 4 杯，能将甲种酒全部挑出来，算是试验成功一次，求：

（1）某人随机地去挑，试验成功一次的概率；

（2）某人声称他通过品尝能区分两种酒. 他连续试验 10 次，成功 3 次. 试推断他是猜对的，还是他确有区分的能力(设各次试验是相互独立的).

8. 某一公安局在长度为 t 的时间间隔内收到的紧急呼救次数 X 服从参数为$(1/2)t$ 的泊松分布，而与时间间隔的起点无关(时间以小时计)，求：

（1）某一天中午 12 时~下午 3 时没有收到紧急呼救的概率；

（2）某一天中午 12 时~下午 5 时至少收到 1 次紧急呼救的概率.

9. 设有同类型设备 200 台，各台设备工作相互独立，发生故障的概率均为 0.005，通常一台设备的故障可由一个人来排除.

（1）至少配备多少维修工人，才能保证设备发生故障而不能及时排除的概率不大于 0.01?

(2) 若一人包干 40 台设备，求设备发生故障而不能及时排除的概率；

(3) 若由 2 人共同负责维修 100 台设备，求设备发生故障而不能及时排除的概率.

10. 设随机变量 X 在[2，5] 上服从均匀分布，现对 X 进行三次独立观测，试求至少有两次观测值大于 3 的概率.

11. 对球的直径作测量，设其均匀地分布在 20 ~ 22cm，求直径在 20.1 ~ 20.5cm 的概率.

12. 设随机变量 X 的分布函数为

$$F(x)=\begin{cases}0, & x<1,\\ \ln x, & 1\leqslant x<\mathrm{e},\\ 1, & x\geqslant \mathrm{e}.\end{cases}$$

(1) 求概率 $P\{X<2\}$，$P\{0<X\leqslant 3\}$，$P\{2<X<\dfrac{5}{2}\}$；

(2) 求 X 的密度函数 $f(x)$.

13. 设连续型随机变量 X 的分布函数为

$$F(x)=\begin{cases}a+b\mathrm{e}^{-\frac{x^2}{2}}, & x\geqslant 0,\\ 0, & x<0.\end{cases}$$

(1) 求常数 a 和 b；

(2) 求 X 的密度函数 $f(x)$；

(3) 求概率 $P\{\sqrt{\ln 4}<X<\sqrt{\ln 16}\}$.

14. 设随机变量 X 的密度函数为

(1) $f(x)=\begin{cases}2\left(1-\dfrac{1}{x^2}\right), & 1\leqslant x\leqslant 2,\\ 0, & \text{其他};\end{cases}$

(2) $f(x)=\begin{cases}x, & 0\leqslant x<1,\\ 2-x, & 1\leqslant x<2,\\ 0, & \text{其他}.\end{cases}$

求 X 的分布函数 $F(x)$.

15. 某机构研究了英格兰在 1875－1951 年期间，在矿山发生导致 10 人以上死亡的事故的频繁程度，得知相继两次事故之间的时间 T 服从指数分布，其概率为

$$f_T(t)=\begin{cases}\dfrac{1}{241}\mathrm{e}^{-\frac{t}{241}}, & t>0,\\ 0, & \text{其他}.\end{cases}$$

求 T 的分布函数 $F_T(t)$，并求概率 $P\{50<T<100\}$.

16. 某种型号器件的寿命 X(以小时计) 具有以下密度函数

$$f(x)=\begin{cases}\dfrac{1\ 000}{x^2}, & x>1\ 000,\\ 0, & \text{其他}.\end{cases}$$

现有一大批此种器件(设各器件损坏与否相互独立)，任取 5 只，求其中至少有 2 只寿命

大于 1 500 小时的概率.

17. 设顾客在某银行的窗口等待服务的时间 X(以分计) 服从指数分布，其密度函数为

$$f(x)=\begin{cases}\dfrac{1}{5}\mathrm{e}^{-\frac{x}{5}}, & x>0,\\ 0, & \text{其他.}\end{cases}$$

某顾客在窗口等待服务，若超过 10 分钟，他就离开. 他一个月要到银行 5 次，以 Y 表示一个月内他未等到服务而离开窗口的次数. 试求 Y 的分布律及概率 $P\{Y\geqslant 1\}$.

18. 设 $X\sim N(3,2^2)$，(1) 求概率 $P\{2<X\leqslant 5\}$，$P\{-4<X\leqslant 10\}$，$P\{|X|>2\}$，$P\{X>3\}$；(2) 确定 c 使得 $P\{X>c\}=P\{X\leqslant c\}$；(3) 设 d 满足 $P\{X>d\}\geqslant 0.9$，问 d 至多为多少?

19. 设 $X\sim N(160,\sigma^2)$，若要求 X 落在区间(120，200) 内的概率不小于 0.80，则应允许 σ 最大为多少?

20. 某地区 18 岁女青年的血压(收缩压，以 mmHg 计) 服从正态分布 $N(110,12^2)$，在该地区任选一个 18 岁女青年，测量她的血压 X，试确定最小 x，使得 $P\{X>x\}\leqslant 0.05$.

21. 某地抽样调查结果表明，考生的外语成绩(百分制) 近似地服从正态分布，平均成绩为 72 分，96 分以上占考生总数的 2.3%，试求考生的外语成绩在 60 分至 84 分之间的概率.

22. 公共汽车的车门高度是按成年男性与车门碰头的机会不超过 0.01 设计的，设成年男性的身高 X(单位：厘米) 服从正态分布 $N(170,6^2)$，问车门的最低高度应为多少?

23. 设随机变量 X 的分布律为

X	0	$\dfrac{\pi}{2}$	π	$\dfrac{3\pi}{2}$
0.3	0.3	0.2	0.4	0.1

求下列随机变量 Y 的分布律.

(1) $Y=(2X-\pi)^2$；　　(2) $Y=\cos(2X-\pi)$.

24. 设随机变量 X 的分布函数为

$$F(x)=\begin{cases}0, & x<-1,\\ 0.3, & -1\leqslant x<1,\\ 0.8, & 1\leqslant x<2,\\ 1, & x\geqslant 2.\end{cases}$$

(1) 求 X 的分布律；(2) 求 $Y=|X|$的分布律.

25. 设 $X\sim N(0,1)$，求下列随机变量 Y 的密度函数.

(1) $Y=2X-1$；　　(2) $Y=\mathrm{e}^{-X}$；　　(3) $Y=X^2$.

26. 设随机变量 $X\sim U(0,\pi)$，求下列随机变量 Y 的密度函数.

(1) $Y=2\ln X$；　　(2) $Y=\cos X$；　　(3) $Y=\sin X$.

第三章　多维随机变量及其分布

在许多实际问题中，随机试验的结果往往需要用两个或两个以上的随机变量来描述. 例如，在打靶射击中，炮弹弹着点的位置需要由它的横坐标 X 及纵坐标 Y 来确定，这就涉及到两个随机变量 X 和 Y. 又例如，在研究分子运动中，考虑分子运动的速度 V，这里 $V=(X, Y, Z)$ 由三个分量组成，当分子自由运动时，其速度 V 的三个分量 X、Y 和 Z 都是随机变量.

一般地，设 $X_1, X_2, \cdots, X_n$ 为 n 个随机变量，称这 n 个随机变量构成的一组随机变量 $(X_1, X_2, \cdots, X_n)$ 为 n 维随机变量，或称为 n 维随机向量. 例如，打靶射击中弹着点的位置 (X, Y) 是一个二维随机变量，自由运动的分子速度 $V=(X, Y, Z)$ 是一个三维随机变量.

由于对于 $n(>2)$ 维随机变量的研究与二维随机变量类似，基本没有本质上的差异，故本章主要讨论二维随机变量，所有结论都可以平行推广到 n 维随机变量.

3.1　二维随机变量的联合分布

二维随机变量 (X, Y) 的性质不仅与 X 的性质及 Y 的性质有关，而且还依赖于这两个随机变量的相互关系，因此，仅仅逐个研究 X 和 Y 的性质是不够的，必须把 (X, Y) 作为一个整体加以研究.

首先引入 (X, Y) 的分布函数的概念.

定义 3.1.1　设 (X, Y) 是二维随机变量，对于任意实数 x，y，称二元函数

$$F(x, y)=P\{X \leqslant x, Y \leqslant y\} \tag{3.1.1}$$

为二维随机变量 (X, Y) 的联合分布函数，简称为分布函数.

分布函数 $F(x, y)$ 表示事件 $\{X \leqslant x\}$ 和事件 $\{Y \leqslant y\}$ 同时发生的概率，即它们的积事件的概率. 如果把 (X, Y) 看成是平面上随机点的坐标，则分布函数 $F(x, y)$ 在点 (x, y) 处的函数值就是随机点 (X, Y) 落在平面上的以点 $n-i$ 为顶点而位于该点左下的无穷矩形区域内的概率，如图 3.1 所示.

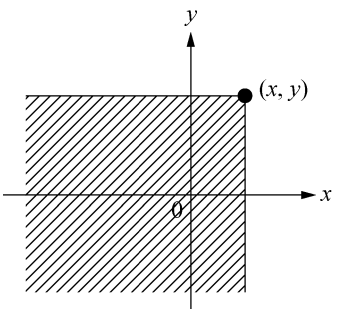


图 3.1　无穷矩形区域

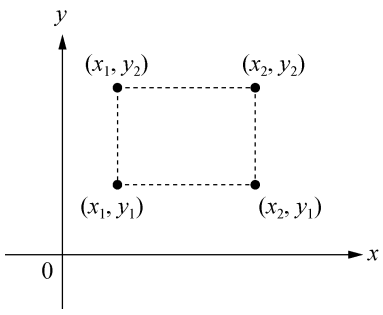


图 3.2　一般矩形区域

由分布函数 $F(x, y)$ 的上述几何解释容易知道，随机点 (X, Y) 落在图 3.2 中矩形区域 $\{x_1 < X \leqslant x_2, y_1 < Y \leqslant y_2\}$ 内的概率可以表示为

$$P\{x_1 < X \leqslant x_2, y_1 < Y \leqslant y_2\} = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \tag{3.1.2}$$

分布函数 $F(x, y)$ 具有下列三个基本性质.

性质 1　$F(x, y)$ 分别关于 x 和 y 单调不减，即对于任意固定的 y，当 $x_1 < x_2$ 时，$F(x_1, y) \leqslant F(x_2, y)$；对于任意固定的 x，当 $y_1 < y_2$ 时，$F(x, y_1) \leqslant F(x, y_2)$.

这里仅对固定 y 时的情况加以证明. 事实上，由(3.1.1) 式可得

$$\begin{aligned} F(x_2, y) - F(x_1, y) &= P\{X \leqslant x_2, Y \leqslant y\} - P\{X \leqslant x_1, Y \leqslant y\} \\ &= P\{x_1 < X \leqslant x_2, Y \leqslant y\} \geqslant 0 \end{aligned}$$

性质 2　$F(x, y)$ 关于 x 右连续，关于 y 也右连续，即

$$F(x+0, y) = F(x, y), \quad F(x, y+0) = F(x, y).$$

性质 3　$0 \leqslant F(x, y) \leqslant 1$，且对于任意固定的 y，

$$F(-\infty, y) = \lim_{x \to +\infty} F(x, y) = 0,$$

对于任意固定的 x，

$$F(x, -\infty) = \lim_{y \to -\infty} F(x, y) = 0,$$

及

$$F(-\infty, -\infty) = \lim_{\substack{x \to -\infty \\ y \to -\infty}} F(x, y) = 0,$$

$$F(+\infty, +\infty) = \lim_{\substack{x \to +\infty \\ y \to +\infty}} F(x, y) = 1.$$

上面四个式子的意义可以从几何上直观解释. 例如，若在图 3.1 中将无穷矩形的右边界向左无限移动(即令 $x \to -\infty$)，则随机变量 (X, Y) 落在这个矩形内这一事件趋于不可能事件，其概率趋于零，即有 $F(-\infty, y) = 0$. 又当 $x \to +\infty$，$y \to +\infty$ 时，图 3.1 中的无穷矩形扩展到全平面，随机变量 (X, Y) 落在这个矩形内这一事件趋于必然事件，其概率趋于 1，即有 $F(+\infty, +\infty) = 1$.

与一维随机变量一样，二维随机变量也有离散型与连续型之分，下面分别讨论它们.

3.2　二维离散型随机变量

如果二维随机变量 (X, Y) 的每个分量都是离散型的，则称 (X, Y) 是二维离散型随机变量. 因为离散型随机变量只能取有限或可列无穷个值，因此二维离散型随机变量 (X, Y) 的所有可能取的值也是有限的或可列无穷的.

定义 3.2.1　设二维离散型随机变量 (X, Y) 的所有可能取值为 (x_i, y_j)，$i, j = 1, 2, 3, \cdots$，记这些基本事件的概率为

$$P\{X = x_i, Y = y_j\} = p_{ij}, \quad i, j = 1, 2, 3, \cdots \tag{3.2.1}$$

称式(3.2.1) 为离散型随机变量 (X, Y) 的联合概率分布或联合分布律，简称为分布律.

(X, Y) 的分布律也可用如下的表格来表示.

X \ Y	y_1	y_2	…	y_j	…
x_1	p_{11}	p_{12}	…	p_{1j}	…
x_2	p_{21}	p_{22}	…	p_{2j}	…
…	…	…	…	…	…
x_i	p_{i1}	p_{i2}	…	p_{ij}	…
…	…	…	…	…	…

由概率的性质知，p_{ij} 具有如下性质：

性质 1　$p_{ij} \geqslant 0$，i，$j = 1$，2，3，…

性质 2　$\sum\limits_i \sum\limits_j p_{ij} = 1.$

离散型随机变量(X, Y) 的分布函数为

$$F(x, y) = P\{X \leqslant x, Y \leqslant y\} = \sum_{x_i \leqslant x} \sum_{y_i \leqslant y} p_{ij} \tag{3.2.2}$$

其中和式对一切满足 $x_i \leqslant x$，$y_i \leqslant y$ 的 i 和 j 求和.

例 3.2.1　设随机变量 X 在 1，2，3，4 四个整数中等可能地取一个值，另一个随机变量 Y 在 $1-X$ 之间等可能地取一整数值. 试求(X, Y) 的分布律.

解　易知$\{X = i, Y = j\}$ 的取值情况是：$i = 1$，2，3，4，$j \leqslant i$. 由概率的乘法公式，

$$P\{X = i, Y = j\} = P\{Y = j \mid X = i\} P\{X = i\} = \frac{1}{i} \cdot \frac{1}{4}, \quad i = 1, 2, 3, 4, \ j \leqslant i.$$

于是(X, Y) 的分布律为

X \ Y	1	2	3	4
1	1/4	0	0	0
2	1/8	1/8	0	0
3	1/12	1/12	1/12	0
4	1/16	1/16	1/16	1/16

例 3.2.2　为了研究抽烟与肺癌的关系，随机调查了 23 000 个 40 岁以上的人，其结果见下表. 表中的数字“3”表示既抽烟又患肺癌的人数，“4 597”表示抽烟但未患肺癌的人数，其余类似.

吸烟 \ 患肺癌	是	否	
是	3	4 597	4 600
否	1	18 399	18 4000
	4	22 996	23 000

为了进一步研究这个问题，引进二维随机变量(X, Y)，其中

$$X=\begin{cases}1, & \text{若被调查者不抽烟},\\ 0, & \text{若被调查者抽烟};\end{cases}$$

$$Y=\begin{cases}1, & \text{若被调查者未患肺癌},\\ 0, & \text{若被调查者患肺癌}.\end{cases}$$

从原始数据表中每一种情况出现的次数计算它们出现的频率作为概率的估计，即有分布律

$$P\{X=0,\ Y=0\}=0.000\ 13,$$
$$P\{X=1,\ Y=0\}=0.000\ 04,$$
$$P\{X=0,\ Y=1\}=0.199\ 87,$$
$$P\{X=1,\ Y=1\}=0.799\ 96.$$

可以看出，既抽烟又患肺癌的概率是 0.000 13，而不抽烟患肺癌的概率是 0.000 04 等. 上述分布律也可由下表给出.

X \ Y	0	1
0	0.000 13	0.199 87
1	0.000 04	0.799 96

3.3　二维连续型随机变量

与一维连续型随机变量定义类似，二维连续型随机变量有如下定义.

定义 3.3.1　设二维随机变量(X, Y)的联合分布函数为$F(x, y)$，如果存在非负函数$f(x, y)$，使得对于任意实数x，y，有

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

则称(X, Y)为二维连续型随机变量，并称函数$f(x, y)$为(X, Y)的联合概率密度函数，简称为联合密度函数，或密度函数.

密度函数$f(x, y)$具有以下性质：

性质 1　$f(x, y)\geqslant 0$，$-\infty<x<+\infty$，$-\infty<y<+\infty$；

性质 2　$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}f(x, y)\,\mathrm{d}x\mathrm{d}y=F(+\infty, +\infty)=1$；

性质 3　如果$f(x, y)$在点(x, y)连续，则有

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

性质 4　设D是xoy平面上的一个平面区域，则(X, Y)落在D内的概率为

$$P\{(X, Y)\in D\}=\iint_D f(x, y)\,\mathrm{d}x\mathrm{d}y \tag{3.3.1}$$

密度函数$z=f(x, y)$的图形是一个空间曲面，称之为分布曲面. 性质 2 的几何意义是，介于分布曲面$z=f(x, y)$和xoy平面之间的全部体积等于 1. 性质 4 中概率$P\{(X, Y)\in D\}$的几何意义是，它等于以D为底，以曲面$z=f(x, y)$为顶的曲顶柱体的体积，如图 3.3 所示.

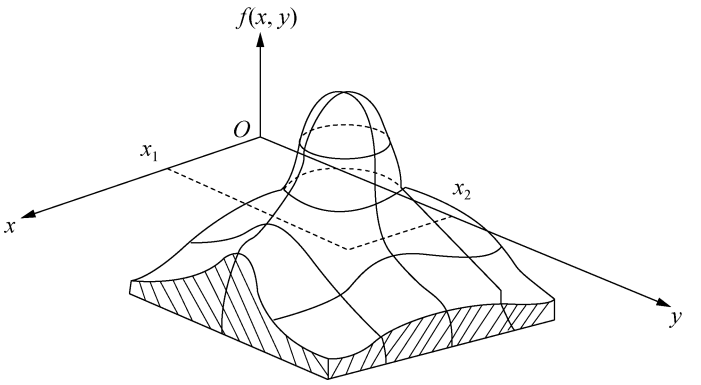


图 3.3　$P\{(X, Y) \in D\}$ 的几何意义

例 3.3.1　设二维随机变量(X, Y) 的密度函数为

$$f(x, y) = \begin{cases} k\mathrm{e}^{-(2x+3y)}, & x > 0, \ y > 0, \\ 0, & \text{其他.} \end{cases}$$

(1) 确定常数 k;

(2) 求(X, Y) 的分布函数;

(3) 求 $P\{0 < X \leqslant 4, \ 0 < Y \leqslant 1\}$;

(4) 求 $P\{X < Y\}$.

解　(1) 由性质 2，有

$$\begin{aligned} 1 &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x, y)\,\mathrm{d}x\mathrm{d}y = \int_{0}^{+\infty}\int_{0}^{+\infty} k\mathrm{e}^{-(2x+3y)}\,\mathrm{d}x\mathrm{d}y \\ &= k\int_{0}^{+\infty}\mathrm{e}^{-2x}\mathrm{d}x\int_{0}^{+\infty}\mathrm{e}^{-3y}\mathrm{d}y \\ &= k\left[-\frac{1}{2}\mathrm{e}^{-2x}\right]_{0}^{+\infty}\left[-\frac{1}{3}\mathrm{e}^{-3y}\right]_{0}^{+\infty} \\ &= \frac{k}{6}, \end{aligned}$$

即得 $k = 6$.

(2) 由定义 3.3.1，有

$$\begin{aligned} F(x, y) &= \int_{-\infty}^{y}\int_{-\infty}^{x} f(u, v)\,\mathrm{d}u\mathrm{d}v \\ &= \begin{cases} \int_{0}^{y}\int_{0}^{x} 6\mathrm{e}^{-(2u+3v)}\,\mathrm{d}u\mathrm{d}v, & x > 0, \ y > 0, \\ 0, & \text{其他.} \end{cases} \\ &= \begin{cases} (1-\mathrm{e}^{-2x})(1-\mathrm{e}^{-3y}), & x > 0, \ y > 0, \\ 0, & \text{其他.} \end{cases} \end{aligned}$$

(3) 所求概率为

$$P\{0 < X \leqslant 4, \ 0 < Y \leqslant 1\} = \int_{0}^{1}\int_{0}^{4} 6\mathrm{e}^{-(2x+3y)}\,\mathrm{d}x\mathrm{d}y = (1-\mathrm{e}^{-8})(1-\mathrm{e}^{-3}) \approx 0.95.$$

该小题也可按公式(3.1.2) 来求，请读者自己验算.

（4）所求概率为

$$P\{X<Y\}=\iint_D f(x,\ y)\,\mathrm{d}x\mathrm{d}y=\iint_{x<y} f(x,\ y)\,\mathrm{d}x\mathrm{d}y$$
$$=\int_0^{+\infty}\left[\int_0^y 6\mathrm{e}^{-(2x+3y)}\,\mathrm{d}x\right]\mathrm{d}y=\int_0^{+\infty}3\mathrm{e}^{-3y}[1-\mathrm{e}^{-2y}]\,\mathrm{d}y$$
$$=\int_0^{+\infty}3\mathrm{e}^{-3y}\mathrm{d}y-\int_0^{+\infty}3\mathrm{e}^{-5y}\mathrm{d}y=1-\frac{3}{5}=\frac{2}{5}.$$

3.4　常见多维随机变量

3.4.1　多项分布

多项分布是重要的多维离散分布，它是二项分布的推广.

进行 n 次独立重复试验，如果每次试验有 r 个互不相容的结果 A_1，A_2，…，A_r 之一发生，且每次试验中 A_i 发生的概率为 $p_i=P(A_i)$，$i=1$，2，…，r，且 $p_1+p_2+\cdots+p_r=1$. 记 X_i 为 n 次独立重复试验中 A_i 出现的次数，$i=1$，2，…，r. 则 $(X_1,\ X_2,\ \cdots,\ X_r)$ 取值为 $(n_1,\ n_2,\ \cdots,\ n_r)$ 的概率，即 A_1 出现 n_1 次，A_2 出现 n_2 次，…，A_r 出现 n_r 次的概率为

$$P(X_1=n_1,\ X_2=n_2,\ \cdots,\ X_r=n_r)=\frac{n!}{n_1!\ n_2!\ \cdots n_r!}p_1^{n_1}p_2^{n_2}\cdots p_r^{n_r}\qquad(3.4.1)$$

其中 $n=n_1+n_2+\cdots+n_r$.

这个联合分布律称为多项分布，记为 $M(n,\ p_1,\ p_2,\ \cdots,\ p_r)$. 式(3.4.1)右侧给出的概率是多项式 $(p_1+p_2+\cdots+p_r)^n$ 展开式中的一般项，故其和为1. 当 $r=2$ 时，多项分布即为二项分布.

3.4.2　多维均匀分布

定义 3.4.1　设 D 是平面上的有界区域，其面积为 d，若二维随机变量 $(X,\ Y)$ 的密度函数为

$$f(x,\ y)=\begin{cases}\dfrac{1}{d},&(x,\ y)\in D,\\ 0,&\text{其他.}\end{cases}$$

则称 $(X,\ Y)$ 在 D 上服从均匀分布.

与第2章中随机变量服从的均匀分布相类似，在区域 D 上服从均匀分布的 $(X,\ Y)$ 落在 D 中某一区域 A 内的概率 $P\{(X,\ Y)\in A\}$ 与 A 的面积成正比，而与 A 的位置和形状无关.

更为一般的均匀分布如下定义：设 D 为 R^n 中的一个有界区域，其度量为 $\mu(D)$，如果多维随机变量 $(X_1,\ X_2,\ \cdots,\ X_n)$ 的联合密度函数为

$$f(x_1,\ x_2,\ \cdots,\ x_n)=\begin{cases}\dfrac{1}{\mu(D)},&\text{当}(x_1,\ x_2,\ \cdots,\ x_n)\in D,\\ 0,&\text{其他.}\end{cases}\qquad(3.4.2)$$

则称$(X_1, X_2, \cdots, X_n)$服从D上的多维均匀分布，记为$(X_1, X_2, \cdots, X_n) \sim U(D)$.

例 3.4.1　设(X, Y)在圆域$x^2+y^2 \leqslant 4$上服从均匀分布，计算$P\{(X, Y) \in A\}$，这里区域A是由$x=0$、$y=0$和$x+y=1$三条直线所围成的三角形区域，即图 3.4 中阴影部分区域.

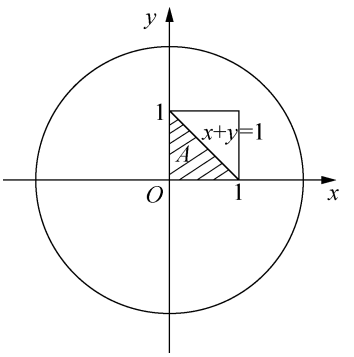


图 3.4　积分区域

解　圆域$x^2+y^2 \leqslant 4$的面积$d=4\pi$，因此(X, Y)的密度函数为

$$f(x, y)=\begin{cases}\dfrac{1}{4\pi}, & x^2+y^2 \leqslant 4,\\ 0, & \text{其他.}\end{cases}$$

区域A包含在圆域$x^2+y^2 \leqslant 4$之内，其面积为$\dfrac{1}{2}$，于是由(3.3.1)式得

$$P\{(X, Y) \in A\}=\iint\limits_A \frac{1}{4\pi}\mathrm{d}x\mathrm{d}y=\frac{1}{4\pi}\iint\limits_A \mathrm{d}x\mathrm{d}y=\frac{1}{8\pi}.$$

3.4.3　多维正态分布

定义 3.4.2　若二维随机变量(X, Y)的密度函数为

$$f(x, y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\mathrm{e}^{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right]},\quad -\infty<x, y<\infty$$

其中$\sigma_1>0$，$\sigma_2>0$，$-1<\rho<1$，则称(X, Y)服从参数为μ_1，μ_2，σ_1，σ_2，ρ的**二维正态分布**.

更一般的多为正态分布定义如下：如果多维随机变量$X=(X_1, X_2, \cdots, X_n)'$的联合密度函数为

$$f(x)=\frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(x-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})\right\},$$

其中$x=(x_1, x_2, \cdots, x_n)'$，$\boldsymbol{\mu}=(\mu_1, \mu_2, \cdots, \mu_n)'$，$\boldsymbol{\Sigma}=(\sigma_{ij})$为$n$阶正定矩阵，$|\boldsymbol{\Sigma}|$为$\boldsymbol{\Sigma}$的行列式，$\boldsymbol{\Sigma}^{-1}$为$\boldsymbol{\Sigma}$的逆矩阵，则称$X=(X_1, X_2, \cdots, X_n)'$服从均值向量为$\boldsymbol{\mu}$，协方差矩阵为$\boldsymbol{\Sigma}$的$n$维正态分布，记为$X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

3.5　边缘分布

3.5.1　边缘分布函数

二维随机变量(X, Y)作为一个整体，具有联合分布函数$F(x, y)$，而X和Y又都是一维随机变量，自然都有各自的分布函数，分别记为$F_X(x)$和$F_Y(y)$，依次称它们为二维随机变量(X, Y)关于X和关于Y的**边缘分布函数**.

需要指出的是，边缘分布函数$F_X(x)$和$F_Y(y)$就是一维随机变量X和Y的分布函数，

即 $F_X(x)=P\{X\leqslant x\}$，$F_Y(x)=P\{Y\leqslant y\}$. 之所以称它们为边缘分布是相对于$(X,\ Y)$ 的联合分布而言的，或者它们可以由联合分布函数 $F(x,\ y)$ 来确定. 事实上，对于一个二维随机变量$(X,\ Y)$，事件$\{X\leqslant x\}$ 是由所有横坐标小于或等于 x 的点组成的集合，在平面上就是事件$\{X\leqslant x,\ Y<+\infty\}$. 因此，边缘分布函数 $F_X(x)$ 可以由$(X,\ Y)$ 的联合分布函数 $F(x,\ y)$ 来确定，即有

$$F_X(x)=P\{X\leqslant x\}=P\{X\leqslant x,\ Y<+\infty\}=F(x,\ +\infty) \tag{3.5.1}$$

同理有

$$F_Y(y)=P\{Y\leqslant y\}=P\{X<+\infty,\ Y\leqslant y\}=F(+\infty,\ y) \tag{3.5.2}$$

3.5.2　离散型随机变量的边缘分布

设$(X,\ Y)$ 为二维离散型随机变量，若已知其联合分布律为

$$P\{X=x_i,\ Y=y_j\}=p_{ij},\ i,\ j=1,\ 2,\ 3,\ \cdots,$$

由式(3.5.1) 和式(3.5.2) 可得

$$F_X(x)=F(x,\ +\infty)=\sum_{x_i\leqslant x}\sum_{y_j<+\infty}p_{ij}=\sum_{x_i\leqslant x}\sum_{j}p_{ij}.$$

对照一维随机变量 X 的分布函数定义

$$F_X(x)=P\{X\leqslant x\}=\sum_{x_i\leqslant x}P\{X=x_i\}$$

可知 X 的分布律为

$$P\{X=x_i\}=\sum_{j=1}^{\infty}p_{ij},\ i=1,\ 2,\ \cdots,$$

这里事件$\{X=x_i\}$ 即所有横坐标为 x_i 的样本点集合$\{(x_i,\ y_j)\mid j=1,\ 2,\ \cdots\}$ 构成的事件. 同理，Y 的分布律为

$$P\{Y=y_j\}=\sum_{i=1}^{\infty}p_{ij},\ j=1,\ 2,\ \cdots,$$

记

$$p_{i\cdot}=P\{X=x_i\}=\sum_{j}p_{ij},\ i=1,\ 2,\ \cdots \tag{3.5.3}$$

$$p_{\cdot j}=P\{Y=y_j\}=\sum_{i}p_{ij},\ j=1,\ 2,\ \cdots \tag{3.5.4}$$

分别称 $p_{i\cdot}$ 和 $p_{\cdot j}(i,\ j=1,\ 2,\ 3,\ \cdots)$ 为 X 和 Y 的**边缘分布律**.

X 的边缘分布律可以用表格形式表示为

X	x_1	x_2	$\cdots$	x_i	$\cdots$
$p_{i\cdot}$	$p_{1\cdot}$	$p_{2\cdot}$	$\cdots$	$p_{i\cdot}$	$\cdots$

类似地，Y 的边缘分布律可表示为为

Y	y_1	y_2	$\cdots$	y_j	$\cdots$
$p_{\cdot j}$	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot j}$	$\cdots$

例 3.5.1 求例 3.2.1 中(X, Y)的分量 X 和 Y 的边缘分布律.

解 X 所有可能取的值为 1，2，3，4，分别记为 x_1，x_2，x_3 和 x_4；Y 所有可能取的值也是 1，2，3，4，分别记为 y_1，y_2，y_3 和 y_4. 由式(3.5.3)得到 X 的边缘分布律

$$P\{X=1\}=p_{1\cdot}=p_{11}+p_{12}+p_{13}+p_{14}=0+0+0+\frac{1}{4}=\frac{1}{4},$$

$$P\{X=2\}=p_{2\cdot}=p_{21}+p_{22}+p_{23}+p_{24}=0+0+\frac{1}{8}+\frac{1}{8}=\frac{1}{4},$$

$$P\{X=3\}=p_{3\cdot}=p_{31}+p_{32}+p_{33}+p_{34}=0+\frac{1}{12}+\frac{1}{12}+\frac{1}{12}=\frac{1}{4},$$

$$P\{X=4\}=p_{4\cdot}=p_{41}+p_{42}+p_{43}+p_{44}=\frac{1}{16}+\frac{1}{16}+\frac{1}{16}+\frac{1}{16}=\frac{1}{4}.$$

由式(3.5.4)得到 Y 的边缘分布律

$$P\{Y=1\}=p_{\cdot 1}=p_{11}+p_{21}+p_{31}+p_{41}=\frac{1}{4}+\frac{1}{8}+\frac{1}{12}+\frac{1}{16}=\frac{25}{48},$$

$$P\{Y=2\}=p_{\cdot 2}=p_{12}+p_{22}+p_{32}+p_{42}=0+\frac{1}{8}+\frac{1}{12}+\frac{1}{16}=\frac{13}{48},$$

$$P\{Y=3\}=p_{\cdot 3}=p_{13}+p_{23}+p_{33}+p_{43}=0+0+\frac{1}{12}+\frac{1}{16}=\frac{7}{48},$$

$$P\{Y=4\}=p_{\cdot 4}=p_{14}+p_{24}+p_{34}+p_{44}=0+0+0+\frac{1}{16}=\frac{3}{48}.$$

(X, Y) 的联合和边缘分布律可由下表给出，该表的中间部分是(X, Y)的联合分布律，最右边一列给出 X 的边缘分布律，最下面一行给出 Y 的边缘分布律.

X \ Y	1	2	3	4	$p_{i\cdot}$
1	1/4	0	0	0	1/4
2	1/8	1/8	0	0	1/4
3	1/12	1/12	1/12	0	1/4
4	1/16	1/16	1/16	1/16	1/4
$p_{\cdot j}$	25/48	13/48	7/48	3/48	

例 3.5.2 对例 3.2.2 中二维随机变量(X, Y)，分别求 X 和 Y 的边缘分布律.

解 由式(3.5.3)，得到 X 的边缘分布律

$$\begin{aligned}P\{X=0\}&=P\{X=0,\ Y=0\}+P\{X=0,\ Y=1\}\\&=0.000\ 13+0.199\ 87=0.2,\\P\{X=1\}&=P\{X=1,\ Y=0\}+P\{X=1,\ Y=1\}\\&=0.000\ 04+0.799\ 96=0.8.\end{aligned}$$

由此可知，随机抽取一个人，他吸烟的概率为 0.2，不吸烟的概率为 0.8.

同样地，由式(3.5.4)，得到 Y 的边缘分布律

$$\begin{aligned}P\{Y=0\}&=P\{X=0,\ Y=0\}+P\{X=1,\ Y=0\}\\&=0.000\ 13+0.000\ 04=0.000\ 17,\end{aligned}$$

$$P\{Y=1\}=P\{X=0,\ Y=1\}+P\{X=1,\ Y=1\}$$
$$=0.199\ 87+0.799\ 96=0.999\ 83.$$

由此可知，随机抽取一个人，他患肺癌的概率为 0.000 17，不患肺癌的概率为 0.999 83.

$(X,\ Y)$ 的联合分布和边缘分布律可由下表给出.

X \ Y	0	1	$p_{i\cdot}$
0	0.000 13	0.199 87	0.200 00
1	0.000 04	0.799 96	0.800 00
$p_{\cdot j}$	0.000 17	0.999 83	

例 3.5.3　证明：多项分布的一维边缘分布为二项分布.

证　下面只证明三项分布的边缘分布为二项分布. 设二维随机变量$(X,\ Y)$ 服从三项分布 $M(n,\ p_1,\ p_2,\ p_3)$，其中 $p_3=1-p_1-p_2$，$(X,\ Y)$ 的联合分布律为

$$P(X=i,\ Y=j)=\frac{n!}{i!\ j!\ (n-i-j)!}p_1^i p_2^j(1-p_1-p_2)^{n-i-j},$$

对上式分别乘以和除以$(1-p_1)^{n-i}/(n-i)!$，再对j从0到$n-i$求和，并记$p_2'=p_2/(1-p_1)$，则可得

$$P(X=i)=\sum_{j=1}^{n-i}P(X=i,\ Y=j)$$
$$=\frac{n!}{i!\ (n-i)!}p_1^i(1-p_1)^{n-i}\sum_{j=0}^{n-i}C_{n-i}^j(p_2')j(1-p_2')^{n-j}$$
$$=\frac{n!}{i!\ (n-i)!}p_1^i(1-p_1)^{n-i}[p_2'+(1-p_2')]^{n-i}$$
$$=\frac{n!}{i!\ (n-i)!}p_1^i(1-p_1)^{n-i}.$$

所以 $X\sim B(n,\ p_1)$，同理可证 $Y\sim B(n,\ p_2)$.

用类似方法可以证明：若多维随机变量$(X_1,\ X_2,\ \cdots,\ X_m)\sim M(n,\ p_1,\ p_2,\ \cdots,\ p_m)$，则 $X_i\sim B(n,\ p_i)$，$i=1,\ 2,\ \cdots,\ m$.

3.5.3　连续型随机变量的边缘分布

设$(X,\ Y)$ 为二维连续型随机变量，其联合密度为$f(x,\ y)$，由式(3.5.1) 和定义 3.3.1 得到 X 的边缘分布函数

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

因此 X 是一个连续型随机变量，其密度函数为

$$f_X(x)=\int_{-\infty}^{+\infty}f(x,\ y)\mathrm{d}y \tag{3.5.5}$$

同理可知，Y 也是一个连续型随机变量，其密度函数为

$$f_Y(y)=\int_{-\infty}^{+\infty}f(x,\ y)\mathrm{d}x \tag{3.5.6}$$

分别称$f_X(x)$和$f_Y(y)$为X和Y的**边缘密度函数**.

在$f_X(x)$和$f_Y(y)$的连续点处，有

$$\frac{\mathrm{d}F_X(x)}{\mathrm{d}x}=f_X(x),\ \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y}=f_Y(y).$$

例3.5.4　设二维随机变量$(X,\ Y)$在区域G上服从均匀分布，其中G是由直线$x=0$，$y=0$和$\frac{x}{2}+y=1$所围区域，即图3.5中阴影部分. 求$(X,\ Y)$的联合密度函数和边缘密度函数.

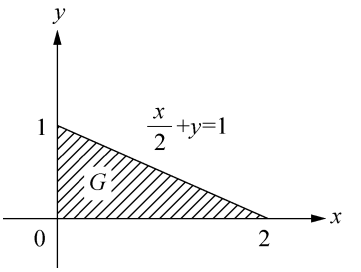


图3.5　三角形中的均匀分布

解　因为$(X,\ Y)$在区域G上服从均匀分布，且G的面积$d=\frac{1}{2}\times2\times1=1$，因此$(X,\ Y)$有联合密度函数

$$f(x,\ y)=\begin{cases}1,\ (x,\ y)\in G,\\0,\ \text{其他}.\end{cases}$$

当$0\leqslant x\leqslant2$时，

$$f_X(x)=\int_{-\infty}^{+\infty}f(x,\ y)\mathrm{d}y=\int_{-\infty}^{0}0\mathrm{d}y+\int_{0}^{1-\frac{x}{2}}1\mathrm{d}y+\int_{1-\frac{x}{2}}^{+\infty}0\mathrm{d}y=1-\frac{x}{2},$$

当$x<0$或$x>2$时，$f(x,\ y)=0$，所以$f_X(x)=0$，即X的边缘密度为

$$f_X(x)=\int_{-\infty}^{\infty}f(x,\ y)\mathrm{d}y=\begin{cases}1-\frac{x}{2},\ 0\leqslant x\leqslant2,\\0,\qquad\text{其他}.\end{cases}$$

同理可得Y的边缘密度

$$f_Y(y)=\int_{-\infty}^{+\infty}f(x,\ y)\mathrm{d}x=\begin{cases}2(1-y),\ 0\leqslant y\leqslant1,\\0,\qquad\text{其他}.\end{cases}$$

例3.5.5　设$(X,\ Y)$服从参数为$\mu_1=\mu_2=0$，$\sigma_1=\sigma_2=1$，ρ的二维正态分布，分别求X和Y的边缘密度.

解　该二维正态分布的联合密度函数为

$$f(x,\ y)=\frac{1}{2\pi\sqrt{1-\rho^2}}\mathrm{e}^{\frac{-1}{2(1-\rho^2)}[x^2+y^2-2\rho xy]},\quad -\infty<x,\ y<+\infty.$$

由于$x^2+y^2-2\rho xy=(y-\rho x)^2+(1-\rho^2)x^2$，于是

$$f_X(x)=\frac{1}{2\pi}\mathrm{e}^{-\frac{x^2}{2}}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{1-\rho^2}}\mathrm{e}^{-\frac{1}{2(1-\rho^2)}(y-\rho x)^2}\mathrm{d}y.$$

令$t=\frac{1}{\sqrt{1-\rho^2}}(y-\rho x)$，则有

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

即$X\sim N(0,\ 1)$. 同理有$Y\sim N(0,\ 1)$，其密度函数为

$$f_Y(y)=\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}},\quad -\infty<y<+\infty.$$

由例3.5.5可知，二维正态分布的边缘分布是一维正态分布，且不依赖于参数ρ. 事实上，通过类似的计算可以证明，若$(X,\ Y)$服从参数为μ_1，μ_2，σ_1，σ_2，ρ的一般二维正态分布，则它们的边缘分布是一维正态分布，即$X\sim N(\mu_1,\ \sigma_1^2)$，$Y\sim N(\mu_2,\ \sigma_2^2)$. 这一事实也说明，边缘分布只考虑了单个分量的情况，未涉及X和Y之间的关系，X和Y之间的关系信息包含在$(X,\ Y)$的联合分布之内. 一般来说，仅由X和Y的边缘分布不能确定二维随机变量$(X,\ Y)$的联合分布.

*3.6　条件分布

在第1章，我们讨论了某事件发生条件下另一事件的条件概率，它是对随机事件而言的. 在本节中，我们讨论随机变量的条件分布. 设有两个随机变量X和Y，在已知Y取定某个值或某些值的条件下，X的分布称为X的条件分布. 类似地，也可以定义Y的条件分布.

例如，从一大群人中随机挑出一个人，分别用X和Y记这个人的体重和身高，则X和Y都是随机变量，它们都有自己的分布. 现在如果限制Y的取值在1.5~1.6米之间，即$1.5\leqslant Y\leqslant 1.6$，在这个限制下求$X$的条件分布，就意味着要从这一大群人中把身高从1.5~1.6米之间的那些人都挑出来，然后在挑出的人群中求其体重的分布. 易知，这个分布与不设限制的无条件分布是不同的，因为我们是把身高Y限制在比较低的人群中来考虑体重X的分布的. 类似地可以考虑限制X取某些值条件下，求Y的条件分布.

3.6.1　离散型随机变量的条件分布

设$(X,\ Y)$是二维离散型随机变量，其分布律为

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

$(X,\ Y)$的分量X和Y的边缘分布律分别为

$$p_{i\cdot}=P\{X=x_i\}=\sum_j p_{ij},\ i=1,\ 2,\ \cdots,$$

$$p_{\cdot j}=P\{Y=y_j\}=\sum_i p_{ij},\ j=1,\ 2,\ \cdots,$$

设$p_{i\cdot}>0$，$p_{\cdot j}>0$，$i,\ j=1,\ 2,\ \cdots$. 现在考虑事件$\{Y=y_j\}$发生条件下X的条件分布，即在$\{Y=y_j\}$发生的条件下，对$i=1,\ 2,\ \cdots$，求事件$\{X=x_i\}$的概率$P\{X=x_i\mid Y=y_j\}$，$i=1,\ 2,\ \cdots$. 由条件概率的定义，

$$P\{X=x_i\mid Y=y_j\}=\frac{P\{X=x_i,\ Y=y_j\}}{P\{Y=y_j\}}=\frac{p_{ij}}{p_{\cdot j}},\ i=1,\ 2,\ \cdots.$$

容易看出，上述条件概率具有离散型随机变量分布律的两条性质.

1. $P\{X=x_i\mid Y=y_j\}\geqslant 0$，$i=1,\ 2,\ \cdots$；

2. $\sum\limits_i P\{X=x_i \mid Y=y_j\}=1$.

类似地可以讨论 $P\{Y=y_j \mid X=x_i\}$，$j=1,2,\cdots$.

定义 3.6.1 设(X,Y)为二维离散型随机变量. 对于固定的j，若$p_{\cdot j}>0$，则称

$$P\{X=x_i \mid Y=y_j\}=\frac{p_{ij}}{p_{\cdot j}},\quad i=1,2,\cdots \tag{3.6.1}$$

为$Y=y_j$条件下X的条件分布律. 对于固定的i，若$p_{i\cdot}>0$，则称

$$P\{Y=y_j \mid X=x_i\}=\frac{p_{ij}}{p_{i\cdot}},\quad j=1,2,\cdots \tag{3.6.2}$$

为$X=x_i$条件下Y的条件分布律.

例 3.6.1 设(X,Y)的联合分布律如在例3.2.1中，求

(1) 在$Y=1$条件下，X的条件分布律；

(2) 在$X=1$条件下，Y的条件分布律.

解 (1) X和Y的边缘分布律在例3.5.1中已求出，由式(3.6.1)可得

$$P\{X=1 \mid Y=1\}=\frac{1}{4}\Big/\frac{25}{48}=\frac{12}{25},$$

$$P\{X=2 \mid Y=1\}=\frac{1}{8}\Big/\frac{25}{48}=\frac{6}{25},$$

$$P\{X=3 \mid Y=1\}=\frac{1}{12}\Big/\frac{25}{48}=\frac{4}{25},$$

$$P\{X=4 \mid Y=1\}=\frac{1}{16}\Big/\frac{25}{48}=\frac{3}{25}.$$

即在$Y=1$条件下，X的条件分布律为

X	1	2	3	4
$\frac{p_{i1}}{p_{\cdot 1}}$	$\frac{12}{25}$	$\frac{6}{25}$	$\frac{4}{25}$	$\frac{3}{25}$

(2) 同理可得，在$X=1$的条件下，Y的条件分布律为

Y	1	2	3	4
$\frac{p_{1j}}{p_{1\cdot}}$	1	0	0	0

类似地，可以分别求出$Y=2$，3，4条件下X的条件分布律，以及$X=2$，3，4条件下Y的条件分布律(请自己练习).

3.6.2 连续型随机变量的条件分布

设(X,Y)是二维连续型随机变量. 由于对任意x，y，$P\{X=x\}=0$，$P\{Y=y\}=0$，因此不能够像离散型那样引入条件分布. 下面我们用求极限的方法来推导条件分布函数.

给定y，设对任意的$\varepsilon>0$，概率$P\{y-\varepsilon<Y\leqslant y+\varepsilon\}>0$，于是对于任意的$x$，

$$P\{X \leqslant x \mid y-\varepsilon < Y \leqslant y+\varepsilon\} = \frac{P\{X \leqslant x,\ y-\varepsilon < Y \leqslant y+\varepsilon\}}{P\{y-\varepsilon < Y \leqslant y+\varepsilon\}},$$

它是在条件 $y-\varepsilon < Y \leqslant y+\varepsilon$ 下 X 的条件分布函数. 若 $\varepsilon \to 0$ 时上式极限存在，则称该极限为在条件 $Y=y$ 下 X 的条件分布函数，记为 $P\{X \leqslant x \mid Y=y\}$ 或 $F_{X|Y}(x \mid y)$，即

$$F_{X|Y}(x \mid y) = \lim_{\varepsilon \to 0} \frac{P\{X \leqslant x,\ y-\varepsilon < Y \leqslant y+\varepsilon\}}{P\{y-\varepsilon < Y \leqslant y+\varepsilon\}} \tag{3.6.3}$$

在条件 $Y=y$ 下，若存在 x 的函数 $f_{X|Y}(x \mid y) \geqslant 0$，使得

$$F_{X|Y}(x \mid y) = \int_{-\infty}^{x} f_{X|Y}(u \mid y)\,\mathrm{d}u,$$

则称 $f_{X|Y}(x \mid y)$ 为在条件 $Y=y$ 下 X 的条件密度函数，简称为条件密度.

定理 3.6.1　设二维连续型随机变量 (X, Y) 的联合密度为 $f(x, y)$，Y 的边缘密度为 $f_Y(y)$. 若 $f(x, y)$ 在点 (x, y) 处连续，$f_Y(y)$ 在点 y 处连续，且 $f_Y(y) > 0$，则在条件 $Y=y$ 下 X 的条件密度函数为

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)} \tag{3.6.4}$$

证　设 (X, Y) 的联合分布函数为 $F(x, y)$，Y 的边缘分布函数为 $F_Y(y)$，由式 (3.6.3) 有

$$\begin{aligned} F_{X|Y}(x \mid y) &= \lim_{\varepsilon \to 0} \frac{P\{X \leqslant x,\ y-\varepsilon < Y \leqslant y+\varepsilon\}}{P\{y-\varepsilon < Y \leqslant y+\varepsilon\}} = \lim_{\varepsilon \to 0} \frac{F(x,\ y+\varepsilon)-F(x,\ y-\varepsilon)}{F_Y(y+\varepsilon)-F_Y(y-\varepsilon)} \\ &= \frac{\lim\limits_{\varepsilon \to 0} \dfrac{F(x,\ y+\varepsilon)-F(x,\ y-\varepsilon)}{2\varepsilon}}{\lim\limits_{\varepsilon \to 0} \dfrac{F_Y(y+\varepsilon)-F_Y(y-\varepsilon)}{2\varepsilon}} = \frac{\dfrac{\partial F(x,\ y)}{\partial y}}{\dfrac{\mathrm{d}F_Y(y)}{\mathrm{d}y}} \\ &= \frac{\int_{-\infty}^{x} f(u,\ y)\,\mathrm{d}u}{f_Y(y)} = \int_{-\infty}^{x} \frac{f(u,\ y)}{f_Y(y)}\mathrm{d}u, \end{aligned}$$

从而

$$f_{X|Y}(x \mid y) = \frac{f(x,\ y)}{f_Y(y)}.$$

类似地可以定义 $F_{Y|X}(y \mid x)$ 和 $f_{Y|X}(y \mid x)$，且可证明，当 $f(x, y)$ 在点 (x, y) 处连续，$f_X(x)$ 在 x 处连续，且 $f_X(x) > 0$ 时，

$$f_{Y|X}(y \mid x) = \frac{f(x,\ y)}{f_X(x)}. \tag{3.6.5}$$

例 3.6.2　设二维随机变量 (X, Y) 服从单位圆域 $x^2+y^2 \leqslant 1$ 上的均匀分布，求条件密度 $f_{X|Y}(x \mid y)$ 和 $f_{Y|X}(y \mid x)$.

解　(X, Y) 的联合密度函数为

$$f(x,\ y) = \begin{cases} \dfrac{1}{\pi}, & x^2+y^2 \leqslant 1, \\ 0, & \text{其他}. \end{cases}$$

Y 的边缘密度为

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

$$=\begin{cases}\dfrac{1}{\pi}\displaystyle\int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}}\mathrm{d}x, & -1\leqslant y\leqslant 1,\\ 0, & \text{其他.}\end{cases}=\begin{cases}\dfrac{2}{\pi}\sqrt{1-y^2}, & -1\leqslant y\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

于是当 $-1<y<1$ 时，$f_Y(y)>0$. 因此由式(3.6.4)，当 $-1<y<1$ 时，

$$f_{X|Y}(x\mid y)=\begin{cases}\dfrac{1/\pi}{(2/\pi)\sqrt{1-y^2}}, & -\sqrt{1-y^2}\leqslant x\leqslant\sqrt{1-y^2},\\ 0, & \text{其他.}\end{cases}$$

$$=\begin{cases}\dfrac{1}{2\sqrt{1-y^2}}, & -\sqrt{1-y^2}\leqslant x\leqslant\sqrt{1-y^2},\\ 0, & \text{其他.}\end{cases}$$

这里条件 $-\sqrt{1-y^2}\leqslant x\leqslant\sqrt{1-y^2}$ 是由 $f(x,\ y)=\dfrac{1}{\pi}$ 的条件 $x^2+y^2\leqslant 1$ 确定的. 因为现在给定了 $y(-1<y<1)$，于是 x 满足 $x^2+y^2\leqslant 1$ 等价于 $-\sqrt{1-y^2}\leqslant x\leqslant\sqrt{1-y^2}$.

同理可得，当 $-1<x<1$ 时，$f_X(x)>0$，且有条件密度

$$f_{Y|X}(y\mid x)=\begin{cases}\dfrac{1}{2\sqrt{1-x^2}}, & -\sqrt{1-x^2}\leqslant y\leqslant\sqrt{1-x^2},\\ 0, & \text{其他.}\end{cases}$$

特别地，当 $y=0$ 时，X 的条件密度为

$$f_{X|Y}(x\mid 0)=\begin{cases}\dfrac{1}{2}, & -1\leqslant x\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

即 X 的条件分布为$[-1,\ 1]$ 上的均匀分布.

例 3.6.3 设二维随机变量$(X,\ Y)$ 的联合密度函数为

$$f(x,\ y)=\begin{cases}\dfrac{21}{4}x^2y, & x^2\leqslant y\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

求条件密度 $f_{X|Y}(x\mid y)$，$f_{Y|X}(y\mid x)$ 和条件概率 $P\left\{Y>\dfrac{3}{4}\,\middle|\,X=\dfrac{1}{2}\right\}$.

解 X 和 Y 的边缘密度分别为

$$f_X(x)=\int_{-\infty}^{\infty}f(x,\ y)\,\mathrm{d}y=\begin{cases}\dfrac{21}{8}x^2(1-x^4), & -1\leqslant x\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

$$f_Y(y)=\int_{-\infty}^{\infty}f(x,\ y)\,\mathrm{d}x=\begin{cases}\dfrac{7}{2}y^{\frac{5}{2}}, & 0\leqslant y\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

从而当 $0<y\leqslant 1$ 时，$f_Y(y)\neq 0$，此时 X 的条件密度为

$$f_{X|Y}(x\mid y)=\frac{f(x,\ y)}{f_Y(y)}=\begin{cases}\dfrac{3}{2}x^2y^{-\frac{3}{2}}, & -\sqrt{y}\leqslant x\leqslant\sqrt{y},\\ 0, & \text{其他.}\end{cases}$$

当 $-1<x<1$ 时，$f_X(x)\neq 0$，此时 Y 的条件密度为

$$f_{Y|X}(y\mid x)=\frac{f(x,\ y)}{f_X(x)}=\begin{cases}\dfrac{2y}{1-x^4}, & x^2\leqslant y\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

特别地，对 $x=\dfrac{1}{2}$，有

$$f_{Y|X}\left(y\,\middle|\,\frac{1}{2}\right)=\begin{cases}\dfrac{32}{15}y, & \dfrac{1}{4}\leqslant y\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

从而

$$P\left\{Y>\frac{3}{4}\,\middle|\,X=\frac{1}{2}\right\}=\int_{\frac{3}{4}}^{1}f_{Y|X}\left(y\,\middle|\,\frac{1}{2}\right)\mathrm{d}y=\int_{\frac{3}{4}}^{1}\frac{32}{15}y\mathrm{d}y=\frac{7}{15}.$$

3.7　随机变量的独立性

下面我们利用第一章中两个随机事件的相互独立性概念，引出随机变量的相互独立性概念.

定义 3.7.1　设 $F(x,\ y)$ 及 $F_X(x)$，$F_Y(y)$ 分别是二维随机变量 $(X,\ Y)$ 的联合分布函数及边缘分布函数，若对任意的实数 x，y，都有

$$F(x,\ y)=F_X(x)F_Y(y)\tag{3.7.1}$$

则称随机变量 X 与 Y 相互独立.

由分布函数的定义，(3.7.1) 式可以写为

$$P\{X\leqslant x,\ Y\leqslant y\}=P\{X\leqslant x\}P\{Y\leqslant y\}\tag{3.7.2}$$

因此，随机变量 X 与 Y 相互独立是指对任意实数 x，y，随机事件 $\{X\leqslant x\}$ 与 $\{Y\leqslant y\}$ 相互独立.

设 $(X,\ Y)$ 是二维离散型随机变量，其所有可能取的值为 $(x_i,\ y_j)$，$i,\ j=1,\ 2,\ \cdots$，则 X 与 Y 相互独立的条件可以写为

$$P\{X=x_i,\ Y=y_j\}=P\{X=x_i\}P\{Y=y_j\},\ i,\ j=1,\ 2,\ \cdots\tag{3.7.3}$$

或者

$$p_{ij}=p_{i\cdot}p_{\cdot j},\ i,\ j=1,\ 2,\ \cdots\tag{3.7.4}$$

该等式是判断二维离散型随机变量 $(X,\ Y)$ 的分量 X 与 Y 是否独立的常用方法，即当等式 $p_{ij}=p_{i\cdot}p_{\cdot j}$ 对所有 $i,\ j=1,\ 2,\ \cdots$ 都成立时，则可判断 X 与 Y 独立，否则若存在某 $(i,\ j)$，使 $p_{ij}\neq p_{i\cdot}p_{\cdot j}$，则 X 与 Y 不独立.

下面定理给出判断连续型随机变量 $(X,\ Y)$ 的分量 X 与 Y 是否独立的充分必要条件.

定理 3.7.1　设 $(X,\ Y)$ 是二维连续型随机变量，$f(x,\ y)$ 和 $f_X(x)$，$f_Y(y)$ 分别是 $(X,\ Y)$

的联合密度函数和边缘密度函数，则 X 与 Y 相互独立的充分必要条件是，在 $f(x, y)$ 的连续点 (x, y) 处均有

$$f(x, y) = f_X(x) f_Y(y) \tag{3.7.5}$$

证明　设式(3.7.1)成立，即 X 与 Y 独立. 将等式(3.7.1)在 $f(x, y)$ 的连续点 (x, y) 处关于 x 和 y 分别求导数，由二维分布函数与密度函数的关系及一维分布函数与密度函数的关系，即得式(3.7.5). 反之，若式(3.7.5)在 $f(x, y)$ 的连续点 (x, y) 处均成立，则对该等式两边求积分得

$$\begin{aligned} F(x, y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f(s, t)\,\mathrm{d}s\mathrm{d}t \\ &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(s) f_Y(t)\,\mathrm{d}s\mathrm{d}t \\ &= \int_{-\infty}^{x} f_X(s)\,\mathrm{d}s \int_{-\infty}^{x} f_Y(t)\,\mathrm{d}t \\ &= F_X(x) F_Y(y), \end{aligned}$$

即式(3.7.1)成立，因此 X 与 Y 独立. 证毕.

式(3.7.5)是判断二维连续型随机变量 (X, Y) 的分量 X 与 Y 是否独立的基本方法，即当 $f(x, y) = f_X(x) f_Y(y)$ 在 $f(x, y)$ 的连续点处均成立时，则可判断 X 与 Y 独立. 否则，若该等式不成立，则可判断 X 与 Y 不独立.

例 3.7.1　考察例 3.2.2(即吸烟与得肺癌关系的研究)中随机变量的独立性.

解　由例3.5.2，$P\{X=0\}=0.2$，$P\{Y=0\}=0.000\,17$，而 $P\{X=0, Y=0\}=0.000\,13$，显然 $P\{X=0, Y=0\} \neq P\{X=0\}P\{Y=0\}$，因此 X 与 Y 不独立.

例 3.7.2　设随机变量 X 与 Y 相互独立，下表列出了二维随机变量 (X, Y) 联合分布律及关于 X 和关于 Y 的边缘分布律中的部分数值，试将其余数值填入表中的空白处.

X \ Y	y_1	y_2	y_3	$P\{X=x_i\}=P_i$
x_1		1/8		
x_2	1/8			
$P\{y=y_j\}=p_j$	1/6			1

解　首先由 $P\{Y=y_1\}=P\{X=x_1, Y=y_1\}+P\{X=x_2, Y=y_1\}$ 得

$$P\{X=x_1, Y=y_1\}=P\{Y=y_1\}-P\{X=x_2, Y=y_1\}=1/6-1/8=1/24.$$

又由于 X 与 Y 相互独立，可知有 $P\{X=x_1\}P\{Y=y_1\}=P\{X=x_1, Y=y_1\}$，所以

$$P\{X=x_1\}=\frac{P\{X=x_1, Y=y_1\}}{P\{Y=y_1\}}=\frac{1/24}{1/6}=\frac{1}{4}.$$

其他数值可类似地求出，例如

$$P\{X=x_1, Y=y_3\}=\frac{1}{4}-\frac{1}{24}-\frac{1}{8}=\frac{1}{12},$$

$$P\{X=x_2\}=1-\frac{1}{4}=\frac{3}{4},\quad P\{Y=y_2\}=\frac{1/8}{1/4}=\frac{1}{2}.$$

将所有数值填入表中空白处后得到下表.

X \ Y	y_1	y_2	y_3	$P\{X=x_i\}=P_i$
x_1	1/24	1/8	1/12	1/4
x_2	1/8	3/8	1/4	3/4
$P\{y=y_j\}=p_j$	1/6	1/2	1/3	1

例 3.7.3　设 X 与 Y 相互独立，它们的密度函数分别为

$$f_X(x)=\begin{cases}e^{-x}, & x>0,\\ 0, & x\leqslant 0;\end{cases}\qquad f_Y(y)=\begin{cases}e^{-y}, & y>0,\\ 0, & y\leqslant 0.\end{cases}$$

求二维随机变量(X, Y) 的联合密度函数.

解　由式(3.7.5)，(X, Y) 的联合密度为

$$f(x, y)=f_X(x)f_Y(y)=\begin{cases}e^{-(x+y)}, & x>0,\ y>0,\\ 0, & \text{其他}.\end{cases}$$

例 3.7.4　设(X, Y) 的联合密度为

(1) $f(x, y)=\begin{cases}xe^{-(x+y)}, & x>0,\ y>0,\\ 0, & \text{其他};\end{cases}$

(2) $f(x, y)=\begin{cases}2, & 0<x<y,\ 0<y<1,\\ 0, & \text{其他}.\end{cases}$

问 X 与 Y 是否独立?

解　(1) 由于

$$f_X(x)=\int_{-\infty}^{\infty}f(x, y)\,dy=\begin{cases}\int_0^{+\infty}xe^{-(x+y)}\,dy, & x>0,\\ 0, & x\leqslant 0,\end{cases}=\begin{cases}xe^{-x}, & x>0,\\ 0, & x\leqslant 0,\end{cases}$$

$$f_Y(x)=\int_{-\infty}^{\infty}f(x, y)\,dx=\begin{cases}\int_0^{+\infty}xe^{-(x+y)}\,dx, & y>0,\\ 0, & y\leqslant 0,\end{cases}=\begin{cases}e^{-y}, & y>0,\\ 0, & y\leqslant 0,\end{cases}$$

可知$f(x, y)=f_X(x)f_Y(y)$，故 X 与 Y 独立.

(2) 由于

$$f_X(x)=\int_{-\infty}^{\infty}f(x, y)\,dy=\begin{cases}\int_x^1 2\,dy, & 0<x<1,\\ 0, & \text{其他},\end{cases}=\begin{cases}2(1-x), & 0<x<1,\\ 0, & \text{其他},\end{cases}$$

$$f_Y(x)=\int_{-\infty}^{\infty}f(x, y)\,dx=\begin{cases}\int_0^y 2\,dx, & 0<y<1,\\ 0, & \text{其他},\end{cases}=\begin{cases}2y, & 0<y<1,\\ 0, & \text{其他},\end{cases}$$

可知$f(x, y)\neq f_X(x)f_Y(y)$，故 X 与 Y 不独立.

*3.8 随机变量函数的分布

在第二章中我们讨论了一维随机变量函数的分布，现在我们讨论二维随机变量函数的分布. 对于二维随机变量(X, Y)，其两个分量X和Y的函数$Z=g(X, Y)$是一个一维随机变量，我们希望通过(X, Y)的分布求Z的分布.

3.8.1 离散型随机变量函数的分布

设(X, Y)是二维离散型随机变量，其分布律为

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

则$Z=g(X, Y)$是一维离散型随机变量，它的分布律为

$$P\{Z=g(x_i, y_j)\}=p_{ij}, \ i, j=1, 2, \cdots \tag{3.8.1}$$

若对于不同的(x_i, y_j)，$g(x, y)$有相同的值，则Z取这些相同值的概率必须合并. 当Z的可能取值不多时，为了求它的分布律，可以将(X, Y)的取值列成一行，其相应概率列成另一行，然后将$g(X, Y)$的值列成一行，即可根据相应的概率求出Z的分布律.

例 3.8.1 已知(X, Y)的分布律为

X \ Y	-1	1	2
-1	$\frac{5}{20}$	$\frac{2}{20}$	$\frac{6}{20}$
2	$\frac{3}{20}$	$\frac{3}{20}$	$\frac{1}{20}$

求(1) $Z=2X-Y$的分布律；(2) $Z=X+Y$的分布律.

解 首先由式(3.7.1)，可得下表，并由此得到有关函数Z的分布律.

$2X-Y$	-1	-3	-4	5	3	2
$X+Y$	-2	0	1	1	3	4
(X, Y)	$(-1, -1)$	$(-1, 1)$	$(-1, 2)$	$(2, -1)$	$(2, 1)$	$(2, 2)$
p_{ij}	$\frac{5}{20}$	$\frac{2}{20}$	$\frac{6}{20}$	$\frac{3}{20}$	$\frac{3}{20}$	$\frac{1}{20}$

(1) $Z=2X-Y$的分布律为

$2X-Y$	-4	-3	-1	2	3	5
p_i	$\frac{6}{20}$	$\frac{2}{20}$	$\frac{5}{20}$	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{3}{20}$

(2) $Z=X+Y$的分布律为

$X+Y$	-2	0	1	3	4
p_i	$\frac{5}{20}$	$\frac{2}{20}$	$\frac{9}{20}$	$\frac{3}{20}$	$\frac{1}{20}$

例 3.8.2　设 X 与 Y 独立，它们分布律分别为

$$P\{X=k\}=p(k),\ k=0,\ 1,\ 2,\ \cdots,$$
$$P\{Y=r\}=q(r),\ r=0,\ 1,\ 2,\ \cdots,$$

试求 $Z=X+Y$ 的分布律.

解　由基本事件的互斥性和 X 与 Y 的独立性得

$$\begin{aligned}P\{Z=i\}&=P\{X+Y=i\}\\&=P\{\{X=0,\ Y=i\}\cup\{X=1,\ Y=i-1\}\cup\cdots\cup\{X=i,\ Y=0\}\}\\&=P\{X=0,\ Y=i\}+P\{X=1,\ Y=i-1\}+\cdots P\{X=i,\ Y=0\}\\&=\sum_{k=0}^{i}P\{X=k,\ Y=i-k\}\\&=\sum_{k=0}^{i}P\{X=k\}P\{Y=i-k\}\\&=\sum_{k=0}^{i}p(k)q(i-k),\ i=0,\ 1,\ 2,\ \cdots.\end{aligned}$$

即 $Z=X+Y$ 的分布律为

$$P\{Z=i\}=\sum_{k=0}^{i}p(k)q(i-k),\ i=0,\ 1,\ 2,\ \cdots \tag{3.8.2}$$

式(3.8.2)称为**离散卷积公式**.

例 3.8.3　设 X 与 Y 独立，它们分别服从参数为 λ_1 和 λ_2 的泊松分布，即 $X\sim P(\lambda_1)$ 和 $Y\sim P(\lambda_2)$，证明 $Z=X+Y$ 服从参数为 $\lambda=\lambda_1+\lambda_2$ 的泊松分布.

证　X 和 Y 的分布律分别为

$$P\{X=k\}=\frac{\lambda_1^k}{k!}e^{-\lambda_1},\ k=0,\ 1,\ 2,\ \cdots;$$
$$P\{Y=r\}=\frac{\lambda_2^r}{r!}e^{-\lambda_2},\ r=0,\ 1,\ 2,\ \cdots.$$

利用离散卷积公式(3.8.2)得到

$$\begin{aligned}P\{Z=i\}&=P\{X+Y=i\}=\sum_{k=0}^{i}\frac{\lambda_1^k}{k!}e^{-\lambda_1}\cdot\frac{\lambda_2^{i-k}}{(i-k)!}e^{-\lambda_2}\\&=\frac{1}{i!}e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{i}\frac{i!\ \lambda_1^k\lambda_2^{i-k}}{k!\ (i-k)!}\\&=\frac{1}{i!}e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{i}C_i^k\lambda_1^k\lambda_2^{i-k}\\&=\frac{(\lambda_1+\lambda_2)^i}{i!}e^{-(\lambda_1+\lambda_2)},\ i=0,\ 1,\ 2,\ \cdots.\end{aligned}$$

所以 $Z\sim P(\lambda_1+\lambda_2)$. 证毕.

3.8.2 连续型随机变量函数的分布

设(X, Y)是二维连续型随机变量，其联合密度函数为$f(x, y)$. 若$Z=g(X, Y)$是一个连续函数，则一般来说它也是一个连续型随机变量，我们希望求它的密度函数$f_Z(z)$. 可用类似于求一维随机变量函数分布的方法求$f_Z(z)$. 基本步骤如下：

(a) 求分布函数$F_Z(z)$.

$$F_Z(z)=P\{Z\leqslant z\}=P\{g(X, Y)\leqslant z\}=P\{(X, Y)\in D_Z\}=\iint_{D_Z} f(x, y)\,\mathrm{d}x\mathrm{d}y.$$

其中，$D_Z=\{(x, y)\mid g(x, y)\leqslant z\}$.

(b) 求Z的密度函数$f_Z(z)=\dfrac{\mathrm{d}F_Z(z)}{\mathrm{d}z}$.

下面我们讨论两种特殊函数的分布.

1. $Z=X+Y$的分布

$Z=X+Y$的分布函数为$F_Z(z)=P\{X+Y\leqslant z\}=P\{(X, Y)\in D\}$，其中$D$可由图3.6中阴影部分区域直观表示. 由二维连续型随机变量的性质，有

$$F_Z(z)=P\{X+Y\leqslant z\}=\iint_D f(x, y)\,\mathrm{d}x\mathrm{d}y$$

$$=\iint_{x+y\leqslant z} f(x, y)\,\mathrm{d}x\mathrm{d}y=\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{z-x} f(x, y)\,\mathrm{d}y\right]\mathrm{d}x.$$

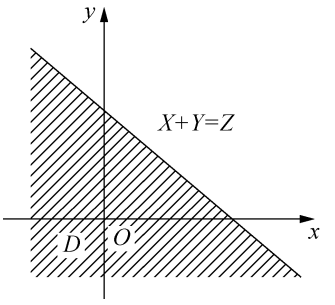


图 3.6 和函数积分区域

在积分$\int_{-\infty}^{z-x} f(x, y)\,\mathrm{d}y$中，作变量代换，令$u=y+x$，得

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

于是，

$$F_Z(z)=\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{z} f(x, u-x)\,\mathrm{d}u\right]\mathrm{d}x=\int_{-\infty}^{z}\left[\int_{-\infty}^{+\infty} f(x, u-x)\,\mathrm{d}x\right]\mathrm{d}u,$$

上式关于z求导数，或由分布函数与密度函数的关系，即得Z的密度函数

$$f_Z(z)=\int_{-\infty}^{+\infty} f(x, z-x)\,\mathrm{d}x \tag{3.8.3}$$

由X和Y地位的对称性可知，Z的密度函数也可以由下式给出

$$f_Z(z)=\int_{-\infty}^{+\infty} f(z-y, y)\,\mathrm{d}y \tag{3.8.4}$$

特别地，若X与Y独立，设X和Y的边缘密度分别为$f_X(x)$和$f_Y(y)$，则式(3.8.3)和式(3.8.4)可分别写成

$$f_Z(z)=\int_{-\infty}^{+\infty} f_X(x)f_Y(z-x)\,\mathrm{d}x \tag{3.8.5}$$

$$f_Z(z)=\int_{-\infty}^{+\infty} f_X(z-y)f_Y(y)\,\mathrm{d}y \tag{3.8.6}$$

式(3.8.5)和式(3.8.6)称为**连续卷积公式**.

例 3.8.4(正态分布的可加性)　设 X 与 Y 独立，且它们同服从标准正态分布 $N(0,1)$，求 $Z=X+Y$ 的密度函数.

解　X 和 Y 的密度函数分别为

$$f_X(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}},\ f_Y(y)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}.$$

由 X 与 Y 的独立性，应用式(3.8.5)得 $Z=X+Y$ 的密度函数

$$\begin{aligned}f_Z(z)&=\int_{-\infty}^{+\infty}f_X(x)f_Y(z-x)\,\mathrm{d}x\\&=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{(z-x)^2}{2}}\mathrm{d}x\\&=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{x^2+(z-x)^2}{2}}\mathrm{d}x,\end{aligned}$$

因为

$$\frac{x^2+(z-x)^2}{2}=\frac{z^2}{4}+\left(x-\frac{z}{2}\right)^2,$$

于是

$$\begin{aligned}f_Z(z)&=\frac{1}{2\pi}\mathrm{e}^{-\frac{z^2}{4}}\int_{-\infty}^{+\infty}\mathrm{e}^{-(x-\frac{z}{2})^2}\mathrm{d}x\\&=\frac{\frac{1}{\sqrt{2}}}{\sqrt{2\pi}}\mathrm{e}^{-\frac{z^2}{4}}\int_{-\infty}^{+\infty}\frac{\sqrt{2}}{\sqrt{2\pi}}\mathrm{e}^{-(x-\frac{z}{2})^2}\mathrm{d}x\\&=\frac{1}{\sqrt{2\pi}\sqrt{2}}\mathrm{e}^{-\frac{z^2}{4}},\end{aligned}$$

其中第 2 个等号后面的积分值等于 1，因为被积函数是正态分布 $N\left(\frac{z}{2},\frac{1}{2}\right)$ 的密度函数，最后得到的函数恰为正态分布 $N(0,2)$ 的密度函数，即 $Z=X+Y\sim N(0,2)$.

一般地，若 X 和 Y 相互独立，且 $X\sim N(\mu_1,\sigma_1^2)$，$Y\sim N(\mu_2,\sigma_2^2)$，可以证明，$Z=X+Y$ 服从正态分布 $N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$. 事实上，可以证明，两个正态随机变量的线性组合仍然服从正态分布.

例 3.8.5　设某种商品一周的需求量 X 是一个连续型随机变量，其密度函数为

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

如果各周的需求量之间相互独立，求两周需求量的密度函数.

解　分别用 X 和 Y 表示第一、二周的需求量，则它们相互独立，其密度函数分别为

$$f_X(x)=\begin{cases}x\mathrm{e}^{-x}, & x>0,\\0, & x\leqslant 0;\end{cases}$$

$$f_Y(y)=\begin{cases}y\mathrm{e}^{-y}, & y>0,\\0, & y\leqslant 0.\end{cases}$$

两周需求量为 $Z=X+Y$，由式(3.8.5)，其密度函数 $f_Z(z)=\int_{-\infty}^{+\infty}f_X(x)f_Y(z-x)\,\mathrm{d}x$.

当 $z\leqslant 0$ 时，若 $x>0$，则 $z-x<0$，$f_Y(z-x)=0$；若 $x\leqslant 0$，则 $f_X(x)=0$，从而 $f_Z(z)=0$. 当 $z>0$ 时，若 $x\leqslant 0$，则 $f_X(x)=0$；若 $x>0$ 且 $z-x\leqslant 0$，即 $z\leqslant x$，则 $f_Y(z-x)=0$，因此只有当 $0<x<z$ 时被积函数才可能非零，即有

$$f_Z(z)=\int_{-\infty}^{+\infty}f_X(x)f_Y(z-x)\,\mathrm{d}x=\int_0^z x\mathrm{e}^{-x}(z-x)\,\mathrm{e}^{-(z-x)}\,\mathrm{d}x=\frac{z^3}{6}\mathrm{e}^{-z},$$

从而

$$f_Z(z)=\begin{cases}\dfrac{z^3}{6}\mathrm{e}^{-z}, & z>0,\\ 0, & z\leqslant 0.\end{cases}$$

习题三

1. 设二维随机变量 $(X,\ Y)$ 的分布函数为

$$F(x,\ y)=\begin{cases}1-2^{-x}-2^{-y}+2^{-x-y}, & x\geqslant 0,\ y\geqslant 0,\\ 0, & \text{其他.}\end{cases}$$

求概率 $P\{1<X\leqslant 2,\ 3<Y\leqslant 5\}$.

2. 袋中有 5 只球(2 只白球，3 只红球)，现进行有放回与无放回抽球两次，每次抽一只，定义随机变量

$$X=\begin{cases}0, & \text{第一次抽到红球},\\ 1, & \text{第一次抽到白球}.\end{cases}$$

$$Y=\begin{cases}0, & \text{第二次抽到红球},\\ 1, & \text{第二次抽到白球}.\end{cases}$$

试就有放回和无放回摸球情况分别求 $(X,\ Y)$ 的联合分布律.

3. 求习题 2 中的 $(X,\ Y)$ 的边缘分布律，并就所得结果讨论联合分布律与边缘分布律的关系.

4. 设二维随机变量 $(X,\ Y)$ 只能取 $(-1,\ 0)$，$(0,\ 0)$ 和 $(0,\ 1)$ 三对数，且取这些数的概率分别是 $\frac{1}{2}$，$\frac{1}{3}$ 和 $\frac{1}{6}$.

(1) 写出 $(X,\ Y)$ 的联合分布律；

(2) 求联合分布函数 $F(x,\ y)$.

5. 设 X 与 Y 独立，它们的分布律分别由下表给出，求 $(X,\ Y)$ 的联合分布律.

X	-2	-1	0	$\frac{1}{2}$
$p_{i\cdot}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{3}$

Y	$-\frac{1}{2}$	1	3
$p_{\cdot j}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

6. 设二维随机变量(X, Y)等可能地取下列值：(1, 1)，(1, 3)，(2, 2)，(2, 4)，(3, 1)，(4, 2)，试求

(1) 在$Y=4$条件下，X的条件分布律；

(2) 在$X=2$条件下，Y的条件分布律.

7. 设二维随机变量(X, Y)的密度函数为

$$f(x, y)=\begin{cases}a(6-x-y), & 0\leqslant x\leqslant 1,\ 0\leqslant y\leqslant 2,\\ 0, & \text{其他},\end{cases}$$

(1) 确定常数a；

(2) 求概率$P\{X\leqslant 0.5,\ Y\leqslant 1.5\}$；

(3) 求概率$P\{(X, Y)\in D\}$，这里D是由$x=0$、$y=0$和$x+y=1$这3条直线所围成的三角形区域.

8. 设二维随机变量(X, Y)的密度函数为

$$f(x, y)=\begin{cases}2e^{-(2x+y)}, & x>0,\ y>0,\\ 0, & \text{其他}.\end{cases}$$

(1) 求分布函数$F(x, y)$；

(2) 求概率$P\{Y\leqslant X\}$.

9. 向一个无限平面靶射击，设命中点(X, Y)的密度函数为

$$f(x, y)=\frac{1}{\pi(1+x^2+y^2)^2},\quad -\infty<x,\ y<\infty.$$

求命中点与靶心(坐标原点)的距离不超过a的概率.

10. 设二维随机变量(X, Y)在区域B上服从均匀分布，B是由x轴、y轴及直线$y=2x+1$所围成的三角形区域，试求

(1) (X, Y)的密度函数$f(x, y)$；

(2) (X, Y)的分布函数$F(x, y)$.

11. 分别求习题10中随机变量X和Y的边缘密度函数，并求条件概率$P\left\{-\frac{1}{4}<X\leqslant 0\,\middle|\,\frac{1}{2}<Y\leqslant 1\right\}$.

12. 设二维随机变量(X, Y)的密度函数为

$$f(x, y)=\begin{cases}\frac{3}{2}xy^2, & 0\leqslant x\leqslant 2,\ 0\leqslant y\leqslant 1,\\ 0, & \text{其他}.\end{cases}$$

求边缘密度函数$f_X(x)$和$f_Y(y)$.

13. 设二维随机变量(X, Y)的密度函数为

$$f(x,\ y)=\begin{cases}4.8y(2-x), & 0\leqslant x\leqslant 1,\ 0\leqslant y\leqslant x,\\ 0, & \text{其他.}\end{cases}$$

求边缘密度函数$f_X(x)$和$f_Y(y)$.

14. 设二维随机变量$(X,\ Y)$在区域D上服从均匀分布，试求当$X=x(0\leqslant x<1)$时，Y的条件分布密度. 其中D是由x轴、y轴及直线$y=2(1-x)$所围成的区域.

15. 设二维随机变量$(X,\ Y)$的密度函数为

$$f(x,\ y)=\begin{cases}x^2+\dfrac{xy}{3}, & 0\leqslant x\leqslant 1,\ 0\leqslant y\leqslant 2,\\ 0, & \text{其他.}\end{cases}$$

求条件密度$f_{X|Y}(x\mid y)$和$f_{Y|X}(y\mid x)$及概率$P\left\{Y<\dfrac{1}{2}\,\middle|\,X=\dfrac{1}{2}\right\}$.

16. 判断习题3中随机变量X与Y的独立性.

17. 二维随机变量$(X,\ Y)$的分布律由下表给出.

X \ Y	1	2	3
1	$\frac{1}{6}$	$\frac{1}{9}$	$\frac{1}{18}$
2	$\frac{1}{3}$	$\frac{1}{a}$	$\frac{1}{b}$

问当a，b取何值时，X与Y独立?

18. 判断习题12和习题13中随机变量X与Y的独立性.

19. 设随机变量X与Y独立，它们均服从$[-1,\ 1]$上的均匀分布，求二次方程$t^2+Xt+Y=0$有实根的概率.

20. 设二维随机变量$(X,\ Y)$的密度函数为

$$f(x,\ y)=\begin{cases}\dfrac{xe^{-x}}{(1+y)^2}, & x>0,\ y>0,\\ 0, & \text{其他.}\end{cases}$$

讨论X与Y的独立性.

21. 设二维随机变量$(X,\ Y)$的分布函数为

$$F(x,\ y)=\begin{cases}1-e^{-x}-e^{-y}+e^{-(x+y)}, & x\geqslant 0,\ y\geqslant 0,\\ 0, & \text{其他.}\end{cases}$$

讨论X与Y的独立性.

22. 设X与Y独立，它们的密度函数分别为

$$f_X(x)=\begin{cases}1, & 0\leqslant x\leqslant 1,\\ 0, & \text{其他.}\end{cases}\qquad f_Y(y)=\begin{cases}e^{-y}, & y>0,\\ 0, & y\leqslant 0.\end{cases}$$

求$Z=X+Y$的密度函数.

23. 设二维随机变量$(X,\ Y)$的密度函数为

$$f(x,\ y)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}},\quad -\infty<x<+\infty,\quad -\infty<y<+\infty,$$

求 $Z = X^2 + Y^2$ 的密度函数.

24. 设二维随机变量(X, Y) 的密度函数为

$$f(x, y) = \begin{cases} e^{-(x+y)}, & 0 > 0, \ y > 0, \\ 0, & \text{其他.} \end{cases}$$

求 $Z = \dfrac{X+Y}{2}$ 的密度函数.

25. 在一个简单电路中，两个电阻 R_1 和 R_2 以串联方式联接. 设 R_1 和 R_2 相互独立同分布，其密度函数均为

$$f(x) = \begin{cases} \dfrac{10-x}{50}, & 0 \leqslant x \leqslant 10, \\ 0, & \text{其他.} \end{cases}$$

求总电阻 $R = R_1 + R_2$ 的密度函数.

第四章　随机变量的数字特征

前面两章分别讨论了一维和二维随机变量. 我们知道，随机变量的分布是对随机变量概率性质的完整刻画，它描述了随机变量的统计规律性. 然而，在实际应用中，有时不需要完全知道随机变量的分布，而只须知道它的某些特征就够了. 另一方面，大部分重要分布可由这些特征完全确定.

描述随机变量平均值的特征和取值分散程度的特征是最重要的两个特征. 例如，对一射手进行技术评定时，除了考察射击命中环数的平均值，还要了解其成绩的稳定性，即命中环数的分散程度. 又如，检验一批棉花的质量时，除了关心棉花纤维的平均长度，还要考虑纤维的长度与平均长度的偏离程度等.

这种由随机变量的分布所确定，能刻画随机变量某些方面特征的数量统称为随机变量的数字特征，它们在理论和应用上都有重要意义. 描述随机变量平均值的数字特征称为数学期望，描述随机变量取值分散程度的数字特征称为方差. 数学期望和方差是刻画随机变量性质的两个重要的数字特征. 另外，对于两个随机变量之间相互关系的数字特征，由它们的协方差和相关系数来描述.

4.1　随机变量的数学期望

4.1.1　离散型随机变量的数学期望

定义 4.1.1　设 X 为离散型随机变量，其分布律为

$$P\{X=x_k\}=p_k,\ k=1,2,\cdots.$$

若级数 $\sum\limits_{i=1}^{\infty}x_ip_i$ 绝对收敛，则称 $\sum\limits_{i=1}^{\infty}x_ip_i$ 为 X 的数学期望(或均值)，记为 $E(X)$. 即 $E(X)=\sum\limits_{i=1}^{\infty}x_ip_i$.

换言之，离散型随机变量 X 的数学期望就是 X 的所有可能取值的加权平均，其中每一个取值的权重等于 X 取这个值的概率.

例 4.1.1　投掷一颗均匀的骰子，观察出现的点数，其结果是一个随机变量，记为 X. 试求 $E(X)$.

解　X 的分布律为

$$P\{X=i\}=\frac{1}{6},\ i=1,2,\cdots,6.$$

所以

$$E(X)=1\times\frac{1}{6}+2\times\frac{1}{6}+3\times\frac{1}{6}+4\times\frac{1}{6}+5\times\frac{1}{6}+6\times\frac{1}{6}=\frac{7}{2}.$$

例 4.1.2　在某地区进行某种疾病普查，为此要检验每个人的血液. 如果当地有 N 个人，考虑用两种检验方法：(1) 检验每个人的血液，这就需要检验 N 次. (2) 先把受检验者分组，假设每个组有 k 个人，把这 k 个人的血液混合在一起检验. 若检验的结果为阴性，这说明 k 个人的血液都是阴性，因而这 k 个人只须检验一次就够了. 若结果呈阳性，为了明确 k 个人中究竟哪个人为阳性，就需要对这 k 个人逐一检验，此时这 k 个人的检验次数为 $k+1$ 次，检验的工作量反而增加. 假设每个人血液检验呈阳性的概率为 p，且试验是相互独立的. 试说明当 p 较小时，按第二种方法可以减少检验次数.

解　显然，若采用第二种方法，则 k 个人需要的检验次数可能是 1 次，也可能是 $k+1$ 次，由于各人的试验是相互独立的，并且每个人检验呈阳性的概率均为 p，呈阴性的概率为 $q=1-p$，因此 k 个人一组的混合血液为阴性的概率为 q^k，呈阳性的概率为 $1-q^k$.

令 X 表示 k 个人为一组时，每人所需的平均检验次数，则 X 的分布律为

X	$\frac{1}{k}$	$\frac{k+1}{k}$
p_k	q^k	$1-q^k$

每个人所需检验次数的均值为

$$E(X)=\frac{1}{k}\times q^k+\left(1+\frac{1}{k}\right)\times(1-q^k)=1-q^k+\frac{1}{k}.$$

按第一种方法每人应检验 1 次，所以当

$$1-q^k+\frac{1}{k}<1,\ \text{即}\ q^k=(1-p)^k>\frac{1}{k}$$

时，即当 p 较小时，用分组方法可减少检验次数.

下面讨论几个常见离散型随机变量的数学期望.

例 4.1.3　设 X 服从 0-1 分布 $B(1,\ p)$，求 $E(X)$.

解　X 的分布律为

X	0	1
p_k	$1-p$	p

所以

$$E(X)=0\times(1-p)+1\times p=p.$$

例 4.1.4　设 X 服从二项分布 $B(n,\ p)$，求 $E(X)$.

解　X 的分布律为

$$p_k=P\{X=k\}=C_n^k p^k q^{n-k},\ k=0,\ 1,\ 2,\ \cdots,\ n.$$

所以

$$E(X)=\sum_{k=0}^{n}kC_n^k p^k q^{n-k}=\sum_{k=1}^{n}kC_n^k p^k q^{n-k}=np\sum_{k=1}^{n}C_{n-1}^{k-1}p^{k-1}q^{(n-1)-(k-1)}$$

$$=np\sum_{j=0}^{n-1}C_{n-1}^{j}p^j q^{(n-1)-j}=np(p+q)^{n-1}=np.$$

例 4.1.5 设 X 服从泊松分布 $P(\lambda)$，求 $E(X)$.

解 X 的分布律为

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

所以

$$E(X)=\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}e^{-\lambda}=\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=\lambda e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!}=\lambda e^{-\lambda}\cdot e^{\lambda}=\lambda.$$

例 4.1.6 设 X 服从几何分布 $Ge(p)$，求 $E(X)$.

解 X 的分布律为

$$P\{X=k\}=q^{k-1}p,\ k=1,2,\cdots,$$

其中 $0<p<1$，$q=1-p$. 利用逐项微分可得 X 的数学期望为

$$E(X)=\sum_{k=1}^{\infty}kq^{k-1}p=p\sum_{k=0}^{\infty}kq^{k-1}=p\frac{\mathrm{d}}{\mathrm{d}q}\left(\sum_{k=1}^{\infty}q^k\right)$$

$$=p\frac{\mathrm{d}}{\mathrm{d}q}\left(\frac{q}{1-q}\right)=p\frac{1}{(1-q)^2}=\frac{1}{p}.$$

若 X 为离散型随机变量，其分布律已知，则 X 的函数 $g(X)$ 也是离散型随机变量，$g(X)$ 的分布律可由 X 的分布律计算得到，于是就可以根据数学期望的定义，求出 $E[g(X)]$.

例 4.1.7 设 X 的分布律为

X	-1	0	1
p_k	0.2	0.5	0.3

求 $E(X^2)$.

解 令 $Y=X^2$，则 Y 的分布律为

Y	0	1
p_k	0.5	0.5

所以

$$E(X^2)=E(Y)=0\times0.5+1\times0.5=0.5.$$

例 4.1.8 某商店对某种家用电器的销售采用先使用后付款的方式. 记使用寿命为 X(以年计)，规定：

$$\begin{aligned}X\leqslant 1,&\quad \text{一台付款 1500 元},\\ 1<X\leqslant 2,&\quad \text{一台付款 2000 元},\\ 2<X\leqslant 3,&\quad \text{一台付款 2500 元},\\ X>3,&\quad \text{一台付款 3000 元}.\end{aligned}$$

设寿命 X 服从指数分布，其密度函数为

$$f(x)=\begin{cases}\dfrac{1}{10}e^{-x/10}, & x>0,\\ 0, & x\leqslant 0.\end{cases}$$

试求该商店销售一台电器收费 Y 的数学期望.

解　先求出寿命 X 落在各个时间区间的概率,

$$P\{X \leqslant 1\} = \int_0^1 \frac{1}{10} e^{-x/10} dx = 1 - e^{-0.1} = 0.095\ 2,$$

$$P\{1 < X \leqslant 2\} = \int_1^2 \frac{1}{10} e^{-x/10} dx = e^{-0.1} - e^{-0.2} = 0.086\ 1,$$

$$P\{2 < X \leqslant 3\} = \int_2^3 \frac{1}{10} e^{-x/10} dx = e^{-0.2} - e^{-0.3} = 0.077\ 9,$$

$$P\{X > 3\} = \int_3^{\infty} \frac{1}{10} e^{-x/10} dx = e^{-0.3} = 0.740\ 8.$$

因此 Y 的分布律为

Y	1 500	2 000	2 500	3 000
p_k	0.095 2	0.086 1	0.077 9	0.740 8

Y 的数学期望为 $E(Y) = 2\ 732.15$.

尽管用上述方法可以求出 X 的函数 $g(X)$ 的数学期望,但有时计算会比较麻烦. 下面两个定理给出求随机变量函数数学期望的常用方法.

定理 4.1.1　设 X 为离散型随机变量,其分布律为

$$P\{X = x_k\} = p_k,\ k = 1,\ 2,\ \cdots.$$

对任一实值函数 $g(x)$,若级数 $\sum_{k=1}^{\infty} g(x_k) p_k$ 绝对收敛,则有

$$E[g(X)] = \sum_{k=1}^{\infty} g(x_k) p_k.$$

定理 4.1.2　设 $(X,\ Y)$ 是离散型随机变量,其分布律为

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

若级数 $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} g(x_i,\ y_j) p_{ij}$ 绝对收敛,则有

$$E[g(X,\ Y)] = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} g(x_i,\ y_j) p_{ij}.$$

这两个定理的意义在于,计算 $Z = g(X)$ 或 $Z = g(X,\ Y)$ 的数学期望时,不需要计算 Z 的分布律,直接利用 X 或 $(X,\ Y)$ 的已知分布律即可直接得到 Z 的数学期望.

例 4.1.9　设 $(X,\ Y)$ 的分布律为

X \ Y	0	1	2	3
1	0	3/8	3/8	0
3	1/8	0	0	1/8

求 $E(Y^2)$,$E(XY)$.

解　由定理 4.1.2 得

$$E(Y^2)=1^2\times\left[0+\frac{3}{8}+\frac{3}{8}+0\right]+3^2\times\left[\frac{1}{8}+0+0+\frac{1}{8}\right]=3,$$

$$E(XY)=(1\times0)\times0+(1\times1)\times\frac{3}{8}+(1\times2)\times\frac{3}{8}+(1\times3)\times0+(3\times0)\times\frac{1}{8}$$

$$+(3\times1)\times0+(3\times2)\times0+(3\times3)\times\frac{1}{8}=\frac{9}{4}.$$

4.1.2 连续型随机变量的数学期望

定义 4.1.2 设 X 为连续型随机变量，其密度函数为 $f(x)$，若积分 $\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x$ 绝对收敛，则称 $\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x$ 为 X 的数学期望(或均值)，记为 $E(X)$. 即

$$E(X)=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x.$$

例 4.1.10 设 X 的密度函数为

$$f(x)=\begin{cases}2x, & 0<x<1,\\ 0, & \text{其他.}\end{cases}$$

求 $E(X)$.

解 $E(X)=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x=\int_0^1x\cdot2x\mathrm{d}x=\frac{2}{3}.$

下面讨论几个常见连续型随机变量的数学期望.

例 4.1.11 设 X 服从均匀分布 $U(a,b)$，求 $E(X)$.

解 X 的密度函数为

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

所以

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

例 4.1.12 设 X 服从指数分布 $E(\lambda)$，求 $E(X)$.

解 X 的密度函数为

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

所以

$$E(X)=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x=\int_0^{+\infty}x\lambda\mathrm{e}^{-\lambda x}\mathrm{d}x$$

$$=-x\mathrm{e}^{-\lambda x}\Big|_0^{+\infty}+\int_0^{+\infty}\mathrm{e}^{-\lambda x}\mathrm{d}x$$

$$=0-\frac{1}{\lambda}\mathrm{e}^{-\lambda x}\Big|_0^{+\infty}=\frac{1}{\lambda}.$$

例 4.1.13　设服从正态分布 $N(\mu, \sigma^2)$，求 $E(X)$.

解　X 的密度函数为

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

所以

$$\begin{aligned}E(X)&=\int_{-\infty}^{+\infty}xf(x)\,dx=\int_{-\infty}^{+\infty}x\frac{1}{\sqrt{2\pi}\cdot\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx\\&=\int_{-\infty}^{+\infty}(x-\mu)\frac{1}{\sqrt{2\pi}\cdot\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx+\mu\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\cdot\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx\\&\xlongequal{t=x-\mu}\int_{-\infty}^{+\infty}t\frac{1}{\sqrt{2\pi}\cdot\sigma}e^{-\frac{t^2}{2\sigma^2}}dt+\mu=\mu.\end{aligned}$$

对于连续型随机变量的函数的数学期望，有与定理 4.1.1 和定理 4.1.2 类似的结论.

定理 4.1.3　设随机变量 X 的密度函数为 $f(x)$，若积分 $\int_{-\infty}^{+\infty}g(x)f(x)\,dx$ 绝对收敛，则有

$$E[g(X)]=\int_{-\infty}^{+\infty}g(x)f(x)\,dx.$$

定理 4.1.4　设 (X, Y) 是连续型随机变量，其密度函数为 $f(x, y)$，若积分 $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}g(x, y)f(x, y)\,dxdy$ 绝对收敛，则有

$$E[g(X, Y)]=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}g(x, y)f(x, y)\,dxdy.$$

定理 4.1.1 至定理 4.1.4 是本章的基本定理，它们提供了求一维和二维随机变量及其函数的数学期望的基本方法. 下面两节中关于随机变量的方差和协方差及相关系数的计算实际上也都归结为求随机变量及其函数的数学期望的计算.

例 4.1.14　设随机变量 X 在 $[0, \pi]$ 上服从均匀分布，求 $E(X)$，$E(\sin X)$，$E(X^2)$ 及 $E[X-E(X)]^2$.

解　由定理 4.1.3，有

$$E(X)=\int_{-\infty}^{+\infty}xf(x)\,dx=\int_0^{\pi}x\cdot\frac{1}{\pi}dx=\frac{\pi}{2},$$

$$E(\sin X)=\int_{-\infty}^{+\infty}\sin xf(x)\,dx=\int_0^{\pi}\frac{1}{\pi}\sin x\,dx=\frac{1}{\pi}(-\cos x)\Big|_0^{\pi}=\frac{2}{\pi},$$

$$E(X^2)=\int_{-\infty}^{+\infty}x^2f(x)\,dx=\int_0^{\pi}x^2\cdot\frac{1}{\pi}dx=\frac{\pi^2}{3},$$

$$E[X-E(X)]^2=E\left(X-\frac{\pi}{2}\right)^2=\int_0^{\pi}\left(x-\frac{\pi}{2}\right)^2\cdot\frac{1}{\pi}dx=\frac{\pi^2}{12}.$$

例 4.1.17　设国际市场上对我国某种出口商品的每年需求量是随机变量 X（单位：吨），它服从区间 $[2\,000, 4\,000]$ 上的均匀分布，每销售出一吨商品，可为国家赚取外汇 3 万元；若销售不出，则每吨商品需要贮存费 1 万元，问应组织多少货源，才能使国家收益最大？

解 设组织货源 t 吨，显然应要求 $2\,000 \leqslant t \leqslant 4\,000$，国家收益 Y(单位：万元) 是 X 的函数，由题意知

$$Y = g(X) = \begin{cases} 3t, & X \geqslant t, \\ 3X - (t - X), & X < t. \end{cases} = \begin{cases} 3t, & X \geqslant t, \\ 4X - t, & X < t. \end{cases}$$

X 的密度函数为

$$f(x) = \begin{cases} 1/2\,000, & 2000 \leqslant x \leqslant 4000, \\ 0, & \text{其他}. \end{cases}$$

于是 Y 的期望为

$$E(Y) = \int_{-\infty}^{+\infty} g(x) f(x)\,\mathrm{d}x = \int_{2\,000}^{4\,000} \frac{1}{2\,000} g(x)\,\mathrm{d}x$$

$$= \frac{1}{2\,000}\left[\int_{2\,000}^{t} (4x - t)\,\mathrm{d}x + \int_{t}^{4\,000} 3t\,\mathrm{d}x\right] = \frac{1}{2\,000}(-2t^2 + 14\,000t - 8 \times 10^6).$$

考虑 t 的取值使 $E(Y)$ 达到最大，易得 $t^* = 3\,500$，因此组织 3 500 吨商品为好.

例 4.1.18 设随机变量 (X, Y) 的密度函数为

$$f(x, y) = \begin{cases} \mathrm{e}^{-x-y}, & x > 0, \ y > 0, \\ 0, & \text{其他}. \end{cases}$$

求 $E(Y)$ 和 $E(XY)$.

解 由定理 4.1.4 可得

$$E(Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y f(x, y)\,\mathrm{d}x\mathrm{d}y = \int_{0}^{+\infty}\int_{0}^{+\infty} y\mathrm{e}^{-xy}\,\mathrm{d}x\mathrm{d}y$$

$$= \int_{0}^{+\infty} \mathrm{e}^{-x}\,\mathrm{d}x \int_{0}^{+\infty} y\mathrm{e}^{-y}\,\mathrm{d}y = 1.$$

$$E(XY) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy f(x, y)\,\mathrm{d}x\mathrm{d}y = \int_{0}^{+\infty}\int_{0}^{+\infty} xy\mathrm{e}^{-xy}\,\mathrm{d}x\mathrm{d}y$$

$$= \int_{0}^{+\infty} x\mathrm{e}^{-x}\,\mathrm{d}x \int_{0}^{+\infty} y\mathrm{e}^{-y}\,\mathrm{d}y = 1.$$

4.1.3 数学期望的性质

数学期望有下列性质：

性质 1 设 C 为常数，则 $E(C) = C$.

性质 2 设 C 为常数，则 $E(CX) = CE(X)$.

性质 3 $E(X+Y) = E(X) + E(Y)$.

该性质可推广为：

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n).$$

性质 4 设 X 与 Y 独立，则 $E(XY) = E(X)E(Y)$.

性质 1 和性质 2 的证明比较容易，由读者自己完成. 对于性质 3 和性质 4，我们只对连续型随机变量情况加以证明，离散型情况的证明与连续型类似.

性质 3 的证明：设二维随机变量 (X, Y) 的密度函数为 $f(x, y)$，则有

$$E(X+Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x+y)f(x,\ y)\,\mathrm{d}x\mathrm{d}y$$

$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xf(x,\ y)\,\mathrm{d}y\mathrm{d}x+\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}yf(x,\ y)\,\mathrm{d}x\mathrm{d}y$$

$$=\int_{-\infty}^{+\infty}xf_X(x)\,\mathrm{d}x+\int_{-\infty}^{+\infty}yf_Y(y)\,\mathrm{d}y=E(X)+E(Y).$$

性质 4 的证明：因为 X 与 Y 独立，故有 $f(x,\ y)=f_X(x)f_Y(y)$，于是

$$E(XY)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xyf(x,\ y)\,\mathrm{d}x\mathrm{d}y=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xyf_X(x)f_Y(y)\,\mathrm{d}x\mathrm{d}y$$

$$=\int_{-\infty}^{+\infty}xf_X(x)\,\mathrm{d}x\cdot\int_{-\infty}^{+\infty}yf_Y(y)\,\mathrm{d}y=E(X)\cdot E(Y).$$

例 4.1.19　设 X 服从参数为 n 和 p 的二项分布 $B(n,\ p)$，求 $E(X)$.

解　由于 X 表示 n 次独立试验中成功的次数，每次成功的概率为 p，我们有

$$X=X_1+X_2+\cdots+X_n,$$

其中 X_1，X_2，…，X_n 相互独立，

$$X_i=\begin{cases}1,\ \text{若第 } i \text{ 次试验成功},\\ 0,\ \text{若第 } i \text{ 次试验不成功}.\end{cases}$$

因此，每个 X_i 服从 0－1 分布，且有 $E(X_i)=p$. 于是

$$E(X_1+X_2+\cdots+X_n)=E(X_1)+E(X_2)+\cdots+E(X_n)=np.$$

例 4.1.19 的解题方法要比例 4.1.4 简单的多. 事实上，通过把一个随机变量分解为若干个相互独立同服从某 0 － 1 分布的随机变量之和的方法可以在很多场合下采用，并能使问题得到简单解决.

例 4.1.20　有 N 个人各自把他们的帽子抛向屋子的中央，将帽子充分混合后，每人随机地从中取出一顶，求刚好拿到自己帽子的人数的数学期望.

解　设 X 为刚好拿到自己帽子的人数，它可表示为

$$X=X_1+X_2+\cdots+X_N,$$

其中 X_1，X_2，…，X_n 相互独立，

$$X_i=\begin{cases}1,\ \text{第 } i \text{ 个人拿到自己的帽子},\\ 0,\ \text{第 } i \text{ 个人未拿到自己的帽子}.\end{cases}$$

由于第 i 个人等可能地从 N 个帽子中取出一顶，即 X_i 有分布律

$$P\{X_i=1\}=\frac{1}{N},\ P\{X_i=0\}=1-\frac{1}{N}.$$

所以有 $E(X_i)=P\{X_i=1\}=\dfrac{1}{N}$，$i=1,\ 2,\ \cdots,\ n$，于是有

$$E(X)=E(X_1)+E(X_2)+\cdots+E(X_N)=N\cdot\frac{1}{N}=1.$$

即平均来说，刚好有一个人取到自己的帽子.

例 4.1.21　设 X 服从负二项分布 $Nb(r,\ p)$，求 $E(X)$.

解　由于负二项分布的随机变量可以表达为 r 个独立同服从几何分布的随机变量之和，即 $X=X_1+X_2+\cdots+X_r$，这里 $X_i\sim Ge(p)$，故由例 4.1.6 的结果和数学期望性质可得

$$E(X)=E(X_1)+E(X_2)+\cdots+E(X_r)=\frac{r}{p}.$$

例 4.1.22 设 X 和 Y 是两个相互独立的连续型随机变量，它们的密度函数分别为

$$f_X(x)=\begin{cases}2x, & 0<x<1,\\ 0, & \text{其他.}\end{cases}\qquad f_Y(y)=\begin{cases}\dfrac{y^2}{9}, & 0<y<3,\\ 0, & \text{其他.}\end{cases}$$

试求 $E(XY)$.

解 因为 X 与 Y 独立，所以有

$$E(XY)=E(X)\cdot E(Y)=\int_{-\infty}^{+\infty}xf_X(x)\,\mathrm{d}x\cdot\int_{-\infty}^{+\infty}yf_Y(y)\,\mathrm{d}y$$

$$=\int_0^1 x\cdot 2x\,\mathrm{d}x\cdot\int_0^3 y\cdot\frac{y^2}{9}\mathrm{d}y=\frac{3}{2}.$$

4.2 随机变量的方差

随机变量的数学期望是刻画随机变量平均取值的数字特征. 在本章开始，我们已经指出，方差是随机变量的又一个重要数字特征，它刻画了随机变量取值的分散程度，也就是随机变量取值与平均值的偏离程度.

定义 4.2.1 设 X 是一个随机变量，若 $E[X-E(X)]^2$ 存在，则称其为 X 的方差，记为 $D(X)$ 或 $Var(X)$. 即

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

称 $D(X)$ 的算术平方根 $\sqrt{D(X)}$ 为 X 的标准差.

若 X 为离散型随机变量，其分布律为

$$P\{X=x_k\}=p_k,\quad k=1,2,\cdots.$$

则

$$D(X)=\sum_{k=1}^{\infty}[x_k-E(X)]^2p_k.$$

若 X 为连续型随机变量，其密度函数为 $f(x)$，则

$$D(X)=\int_{-\infty}^{+\infty}[x-E(X)]^2f(x)\,\mathrm{d}x.$$

因为

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

所以，常采用以下公式计算方差

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

例 4.2.1 抛掷一颗均匀的骰子，观察出现的点数，其结果是一个随机变量，记为 X. 试求 $D(X)$.

解　由例 4.1.1 知，X 的数学期望为 $E(X)=\dfrac{7}{2}$，由于

$$E(X^2)=1^2\times\frac{1}{6}+2^2\times\frac{1}{6}+3^2\times\frac{1}{6}+4^2\times\frac{1}{6}+5^2\times\frac{1}{6}+6^2\times\frac{1}{6}=\frac{91}{6},$$

因此，

$$D(X)=E(X^2)-[E(X)]^2=\frac{91}{6}-\left(\frac{7}{2}\right)^2=\frac{35}{12}.$$

下面介绍几个常见随机变量的方差.

例 4.2.2　设 X 服从 0－1 分布 $B(1,\ p)$，求 $D(X)$.

解　X 的分布律为

X	0	1
p_k	$1-p$	p

由例 4.1.3 知，X 的数学期望为 $E(X)=p$，由于

$$E(X^2)=1^2\cdot p+0^2\cdot(1-p)=p,$$

因此

$$D(X)=E(X^2)-[E(X)]^2=p-p^2=p(1-p).$$

例 4.2.3　设 X 服从泊松分布 $P(\lambda)$，求 $D(X)$.

解　X 的分布律为

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

由例 4.1.5 知，$E(X)=\lambda$，又由于

$$\begin{aligned}E(X^2)&=\sum_{k=0}^{\infty}k^2\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}=\sum_{k=1}^{\infty}k^2\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}\\&=\lambda\mathrm{e}^{-\lambda}\sum_{k=1}^{\infty}k\frac{\lambda^{k-1}}{(k-1)!}\\&=\lambda\mathrm{e}^{-\lambda}\sum_{k=1}^{\infty}(k-1)\frac{\lambda^{k-1}}{(k-1)!}+\lambda\mathrm{e}^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}\\&=\lambda\mathrm{e}^{-\lambda}\sum_{k=2}^{\infty}(k-1)\frac{\lambda^{k-1}}{(k-1)!}+\lambda\mathrm{e}^{-\lambda}\cdot\mathrm{e}^{\lambda}\\&=\lambda^2\mathrm{e}^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}+\lambda=\lambda^2+\lambda.\end{aligned}$$

因此

$$D(X)=E(X^2)-[E(X)]^2=\lambda^2+\lambda-\lambda^2=\lambda.$$

例 4.2.4　设 X 服从几何分布 $Ge(p)$，求 $D(X)$.

解　X 的分布律为

$$P\{X=k\}=q^{k-1}p,\ k=1,\ 2,\ \cdots,$$

其中 $0<p<1$，$q=1-p$. 由例 4.1.6 知，$E(X)=1/p$，又

$$E(X^2)=\sum_{k=1}^{\infty}k^2q^{k-1}p=\left[p\sum_{k=1}^{\infty}k(k-1)q^{k-1}+\sum_{k=1}^{\infty}kq^{k-1}p\right]$$
$$=pq\sum_{k=1}^{\infty}k(k-1)q^{k-2}+\frac{1}{p}=pq\frac{d^2}{dq^2}\left(\sum_{k=1}^{\infty}q^k\right)+\frac{1}{p}$$
$$=pq\frac{d^2}{dq^2}\left(\frac{q}{1-q}\right)+\frac{1}{p}=pq\frac{2}{(1-q)^3}+\frac{1}{p}=\frac{2q}{p^2}+\frac{1}{p}.$$

由此得 X 的方差为

$$D(X)=E(X^2)-[E(X)]^2=\frac{2q}{p^2}+\frac{1}{p}-\frac{1}{p^2}=\frac{1-p}{p^2}.$$

例 4.2.5 设 $X\sim U(a,b)$，求 $D(X)$.

解 X 的密度函数为

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

由例 4.1.11 知，$E(X)=\dfrac{b+a}{2}$，又由于

$$E(X^2)=\int_{-\infty}^{+\infty}x^2f(x)\,\mathrm{d}x=\int_a^b x^2\frac{1}{b-a}\mathrm{d}x=\frac{b^2+ab+a^2}{3}.$$

可得 X 的方差

$$D(X)=E(X^2)-[E(X)]^2=\frac{b^2+ab+a^2}{3}-\left(\frac{a+b}{2}\right)^2=\frac{(b-a)^2}{12}.$$

例 4.2.6 设 $X\sim E(\lambda)$，求 $D(X)$.

解 X 的密度函数为

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

由例 4.1.12 知，$E(X)=\dfrac{1}{\lambda}$，又由

$$E(X^2)=\int_{-\infty}^{+\infty}x^2f(x)\,\mathrm{d}x=\int_0^{+\infty}x^2\lambda\mathrm{e}^{-\lambda x}\mathrm{d}x$$
$$=-x^2\mathrm{e}^{-\lambda x}\Big|_0^{+\infty}+\int_0^{+\infty}2x\mathrm{e}^{-\lambda x}\mathrm{d}x=\frac{2}{\lambda^2}.$$

可得

$$D(X)=E(X^2)-[E(X)]^2=\frac{2}{\lambda^2}-\left(\frac{1}{\lambda}\right)^2=\frac{1}{\lambda^2}.$$

例 4.2.7 设 $X\sim N(\mu,\sigma^2)$，求 $D(X)$.

解 X 的密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\cdot\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad -\infty<x<+\infty.$$

由例 4.1.13 知，$E(X)=\mu$，因此

$$D(X)=\int_{-\infty}^{+\infty}[x-E(X)]^2f(x)\mathrm{d}x=\int_{-\infty}^{+\infty}(x-\mu)^2\frac{1}{\sqrt{2\pi}\cdot\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x.$$

令$\frac{x-\mu}{\sigma}=t$，则

$$\begin{aligned}D(X)&=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}t^2\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\frac{\sigma^2}{\sqrt{2\pi}}(-t\mathrm{e}^{-\frac{t^2}{2}})\Big|_{-\infty}^{+\infty}+\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t\\&=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\sigma^2.\end{aligned}$$

由例4.1.13和例4.2.7可知，正态分布$N(\mu,\sigma^2)$的参数μ和σ^2分别是该分布的数学期望和方差. 特别地，若$X\sim N(0,1)$，则$E(X)=0$，$D(X)=1$.

下面是方差的几个重要性质.

性质1　设C为常数，则$D(C)=0$.

证明　$D(C)=E[C-E(C)]^2=0$.

性质2　设C为常数，则有

$$D(CX)=C^2D(X),\quad D(X+C)=D(X).$$

证明　$D(CX)=E[CX-E(CX)]^2=C^2E[X-E(X)]^2=C^2D(X)$.

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

性质3　设X，Y是两个随机变量，则有

$$D(X+Y)=D(X)+D(Y)+2E[(X-E(X))(Y-E(Y))].$$

若X和Y相互独立，则有

$$D(X+Y)=D(X)+D(Y).$$

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

若X和Y相互独立，则$E(XY)=E(X)E(Y)$，于是

$$\begin{aligned}E[(X-E(X))(Y-E(Y))]&=E[XY-XE(Y)-YE(X)+E(X)E(Y)]\\&=E(XY)-E(X)E(Y)-E(Y)E(X)+E(X)E(Y)\\&=0.\end{aligned}$$

即$D(X+Y)=D(X)+D(Y)$.

上述性质可以推广到n个随机变量的情况. 特别地，若随机变量X_1，X_2，…，X_n相互独立，则有

$$D(X_1+X_2+\cdots+X_n)=D(X_1)+D(X_2)+\cdots+D(X_n),$$

且有

$$D(c_1X_1+c_2X_2+\cdots+c_2X_n)=c_1^2D(X_1)+c_2^2D(X_2)+\cdots+c_n^2D(X_n),$$

其中c_1，c_2，…，c_n都是常数. 例如，若X与Y独立，则有

$$D(X-Y)=D(X)+D(-Y)=D(X)+(-1)^2D(Y)=D(X)+D(Y),$$

即相互独立随机变量之差的方差等于它们的方差之和.

例 4.2.8 设 X 服从参数为 n 和 p 的二项分布 $B(n, p)$，求 $D(X)$.

解 由于 X 表示 n 次独立试验中成功的次数，每次成功的概率为 p，我们有

$$X = X_1 + X_2 + \cdots + X_n,$$

其中 X_1，X_2，…，X_n 相互独立，

$$X_i = \begin{cases} 1, & \text{若第 } i \text{ 次试验成功}, \\ 0, & \text{若第 } i \text{ 次试验不成功}. \end{cases}$$

由例 4.2.2 知，$D(X_i) = pq$，这里 $q = 1 - p$. 由 X_1，X_2，…，X_n 的相互独立性可得

$$D(X_1 + X_2 + \cdots + X_n) = D(X_1) + D(X_2) + \cdots + D(X_n) = npq.$$

例 4.2.9 设 X 服从负二项分布 $Nb(r, p)$，求 $D(X)$.

解 由于负二项分布的随机变量可以表达为 r 个独立同服从几何分布的随机变量之和，即 $X = X_1 + X_2 + \cdots + X_r$，这里 $X_i \sim Ge(p)$，由 X_1，X_2，…，X_n 相互独立及例 4.2.4 结果得

$$D(X) = D(X_1) + D(X_2) + \cdots + D(X_r) = \frac{r(1-p)}{p^2}.$$

例 4.2.10 设 $X \sim N(0, 1)$，$Y \sim N(-1, 4)$，且 X 和 Y 相互独立，求 $D(2X - 3Y)$.

解 由于 X 与 Y 独立，所以

$$D(2X - 3Y) = 2^2 D(X) + 3^2 D(Y) = 4 + 36 = 40.$$

事实上，若 $X_i \sim N(\mu_i, \sigma_i^2)$，$i = 1, 2, \cdots, n$，且它们相互独立，则它们的线性组合 $\sum\limits_{i=1}^{n} C_i X_i$ 仍然服从正态分布. 由数学期望和方差的性质知

$$E\left(\sum_{i=1}^{n} C_i X_i\right) = \sum_{i=1}^{n} C_i \mu_i,$$

$$D\left(\sum_{i=1}^{n} C_i X_i\right) = \sum_{i=1}^{n} C_i^2 D(X_i) = \sum_{i=1}^{n} C_i^2 \sigma_i^2.$$

于是有

$$\sum_{i=1}^{n} C_i X_i \sim N\left(\sum_{i=1}^{n} C_i \mu_i, \sum_{i=1}^{n} C_i^2 \sigma_i^2\right).$$

例 4.2.11 设 $X \sim N(\mu, \sigma^2)$，求 $\dfrac{X-\mu}{\sigma}$ 的分布.

解 由于 $X \sim N(\mu, \sigma^2)$，所以 $\dfrac{X-\mu}{\sigma}$ 也服从正态分布，其数学期望和方差分别为

$$E\left(\frac{X-\mu}{\sigma}\right) = \frac{E(X) - \mu}{\sigma} = 0,$$

$$D\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2} D(X - \mu) = \frac{1}{\sigma^2} D(X) = 1.$$

因此 $\dfrac{X-\mu}{\sigma} \sim N(0, 1)$，也就是说，对于任意一个正态随机变量，总可以通过变换 $Z = \dfrac{X-\mu}{\sigma}$，使其化为标准正态分布.

例 2.2.12　设随机变量(X, Y)在以点$(0, 1)$，$(1, 0)$，$(1, 1)$为顶点的三角形区域G上服从均匀分布，试求随机变量$Z = X+Y$的数学期望和方差.

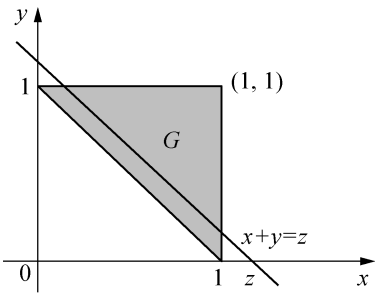


图 4.1　三角形区域的均匀分布

解　三角形区域G如图4.1所示，G的面积为1/2，所以(X, Y)的密度函数为

$$f(x, y) = \begin{cases} 2, & (x, y) \in G, \\ 0, & (x, y) \in G. \end{cases}$$

于是有

$$E(X+Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x+y)f(x, y)\,\mathrm{d}x\mathrm{d}y$$

$$= \int_0^1 \mathrm{d}x\int_{1-x}^1 2(x+y)\,\mathrm{d}y = \int_0^1 (x^2+2x)\,\mathrm{d}x = \left(\frac{x^3}{3}+x^2\right)\Big|_0^1 = \frac{4}{3},$$

$$E[(X+Y)^2] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x+y)^2 f(x, y)\,\mathrm{d}x\mathrm{d}y = \int_0^1 \mathrm{d}x\int_{1-x}^1 2(x+y)^2\mathrm{d}y$$

$$= \frac{2}{3}\int_0^1 (x^3+3x^2+3x)\,\mathrm{d}x = \frac{11}{6},$$

$$D(X+Y) = E[(X+Y)^2] - [E(X+Y)]^2 = \frac{1}{18}.$$

以下我们将常用分布及其期望和方差以表格形式放在一起.

表 4.1　常用分布及其数学期望和方差

分布	分布律p_k或密度函数$f(x)$	期望	方差
0-1 分布	$p_k = p^k(1-p)^{1-k}$, $k = 0, 1$	p	$p(1-p)$
二项分布 $B(n, p)$	$p_k = C_n^k p^k (1-p)^{n-k}$, $k = 1, 2, \cdots, n$	np	$np(1-p)$
泊松分布 $P(\lambda)$	$p_k = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$, $k = 0, 1, 2, \cdots$	λ	λ
超几何分布 $h(n, N, M)$	$p_k = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$, $k = 0, 1, \cdots, r$, $r = \min\{M, n\}$	$\frac{nM}{N}$	$\frac{nM(N-M)(N-n)}{N^2(N-1)}$
几何分布 $Ge(p)$	$p_k = (1-p)^{k-1}p$, $k = 1, 2, \cdots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
负二项分布 $Nb(r, p)$	$p_k = C_{k-1}^{r-1}(1-p)^{k-r}p^r$, $k = r, r+1, \cdots$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$
正态分布 $N(\mu, \sigma^2)$	$f(x) = \frac{1}{\sqrt{2\pi}\sigma}exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$, $-\infty<x<\infty$	μ	σ^2
均匀分布 $U(a, b)$	$\frac{a+b}{2}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
指数分布 $E(p)$	$f(x) = \lambda\mathrm{e}^{-\lambda x}$, $x>0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

注：表中仅列出密度函数的非零区域.

*4.3 协方差和相关系数

设(X, Y)为二维随机变量，本节讨论X和Y之间相互关系的数字特征.

如果X和Y相互独立，则容易证明$E\{[X-(EX)][Y-E(Y)]\}=0$. 若$E\{[X-(EX)][Y-E(Y)]\}\neq 0$，则说明它们之间存在一定的相关性，我们用$X$和$Y$的协方差和相关系数来描述它们之间相关性.

定义 4.3.1 若$E[X-(EX)][Y-E(Y)]$存在，称其为随机变量X和Y的协方差，记为$Cov(X, Y)$，即

$$Cov(X, Y)=E\{[X-(EX)][Y-E(Y)]\}.$$

当$D(X)>0$，$D(Y)>0$时，称

$$\rho_{XY}=\frac{Cov(X, Y)}{\sqrt{D(X)}\cdot\sqrt{D(Y)}}$$

为X和Y的相关系数.

将协方差的定义式展开，得

$$\begin{aligned}Cov(X, Y)&=E\{[X-(EX)][Y-E(Y)]\}\\&=E\{XY-XE(Y)-YE(X)+E(X)E(Y)\}\\&=E(XY)-E(X)E(Y)-E(Y)E(X)+E(X)E(Y)\\&=E(XY)-E(X)E(Y),\end{aligned}$$

即$Cov(X, Y)=E(XY)-E(X)E(Y)$，这是计算协方差的常用公式.

协方差有下列性质：

性质 1 $Cov(X, Y)=Cov(Y, X)$；

性质 2 $Cov(X, X)=D(X)$；

性质 3 $D(X\pm Y)=D(X)+D(Y)\pm 2Cov(X, Y)$；

该性质可以推广到n个随机变量，即有

$$D\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}D(X_i)+2\sum_{i<j}\sum Cov(X_i, X_j);$$

性质 4 $Cov(aX, bY)=abCov(X, Y)$；

性质 5 $Cov(X_1+X_2, Y)=Cov(X_1, Y)+Cov(X_2, Y)$.

若随机变量X和Y相互独立，则$E(XY)=E(X)E(Y)$，从而$Cov(X, Y)=0$. 该命题的逆命题不成立，即两个随机变量的协方差为零，并不能说明它们相互独立.

例 4.3.1 设(X, Y)分布律为

X \ Y	-1	0	1	$P\{X=i\}$
0	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{2}{3}$
1	0	$\frac{1}{3}$	0	$\frac{1}{3}$
$P\{Y=j\}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	

求 $Cov(X, Y)$.

解　由于 $E(Y)=(-1)\times\frac{1}{3}+0\times\frac{1}{3}+1\times\frac{1}{3}=0$，且 $E(XY)=0$，所以

$$Cov(X, Y)=E(XY)-E(X)E(Y)=0.$$

然而，易见 X 与 Y 不独立.

例 4.3.2　设 (X, Y) 在单位圆 $X^2+Y^2\leqslant 1$ 上具有均匀分布，求 $Cov(X, Y)$.

解　(X, Y) 的联合密度函数为

$$f(x, y)=\begin{cases}\dfrac{1}{\pi}, & x^2+y^2\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

X 和 Y 的边缘密度函数为

$$f_X(x)=\begin{cases}\dfrac{2}{\pi}\sqrt{1-x^2}, & -1\leqslant x\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

$$f_Y(y)=\begin{cases}\dfrac{2}{\pi}\sqrt{1-y^2}, & -1\leqslant y\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

因此有

$$E(X)=\int_{-\infty}^{+\infty}xf_X(x)\,\mathrm{d}x=\int_{-1}^{+1}x\,\frac{2}{\pi}\sqrt{1-x^2}\,\mathrm{d}x=0.$$

同理可得 $E(Y)=0$. 而

$$\begin{aligned}E(XY)&=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xyf(x, y)\,\mathrm{d}x\mathrm{d}y\\&=\frac{1}{\pi}\int_0^{2\pi}\mathrm{d}\theta\int_0^1 r^2\sin\theta\cos\theta r\mathrm{d}r\\&=0.\end{aligned}$$

于是有

$$Cov(X, Y)=E(XY)-E(X)E(Y)=0.$$

然而，易见随机变量 X 和 Y 不独立，因为 $f(x, y)\neq f_X(x)f_Y(x)$.

例 4.3.3　设 (X, Y) 的密度函数为

$$f(x, y)=\begin{cases}8xy, & 0\leqslant x\leqslant y\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

求 $Cov(X, Y)$，$D(X+Y)$ 和 ρ_{XY}.

解　X 和 Y 的边缘密度函数为

$$f_X(x)=\begin{cases}4x(1-x^2), & 0\leqslant x\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

$$f_Y(y)=\begin{cases}4y^3, & 0\leqslant y\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

于是

$$E(X)=\int_{-\infty}^{+\infty}xf_X(x)\,\mathrm{d}x=\int_0^1 x\cdot 4x(1-x^2)\,\mathrm{d}x=\frac{8}{15},$$

$$E(Y)=\int_{-\infty}^{+\infty}yf_Y(y)\,\mathrm{d}y=\int_0^1 y\cdot 4y^3\,\mathrm{d}y=\frac{4}{5},$$

$$E(XY)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xyf(x,\ y)\,\mathrm{d}x\mathrm{d}y=\int_0^1\mathrm{d}x\int_x^1 xy\cdot 8xy\cdot\mathrm{d}y=\frac{4}{9}.$$

从而

$$Cov(X,\ Y)=E(XY)-E(X)E(Y)=\frac{4}{225}.$$

又

$$E(X^2)=\int_{-\infty}^{+\infty}x^2f_X(x)\,\mathrm{d}x=\int_0^1 x^2\cdot 4x(1-x^2)\,\mathrm{d}x=\frac{1}{3},$$

$$E(Y^2)=\int_{-\infty}^{+\infty}y^2f_Y(y)\,\mathrm{d}y=\int_0^1 y^2\cdot 4y^3\,\mathrm{d}y=\frac{2}{3},$$

所以有

$$D(X)=E(X^2)-[E(X)]^2=\frac{11}{225},$$

$$D(Y)=E(Y^2)-[E(Y)]^2=\frac{6}{225}=\frac{2}{75},$$

于是可得

$$D(X+Y)=D(X)+D(Y)+2Cov(X,\ Y)=\frac{1}{9},$$

$$\rho_{XY}=\frac{Cov(X,\ Y)}{\sqrt{D(X)}\cdot\sqrt{D(Y)}}=\frac{4}{\sqrt{33}}.$$

下面定理给出相关系数ρ_{XY}的两条重要性质.

定理 4.3.1 随机变量 X 和 Y 的相关系数ρ_{XY}满足:

(1) $|\rho_{XY}|\leqslant 1$.

(2) $|\rho_{XY}|=1$ 的充要条件是,存在常数 a, b 使

$$P\{Y=aX+b\}=1.$$

证明略.

由定理 4.3.1 可知,当$|\rho_{XY}|=1$时,X 和 Y 之间以概率 1 存在线性关系. 特别地,当$\rho_{XY}=1$时,称为正线性相关;当$\rho_{XY}=-1$时,称为负线性相关. 且当$|\rho_{XY}|$较小时,X 和 Y 之间的线性相关程度较弱;当$|\rho_{XY}|$较大时,X 和 Y 之间的线性相关程度较强. 当$\rho_{XY}=0$时,X 与 Y 不相关.

习题四

1. 设随机变量 X 的分布律为

X	0	1	2
p_k	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

求 $E(X)$，$E(X^2+2)$ 及 $D(X)$.

2. 把 4 个球随机地投入 4 个盒子中，设 X 表示空盒子的个数，求 $E(X)$ 和 $D(X)$.

3. 设随机变量 X 的密度函数为

$$f(x)=\begin{cases}2(1-x), & 0<x<1,\\ 0, & \text{其他.}\end{cases}$$

求 $E(X)$ 和 $D(X)$.

4. 设随机变量 X 的密度函数为

$$f(x)=\begin{cases}1+x, & -1\leqslant x\leqslant 0,\\ 1-x, & 0<x\leqslant 1,\\ 0, & \text{其他.}\end{cases}$$

求 $E(X)$ 和 $D(X)$.

5. 设 X 表示 10 次独立重复射击命中目标的次数，每次命中目标的概率为 0.4，求 $E(X^2)$.

6. 已知随机变量 X 服从参数为 2 的泊松分布，求 $E(3X-2)$.

7. 设一部机器在一天内发生故障的概率为 0.2，一周 5 个工作日. 若无故障，可获利润 10 万元；发生一次故障仍可获利润 5 万元；若发生两次故障，获利润 0 元；若发生 3 次或 3 次以上故障就要亏损 2 万元. 求一周利润的数学期望.

8. 设某工厂生产的圆盘，其直径在区间$(a,\ b)$ 上服从均匀分布，求该圆盘面积的数学期望.

9. 设随机变量 X 的密度函数为

$$f(x)=\begin{cases}\mathrm{e}^{-x}, & x>0,\\ 0, & \text{其他.}\end{cases}$$

求（1）$Y=2X$ 的数学期望；（2）$Y=\mathrm{e}^{-2X}$ 的数学期望.

10. 设二维随机变量$(X,\ Y)$ 的联合密度为

$$f(x,\ y)=\begin{cases}\frac{1}{8}(x+y), & 0<x<2,\ 0<y<2,\\ 0, & \text{其他.}\end{cases}$$

求 $E(X)$，$E(Y)$，$E(XY)$ 和 $E(X^2+Y^2)$.

11. 设随机变量 X，Y 分别服从参数为 2 和 4 的指数分布，

（1）求 $E(X+Y)$，$E(2X-3Y^2)$；

(2) 设 X，Y 相互独立，求 $E(XY)$，$D(X+Y)$.

12. 设 $X\sim N(1,2)$，$Y\sim N(0,1)$，且 X 和 Y 相互独立，求随机变量 $Z=2X-Y+3$ 的密度函数.

13. 设有 10 个猎人正等着野鸭飞过来，当一群野鸭飞过头顶时，他们同时开了枪，但他们每个人都是随机地彼此独立地选择自己的目标. 如果每个猎人独立地射中其目标的概率均为 p，试求当 10 只野鸭飞来时，没有被击中而飞走的野鸭数的数学期望.

14. 一个骰子掷 10 次，求得到的总点数的数学期望.

15. 设随机变量 X 和 Y 的联合分布律为

X \ Y	-1	0	1
0	0.07	0.18	0.15
1	0.08	0.32	0.20

求 $E(X)$，$E(Y)$，$Cov(X,Y)$.

16. 设随机变量 (X,Y) 的密度函数为

$$f(x,y)=\begin{cases}1, & |y|<x,\ 0<x<1,\\ 0, & \text{其他.}\end{cases}$$

求 $E(X)$，$E(Y)$，$Cov(X,Y)$.

17. 设随机变量 X 服从拉普拉斯分布，其密度函数为

$$f(x)=\frac{1}{2}e^{-|x|},\quad -\infty<x<+\infty.$$

(1) 求 $E(X)$ 和 $D(X)$；

(2) 求 X 与 $|X|$ 的协方差，并判断 X 与 $|X|$ 的相关性；

(3) 问 X 与 $|X|$ 是否相互独立？

18. 已知随机变量 X 服从二项分布，且 $E(X)=2.4$，$D(X)=1.44$，求此二项分布的参数 n，p 的值.

19. 某流水生产线上每个产品不合格的概率为 $p(0<p<1)$，各产品合格与否相互独立，当出现一个不合格品时即停机检修. 设开机后第一次停机时已生产的产品个数为 X，求 $E(X)$ 和 $D(X)$.

20. 设随机变量 X 在区间 $(-1,1)$ 上服从均匀分布，随机变量

$$Y=\begin{cases}-1, & X<0,\\ 0, & X=0,\\ 1, & X>0.\end{cases}$$

求 $E(Y)$ 和 $D(Y)$.

21. 设随机变量 X 的密度函数为

$$f(x)=\begin{cases}\frac{1}{2}\cos\frac{x}{2}, & 0<x<\pi,\\ 0, & \text{其他.}\end{cases}$$

对 X 独立地观察 4 次，用 Y 表示观察值大于 $\frac{\pi}{3}$ 的次数，求 Y^2 的数学期望.

22. 设随机变量 Y 服从参数为 1 的指数分布，随机变量

$$X_k=\begin{cases}1, & Y>k,\\ 0, & Y\leqslant k,\end{cases}\qquad (k=1,\ 2)$$

求(1) $(X_1,\ X_2)$ 的分布律；(2)$E(X_1+X_2)$.

23. 设 X 和 Y 是两个相互独立且均服从正态分布 $N(0,\ 0.5)$ 的随机变量，求 $E[\,|X-Y|\,]$.

24. 已知 $X\sim N(1,\ 9)$，$Y\sim N(0,\ 16)$，X 和 Y 的相关系数为 $\rho_{XY}=-\dfrac{1}{2}$. 设 $Z=\dfrac{X}{3}+\dfrac{Y}{2}$.

(1) 求 $E(Z)$ 和 $D(Z)$；

(2) 求 X 和 Z 的相关系数.

25. 设 A，B 为随机事件，且 $P(A)=\dfrac{1}{4}$，$P(B\mid A)=\dfrac{1}{3}$，$P(A\mid B)=\dfrac{1}{2}$，令

$$X=\begin{cases}1, & A\text{ 生},\\ 0, & A\text{ 不生}.\end{cases}\qquad Y=\begin{cases}1, & B\text{ 生},\\ 0, & B\text{ 不生}.\end{cases}$$

求(1) 二维随机变量 $(X,\ Y)$ 的分布律；(2) X 和 Y 的相关系数.

26. 将一枚硬币重复掷 n 次，以 X 和 Y 分别表示正面向上和反面向上的次数，求 X 和 Y 的相关系数.

第五章　极限定理

极限定理是概率论的基本定理，在理论研究和实际应用中起着十分重要的作用. 本章介绍关于随机变量序列的两类极限定理，即大数定律和中心极限定理.

概率论早期发展的目的在于揭示由于大量随机因素产生影响而呈现的规律性. 伯努利首先认识到研究无穷随机试验序列的重要性，并建立了概率论的第一个极限定理 —— 大数定律，清楚地刻画了事件的概率与它发生的频率之间的关系，即频率趋于事件的概率. 这里是指试验的次数无限增大时，频率在某种收敛意义下逼近某一个定数. 这就是最早的一个大数定律. 一般的大数定律讨论 n 个随机变量的平均值的稳定性，对上述情况从理论的高度给予概括和论证.

棣莫弗和拉普拉斯提出将观察的误差看作大量独立微小误差的累加，证明了观察误差的分布一定渐近正态的结论 —— 中心极限定理. 随后，出现了许多各种意义下的极限定理. 这类定理证明了在某些一般性条件下，大量的随机变量之和的分布逼近于正态分布. 因此，中心极限定理不仅提供了计算独立随机变量之和的近似概率的方法，而且有助于解释为什么很多观察数据的经验频率呈现正态曲线这一值得注意的事实. 利用这些定理，许多复杂随机变量的分布可以用正态分布近似，而正态分布有着许多完美的性质. 这些结果和研究方法对概率论与数理统计及其应用的许多领域有着重大影响.

*5.1　大数定律

5.1.1　切比雪夫不等式

首先由下面定理给出一个重要不等式，称为切比雪夫(Chebyshev) 不等式.

定理 5.1.1　设随机变量 X 具有数学期望 $E(X)$ 和方差 $D(X)$，则对于任意正数 ε，不等式

$$P\{|X-E(X)|\geqslant\varepsilon\}\leqslant\frac{D(X)}{\varepsilon^2} \tag{5.1.1}$$

成立.

证明　我们只就连续型随机变量的情况来证明. 设 X 的密度函数为 $f(x)$，则有

$$\begin{aligned}P\{|X-E(X)|\geqslant\varepsilon\}&=\int_{|X-E(X)|\geqslant\varepsilon}f(x)\,\mathrm{d}x\leqslant\int_{|X-E(X)|\geqslant\varepsilon}\frac{(x-E(X))^2}{\varepsilon^2}f(x)\,\mathrm{d}x\\&\leqslant\int_{-\infty}^{+\infty}\frac{(x-E(X))^2}{\varepsilon^2}f(x)\,\mathrm{d}x=\frac{1}{\varepsilon^2}\int_{-\infty}^{+\infty}(x-E(X))^2f(x)\,\mathrm{d}x\\&=\frac{1}{\varepsilon^2}E[X-E(X)]^2=\frac{D(X)}{\varepsilon^2}.\end{aligned}$$

定理证毕.

不等式(5.1.1)等价于

$$P\{|X-E(X)|<\varepsilon\} \geqslant 1-\frac{D(X)}{\varepsilon^2} \tag{5.1.2}$$

两个不等式都称为切比雪夫不等式.

切比雪夫不等式表明，在随机变量 X 的分布未知的情况下，可利用 X 的期望 $E(X)=\mu$ 及方差 $D(X)=\sigma^2$ 对 X 的概率分布进行估计. 从切比雪夫不等式(5.1.1)可以看出，当方差越小时，事件 $\{|X-E(X)|\geqslant\varepsilon\}$ 的概率越小，从这里也可以看出方差是描述随机变量与其期望值离散程度的一个量，这与我们以前的理解完全一致.

例 5.1.1　设在每次试验中事件 A 发生的概率为 0.5，利用切比雪夫不等式估计：在 1 000 次独立试验中，事件 A 发生的次数在 450 ~ 550 次之间的概率.

解　设 X 为事件 A 在 1 000 次试验中发生的次数，则 $X\sim B(1\,000,\ 0.5)$，所以

$$E(X)=1\,000\times 0.5=500,\quad Var(X)=1\,000\times 0.5\times(1-0.5)=250,$$

由切比雪夫不等式，得所求概率为

$$P(450\leqslant X\leqslant 550)=P(|X-EX|\leqslant 50)\geqslant 1-\frac{250}{50^2}=0.9.$$

例 5.1.2　设随机变量 X 和 Y 的数学期望都是 2，方差分别为 2 和 4，且 X 和 Y 相互独立，试根据切比雪夫不等式估计 $P\{|X-Y|<6\}$.

解　$X-Y$ 的数学期望为 $E(X-Y)=E(X)-E(Y)=2-2=0$，由于 X 与 Y 独立，所以其方差为

$$D(X-Y)=D(X)+D(Y)=2+4=6.$$

由切比雪夫不等式，得

$$\begin{aligned}P\{|X-Y|<6\}&=P\{|(X-Y)-E(X-Y)|<6\}\\&\geqslant 1-\frac{D(X-Y)}{6^2}=1-\frac{1}{6}=\frac{5}{6}.\end{aligned}$$

5.1.2　大数定律

首先给出依概率收敛的定义.

定义 5.1.1　设 X_1，X_2，…，X_n，… 是随机变量序列，μ 是一个常数，若对于任意给定的正数 ε，有

$$\lim_{n\to\infty}P\{|X_n-\mu|<\varepsilon\}=1,$$

或等价地有

$$\lim_{n\to\infty}P\{|X_n-\mu|\geqslant\varepsilon\}=0,$$

则称随机变量序列 $\{X_n\}$ 依概率收敛于 μ，记为 $X_n\xrightarrow{P}\mu$.

若记 $a_n=P\{|X_n-\mu|<\varepsilon\}$，则 $X_n\xrightarrow{P}\mu\Leftrightarrow\lim\limits_{n\to\infty}a_n=1$，由此可见，依概率收敛仍然是用数列收敛的概念来定义的.

依概率收敛的序列具有以下性质.

设 $X_n \xrightarrow{P} \mu$，$Y_n \xrightarrow{P} v$，又设函数在 $g(x, y)$ 点 (u, v) 连续，则

$$g(X_n, Y_n) \xrightarrow{P} g(\mu, v).$$

定理 5.1.2(切比雪夫大数定律) 设随机变量序列 X_1，X_2，… 相互独立，若存在常数 $c>0$ 使得 $DX_i \leqslant c$，$i=1, 2, \cdots$，则对于任意给定的正数 ε，有

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} EX_i\right| < \varepsilon\right\} = 1.$$

证明 记 $Y_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$，则

$$EY_n = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i,$$

$$DY_n = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} DX_i \leqslant \frac{c}{n}.$$

由切比雪夫不等式，得

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} EX_i\right| < \varepsilon\right\} \geqslant 1 - \frac{c}{n\varepsilon^2},$$

故

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} EX_i\right| < \varepsilon\right\} \geqslant 1.$$

又因为任何事件的概率不大于 1，所以有

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} EX_i\right| < \varepsilon\right\} = 1.$$

定理证毕.

这个结果在 1866 年被俄国数学家切比雪夫所证明，它是关于大数定律的一个相当普遍的结论，许多大数定律的古典结果是它的特例. 此外，证明这个定律所用的方法也很有创造性，在这个基础上发展起来的一系列不等式是研究各种极限定理的有力工具.

下面，我们给出切比雪夫大数定律的特殊情况.

首先，我们来回答频率与概率的关系问题. 在 n 重伯努利试验中，设事件 A 发生的次数为随机变量 X，p 是事件 A 在每次试验中发生的概率，记

$$X_i = \begin{cases} 1, & \text{第 } i \text{ 次试验中事件 } A \text{ 发生}, \\ 0, & \text{第 } i \text{ 次试验中事件 } A \text{ 不发生}. \end{cases} \quad i = 1, 2, \cdots,$$

则 $X = \sum_{i=1}^{n} X_i$. 由于 X_i 只依赖于第 i 次试验，而各次试验是相互独立的，因此 X_1，X_2，…，X_i，… 相互独立，并且都服从 0-1 分布，故有

$$EX_i = p,\ DX_i = p(1-p) \leqslant \frac{1}{4},\ i = 1, 2, \cdots.$$

由切比雪夫大数定律，可知

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| < \varepsilon\right\} = 1,$$

即

$$\lim_{n\to\infty} P\left\{\left|\frac{X}{n}-p\right|<\varepsilon\right\}=1.$$

于是有如下定理.

定理 5.1.3[伯努利(Bernoulli)大数定律]　设 μ_n 是 n 重伯努利试验中事件 A 发生的次数，p 是事件 A 在每次试验中发生的概率，则对于任意给定的正数 ε，有

$$\lim_{n\to\infty} P\left\{\left|\frac{\mu_n}{n}-p\right|<\varepsilon\right\}=1,$$

即 $\frac{\mu_n}{n}\xrightarrow{P} p.$

伯努利大数定律告诉我们，当试验次数 $n\to\infty$ 时，事件 A 发生的频率依概率收敛于该事件的概率，即当 n 充分大时，“事件 A 发生的频率 $\frac{\mu_n}{n}$ 与它的概率 p 的偏差小于任意小的正数 ε”是几乎必定发生的. 因此，我们通常把某事件发生的频率的稳定值作为该事件发生的概率. 可以说，伯努利大数定律给“频率的稳定性”提供了理论依据.

伯努利大数定律建立了在大量重复独立试验中事件出现频率的稳定性，正因为这种稳定性，概率的概念才有客观意义. 伯努利大数定律还提供了通过试验来确定事件概率的方法，即通过大量试验确定某事件频率的稳定值并把它作为该事件概率的估计. 这种估计方法称为参数的矩估计，它是数理统计中研究的参数估计方法之一，矩估计方法的理论根据就是大数定律.

另外，当随机变量序列 X_1，X_2，… 相互独立且服从相同的分布时，切比雪夫大数定律中关于方差存在的条件可以去掉，即有下面定理.

定理 5.1.4[辛钦(Khinchine)大数定律]　若随机变量 X_1，X_2，… 相互独立且服从相同的分布，X_i 的数学期望 $EX_i=\mu$ 存在，$i=1,2,\cdots$，则对于任意给定的正数 ε，有

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i-\mu\right|<\varepsilon\right\}=1,$$

即

$$\frac{1}{n}\sum_{i=1}^{n} X_i\xrightarrow{P}\mu.$$

证明略.

这个定理表明，当 n 充分大时，独立同分布且数学期望存在的随机变量序列 X_1，X_2，… 的算术平均值依概率收敛于它们的期望值 $EX_i=\mu$，这意味着当 n 充分大时，它们的算术平均值几乎变成一个常数(即期望值 μ).

5.2　中心极限定理

在实际问题中，一个现象或试验的结果往往受到大量随机因素的影响，就其单个因素来说，影响常常是微小的，但它们的综合影响常常能呈现出一定的规律性. 例如，自动机

床加工零件时所产生的误差受温度、湿度等随机因素的影响，其中每个因素的影响都独立地作用在零件上引起微小的误差. 现在需要考虑的是所有这些影响的总和会对零件产生什么样的效果. 换句话说，如果假设各随机因素的影响为 X_1，X_2，…，X_n，那么总的影响为 $Y_n = X_1 + X_2 + \cdots + X_n$. 我们感兴趣的问题是，当 n 充分大时，随机变量 Y_n 有什么样的分布？一般而言，这种随机变量往往服从或近似地服从正态分布，这就是中心极限定理的实际背景. 本节介绍三个常用的中心极限定理.

定理 5.2.1[林德伯格-列维(Lindeberg-Levy)中心极限定理]　设随机变量 X_1，X_2，… 相互独立，且具有相同的分布，记

$$EX_i = \mu,\ DX_i = \sigma^2 \neq 0,\ i = 1,\ 2,\ \cdots.$$

则对任意实数 x 有

$$\lim_{n\to\infty} P\left\{\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \leqslant x\right\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(x).$$

其中 $\Phi(x)$ 是标准正态分布的分布函数.

上述定理也常称为独立同分布的中心极限定理. 此定理表示，当 n 充分大时，

$$X = \frac{\sum_{i=1}^{n} X_i - E\left(\sum_{i=1}^{n} X_i\right)}{\sqrt{D\left(\sum_{i=1}^{n} X_i\right)}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \overset{\text{近似地}}{\sim} N(0,\ 1),$$

从而，$\sum_{i=1}^{n} X_i$ 近似服从正态分布 $N(n\mu,\ n\sigma^2)$. 若记 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，则上述结果可解释为，当 n 充分大时，

$$\frac{\overline{X} - E(\overline{X})}{D(\overline{X})} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \overset{\text{近似地}}{\sim} N(0,\ 1),\ \text{或}\ \overline{X} \overset{\text{近似地}}{\sim} N\left(\mu,\ \frac{\sigma^2}{n}\right).$$

这是独立同分布的中心极限定理的另一个形式. 也就是说，当 n 充分大时，均值为 μ，方差为 $\sigma^2 > 0$ 的独立同分布的随机变量 X_1，X_2，…，X_n 的算术平均 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 近似地服从均值为 μ，方差为 σ^2/n 的正态分布.

下面介绍一个很有用的例子.

例 5.2.1(正态随机数的产生)　在蒙特卡罗方法中经常需要产生服从正态分布的随机数，但是一般计算机只有能产生[0，1]区间上均匀分布随机数(实际上是伪随机数)的程序. 怎样通过均匀分布随机数来产生正态分布随机数呢？对这个问题有多种解决途径，其中一种是利用上述定理来实现.

设 X_1，X_2，… 是相互独立、均服从均匀分布 $U[0,\ 1]$ 的随机变量序列，容易验证它们满足定理5.2.1的条件，因此当 n 较大时，$X_1 + X_2 + \cdots + X_n$ 渐近于正态变量. 事实上，n 取不太大的值就可满足实际要求. 在蒙特卡罗方法中，一般取 $n = 12$，并用下式得到随机数序列

$$Y_k = \sum_{i=1}^{12} X_{12(k-1)+i} - 6,\ k = 1,\ 2,\ \cdots.$$

显然$\{Y_k\}$也是独立随机数序列，而且$EY_k=0$，$DY_k=1$. 经过检验证明，这时Y_k的渐近正态性已能满足一般精度要求，即近似地有$Y_k\sim N(0,1)$，$k=1,2,\cdots$.

例 5.2.2　某计算器进行加法时，将每个数舍入至其邻近的整数. 设所有的舍入是独立的，且舍入的误差值服从$[-0.5,0.5)$上的均匀分布.

(1) 若将 1 000 个数相加，求误差总和的绝对值超过 10 的概率；

(2) 问最多可有几个数相加，可使得误差之和的绝对值小于 20 的概率不小于 0.90?

解　设$X_i(i=1,2,\cdots)$为每个加数的舍入误差，由题意可知$X_i(i=1,2,\cdots)$独立且都服从$[-0.5,0.5)$上的均匀分布，因此

$$EX_i=0,\quad DX_i=\frac{[0.5-(-0.5)]^2}{12}=\frac{1}{12},\quad i=1,2,\cdots.$$

(1) 记$X=\sum_{i=1}^{1\,000}X_i$，则$\dfrac{X}{\sqrt{1\,000/12}}$近似地服从标准正态分布$N(0,1)$，所以

$$\begin{aligned}P\{|X|>10\}&=1-P\{-10\leqslant X\leqslant 10\}\\&=1-P\left\{\frac{-10}{\sqrt{1\,000/12}}\leqslant\frac{X}{\sqrt{1\,000/12}}\leqslant\frac{10}{\sqrt{1\,000/12}}\right\}\\&\approx 1-[\Phi(1.095)-\Phi(-1.095)]\\&=2-2\Phi(1.095)=0.275\ 8.\end{aligned}$$

(2) 依题意，记$Y=\sum_{i=1}^{n}X_i$，要使得$P\{|Y|<20\}\geqslant 0.90$，根据定理 5.2.1 有

$$\begin{aligned}P\{|Y|<20\}&=P\{-20<Y<20\}\\&=P\left\{\frac{-20}{\sqrt{n/12}}<\frac{Y}{\sqrt{n/12}}<\frac{20}{\sqrt{n/12}}\right\}\\&\approx\Phi\left(\frac{20}{\sqrt{n/12}}\right)-\Phi\left(\frac{-20}{\sqrt{n/12}}\right)=2\Phi\left(\frac{20}{\sqrt{n/12}}\right)-1\geqslant 0.90.\end{aligned}$$

因此有

$$\Phi\left(\frac{20}{\sqrt{n/12}}\right)\geqslant 0.95=\Phi(1.645),$$

$$\frac{20}{\sqrt{n/12}}\geqslant 1.645,\quad n\leqslant 1773.8.$$

即最多 1 773 个数相加，可使得误差之和的绝对值小于 20 的概率不小于 0.90.

大数定律断言：当$n\to\infty$时，$P\left\{\left|\dfrac{\mu_n}{n}-p\right|<\varepsilon\right\}$趋于 1，即$\dfrac{\mu_n}{n}$接近于$p$. 而下面的棣莫弗－拉普拉斯极限定理则给出了$\mu_n$的渐近分布的更精确表述.

定理 5.2.2[棣莫弗–拉普拉斯(De Moivre－Laplace)定理]　设$X\sim B(n,p)$，则对于任意x有

$$\lim_{n\to\infty}P\left\{\frac{X-np}{\sqrt{np(1-p)}}\leqslant x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\Phi(x).$$

这个定理指出，在n重伯努利试验中，当试验次数n充分大时，二项分布可以用正态

分布近似，即 X 近似服从正态分布 $N(np,\ np(1-p))$. 我们知道，X 为 n 重伯努利试验中事件 A 发生的次数，p 是事件 A 在每次试验中发生的概率. 若记

$$X_i=\begin{cases}1, & \text{第 } i \text{ 次试验中事件 } A \text{ 发生},\\ 0, & \text{第 } i \text{ 次试验中事件 } A \text{ 不发生}.\end{cases}\qquad i=1,\ 2,\ \cdots,$$

则 $X=\sum\limits_{i=1}^{n}X_i$. 因此 X 近似服从正态分布 $N(np,\ np(1-p))$ 等价于频率 μ_n/n 近似服从正态分布 $N(p,\ p(1-p)/n)$，这里 $\mu_n=X$ 为 n 重伯努利试验中事件 A 发生的次数.

例 5.2.3　某工厂生产的某种产品，其次品率为 0.01，今取 500 个装一盒，求一盒中次品数不多于 9 个的概率.

解　设 X 为盒中的次品数，则 $X\sim B(500,\ 0.01)$，$EX=500\times0.01=5$，$DX=500\times0.01\times0.99=4.95$，由定理 5.2.2，得所求概率为

$$\begin{aligned}P\{0\leqslant X\leqslant 9\}&=P\left\{\frac{0-5}{\sqrt{4.95}}\leqslant\frac{X-5}{\sqrt{4.95}}\leqslant\frac{9-5}{\sqrt{4.95}}\right\}\\&\approx\Phi\left(\frac{9-5}{\sqrt{4.95}}\right)-\Phi\left(\frac{0-5}{\sqrt{4.95}}\right)=\Phi(1.8)-\Phi(-2.25)\\&=0.964\ 1-0.012\ 2=0.951\ 5.\end{aligned}$$

该问题也可用泊松分布近似计算. 由 $\lambda=np=500\times0.01=5$，可得

$$P\{0\leqslant X\leqslant 9\}=\sum_{i=0}^{9}C_{500}^{i}\,0.01^{i}\,0.99^{500-i}\approx\sum_{i=0}^{9}\frac{5^{i}\mathrm{e}^{-5}}{k!}=0.968\ 2.$$

第 2 章中我们讨论过用泊松分布近似二项分布的结论. 一般来说，在 np 适中且 p 较小的情况下，用泊松分布近似比较有效，而当 np 较大时，用正态分布来近似二项分布比较好.

例 5.2.4　对于一个学生而言，其家长来参加会议的人数是一个随机变量，设一个学生无家长、1 个家长、2 个家长来参加会议的概率分别为 0.05，0.8，0.15. 若某学校共有 400 名学生，设各学生参加会议的家长人数相互独立，且服从同一分布.

（1）求参加会议的家长总人数 X 超过 450 的概率；

（2）求有 1 名家长来参加会议的学生人数不多于 340 的概率.

解　（1）设 $X_i(i=1,\ 2,\ \cdots,\ 400)$ 为第 i 个学生来参加会议的家长人数，则 X_i 的分布律为

X_i	0	1	2
p_i	0.05	0.8	0.15

易知 $EX_i=1.1$，$DX_i=0.19$，$i=1,\ 2,\ \cdots,\ 400$. 而 $X=\sum\limits_{i=1}^{400}X_i$，由定理 5.2.1，有

$$\frac{\sum\limits_{i=1}^{400}X_i-400\times1.1}{\sqrt{400\times0.19}}=\frac{X-400\times1.1}{\sqrt{400\times0.19}}$$

近似服从正态分布 $N(0,\ 1)$，于是有

$$P\{X>450\}=P\left\{\frac{X-400\times1.1}{\sqrt{400\times0.19}}>\frac{450-400\times1.1}{\sqrt{400\times0.19}}\right\}$$
$$=1-P\left\{\frac{X-400\times1.1}{\sqrt{400\times0.19}}\leqslant1.147\right\}$$
$$\approx1-\Phi(1.147)=0.125\ 1.$$

（2）以 Y 记有一名家长参加会议的学生人数，则 $Y\sim B(400,0.8)$，由定理 5.2.2，有

$$P\{Y\leqslant340\}=P\left\{\frac{Y-400\times0.8}{\sqrt{400\times0.8\times0.2}}\leqslant\frac{340-400\times0.8}{\sqrt{400\times0.8\times0.2}}\right\}$$
$$=P\left\{\frac{Y-400\times0.8}{\sqrt{400\times0.8\times0.2}}\leqslant0.25\right\}$$
$$\approx\Phi(2.5)=0.993\ 8.$$

习题五

1. 设 X 为随机变量，$EX=\mu$，$DX=\sigma^2$，试估计概率 $P\{|X-\mu|<3\sigma\}$.

2. 某路灯管理所有 20 000 只路灯，夜晚每盏路灯开的概率为 0.6，设路灯开关是相互独立的，试用切比雪夫不等式估计夜晚同时开着的路灯数在 11 000 ~ 13 000 盏之间的概率.

3. 在 n 重伯努利试验中，若已知每次试验中事件 A 出现的概率为 0.75，请利用切比雪夫不等式估计 n，使 A 出现的频率在 0.74 ~ 0.76 之间的概率不小于 0.90.

4. 某批产品合格率为 0.6，任取 10 000 件，其中合格品在 5 980 ~ 6 020 件之间的概率是多少？

5. 某保险公司有 3 000 个同一年龄段的人参加人寿保险，在一年中这些人的死亡率为 0.1%. 参加保险的人在一年的开始交付保险费 100 元，死亡时家属可从保险公司领取 10 000 元. 求

（1）保险公司一年获利不少于 240 000 元的概率；

（2）保险公司亏本的概率.

6. 计算器在进行加法时，将每个加数舍入最靠近它的整数，设所有舍入误差相互独立且在(-0.5，0.5）上服从均匀分布.

（1）将 1 500 个数相加，问误差总和的绝对值超过 15 的概率是多少？

（2）最多可有几个数相加使得误差总和的绝对值小于 10 的概率不小于 0.9？

7. 对敌人的防御地带进行 100 次轰炸，每次轰炸命中目标的炸弹数目是一个均值为 2，方差为 1.69 的随机变量. 求在 100 次轰炸中有 180 ~ 220 颗炸弹命中目标的概率.

8. 有一批建筑房屋用的木柱，其中 80% 的长度不小于 3 米，现从这批木柱中随机地取 100 根，求其中至少有 30 根短于 3 米的概率.

9. 分别用切比雪夫不等式与棣莫弗 -拉普拉斯定理确定：当掷一枚硬币时，需要掷多少次才能保证出现正面的频率在 0.4 ~ 0.6 之间的概率不少于 0.9？

10. 已知在某十字路口一周内事故发生数的数学期望为 2.2，标准差为 1.4.

（1）以 $\overline{X}$ 表示一年内（以 52 周计）此十字路口事故发生数的算术平均，使用中心极限定理求 $\overline{X}$ 的近似分布，并求 $P\{\overline{X}<2\}$；

（2）求一年内事故发生数小于 100 的概率.

11. 为检验一种新药对某种疾病的治愈率为 80% 是否可靠，给 10 个患该疾病的病人同时服药，结果治愈人数不超过 5 人，试判断该药的治愈率为 80% 是否可靠.

12. 一公寓有 200 个住户，一个住户拥有汽车辆数 X 的分布律为

X	0	1	2
p_k	0.1	0.6	0.3

问需要多少车位，才能使每辆汽车都有一个车位的概率至少为 0.95？

13. 甲、乙两个戏院在竞争 1 000 名观众，假设每个观众可随意选择戏院，观众之间相互独立，问每个戏院应该设有多少座位才能保证因缺少座位而使观众离去的概率小于 1%？

第六章　抽样分布理论

前面几章内容属于概率论的范畴. 从本章开始，我们介绍数理统计的基本概念和基本方法. 概率论是研究随机现象的模型，即概率分布. 数理统计是研究随机现象的数据分析与处理方法.

在概率论中，一般假定概率分布是已知的. 然而，在实际中，随机变量的分布往往是未知的，或分布形式虽然已知，但其中的参数是未知的. 例如，某条道路某天的交通事故数所服从的分布是什么？某商场每天服务的顾客数所服从的分布是什么等. 又如，一般认为学生的考试成绩近似服从正态分布，但其数学期望(平均成绩)、方差等参数是事先未知的. 如何判断随机变量服从的分布？以及如果已知它具有某种分布形式，如何确定分布中的参数等，这些问题都是数理统计要研究的内容. 数理统计是以概率论为基础，研究如何收集数据，如何依据数据对总体做出合理推断的学科.

随着计算机技术的发展，数理统计方法得到了迅速的发展，研究内容非常丰富，且形成了许多学科分支和交叉学科. 基本的数理统计方法主要有两大类，即描述性统计方法和推断统计方法. 本书主要介绍推断统计方法的一些基本内容，包括：抽样分布、参数估计、假设检验、方差分析和回归分析等.

6.1　样本与统计量

6.1.1　总体与样本

在一个统计问题中，我们把研究对象的全体称为**总体**，构成总体的每一个单元称为**个体**. 总体按照个体的数量多少可以分为**有限总体**和**无限总体**. 在大多数实际问题中，总体中的个体都是实实在在的人和物. 例如，研究某一年龄段学生的身体状况，则该年龄段全体学生就是总体，而每个学生都是个体. 但实际上，我们真正关心的并不一定是总体或个体本身，而真正关心的是总体或个体的某项数量指标 X(或几个数量指标). 例如，研究某一年龄段学生的身体状况时，研究者主要关心的是该年龄段学生的身高和体重等指标. 研究中，有时也将总体理解为那些研究对象的某项数量指标的全体，将总体的某项数量指标 X 可能取值的全体组成的集合视为总体. 例如，研究某品牌电视机的彩色浓度时，该品牌电视机的彩色浓度数据的集合就是一个总体，而该品牌每台电视机的彩色浓度都是个体.

由于大量随机现象必然呈现出某种规律性，因而从理论上讲，只要对随机现象进行足够多的观察，随机现象的规律性就一定能够清楚地呈现出来. 但是，如同研究某饮料的合格情况，不必要对每一瓶饮料进行检验，研究某型号的导弹的威力，不可能把该型号所有导弹都打光一样，客观上只允许我们对随机现象进行有限次数的观察或试验. 换句话说，我们只能从总体中抽取部分样本，利用样本的数据信息推断出总体的规律. 在统计工作

中，通常对某个总体抽取一部分个体进行观测，这个过程称为**抽样**. 从总体中抽取的个体必须满足随机性，可采取一些抽样方法来保证. 每次抽取的个体数量 n 称为**样本容量**，抽取的 n 个体 X_1，X_2，…，X_n 称为总体的一个**样本**，样本会随着每次抽样观测的不同而随之变化，且不能预测，故样本 X_1，X_2，…，X_n 是随机变量，但是一旦具体抽取之后，就得到一组确定的数值，用 x_1，x_2，…，x_n 来表示，称为**样本观测值**.

在抽取样本的过程中，为了能让样本能够较好反映总体的特性，一般采取简单随机抽样方法，它有如下两个特征.

（1）代表性：样本 X_1，X_2，…，X_n 中每个分量 X_i 与总体 X 具有相同的概率分布.

（2）独立性：样本 X_1，X_2，…，X_n 相互独立.

采用简单随机抽样方法得到的样本称之为**简单随机样本**，今后如不特殊说明，所说的样本均指简单随机样本.

若简单随机样本 X_1，X_2，…，X_n 来自某总体 X，并假设总体 X 具有连续型分布，其密度函数为 $f(x)$，由于 X_1，X_2，…，X_n 与总体 X 具有相同的分布，可知样本 X_1，X_2，…，X_n 的联合密度函数为

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

例如，为估计一物件的重量 μ，用一架天平重复测量 n 次，得到样本 X_1，X_2，…，X_n，由于是独立重复测量，X_1，X_2，…，X_n 是简单随机样本. 假设物件的重量 $X\sim N(\mu,\ \sigma^2)$，其密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},$$

则样本 $(X_1,\ X_2,\ \cdots,\ X_n)$ 的联合密度函数为

$$f(x_1,\ x_2,\ \cdots,\ x_n)=\prod_{i=1}^{n}f(x_i)=(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}.$$

对离散型随机变量，假设总体 X 具有分布律 $p(x)=P(X=x)$，则样本 $(X_1,\ X_2,\ \cdots,\ X_n)$ 的联合分布律为

$$p(x_1,\ x_2,\ \cdots,\ x_n)=\prod_{i=1}^{n}p(x_i).$$

例 6.1.1 设某电话交换台一小时内的呼唤次数 X 服从泊松分布 $P(\lambda)$，求来自这一总体的样本 X_1，X_2，…，X_n 的联合分布律.

解 X 的分布律为

$$P_X(x)=P(X=x)=\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}(\lambda>0),\ x=0,\ 1,\ 2,\ \cdots,$$

从而得到样本 X_1，X_2，…，X_n 的联合分布律为

$$p(x_1,\ x_2,\ \cdots,\ x_n)=\prod_{i=1}^{n}p(x_i)=\frac{\lambda^{\sum\limits_{i=1}^{n}x_i}}{\prod\limits_{i=1}^{n}x_i!}\mathrm{e}^{-n\lambda}.$$

6.1.2　统计量

样本是总体的代表和反映，但在统计推断中，往往不是直接使用样本本身，而是根据具体问题的特点对样本进行“加工”或“提炼”，把样本中包含的关于所研究问题的信息集中起来，这个过程就是针对不同的问题构造样本的适当函数. 然后利用这些样本函数进行统计推断. 为此，下面引进统计量的概念.

定义 6.1.1　设 X_1，X_2，…，X_n 是来自总体 X 的一个样本，$g(X_1, X_2, \cdots, X_n)$ 为 X_1，X_2，…，X_n 的一个函数，若该函数中不含任何未知参数，则称 $g(X_1, X_2, \cdots, X_n)$ 是一个统计量.

例如，从总体 $X \sim (\mu, \sigma^2)$ 中抽取样本 X_1，X_2，…，X_n，其中 μ 为已知参数，σ 为未知参数，则

(1) $\dfrac{1}{n}(X_1 + X_2 + \cdots + X_n) - \mu$ 是统计量，因为它不含未知参数；

(2) $\dfrac{X_1 + X_2 + \cdots + X_n - nu}{\sigma}$ 不是统计量，因为它包含未知参数 σ.

下面介绍一些常用的统计量：样本均值、样本方差、样本标准差和极差等.

设 X_1，X_2，…，X_n 是来自总体 X 的一个样本，则统计量

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

称为样本均值；统计量

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

称为样本方差；统计量 $S = \sqrt{S^2}$ 称为样本标准差；统计量

$$R = \max_{1 \leqslant i \leqslant n} \{X_i\} - \min_{1 \leqslant i \leqslant n} \{X_i\}$$

称为样本极差；统计量

$$C_v = \frac{S}{|\overline{X}|}$$

称为样本变异系数.

*6.1.3　经验分布函数

在很多理论和实际问题中，总体 X 的分布函数 $F(x)$ 是未知的，能否根据样本的观测值来估计总体的分布函数呢？经验分布函数是样本观测值的函数，可用于对总体 X 的分布函数的估计. 下面，我们给出经验分布函数的概念.

定义 6.1.2　设总体 X 的样本 X_1，X_2，…，X_n 的观测值为 x_1，x_2，…，x_n，将这组值由小到大排列成 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$，令

$$F_n(x)=\begin{cases}0, & \text{当 } x<x_{(1)},\\ \dfrac{k}{n}, & \text{当 } x_{(k)}\leqslant x<x_{(k+1)}, \quad k=1,2,\cdots,n-1,\\ 1, & \text{当 } x\geqslant x_{(n)}.\end{cases}$$

则称 $F_n(x)$ 为该样本的经验分布函数.

经验分布函数 $F_n(x)$ 在点 x 的函数值其实就是观测值 x_1，x_2，…，x_n 中小于或等于 x 的频率. 易得 $F_n(x)$ 的如下性质.

(1) $F_n(x)$ 单调非降且右连续;

(2) $0\leqslant F_n(x)\leqslant 1$;

(3) $F_n(-\infty)=0$，$F_n(+\infty)=1$.

由上述性质可知，经验分布函数 $F_n(x)$ 具有分布函数 $F(x)$ 的性质.

值得注意的是，对于固定的 x，经验分布函数是依赖于样本观测值的，由于样本观测值的抽取是随机的，因而 $F_n(x)$ 是样本的函数，故它是一个统计量.

例 6.1.2 有 10 位儿童身高如下(单位：厘米).

148，139，145，149，148，140，145，151，139，145

求对应于上述数据的经验分布函数 $F_{10}(x)$.

解 由所给数据可得如下频数表.

身高值	139	140	145	148	149	151
频数	2	1	3	2	1	1

由此得到经验分布函数为

$$F_{10}(x)=\begin{cases}0, & x<139,\\ 0.2, & 139\leqslant x<140,\\ 0.3, & 140\leqslant x<145,\\ 0.6, & 145\leqslant x<148,\\ 0.8, & 148\leqslant x<149,\\ 0.9, & 149\leqslant x<151,\\ 1, & x\geqslant 151.\end{cases}$$

由大数定律可知，随机事件发生的频率依概率收敛于该事件发生的概率. 因此可以用事件 $\{X\leqslant x\}$ 发生的频率 $\dfrac{k}{n}$ 来估计概率 $P\{X\leqslant x\}$，即用经验分布函数 $F_n(x)$ 来估计总体 X 的理论分布函数 $F(x)=P\{X\leqslant x\}$. 格里汶科(W. Glivenko) 于 1933 年从理论上证明了以下结论.

定理 6.1.1 设总体 X 的分布函数为 $F(x)$，经验分布函数为 $F_n(x)$，则当 $n\to\infty$ 时，$F_n(x)$ 以概率 1 一致收敛于分布函数 $F(x)$，即

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

定理 6.1.1 表明，对于任一实数 x，当 n 充分大时，经验分布函数 $F_n(x)$ 收敛于总体 X 的真实分布函数 $F(x)$，因而在实际中可以用 $F_n(x)$ 来估计 $F(x)$.

*6.1.4　数据的简单处理与显示

原始数据一般通过抽样调查或试验得到，通过整理后才能显示出其分布特征. 数据的简单处理包括数据整理与计算样本特征数两方面，下面通过一个实例来说明它的用法.

例 6.1.3　为对某小麦杂交组合 F_2 代的株高 X 进行研究，抽取容量为 100 的样本，测试的原始数据记录如下(单位：厘米).

87	88	111	91	73	70	92	98	105	94
99	91	98	110	98	97	90	83	92	88
86	94	102	99	89	104	94	94	92	96
87	94	92	86	102	88	75	90	90	80
84	91	82	94	99	102	91	96	94	94
85	88	80	83	81	69	95	80	97	92
96	109	91	80	80	94	102	80	86	91
90	83	84	91	87	95	76	90	91	77
103	89	88	85	95	92	104	92	95	83
86	81	86	91	89	83	96	86	75	92

为了研究随机变量 X 的分布状况，可根据以上数据，进行简单处理与显示分析，具体步骤如下.

第一步：整理原始数据，将其加工为分组资料，制作频率分布表，并画出直方图.

(1) 找出数据中的最小值 m、最大值 M 和极差 $R=M-m$. 本例中，$m=69$，$M=111$，$R=42$.

(2) 数据分组. 根据样本容量 n 的大小，决定分组数 k. 本例中，$n=100$，取 $k=9$. 一般采取等距分组(也可采用不等距分组) 方法，组距等于或略大于极差与组数之比. 本例中，我们采取等距分组方法，组距取为 5.

(3) 计算出各组频数，制作频数、频率分布表(见表6.1)，其中频数 v_j 为落在第 j 个小区间的数据个数.

表 6.1　频数、频率分布表

组序	区间范围	频数 v_j	频率 $f_j=v_j/n$	累计频率 F_j
1	[67.5，72.5]	2	0.02	0.02
2	[72.5，77.5]	5	0.05	0.07
3	[77.5，82.5]	10	0.10	0.17
4	[82.5，87.5]	18	0.18	0.35
5	[87.5，92.5]	30	0.30	0.65
6	[92.5，97.5]	18	0.18	0.83
7	[97.5，102.5]	10	0.10	0.93
8	[102.5，107.5]	4	0.04	0.97
9	[107，5，112.5]	3	0.03	1.00

(4) 画出频率直方图，即以分组区间为底，频率 / 组距为高，画出一系列矩形(见图 6.1).

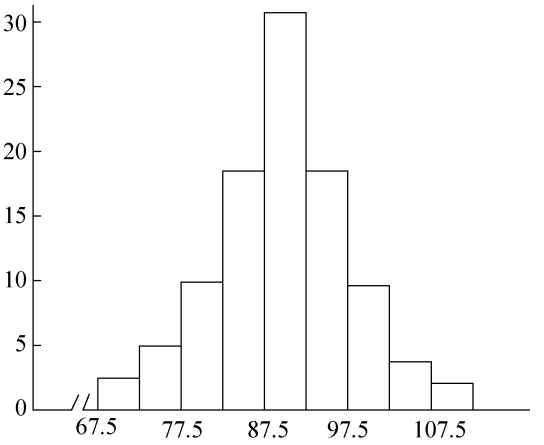


图 6.1 频率直方图

在频率直方图中，每个矩形面积恰好等于样本落在该矩形对应的分组区间内的频率 f_j. 因此所有矩形面积之等于 1. 频率直方图中的小矩形的面积近似反映样本数据落在某个区间内的概率，故它可近似描述 X 的分布状况.

第二步：计算两个方面的样本特征数.

(1) 反映集中趋势的特征数. 此类特征数有样本均值、中位数、众数等，其中样本均值为

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{100}(87+88+\cdots+92) = 90.30.$$

若把原始数据按大小顺序排列成 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$，则当 n 为奇数时，居中的那个数即为中位数；当 n 为偶数时，位于中间位置的两个数的平均值为中位数，本例中，$n = 100$，故中位数取为

$$M_e = \frac{1}{2}(x_{(50)} + x_{(51)}) = 91.00.$$

众数为样本中出现频数最多的那个数，本例中众数为 $M_o = 91.00$.

(2) 反映分散程度的特征数. 此类特征数有样本方差、样本标准差、极差、变异系数等，其中样本方差为

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{99}[(87-90.3)^2 + \cdots + (92-90.3)^2] \approx 68.69,$$

样本标准差为 $S = \sqrt{68.69} = 8.288$，极差为

$$R = \max_{1\leqslant i\leqslant n}\{X_i\} - \min_{1\leqslant i\leqslant n}\{X_i\} = M - m = 111 - 69 = 42,$$

变异系数为

$$C_v = \frac{S}{|\overline{X}|} = \frac{8.288}{90.3} \approx 0.092.$$

上述特征统计量的值越小，表示离散程度越小.

6.2　抽样分布

统计量是随机变量，统计量的分布称为抽样分布. 研究统计量的分布是统计推断的重要内容. 这一节主要介绍以标准正态分布为基石而构造的三个著名统计量的分布，它们被称为统计中三大抽样分布，即χ^2 分布、t 分布和 F 分布.

首先引入分位数的定义.

定义 6.2.1　对于总体 X 和给定的正数 $\alpha(0<\alpha<1)$，若存在实数 x_α，满足

$$P(X \geqslant x_\alpha) = \alpha \tag{6.2.1}$$

则称 x_α 为 X 的上侧 α 分位数. 若 X 为连续型随机变量，其密度函数为 $f(x)$，则 X 的上侧 α 分位数 x_α 满足

$$P(X \geqslant x_\alpha) = \int_{x_\alpha}^{+\infty} f(x)\,\mathrm{d}x = \alpha,$$

对于标准正态分布变量 $U \sim N(0,1)$，上侧 α 分位数 z_α 由

$$P\{U \geqslant u_\alpha\} = \int_{u_\alpha}^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t = \alpha \tag{6.2.2}$$

确定. 其几何意义如图 6.2 所示，其中阴影部分的面积为 α.

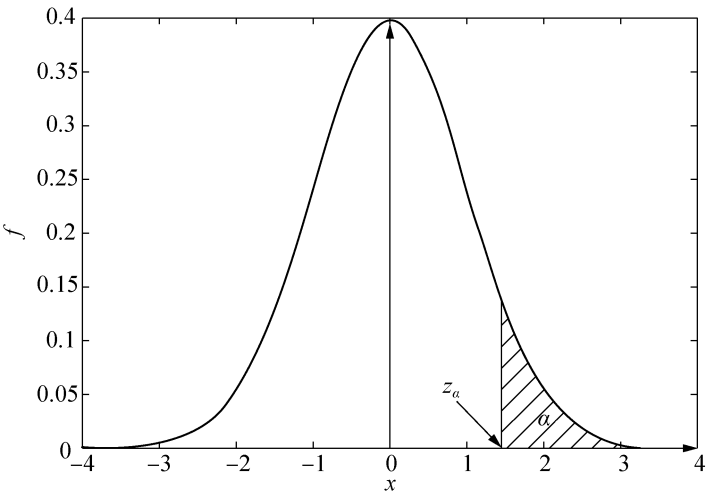


图 6.2　标准正态分布的上侧 α 分位数 z_α

由于标准正态分布的密度函数关于纵坐标轴对称，因此容易知道 $P\{U<-u_\alpha\}=\alpha$，于是有 $P\{U \geqslant -u_\alpha\}=1-\alpha$，因此标准正态分布的上侧分位数满足 $u_{1-\alpha}=-u_\alpha$. 例如，若 $\alpha=0.05$，则由于 $P\{U \geqslant 1.645\}=0.05$，知 $u_\alpha=1.645$. 而 $u_{1-\alpha}=u_{0.95}=-u_{0.05}=-1.645$ 满足 $P(U \geqslant -1.645)=1-\alpha=0.95$. 一般来说，若某随机变量 X 的密度曲线 $f(x)$ 关于纵坐标轴对称，则其上侧分位数满足 $x_{1-\alpha}=-x_\alpha$.

6.2.1 χ^2 分布

定义 6.2.2 设 X_1，X_2，…，X_n 为来自标准正态分布 $X\sim N(0,1)$ 的样本，则称随机变量 $Y=X_1^2+X_2^2+\cdots+X_n^2$ 所服从的分布为自由度为 n 的 χ^2 分布，记为 $Y\sim\chi^2(n)$.

$\chi^2(n)$ 分布的密度函数为

$$f(y)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)}y^{\frac{n}{2}-1}\mathrm{e}^{-\frac{y}{2}}, & y\geqslant 0,\\ 0, & y<0.\end{cases}\tag{6.2.3}$$

将密度函数(6.2.3)与第二章中伽玛分布的密度函数(2.4.15)对照可知 $\chi^2(n)$ 分布是一个特殊的伽玛分布，即 $\chi^2(n)=Ga\left(\dfrac{n}{2},\ \dfrac{1}{2}\right)$.

根据伽玛分布的可加性质，见例 3.8.5 和式(3.8.7)，及伽玛分布的期望和方差结果可知，χ^2 具有下列两条重要结论.

(1) 可加性：若 $Y_1\sim\chi^2(n_1)$，$Y_2\sim\chi^2(n_2)$，且 Y_1 与 Y_2 独立，则 $Y_1+Y_2\sim\chi^2(n_1+n_2)$.

(2) 若 $Y\sim\chi^2(n)$，则 $E(Y)=n$，$D(Y)=2n$，即 χ^2 分布的期望等于其自由度，方差等于其自由度的 2 倍.

由定义 6.2.1 可知，对于给定的正数 $\alpha(0<\alpha<1)$，满足下列条件

$$P(\chi^2(n)\geqslant\chi_\alpha^2(n))=\int_{\chi_\alpha^2(n)}^{+\infty}f(y)\,\mathrm{d}y=\alpha\tag{6.2.4}$$

的实数 $\chi_\alpha^2(n)$ 为 $\chi^2(n)$ 分布的上侧 α 分位数，其几何意义如图 6.3 所示.

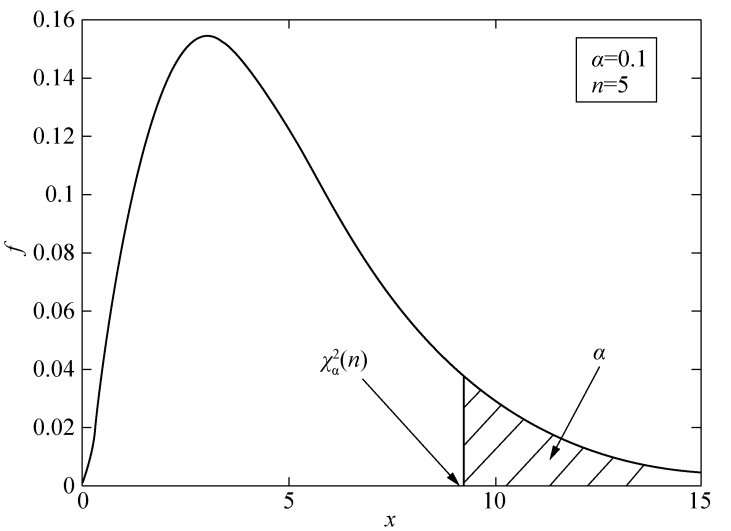


图 6.3 χ^2 分布的上侧 α 分位数

给定自由度 n 后，分位数 $\chi_\alpha^2(n)$ 的值取决于正数 $\alpha(0<\alpha<1)$. 对应于不同的 n 和 α，$\chi_\alpha^2(n)$ 的值可通过查表得到(见附表 3).

6.2.2　t 分布

定义 6.2.3　设随机变量 $X \sim N(0,\ 1)$，$Y \sim \chi^2(n)$，且 X 与 Y 独立，则称随机变量

$$t = \frac{X}{\sqrt{Y/n}} \tag{6.2.5}$$

所服从的分布为自由度为 n 的 t 分布，记为 $t \sim t(n)$.

t 分布又称为学生分布(student distribution)，其密度函数为

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}},\quad -\infty < t < +\infty \tag{6.2.6}$$

其图形如图 6.4 所示，形状类似于标准正态分布 $N(0,\ 1)$ 的密度函数图.

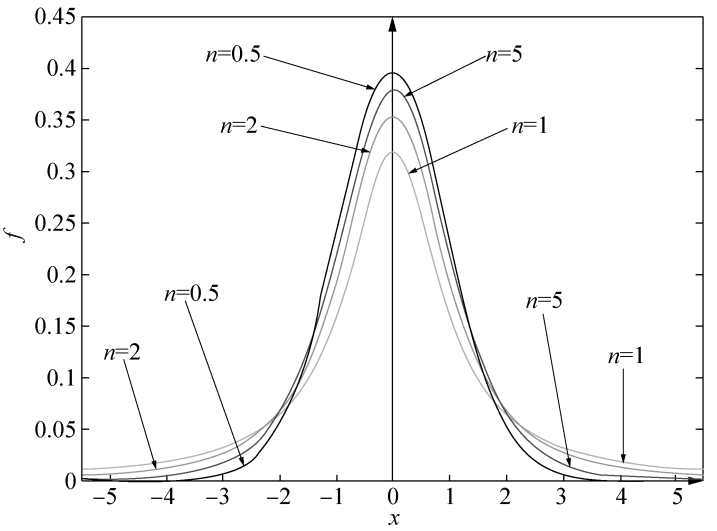


图 6.4　t 分布密度函数图

从图 6.4 不难看出，t 分布的密度关于纵轴对称，即该密度函数是一个偶函数. 对一切 $n = 2,\ 3\cdots$，t 分布的数学期望为 $E(t) = 0$，但当 $n = 1$ 时 t 分布的数学期望不存在. 对一切 $n = 3,\ 4\cdots$，t 分布的方差为 $D(t) = \dfrac{n}{n-2}$，但当 $n = 1,\ 2$ 时 t 分布的方差不存在.

设 $t \sim t(n)$，则由定义 6.2.1 可知，对于给定的 $\alpha(0 < \alpha < 1)$，满足条件

$$P(t(n) \geqslant t_\alpha(n)) = \int_{t_\alpha(n)}^{+\infty} f(t)\,\mathrm{d}t = \alpha$$

的实数 $t_\alpha(n)$ 为 t 分布的上侧 α 分位数，其几何意义如图 6.5 所示. 由 t 分布的对称性可知，$t_{1-\alpha}(n) = -t_\alpha(n)$.

在附表 4 中给出了 t 分布的临界值表，例如，当 $\alpha = 0.05$，$n = 15$ 时，查 t 分布临界值表有

$$t_{0.05}(15) = 1.753,\ t_{\frac{0.05}{2}}(15) = 2.131.$$

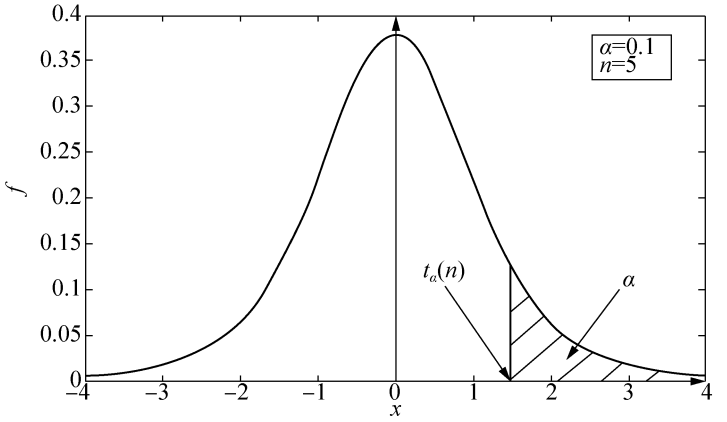


图 6.5　t 分布的上侧 α 分位数

6.2.3　F 分布

定义 6.2.4　设随机变量 $Y_1 \sim \chi^2(n_1)$，$Y_2 \sim \chi^2(n_2)$，且 Y_1 与 Y_2 独立，则称随机变量

$$F = \frac{Y_1/n_1}{Y_2/n_2} \tag{6.2.7}$$

所服从的分布为第一自由度为 n_1，第二自由度为 n_2 的 F 分布，记作 $F \sim F(n_1, n_2)$.

F 分布的密度函数为

$$f(y) = \begin{cases} \dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)}\left(\dfrac{n_1}{n_2}\right)^{\frac{n_1}{2}} y^{\frac{n_1}{2}-1}\left(1+\dfrac{n_1}{n_2}y\right)^{-\frac{n_1+n_2}{2}}, & y \geqslant 0, \\ 0, & y < 0, \end{cases} \tag{6.2.8}$$

其图形如图 6.6 所示.

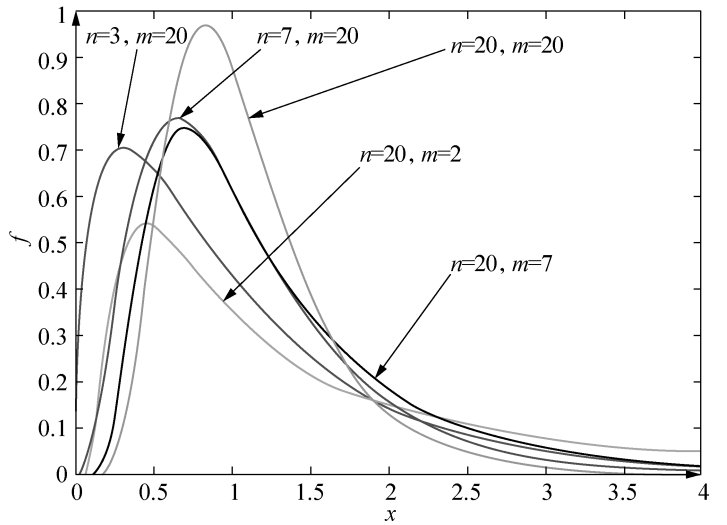


图 6.6　F 分布密度函数图

由 F 分布的定义易知，若 $F \sim F(n_1, n_2)$，则 $\frac{1}{F} \sim F(n_2, n_1)$.

类似于 χ^2 分布和 t 分布，F 分布的上侧 α 分位数是指满足条件

$$P(F(n_1, n_2) > F_\alpha(n_1, n_2)) = \int_{F_\alpha(n_1, n_2)}^{+\infty} f(y)\,\mathrm{d}y = \alpha$$

的 $F_\alpha(n_1, n_2)$，其几何意义如图 6.7 所示.

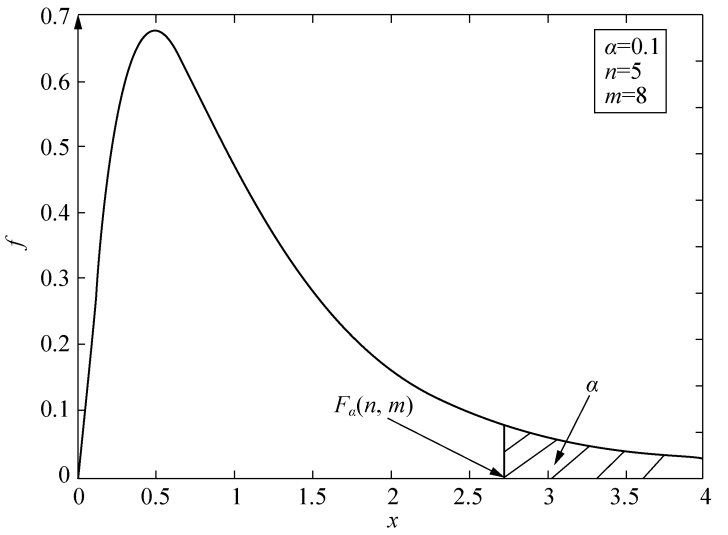


图 6.7　F 分布的上侧 α 分位数

$F_\alpha(n_1, n_2)$ 的值可由附表 5 查得，例如 $n_1 = 10$，$n_2 = 15$，$\alpha = 0.01$，有 $F_{0.01}(10, 15) = 3.80$. 由 F 分布的性质，可得以下关系式

$$F_{1-\alpha}(n_1, n_2) = \frac{1}{F_\alpha(n_2, n_1)} \tag{6.2.9}$$

例如 $F_{0.90}(10, 15) = \dfrac{1}{F_{0.10}(15, 10)} = \dfrac{1}{2.24} \approx 0.446$.

6.3　样本均值和样本方差的分布

6.3.1　大样本情况下样本均值的分布

首先，对于一般的总体，我们利用中心极限定理求样本均值的近似分布. 由于所得结论只有当样本容量 n 充分大时才成立，所以也称为大样本分布.

定理 6.3.1　设 X_1，X_2，…，X_n 是来自于均值为 μ，方差为 σ^2 的某总体的一组样本，当 n 充分大时，则近似地有

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

证　因为 X_1，X_2，…，X_n 是来自于均值为 μ，方差为 σ^2 的总体的样本是独立同分布的，

且 $E(X_i)=\mu$，$Var(X_i)=\sigma^2$，$i=1,2,\cdots n$，由中心极限定理，对于充分大的 n，近似地有

$$\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1),$$

即对于充分大的 n，近似地有 $\bar{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$.

该定理说明，不论总体是何种分布，只要其均值是 μ，方差是 σ^2，则来自这个总体的样本均值 $\bar{X}$ 就近似地服从均值为 μ、方差为 $\frac{\sigma^2}{n}$ 的正态分布总体 $N\left(\mu,\frac{\sigma^2}{n}\right)$.

例 6.3.1　设某种零件的平均长度为 0.50cm，标准差为 $\sigma=0.04$cm，从该种零件的总体中随机抽出 100 个样本，问这 100 个零件的平均长度小于 0.49cm 的概率是多少？

解　虽然我们不知道该种零件的总体分布是什么，但由定理 6.3.1 可知，样本均值 $\bar{X}$ 近似服从正态分布，即 $\bar{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$，$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$，因此所求概率为

$$P\{\bar{X}<0.49\}=P\left\{\frac{\bar{X}-\mu}{\sigma/n}<\frac{0.49-\mu}{\sigma/n}\right\}=P\left\{\frac{\bar{X}-100}{0.004}<\frac{0.49-100}{0.004}\right\}$$

$$\approx\Phi\left(\frac{x-100}{0.004}\right)=\Phi(-2.5)=1-\Phi(2.5)=0.006\ 2.$$

6.3.2　正态总体的样本均值和样本方差的分布

对于正态总体来说，关于样本均值和样本方差以及某些重要统计量的抽样分布具有非常完美的理论结果，它们在假设检验和参数估计的讨论中具有非常重要的作用. 这些内容可归结为下列的一些定理结论.

定理 6.3.2　设 X_1，X_2，…，X_n 是来自正态分布总体 $N(\mu,\sigma^2)$ 的样本，则

(1) $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\sim N\left(\mu,\frac{\sigma^2}{n}\right)$；

(2) $\frac{(n-1)S^2}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2\sim\chi^2(n-1)$；

(3) $\bar{X}$ 和 S^2 相互独立；

(4) $\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$.

证明　(2) 和(3) 的证明方法超出本书范围，我们仅证明(1) 和(4). 事实上，若总体 X 服从正态分布 $N(\mu,\sigma^2)$，其样本均值 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 仍服从正态分布，且其数学期望和方差分别为

$$E(\bar{X})=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_I)=\frac{1}{n}\cdot n\mu=\mu,$$

$$D(\bar{X})=D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)=\frac{1}{n^2}\cdot n\sigma^2=\frac{\sigma^2}{n},$$

即 $\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$. 对于(4)，由(1)可知

$$U = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1),$$

由(2)和(3)可知

$$V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

且 U 与 V 独立. 于是由 t 分布的定义可知

$$t = \frac{U}{\sqrt{V/(n-1)}} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1).$$

定理 6.3.3　设 $X_1, X_2, \cdots X_{n_1}$ 和 $Y_1, Y_2, \cdots, Y_{n_2}$ 是分别来自于正态分布总体 $N(\mu_1, \sigma^2)$ 和 $N(\mu_2, \sigma^2)$ 的样本，且它们相互独立，则

(1) 统计量

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim t(n_1+n_2-2),$$

其中 $S = \sqrt{\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$，$S_1^2$ 和 S_2^2 分别是两总体的样本方差.

(2) 统计量

$$\frac{S_1^2}{S_2^2} \sim F(n_1-1, n_2-1).$$

证明　由定理 6.3.2 可知

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right), \quad \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right),$$

且由于两样本间相互独立，可知 $\overline{X}$ 与 $\overline{Y}$ 独立，因此有

$$\overline{X}-\overline{Y} \sim N\left(\mu_1-\mu_2, \left(\frac{1}{n_1}+\frac{1}{n_2}\right)\sigma^2\right),$$

$$U = \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\sigma} \sim N(0, 1).$$

又由于

$$V_1 = \frac{(n_1-1)S_1^2}{\sigma^2} \sim \chi^2(n_1-1), \quad V_2 = \frac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_2-1),$$

且 V_1 与 V_2 独立，由 χ^2 分布的可加性知 $V = V_1+V_2 \sim \chi^2(n_1+n_2-2)$，又由样本间的独立性可知 U 与 V 独立. 于是由 t 分布的定义可知

$$t = \frac{U}{\sqrt{V/(n_1+n_2-2)}} = \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}S} \sim t(n_1+n_2-2).$$

又由上述结论可知

$$F=\frac{V_1/(n_1-1)}{V_2/(n_2-1)}=\frac{S_1^2}{S_2^2}\sim F(n_1-1,\ n_2-1).$$

例 6.3.2　设 X_1，X_2，…，X_n 为来自标准正态分布 $N(0,\ 1)$ 的样本，试问下列各统计量服从什么分布？

(1) $\dfrac{X_1-X_2}{\sqrt{X_3^2+X_4^2}}$　　(2) $\dfrac{\sqrt{n-1}X_1}{\sqrt{\sum\limits_{i=2}^{n}X_i^2}}$　　(3) $\dfrac{\left(\dfrac{n}{4}-1\right)\sum\limits_{i=1}^{4}X_i^2}{\sum\limits_{i=5}^{n}X_i^2}$

解　(1) 因为 $X_i\sim N(0,\ 1)$，$i=1,\ 2,\ \cdots,\ n$ 且它们之间相互独立. 故 $X_1-X_2\sim N(0,\ 2)$，对其标准化得 $\dfrac{X_1-X_2}{\sqrt{2}}\sim N(0,\ 1)$，而 $X_3^2+X_4^2\sim\chi^2(2)$，且两者相互独立. 因此由 t 分布的定义得

$$\frac{X_1-X_2}{\sqrt{X_3^2+X_4^2}}=\frac{\dfrac{X_1-X_2}{\sqrt{2}}}{\sqrt{\dfrac{X_3^2+X_4^2}{2}}}\sim t(2).$$

(2) 由于 $X_1\sim N(0,\ 1)$，及 $\sum\limits_{i=2}^{n}X_i^2\sim\chi^2(n-1)$，故由 t 分布的定义得

$$\frac{\sqrt{n-1}X_1}{\sqrt{\sum\limits_{i=2}^{n}X_i^2}}=\frac{X_1}{\sqrt{\dfrac{\sum\limits_{i=2}^{n}X_i^2}{n-1}}}\sim t(n-1).$$

(3) 由条件可知 $\sum\limits_{i=1}^{4}X_i^2\sim\chi^2(4)$，及 $\sum\limits_{i=5}^{n}X_i^2\sim\chi^2(n-4)$，且两者相互独立，因此由 F 分布定义得

$$\frac{\left(\dfrac{n}{4}-1\right)\sum\limits_{i=1}^{4}X_i^2}{\sum\limits_{i=5}^{n}X_i^2}=\frac{\left(\dfrac{n-4}{4}\right)\sum\limits_{i=1}^{4}X_i^2}{\sum\limits_{i=5}^{n}X_i^2}=\frac{\sum\limits_{i=1}^{4}X_i^2/4}{\sum\limits_{i=5}^{n}X_i^2/(n-4)}\sim F(4,\ n-4).$$

例 6.3.3　若 $t\sim t(n)$，试问 t^2 服从什么分布？

解　由 $t\sim t(n)$，根据 t 分布定义，可知有表达式

$$t=\frac{X}{\sqrt{Y/n}},$$

其中 $X\sim N(0,\ 1)$，$Y\sim\chi^2(n)$，且 X 和 Y 相互独立. 于是

$$t^2=\frac{X^2}{Y/n},$$

而 $X^2\sim\chi^2(1)$，故由 F 分布定义知，$t^2\sim F(1,\ n)$.

例 6.3.4　设某种小麦的产量(公斤/单位面积)服从正态分布 $N(\mu, \sigma^2)$，这里 $\sigma = 10$ 公斤，现在共收割了 25 块试验田，每块试验田均为同一单位面积，用 S^2 表示这 25 块试验田产量的样本方差，试求 $P(S^2 > 51.67)$.

解　由定义 6.2.2 可知

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

于是

$$P(S^2 > 51.67) = P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{51.67(n-1)}{\sigma^2}\right)$$
$$= P\left(\chi^2(24) > \frac{51.67 \times 24}{100}\right) = P(\chi^2(24) > 12.4) = 0.975.$$

习题六

1. 在某总体抽取一个容量为 5 的样本，测得样本值为：98.5，98.3，99，100.6，95.8，求其样本均值和样本方差.

2. 查表计算：

(1) $\chi^2_{0.05}(15)$，$\chi^2_{0.95}(9)$，$\chi^2_{0.025}(3)$；

(2) $t_{0.05}(12)$，$t_{0.025}(2)$，$t_{0.01}(54)$；

(3) $F_{0.05}(12, 15)$，$F_{0.95}(12, 15)$，$F_{0.025}(12, 15)$.

3. 抽水机每天的的停机时间服从正态分布 $N(4, 0.64)$，求

(1) 一个月(30 天)每天的平均停机时间在 1 ~ 5 小时之间的概率.

(2) 一个月(30 天)，总的停机时间小于 115 个小时的概率.

4. 若总体 $X \sim N(\mu, \sigma^2)$，X_1，X_2，…，X_n 为其简单随机样本，$\overline{X}$ 为样本均值，S^2 为样本方差. 试问

(1) 统计量 $U = n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2$ 服从何种分布?

(2) 统计量 $V = n\left(\frac{\overline{X}-\mu}{S}\right)^2$ 服从何种分布?

5. 若 $T \sim t(n)$，$n \geqslant 2$，证明：$E(T) = 0$.

6. 设 X_1，X_2，…，X_{10} 为来自于总体 $N(\mu, \sigma^2)$ 的一个样本，试问

(1) 若 $\mu = 0$，$\sigma = 0.3$，求 $P(\sum_{i=1}^{10} X_i^2 > 1.44)$.

(2) 若 $\sigma = 4$，而 μ 未知，S^2 为样本方差且满足 $P(S^2 > A) = 0.1$，求 A.

7. 设某县农民人均收入(单位：万元)服从正态分布 $N(1.5, 0.25)$，现随机调查了 n 个人，若这 n 个人的人均收入不超过 1.6 万元的概率为 0.9，求至少调查多少人?

第七章　参数估计

总体是由总体分布来刻画的，但在实际问题中，总体分布中的参数往往是未知的. 例如，已知某总体是正态分布类型，但该正态分布的数学期望和方差未知，需要根据样本来估计这些未知参数，这类问题称为参数估计问题. 参数估计问题是数理统计中的基本问题之一，也是统计推断的一种重要形式.

假设总体的 X 分布函数为 $F(x, \theta)$，其中 θ 是未知参数，$X_1, X_2, \cdots, X_n$ 为来自该总体的样本. 若构造统计量 $\hat{\theta}=\hat{\theta}(X_1, X_2, \cdots, X_n)$ 来估计参数 θ，则称 $\hat{\theta}$ 为参数 θ 的估计量. 将样本观测值 $x_1, x_2, \cdots, x_n$ 代入 $\hat{\theta}$ 得到一个具体值 $\hat{\theta}=\hat{\theta}(x_1, x_2, \cdots, x_n)$，该数值称为 θ 的估计值.

参数估计问题主要分为两类：点估计和区间估计. 假如构造一个统计量 $\hat{\theta}(X_1, X_2, \cdots, X_n)$ 作为未知参数 θ 的估计量，那么 $\hat{\theta}$ 就称为参数 θ 的**点估计**；假如构造两个统计量 $\hat{\theta}_1=\hat{\theta}_1(X_1, X_2, \cdots, X_n)$ 和 $\hat{\theta}_2=\hat{\theta}_2(X_1, X_2, \cdots, X_n)$，而用区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 作为未知参数 θ 可能取值范围的一种估计，那么区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 就称为参数 θ 的**区间估计**或称为参数 θ 的**置信区间**. 如果 $x_1, x_2, \cdots, x_n$ 是样本的一组观测值，把它代入上述估计量，就可得到未知参数 θ 的一个确定的估计值 $\hat{\theta}$ 或 θ 的一个确定的估计区间 $[\hat{\theta}_1, \hat{\theta}_2]$.

7.1　参数的点估计

7.1.1　样本数字特征法

既然样本来自总体，样本的特性在一定程度上就反映了总体的特性. 为此，最经常用的点估计方法就是样本数字特征法. 该方法简单、易行、直观，不必考虑总体的分布类型，具有普遍性. 样本数字特征法就是以样本的数字特征作为相应总体数字特征的估计量，常用的估计量有样本均值和样本方差.

1. 以样本均值 $\overline{X}$ 作为总体均值 μ 的点估计量，即

$$\hat{\mu}=\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{7.1.1}$$

相应地以 $\hat{\mu}=\bar{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$ 作为 μ 的点估计值.

2. 以样本方差 S^2 作为总体方差 σ^2 的点估计量，即

$$\hat{\sigma}^2=S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{7.1.2}$$

相应地以 $\hat{\sigma}^2 = s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ 作为 σ^2 的点估计值.

例 7.1.1　随机抽取某村农民 10 人，调查其 20 ×× 年收入(元) 如下：

24 445，27 967，28 265，22 943，25 688，28 761，26 309，25 798，31 587，29 860.

试估计该村农民人均年收入及其方差.

解　该村农民人均年收入为

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(24445 + 27967 + \cdots + 29860) = 27162,$$

该村农民年收入的方差为

$$\begin{aligned}\hat{\sigma}^2 = s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{9}[(24\,445 - 27\,612)^2 + (27\,967 - 27\,612)^2 + \cdots + (29\,860 - 27\,612)^2] \\ &= 6\,802\,677.\end{aligned}$$

例 7.1.2　设 n 次独立试验中事件 A 发生 μ 次，试用样本数字特征法求事件 A 的概率 $P(A)$ 的估计量.

解　构造总体 X 如下：

$$X = \begin{cases} 1, & \text{第 } i \text{ 次试验 } A \text{ 发生}, \\ 0, & \text{第 } i \text{ 次试验 } A \text{ 不发生}. \end{cases}$$

则 $P(X=1) = P(A) = p$，$P(X=0) = P(\bar{A}) = 1-p$，且 $E(X) = p$，令 X_1，X_2，…，X_n 为来自总体 X 的样本，则 p 的估计量为 $\dfrac{1}{n}\sum_{i=1}^{n} X_i = \dfrac{\mu}{n}$.

样本数字特征法用途广泛，不限于均值和方差，也不限于一个总体，比如设两个总体 X 和 Y 的均值分别为未知参数 μ_1 和 μ_2，则样本均值之差 $\bar{X} - \bar{Y}$ 可作为为 $\mu_1 - \mu_2$ 的估计量.

7.1.2　矩估计法

定义 7.1.1　设 X 为随机变量，若 $E|X|^k$ 存在，则称 $E(X^k)$ 为总体 X 的 k **阶原点矩**，记为 $a_k = E(X^k)$；若 $E|X - E(X)|^k$ 存在，则称 $E(X - E(X))^k$ 为总体 X 的 k **阶中心矩**，记为 $b_k = E(X - E(X))^k$.

下面给出与总体矩相对的样本矩的定义.

定义 7.1.2　设 X_1，X_2，…，X_n 为总体 X 的样本，则称 $A_k = \dfrac{1}{n}\sum_{i=1}^{n} X_i^k$ 为 k **阶样本原点矩**，称 $B_k = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$ 为 k **阶样本中心矩**.

矩估计是基于大数定律而提出的，该方法的基本思想是利用 k 阶样本原点矩(或 k 阶样本中心矩) 来估计对应的总体的 k 阶原点矩(或 k 阶中心矩). 由于来自总体 X 的样本 X_1，X_2，…，X_n 是独立同分布的，从而 X_1^k，X_2^k，…，X_n^k 也是独立同分布，因此 $E(X_1^k) = E(X_2^k) = \cdots = E(X_n^k) = a_k$，由大数定律可知，样本原点矩 A_k 作为 X_1^k，X_2^k，…，X_n^k 的算数平

均值依概率收敛到期望值 $a_k = E(X_i^k)$，即当 $n \to +\infty$ 时，

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} a_k.$$

因此对充分大的 n，近似的有 $a_k \approx A_k$，将近似等于号换成等号就是矩估计的表达式，即 $\hat{a}_k = A_k$，$\hat{b}_k = B_k$.

若总体 X 的分布律(或密度函数)中包 m 个未知参数 θ_1，θ_2，$\cdots$，θ_m，则总体 X 的 k 阶矩 $a_k = E(X^k)$ 为 θ_1，θ_2，$\cdots$，θ_m，的函数，即 $a_k = a_k(\theta_1, \cdots, \theta_m)$，以样本矩 A_k 作为总体矩 a_k 的估计，即

$$a_k(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_m) = A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k,\ k = 1, 2, \cdots, m \tag{7.1.3}$$

或者

$$b_k(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_m) = B_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^k,\ k = 1, 2, \cdots, m \tag{7.1.4}$$

这是由 m 个方程构成的方程组，从中可以解出 $\hat{\theta}_1$，$\hat{\theta}_2$，$\cdots$，$\hat{\theta}_m$，它们都是样本的函数，即 $\hat{\theta}_k = \hat{\theta}_k(X_1, X_2, \cdots, X_n)$，$k = 1, 2, \cdots, m$. $\hat{\theta}_k$ 就是参数 θ_k 的估计量，称为**矩估计**. 对一次具体抽取的样本值 x_1，x_2，$\cdots$，x_n，$\hat{\theta}_k(x_1, x_2, \cdots, x_n)$ 称为 θ_k 的**矩估计值**.

例 7.1.3 设总体 X 有数学期望 $E(X) = \mu$，方差 $D(X) = \sigma^2$，它们都是未知参数，设 X_1，X_2，$\cdots$，X_n 为来自其总体 X 的样本. 求 μ 和 σ^2 的矩估计量.

解 这里 $a_1 = E(X) = \mu$，$a_2 = E(X^2) = D(X) + E^2(X) = \sigma^2 + \mu^2$，由矩法估计得

$$\begin{cases} A_1 = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i = \hat{\mu}, \\ A_2 = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i^2 = \hat{\sigma}^2 + \hat{\mu}^2, \end{cases}$$

由此解出 μ 和 σ^2 的矩估计分别为

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}, \tag{7.1.5}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2. \tag{7.1.6}$$

从上例我们看出，不管总体 X 服从什么分布，样本均值 $\overline{X}$ 都是总体 X 的期望 $E(X)$ 的矩估计量，样本的二阶中心矩都是总体 X 的方差 $Var(X)$ 的矩估计量.

例 7.1.4 设 X_1，X_2，$\cdots$，X_n 是来自该正态总体 $N(\mu, \sigma^2)$ 的样本，求 μ 和 σ^2 的矩估计.

解 对于正态总体 $N(\mu, \sigma^2)$，μ 和 σ^2 分别为其均值和方差，故由例 7.1.3 可知 μ 和 σ^2 的矩估计分别为

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X},\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

例 7.1.5 设总体 $X \sim U[0, \theta]$，其中 θ 为未知参数，X_1，X_2，$\cdots$，X_n 为来自该总体的

样本，求 θ 的矩估计量.

解　均匀分布 $U[0,\ \theta]$ 的数学期望和方差分别为

$$\mu = E(X) = \frac{\theta}{2},\ \sigma^2 = D(X) = \frac{\theta^2}{12},$$

此时可求出 θ 的下列两种不同的矩估计.

（1）用样本均值估计总体均值，即 $\hat{\mu} = \overline{X} = \frac{1}{2}\hat{\theta}$，得到 θ 的矩估计量

$$\hat{\theta} = 2\overline{X} = \frac{2}{n}\sum_{i=1}^{n} X_i$$

（2）用样本方差来估计总体方差，即

$$\hat{\theta}^2 = B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{\hat{\theta}^2}{12},$$

得到 θ 的另一个矩估计量

$$\hat{\theta} = \sqrt{12B_2} = \sqrt{\frac{12}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

注：从上例我们看到，矩估计有可能不唯一.

例 7.1.6　设 X_1，X_2，…，X_n 为来自均匀分布 $U[a,\ b]$ 的一个样本，试求 a 和 b 的矩估计.

解　由总体 X 均匀分布的性质可知 $E(X) = \frac{a+b}{2}$，$D(X) = \frac{(b-a)^2}{12}$，由此根据例 7.1.3 的结论得到

$$\begin{cases} \dfrac{a+b}{2} = \overline{X}, \\ \dfrac{(b-a)^2}{12} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2, \end{cases}$$

解此方程组可得 a 和 b 的矩估计

$$\hat{a} = \overline{X} - \sqrt{3}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

$$\hat{b} = \overline{X} + \sqrt{3}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

7.1.3　最大似然法

设连续型总体 X 的密度函数为 $f(x;\ \theta)$，θ 是未知参数，则样本 X_1，X_2，…，X_n 的联合密度函数为

$$f^*(x_1,\ x_2,\ \cdots,\ x_n;\ \theta)\prod_{i=1}^{n} f(x_i;\ \theta) \tag{7.1.7}$$

对于样本的一组观察值 x_1，x_2，…，x_n，它是 θ 的函数，记为

$$L(\theta)=L(x_1,\ x_2,\ \cdots,\ x_n;\ \theta)=\prod_{i=1}^{n}f(x_i;\ \theta) \tag{7.1.8}$$

称 $L(\theta)$ 为 θ 的**似然函数**.

最大似然法就是选取这样一个统计量 $\hat{\theta}$ 作为参数 θ 的估计值，它使得样本落在观察值 $(x_1,\ x_2,\ \cdots,\ x_n)$ 的邻域里的概率达到最大，由于该概率与 $\prod_{i=1}^{n}f(x_i;\ \theta)$ 成正比，因此对固定的 $(x_1,\ x_2,\ \cdots,\ x_n)$，可选取 $\hat{\theta}$ 使 $L(\theta)$ 达到最大.

定义 7.1.3 对固定的样本值 $(x_1,\ x_2,\ \cdots,\ x_n)$，若存在 $\hat{\theta}(x_1,\ x_2,\ \cdots,\ x_n)$ 使得

$$L(x_1,\ x_2,\ \cdots,\ x_n;\ \hat{\theta})=\max L(x_1,\ x_2,\ \cdots,\ x_n;\ \theta) \tag{7.1.9}$$

则称 $\hat{\theta}$ 是**参数 θ 的最大似然估计值**，相应地 $\hat{\theta}(x_1,\ x_2,\ \cdots,\ x_n)$ 是**参数 θ 的最大似然估计量**.

因为对数函数 $\ln(x)$ 是单调函数，所以当 $L(\theta)$ 达到最大时，$\ln L(\theta)$ 同时也达到最大. 若 $L(\theta)$ 是可微函数，则最大似然估计可以通过求函数最大值点的微积分法来求 $\ln L(\theta)$ 的最大值点. 具体步骤如下.

(1) 求似然函数 $L(\theta)=\prod_{i=1}^{n}f(x_i;\ \theta)$ 并对其进行化简；

(2) 对 $L(\theta)$ 计算其对数得到对数似然函数 $\ln L(\theta)$；

(3) 由极值的必要条件，求似然方程 $\dfrac{\mathrm{d}\ln L(\theta)}{\mathrm{d}\theta}=0$ 的解 $\theta=\hat{\theta}(x_1,\ x_2,\ \cdots,\ x_n)$；

(4) 将解 $\hat{\theta}(x_1,\ x_2,\ \cdots,\ x_n)$ 中 $x_1,\ x_2,\ \cdots,\ x_n$ 换成 $X_1,\ X_2,\ \cdots,\ X_n$，则 $\hat{\theta}=\hat{\theta}(X_1,\ X_2,\ \cdots,\ X_n)$ 就是 θ 的最大似然估计量.

一般来说，需要进一步用微分法判断上述方法得到的估计量是否是 θ 的最大似然估计量. 但上述方法得到估计量在相当广泛的情况下就是参数 θ 的最大似然估计量. 另外，求最大似然估计必须知道总体分布形式，否则将无法进行.

如果总体是离散型的，上述方法同样适用，此时只需用分布律 $p(x;\ \theta)$ 替换密度函数 $f(x;\ \theta)$. 如果总体分布中含有 k 个未知参数 $\theta_1,\ \theta_2,\ \cdots,\ \theta_k$，那么似然函数就是这些未知参数的函数，记为 $L(x_1,\ x_2,\ \cdots,\ x_n;\ \theta_1,\ \theta_2,\ \cdots,\ \theta_k)$，此时只需求出似然方程组 $\dfrac{\partial \ln L(x_1,\ x_2,\ \cdots,\ x_n;\ \theta_1,\ \theta_2,\ \cdots,\ \theta_k)}{\partial \theta_i}=0,\ i=1,\ 2,\ \cdots,\ k$ 的解，即可得到 $\theta_1,\ \theta_2,\ \cdots,\ \theta_k$ 的最大似然估计量 $\hat{\theta}_1,\ \hat{\theta}_2,\ \cdots,\ \hat{\theta}_k$.

例 7.1.7 求种子发芽率 p 的最大似然估计.

解 设 $X_1,\ X_2,\ \cdots,\ X_n$ 为来自总体 $X\sim B(1,\ p)$ 的一个样本，则 $X_i\sim B(1,\ p)$，似然函数为

$$L(p)=\prod_{i=1}^{n}f(x_i,\ p)=p^{\sum\limits_{i=1}^{n}x_i}(1-p)^{n-\sum\limits_{i=1}^{n}x_i},$$

对数似然函数为

$$\ln L(p)=\sum_{i=1}^{n}x_i\ln p+\left(n-\sum_{i=1}^{n}x_i\right)\ln(1-p),$$

解方程

$$\frac{\mathrm{d}\ln L(p)}{\mathrm{d}p}=\frac{\sum_{i=1}^{n}x_i}{p}-\frac{\left(n-\sum_{i=1}^{n}x_i\right)}{1-p}=0,$$

得到 $p=\bar{x}$，于是 $\hat{p}=\overline{X}$ 就是种子发芽率 p 的最大似然估计量.

例 7.1.8　设总体 X 服从指数分布，其密度函数为 $f(x;\lambda)=\begin{cases}\lambda \mathrm{e}^{-\lambda x}, & x>0,\\ 0, & x\leqslant 0,\end{cases}$ 其中 $\lambda>0$，求 λ 的最大似然估计量.

解　设 X_1，X_2，…，X_n 为总体 X 的一个样本，x_1，x_2，…，x_n 为对应的样本值. 似然函数为

$$L(\lambda)=\prod_{i=1}^{n}f(x_i;\lambda)=\lambda^n \mathrm{e}^{-\lambda\sum_{i=1}^{n}x_i},$$

对数似然函数为

$$\ln L(\lambda)=n\ln\lambda-\lambda\sum_{i=1}^{n}x_i,$$

解方程

$$\frac{\mathrm{d}\ln L(\lambda)}{\mathrm{d}\lambda}=\frac{n}{\lambda}-\sum_{i=1}^{n}x_i=0,$$

得到解 $\lambda=\dfrac{n}{\sum_{i=1}^{n}x_i}=\dfrac{1}{\bar{x}}$，故 λ 的最大似然估计量为 $\hat{\lambda}=\dfrac{1}{X}$.

例 7.1.9　设 $X\sim N(\mu,\sigma^2)$，求参数 μ，σ^2 的最大似然估计量.

解　设 X_1，X_2，…，X_n 为来自总体 X 的一个样本，x_1，x_2，…，x_n 为对应的样本，似然函数为

$$L(\mu,\sigma^2)=\frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right],$$

对数似然函数为

$$\ln L(\mu,\sigma^2)=-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2,$$

分别对 μ，σ^2 求偏导数，并导数等于 0，得

$$\begin{cases}\dfrac{\partial \ln L(\mu,\sigma^2)}{\partial\mu}=\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0,\\ \dfrac{\partial \ln L(\mu,\sigma^2)}{\partial\sigma^2}=-\dfrac{n}{2}\dfrac{1}{\sigma^2}+\dfrac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0,\end{cases}$$

解出

$$\mu=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x},\quad \sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

相应的最大似然估计量为

$$\hat{\mu}=\overline{X},\ \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

上述最大似然估计量与 μ，σ^2 的矩估计量相同.

注：并非每个最大似然估计量都可以通过上述微积分方法得到，如下例.

例 7.1.10 设 $X\sim U[0,\theta]$，求参数 θ 的最大似然估计量.

解 设 X_1，X_2，…，X_n 为来自总体 X 的一个样本，x_1，x_2，…，x_n 为对应的样本值. 均匀分布 $U[0,\theta]$ 的密度函数为

$$f(x;\theta)=\begin{cases}\dfrac{1}{\theta}, & 0<x<\theta,\\ 0, & \text{其他.}\end{cases}$$

似然函数为

$$L(\theta)=\prod_{i=1}^{n}f(x_i;\theta)=\begin{cases}\dfrac{1}{\theta^n}, & 0<x_i<\theta,\ i=1,2,\cdots,n,\\ 0, & \text{其他,}\end{cases}$$

$$=\begin{cases}\dfrac{1}{\theta^n}, & 0\leqslant\max\{x_1,x_2,\cdots,x_n\}\leqslant\theta,\\ 0, & \text{其他.}\end{cases}$$

对于此似然函数，不能通过求解似然方程 $\dfrac{\mathrm{d}\ln L(\theta)}{\mathrm{d}\theta}=0$ 的方法得到 θ 的最大似然估计量，即微分方法失效. 此时必须通过具体分析的方法求似然函数 $L(\theta)$ 的最大值点. 观察上述似然函数 $L(\theta)$，要使它不等于 0，θ 不能小于 $\max\{x_1,x_2,\cdots,x_n\}$，在此条件下要使 $L(\theta)$ 达到最大，应使 θ 尽可能地小，θ 的最小值为 $\max\{x_1,x_2,\cdots,x_n\}$. 因此 θ 的最大似然估计量为 $\hat{\theta}=\max\{X_1,X_2,\cdots,X_n\}$.

7.2 估计量的优良性准则

在参数点估计的讨论中我们看到，同一个未知参数可能有不同的估计量，这样就存在估计量的筛选问题. 那么，比较估计量的评价标准是什么？本节我们将讨论估计量的三个常用的评价准则，即无偏性、有效性和均方误差准则.

7.2.1 无偏性

定义 7.2.1 设总体分布的未知参数为 θ，$\hat{\theta}(X_1,X_2,\cdots,X_n)$ 是 θ 的一个估计量，若满足

$$E[\hat{\theta}(X_1,X_2,\cdots,X_n)]=\theta \tag{7.2.1}$$

则称 $\hat{\theta}$ 是 θ 的**无偏估计**，否则称 $\hat{\theta}$ 是 θ 的**有偏估计**.

无偏性的意义在于，用无偏估计 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 去估计未知参数 θ 时，虽然存在

随机性，但是其平均值等于未知参数 θ.

定理 7.2.1　设总体 X 的均值为 μ，方差为 σ^2，X_1，X_2，…，X_n 是总体 X 的一个样本，$\overline{X}$ 是样本均值，$S^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2$ 是样本方差，则 $\overline{X}$ 和 S^2 分别是 μ 和 σ^2 的无偏估计.

证明　因为 X_1，X_2，…，X_n 是总体 X 的一个样本，所以 $E(X_i)=\mu$，从而得

$$E(\overline{X})=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}\cdot n\mu=\mu.$$

又因为

$$\sum_{i=1}^{n}(X_i-\overline{X})^2=\sum_{i=1}^{n}X_i^2-2\left(\sum_{i=1}^{n}X_i\right)\overline{X}+n\overline{X}^2=\sum_{i=1}^{n}X_i^2-n\overline{X}^2,$$

且

$$E(X_i^2)=Var(X_i)+[E(X_i)]^2=\mu^2+\sigma^2,$$

$$E(\overline{X}^2)=Var(\overline{X})+[E(\overline{X})]^2=\frac{\sigma^2}{n}+\mu^2,$$

从而有

$$\begin{aligned}E(S^2)&=\frac{1}{n-1}\left[E\left(\sum_{i=1}^{n}X_i^2\right)-nE(\overline{X}^2)\right]\\&=\frac{1}{n-1}[n(\mu^2+\sigma^2)-(\sigma^2+n\mu^2)]=\sigma^2.\end{aligned}$$

由定理 7.2.1 可知，无论总体服从什么分布，只要其均值和方差存在，那么样本均值和样本方差分别是总体数学期望和总体方差的无偏估计.

当未知参数的无偏估计不止一个的时候，需要确定这些无偏估计中孰优孰劣. 因此需要有另外的评判标准在无偏估计中进行选择，有效性就是比较无偏估计优劣的一个标准.

7.2.2　有效性

定义 7.2.2　设总体分布的未知参数为 θ，$\hat{\theta}_1$ 和 $\hat{\theta}_2$ 是 θ 的两个无偏估计量，如果

$$Var(\hat{\theta}_1)\leqslant Var(\hat{\theta}_2)\tag{7.2.2}$$

则称估计量 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效. 若在 θ 的所有无偏估计类中，$\hat{\theta}$ 的方差达到最小，则称 $\hat{\theta}$ 是 θ 的**一致最小方差无偏估计**.

例 7.2.1　设总体 X 的均值为 μ，方差为 σ^2，X_1，X_2，…，X_n 是来自该总体 X 的一个样本. 试证明(1) $\hat{\mu}_1=X_i(i=1,2,\cdots,n)$ 和 $\hat{\mu}_2=\overline{X}$ 均为 μ 的无偏估计量；(2) $\hat{\mu}_2=\overline{X}$ 比 $\hat{\mu}_1=X_i(i=1,2,\cdots,n)$ 有效.

证明　(1) 由于 X_1，X_2，…，X_n 与总体 X 同分布，则

$$E(X_i)=\mu\quad,\ i=1,2,\cdots,n,\ \text{且}\quad E(\overline{X})=\mu,$$

故 $\hat{\mu}_1=X_i(i=1,2,\cdots,n)$ 及 $\hat{\mu}_2=\overline{X}$ 均为 μ 的无偏估计量.

（2）由于 X_1，X_2，…，X_n 独立且与总体 X 同分布，则

$$Var(\bar{X}) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

易见只要 $n>1$，则 $Var(\bar{X}) \leqslant Var(X_i)$，$i=1$，$2$，…，$n$. 所以以 $\hat{\mu}_2 = \bar{X}$ 比 $\hat{\mu}_1 = X_i (i=1, 2, \cdots, n)$ 有效，这表明用全部数据的平均估计总体均值要比只使用局部值更有效.

例 7.2.2　设 X_1，X_2 是总体 X 的一个样本，且 X 的均值为 μ，方差为 σ^2，试比较 $\hat{\mu}_1 = \frac{1}{2}X_1 + \frac{1}{2}X_2$ 与 $\hat{\mu}_1 = \frac{1}{3}X_1 + \frac{2}{3}X_2$ 哪个更有效.

解　计算它们的均值，得

$$E(\hat{\mu}_1) = E\left(\frac{1}{2}X_1 + \frac{1}{2}X_2\right) = E\left(\frac{1}{2}X_1\right) + E\left(\frac{1}{2}X_2\right) = \frac{1}{2}\mu + \frac{1}{2}\mu = \mu,$$

$$E(\hat{\mu}_2) = E\left(\frac{1}{3}X_1 + \frac{2}{3}X_2\right) = E\left(\frac{1}{3}X_1\right) + E\left(\frac{2}{3}X_2\right) = \frac{1}{3}\mu + \frac{2}{3}\mu = \mu,$$

故 $\hat{\mu}_1$ 与 $\hat{\mu}_2$ 都是 μ 的无偏估计. 计算它们的方差，得

$$Var(\hat{\mu}_1) = Var\left(\frac{X_1 + X_2}{2}\right) = \frac{1}{2^2}Var(X_1 + X_2) = \frac{\sigma^2 + \sigma^2}{4} = \frac{\sigma^2}{2},$$

$$Var(\hat{\mu}_2) = Var\left(\frac{1}{3}X_1 + \frac{2}{3}X_2\right) = \frac{1}{3^2}Var(X_1) + \frac{2^2}{3^2}Var(X_2) = \frac{1}{9}\sigma^2 + \frac{4}{9}\sigma^2 = \frac{5\sigma^2}{9},$$

因为 $Var(\hat{\mu}_1) < Var(\hat{\mu}_2)$，故 $\hat{\mu}_1$ 比 $\hat{\mu}_2$ 有效.

*7.2.3　均方误差准则

虽然无偏性是估计的一个优良性质，但是不能简单认为有偏估计就一定是好估计. 在有一些场合下，有偏估计可能比无偏估计更优. 如何对有偏估计进行评价，一般而言，样本量一定的前提下，使用的度量指标总是估计值 $\hat{\theta}$ 与参数值 θ 的距离的函数，常用的就是距离的平方，由于 $\hat{\theta}$ 的随机性，可以对该函数求数学期望，即下面给出的均方误差

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2.$$

均方误差是评价点估计的最一般标准，当然希望均方误差越小越好. 均方误差可以分解为点估计的方差和偏差的平方两部分，事实上有

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 = E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 + 2E\{[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta]\} \\ &= Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2. \end{aligned}$$

由上述结果可知，若估计量 $\hat{\theta}$ 是参数值 θ 的无偏估计，则 $MSE(\hat{\theta}) = Var(\hat{\theta})$，即在无偏估计中，均方误差与方差相等，均方误差评价准则与方差评价准则是一致的. 这也说明了用方差的大小考察无偏估计优劣的正确性. 若估计量 $\hat{\theta}$ 是参数值 θ 的有偏估计，则不仅看

其方差大小，还要看其偏差大小.

例 7.2.3　在例 7.1.10 中，$X \sim U[0, \theta]$，参数 θ 的最大似然估计量为

$$\hat{\theta}_1 = \max\{X_1, X_2, \cdots, X_n\},$$

可以求得 $\hat{\theta}_1$ 的密度函数为

$$f_1(x) = \begin{cases} \dfrac{n}{\theta^n} x^{n-1}, & 0 \leqslant x \leqslant \theta, \\ 0, & \text{其他}, \end{cases}$$

并因此可得 $\hat{\theta}_1$ 的均值 $E(\hat{\theta}_1) = \dfrac{n}{n+1}\theta$ 和方差 $Var(\hat{\theta}_1) = \dfrac{n}{(n+1)^2(n+2)}\theta^2$. 所以 $\hat{\theta}_1$ 不是参数 θ 的的无偏估计. 但容易知道，对其修正后的估计量 $\hat{\theta}_2 = \dfrac{n+1}{n}\hat{\theta}_1$ 是 θ 的无偏估计，而另一个估计量 $\hat{\theta}_3 = \dfrac{n+2}{n+1}\hat{\theta}_1$ 是 θ 的有偏估计. 下面比较 $\hat{\theta}_2$ 和 $\hat{\theta}_3$ 的均方误差.

$$MSE(\hat{\theta}_2) = Var(\hat{\theta}_2) + [E(\hat{\theta}_2) - \theta]^2 = \left(\frac{n+1}{n}\right)^2 Var(\hat{\theta}_1) = \frac{\theta^2}{n(n+2)},$$

$$MSE(\hat{\theta}_3) = Var(\hat{\theta}_3) + [E(\hat{\theta}_3) - \theta]^2 = \left(\frac{n+2}{n+1}\right)^2 Var(\hat{\theta}_1) + \left(\frac{n(n+2)}{(n+1)^2}\theta - \theta\right)^2 = \frac{\theta^2}{(n+1)^2}.$$

显然，当 $n > 1$ 时，有 $MSE(\hat{\theta}_2) > MSE(\hat{\theta}_3)$，这说明有偏估计 $\hat{\theta}_3$ 在均方误差准则下优于无偏估计 $\hat{\theta}_2$.

评判估计量的优良标准除了无偏性、有效性、均方误差准则之外，还有其他的一些标准，如充分性、完备性等，出于篇幅限制，本书就不做一一介绍了.

7.3　区间估计

在参数的点估计中，估计值只是一个数值，其优点是可直接告诉人们未知参数的值大致是多少，其缺点是不能反映出估计的误差范围(精度)，难以判断这个点估计和参数真值之间的差距，故在使用上还有不尽人意之处，为了弥补这个缺陷，参数的区间估计就应运而生了. 例如，要了解某地区的气温情况，一般都会给出该地区的最低气温、最高气温以及平均气温，这类估计称为区间估计，这种估计采用一个区间而不是用一个点来估计未知参数.

定义 7.3.1　设总体的分布中含有一个未知参数 θ，对给定的 $\alpha(0 < \alpha < 1)$，如果由样本 $X_1, X_2, \cdots, X_n$ 确定两个统计量 $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \cdots, X_n)$ 和 $\hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \cdots, X_n)$，使 $P(\hat{\theta}_1 \leqslant \theta \leqslant \hat{\theta}_2) = 1-\alpha$，则称随机区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 为参数 θ 的置信度(或置信水平)为 $1-\alpha$ 的**置信区间**. $\hat{\theta}_1$ 为(双侧)置信下限，$\hat{\theta}_2$ 为(双侧)置信上限.

关于区间估计需要说明以下几点.

(1) 对给定的一个样本 $X_1, X_2, \cdots, X_n$，随机区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 可能包含也可能不包含

未知参数 θ，这里置信度 $1-\alpha$ 的含义是指在大量使用该置信区间时，至少有大致 $100(1-\alpha)\%$ 的区间包含参数 θ.

（2）对于不同的置信水平，得到的置信区间是不相同的. 求置信区间时，α 是事先给定的，一般在应用上取 $\alpha=0.05$，这时置信水平为 $1-\alpha=95\%$，或取 $\alpha=0.01$，这时置信水平为 $1-\alpha=99\%$.

（3）置信区间的长度越短，估计精度越高；置信水平 $1-\alpha$ 越高，估计的可靠性越大.

（4）对于固定的样本容量 n，我们不可能同时做到置信区间短且可靠程度高，因为当区间长度变小时，置信度会随之变低. 如果不降低可靠度，又要缩短置信区间长度，则只有加大样本容量 n.

7.3.1 单个正态总体的区间估计

1. 正态总体均值 μ 的置信区间

（1）$\sigma^2=\sigma_0^2$ 已知时 μ 的置信区间

设总体 $X\sim N(\mu,\ {\sigma_0}^2)$，$X_1$，$X_2$，…，$X_n$ 是来自该总体的样本，$\overline{X}$ 是样本均值，则 $\overline{X}\sim N(\mu,\ \sigma^2/n)$，因此可构造下列标准正态分布

$$U=\frac{\overline{X}-\mu}{\sigma_0/\sqrt{n}}\sim N(0,\ 1).$$

对给定的 α，查表得双侧分位数 $u_{\frac{\alpha}{2}}$ 满足

$$P\{\,|U|<u_{\alpha/2}\}=P\left\{\left|\frac{\overline{X}-\mu}{\sigma_0/\sqrt{n}}\right|\leqslant u_{\alpha/2}\right\}=1-\alpha,$$

即等价地有

$$P\left\{\overline{X}-u_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\leqslant\mu\leqslant\overline{X}+u_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right\}=1-\alpha.$$

因此得 μ 的置信度为 $1-\alpha$ 的置信区间

$$\left[\overline{X}-u_{\alpha/2}\frac{\sigma_0}{\sqrt{n}},\ \overline{X}+u_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right]\tag{7.3.1}$$

（2）σ^2 未知时 μ 的置信区间

设总体 $X\sim N(\mu,\ \sigma^2)$，X_1，X_2，…，X_n 是来自总体 X 的样本，$\overline{X}$ 是样本均值，S^2 是样本方差，此时可构造下列 t 分布

$$T=\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1).$$

对于给定的 α，查表得双侧 t 分位数 $t_{\alpha/2}(n-1)$，使得

$$P\{\,|T|\leqslant t_{\alpha/2}(n-1)\}=P\left\{\left|\frac{\overline{X}-\mu}{S/\sqrt{n}}\right|\leqslant t_{\alpha/2}(n-1)\right\}=1-\alpha,$$

即等价地有

$$P\left\{\overline{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\leqslant\mu\leqslant\overline{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right\}=1-\alpha.$$

因此 μ 的置信度为 $1-\alpha$ 的置信区间为

$$\left[\bar{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\ \bar{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right] \tag{7.3.2}$$

例 7.3.1　经验表明，60 日龄的雄鼠体重服从正态分布，且标准差 $\sigma=2.1$ 克. 今从经过 X 射线照射处理过的 60 日龄的雄鼠中随机抽取 16 只测其体重，得数据如下（单位：克）.

20.3，21.5，22.0，19.8，22.5，23.7，25.4，24.0，
23.2，26.8，18.7，21.9，24.4，22.8，26.2，21.4.

求经过 X 射线照射处理过的 60 日龄的雄鼠的体重均值 μ 的置信水平为 95% 的置信区间.

解　计算得 $\bar{x}=22.\ 806\ 3$，$\alpha=0.05$，$u_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}=1.96\dfrac{2.1}{\sqrt{16}}=1.029$，因此由式(7.3.1)式得到 μ 的置信水平为 95% 的置信区间为(21.777，23.835)，即经过 X 射线照射处理过的 60 日龄的雄鼠的体重均值在 21.777 ~ 23.835 克之间的可靠度约为 95%.

例 7.3.2　从一批钉子中随机抽取 16 个，测得其重量(单位：千克) 为

2.14，2.10，2.13，2.15，2.12，2.13，2.13，2.10，
2.15，2.12，2.14，2.10，2.13，2.11，2.14，2.11.

假设钉子的长度服从正态分布 $N(\mu,\ \sigma^2)$，且 σ^2 未知，求总体均值 μ 的置信度为 0.90 的置信区间.

解　对于 $\alpha=0.10$，$n=16$，查表得 $t_{\alpha/2}(n-1)=t_{0.05}(15)=1.753$，计算可得

$$\bar{x}=\frac{1}{16}\sum_{i=1}^{16}x_i=2.\ 125,\ s=\sqrt{s^2}=\sqrt{\frac{1}{16}\sum_{i=1}^{n}(x_i-\bar{x})^2}=0.017\ 13,$$

即 $t_{0.05}(15)\cdot\dfrac{s}{\sqrt{n}}=1.753\cdot\dfrac{0.017\ 13}{4}=0.005$，由式(7.3.2) 得均值 μ 的置信度为 0.90 的置信区间为[2.12，2.13].

2. 正态总体方差 σ^2 的置信区间

(1) $\mu=\mu_0$ 已知时 σ^2 的置信区间

设总体 $X\sim N(\mu_0,\ \sigma^2)$，其中 μ_0 已知，X_1，X_2，…，X_n 是来自该总体的样本，此时可构造分布

$$\chi^2=\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\sigma^2}\sim\chi^2(n).$$

对于给定的置信度 $1-\alpha$，查表得双侧分位数 $\chi^2_{1-\alpha/2}(n)$ 和 $\chi^2_{\varepsilon/2}(n)$(见图 7.1)，
使得

$$P\{\chi^2_{1-\alpha/2}(n)\leqslant\chi^2\leqslant\chi^2_{\alpha/2}(n)\}=1-\alpha,$$

等价地有

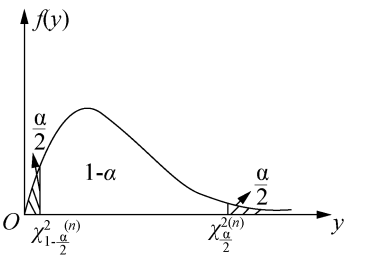


图 7.1　χ^2 分布的分位数

$$P\left\{\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\chi^2_{\alpha/2}(n)}\leqslant\sigma^2\leqslant\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\chi^2_{1-\alpha/2}(n)}\right\}=1-\alpha,$$

故 σ^2 的置信度为 $1-\alpha$ 的置信区间为

$$\left[\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\chi^2_{\alpha/2}(n)},\ \frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\chi^2_{1-\alpha/2}(n)}\right] \tag{7.3.3}$$

(2) μ 未知时 σ^2 的置信区间

设总体 $X\sim N(\mu,\ \sigma^2)$，X_1，X_2，…，X_n 是来自该总体的样本，S^2 是样本方差，此时可构造分布

$$\chi^2=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1).$$

对于给定的置信度 $1-\alpha$，查表得双侧分位数 $\chi^2_{1-\alpha/2}(n-1)$ 和 $\chi^2_{\alpha/2}(n-1)$，使得

$$P\{\chi^2_{1-\alpha/2}(n-1)\leqslant\chi^2\leqslant\chi^2_{\alpha/2}(n-1)\}=1-\alpha,$$

即有

$$P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\leqslant\sigma^2\leqslant\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right\}=1-\alpha.$$

故 σ^2 的置信度为 $1-\alpha$ 的置信区间为

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)},\ \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right] \tag{7.3.4}$$

例 7.3.3 测得一批树苗的直径为(单位：mm)如下：

12.15，12.12，12.01，12.08，12.09，12.16，12.03，12.01，
12.06，12.13，12.07，12.11，12.08，12.01，12.03，12.06.

假设树苗直径服从正态分布 $N(\mu,\ \sigma^2)$，求树苗直径的标准差 σ 的置信度为 0.99 的置信区间.

解 由公式(7.3.4)可知，树苗直径的标准差 σ 的置信度为 $1-\alpha$ 的置信区间为

$$\left[\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}},\ \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}}\right].$$

查表得 $\chi^2_{\alpha/2}(n-1)=\chi^2_{0.005}(15)=32.8$，$\chi^2_{1-\alpha/2}(n)=\chi^2_{0.995}(15)=4.60$，经计算得样本方差为 $S^2=0.002\ 44$，代入上式后得到 σ 的置信度为 0.99 的置信区间

$$\left[\sqrt{\frac{15\times0.002\ 44}{32.8}},\ \sqrt{\frac{15\times0.002\ 44}{4.60}}\right]=[0.033\ 4,\ 0.089\ 2].$$

*7.3.2 两个正态总体的区间估计

在实际应用中，经常会遇到两个正态总体的比较问题. 例如，考察一项新的小麦栽培技术对小麦产量的影响，若将实施新栽培技术的小麦亩产量看成正态总体 $X\sim N(\mu_1,\ \sigma_1^2)$，把原栽培技术的小麦亩产量看成正态总体 $Y\sim N(\mu_2,\ \sigma_2^2)$. 于是评价新栽培技术的效果问题就归

结为研究两个正态总体均值之差 $\mu_1-\mu_2$ 和两个正态总体方差之比 $\dfrac{\sigma_1^2}{\sigma_2^2}$ 的问题.

(一) 两个正态总体均值差的估计

设有两个正态总体 $X\sim N(\mu_1,\ \sigma_1{}^2)$，$Y\sim N(\mu_2,\ \sigma_2^2)$，而 $X_1,\ X_2,\ \cdots,\ X_{n_1}$ 及 $Y_1,\ Y_2,\ \cdots,\ Y_{n_2}$ 分别是从总体 X 和总体 Y 中抽取的两个独立样本，$\overline{X}$，$\overline{Y}$和 S_1^2，S_2^2分别为两个样本的均值和方差，下面求 $\mu_1-\mu_2$ 的 $1-\alpha$ 置信区间.

1. 方差 σ_1^2，σ_2^2 已知时，$\mu_1-\mu_2$ 的区间估计

因为 $\overline{X}\sim N\left(\mu_1,\ \dfrac{\sigma_1^2}{n_1}\right)$，$\overline{Y}\sim N\left(\mu_2,\ \dfrac{\sigma_2^2}{n_2}\right)$，两个样本相互独立，故有

$$\overline{X}-\overline{Y}\sim N\left(\mu_1-\mu_2,\ \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right),$$

从而

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\sim N(0,\ 1) \tag{7.3.5}$$

对于给定的 α，由正态分布函数表查得 $u_{\alpha/2}$，使得

$$P\left\{\frac{|\overline{X}-\overline{Y}-(\mu_1-\mu_2)|}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\leqslant u_{\alpha/2}\right\}=1-\alpha,$$

即

$$P\{\overline{X}-\overline{Y}-u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\leqslant\mu_1-\mu_2\leqslant\overline{X}-\overline{Y}+u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\}=1-\alpha,$$

从而得到 $\mu_1-\mu_2$ 的 $1-\alpha$ 置信区间为

$$\left[\overline{X}-\overline{Y}-u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\ \overline{X}-\overline{Y}+u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right] \tag{7.3.6}$$

2. 方差 $\sigma_1^2=\sigma_2^2=\sigma^2$，$\sigma^2$ 未知时，$\mu_1-\mu_2$ 的区间估计

设

$$\overline{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i,\ S_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2;\qquad \overline{Y}=\frac{1}{n_2}\sum_{i=1}^{n_2}Y_i,\ S_2^2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2$$

$$S_w^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

由抽样分布知

$$T=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2) \tag{7.3.7}$$

对于给定的 $\alpha(0<\alpha<1)$，由 t 分布表查得 $t_{\alpha/2}(n_1+n_2-2)$，使得

$$P\{|T|\leqslant t_{\alpha/2}(n_1+n_2-2)\}=1-\alpha$$

把 T 统计量代入上式，并解其中的不等式得 $\mu_1-\mu_2$ 的 $1-\alpha$ 置信区间为

$$\left[\bar{X}-\bar{Y}-t_{\alpha/2}(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\ t_{\alpha/2}(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right] \tag{7.3.8}$$

例 7.3.4 设有两个独立正态总体 $X\sim N(\mu_1,\ 2^2)$，$Y\sim N(\mu_2,\ 3^2)$，从中各抽取容量为 25 的样本，测得样本均值分别为 $\bar{X}=90$，$\bar{Y}=89$，试求 $\mu_1-\mu_2$ 的置信区间(取 $\alpha=0.05$).

解 由 $\alpha=0.05$，查标准正态分布函数表得 $u_{\alpha/2}=u_{0.025}=1.96$，进一步计算得

$$\bar{X}-\bar{Y}-u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}=90-89-1.96\sqrt{\frac{9}{25}+\frac{16}{25}}=-0.96,$$

$$\bar{X}-\bar{Y}+u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}=90-89+1.96\sqrt{\frac{9}{25}+\frac{16}{25}}=2.96.$$

从而由公式(7.3.6) 得到 $\mu_1-\mu_2$ 的置信区间$[-0.96,\ 2.96]$. 由于该置信区间包含了零，因此可以认为 μ_1 与 μ_2 很接近，两者无显著的差别.

(二) 两个正态总体方差比的置信区间

设有两个正态总体 $X\sim N(\mu_1,\ \sigma_1{}^2)$，$Y\sim N(\mu_2,\ \sigma_2^2)$，而 X_1，X_2，$\cdots$，X_{n_1} 和 Y_1，Y_2，$\cdots$，Y_{n_2} 分别是从总体 X 和总体 Y 中抽取的两个独立样本，$\bar{X}$，$\bar{Y}$和 S_1^2，S_2^2分别为两个样本的均值和方差，下面求$\dfrac{\sigma_1^2}{\sigma_2^2}$ 的置信度为 $1-\alpha$ 置信区间.

由抽样分布知：$\dfrac{(n_1-1)S_1^2}{\sigma_1^2}\sim\chi^2(n_1-1)$，$\dfrac{(n_2-1)S_2^2}{\sigma_2^2}\sim\chi^2(n_2-1)$，又两个样本相互独立，由 F 分布的定义有

$$F=\frac{\dfrac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\dfrac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)}=\frac{\sigma_2^2S_1^2}{\sigma_1^2S_2^2}\sim F(n_1-1,\ n_2-1) \tag{7.3.9}$$

对于给定的 $\alpha(0<\alpha<1)$，查 F 分布表得分位数 $F_{\alpha/2}(n_1-1,\ n_2-1)$，$F_{1-\alpha/2}(n_1-1,\ n_2-1)$ 使得

$$P\left\{F_{1-\alpha/2}(n_1-1,\ n_2-1)\leqslant\frac{\sigma_2^2S_1^2}{\sigma_1^2S_2^2}\leqslant F_{\alpha/2}(n_1-1,\ n_2-1)\right\}=1-\alpha,$$

$$P\left\{\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2}(n_1-1,\ n_2-1)}\leqslant\frac{\sigma_1^2}{\sigma_2^2}\leqslant\frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2}(n_1-1,\ n_2-1)}\right\}=1-\alpha,$$

从而得到$\dfrac{\sigma_1^2}{\sigma_2^2}$ 的 $1-\alpha$ 置信区间为

$$\left[\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2}(n_1-1,\ n_2-1)},\ \frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2}(n_1-1,\ n_2-1)}\right] \tag{7.3.10}$$

例 7.3.5 已知 A，B 两个不同型号的机器重量服从正态分布，测得其样本数据如下(单位：kg).

A：79.98　80.04　80.02　80.04　80.03　80.04　80.03　79.97　80.05　80.03
　　80.02　80.00　80.02

B：80.02　79.94　79.98　79.97　80.03　79.97　79.95　79.97

在均值未知情况下，求两个正态总体方差比的95%置信区间.

解　由题中数据得 $n_1 = 13$，$n_2 = 8$，

$$\bar{x}_1 = \frac{1}{13}(79.98 + 80.04 + \cdots + 80.02) = 80.020\ 77,$$

$$\bar{x}_2 = \frac{1}{13}(80.02 + 79.94 + \cdots + 79.97) = 79.978\ 75,$$

$$S_1^2 = \frac{1}{12}\sum_{i=1}^{13}(x_{1i} - \bar{x}_1)^2 = 0.000\ 57,$$

$$S_2^2 = \frac{1}{7}\sum_{i=1}^{8}(x_{2i} - \bar{x}_2)^2 = 0.000\ 98,$$

由 $\alpha = 0.05$，查表得

$$F_{0.025}(12,\ 7) = 4.67,\ F_{0.975}(12,\ 7) = \frac{1}{F_{0.025}(7,\ 12)} = \frac{1}{3.61} = 0.277,$$

由(7.3.10)式得两个方差比的置信度为95% 的置信区间为$[0.125,\ 2.100]$.

习题七

1. 从某地区小学11岁的男生中随机抽取9人，测得其身高和体重值格式如(身高，体重)，数据如下：

(160，43)，(157，40)，(153，42)，(158，49)，(157，45)
(154，42)，(154，41)，(163，46)，(154，45).

用数字特征法分别对身高 X 和体重 Y 的均值和方差进行估计.

2. 若总体 X 为参数为 λ 的指数分布，X_1，X_2，…，X_n 是总体的一个样本，求未知参数 λ 的一个矩估计量.

3. 若总体 X 为几何分布

$$P(X = k) = (1-p)^{k-1}p,\ \text{其中}\ 0 < p < 1,\ k = 1,\ 2\cdots,$$

X_1，X_2，…，X_n 是总体 X 的一个容量为 n 的样本，求未知参数 p 的矩估计量和最大似然估计量.

4. 若总体 X 为二项分布 $B(N,\ p)$，其中 N 已知，X_1，X_2，…，X_n 是总体 X 的一个容量为 n 的样本，求未知参数 p 的矩估计和最大似然估计量.

5. 设 X_1，X_2，…，X_n 是从总体 X 抽得的一个简单随机样本，总体 X 的概率密度函数为

$$p(x,\ \theta) = \begin{cases} \theta x^{\theta-1}, & 0 < x < 1, \\ 0, & \text{其他}, \end{cases}$$

其中 $\theta > 0$ 未知参数. 试用矩估计和最大似然法估计总体未知参数 θ.

6. 设 X_1，X_2，…，X_n 是从总体 X 中抽得的一个简单随机样本，总体 X 的概率密度函数为

$$p(x,\ \theta)=\begin{cases}\dfrac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}},\ x>0,\ \theta>0,\\ 0,\qquad \text{其他},\end{cases}$$

试用最大似然法估计总体的未知参数 θ.

7. 设总体 X 为正态分布 $N(\mu,\ 1)$，X_1，X_2，X_3 为其一个样本，下列三个统计量中哪些是无偏估计量，并说明无偏估计中哪个方差最小.

(1) $\hat{\mu}_1=\dfrac{1}{5}X_1+\dfrac{3}{10}X_2+\dfrac{1}{2}X_3$;

(2) $\hat{\mu}_2=\dfrac{1}{3}X_1+\dfrac{1}{4}X_2+\dfrac{5}{12}X_3$;

(3) $\hat{\mu}_3=\dfrac{1}{3}X_1+\dfrac{1}{5}X_2+\dfrac{1}{12}X_3$.

8. 设某种水稻的亩产量服从正态分布 $N(\mu,\ \sigma^2)$，随机抽取 9 亩试验田，测得其重量(单位：克) 如下.

510，485，505，505，490，495，520，515，490.

试求均值 μ 置信度为 99% 的置信区间.

9. 设某种电子管的使用寿命服从正态分布，从中随机抽取 16 个进行检验，得平均使用寿命为 1 950 小时，标准差 s = 300 小时，试分别求

(1) 整批电子管平均使用寿命置信度为 95% 的置信区间;

(2) 使用寿命的标准差置信度为 95% 的置信区间.

10. 设总体 $X\sim N(\mu,\ 100)$，若置信度为 95% 时，μ 的置信区间长度为 5，则样本容量 n 至少为多少？若置信度为 99% 时，样本容量 n 至少为多少？

11. 假定某地一旅游者的消费额 X 服从正态分布 $N(\mu,\ \sigma^2)$，且标准差 $\sigma=500$ 元，μ 未知，今要对该地旅游者的平均消费额 μ 加以估计，为了能以 95% 的置信度来相信这种估计绝对误差小于 50 元，问至少要调查多少名游客？

12. 已知某种果树产量服从正态分布 $N(218,\ \sigma^2)$，随机抽取 6 棵计算其产量(单位：千克) 为

221，191，202，205，256，236.

试以 95% 的置信水平估计果树产量的方差.

13. 已知某种木材横纹抗压力的实验值服从正态分布，对 10 个试件作横纹抗压力试验得到数据(单位：$\mathrm{kg/cm^2}$) 如下;

482，493，457，510，446，435，418，394，496，480.

试以 95% 的置信度对该木材横纹抗压力的方差进行区间估计.

14. 比较棉花品种的优劣：假设用甲、乙两种棉花纺出的棉纱强度分别为 $X\sim N(\mu_1,\ 2.18^2)$ 和 $Y\sim N(\mu_2,\ 1.76^2)$. 试验者从这两种棉纱中分别抽取样本 X_1，X_2，…，X_{200} 和 Y_1，Y_2，…，Y_{100}，样本均值分别为：$\overline{X}=5.32$，$\overline{Y}=5.76$，求 $\mu_1-\mu_2$ 的置信系数为 95% 的区间估计.

15. 设有两个正态总体：$X \sim N(\mu_1, {\sigma_1}^2)$，$Y \sim N(\mu_2, \sigma_2^2)$. 分别从 X 和 Y 抽取容量为 $n_1 = 25$ 和 $n_2 = 8$ 的两个样本，并求得 $S_1 = 8$，$S_2 = 7$. 试求两正态总体方差比$\dfrac{{\sigma_1}^2}{\sigma_2^2}$的置信度为 0.98 的置信区间.

16. 某公司利用两条自动化流水线灌装矿泉水. 设这两条流水线所装矿泉水的体积(单位：毫升)$X \sim N(\mu_1, \sigma_1^2)$ 和 $Y \sim N(\mu_2, \sigma_2^2)$. 现从生产线上分别抽取 X_1，X_2，…，X_{12} 和 Y_1，Y_2，…，Y_{17}，其样本均值与样本方差分别为：$\overline{X} = 501.1$，$S_1^2 = 2.4$；$\overline{Y} = 499.7$，$S_2^2 = 4.7$. 求 $\mu_1 - \mu_2$ 的置信系数为 95% 的区间估计.

第八章　假设检验

假设检验是统计推断中不同于参数估计的另一重要内容. 统计假设是指对总体分布的形式或对总体分布中某些参数作某种假定. 这些假定可能是针对实际的观察提出的，也可能是对理论分析提出的. 假设检验就是根据样本提供的信息对所提出的假设作出判断：是接受，还是拒绝. 这种判断通常称为假设检验. 如果统计假设是对总体的某些参数提出的，则称为参数假设检验；如果统计假设是总体分布的形式提出的，则称为非参数假设检验. 本章我们讨论参数假设检验和简单的非参数假设检验问题.

8.1　假设检验的基本概念

首先通过一个例子来引出假设检验的一些基本概念.

例 8.1.1　已知某企业自动化灌装矿泉水，每瓶矿泉水的容量服从正态分布 $N(300, 12^2)$，现在随机抽查 10 瓶矿泉水的水量，测得矿泉水的平均水量为 320 毫升，问该自动化灌装线是否正常工作?

显然不能直接得出灌装线不正常工作的结论. 因为 320 − 300 = 20 毫升只是试验的表面差异，它既可能是统计假设造成的，也可能是由抽样误差所造成. 如何正确地用来自总体的样本来推断总体的特征，是统计假设检验要解决的问题.

一般地，统计假设要针对研究的问题提出“一对”相互对立的假设，其中之一是人们特别关心的、经过精心研究确定下来的，称之为**原假设**(或零假设)，用 H_0 表示；而另一个往往是与前一个对立的假设，称之为**备择假设**(或**对立假设**)，用 H_1 表示. 故例 8.1.1 中的原假设为 H_0：$\mu = 300$，备择假设为 H_1：$\mu \neq 300$. 对总体作出原假设与备择假设后，就要根据数据提供的信息对所提出的假设加以检验，并作出是接受原假设 H_0 还是拒绝 H_0(即接受 H_1) 的判决. 接受原假设就意味着拒绝备择假设，拒绝原假设就意味着接受备择假设. 上述过程我们称为统计假设检验.

由样本对原假设进行判断总是通过一个统计量来完成的，该统计量称为**检验统计量**. 使得原假设被拒绝的样本观测值所在的区域称为该检验的**拒绝域**，记作 W，将 $\overline{W}$ 称为**接受域**. 当拒绝域确定时，也就确定检验统计量的接受域与拒绝域的临界点，这个临界点称为**临界值**. 检验的判断准则也随之定下. 假设检验的基本原则就是在一次试验中小概率事件是不可能发生的，称为“小概率事件原理”. 若数据提供的信息有足够证据否定原假设，则给出拒绝原假设 H_0 的判决；若证据不足以否定原假设 H_0，则给出接受原假设 H_0 的判决. 检验的结果与真实情况可能吻合，也有可能不吻合，因此假设检验根据小概率事件原理得出的判决结果，可能是错误的，不过犯错误的概率一般很小. 因为小概率事件的发生并不能说明事件不会发生，仅仅是发生的概率很小很小罢了. 归纳起来，所犯的错误分为两类：第 Ⅰ 类错误和第 Ⅱ 类错误.

第Ⅰ类错误：原假设 H_0 正确，而由于样本观测值落在了拒绝域，而作出拒绝 H_0 的判决而犯的错误，称为**弃真错误**. 假如犯这类错误的概率为 α，则有

$$P(\text{拒绝 } H_0 \mid H_0 \text{ 为真}) = \alpha.$$

概率 α 也被称为检验的**显著性水平**.

第Ⅱ类错误：原假设 H_0 不正确，而由于样本观测值落在了接受域，而作出接受 H_0 的判决而犯的错误，称为**取伪错误**. 假如犯这类错误的概率为 β，则有

$$P(\text{接受 } H_0 \mid H_0 \text{ 为假}) = \beta.$$

当然，我们希望一个假设检验犯上述两类错误的概率都很小. 但是在样本容量 n 固定的情况下是无法达到两全其美的. 因为理论上已经证明：当 α 在减少时，β 就会增大；反之，当 β 在减少时，α 就会增大. 因此无法实现两者平衡，往往采用“保护零假设的原则”来构造假设检验，即先限制犯第Ⅰ类错误的概率不超过显著性水平 α，并在此条件下考虑如何使犯第Ⅱ类错误的概率 β 尽可能小. 而要使它们同时减少，则一般采用增加样本容量的方法.

假设检验的步骤主要分为下面四步.

（1）根据实际问题，提出原假设 H_0 和备择假设 H_1；

（2）在原假设 H_0 成立条件下，构造不依赖任何未知参数且其分布已知的统计量；

（3）给定显著性检验水平 α，根据备择假设 H_1 的具体情况，确定相应的拒绝域；

（4）从子样观察值计算出统计量的观察值. 如果该值落在拒绝域中，则拒绝原假设 H_0，否则就接受原假设 H_0.

下面用例 8.1.1 中的问题来说明上述步骤的具体实现.

步骤(1) 提出原假设和备择假设：$H_0: \mu = 300 \leftrightarrow H_1: \mu \neq 300$.

步骤(2) 构造检验统计量 $U = \dfrac{\overline{X} - \mu_0}{\sigma_0 / \sqrt{n}}$，在原假设 H_0 成立条件下 $U \sim N(0, 1)$.

步骤(3) 确定拒绝域. 由于 H_0 成立时 $\overline{X}$ 是 μ_0 的无偏估计，因此若 H_0 为真，则检验统计量 U 的绝对值 $|U|$ 不应该太大，所以若 $|U|$ 太大就有理由怀疑 H_0 的真实性，即应该拒绝它. 当原假设 H_0 成立时，$P\left(|U| = \left| \dfrac{\overline{X} - \mu_0}{\sigma_0 / \sqrt{n}} \right| \geqslant u_{\alpha/2} \right) = \alpha$，由小概率事件原理确定的拒绝域为 $|U| = \left| \dfrac{\overline{X} - \mu_0}{\sigma_0 / \sqrt{n}} \right| \geqslant u_{\alpha/2}$.

步骤(4) 根据观测数据是否落入拒绝域做出拒绝或接受原假设 H_0 的判决. 在例 8.1.1 中，由于 $\mu_0 = 300$，$\sigma_0 = 12$，$\bar{x} = 320$，$n = 10$，代入统计量 U 计算得 $U = \dfrac{320 - 310}{12 / \sqrt{10}} = 5.270$.

对于给定的显著性检验水平 $\alpha = 0.05$，查标准正态分布函数值表得 $u_{\alpha/2} = u_{0.025} = 1.96$，由于 $|U| > u_{0.025} = 1.96$，即观察值落在拒绝域中，因此应拒绝原假设 H_0，即认为矿泉水灌装线工作不正常.

8.2　正态总体参数的假设检验

在实际应用中，许多变量都可以近似地用正态总体来刻画，所以关于正态总体参数的

检验会经常遇到，下面讨论正态总体参数的假设检问题.

8.2.1　单个正态总体的假设检验

1. 正态总体 $N(\mu, \sigma^2)$ 的均值 μ 的假设设检验

（1）当方差 $\sigma^2 = \sigma_0^2$ 已知，总体均值 μ 的假设检验（U 检验）

假设原假设为 H_0：$\mu = \mu_0$，设 X_1，X_2，…，X_n 是从正态总体 $N(\mu, \sigma_0{}^2)$ 中抽取的样本. 若 H_0 成立，则样本的均值 $\overline{X}$ 服从正态分布 $N(\mu_0, \sigma_0^2/n)$，于是统计量 $U = \dfrac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}}$ 服从标准正态分布 $N(0, 1)$. 取检验水平为 α，由前面讨论可知有

$$P\left(|U| = \left| \frac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}} \right| \geqslant u_{\alpha/2} \right) = \alpha.$$

若原假设为 H_0：$\mu = \mu_0$，备择假设为 H_1：$\mu \neq \mu_0$，则称为双侧假设经验问题，此时拒绝域为 $\left\{ |U| = \left| \dfrac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}} \right| \geqslant u_{\alpha/2} \right\}$，根据样本值算出 U 的观察值，如果 $|U| \geqslant u_{\alpha/2}$，则拒绝原假设 H_0，如果 $|U| < u_{\alpha/2}$，则接受原假设 H_0.

由于上述检验选取服从标准正态分布的 U 统计量，因此称为 U **检验**，同时该检验的拒绝域在两边，故称为**双侧检验**. 其他假设经验问题，如 H_0：$\mu \geqslant \mu_0$，H_1：$\mu < \mu_0$ 或 H_0：$\mu \leqslant \mu_0$，H_1：$\mu > \mu_0$ 称为单边假设检验. 可以证明，当原假设为 H_0：$\mu = \mu_0$、H_0：$\mu \geqslant \mu_0$ 或 H_0：$\mu \leqslant \mu_0$ 时检验统计量不变，即为 $U = \dfrac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}}$，但由对立假设确定的拒绝域有所不同，下面由表 8.1 给出各种假设检验的检验统计量和拒绝域.

表 8.1　方差已知时单个正态总体均值的检验统计量及拒绝域

原假设	备择假设	检验统计量及其分布	临界值	拒绝域
H_0：$\mu = \mu_0$	H_1：$\mu \neq \mu_0$	$U = \dfrac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0, 1)$	$u_{\alpha/2}$	$\lvert U \rvert \geqslant u_{\alpha/2}$
H_0：$\mu \geqslant \mu_0$	H_1：$\mu < \mu_0$	$U = \dfrac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0, 1)$	u_{α}	$U \leqslant -u_{\alpha}$
H_0：$\mu \leqslant \mu_0$	H_1：$\mu > \mu_0$	$U = \dfrac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0, 1)$	u_{α}	$U \geqslant u_{\alpha}$

例 8.2.1　某切割机在正常工作时，切割每段金属棒的平均长度为 10.5cm，标准差是 0.15cm. 今从一批产品中随机地抽取 15 段进行测量，其结果如下（单位：cm）.

10.4，10.6，10.1，10.4，10.5，10.3，10.3，10.2，
10.9，10.6，10.8，10.5，10.7，10.2，10.7.

试在显著性水平 $\alpha = 0.05$ 下检验该切割机工作是否正常.

解　由题意可知，需检验假设

$$H_0：\mu = 10.5 \leftrightarrow H_1：\mu \neq 10.5,$$

总体的标准差为 $\sigma_0 = 0.15$，故采用 U 检验，检验统计量为 $U = \dfrac{\bar{X} - u_0}{\sigma_0/\sqrt{n}}$，拒绝域为 $|U| \geqslant u_{\alpha/2}$. 由 $\alpha = 0.05$ 查表得临界值 $u_{\alpha/2} = u_{0.025} = 1.96$. 将 $\mu_0 = 10.5$，$n = 15$，$u_{\alpha/2} = 1.96$，$\sigma_0 = 0.15$，$\bar{x} = 10.48$ 代入检验统计量得 $|U| = 0.5164 < 1.96$，所以接受原假设 H_0，即可认为该切割机工作正常.

例 8.2.2　某电子元件要求其平均寿命为 1 000 小时，现随机抽取 25 件，测得其平均寿命为 950 小时. 已知该元件寿命服从 $\sigma = 100$ 小时的正态分布，试在显著性水平 $\alpha = 0.05$ 下检验这批电子元件是否合格.

解　根据题意，电子元件寿命 $X \sim N(\mu, 100^2)$，由于电子元件的寿命应该是越大越好，故要检验的假设为

$$H_0: \mu \geqslant 1\,000 \leftrightarrow H_1: \mu < 1\,000,$$

这个方差已知条件下均值的单边检验问题，根据表 8.1，检验统计量为 $U = \dfrac{\bar{X} - u_0}{\sigma_0/\sqrt{n}}$，拒绝域为 $U \leqslant -u_\alpha$. 由 $\alpha = 0.05$ 查表得临界值 $u_\alpha = u_{0.05} = 1.645$. 检验统计量 U 的观察值为 $U = \dfrac{950 - 1\,000}{100/\sqrt{25}} = -2.5$，它落入拒绝域中，故拒绝原假设 H_0，即认为该批电子元件不合格.

由于在实际应用中存在大量关于比例 p 的检验问题，例如检验某人群中男人所占比例、某产品的次品率等. 下面我们利用正态总体方差已知情况下对其均值的假设检验方法对此类问题作简单的讨论. 由于比例 p 可看作某事件发生的概率，即可看作两点分布 $B(1, p)$ 中参数 p. 假设进行 n 次独立试验，以 X 表示该事件 A 在 n 次独立试验中发生的次数，$p = P(A)$，则 $X \sim B(n, p)$. 类似于 p 的置信区间的讨论，当 n 充分大时，可对比例 p 作假设检验 $H_0: p = p_0 (n \geqslant 30)$，当 H_0 成立时，近似地有

$$U = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}} \sim N(0, 1).$$

因此可用 U 检验法对 H_0 进行检验，我们把比例 p 的检验看成在大样本条件下 U 检验法的一个推广.

例 8.2.3　一名研究者声称他所在地区至少有 80% 的观众对电视剧中播广告表示厌烦，为此随机询问了 120 位观众，有 70 人赞成他的观点，在显著水平为 $\alpha = 0.05$ 下该样本是否支持这位研究者的观点？

解　设 p 为对电视剧中插播广告表示厌烦的观众的比例，则所要检验的假设为

$$H_0: p \geqslant 0.8 \leftrightarrow H_1: p < 0.8.$$

设 X 为 120 位观众中赞成其观点的人数，则 $X \sim B(120, p)$. 选取统计量

$$U = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}},$$

其中 $p_0 = 0.8$. 拒绝域为 $U \leqslant -u_\alpha$，当 $\alpha = 0.05$ 查表得临界值 $u_\alpha = u_{0.05} = 1.645$，统计量 U 的观察值 $U = \dfrac{70 - 120 \times 0.8}{\sqrt{120 \times 0.8 \times 0.2}} \approx -5.933 < -1.645$，故拒绝原假设 H_0，因此在 $\alpha = 0.05$ 水平下不支持该研究者的观点.

（2）当方差 σ^2 未知，总体均值 $\mu=\mu_0$ 的检验（t 检验）

设 X_1，X_2，…，X_n 是从正态总体 $N(\mu,\sigma^2)$ 中抽取的样本，σ^2 未知. 若原假设为 $H_0:\mu=\mu_0$，备择假设为 $H_1:\mu\neq\mu_0$，由于 σ^2 未知，故 U 检验是不能采用的，此时可采用下列 t 检验统计量 T，当 $H_0:\mu=\mu_0$ 成立时，

$$T=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\sim t(n-1) \tag{8.2.1}$$

对于给定的检验水平 α，通过查 t 分布表得临界值 $t_{\alpha/2}(n-1)$，使得 $P\{|T|\geqslant t_{\alpha/2}(n-1)\}=\alpha$，即 $\{|T|\geqslant t_{\alpha/2}(n-1)\}$ 是小概率事件.

根据样本计算 T 的值，如果 $|T|\geqslant t_{\alpha/2}(n-1)$ 时，拒绝原假设 H_0；当 $|t|<t_{\alpha/2}(n-1)$ 时，接受原假设 H_0. 这里采用的统计量为 T 统计量，故称为 t **检验法**.

对于单边检验 $H_0:\mu\geqslant\mu_0$，$H_1:\mu<\mu_0$ 或 $H_0:\mu\leqslant\mu_0$，$H_1:\mu>\mu_0$，依照前面类似的讨论，检验统计量不变，即为 $T=\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}$，拒绝域可根据相应的对立假设来确定，表 8.2 给出了各种假设检验问题的检验统计量和拒绝域.

表 8.2　方差未知时单个正态总体均值的检验统计量及拒绝域验

原假设	备择假设	统计量及其分布	临界值	拒绝域
$H_0:\mu=\mu_0$	$H_1:\mu\neq\mu_0$	$T=\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}\sim t(n-1)$	$t_{\alpha/2}(n-1)$	$\lvert T\rvert\geqslant t_{\alpha/2}(n-1)$
$H_0:\mu\geqslant\mu_0$	$H_1:\mu<\mu_0$	$T=\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}\sim t(n-1)$	$t_{\alpha}(n-1)$	$T\leqslant -t_{\alpha}(n-1)$
$H_0:\mu\leqslant\mu_0$	$H_1:\mu>\mu_0$	$T=\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}\sim t(n-1)$	$t_{\alpha}(n-1)$	$T\geqslant t_{\alpha}(n-1)$

假设检验的结论通常是简单的，就是能否拒绝原假设. 然而，在检验中会出现这样的情况，在一个较大的显著性检验水平（如 $\alpha=0.05$）下得到拒绝原假设的结论，而在一个较小的显著性检验水平（如 $\alpha=0.01$）下得到的却是接受原假设的结论. 也就是说，由于显著性水平的不同，可能得到两种截然相反的结论，这情况的出现会造成判决上的困惑. 事实上，在假设检验中，还有一个与检验水平密切相关的概念，即**检验的** p **值**. 下面我们以例 8.2.1 为例说明 p 值的计算. 在该例中，拒绝域为 $\{U\leqslant -u_\alpha\}$，其概率 $P\{U\leqslant -u_\alpha\}$ 为犯第一类错误的概率. 由观测数据得到 U 的观测值为 $u=\dfrac{950-1\,000}{100/\sqrt{25}}=-2.5$. 若用该观测值 u 代替 $P\{U\leqslant -u_\alpha\}$ 中的临界值 u_α，所得的概率即为检验的 p 值. 因此，例 8.2.1 中检验的 p 值为 $p=P\{U\leqslant -2.5\}\approx 0.007\,3$. 对一般的假设检验问题，我们用检验统计量的观察值代替拒绝域中的临界值后所对应事件在原假设 H_0 成立（或 H_0 中等号成立）条件下的概率即为相应检验的 p 值.

关于 p 值的统计意义，一种简单的解释是，它的大小与原假设是否被接受是一致的，即当 p 值较大时，应接受原假设，否则应拒绝原假设. 一般认为，当 p 值大于 0.05 时应接受原假设，当 p 值小于 0.01 时应拒绝原假设. 因此 p 值的大小可以提供对原假设是否被接

受，以及接受程度的一个较客观的数量依据.

例 8.2.4　已知健康人的红血球直径服从均值为 7.2μm 的正态分布，今在某患者血液中随机测得 9 个红血球的直径如下：

$$7.8,\ 9.0,\ 7.1,\ 7.6,\ 8.5,\ 7.7,\ 7.3,\ 8.1,\ 8.0.$$

试在显著性检验水平 $\alpha = 0.05$ 下判断该患者的红血球与健康人的红血球是否有显著差异.

解　检验的假设为

$$H_0:\ \mu = \mu_0 = 7.2,\ H_1:\ \mu \neq 7.2$$

由于方差未知，故采用 t 检验. 检验统计量为 $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$，拒绝域为 $|T| \geqslant t_{\alpha/2}(n-1)$. 当 $\alpha = 0.05$ 时，查 t 分布表得 $t_{\alpha/2}(n-1) = t_{0.025}(8) = 2.306$. 由观测数据计算得到检验统计量的观察值为

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7.9 - 7.2}{0.346/3} = 6.071,$$

它落在拒绝域内，所以拒绝原假设 H_0，即在检验水平 $\alpha = 0.05$ 下认为该患者的红血球平均值与健康人之间有显著差异.

例 8.2.5　已知某品种花卉株高服从正态分布 $N(52,\ \sigma^2)$，现在改变施肥的配方，并从利用新配方施肥获得的花卉中随机挑选 7 株，测得其株高为

$$52.45,\ 48.51,\ 56.02,\ 49.02,\ 53.38,\ 54.04,\ 53.21$$

问在显著性水平 $\alpha = 0.05$ 下，使用新配方获得的花卉株高是否有明显提高？

解　检验的假设为

$$H_0:\ \mu \leqslant 52,\ H_1:\ \mu > 52$$

由于方差未知，故采用 t 检验. 检验统计量为 $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$，拒绝域为 $T \geqslant t_{\alpha}(n-1)$. 查表得临界值 $t_{\alpha}(n-1) = t_{0.05}(6) = 1.943\ 2$. 由样本计算得到

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 52.38,\ s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = 7.328,$$

由此得统计量的观测值 $t = \dfrac{52.38 - 52}{\sqrt{7.328/7}} = 0.371\ 4$，它未落在拒绝域中，故应接受原假设 H_0，即认为利用新配方获得的花卉株高没有显著提高.

2. 单个正态总体 $N(\mu,\ \sigma^2)$ 的方差 σ^2 的假设检验

（1）均值 μ 已知时，总体方差 σ^2 的检验（χ^2 检验）

设 $X_1,\ X_2,\ \cdots,\ X_n$ 是从正态总体 $N(\mu,\ \sigma^2)$ 中抽取的样本，均值 μ 已知，检验的假设为

$$H_0:\ \sigma^2 = \sigma_0^2,\ H_1:\ \sigma^2 \neq \sigma_0^2,$$

在原假设 H_0 成立条件下，检验统计量及其分布为

$$\chi^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma_0^2} \sim \chi^2(n) \tag{8.2.2}$$

对于给定的检验水平 α，由 $P\{\chi^2 \leqslant \chi^2_{1-\alpha/2}(n)\} = \alpha/2$，及 $P\{\chi^2 \geqslant \chi^2_{\alpha/2}(n)\} = \alpha/2$ 确定临界值 $\chi^2_{1-\alpha/2}(n)$ 和 $\chi^2_{\alpha/2}(n)$. 当 $\chi^2 \leqslant \chi^2_{1-\alpha/2}(n)$ 或 $\chi^2 \geqslant \chi^2_{\alpha/2}(n)$ 时，拒绝原假设 H_0；当 $\chi^2_{1-\alpha/2}(n) < \chi^2 < \chi^2_{\alpha/2}(n)$ 时，接受原假设 H_0.

对于单边检验 $H_0: \sigma^2 \geqslant \sigma_0^2$，$H_1: \sigma^2 < \sigma_0^2$ 或 $H_0: \sigma^2 \leqslant \sigma_0^2$，$H_1: \sigma^2 > \sigma_0^2$，依照前面类似的讨论，检验统计量不变，即为 $\chi^2 = \dfrac{1}{\sigma_0^2}\sum\limits_{i=1}^{n}(X_i - \mu)^2$，检验的拒绝域可根据对立假设来确定. 给定的检验水平 α，由 $P\{\chi^2 \leqslant \chi^2_{1-\alpha}(n)\} = \alpha$（或 $P\{\chi^2 \geqslant \chi^2_{\alpha}(n)\} = \alpha$）可确定临界值 $\chi^2_{1-\alpha}(n)$（或 $\chi^2_{\alpha}(n)$），当 $\chi^2 \leqslant \chi^2_{1-\alpha}(n)$（或 $\chi^2 \geqslant \chi^2_{1-\alpha}(n)$）时拒绝 H_0.

（2）均值 μ 未知时，总体方差 σ^2 的检验（χ^2 检验）

若 X_1，X_2，…，X_n 是来自正态总体 X 的样本，若 $H_0: \sigma^2 = \sigma_0^2$ 成立，则统计量

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(n-1) \tag{8.2.3}$$

若对双边假设检验，则对立假设为 $H_1: \sigma^2 \neq \sigma_0^2$. 此时，对给定的检验水平 α，由 $P\{\chi^2 \leqslant \chi^2_{1-\alpha/2}(n-1)\} = \alpha/2$，及 $P\{\chi^2 \geqslant \chi^2_{\alpha/2}(n-1)\} = \alpha/2$ 确定临界值 $\chi^2_{1-\alpha/2}(n-1)$ 和 $\chi^2_{\alpha/2}(n-1)$. 当 $\chi^2 \leqslant \chi^2_{1-\alpha/2}(n-1)$ 或 $\chi^2 \geqslant \chi^2_{\alpha/2}(n-1)$ 时拒绝原假设 H_0；当 $\chi^2_{1-\alpha/2}(n-1) < \chi^2 < \chi^2_{\alpha/2}(n-1)$ 时接受原假设 H_0.

下面将单个总体下关于方差的 χ^2 检验方法总结在表 8.3 中.

表 8.3　单个正态总体方差的 χ^2 检验

条件	原假设	备择假设	统计量及其分布	临界值	拒绝域
μ 已知	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2 = \dfrac{1}{\sigma_0^2}\sum\limits_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2(n)$	$\chi^2_{1-\alpha/2}(n)$ 和 $\chi^2_{\alpha/2}(n)$	$\chi^2 \leqslant \chi^2_{1-\alpha/2}(n)$ 或 $\chi^2 \geqslant \chi^2_{\alpha/2}(n)$
	$\sigma^2 \geqslant \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\chi^2 = \dfrac{1}{\sigma_0^2}\sum\limits_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2(n)$	$\chi^2_{1-\alpha}(n)$	$\chi^2 \leqslant \chi^2_{1-\alpha}(n)$
	$\sigma^2 \leqslant \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2 = \dfrac{1}{\sigma_0^2}\sum\limits_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2(n)$	$\chi^2_{\alpha}(n)$	$\chi^2 \geqslant \chi^2_{\alpha}(n)$
μ 已知	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$	$\chi^2_{1-\alpha/2}(n-1)$ 和 $\chi^2_{\alpha/2}(n-1)$	$\chi^2 \leqslant \chi^2_{1-\alpha/2}(n-1)$ 或 $\chi^2 \geqslant \chi^2_{\alpha/2}(n-1)$
	$\sigma^2 \geqslant \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$	$\chi^2_{1-\alpha}(n-1)$	$\chi^2 \leqslant \chi^2_{1-\alpha}(n-1)$
	$\sigma^2 \leqslant \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$	$\chi^2_{\alpha}(n-1)$	$\chi^2 \geqslant \chi^2_{\alpha}(n-1)$

例 8.2.6　（续例 8.2.5）已知某品种花卉株高服从正态分布 $N(\mu, \sigma^2)$，现在改变施肥的配方，从利用新配方施肥得到的花卉中随机挑选 7 株，测得其株高为

52.45，48.51，56.02，49.02，53.38，54.04，53.21

问在显著性水平 $\alpha = 0.05$ 下，能否认为该花卉株高的方差为 $\sigma^2 = 2.5^2$？

解　检验的假设为

$$H_0: \sigma^2 = \sigma_0^2 = 2.5^2,\ H_1: \sigma^2 \neq 2.5^2,$$

由于均值未知，应采用自由度为 $n-1$ 的 χ^2 检验. 检验统计量为 $\chi^2=\dfrac{(n-1)S^2}{\sigma_0^2}\sim\chi^2(n-1)$，拒绝域为 $\chi^2\leqslant\chi^2_{1-\alpha/2}(n)$ 或 $\chi^2\geqslant\chi^2_{\alpha/2}(n)$. 当 $\alpha=0.05$，$n=7$ 时，查 χ^2 分布临界值得 $\chi^2_{1-\alpha/2}(n-1)=\chi^2_{0.975}(6)=1.237$，$\chi^2_{\alpha/2}(n-1)=\chi^2_{0.025}(6)=14.449$. 由样本观测值得到统计量的观察值为

$$\chi^2=\frac{(n-1)S^2}{\sigma_0^2}=\frac{(7-1)\times 7.328}{2.5^2}=7.035,$$

它未落入拒绝域中，故应接受原假设，即认为花卉株高的方差是 $\sigma^2=2.5^2$.

例 8.2.7　一个混杂的小麦品种，其株高的标准差为 14 厘米，经提纯后随机抽取 10 株，它们的株高为：90，105，101，95，100，100，101，105，93，97(厘米). 试问经提纯后的群体是否比原群体整齐($\alpha=0.01$)？

解　要检验提纯后的群体是否比原群体整齐，即检验假设

$$H_0:\ \sigma^2\geqslant 14^2,\ H_1:\ \sigma^2<14^2,$$

检验统计量为 $\chi^2=\dfrac{(n-1)S^2}{\sigma_0^2}\sim\chi^2(n-1)$，拒绝域为 $\chi^2\leqslant\chi^2_{1-\alpha}(n-1)$. 对于 $\alpha=0.01$，查表得临界值 $\chi^2_{1-\alpha}(n-1)=\chi^2_{0.99}(9)=2.087\ 9$. 检验统计量的观察值为

$$\chi^2=\frac{(n-1)S^2}{\sigma_0^2}=\frac{9\times 13.73^2}{14^2}=1.126\ 7.$$

因为 $\chi^2<\chi^2_{0.99}(9)$，它落在拒绝域中，故应拒绝原假设 H_0，即认为提纯后的群体比原群体整齐.

*8.2.2　两个正态总体的假设检验

在实际工作中，有时需要对两个正态分布的参数进行比较，本节介绍几种常用的检验方法.

1. 两个总体的方差 σ_X^2，σ_Y^2 已知时，均值的检验

设有两个正态总体 $X\sim N(\mu_1,\ \sigma_2^2)$，$Y\sim N(\mu_2,\ \sigma_2^2)$，分别从两个总体中随机抽取两组相互独立的样本 X_1，X_2，…，X_{n_1} 和 Y_1，Y_2，…，Y_{n_2}. 设 $\overline{X}$、S_1^2 分别为总体 X 的样本均值和样本方差，$\overline{Y}$、S_2^2 分别为总体 Y 的样本均值和样本方差.

若检验双边假设

$$H_0:\ \mu_1=\mu_2\leftrightarrow H_1:\ \mu_1\neq\mu_2,$$

由于

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\sim N(0,\ 1),$$

在原假设 H_0 成立的条件下，检验统计量为

$$U=\frac{(\overline{X}-\overline{Y})}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\tag{8.2.4}$$

其中 n_1、n_2 分别为来自总体 X、Y 的样本容量，在原假设 H_0 成立的条件下 $U \sim N(0,1)$. 对给定的显著性水平 α，查标准正态分布表，得临界值 $u_{\alpha/2}$，使 $P\{|U| \geqslant u_{\alpha/2}\} = \alpha$. 由样本观测值可算得 U 的观察值 U，若满足 $|U| \geqslant u_{\alpha/2}$，则拒绝 H_0，否则接受 H_0.

例 8.2.8　已知甲乙两葡萄品种的含糖量分别服从正态分布 $X \sim N(\mu_1, (5.5)^2)$ 和 $Y \sim N(\mu_2, (5.2)^2)$，从甲品种测得 150 个果穗得平均含糖量 $\bar{X} = 15$，从乙品种测定 100 个果穗得平均含糖量 $\bar{Y} = 13$，试在显著性水平 $a = 0.01$ 下检验这两个品种含糖量的差异是否显著.

解　由于方差 $\sigma_1^{\ 2}$ 和 $\sigma_2^{\ 2}$ 已知，故采用 U 检验方法. 检验的假设为

$$H_0: \mu_1 = \mu_2 \leftrightarrow H_1: \mu_1 \neq \mu_2,$$

检验统计量为式(8.2.4)，绝域为 $|U| \geqslant u_{\alpha/2} = u_{0.005} = 2.58$. 由于

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{5.5^2}{150} + \frac{5.2^2}{100}} = 0.687,$$

计算得到 U 的观察值 $U = \dfrac{\bar{x} - \bar{y}}{0.687} = \dfrac{15 - 13}{0.687} = 2.911$，它落在拒绝域中，故拒绝 H_0，检验结果认为两品种含糖量有显著差异.

若检验单边假设

$$H_0: \mu_1 \geqslant \mu_2 \leftrightarrow H_1: \mu_1 < \mu_2,$$

则检验统计量仍为式(8.2.4)，但此时拒绝域为 $U \leqslant -u_\alpha$. 类似地，若检验单边假设

$$H_0: \mu_1 \leqslant \mu_2 \leftrightarrow H_1: \mu_1 > \mu_2,$$

则检验统计量为式(8.2.4)，拒绝域为 $U \geqslant u_\alpha$.

2. 两个总体的方差 σ_X^2，σ_Y^2 未知，但 $\sigma_X^2 = \sigma_Y^2$ 时，均值的检验

设 $X \sim N(\mu_1, \sigma_1^2)$，$Y \sim N(\mu_1, \sigma_2^2)$，且 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知，分别从两个总体中随机抽取两组相互独立的样本 X_1，X_2，…，X_{n_1} 和 Y_1，Y_2，…，Y_{n_2}，$\bar{X}$ 和 S_1^2 分别为从总体 X 的样本中计算得到的样本均值和样本方差，$\bar{Y}$ 和 S_2^2 分别为从总体 Y 的样本中得到的样本均值和样本方差.

若检验双边假设

$$H_0: \mu_1 = \mu_2 \leftrightarrow H_1: \mu_1 \neq \mu_2,$$

检验统计量为

$$T = \frac{(\bar{X} - \bar{Y})}{\sqrt{\dfrac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{8.2.5}$$

其中 n_1、n_2 分别为从总体 X、Y 中抽取的样本的容量，由定理 6.3.3 可知，当假设 H_0 成立时，$T \sim t(n_1 + n_2 - 2)$. 对给定的检验水平 α 可查自由度为 $n_1 + n_2 - 2$ 的 t 分布，得临界值 $t_{\alpha/2}(n_1 + n_2 - 2)$，使 $P\{|T| \geqslant t_{\alpha/2}(n_1 + n_2 - 2)\} = \alpha$，由样本观测值计算得 T 的观察值 T，若满足 $|T| \geqslant t_{\alpha/2}(n_1 + n_2 - 2)$，则拒绝 H_0，否则接受 H_0.

例 8.2.9　设有种植玉米的甲、乙两个农业试验区，各分为 10 个小区，各小区的面积相同，除甲区各小区增施磷肥外，其他试验条件均相同，两个试验区的玉米产量(单位：公斤)如下.(假设玉米产量服从正态分布，且有相同的方差)

甲区：65　60　62　57　58　63　60　57　60　58

乙区：59　56　56　58　57　57　55　60　57　55

试判别磷肥对玉米产量有无显著性影响($\alpha = 0.05$).

解　检验假设为

$$H_0: \mu_1 = \mu_2,\ H_1: \mu_1 \neq \mu_2,$$

检验统计量为式(8.2.5)，拒绝区域为$|T| \geqslant t_{\alpha/2}(n_1+n_2-2)$. 对于给定的$\alpha = 0.05$，查自由度为$10+10-2=18$的$t$分布表，得$t_{\alpha/2}(n_1+n_2-2) = t_{0.025}(18) = 2.101$. 检验统计量的观测值为

$$T = \frac{(\bar{X}-\bar{Y})}{\sqrt{\dfrac{S_X^2(n_1-1)+S_Y^2(n_2-1)}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} = \frac{60-57}{\sqrt{\dfrac{64+24}{10+10-2}}\sqrt{\dfrac{1}{10}+\dfrac{1}{10}}} = 3.03.$$

它落在拒绝域中，故拒绝原假设H_0，即可认为磷肥对玉米产量有显著性的影响.

若检验单边假设

$$H_0: \mu_1 \geqslant \mu_2 \leftrightarrow H_1: \mu_1 < \mu_2,$$

则检验统计量仍为式(8.2.5)，但此时拒绝域为$T \leqslant -t_\alpha(n_1+n_2-2)$. 类似地，若检验单边假设

$$H_0: \mu_1 \leqslant \mu_2 \leftrightarrow H_1: \mu_1 > \mu_2,$$

则检验统计量为式(8.2.5)，拒绝域为$T \geqslant t_\alpha(n_1+n_2-2)$.

3. 两个总体的方差的假设检验

设有两个正态总体$X \sim N(\mu_1, \sigma_2^2)$，$Y \sim N(\mu_1, \sigma_2^2)$，分别从两总体中随机抽取两组相互独立的样本$X_1, X_2, \cdots, X_{n_1}$和$Y_1, Y_2, \cdots, Y_{n_2}$，设$S_X^2$和$S_Y^2$分别为来自这两个样本的容量分别为$n_1$，$n_2$的样本方差.

若检验假设

$$H_0: \sigma_1^2 = \sigma_2^2 \leftrightarrow H_1: \sigma_1^2 \neq \sigma_2^2,$$

则检验统计量为

$$F = \frac{S_X^2}{S_Y^2} \tag{8.2.6}$$

由定理6.3.3可知，当H_0为真时，$F \sim F(n_1-1, n_2-1)$. 给定的检验水平α，查F分布的临界值$F_{1-\alpha/2}(n_1-1, n_2-1)$和$F_{\alpha/2}(n_1-1, n_2-1)$满足

$$P\{F \leqslant F_{1-\alpha/2}(n_1-1, n_2-1)\} = P\{F \geqslant F_{\alpha/2}(n_1-1, n_2-1)\} = \alpha/2.$$

若$F \leqslant F_{1-\alpha/2}(n_1-1, n_2-1)$或$F \geqslant F_{\alpha/2}(n_1-1, n_2-1)$，则拒绝原假设$H_0$；否则接受原假设$H_0$.

例 8.2.10　对某种羊毛纤维在处理前与处理后分别抽样分析，其含脂率如下.

处理前x_i：0.19　0.18　0.21　0.30　0.41　0.12　0.27

处理后y_i：0.15　0.13　0.07　0.24　0.19　0.06　0.08　0.12

假设处理前后的含脂率都是服从正态分布的，检验处理前后的含脂率的方差是否有显著性差异($\alpha = 0.05$).

解　根据题意，提出假设

$$H_0: \sigma_1^2 = \sigma_2^2 \leftrightarrow H_1: \sigma_1^2 \neq \sigma_2^2,$$

检验统计量为 $F=\dfrac{S_X^2}{S_Y^2}$，拒绝域为 $F\leqslant F_{1-\alpha/2}(n_1-1,\ n_2-1)$ 或 $F\geqslant F_{\alpha/2}(n_1-1,\ n_2-1)$. 对于给定的检验水平 $\alpha=0.05$，查 F 分布表得

$$F_{0.975}(6,\ 7)=\frac{1}{F_{0.025}(7,\ 6)}=\frac{1}{5.70}\approx 0.175,\ F_{0.025}(6,\ 7)=5.12,$$

由样本观测值计算得到 $\dfrac{s_X^2}{s_Y^2}=\dfrac{0.009\ 1}{0.003\ 9}=2.33$，它未落在拒绝域中，故接受原假设 H_0，即在检验水平 $\alpha=0.05$ 下，认为处理前后的含脂率的方差之间无显著差异.

对于单边 F 检验，若检验假设

$$H_0:\ \sigma_1^2\geqslant\sigma_2^2\leftrightarrow H_1:\ \sigma_1^2<\sigma_2^2,$$

则检验统计量仍为式(8.2.6)，但此时拒绝域为 $F\leqslant F_{1-\alpha}(n_1-1,\ n_2-2)$. 类似地，若检验单边假设

$$H_0:\ \sigma_1^2\leqslant\sigma_2^2\leftrightarrow H_1:\ \sigma_1^2>\sigma_2^2,$$

则检验统计量为式(8.2.6)，拒绝域为 $F\geqslant F_{\alpha}(n_1-1,\ n_2-2)$.

习题八

1. 某公司用包装机包装肥料，包装机在正常工作时，装包量 $X\sim N(500,\ 2^2)$(单位：g)，每天开工后，需先检验包装机工作是否正常. 某天开工后，在桩号的肥料中任取 9 袋，其重量如下.

505，499，502，506，498，498，497，510，503.

假设总体标准差 σ 不变，即 $\sigma=2$，试问这天包装机工作是否正常($\alpha=0.05$)？

2. 某批农药的 5 个样品中的含磷量，经测定分别为(%)

3.25，3.27，3.24，3.26，3.24.

设测定值总体服从正态分布，问在 $\alpha=0.01$ 时，能否认为这批农药的含磷量的均值为 3.25？

3. 从过去的资料知道某城市高中男生的身高服从正态分布，平均值为 1.67m，标准差为 $\sigma=0.10$m，现在抽查了 100 名高中男学生，其平均身高为 $\bar{x}=1.69$m，如果标准差没有变化，能否认为现在男生身高上的变化有统计意义($\alpha=0.05$)？

4. 在原木中抽出 100 根，测其直径得到样本的均值 $\bar{x}=11.2$ 厘米，标准差 $s=2.6$ 厘米. 问该批原木的平均直径能否认为不低于 12 厘米($\alpha=0.05$)？

5. 现有种植一批某品种葡萄，该品种一串成熟葡萄的重量服从正态分布 $N(450,\ 70^2)$，现在随机采摘了 9 串葡萄进行称重，数据如下(单位：克).

424，497，582，463，503，430，481，402，389.

问在检验水平 $\alpha=0.05$ 时，能否认为该葡萄的方差未发生变化？

6. 某品种水稻的亩产量服从正态分布 $N(320,\ 40^2)$，现在随机抽取 8 亩该水稻，测得其亩产量分别为

309，321，278，289，367，342，314，338.

能否在检验水平 $\alpha = 0.05$ 时，能否认为该水稻亩产量的方差变小？

7. 某炼铁厂的铁水含碳量 X 在正常情况下服从正态分布，现对工艺进行了某些改进，从中抽取五炉铁水测得含碳量如下.

4.421，4.052，4.357，4.287，4.683.

据此是否可以认为新工艺炼出的铁水含碳量的方差仍为 0.108^2（$\alpha = 0.05$）？

8. 某工厂生产一批产品，质量要求：当次品率 $P \leqslant 0.05$ 时，产品才能出厂. 今从生产出的产品中随机抽查 100 件，发现 8 个次品，试问这批产品是否可以出厂？（$\alpha = 0.05$）

9. 某品种石榴籽的含水率(%) 服从正态分布 $N(15, 5^2)$，现在对此品种进行改良，对改良前后的含水率进行对比，数据如下.

改良前：13.2，13.4，13.6，13.3，13.7，13.1，14.0，13.0；

改良后：14.5，15.5，14.2，13.5，13.8，13.6，14.

假设改良后的方差不变，试问在检验水平 $\alpha = 0.05$ 能否认为改良前后石榴籽的含水率发生的变化有统计意义？

10. 设有两组来自相互独立的不同正态总体 X，Y 的样本观测值为

X：-4.4，4.0，2.0，-4.8；

Y：6.0，1.0，3.2，-0.4.

试着在检验水平 $\alpha = 0.05$ 下检验两个总体的方差是否相同.

第九章　方差分析

在科学实验和生产实践中，人们总是希望通过各种试验来观察各种因素对试验结果的影响，而影响结果的因素往往有多种，影响的大小也不等. 例如，在农业生产中，肥料、土壤、品种等对农作物有不同程度的影响；在化工生产中，原料成份、原料剂量、反应温度、溶液浓度、催化剂、反应时间和压力等因素对产品有不同程度的影响；在市场调查中，不同地区、不同季节对商品需求量的变化也存在不同程度的影响等. 方差分析就是研究一种或多种因素的变化对试验结果的观测值是否有显著影响的一种常用数理统计方法，利用该方法人们可以找出较优的实验条件或生产条件.

人们将要考察的指标称为试验指标，影响指标的条件称为因素. 因素的不同状态称为水平. 在一项试验中，只有一个因素在改变，其他因素控制不变的试验称为单因素试验；如果多于一个因素在改变，则称为多因素试验. 实验过程中，由于处理方式不同或条件不同引起观测值不同的，称作因素效应(或处理效应、条件变异)；由于偶然性因素的干扰或观测误差所导致的误差，称作试验误差. 方差分析主要是将总变异分解为因素效应和试验误差，并对其作出数量分析，比较各种原因在总变异中所占的重要程度，作为统计推断的依据，由此确定进一步的工作方向.

9.1　单因素方差分析

下面从一个实例出发说明单因素方差分析的基本思想.

例 9.1.1　在饲料养鸡增肥的研究中，某研究所提出三种饲料配方：A_1 是以鱼粉为主的饲料，A_2 是以槐树粉为主的饲料，A_3 是以苜蓿粉为主的饲料. 为比较三种饲料的效果，特选 24 只相似的小鸡随机均分为三组，每组各喂一种饲料，60 天后观察它们增加的重量. 试验结果数据如表 9.1 所示，试问不同的饲料对小鸡体重的增加有无显著差异?

表 9.1　小鸡体重数据

饲料 A	体重的增量/g							
A_1	1 073	1 009	1 060	1 001	1 002	1 012	1 009	1 028
A_2	1 107	1 092	990	1 109	1 090	1 074	1 122	1 001
A_3	1 093	1 029	1 080	1 021	1 022	1 032	1 029	1 048

本例中饲料的配方是影响小鸡体重增量的因素，三种不同的配方表明因素处于 3 种状态，称为3种水平，这样的试验称为单因素3水平试验. 由表9.1的数据可知，不仅不同的饲料配方喂养出的小鸡的体重不同，而且同一配方喂养出的小鸡体重也不用. 分析数据波动的原因主要来自两方面.

(1) 同一配方喂养的小鸡，喂养条件大致相同，因此，数据的波动是由于其他随机因

素的干扰引起的. 可以想象，同一配方下小鸡的体重增量应该有一个理论上的均值，而实测增值数据与均值的偏离即为随机误差，一般假定随机误差服从正态分布.

（2）在不同的配方下，体重增值有不同的均值，它导致不同组的元件间体重增值数据不同.

对于一般情况，设试验只有一个因素 A 在变化，其他因素都不变，A 有 r 个水平 A_1，…，A_r，在水平 A_i 下进行 n_i 次独立观测，得到试验指标如表 9.2 所示.

表 9.2　单因素方差分析数据

水平	观测值				总体
A_1	x_{11}	x_{12}	…	x_{1n_1}	$N(\mu_1, \sigma_1^2)$
A_2	x_{21}	x_{22}	…	x_{2n_1}	$N(\mu_2, \sigma_2^2)$
⋮	⋮	⋮	…	⋮	⋮
A_r	x_{r1}	x_{r2}	…	x_{rn_r}	$N(\mu_r, \sigma_r^2)$

其中 x_{ij} 表示在因素 A 的第 i 个水平下的第 j 次试验的试验结果.

9.1.1　数学模型

假设在单因素试验中，因素 A 有 r 个水平，记为 A_1，…，A_r，在水平 A_i 下的总体为 X_i，并假设 $X_i \sim N(\mu_i, \sigma^2)$，$i = 1, 2, \cdots r$，其中 μ_i，σ^2 均未知，且各总体之间相互独立，考虑线性统计模型

$$\begin{cases} x_{ij} = \mu_i + \varepsilon_{ij}, \ i = 1, \cdots, r, \ j = 1, \cdots, n_i \\ \varepsilon_{ij} \sim^{iid} N(0, \sigma^2) \end{cases} \tag{9.1.1}$$

其中 μ_i 是第 i 个总体的均值，ε_{ij} 是相应的随机误差.

因素 A 的 r 个水平的比较归结为这 r 个均值的比较，即检验假设

$$H_0: \mu_1 = \cdots = \mu_r \leftrightarrow H_1: \mu_1, \mu_2, \cdots, \mu_r \text{ 不全相等} \tag{9.1.2}$$

记

$$\mu = \frac{1}{n}\sum_{i=1}^{r} n_i\mu_i, \ n = \sum_{i=1}^{r} n_i, \ \alpha_i = \mu_i - \mu,$$

这里 μ 表示总的期望均值，α_i 为水平 A_i 对指标的效应，容易验证 $\sum_{i=1}^{r} n_i\alpha_i = 0$.

模型式(9.1.1)可等价的写成

$$\begin{cases} x_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \ i = 1, \cdots, r, \ j = 1, \cdots, n_i \\ \varepsilon_{ij} \sim^{iid} N(0, \sigma^2) \\ \sum_{i=1}^{r} n_i\alpha_i = 0 \end{cases} \tag{9.1.3}$$

称式(9.1.3)为单因素方差分析的数学模型，它是一种线性模型.

9.1.2 单因素方差分析表

不难推断，假设式(9.1.2)等价于

$$H_0: \alpha_1 = \cdots = \alpha_r = 0,\ H_1: \alpha_1, \alpha_2, \cdots, \alpha_r \text{ 不全为零} \tag{9.1.4}$$

若 H_0 被拒绝，则说明因素 A 的各个水平之间有显著差异；否则，无明显差异.

引入总离差平方和

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2,\ \bar{x} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}x_{ij},\ n = n_1+n_2+\cdots+n_r,$$

这里 $\bar{x}$ 是数据的总平均值. S_T 能反映全部试验数据之间的差异，因此又称为总变差.

经计算可将 S_T 分解成

$$S_T = S_E + S_A \tag{9.1.5}$$

其中

$$S_E = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i\cdot})^2,\ \bar{x}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i}x_{ij},\ i = 1, 2, \cdots, r,$$

$$S_A = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{x}_{i\cdot}-\bar{x})^2 = \sum_{i=1}^{r}n_i(\bar{x}_{i\cdot}-\bar{x})^2 = \sum_{i=1}^{r}n_i\bar{x}^2{}_{i\cdot} - n\bar{x}^2.$$

上面 S_E 的各项 $(x_{ij}-\bar{x}_{i\cdot})^2$ 表示在水平 A_i 下样本观察值与样本均值的差异，它反映了试验过程中各种随机因素引起的试验误差，称 S_E 为**组内平方和**或**误差平方和**. S_A 的各项 $n_i(\bar{x}_{i\cdot}-\bar{x})^2$ 表示水平 A_i 下的样本均值与数据总平均的差异，这是由水平 A_i 的效应的差异引起的，它是由因素 A 的不同水平所引起的系统误差，故称 S_A 为**组间平方和**或因素 A 的**效应平方和**.

式(9.1.5)表明，总离差平方和 S_T 可分解为两部分，一部分是误差平方和 S_E，它是由随机误差引起的；另一部分是组间平方和 S_A，它是由因素 A 的各水平的差异引起的.

若 H_0 为真，即 $\mu_1 = \mu_2 = \cdots = \mu_r$，则所有样本来自同一正态总体 $N(\mu, \sigma^2)$，即 $x_{ij} \sim N(\mu, \sigma^2)$，$j = 1, 2, \cdots, n_i$，$i = 1, 2, \cdots, r$，且它们之间相互独立. 由定理 6.3.3 可知，此时

$$\frac{S_T}{\sigma^2} \sim \chi^2(n-1),\ \frac{S_i}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i\cdot})^2 \sim \chi^2(n_i-1),\ i = 1, 2, \cdots, r.$$

因为 $S_1, S_2, \cdots, S_r$ 相互独立，由 χ^2 分布的可加性得

$$\frac{S_E}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{r}S_i = \frac{1}{\sigma^2}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i\cdot})^2 \sim \chi^2(n-r).$$

进一步可以证明，S_E 与 S_A 相互独立，且当 H_0 为真时，

$$\frac{S_A}{\sigma^2} \sim \chi^2(r-1).$$

于是，在 H_0 为真的条件下，有

$$F = \frac{S_A/(r-1)}{S_E/(n-r)} \sim F(r-1, n-r) \tag{9.1.6}$$

因此，F 统计量可作为 H_0 的检验统计量. 对于给定的显著性水平 α，查 F 分布表，得临界值 $F_\alpha(r-1,\ n-r)$. 由样本值计算统计量 F 的观测值，若 $F \geqslant F_\alpha(r-1,\ n-r)$，则拒绝 H_0，否则接受 H_0.

上述分析结果可以排成表 9.3 的形式，称为方差分析表.

表 9.3　单因素方差分析表

方差来源	平方和	自由度	均方和	F 值	P 值
因素 A	S_A	$r-1$	$\bar{S}_A = \dfrac{S_A}{r-1}$	$F = \dfrac{\bar{S}_A}{\bar{S}_E}$	p
误差	S_E	$n-r$	$\bar{S}_E = \dfrac{S_E}{n-r}$		
总和	S_T	$n-1$			

9.1.3　应用举例

下面考虑用 R 软件计算方差分析表. 在 R 软件中，可用函数 aov( )计算方差分析表，aov( )的使用方法为

aov(formula，data = NULL，projections = FALSE，qr =TRUE，contrast = NULL，…)

其中 formula 是方差分析的公式，data 是数据框，其他可参见在线帮助.

另外，可用 summary( )列出方差分析表的详细信息.

例 9.1.2　茶叶中的叶酸是 B 族维生素的一种，如今研究 4 种不同的茶叶，分别记作 A_1，A_2，A_3，A_4，它们的叶酸含量数据见表 9.4. 用 *R* 软件计算该例子.

表 9.4　叶酸含量数据

茶叶	叶酸含量										
A_1	7.9	6.2	6.6	8.6	8.9	10.1	9.6	8.3	7.5	9.0	8.4	
A_2	5.7	7.5	9.8	6.1	8.4	7.1	7.6	6.9	8.1	7.5		
A_3	6.4	7.1	7.9	4.5	5.0	4.0	6.0	5.6	5.9	5.7	5.1	6.5
A_4	6.8	7.5	5.0	5.3	6.1	7.4	6.3	5.9	6.4	6.8		

解　先采用数据框的格式输入数据，然后调用 aov( )函数进行方差分析，用 summary( )函数提取方差分析的结果. 具体操作如下.

```
tea<-data.frame(X=c(7.9,6.2,6.6,8.6,8.9,10.1,9.6,8.3,7.5,9.0,8.4,5.7,7.5,9.8,6.1,8.4,7.1,
    7.6,6.9,8.1,7.5,6.4,7.1,7.9,4.5,5.0,4.0,6.0,5.6,5.9,5.7,5.1,6.5,6.8,7.5,5.0,5.3,6.1,
    7.4,6.3,5.9,6.4,6.8),
A=factor(rep(1:4,c(11,10,12,10))))
tea.aov<-aov(X~A,data=tea)
summary(tea.aov)
```

分析结果如下：

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	3	41.46	13.818	11.9	1.13e-05 ***
Residuals	39	45.29	1.161		

将上面结果填在方差分析表中，如表 9.5 所示.

表 9.5 叶酸含量试验的方差分析表

方差来源	平方和	自由度	均方和	F 值	P 值
因素 A	41.46	3	13.818	11.9	0.000 011 3
误差	45.29	39	1.162		
总和	86.75	42			

从上面的数据可以看出，方差分析中检验的 p 值为 1.13e-05<0.05，所以拒绝原假设，即认为四种不同的茶叶的叶酸含量具有显著性差异.

例 9.1.3 小白鼠在接种了 3 种不同菌型的伤寒杆菌后的存活天数如表 9.6 所示. 判断小白鼠被注射 3 种菌型后的平均存活天数有无显著差异?

表 9.6 白鼠试验数据

菌型	存活天数											
1	2	4	3	2	4	7	7	2	2	5	4	
2	5	6	8	5	10	7	12	12	6	6		
3	7	11	6	6	7	9	5	5	10	6	3	10

解 设小白鼠被注射的伤寒杆菌为因素 A，3 种不同的菌型为 3 个水平，接种后的存活天数视作来自 3 个正态分布总体 $N(\mu_i, \sigma^2)$，$i=1, 2, 3$ 的样本观测值.

问题归结为检验假设

$$H_0: \mu_1=\mu_2=\mu_3 \leftrightarrow H_1: \mu_1, \mu_2, \mu_3 \text{ 不全相等.}$$

R 软件计算过程与计算结果如下：

```
mouse <-data.frame(X=c(2,4,3,2,4,7,7,2,2,5,4,5,6,8,5,10,7,12,12,6,6,7,11,6,6,7,9,5,5,10,
      6,3,10),
A=factor(rep(1:3,c(11,10,12))))
mouse.aov<-aov(X~A,data=mouse)
summary(mouse.aov)
```

分析结果如下：

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	2	94.256	47.128	8.4827	0.001202 **
Residuals	30	166.653	5.555		

由上面数据可知，方差分析中检验的 p 值远小于 0.01，应拒绝原假设，即认为小白鼠在接种 3 种不同菌型的伤寒杆菌后的存活天数有显著的差异.

*9.1.4　均值的多重比较

若 F 检验的结果是拒绝 H_0，则说明因素 A 的 r 个水平存在显著差异，也就是 r 个均值存在显著差异. 但是这并不意味着所有均值都存在差异，此时有必要对每对 μ_i 和 μ_j 作一对一的比较，即多重比较.

多重比较的方法很多，以下主要介绍 3 种常见的方法.

1. 多重 t 检验法

此方法其实就是针对每组数据进行的 t 检验，其不同之处是该方法估计方差时利用的是全体数据，因此自由度变大. 具体地说，要比较第 i 组与第 j 组平均数，即检验

$$H_0:\ \mu_i \neq \mu_j,\ i \neq j,\ i,\ j = 1,\ \cdots,\ r.$$

方法是采用两正态总体均值的 t 检验，取检验统计量

$$t_{ij} = \frac{\bar{x}_{i\cdot} - \bar{x}_{j\cdot}}{\sqrt{\bar{S}_E\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}},\ i \neq j,\ i,\ j = 1,\ \cdots,\ r \tag{9.1.7}$$

当 H_0 成立时，$t_{ij} \sim t(n-r)$. 所以当

$$|t_{ij}| > t_{\alpha/2}(n-r) \tag{9.1.8}$$

时，说明 μ_i 和 μ_j 差异显著. 定义相应的 p 值

$$p_{ij} = P\{t(n-r) > |t_{ij}|\} \tag{9.1.9}$$

即服从分布 $t(n-r)$ 的随机变量大于 $|t_{ij}|$ 的概率. 上述方法等价于当 $p_{ij} < \alpha$ 时，μ_i 和 μ_j 的差异显著.

该方法使用方便，但是若因素的水平较多，检验又是同时进行时，多次重复使用 t 检验会增大犯第一类错误的概率，所得到的“有显著差异”的结论不一定可靠.

2. p 值的修正

统计学家们提出了许多更有效的方法来调整 p 值，以便于克服多重 t 检验方法的缺点，默认值是 Holm 方法，具体调整方法的名称和参数见表 9.7.

表 9.7　p 值的调整方法

调整方法	R 软件中的参数
Bonferroni	“bonferroni”
Holm(1979)	“holm”
Hochberg(1988)	“Hochberg”
Hommel(1988)	“hommel”
Berjamini & Hockberg(1995)	“BH”
Berjamini & Yekutieli(2001)	“BY”

3. 均值的多重比较的计算

R 软件中的 pairwise. t. test( ) 函数可以得到多重比较的 p 值，其使用方法如下：

pairwise. t. test(x, g, p. adjust. method = p. adjust. methods, pool. sd = TRUE, …), 其中 x 是响应向量, g 是因子向量, p. adjust. method 是 p 值的调整方法, 其方法由函数 p. adjust( ) 给出, 参数值由表 9.6 所示. 如果 p. adjust. method = "none", 表示 p 值是由式(9.1.7)和式(9.1.9)计算出的, 不作任何调整, 默认值按 Holm 方法("holm")作调整.

例 9.1.4 (续例 9.1.3)由于在例 9.1.3 中 F 检验的结论是拒绝 H_0, 应进一步检验

$$H_0: \mu_i = \mu_j,\ i, j = 1, 2, 3$$

解 首先计算三个因子的均值, 然后利用 pairwise. t. test( ) 函数进行多重 t 检验.

利用 R 软件进行分析, 代码如下:

```
Accach(mouse)
mu<-c(mean(X[A==1]),mean(X[A==2]),mean(X[A==3]));mu
pairwise.t.test(X,A,p.adjust.method = "holm")
```

分析结果如下:

```
3.818182  7.700000  7.083333
Pairwise comparisons using t tests with pooled SD
data:  X and A
     1      2
2  0.0021 -
3  0.0048 0.5458
P value adjustment method: holm
```

这里我们对 p 值作 holm 调整, 在一定程度上可以克服多重 t 检验方法的缺点. 从计算结果可以看出, μ_1 与 μ_2, μ_1 与 μ_3 均有显著差异, 而 μ_2 与 μ_3 没有显著差异. 即小白鼠所接种的三种不同菌型的伤寒杆菌中, 第一种与后两种使得小白鼠的平均存活天数有显著差异, 而后两种差异不显著.

9.1.5 单因素方差齐次性检验

对模型进行方差分析时, 该模型应当具备以下三个条件.

(1) 可加性: 假设模型是线性可加模型, 每个处理效应与随机误差是可以叠加的, 即

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

(2) 独立正态性: 试验误差应当服从正态分布, 而且相互独立.

(3) 方差齐性: 不同水平下的方差是一致的, 即满足假设

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 \tag{9.1.10}$$

对于常用的试验来说, 大都能满足上述三个条件. 对于有些不满足条件的试验, 可以先进行数据变换再进行方差分析.

面对实验结果, 如果对误差的正态性和方差齐性没有把握, 则应进行检验.

1. 误差的正态性检验

误差的正态性检验本质上就是数据的正态性检验. 可以用 W 检验(Shapiro. test( ))方

法做正态性检验.

例 9.1.5　对例 9.1.2 的数据作正态性检验.

解利用 R 软件进行分析，代码如下：

```
attach(tea)
shapiro.test(X[A==1])   #水平 1
shapiro.test(X[A==2])   #水平 2
shapiro.test(X[A==3])   #水平 3
```

分析结果如下：

```
Shapiro-Wilk normality test
data:  X[A == 1]
W = 0.96991, p-value = 0.8857
Shapiro-Wilk normality test
data:  X[A == 2]
W = 0.96442, p-value = 0.8348
Shapiro-Wilk normality test
data:  X[A == 3]
W = 0.9892, p-value = 0.9996
Shapiro-Wilk normality test
data:  X[A == 4]
W = 0.9624, p-value = 0.8129
```

从计算结果可见 p 值均大于 0.05，故例 9.1.2 中的数据在 4 个水平下都是正态的.

2. 方差齐性检验

所谓方差齐性检验，就是检验数据在不同水平下的方差是否相同. 方差齐性的检验方法有 Hartley 检验、Cochran 检验和 Bartlett 检验，最常用的方法是 Bartlett 检验. 当各水平下的数据较多时，令

$$S_i^2=\frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i\cdot})^2,\quad S^2=\frac{1}{n-r}\sum_{i=1}^{r}(n_i-1)S_i^2,$$

$$c=1+\frac{1}{3(r-1)}\left[\sum_{i=1}^{r}(n_i-1)^{-1}-(n-r)^{-1}\right],$$

其中 $n=n_1+n_2+\cdots+n_r$. 若假设式(9.1.10)成立，统计量

$$K=\frac{2.3026}{c}\left[(n-r)\ \ln S^2-\sum_{i=1}^{r}(n_i-1)\ln S_i^2\right] \tag{9.1.11}$$

近似服从自由度为 $r-1$ 的 χ^2 分布. 当

$$K^2>\chi_\alpha^2(r-1)\ \text{或}\ P\{\chi^2>K^2\}<\alpha$$

成立时，拒绝 H_0，即可认为至少有两个水平下的数据的方差不相等；否则，认为数据满足方差齐性的要求.

R 软件中，bartlett. test( )函数提供的是 Bartlett 检验，其使用格式为

```
bartlett.test(x,g,…),
bartlett.test(formula,data,subset,na.action,…),
```

其中 x 是数据向量或数据列表，g 是由因子构成的向量，当 x 是列表时，此项无效，formula 是方差分析的公式，data 是数据框. 其余可参见在线帮助.

例 9.1.6 对例 9.1.2 的数据作 Bartlett 方差齐性检验.

解 利用 R 软件分析代码为

```
bartlett.test(X~A,data=tea)
```

分析结果如下：

```
Bartlett test of homogeneity of variances
data: X by A
Bartlett's K-squared = 1.3837,df = 3, p-value = 0.7094
```

从计算结果可见 p 值大于 0.05，故接受原假设，认为各处理组的数据是等方差的.

*9.2 双因素方差分析

实际问题中，影响试验结果的因素往往有很多个. 例如，在化工生产中，产量既受反应温度的影响，又受反应时间、催化剂等因素的影响. 前面所介绍的单因素方差分析无法解决此类问题. 因此，有必要讨论多因素的方差分析问题. 本节介绍双因素方差分析方法.

例 9.2.1 在一个农业试验中，考虑 4 种不同的种子品种 A_1，A_2，A_3，A_4 和 3 种不同的施肥方法 B_1，B_2，B_3 得到产量数据如表 9.8 所示. 试分析种子与施肥对产量有无显著影响.

表 9.8 农业试验数据

	B_1	B_2	B_3
A_1	325	292	316
A_2	317	310	318
A_3	310	320	318
A_4	330	370	365

该试验有两个因素，其中因素 A(种子)有 4 个水平，因素 B(施肥)有 3 个水平. 下面用双因素方差分析法来解决上述问题.

9.2.1 不考虑交互作用

1. 数学模型

设双因素方差分析中，有 A、B 两因素，因素 A 有 r 个水平 A_1，A_2，…，A_r，因素 B 有 s 个水平 B_1，B_2，…，B_s. 在因素 A、B 的每种水平搭配下相互独立地进行一次试验，试验

结果如表 9.9 所示.

表 9.9　无重复试验的双因素方差分析数据

因素 B / 因素 A	B_1	B_2	…	B_s
A_1	x_{11}	x_{12}	…	x_{1s}
A_2	x_{21}	x_{22}	…	x_{2s}
⋮	⋮	⋮		⋮
A_r	x_{r1}	x_{r2}	…	x_{rs}

其中 x_{ij} 表示$(A_i,\ B_j)$，$i=1,\ 2,\ \cdots,\ r$，$j=1,\ 2,\ \cdots,\ s$ 条件下的试验结果.

类似于单因素的方差分析，假定 x_{ij} 服从具有相同方差的正态分布 $N(\mu_{ij},\ \sigma^2)$，$i=1,\ 2,\ \cdots,\ r$，$j=1,\ 2,\ \cdots,\ s$，且各 x_{ij} 相互独立. 不考虑两因素的交互作用，数据可以分解为

$$\begin{cases} x_{ij}=\mu+\alpha_i+\beta_j+\varepsilon_{ij},\ i=1,\ 2,\ \cdots,\ r,\ j=1,\ 2,\ \cdots,\ s \\ \varepsilon_{ij}\sim N(0,\ \sigma^2),\ \text{且各 } \varepsilon_{ij} \text{ 相互独立} \\ \sum\limits_{i=1}^{r}\alpha_i=0,\ \sum\limits_{j=1}^{s}\beta_j=0 \end{cases} \tag{9.2.1}$$

其中 $\mu=\dfrac{1}{rs}\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{s}\mu_{ij}$ 为总平均，α_i 为因素 A 的第 i 个水平的效应，β_j 为因素 B 的第 j 个水平的效应.

2. 方差分析

由于两因素间没有交互作用，因此，在给定显著性水平 α 下，要检验因素 A 是否有影响，就是要检验假设

$$H_{01}:\ \alpha_1=\alpha_2=\cdots=\alpha_r.$$

同样要判断因素 B 是否有影响，等于检验假设

$$H_{02}:\ \beta_1=\beta_2=\cdots=\beta_r.$$

与单因素试验方差分析类似，检验假设 H_{01} 和 H_{02} 的方法也是建立在平方和的分解上，引入以下记号

$$S_T=S_E+S_A+S_B,$$

其中

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij}-\bar{x})^2,\ \bar{x}=\frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}x_{ij},$$

$$S_A=s\sum_{i=1}^{r}(\bar{x}_{i\cdot}-\bar{x})^2,\ \bar{x}_{i\cdot}=\frac{1}{s}\sum_{j=1}^{s}x_{ij},\ i=1,\ 2,\ \cdots,\ r,$$

$$S_B=r\sum_{j=1}^{s}(\bar{x}_{\cdot j}-\bar{x})^2,\ \bar{x}_{\cdot j}=\frac{1}{r}\sum_{j=1}^{r}x_{ij},\ j=1,\ 2,\ \cdots,\ s,$$

$$S_E=\sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})^2.$$

S_T 称为总平方和，S_A 称为因素 A 的组间平方和，S_B 称为因素 B 的组间平方和，S_A 和 S_B 分别反映了因素 A 及因素 B 不同水平所引起的系统误差，S_E 称为误差平方和，它反映了由于各种随机因素所引起的试验误差.

可以证明，当 H_{01} 成立时，$\frac{S_A}{\sigma^2}$ 与 $\frac{S_E}{\sigma^2}$ 相互独立，且

$$\frac{S_A}{\sigma^2} \sim \chi^2(r-1), \quad \frac{S_E}{\sigma^2} \sim \chi^2[(r-1)(s-1)].$$

于是，当 H_{01} 成立时，

$$F_A = \frac{S_A/(r-1)}{S_E/[(r-1)(s-1)]} \sim F(r-1, (r-1)(s-1))$$

类似地，当 H_{02} 成立时，

$$F_B = \frac{S_B/(s-1)}{S_E/[(r-1)(s-1)]} \sim F(s-1, (r-1)(s-1))$$

于是，对于给定的显著性水平 α，查 F 分布表得临界值 $F_\alpha[(r-1), (r-1)(s-1)]$ 与 $F_\alpha[(s-1), (r-1)(s-1)]$，由样本值计算出 F_A 及 F_B. 若 $F_A > F_\alpha[(r-1), (r-1)(s-1)]$ 则拒绝原假设 H_{01}，否则就接受 H_{01}；若 $F_B > F_\alpha[(s-1), (r-1)(s-1)]$，则拒绝原假设 H_{02}，否则就接受 H_{02}.

与单因素的情况类似，把计算结果汇总在方差分析表中(见表 9.10).

表 9.10 双因素方差分析表

方差来源	平方和	自由度	均方和	F 值	P 值
因素 A	S_A	$r-1$	$\bar{S}_A = \frac{S_A}{r-1}$	$F_A = \frac{\bar{S}_A}{\bar{S}_E}$	P_A
因素 B	S_B	$s-1$	$\bar{S}_B = \frac{S_B}{s-1}$	$F_B = \frac{\bar{S}_B}{\bar{S}_E}$	P_B
误差	S_E	$(r-1)(s-1)$	$\bar{S}_E = \frac{S_E}{(r-1)(s-1)}$		
总和	S_T	$rs-1$			

3. 方差分析表的计算

仍然使用 aov( )函数计算双因素方差分析表 9.10 中的各种统计量.

例 9.2.2 (续例 9.2.1)对例 9.2.1 的数据作双因素方差分析，试确定种子与施肥对产量有无显著影响.

解 输入数据，用 aov( )函数求解. 与单因素方差分析相同，用函数 summary( )得到方差分析表.

分析代码如下：

```
agriculture<-data.frame(
Y=c(325,292,316,317,310,318,310,320,318,330,370,365),
```

```
A=gl(4,3),
B=gl(3,1,12)
)
agriculture.aov<-aov(Y~A+B,data=agriculture)
summary(agriculture.aov)
```

分析结果如下：

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	3	3824.3	1274.75	5.2262	0.04126 *
B	2	162.5	81.25	0.3331	0.72915
Residuals	6	1463.5	243.92		

根据数据结果，说明不同品种(因素 A) 对产量有显著影响，但没有充分理由说明施肥方法(因素 B) 对产量有显著影响.

这一检验结果与实际分析情况有出入. 其原因是该模型遵循着一种假定，即因素 A、B 对指标的效应是可以叠加的，而且认为因素 A 的各水平效应的比较与因素 B 在什么水平无关. 即并没有考虑两因素的各水平组合(A_i，B_j) 的不同给产量带来的影响. 这种影响在很多实际研究中不能忽视，这种影响被称为交互效应. 也就是下面所要介绍的内容.

9.2.2　考虑交互作用

1. 数学模型

设双因素方差分析中，有 A、B 两因素，因素 A 有 r 个水平 A_1，A_2，…，A_r，因素 B 有 s 个水平 B_1，B_2，…，B_s. 每种水平组合(A_i，B_j) 下重复试验 t 次. 记第 k 次的观测值为 X_{ijk}，观测数据如表 9.11 所示.

表 9.11　双因素重复试验数据

因素 A \ 因素 B	B_1	B_2	…	B_s
A_1	$x_{111}x_{112}\cdots x_{11t}$	$x_{121}x_{122}\cdots x_{12t}$	…	$x_{1s1}x_{1s2}\cdots x_{1st}$
A_2	$x_{211}x_{212}\cdots x_{21t}$	$x_{221}x_{222}\cdots x_{22t}$	…	$x_{2s1}x_{2s2}\cdots x_{2st}$
⋮	⋮	⋮		⋮
A_r	$x_{r11}x_{r12}\cdots x_{r1t}$	$x_{r21}x_{r22}\cdots x_{r2t}$	…	$x_{rs1}x_{rs2}\cdots x_{rst}$

假定

$$x_{ijk}\sim N(\mu_{ij},\ \sigma^2),\ i=1,\ 2,\ \cdots,\ r;\ j=1,\ 2,\ \cdots,\ s;\ k=1,\ 2,\ \cdots,\ t,$$

各 x_{ijk} 相互独立. 可以将 x_{ijk} 分解为

$$\begin{cases} x_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk} \\ \varepsilon_{ijk} \sim N(0,\ \sigma^2),\ \text{且各}\ \varepsilon_{ijk}\ \text{相互独立} \\ i = 1,\ 2,\ \cdots,\ r,\ j = 1,\ 2,\ \cdots,\ s,\ k = 1,\ 2,\ \cdots,\ t \end{cases} \tag{9.2.2}$$

其中，α_i 为因素 A 的第 i 个水平的效应，β_j 为因素 B 的第 j 个水平的效应，δ_{ij} 为 A_i 和 B_j 的交互效应，因此有

$$\mu = \frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}\mu_{ij},\ \sum_{i=1}^{r}\alpha_i = 0,\ \sum_{j=1}^{s}\beta_j = 0,\ \sum_{i=1}^{r}\delta_{ij} = \sum_{j=1}^{s}\delta_{ij} = 0$$

2. 方差分析

此时，判断因素 A，B 及交互效应的影响是否显著等价于检验下列假设：

$$H_{01}:\ \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0,$$

$$H_{02}:\ \beta_1 = \beta_2 = \cdots = \beta_r = 0,$$

$$H_{03}:\ \delta_{ij} = 0,\ i = 1,\ 2,\ \cdots,\ r,\ j = 1,\ 2,\ \cdots,\ s$$

这种情况下的方差分析法与前述的方法类似，总离差平方和

$$S_T = S_E + S_A + S_B + S_{A\times B},$$

其中，

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(x_{ijk} - \bar{x})^2,\ \bar{x} = \frac{1}{rst}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}x_{ijk},$$

$$S_A = st\sum_{i=1}^{r}(\bar{x}_{i\cdot\cdot} - \bar{x})^2,\ \bar{x}_{i\cdot\cdot} = \frac{1}{st}\sum_{j=1}^{s}\sum_{k=1}^{t}x_{ijk},\ i = 1,\ 2,\ \cdots,\ r,$$

$$S_B = rt\sum_{j=1}^{s}(\bar{x}_{\cdot j\cdot} - \bar{x})^2,\ \bar{x}_{\cdot j\cdot} = \frac{1}{rt}\sum_{j=1}^{r}\sum_{k=1}^{t}x_{ijk},\ j = 1,\ 2,\ \cdots,\ s,$$

$$S_E = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(x_{ijk} - \bar{x}_{ij\cdot})^2,\ \bar{x}_{ij\cdot} = \frac{1}{t}\sum_{k=1}^{t}x_{ijk},\ i = 1,\ 2,\ \cdots,\ r,\ j = 1,\ 2,\ \cdots,\ s,$$

$$S_{A\times B} = t\sum_{i=1}^{r}\sum_{j=1}^{s}(\bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot} - \bar{x}_{\cdot j\cdot} + \bar{x})^2$$

S_T 称为总离差平方和，S_A 称为因素 A 的组间平方和，S_B 称为因素 B 的组间平方和，S_E 称为误差平方和，$S_{A\times B}$ 称为交互效应平方和.

可以证明，当 H_{01} 成立时，

$$F_A = \frac{S_A/(r-1)}{S_E/[rs(t-1)]} \sim F(r-1,\ rs(t-1));$$

当 H_{02} 成立时，

$$F_B = \frac{S_B/(s-1)}{S_E/[rs(t-1)]} \sim F(s-1,\ rs(t-1));$$

当 H_{03} 成立时，

$$F_{A\times B} = \frac{S_{A\times B}/(r-1)(s-1)}{S_E/[rs(t-1)]} \sim F((r-1)(s-1),\ rs(t-1))$$

分别以 F_A，F_B，$F_{A\times B}$ 作为 H_{01}，H_{02}，H_{03} 的检验统计量，将检验结果列成表 9.12.

表 9.12　有交互效应的双因素方差分析表

方差来源	平方和	自由度	均方和	F 值	P 值
因素 A	S_A	$r-1$	$\bar{S}_A = \dfrac{S_A}{r-1}$	$F_A = \dfrac{\bar{S}_A}{S_E}$	P_A
因素 B	S_B	$s-1$	$\bar{S}_B = \dfrac{S_B}{s-1}$	$F_B = \dfrac{\bar{S}_B}{S_E}$	P_B
交互效应 $A \times B$	$S_{A\times B}$	$(r-1)(s-1)$	$\bar{S}_{A\times B} = \dfrac{S_{A\times B}}{(r-1)(s-1)}$	$F_{A\times B} = \dfrac{\bar{S}_{A\times B}}{S_E}$	$P_{A\times B}$
误差	S_E	$rs(t-1)$	$\bar{S}_E = \dfrac{S_E}{rs(t-1)}$		
总和	S_T	$rst-1$			

例 9.2.3　下表记录了 3 位操作工分别在 4 台不同机器上操作 3 天的日产量：

表 9.13　日产量数据

机器	操作工								
	甲			乙			丙		
A_1	15	15	17	17	19	16	16	18	21
A_2	17	17	17	15	15	15	19	22	22
A_3	15	17	16	18	17	16	18	18	18
A_4	18	20	22	15	16	17	17	17	17

试在显著性水平 $\alpha=0.05$ 下检验

（1）操作工之间有无显著性差异.

（2）机器之间的差异是否显著.

（3）操作工与机器的交互作用是否显著.

解　利用 R 软件进行分析，代码如下：

```
output<-data.frame(
Y=c(15,15,17,17,19,16,16,18,21,17,17,17,15,15,15,19,22,22,15,17,16,18,17,16,18,18,18,18,
    20,22,15,16,17,17,17,17),
A=gl(4,9,36),
B=gl(3,3,36)
)
output.aov <- aov(Y~A+B+A:B,data=output)
anova(output.aov)
```

分析结果如下：

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	3	3.417	1.1389	0.6833	0.5709261

B	2	31.056	15.5278	9.3167	0.0010128 **
A:B	6	67.833	11.3056	6.7833	0.0002675 ***
Residuals	24	40.000	1.6667		

可见在显著性水平 $\alpha=0.05$ 下：

(1) 操作工(因素 B)效应是高度显著的.

(2) 机器(因素 A)效应并不显著.

(3) 操作工与机器的交互作用是极为显著的.

9.2.3 双因素方差齐性检验

类似于单因素方差分析，双因素方差分析的数据也应该满足正态性和方差齐性的要求.

例 9.2.4 检验例 9.2.3 中的数据对于因素 A 和 B 是否是正态的，是否满足方差齐性的要求?

解仍然采用 W 正态检验方法检验数据的正态性，用 Bartlett 检验检验方差齐性.

首先利用 R 软件进行正态性经验，代码如下：

```
attach(output)
shapiro.test(Y[A==1])   #因素 A 水平 1
shapiro.test(Y[A==2])   #因素 A 水平 2
shapiro.test(Y[A==3])   #因素 A 水平 3
shapiro.test(Y[A==4])   #因素 A 水平 4
shapiro.test(Y[B==1])   #因素 B 水平 1
shapiro.test(Y[B==2])   #因素 B 水平 2
shapiro.test(Y[B==3])   #因素 B 水平 3
```

分析结果如下：

```
Shapiro-Wilk normality test
data:  Y[A == 1]
W = 0.9165, p-value = 0.3637
Shapiro-Wilk normality test
data:  Y[A == 2]
W = 0.8322, p-value = 0.04732
Shapiro-Wilk normality test
data:  Y[A == 3]
W = 0.8445, p-value = 0.06476
Shapiro-Wilk normality test
data:  Y[A == 4]
W = 0.8761, p-value = 0.1428
Shapiro-Wilk normality test
```

```
data:  Y[B == 1]
W = 0.8479, p-value = 0.03459
Shapiro-Wilk normality test
data:  Y[B == 2]
W = 0.886, p-value = 0.1045
Shapiro-Wilk normality test
data:  Y[B == 3]
W = 0.8567, p-value = 0.04451
```

下面进行 Bartlett 方差齐性检验：

分析代码如下：

```
bartlett.test(Y~A,data=output)
bartlett.test(Y~B,data=output)
```

分析结果如下：

```
        Bartlett test of homogeneity of variances
data: Y by A
Bartlett's K-squared = 5.6717, df = 3, p-value = 0.1287
        Bartlett test of homogeneity of variances
data: Y by B
Bartlett's K-squared = 2.5768, df = 2, p-value = 0.2757
```

从分析结果看，因素 A 和因素 B 基本上都满足正态性要求和方差齐性要求.

习题九

1. 考察温度对某一化工产品得率的影响，选了 5 种不同的温度，在同一温度下做了 3 次实验，测得其得率如下表 1 所示，试分析温度对得率有无显著影响.

温度对得率试验数据

温度	60	65	70	75	80
得率	90	91	96	84	84
	92	93	96	83	89
	88	92	93	83	82

2. 有 4 种 $A_i(i=1, 2, 3)$ 产品，分别为国内甲、乙、丙三个工厂生产的产品，A_4 为国外同类产品. 现从各厂分别取 10，6，6 和 2 个产品做 300h 连续磨损老化试验，得变化率如下表所示. 假定各厂产品试验变化率服从等方差的正态分布.

磨损老化试验数据

产品	变化率									
A_1	20	18	19	17	15	16	13	18	22	17
A_1	26	19	26	28	23	25				
A_1	24	25	18	22	27	24				
A_1	12	14								

（1）试问 4 个厂生产的产品的变化率是否有显著差异？

（2）若有差异，请作出进一步的检验. ①国内产品与国外产品有无显著差异？②国内各厂家的产品有无显著差异？

3. 利用 4 种不同配方的材料 $A_i(i=1, 2, 3, 4)$ 生产出来的元件，测得其使用寿命如下表所示. 问 4 种不同配方下元件的使用寿命有无显著差异？

元件寿命数据

材料	使用寿命							
A_1	1 600	1 610	1 650	1 680	1 700	1 700	1 780	
A_2	1 500	1 640	1 400	1 700	1 750			
A_3	1 640	1 550	1 600	1 620	1 640	1 600	1 740	1 800
A_1	1 510	1 520	1 530	1 570	1 640	1 600		

4. 为了比较属同一类的 4 种不同食谱的营养效果，将 25 只老鼠随机分为 4 组，每组分别是 8 只、4 只、7 只和 6 只，各采用食谱甲、乙、丙、丁喂养. 假设其他条件均保持相同，12 周后测得体重增加量如下表所示. 对于 A_1，检验各食谱的营养效果是否有显著差异.

12 周后 25 只老鼠的体重增加量

食谱	体重增加值							
A_1	164	190	203	205	206	214	228	257
A_2	185	197	201	231				
A_3	187	212	215	220	248	265	281	
A_1	202	204	207	227	230	276		

5. 24 只小鼠按不同窝别分为 8 个区组，再把每个区组中的观察单位随机分配到 3 种不同的饲料组，喂养一定时间后，测得小鼠肝中铁含量，结果如下表所示. 试分析不同饲料的小鼠肝中的铁含量是否不同.（单位：μg/g）

不同饲料组小鼠肝脏中铁含量

窝别	1	2	3	4	5	6	7	8
饲料 A	1.00	1.01	1.13	1.14	1.70	2.01	2.23	2.63
饲料 B	0.96	1.23	1.54	1.96	2.94	3.68	5.59	6.96
饲料 C	2.07	3.72	4.50	4.90	6.00	6.84	8.23	10.33

6. 为了研究不同的田间管理方法对草莓产量的影响，选择6块不同的地块，每个地块分成3个小区，随机安排3种田间管理方法，试验结果见下表.

试验结果

管理 B / 地块 A	A_1	A_1	A_1
A_1	71	73	77
A_1	90	90	92
A_1	59	70	80
A_1	75	80	82
A_1	65	60	67
A_1	82	86	85

试分析不同的管理方法和不同的地块对草莓产量的影响.

7. 研究树种与地理位置对松树生长的影响，对4个地区的3种同龄松树的直径进行测量得到数据如下表所示. $A_i(i=1，2，3)$ 表示3个不同树种，$B_i(i=1，2，3，4)$ 表示4个不同地区. 对每一种水平组合，进行了5次测量，对此试验结果进行方差分析.

3种同龄松树的直径测量数据

地区 B / 树种 A	B_1			B_2			B_3			B_4		
A_1	23	25	21	20	17	11	16	19	13	20	21	18
	14	15		26	21		16	24		27	24	
A_2	28	30	19	26	24	21	19	18	19	26	26	28
	17	22		25	26		20	25		29	23	
A_3	18	15	23	21	25	12	19	23	22	22	13	12
	18	10		12	22		14	13		22	19	

第十章　回归模型

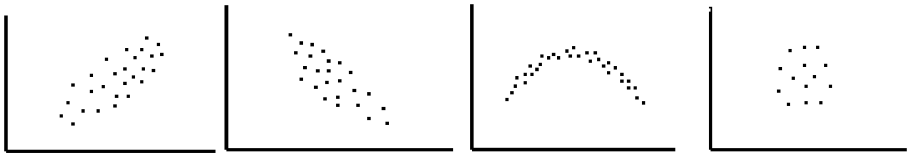
社会经济与自然科学中诸多现象之间始终存在相互制约和相互联系的普遍规律，其中许多问题的研究，往往归结为弄清楚一些变量之间的联系，这种联系一般表现为两种：确定性关系和非确定性关系. 确定性关系就是函数关系 $y=f(x)$，当变量 x（可以是向量）给定之后，y 就会随着函数关系 $y=f(x)$ 唯一确定. 对非确定性关系，变量间的相关关系不能用完全确切的函数形式表示，也就是说变量 y 和 x 有一定的关系，但是没有密切到可以通过 x 唯一决定 y 的程度. 本章的主要内容就是利用回归方法作为工具来研究两个变量或多个变量之间的相关关系.

*10.1　相关分析

10.1.1　散点图

相关分析是研究变量之间是否存在某种相关关系，并对具体有相关关系的现象探讨其相关方向以及相关程度的一种统计方法. 它是回归分析前的重要分析步骤.

散点图是一种识别变量间相关关系的直观工具，就是将变量 X 与 Y 的观察值 $(x_i, y_i)(i=1, 2, \cdots, n)$ 在平面直角坐标系中标出. 散点图通过诸点呈现出的特征，来判断变量之间是否存在相关关系，以及相关的形式、相关的方向和相关的程度等（见图 10.1），为下一步的回归分析做准备. 如果两个变量没有明显的相关关系，则谈不上建立回归模型.



（a）正相关　（b）负相关　（c）非线性相关　（d）零相关

图 10.1　散点图

10.1.2　样本相关系数

散点图虽然有助于直观上识别变量间的相关关系，但它无法对这种关系进行精确的计量. 因此在初步判定变量间存在相关关系的基础上，通常还要计算相关关系的度量指标. 我们已在第四章讨论过用两个总体 X 与 Y 之间的相关系数 $\rho=\dfrac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$ 来度量它

们的线性相关程度，现在的任务是对 n 个独立同分布的观察值$(x_i, y_i)(i = 1, 2, \cdots, n)$，如何对$\rho$作估计. 事实上，用矩法估计可得到总体相关系数的估计，即样本相关系数. 统计量

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

称为随机变量 X 与 Y 的**样本相关系数**，r 的大小能够反映变量 X 与 Y 之间**线性关系的密切程度**.

与ρ一样，样本相关系数的范围在-1到1之间，即$-1 \leqslant r \leqslant 1$；如果变量$Y$与$X$间存在相关关系，则$-1 < r < 1$. 结合散点图分析，我们可得到关于 r 的一些重要结果.

若 $r > 0$，则当 X 增大时，Y 也随之增大，此时称它们之间的相关关系为**正相关**，它表示 X，Y 变化的方向一致(见图 10.1(a)).

若 $r < 0$，则当 X 增大时，Y 随之减小，此时称它们之间的相关关系为**负相关**，它表示 X，Y 变化的方向相反(见图 10.1(b)).

当 $r = 0$ 或 $r \to 0$ 时，n 个点可能毫无规律(见图 10.1(d))，也有可能呈现某种曲线趋势(见图 10.1(c))，此时称 X，Y(线性) 不相关. 特别注意的是，r 反映的关系仅仅是线性相关关系. 因此当$r = 0$或$r \to 0$时，并不一定表示X与Y之间毫无关系，只不过不是线性关系罢了，如图 10.1(c) 显示的情况表明它们之间有密切的抛物线关系.

当 $r = 1$ 或 $r = -1$ 时，n 个点完全在某条直线上，此时称变量 Y 与 X 间存在线性函数关系.

经过计算，我们可得到的样本相关系数的下列简化计算公式.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2\right]\left[\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2\right]}} \quad (10.1.1)$$

10.1.3　相关系数的统计推断

我们通过样本值得到的样本相关系数 r 只不过是总体相关系数ρ的估计值，从同一总体抽取的不同样本会产生不同的样本相关系数，样本相关系数是随机变量存在变异性. 一般地，样本相关系数r的绝对值越大，说明 X 与 Y 的线性相关关系越密切，但到底$|r|$的值要取多大，才算具备线性相关关系，这需要用统计推断来衡量. 另一方面，在实践中，人们常问总体相关系数ρ是否为零，然而，样本相关系数$r \neq 0$未必说明ρ不为零，为此，必须检验

$$H_0: \rho = 0 \leftrightarrow H_1: \rho \neq 0 .$$

为了对相关系数作假设检验，我们首先假设 X 与 Y 服从正态分布，然后找出样本相关系数的分布，从而确定它的拒绝域$\{|r| \geqslant c\}$，c 为临界值. 常用 t 检验方法，可以证明统计

量$\frac{r-\rho}{\sqrt{(1-r^2)/(n-2)}}\sim t(n-2)$. 在假设$H_0$：$\rho=0$成立时，$\frac{r-0}{\sqrt{(1-r^2)/(n-2)}}\sim t(n-2)$，故可用$t$检验法对$H_0$：$\rho=0$作假设检验，以判断$X$与$Y$之间是否存在线性相关关系.

关于样本相关系数应注意下面几点.

（1）样本相关系数r是对变量X与Y的相关关系ρ的估计，它只反映了两个变量间的线性相关关系，而不是反映其他函数关系是否存在.

（2）线性相关关系与因果关系是不同的，相关系数很大，未必表示变量间存在因果关系. 比如，有人对北欧某地的小学生进行一次抽样，测得他们的绘画成绩y_i，同时也测他们的体重x_i，人们发现$(x_i, y_i)(i=1, 2, \cdots, n)$的相关系数超过0.8，表明$x$与$y$呈现高度相关性. 然而，谁也不会得出一个荒谬的结论“想要一个小孩绘画出色吗？你只需要将他喂成一个小胖子就成了”. 其实，它们之所以高度相关，其间涉及一个混杂因素——年龄的影响所造成的，所以研究两个变量相关性时，要注意它们是否同时受第三个变量的影响，避免得出荒唐的结论.

（3）把从逻辑上不存在联系的两个变量放在一起做相关分析没有意义，在统计上称之为“虚假相关”.

（4）相关分析与回归分析在实际应用中有密切关系. 然而在回归分析中，所关心的是一个随机变量Y对另一个（或一组）随机变量X的依赖关系的某种函数形式. 而在相关分析中，所讨论的变量之间地位平等，分析侧重于它们之间的相关特征.

例 10.1.1 今有某种大豆脂肪含量$x(\%)$与蛋白质含量$y(\%)$的检查结果如下.

x	16.5	17.5	18.5	19.5	20.5	21.5	22.5	23.5	24.5
y	43.5	42.6	42.6	40.6	40.3	38.7	37.2	36.0	34.0

试计算X，Y之间的样本相关系数并对其作统计推断.

解 根据表格提供的数据求得

$$\bar{x}=20.5,\ \bar{y}=39.5,\ \sum_{i=1}^{9}x_i^2=3842.3,\ \sum_{i=1}^{9}y_i^2=14128,\ \sum_{i=1}^{9}x_iy_i=7127.3.$$

利用公式(10.1.1)得相关系数

$$r=\frac{\sum_{i=1}^{9}x_iy_i-n\bar{x}\bar{y}}{\sqrt{\left[\sum_{i=1}^{9}x_i^2-n(\bar{x})^2\right]\left[\sum_{i=1}^{9}y_i^2-n(\bar{y})^2\right]}}$$

$$=\frac{7217.3-9\times20.5\times39.5}{\sqrt{[3842.3-9\times20.5^2][14128-9\times39.5^2]}}\approx-0.984.$$

下面对X与Y之间的线性相关性作假设检验，即检验假设

$$H_0:\ \rho=0\leftrightarrow H_1:\ \rho\neq0$$

在分别给定检验水平$\alpha=0.05$与$\alpha=0.01$下，查自由度为$n-2=7$的t分布表，得$t_{0.025}(7)=2.365$，$t_{0.005}(7)=3.409$，而

$$\frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}}=\frac{0.984}{\sqrt{\frac{1-0.984^2}{9-2}}}\approx 14.612>t_{0.005}(7).$$

故拒绝原假设 H_0，即认为 X 与 Y 之间的线性相关性显著.

以上检验方法均拒绝 H_0，即认为在检验水平 $\alpha=0.01$ 下 X 与 Y 之间存在线性相关关系. 因为 α 是犯第一类错误(即弃真错误)的概率，因此在检验水平 $\alpha=0.05$ 下更加倾向于拒绝原假设 H_0.

R 软件的相关分析使用的命令是 cor. test，例 10.1.1 的相关分析代码如下：

```
x<-c(16.5:24.5)
y<-c(43.5,42.6,42.6,40.6,40.3,38.7,37.2,36.0,34.0)
cor.test(x,y)
```

分析结果如下：

```
        Pearson's product-moment correlation
data:  x and y
t = -14.758, df = 7, p-value = 1.57e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.9968126  -0.9245869
sample estimates:
       cor
-0.9843067
```

10.2　一元线性回归分析

回归分析研究的主要对象是客观事物变量间的关系，在对客观事物的大量实验和观察基础上，寻找隐藏在表面上看来是不确定现象之中的统计规律性的统计方法. 回归这个术语是由英国著名统计学家 Galton 在 19 世纪末期研究人类遗传问题时提出来的. Galton 发现身材高的父母，他们的孩子也高. 但这些孩子平均起来并不像他们的父母那样高. 对于比较矮的父母情形也类似，他们的孩子比较矮，但这些孩子的平均身高要高于他们的父母的平均身高. Galton 把这种孩子的身高向中间值靠近的趋势称之为回归效应.

在统计学中，将自变量 X 称为解释变量或协变量，将因变量 Y 称为响应变量. 回归分析中最简单的就是线性回归分析. 按照解释变量个数的不同，分为一元线性回归分析和多元线性回归分析. 为了叙述方便，将把样本的观测值也称为样本，在符号使用上不加以区分，样本或其观察值均用 y_1，y_2，…，y_n 等表示，何时表示定值则由上下文而定.

10.2.1　一元线性回归模型

例 10.2.1　一般来说，人的血压和年龄之间有密切的联系，年龄越大血压通常会越

高. 为了研究血压和年龄的关系，调查 10 个不同年龄人的血压数据如下：

个体编号	1	2	3	4	5	6	7	8	9	10
年龄 x/岁	34	42	47	64	63	46	21	25	39	53
血压 y/mmHg	125	128	160	162	144	142	120	125	120	158

在图 10.2 中看到点$(x_i, y_i)(i = 1, 2, \cdots, 10)$比较接近一条直线，但又不完全在直线上. 导致这些点与直线偏离的原因是在生产过程和试验过程中一些未知的不可控的因素存在，导致实验结果 y_i 在变化.

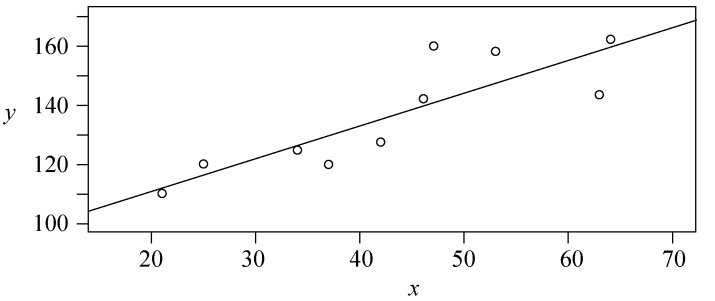


图 10.2　年龄 x 与血压 y 的散点图

现在将试验结果 y_i 看成是由两部分组成的，一部分是由于它与 x_i 的线性关系引起的，记为 $\beta_0 + \beta_1 x_i$；另外一部分则是由随机误差引起的，记作 ε_i. 因此试验结果 y_i 与解释变量 x_i 之间关系可表示为

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{10.2.1}$$

一般地，常常假定误差 ε_i 满足下面三条假设：

(1) 正态性，即

$$\varepsilon_i \sim N(0, \sigma^2),\ i = 1, 2, \cdots, n;$$

(2) 等方差性，即

$$Var(\varepsilon_i) = \sigma^2,\ i = 1, 2, \cdots, n;$$

(3) 独立性，要求不同的试验或观察误差互不相关，即

$$Cov(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j,$$

从而由式(10.2.1)可知

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \tag{10.2.2}$$

一般地，对于可观测的解释变量 x 的每一个取值，对应的响应变量 y 就是一个服从正态分布的随机变量. 将上述结果总结为数学模型如下

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ \varepsilon_i \sim^{i.\ i.\ d} N(0, \sigma^2) \end{cases} \tag{10.2.3}$$

其中 β_0，β_1，σ^2 均为未知参数，符号 $i.\ i.\ d$ 代表独立同分布，式(10.2.3)称为**一元线性回归模型**. 线性函数

$$\mu(x) = E(y) = \beta_0 + \beta_1 x(\text{其中}\beta_1 \neq 0) \tag{10.2.4}$$

称为**回归函数**.

对于一元线性回归模型，主要讨论如下三个问题：

（1）对参数 β_0，β_1 和 σ^2 进行点估计，估计量分别记为 $\hat{\beta}_0$，$\hat{\beta}_1$ 和 $\hat{\sigma}^2$，而 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$ 称为**拟合回归方程**或**经验回归方程**.

（2）在模型(10.2.3)下对 β_0，β_1 作统计推断，并检验 Y 与 X 之间是否线性相关.

（3）利用求得的经验回归方程对 Y 进行预测或控制.

10.2.2　参数估计及其性质

根据观测值 (x_i, y_i)，$i=1, 2, \cdots n$，进行线性回归，目标就是要在无穷条直线 $y=\beta_0+\beta_1 x$ 中寻找一条直线 $y=\hat{\beta}_0+\hat{\beta}_1 x$ 来拟合这些观测值点. 因为要在所有可能的直线中进行挑选，因此首先要确定选择这条直线的标准. 当然标准有很多，结果也不尽相同，这里介绍**最小二乘法**. 最小二乘法是统计学中估计未知参数的一种简单且常用的方法，“二乘”在古汉语中是平方的意思，最小二乘法回归就是要寻找一条直线使得所有的点到该直线的垂直距离的平方和最小，下面具体讨论之.

假设分别用 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 去估计 β_0 和 β_1，从而得到拟合回归直线 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$，则样本点中任意一点到该直线的垂直距离为

$$\hat{\varepsilon}_i=y_i-\hat{y}_i=y_i-\hat{\beta}_0-\hat{\beta}_1 x_i,$$

它刻画了各观测值 (x_i, y_i) 与拟合直线的偏离程度(见图 10.3).

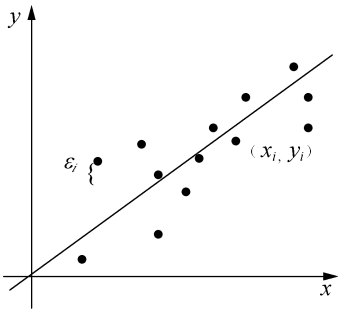


图 10.3　残差示意图

称 $\hat{\varepsilon}_i$ 为**残差**. 我们希望残差尽可能的小，这样拟合直线与观测值就尽可能接近. 显然我们可以用**绝对残差和**

$$\sum_{i=1}^{n}|\hat{\varepsilon}_i|=\sum_{i=1}^{n}|y_i-\hat{\beta}_0-\hat{\beta}_1 x_i| \tag{10.2.5}$$

来刻画拟合程度，但是在数学上绝对值函数处理起来比较繁琐，所以一般考虑**残差平方和**

$$Q(\beta_0, \beta_1)=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2 \tag{10.2.6}$$

而 β_0 和 β_1 的最小二乘估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$，就是使 $Q(\beta_0, \beta_1)$ 达到最小的 β_0 和 β_1，即

$$Q(\hat{\beta}_0, \hat{\beta}_1)=\min_{\beta_0, \beta_1} Q(\beta_0, \beta_1).$$

根据求多元函数最小值点的微分法，为求 $Q(\beta_0, \beta_1)$ 的最小值，对 $Q(\beta_0, \beta_1)$ 分别关于 β_0 和 β_1 求偏导数并令其为零，得到方程组

$$\begin{cases}\sum\limits_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)=0\\ \sum\limits_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)x_i=0\end{cases} \tag{10.2.7}$$

上述方程组称为**正规方程组**，化简得到

$$\begin{cases} n\beta_0 + \left(\sum_{i=1}^{n} x_i\right)\beta_1 = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)\beta_0 + \left(\sum_{i=1}^{n} x_i^2\right)\beta_1 = \sum_{i=1}^{n} x_i y_i \end{cases},$$

解方程组得 β_1 和 β_0 的**最小二乘估计**

$$\begin{cases} \hat{\beta}_1 = \dfrac{L_{xy}}{L_{xx}} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases} \tag{10.2.8}$$

其中

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\ L_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2,\ L_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

由式(10.2.8)中$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$知，回归方程$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$一定通过点$(\bar{x},\ \bar{y})$，该点为各散点的中心. 这正体现 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 在平均意义下的定量关系表达式的初衷. 最小二乘估计之所以被广泛应用，因为它有许多优良性质，下面的定理给出了最小二乘估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 一些重要性质.

对于估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$，由于样本不同而得到的估计量也是不同的，故这些估计量是随机变量，也有其对应分布，可以用来构造检验统计量作统计推断. 我们不加证明的叙述如下事实.

定理 10.2.1 对于一元线性回归模型式(10.2.3)，若$\hat{\beta}_0$和$\hat{\beta}_1$是β_0和β_1的最小二乘估计，则

$$\hat{\beta}_0 \sim N\left(\beta_0,\ \left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right)\sigma^2\right) \tag{10.2.9}$$

$$\hat{\beta}_1 \sim N\left(\beta_1,\ \frac{1}{L_{xx}}\sigma^2\right) \tag{10.2.10}$$

其中 $L_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$.

定理 10.2.1 说明：

(1) $\hat{\beta}_0$，$\hat{\beta}_1$ 是β_0，β_1 的无偏估计.

(2) $\hat{y}_0$ 是 $E(y_0) = \beta_0 + \beta_1 x_0$ 的无偏估计，事实上，

$$E(\hat{y}_0) = E(\hat{\beta}_0) + E(\hat{\beta}_1)x_0 = \beta_0 + \beta_1 x_0 = E(y_0).$$

(3) 要提高 $\hat{\beta}_0$，$\hat{\beta}_1$ 的精度，则要增大 n 和 L_{xx}.

为了今后讨论方便，再引进记号 $L_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$，在实际求回归方程时，常常需要将 L_{xx}，L_{xy}，L_{yy} 改写成便于计算的形式，即

$$L_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2,\ L_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\cdot\bar{y},\ L_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

σ^2 的无偏估计通常用残差 $\hat{\varepsilon}_i$ 来构造，具体的说就是

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{10.2.11}$$

我们不加证明的叙述如下事实.

定理 10.2.2　(1) $\hat{\sigma}^2$ 是 σ^2 的无偏估计；(2) $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2}\sim\chi^2(n-2)$ 并且 $\hat{\sigma}^2$ 与 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 相互独立.

下面，我们利用上述结论继续计算例 10.2.1 中的线性回归方程.

续例 10.2.1　计算 10.2.1 中的回归方程.

解　设年龄为 x，血压为 y，列表格计算数据如下：

编号	x	y	x^2	xy	y^2	残差
1	34	125	1 156	4 250	15 625	−4.70
2	42	128	1 764	5 376	16 384	−9.27
3	47	160	2 209	7 520	25 600	18.01
4	64	162	4 096	10 368	26 244	3.94
5	63	144	3 969	9 072	20 736	−13.12
6	46	142	2 116	6 532	20 164	0.95
7	21	120	441	2 520	14 400	2.59
8	25	125	625	3 125	15 625	3.81
9	37	120	1 369	4 440	14 400	−12.54
10	53	158	2 809	8 374	24 964	10.34
合计	432	1 384	20 554	61 577	194 142	

利用表格求得

$$\bar{x} = \frac{432}{10} = 43.2,\ \bar{y} = \frac{1384}{10} = 138.4,\ L_{xx} = \sum_{i=1}^{7}x_i^2 - n\bar{x}^2 = 20\,554 - 10\times 43.2^2 = 1\,891.6,$$

$$L_{xy} = \sum_{i=1}^{n}x_iy_i - n\bar{x}\cdot\bar{y} = 61\,577 - 10\times 138.4^2 = 1\,788.2$$

由公式(10.2.4)得

$$\hat{\beta}_1 = \frac{L_{xy}}{L_{xx}} = \frac{1\,788.2}{1\,891.6} \approx 0.945\,3,\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 138.4 - 0.945\times 43.2 = 97.561\,4,$$

取 $\hat{\beta}_0 = 97.561\,4$ 和 $\hat{\beta}_1 = 0.945\,3$，因此回归方程为

$$\hat{y} = 97.561\,4 + 0.945\,3x$$

参数 σ^2 的无偏估计为 $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-2} = 113.24.$

R 软件的线性回归模型的命令是 lm，对于上述例子的代码为如下：

```
x<-c(34,42,47,64,63,46,21,25,37,53)
y<-c(125,128,160,162,144,142,120,125,120,158)
fit<-lm(y~x)
summary(fit)
```

R 软件分析结果和上面的结果保持一致,具体结果如下:

```
Coefficients:
             Estimate Std.Error t value Pr(>|t|)
(Intercept)   97.5614   11.0928   8.795  2.19e-05 ***
x              0.9453    0.2447   3.864  0.00478 **
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

10.2.3 回归系数的统计推断

对于任何两个变量x和y，只要对其进行n次观测，得到数据$(x_i, y_i)(i=1, 2, \cdots, n)$后，就可利用式(10.2.8)计算得到拟合回归方程$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$. 但这样得到的回归方程不一定有意义. 故在使用拟合回归方程作进一步的分析之前，首先要对拟合回归方程是否有意义做出判断. 在一元线性模型$y=\beta_0+\beta_1 x$中，Y与X之关系主要是通过参数β_1连结的，参数β_1的意义就是：若解释变量X增加一个单位，响应变量Y的平均值就增加β_1. 如果$\beta_1=0$则说明Y与X之间并无线性关系，如果$\beta_1 \neq 0$说明Y与X存在线性关系，故必须要检验

$$H_0: \beta_1=0 \leftrightarrow H_1: \beta_1 \neq 0$$

下面介绍该假设检验最常用的的检验方法：**F检验法(方差分析法)**.

观测值$y_i(i=1, 2, \cdots, n)$之间的差异主要来自两个方面：一是解释变量不同；二是其他因素的影响. 为了检验哪个是主要因素，就必须把这两部分的差异从总的差异中分解出来. 观测值$y_i(i=1, 2, \cdots, n)$之间的差异，可以用y_i与其均值$\bar{y}$的离差平方和来表示，称为**总离差平方和**，记作

$$SS_T=\sum_{i=1}^{n}(y_i-\bar{y})^2=L_{yy} \tag{10.2.12}$$

和上一章的方差分析中总离差平方和的分解式类似，可以推得

$$SS_T=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{10.2.13}$$

记为

$$SS_T=SS_R+SS_E.$$

其中

$$SS_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\hat{\beta}_1 L_{xy} \tag{10.2.14}$$

称为**回归平方和**，它是由解释变量 x 的变化而引起的差异.

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = L_{yy} - \hat{\beta}_1 L_{xy} = Q(\hat{\beta}_0, \hat{\beta}_1) \tag{10.2.15}$$

称为**误差平方和**，它是由试验误差以及其他未加控制的因素引起的.

通过上述分解，我们把反映观测值 $y_i(i=1, 2, \cdots, n)$ 之间差异的总离差平方和分解为回归平方和与误差平方和之和. 这里 $\frac{SS_E}{\sigma^2} \sim \chi^2(n-2)$，在原假设 H_0：$\beta_1 = 0$ 成立条件下，$\frac{SS_T}{\sigma^2} \sim \chi^2(n-1)$，$\frac{SS_R}{\sigma^2} \sim \chi^2(1)$ 且 $\frac{SS_R}{\sigma^2}$ 与 $\frac{SS_E}{\sigma^2}$ 相互独立. 故构造 F 统计量及其在原假设成立条件下的分布为

$$F = \frac{SS_R}{SS_E/(n-2)} \sim F(1, n-2) \tag{10.2.16}$$

原假设 H_0：$\beta_1 = 0$ 的拒绝域为 $F \geqslant F_\alpha(1, n-2)$. 给定检验水平 α，查 F 分布表得其临界值 $F_\alpha(1, n-2)$，计算检验统计量的值 F，若 $F \geqslant F_\alpha(1, n-2)$ 则拒绝原假设 H_0，否则就接受原假设 H_0. 检验结果可列出方差分析表(见表 10.1).

表 10.1　一元线性回归的方差分析表

来源	平方和	自由度	均方和	F 值	临界值
回归	$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	1	$MS_R = SS_R/1$	$F = \frac{MS_R}{MS_E}$	$F_\alpha(1, n-2)$
残差	$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n-2$	$MS_E = SS_E/(n-2)$		
总和	$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$			

现在用上述方法对于例 10.2.1 的回归方程进行检验，给定检验水平 $\alpha = 0.05$，检验结果见下面的方差分析表：

方差来源	平方和	自由度	均方和	F 值	临界值
回归	1 690.45	1	1 690.45	14.928	5.32
残差	905.95	8	113.25		
总和	2 596.40	9			

因为 $F > F_{0.05}(1, 8)$，故拒绝原假设，即认为线性回归方程成立，也就是说年龄 x 与血压 y 存在线性关系：$\hat{y} = 97.561\,4 + 0.945\,3x$.

也可以采用 t 检验法. 由定理 10.2.1 和定理 10.2.2 我们知道

$$\hat{\beta}_1 \sim N(\beta_1, \frac{1}{L_{xx}}\sigma^2), \quad \frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{SS_E}{\sigma^2} \sim \chi^2(n-2),$$

且两者相互独立，于是当原假设 H_0：$\beta_1 = 0$ 成立时，有

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{L_{xx}}} \sim t(n-2), \tag{10.2.17}$$

其中 $\hat{\sigma}=\sqrt{SS_E/(n-2)}$. 给定检验水平 α，其拒绝域为 $|t|>t_{\alpha/2}(n-2)$，这就是 t 检验法. 事实上，由于 t 分布与 F 分布之间的关系，可知

$$t^2=\frac{\hat{\beta}_1^2}{\hat{\sigma}^2/L_{xx}}=\frac{\hat{\beta}_1 L_{xy}}{\hat{\sigma}^2}=\frac{SS_R}{SS_E/(n-2)}=F\sim F(1,\ n-2).$$

显然 t 检验法与之前的 F 检验法是等价的.

现在引入一个与样本相关系数有关的统计量 R^2，称为决定系数，可以证明 $SS_R=R^2SS_T$，于是有 $R^2=\dfrac{SS_R}{SS_T}$，它建立了相关系数与回归之间的联系，又通过具体数量大小反映了回归的贡献大小，这是回归分析中一个十分有用的统计量. 因检验有时也会犯错，若补充计算一下其决定系数值，如果 R^2 远远小于 0.5，说明由解释变量引起响应变量的变化部分不足 50%，此时若求线性回归方程，就没多大的实用价值了.

10.2.4　预测和控制

建立回归方程的重要目的就是为了用来预测和控制响应变量 y 的值，预测的前提必须是回归方程是有效的. 所谓预测，就是对解释变量的可取范围内的任何一个 x_0 时，对响应变量相应的取值 y_0 的一个估计；所谓控制，就是通过控制解释变量的取值来把响应变量的值限制在指定范围内.

1. 点预测

设回归方程为 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$，对于在可取范围内给定的解释变量值 x_0，用 $\hat{\beta}_0+\hat{\beta}_1 x_0$ 作为相应的响应变量的预测值，记为 $\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1 x_0$，这种方法就是**点预测**. 注意到，若用已建立的回归方程作预测时，如果解释变量值 x_0 在建立回归方程时所用的解释变量数据的范围之外作预测，要特别慎重，一般要求 x_0 点不能外推得太远.

2. 区间预测

所谓区间预测，就是对于在可取范围内给定的解释变量值 x_0，响应变量的取值 y_0 有一个置信度为 $1-\alpha$ 的区间，称为预测区间，即找到包含 y_0 的区间 $(t_1,\ t_2)$，使其满足

$$P(t_1<y_0<t_2)=1-\alpha.$$

对于一元线性回归模型下，可以证明

$$y_0-\hat{y}_0\sim N\left(0,\ \sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}\right)\right) \tag{10.2.18}$$

又由式(10.2.11)和定理10.2.2可知 $\hat{\sigma}^2=\dfrac{SS_E}{n-2}$ 且 $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2}\sim\chi^2(n-2)$，故构造 T 统计量及其分布为

$$T=\frac{y_0-\hat{y}_0}{\hat{\sigma}\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{L_{xx}}}}\sim t(n-2) \tag{10.2.19}$$

其中 $\hat{\sigma}=\sqrt{\dfrac{SS_E}{n-2}}$，于是对于给定的置信度 $1-\alpha$，则有

$$P(|T|<t_{\alpha/2}(n-2))=1-\alpha.$$

即有

$$P\{\hat{y}_0-\delta(x_0)<y_0<\hat{y}_0+\delta(x_0)\}=1-\alpha,$$

其中 $\delta(x_0)=t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{L_{xx}}}$，因此 y_0 的置信度为 $1-\alpha$ 的预测区间为$(\hat{y}_0-\delta(x_0),\ \hat{y}_0+\delta(x_0))$.

对于任意的 x，根据样本的观测值可做出两条曲线方程

$$y_1(x)=\hat{y}(x)-\delta(x) \tag{10.2.20}$$

$$y_2(x)=\hat{y}(x)+\delta(x) \tag{10.2.21}$$

这两条曲线将回归直线 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$ 夹在中间，形成一条宽窄不等的带域，该带域在 $x=\bar{x}$ 处最窄，如图 10.4 所示 .

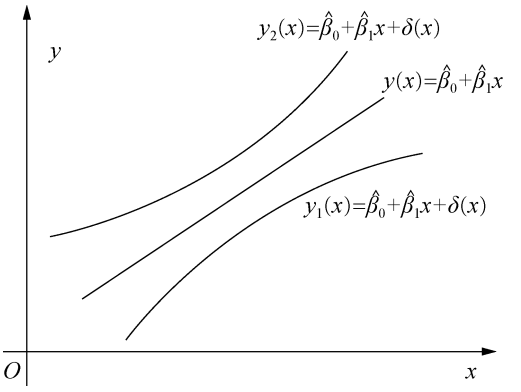


图 10.4　预测区间示意图

续例 10.2.1　当年龄是 $x=47$ 时，求所获得血压 y 的 95% 的预测区间.

解　根据前面讨论的结果，回归方程为 $\hat{y}=97.5614+0.9453x$ 而且该方程显著，故可用来做预测. 当 $x=47$ 时，根据回归方程得到点预测值为 $\hat{y}_0=141.99$. 现在已知 $\alpha=0.05$，$n=10$，查 t 分布表可知 $t_{0.025}(8)=2.306$，而之前已经算得 $\hat{\sigma}^2=113.24$. 对数据计算得

$$L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=1891.6,\quad (x_0-\bar{x})^2=14.44.$$

故根据公式

$$\delta(x_0)=t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}}$$

得 $\delta=2.306\times\sqrt{113.24}\times\sqrt{1+\dfrac{1}{10}+\dfrac{14.44}{1891.6}}=25.83$，故

$$y_1(47)=141.99-25.83=116.16,$$

$$y_2(47)=141.99+25.83=167.81.$$

因此年龄 $x=47$ 时，得血压 y 的 95% 的预测区间为(116.16，167.81).

R 软件计算上述预测区间的代码如下：

```
newx<-data.frame(x=47)
newy<-predict.lm(fit,newx,interval="predict")
```

R 程序的结果和上述计算结果一样，显示如下.

```
        fit       lwr       upr
1 141.9923 116.1658 167.8187
```

在 x_0 与 $\bar{x}$ 比较接近且样本量 n 较大时，可以对式(10.2.20)和式(10.2.21)取近似值. 此时有 $t_{\alpha/2}(n-2) \approx u_{\alpha/2}$ 和 $\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}} \approx 1$，则 y_0 的置信度为 $1-\alpha$ 的置信区间可近似表示为

$$(\hat{y}_0(x)-u_{\alpha/2}\cdot\hat{\sigma},\ \hat{y}_0(x)+u_{\alpha/2}\cdot\hat{\sigma}) \tag{10.2.22}$$

3. 控制

预测的反问题就是控制问题，即如果要求将响应变量 y 的取值范围控制在一定范围内，那么解释变量 x 和的取值应该控制在哪一个范围内？控制的前提也必须满足回归方程显著. 在这里就只考虑样本量较大的情形，而一般的情形亦可用类似的方法讨论.

假设现在需要控制响应变量 y 的取值范围在 (y_1, y_2) 中，可利用近似区间式(10.2.22)，令

$$y_1 = \hat{y}_0 - u_{\frac{\alpha}{2}}\cdot\hat{\sigma} = \hat{\beta}_0+\hat{\beta}_1 x_1 - u_{\frac{\alpha}{2}}\cdot\hat{\sigma},$$

$$y_2 = \hat{y}_0 + u_{\frac{\alpha}{2}}\cdot\hat{\sigma} = \hat{\beta}_0+\hat{\beta}_1 x_2 + u_{\frac{\alpha}{2}}\cdot\hat{\sigma},$$

解出上述方程得

$$x_1 = \frac{y_1-\hat{\beta}_0+u_{\frac{\alpha}{2}}\cdot\hat{\sigma}}{\hat{\beta}_1} \tag{10.2.23}$$

$$x_2 = \frac{y_2-\hat{\beta}_0-u_{\frac{\alpha}{2}}\cdot\hat{\sigma}}{\hat{\beta}_1} \tag{10.2.24}$$

当 $\hat{\beta}_1>0$ 时，解释变量 x 的控制范围为 (x_1, x_2)；当 $\hat{\beta}_1<0$ 时. 解释变量 x 的控制范围为 (x_2, x_1). 在实际应用中要实践控制，必须要满足响应变量 y 的取值范围在 (y_1, y_2) 的区间长度超过 $2u_{\alpha/2}\cdot\hat{\sigma}$，否则控制区间就不存在.

*10.3　多元线性回归分析

在实际问题中，影响一个试验的结果的因素常常不止一个，这就需要研究一个响应变量 y 与多个解释变量 x_1，x_2，…，x_p 之间的相关关系. 研究该问题的常用方法就是多元回归分析方法. 而多元回归分析中最常用的就是多元线性回归分析，它是一元线性回归的推广.

10.3.1　多元线性回归模型

设影响响应变量 y 的解释变量个数为 p 个，分别记为 x_1，x_2，…，x_p，通过 n 次观测取

得样本观测值为$(x_{i1}, x_{i2}, \cdots, x_{ip}, y_i)$，$i=1, 2, \cdots, n$. 所谓多元线性回归模型是指这些解释变量对响应变量的影响是线性的，即

$$Y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon,\ \varepsilon\sim N(0,\ \sigma^2) \tag{10.3.1}$$

其中β_0，β_1，β_2，$\cdots$，β_p，σ^2是与x_1，x_2，$\cdots$，x_p无关的未知参数.

由于样本的观测值分别是$(x_{i1}, x_{i2}, \cdots, x_{ip}, y_i)(i=1, 2, \cdots, n)$，则有

$$\begin{cases} y_1=\beta_0+\beta_1x_{11}+\beta_2x_{12}+\cdots+\beta_px_{1p}+\varepsilon_1 \\ y_2=\beta_0+\beta_1x_{21}+\beta_2x_{22}+\cdots+\beta_px_{2p}+\varepsilon_2 \\ \vdots \\ y_n=\beta_0+\beta_1x_{n1}+\beta_2x_{n2}+\cdots+\beta_px_{np}+\varepsilon_n \end{cases},$$

其中ε_1，ε_2，$\cdots$，ε_n相互独立，且$\varepsilon_i\sim N(0,\ \sigma^2)$，$i=1, 2, \cdots, n$，令

$$Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\ X=\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},\ \beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},\ \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

则上述统计模型可用矩阵形式表示为

$$\begin{cases} Y=X\beta+\varepsilon \\ \varepsilon\sim N(0,\ \sigma^2I_n) \end{cases} \tag{10.3.2}$$

其中$X_{n\times(p+1)}$称为设计矩阵，$Y_{n\times1}$称为响应向量，$\beta_{n\times1}$称为回归系数向量，ε是n维随机误差向量，各分量间相互独立，I_n是n阶单位矩阵.

10.3.2　最小二乘估计

与一元线性回归类似，我们亦采用最小二乘法估计参数β_0，β_1，β_2，$\cdots$，β_p. 引入误差平方和

$$Q(\beta_0,\ \beta_1,\ \cdots,\ \beta_p)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})^2.$$

最小二乘估计就是求$\hat{\beta}=(\hat{\beta}_0,\ \hat{\beta}_1,\ \cdots,\ \hat{\beta}_p)'$，使得

$$Q(\hat{\beta}_0,\ \hat{\beta}_1,\ \cdots,\ \hat{\beta}_p)=\min_{\beta}Q(\beta_0,\ \beta_1,\ \cdots,\ \beta_p).$$

因为$Q(\beta_0,\ \beta_1,\ \cdots,\ \beta_p)$是$\beta_0$，$\beta_1$，$\cdots$，$\beta_p$的非负二次型，故其最小值一定存在. 根据求多元函数最小值点的微分法，令

$$\begin{cases} \dfrac{\partial Q}{\partial\beta_0}=-2\sum\limits_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\cdots-\beta_px_{ip})=0 \\ \dfrac{\partial Q}{\partial\beta_j}=-2\sum\limits_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\cdots-\beta_px_{ip})x_{ij}=0,\ j=1, 2, \cdots, p \end{cases},$$

上述方程组称为**正规方程组**，可用矩阵表示为

$$X^TX\beta=X^TY \tag{10.3.3}$$

在系数矩阵 X′X 可逆条件下，可解得

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{10.3.4}$$

$\hat{\beta}$ 就是 β 的**最小二乘估计**，即 $\hat{\beta}$ 为多元线性回归方程

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \tag{10.3.5}$$

的**回归系数**，其中 $\hat{\beta}_j$ 称为响应变量 y 关于解释变量 x_j 的**偏回归系数**.

10.3.3　多元线性回归模型的有效性检验

1. 多元线性回归方程的统计检验

在求出多元线性回归方程(10.3.5)后，接下来就自然而然的讨论这个方程的回归效果是否显著，即响应变量与解释变量之间是否存在线性相关关系. 因此我们需要多元线性回归模型的假设检验问题. 要检验的原假设与备择假设如下：

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \leftrightarrow H_1: \beta_1, \beta_2, \cdots, \beta_p \text{ 中至少有一个不为零} \tag{10.3.6}$$

与一元线性回归模型的检验类似，我们通过方差分析构造 F 统计量对上述假设作 F 检验. 考察响应变量的观测值 $y_i(i = 1, 2, \cdots, n)$ 的总离差平方和 SS_T，对其进行分解，即

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 SS_T = SS_R + SS_E \tag{10.3.7}$$

其中

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{10.3.8}$$

称为**回归平方和**；

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_i x_{ij})^2 \tag{10.3.9}$$

称为**误差平方和**.

SS_R 反映了回归效应，它是由响应变量与解释变量间的线性相关关系引起的. SS_E 反映了观测值 $y_i(i = 1, 2, \cdots, n)$ 偏离回归直线的程度，它是由随机误差等随机因素造成的. 可以证明 $\frac{SS_E}{\sigma^2} \sim \chi^2(n-p-1)$，如果在原假设 $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ 成立，则有 $\frac{SS_T}{\sigma^2} \sim \chi^2(n-1)$，$\frac{SS_R}{\sigma^2} \sim \chi^2(p)$，且 $\frac{SS_R}{\sigma^2}$ 与 $\frac{SS_E}{\sigma^2}$ 相互独立. 由此构造 F 统计量及其在原假设成立条件下的分布

$$F = \frac{SS_R/p}{SS_E/(n-p-1)} \sim F(p, n-p-1) \tag{10.3.10}$$

如果响应变量 y 与解释变量 $x_i(i = 1, 2, \cdots, p)$ 之间线性关系显著，则 SS_R 值会较大，因此得到的 F 值较大，反之 F 值则较小. 由此根据给定的检验性水平 α，查得 F 分布的临界值 $F_\alpha(p, n-p-1)$. 如果 $F > F_\alpha(p, n-p-1)$，则拒绝原假设，即认为响应变量 y 与解释变量 $x_i(i = 1, 2, \cdots, p)$ 之间线性关系显著，反之则接受原假设，即认为响应变量 y 与解释变量 $x_i(i = 1, 2, \cdots, p)$ 之间不存在显著的线性关系. 上述假设检验可通过下列的方差

分析表表示.

表 10.2　多元线性回归的方差分析表

来源	平方和	自由度	均方	F 值	临界值
回归	$SS_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	p	$MS_R=\dfrac{SS_R}{p}$	$F=\dfrac{MS_R}{MS_E}$	$F_\alpha(p,\ n-p-1)$
残差	$SS_E=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$n-p-1$	$MS_E=\dfrac{SS_E}{(n-p-1)}$		
总和	$SS_T=\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$			

2. 多元线性回归方程回归系数的统计检验

与一元线性回归模型不同的是，在多元线性回归方程显著的时候，并不是所有解释变量对响应变量都有显著影响，此时需要进一步检验每个偏回归系数 $\beta_j(j=1,\ 2,\ \cdots,\ p)$ 是否为零. 为此提出 p 个原假设与备择假设如下：

$$H_{0j}:\ \beta_j=0\leftrightarrow H_{1j}:\ \beta_j\neq 0,\ j=1,\ 2,\ \cdots,\ p \tag{10.3.11}$$

在原假设 H_{0j} 成立的条件下，仿照一元线性回归模型中的 t 检验方法，构造 t 统计量及其分布为

$$T_j=\frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}}\sim t(n-p-1) \tag{10.3.12}$$

其中 c_{jj} 为矩阵 $X'X$ 对角线上的第 j 个元素，$\hat{\sigma}=\sqrt{SS_E/(n-p-1)}$.

给定检验水平 α，若 $|T_j|>t_{\alpha/2}(n-p-1)$，$j=1,\ 2,\ \cdots,\ p$，则拒绝原假设 H_{0j}，即认为解释变量 x_j 对响应变量 y 有统计意义.

3. 决定系数

我们在一元回归分析中已介绍过决定系数的概念，在多元回归中决定系数仍为

$$R^2=\frac{SS_R}{SS_T}.$$

它用以反映多元回归模型能在多大的程度解释响应变量 y 与解释变量 $x_1,\ x_2,\ \cdots,\ x_p$ 之间的因果关系，其取值范围为 $0\leqslant R^2\leqslant 1$，当 $R^2\to 1$ 时，表示样本数据较好地拟合了的回归模型，当 $R^2\to 0$ 时，表示样本数据不能拟合为线性回归模型. 同时我们亦称决定系数 R^2 的算术平方根 R 为复相关系数，它表示了变量 y 与解释变量 $x_1,\ x_2,\ \cdots,\ x_p$ 的线性相关密切程度.

10.3.4　多元线性回归的预测区间

与一元线性回归模型类似，可以利用求得的线性回归方程

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p$$

对响应变量做出预测. 给定解释变量的观测点 $X_0=(x_1^0,\ x_2^0,\ \cdots,\ x_p^0)'$，可计算对应的

点响应变量的预测值$\hat{y}_0$. 此外，还可以计算y_0的预测区间. 给定的检验水平α，查t分布表得临界值为$t_{\alpha/2}(n-p-1)$，则y_0的置信度为$1-\alpha$的预测区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-p-1)\hat{\sigma}\sqrt{1+X_0'(X'X)^{-1}X_0} \tag{10.3.13}$$

其中$\hat{\sigma}=\sqrt{SS_E/(n-p-1)}$.

在建立了线性回归方程$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p$后，在应用上除了作预测外，还要注重回归系数的解释，一般地，β_i的绝对值大则x_i对响应变量y的影响就大，$\beta_i>0$时x_i对响应变量y有正向影响，$\beta_i<0$时x_i对响应变量y有反向影响.

例 10.3.1　研究某土壤内所含植物可给态磷浓度y与土壤内含无机磷浓度x_1，土壤内易溶于碳酸钾溶液并受化合物水解的有机磷浓度x_2和土壤内溶于碳酸钾但不受水解的有机磷浓度x_3之间关系，做试验得到如下数据(见表 10.3)：

(1) 计算出y与x_1，x_2，x_3的线性回归方程.

(2) 在检验水平为$\alpha=0.05$下，对多元线性回归方程进行统计检验.

(3) 若$x_1=50$，$x_2=46$，$x_3=98$，求对应Y值的置信度为95%的预测区间.

表 10.3　土壤试验数据表

y	x_1	x_2	x_3
0.4	53	158	64
0.4	23	163	60
3.1	19	37	71
0.6	34	157	61
4.7	24	59	54
1.7	65	123	77
10.4	44	46	81
10.1	31	117	93
11.6	29	173	93
12.6	58	112	51
10.9	37	111	76
23.1	46	114	96
23.1	50	134	77
21.6	44	73	93
23.1	56	168	95
1.9	36	143	54
26.8	58	202	168
29.9	51	124	99

解(1)　利用上表数据写出设计矩阵

$$X_{18\times4}=\begin{pmatrix}1&53&158&64\\1&23&163&60\\\cdots&\cdots&\cdots&\cdots\\1&58&202&168\\1&51&124&99\end{pmatrix},\quad Y_{18\times1}=\begin{pmatrix}0.4\\0.4\\\cdots\\26.8\\29.9\end{pmatrix},$$

由式(10.3.4)计算得到回归系数向量

$$\hat{\beta} = (X^T X)^{-1} X^T Y = \begin{pmatrix} -12.073\ 1 \\ 0.214\ 9 \\ -0.041\ 7 \\ 0.247\ 9 \end{pmatrix}.$$

故求得的多重回归方程为

$$y = -12.073\ 1 + 0.214\ 9x_1 - 0.041\ 7x_2 + 0.247\ 9x_3$$

(2) 回归方程的检验由表 10.4 的方差分析表给出.

表 10.4　土壤数据的方差分析表

来源	平方和	自由度	均方和	F 值	临界值
回归	987.91	3	329.30	6.06	$F_{0.05}(3, 14) = 3.34$
残差	760.91	14	54.35		
总和	1 748.82	17			

决定系数为 $R^2 = \dfrac{SS_R}{SS_T} = \dfrac{987.92}{1\ 748.82} = 0.564\ 9$，由此可知，该回归方程显著，即 y 与 x_1，x_2，x_3 的线性关系显著.

回归系数的检验如表 10.5 所示，从该表中可以看出，当检验水平为 $\alpha = 0.05$ 时，只有解释变量 x_3 对应的回归系数 $\hat{\beta}_3$ 的 T 值落在拒绝域中，即 x_3 对 y 的影响显著，而 x_1 和 x_2 对 y 的影响不显著，可以将其删除后重新做一个回归方程.

表 10.5　土壤数据的回归系数检验表

解释变量	最小二乘估计 $\hat{\beta}_i$	$\hat{\beta}_i$ 的标准差	T 值	临界值
x_1	0.212 3	0.144 8	1.47	2.145
x_2	-0.039 01	0.043 26	-0.90	2.145
x_3	0.246 75	0.074 34	3.32	2.145

(3) 当 $x_1 = 50$，$x_2 = 46$，$x_3 = 98$ 时，$\hat{y}_0 = 21.049\ 1$，将 $t_{\alpha/2}(n-p-1) = t_{0.025}(14) = 2.145$，$\hat{\sigma} = \sqrt{MSE} = \sqrt{54.35}$，$\sqrt{1 + X_0^T (X^T X)^{-1} X_0} = 1.169$ 代入公式(10.3.13)得 y_0 的预测区间为(2.563 7，39.534 4).

R 软件计算上述例子的代码如下：

```
data<-read.table(file="D://Book1.txt",head=TRUE)
fit<-lm(Y~X1+X2+X3,data)
summary(fit)
anova(fit)
newx<-data.frame(X1=50,X2=46,X3=98)
newy<-predict.lm(fit,newx,interval="predict")
```

计算结果如下.

```
Coefficients:
             Estimate      Std.Error t value    Pr(>|t|)
(Intercept)  -12.07314     7.22382   -1.671     0.11686
X1           0.21492       0.14378   1.495      0.15716
X2           -0.04167      0.04295   -0.970     0.34846
X3           0.24789       0.07381   3.359      0.00468 **
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.372 on 14 degrees of freedom
Multiple R-squared:  0.5649,     Adjusted R-squared:  0.4717
F-statistic: 6.059 on 3 and 14 DF, p-value: 0.007294

       fit       lwr       upr
1 21.04911  2.563763  39.53445
```

习题十

1. 为了调查某农药公司的宣传费用投入与销售收入之间的关系，该公司记录了5个月的宣传投入 x(万元)与销售收入 y(万元)数据(见下表).

月份	1	2	3	4	5
宣传费用 x	1	2	3	4	5
销售收入 y	10	10	20	20	40

(1) 画出散点图，求出其样本相关系数.

(2) 用最小二乘估计求出回归方程.

(3) 检验水平为 $\alpha=0.05$，检验回归方程是否有效?

(4) 当宣传费用为4.2万元时，销售收入将达到多少，并给出置信度为95%的区间估计.

2. 测量不同浓度(x%)葡萄糖溶液在光电比色计上的消光度 y，得到数据如下表所示.

浓度 x	0	5	10	15	20	25	30
消光度 y	0	0.11	0.23	0.34	0.46	0.57	0.71

试根据结果求出回归方程，并预测葡萄糖溶液浓度在 $x=12$ 时的消光度及置信度为95%的预测区间.

3. 测定某肉鸡的生长过程，每两周记录一次肉鸡重量，数据如下表所示.

周数 x	2	4	6	8	10	12	14
重量 y	0.3	0.86	1.73	2.2	2.47	2.67	2.8

由经验已知肉鸡的生长曲线为 S 型曲线，极限生长重量值 $K=2.827$，试求重量对时间周数的 S 型曲线回归方程并检验方程显著性 $\alpha=0.01$.

4. 为了检验 X 射线的杀菌作用，用 200KV 的 X 射线照射杀菌，每次照射 6 分钟，照射次数为 x，照射后剩余细菌数为 y，试验数据如下表所示.

x	1	2	3	4	5	6	7	8	9	10
y	783	621	433	431	287	251	175	154	129	103
x	11	12	13	14	15	16	17	18	19	20
y	72	50	43	31	28	20	16	12	9	7

由经验知道两者的关系为指数函数曲线 $y=ae^{bx}$，

（1）求剩余细菌数 y 与照射次数 x 的曲线回归方程.

（2）在检验水平 $\alpha=0.05$，检验回归方程是否显著.

5. 果园土壤营养含量的高低直接关系到果树的生长、产量和品质的提高. 根据调查分析土壤全氮 N(x_1)，有效磷 P(x_2)，钾 K(x_3)，锌 Zn(x_4)，硼 B(x_5)(单位均为 ppm)，平均亩产量 y(单位 kg)的相关数据如下表.

编号	x_1	x_2	x_3	x_4	x_5	y
1	1.280 0	3 310.5	47.16	6.33	0.530 0	3 000
2	1.158 0	3 510.0	20.99	2.59	0.861 0	3 150
3	0.834 9	258.0	33.71	2.69	0.524 0	2 750
4	0.670 0	2 910.8	76.45	1.97	1.143 0	2 000
5	0.541 5	3 710.0	410.75	2.51	0.733 0	3 250
6	1.900 0	4 410.8	61.55	1.47	0.342 0	3 750
7	0.810 0	3 710.0	58.75	1.63	0.444 0	2 850
8	1.038 0	3 210.8	23.92	5.77	1.210 0	3 100
9	1.489 0	2 910.8	27.10	3.04	0.710 0	2 330
10	1.390 0	3 510.0	98.30	2.84	0.620 0	3 000
11	1.977 0	3 310.5	18.99	1.61	0.692 0	4 000
12	1.250 0	4 310.0	26.03	2.39	0.256 5	4 100
13	1.860 0	3 710.0	46.76	14.69	0.721 0	4 050
14	1.150 0	3 510.0	410.75	5.09	1.144 0	2 500

（1）求出 y 对 x_1，x_2，x_3，x_4，x_5 的线性回归方程.

（2）对上述线性回归方程进行显著性检验($\alpha=0.05$).

（3）对回归系数进行检验($\alpha=0.05$).

（4）预测 $x_1=1$，$x_2=300$，$x_3=60$，$x_4=3.5$，$x_5=0.7$ 时的平均亩产量 y 的点估计及其置信度为 95% 的预测区间.

第十一章　R 软件简介

11.1　R 的概述

R 语言是从 S 语言的基础上发展起来的，是一种为统计计算和绘图而生的语言和环境. 相对于其他的同类软件，它具有有效的数据处理和保存机制，完整的数组和矩阵计算操作符，强大的数据分析工具，优秀的统计制图功能等特点. 用户可以灵活机动地进行数据存储，数据分析和结果分享等工作. 通过 R 语言的许多内嵌统计函数，用户可以很容易学习和掌握 R 语言的语法，从而编制自己的函数来扩展现有的 R 语言，进行相关科研工作.

R 语言是自由软件，不向使用者收取任何费用，在官方网站上可以下载到 R 软件的 Windows 版本，然后按提示安装即可. Linux、Mac 及 OS 也都有相应的编译好的二进制版本.

1. 启动与退出

将 R 在计算机上安装完毕后，系统自动会在 Windows 菜单中和桌面上创建快捷方式. 双击快捷方式即可进入 R 的界面. 若要退出 R，可输入 q( )，或直接单击“关闭”按钮，在退出 R 之前，将会询问你是否保存工作空间.

2. 工作空间

工作空间就是当前 R 的工作环境，它储存着用户定义的对象(向量、矩阵、函数、数据框、列表). 在退出 R 时，可选择保存当前的工作空间到一个镜像中，并在下次启动 R 时自动载入它，从而避免一些不必要的输入，通过上下键查看已输入命令的历史记录.

当前的工作目录是 R 用来读取文件和保存结果的默认目录，使用函数 getwd( )可以查看当前的工作目录；函数 setwd( )可以用来设定当前的工作目录(也可通过单击【文件】，然后单击【改变工作目录】，选择所需要的工作目录)；函数 ls( )表示列出当前工作空间中的对象；函数 rm(objectlist)表示移除一个或多个对象；函数 history(n)表示显示当前工作空间最近使用过的 n 个命令；函数 save. image(" myfile")表示将当前工作空间保存到 myfile. RData 文件中等等.

R 的基本界面是一个交互式命令窗口，命令提示符是“>”，命令运行的结果显示在命令下面. R 语言的命令有两种形式：表达式和赋值运算(用“<-”或“ = ”表示赋值运算符). 在命令提示符后键入一个表达式表示计算表达式并显示结果. 例如：

```
>2 * 4+sqrt(4)
[1] 10
```

在命令提示符后键入“<-”或“ = ”进行赋值运算，把赋值号右边的值计算出来并赋给左边的变量. 例如：

```
>x<- 0:10;x   #如果不需要显示 x 的值,去掉 x 就可以
[1] 0 1 2 3 4 5 6 7 8 9 10
```

这里“0：10”表示间隔差为 1 的整数向量. 符号“#”表示此符号以后的这行语句是注释语句，R 是不会执行的.

若要绘制正弦曲线，则可以使用如下语句.

```
> x1<- 0:100
> x2<- x1 * 2 * pi/100
>y=sin(x2)
>plot(x2,y,type='l')
```

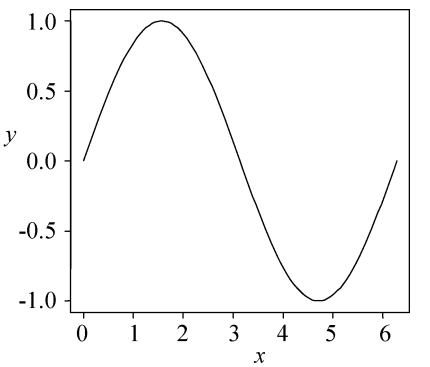


图 11.1　正弦曲线

若要建立新的程序脚本，可先单击【文件】，然后单击【新建程序脚本】打开一个新的 R 程序编辑窗口，输入要编写的 R 程序. 输入完毕后，选择保存，并给文件取名.

若要打开已有的程序脚本，可先单击【文件】，然后单击【打开程序脚本】打开已有的程序，屏幕弹出程序编辑窗口，可以利用这个窗口对已有程序进行编辑或执行.

若要加载程序包，可先单击主窗口中的【程序包】，然后单击【载入程序包】，选择需要调入的程序包，这里面提供了许多开箱即用的功能，包括分析地理数据、处理蛋白质质谱、读取 SPSS 数据的功能. 命令 search( ) 可以告诉你哪些包已加载并可使用.

3. 帮助信息

R 提供了丰富的在线帮助信息，学会如何使用这些帮助文档可有助于你的编程工作. R 的内置帮助系统提供了当前安装包中所有函数的细节、参考文献及使用范例. 例如：函数 help. start( ) 会打开帮助文档首页，我们可在其中查看入门和高级的帮助手册、常见问题集，以及参考材料；函数 help(" ** ") 可以查看函数 ** 的帮助；函数 example(" ** ") 给出函数 ** 的具体使用示例等.

11.2　R 软件的基本操作

11.2.1　向量的赋值与运算

定义向量的最常用方法是使用函数 c( )，也可以使用函数 assign( ). 例如：

```
>x<- c( 1,3,5,7)
>x
```

```
[1] 1 3 5 7
>assign("x",c(1,2,3,4)),x
[1] 1 2 3 4
>x1<-c(1,2)
>x2<-c(3,4)
>x=c(x1,x2)
>x
[1] 1 2 3 4
```

进一步有

```
>y<-c(x1,0,x2)
```

定义变量 y，其中两边分别是 x1 和 x2，中间是 0.

对于向量，可以进行加(+)、减(-)、乘(*)、除(/)和乘方(^)运算，其意义是对向量的每一个对应分量进行运算. 如：

```
>x<-c(1,0,2);y<-c(2,3,5)
>z<- -2*x+y+1;z
[1] 1 4 2
>x*y
[1] 2 0 10
>x/y
[1] 0.5 0.0 0.4
>y^x
[1] 2 1 25
```

另外，用“%/%”表示整除(如 7%/%3 为 2)，用“%%”表示求余数(如 7%%3 为 1).

一些基本初等函数，如 log( )、*exp*( )、cos( )、tan( )、sqrt( )等都可以将向量作为自变量，函数的返回值也将是向量，即向量的每个分量取相应的函数值. 例如：

```
>x<-c(1,4,16)
>sqrt(x)
[1] 1 2 4
```

函数 sqrt( )是对向量的每个分量进行开根处理，但如果要处理复数，应该给出明确的复数部分. 例如：

```
>sqrt(-4+0i)
[1] 0+2i
```

函数 min( )、max( )和 range( )分别用来求向量的最小值、最大值和范围，函数 sum( )、mean( )、var( )和 sd( )分别用来计算向量的和、均值、方差和标准差. 例如：

```
>x<-c(12,6,5,7,3)
>min(x)
[1]3
>mean(x)
[1]6.6
```

11.2.2　产生有规律的序列

如果要产生一个等差数列，只需在 R 里面键入“a：b”，这就产生了一个从 a 开始逐项加 1 或减 1 的序列，直到 b 结束. 例如：

```
>1:4
[1]1 2 3 4
>4:1
[1]4 3 2 1
```

但值得注意的是，在 R 里面冒号的优先级别最高，读者可以认真将“1：4-1”和“1：(4-1)”相互比较一下.

如果要产生一个更一般的等差数列，可以使用 seq( )函数，它产生等距间隔的数列，其基本形式为

```
seq(from=value1,to=value2,by=value3I),
```

即从 value1 开始，到 value2 结束，中间间隔为 value3. 例如：

```
>seq(-2,2,.5)
>-2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0
```

另外，使用

```
>seq(length=9,from=-2.0,by=0.5)
```

也可以产生一样的向量.

还有一个相关的函数是 rep( ). 用 rep(x, times=3)可将向量 x 重复三次，而 rep(x, each=3)是将向量里的每个元素重复 3 次.

11.2.3　矩阵、数组的生成和运算

数组可以看作是带有多个下标类型相同的元素集合，如数值型. 而矩阵是一种特殊的数组(二维数组). 利用 R 可以很容易地创建和处理数组.

产生数组最常用的方法是使用函数 array( )，其基本形式为

```
array(data=NA,dim=length(data),dimnames=NULL),
```

其中 data 表示一组向量，dim 表示数组的维数，默认值为向量的长度，dimnames 为数组维数的名字，默认值为空. 例如：

```
>x<- array(1:12,dim=c(3,4))
>x
```

产生一个 3×4 的二维数组(矩阵). 即

```
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

另外一种构造矩阵的方法是使用函数 matrix( )，其基本形式为

```
matrix(data=NA,nrow=1,ncol=1,byrow=FALSE,dimnames=NULL),
```

其中 data 表示一组向量，nrow 表示矩阵的行数，ncol 表示矩阵的列数，默认值 byrow = FALSE 表示数据按列放置，dimnames 表示数组的名字，默认值为空. 例如：

```
>x<-matrix(1:12,nrow=3,ncol=4,byrow=TRUE)
>x
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
```

如果将语句中的 byrow=TRUE(R 语言区分大小写，这里的 TRUE 要大写)去掉，则数据将按列放置. 如果要取出这个矩阵的第一行元素，可以用 $x[1,]$，如果要取出这个矩阵的第一行第二列的元素，可以使用 $x[1, 2]$.

对于矩阵 A，函数 t(A)表示矩阵的转置，函数 det(A)表示求方阵 A 的行列式，函数 eigen(A)表示求矩阵 A 的特征值和特征向量，函数 solve(A)表示求矩阵 A 的逆，函数 dim(A)表示求矩阵 A 的维数，函数 nrow(A)表示矩阵的行数，函数 ncol(A)表示矩阵的列数，函数 as. vector( )表示将函数进行拉直，转化为向量. 例如：

```
>A<-array(c(1,2,3,4,5,6,7,8,10),dim=c(3,3));A
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6   10
>det(A)
[1] -3
>solve(A)
```

```
          [,1]          [,2]         [,3]
[1,]  -0.6666667  -0.6666667    1
[2,]  -1.3333333   3.6666667   -2
[3,]   1.0000000  -2.0000000    1
>dim(A)
[1] 3 3
>as.vector(A)
[1] 1 2 3 4 5 6 7 8 10
```

对于两个同型矩阵 A 和 B，则 A * B 表示矩阵中对应元素相乘，A%*%B 表示通常意义下的两个矩阵的乘积. 例如：

```
>A<-array(c(1,3,5,7),dim=c(2,2))
>B<- array(4:1,dim=c(2,2))
>C<-A*B;C
     [,1] [,2]
[1,]   4    10
[2,]   9    7
>D<-A%*%B;D
     [,1] [,2]
[1,] 19    7
[2,] 33    13
```

对于矩阵，如果想对每行(或每列)、若干行(或若干列)进行某种计算，可以用函数 apply( )，其一般形式为

```
apply(A,margin,fun,…),
```

其中 A 为矩阵，margin=1 表示对每行计算，margin=2 表示对每列计算，fun 是用来计算的函数. 例如：

```
>A<-array(c(1,3,5,7),dim=c(2,2));A
      [,1]  [,2]
[1,]    1     5
[2,]    3     7
>apply(A,1,sum)
[1] 6 10
>apply(A,2,mean)
[1] 2 6
```

11.2.4　图形的绘制

数据作图是数据分析的一种直观的、有效的的方法. 在 R 里，我们可以利用图形工具显示各种各样的统计图并且创建一些全新的图，常用的作图函数包括 plot( )、pairs( )、hist( )、coplot( )、qqnorm( )、qqline( )和 contour( )等.

函数 plot( )用来绘制数据的散点图、曲线图等. 例如：

```
>Age<-c(13,13,14,14,12,15,11,14,14,14,15,13,12)
>Height<-c(148,150,151,149,143,153,141,149,152,158,153,147,145)
>Weight<-c(41,45,44,43,40,46,38,39,45,50,52,48,43)
>plot(Age,Height)  #绘制 Height 关于 Age 的散点图,如图 11.2 所示
>plot(Age)  #绘制关于 Age 的时间序列图形
>plot(Weight~Age+Height)  #绘制两张散点图,第一张是 Age 与 Weight 的散点图,第二张则是 Height 与 Weight 的
```

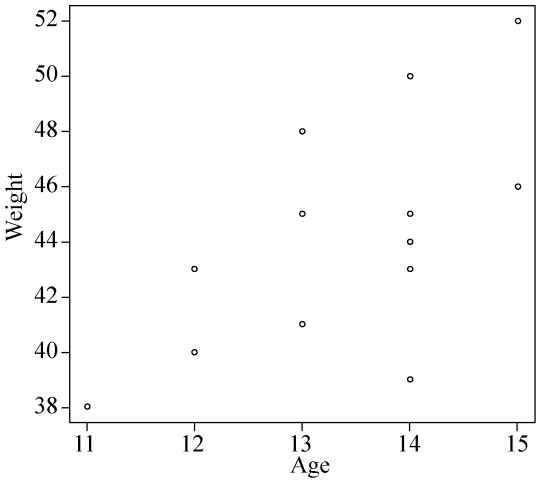


图 11.2　Weight 关于 Age 的散点图

对于多变量数据，例如 A 是一个 13×3 的矩阵，矩阵里的每列元素对应着 Age、Weight、和 Height，则函数 pairs(A)表示绘制关于矩阵各列，即 Age、Weight、和 Height 构成的散点图.

```
>coplot(Weight~Height | Age)  #绘制按年龄段给出的 Weight 与 Height 的散点图,如图 11.3 所示
```

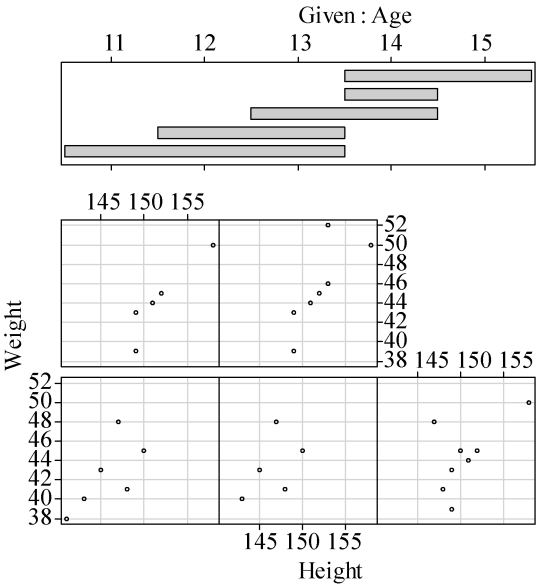


图 11.3　按年龄划分的 Weight 与 Height 的散点图

如果有 4 个变量，还可以用如下命令

```
>coploat(a~b | c+d)   #所有变量具有相同的长度
```

来绘制按 c、d 划分下，a 关于 b 的散点图.

```
>stem(Height)   #绘制 Height 的茎叶图,如图 11.4 所示
```

```
The decimal point is 1 digit(s) to the right of the |
  14 | 13
  14 | 57899
  15 | 01233
  15 | 8
```

图 11.4　Height 的茎叶图

```
>boxplot(Weight)   #绘制 Weight 的箱形图,如图 11.5 所示
```

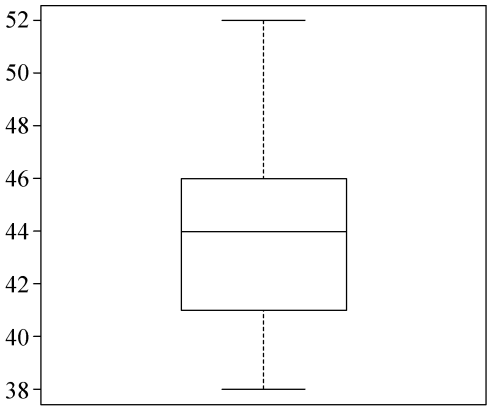


图 11.5　Weight 的箱形图

函数 hist( )表示绘制直方图,例如：

```
>hist(Age)
```

结果如图 11.6 所示 .

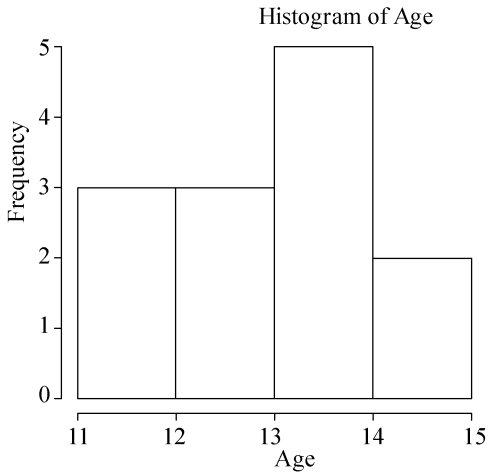


图 11.6　Age 的直方图

在 R 里，函数 qqnorm( )、qqline( )提供了画正态 QQ 图和相应直线的方法. 若正态 QQ 图上的点在一条直线附近，则可以认为样本数据来自正态总体. 例如：

```
>qqnorm(Weight)
>qqline(Weight,col="red")
```

执行后绘出正态 QQ 图，如图 11.7 所示. 从图中可以看出，数据基本上可以看成是自正态总体.

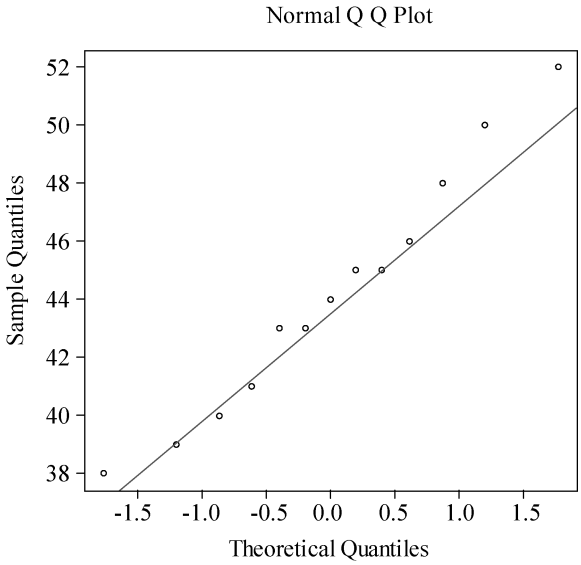


图 11.7　Weight 的正态 QQ 图

R 也可以绘出三维图形的表面曲线、等值线和等值色图，相应的函数分别为 persp( )、contour( )和 image( ). 例如：

```
>x<-y<-seq(-pi,pi,pi/15)
> f<-function(x,y) sin(x) * sin(y)
> z<-outer(x,y,f)   #在函数 f 关系下作外积运算,形成网格
> contour(x,y,z,col="red")
> persp(x,y,z,theta=30,phi=30,expand=0.7,col="lightblue")
>image(x,y,z)
```

结果如图 11-8 所示.

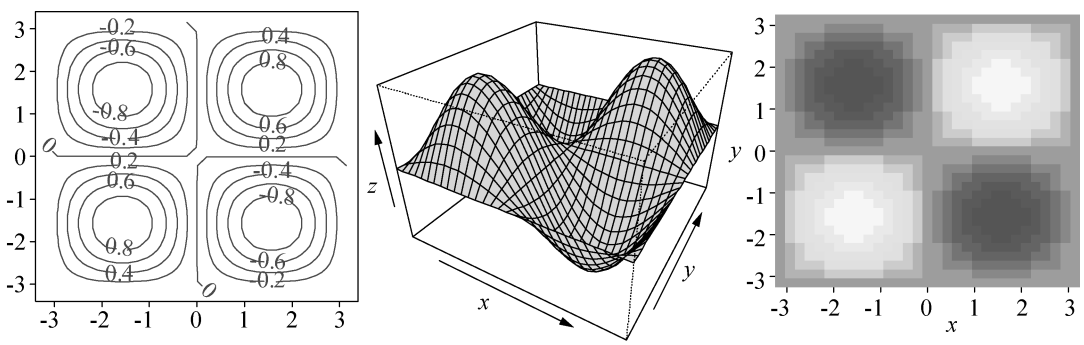


图 11.8　函数 z=sinxsiny 的等值线、网格曲面与等值色图

在绘制函数图形时，可以加一些命令选项，不断丰富和完善图形的内容，或增加一些有用的说明. 例如：

```
>plot(x,y,xlab="时间",ylab="价格",main="某某股票价格图")
```

其中 xlab、ylab 和 main 都是可选参数，用来说明图形的 x 轴、y 轴和标题. 如果没有这些选项，图形也就没有了 x 轴、y 轴和标题的说明.

下表给出一些常见的图形选项.

表 11.1　图形选项

add=TRUE add=FALSE	所绘图在原图上加图 新的图替换原来的图
log="x" log="y" log="xy"	x 轴的数据取对数 y 轴的数据取对数 x 轴与 y 轴的数据取对数
type="p" type="l" type="b" type="o" type="h" type="s" or "S" type="n"	绘散点图(默认值) 绘实线 所有点被实线连接 实线通过所有点 绘出点到 x 轴的实线 绘出阶梯形曲线 不绘出任何点或曲线

如果要在所绘图形上加点、线、标记、说明或其他内容，可以使用函数 points( )，lines( )，text( )，abline( )，polygon( )，legend( )，title( )和 axis( )等. 例如：

(1) points(x，y)表示在已有图形上加点(x，y).

(2) lines(x，y)的功能相对于命令 plot(x，y，type="l").

(3) abline(a，b)表示在已有图形上添加直线 y=a+bx.

(4) polygon(x，y)表示以数据(x，y)为坐标，依次连接所有的点，绘出一多边形.

(5) abline(v=x)表示绘出过所有点的竖直直线.

(6) abline(h=y)表示绘出过所有点的水平直线.

(7) title(main="主标题"，sub="子标题")表示将主标题加在图的顶部，子标题加在图的底部.

11.3　常用统计分析

11.3.1　分布函数或分布律

统计中一些典型分布的分布函数、分布律或概率密度函数，以及分布函数的反函数都可以通过 R 中的函数进行计算. 例如：

```
>dnorm(x,mean=0,sd=1)   #计算标准正态分布在点 x 处的概率密度
```

```
>pnorm(x,mean=0,sd=1)   #计算标准正态分布在点 x 处的分布函数值
>qnorm(1-x,mean=0,sd=1)   #计算标准正态分布的上 x 分位点
>rnorm(n,mean=0,sd=1)   #生成 n 个标准正态分布的随机数
```

其中 mean 表示正态分布的均值，默认值为 0；sd 表示正态分布的标准差，默认值为 1. 如果函数的返回值是对数正态分布，则需要用到逻辑变量 log、log. p. 例如：

```
>dnorm(x,mean=0,sd=1,log=TRUE)
```

如果要计算 $F(x)=P\{X>x\}$，则使用命令

```
> pnorm(x,mean=0,sd=1,lower.tail=FALSE,log.p=FALSE)
```

除了正态分布函数，其他的分布函数也有类似的结果. 下表给出了一些常用的分布，以及在 R 中的名称和调用函数用到的参数.

表 11.2　常用分布在 R 中的函数

分布	R 中的名称	参数
二项分布	binorm	Size，prob
泊松分布	pois	lambda
几何分布	geom	prob
超几何分布	hyper	m，n，k
均匀分布	unif	min，max
指数分布	exp	rate
正态分布	norm	mean，sd
t 分布	t	df，ncp
卡方分布	chisq	df，ncp
F 分布	f	df1，df2，ncp
柯西分布	cauchy	location，scale
伽马分布	gamma	shape，scale
威布尔分布	weibull	shape，scale
贝塔分布	beta	shape1，shape2，ncp

注：在这些函数之前加上一些前缀，就可以进行相应的计算. 例如：前缀 d 表示计算概率密度或分布律；前缀 p 表示计算分布函数；前缀 q 表示计算分布函数的反函数(下分位点)；前缀 r 表示生成相同分布的随机数.

11.3.2　样本的数字特征以及相关性检验

如果样本是来自单个正态总体，那么函数 mean( )、sd( )、skewness( )和 kurtosis( )可以分别求得样本的期望、标准差、偏度和峰度. 对于两个正态总体或多个正态总体，我们除了分析各个分量的取值特点外，还需要去研究各个分量之间的相关关系.

例 11.3.1　现有两个样本的数据如下.

表 11.3　两个样本的数据值

样本 1	16.5	17.5	18.5	19.5	20.5	21.5	22.5	23.5	24.5
样本 2	43.5	42.6	42.6	40.6	40.3	38.7	37.2	36.0	34.0

计算样本的均值、方差、协方差和相关系数.

解　采用数据框方式输入数据，用函数 mean( )计算均值，用函数 cov( )计算协方差，用函数 cor( )计算相关系数.

```
>samp<-data.frame(x=c(16.5,17.5,18.5,19.5,20.5,21.5,22.5,23.5,24.5),y=c(43.5,42.6,42.6,40.6,40.3,38.7,37.2,36.0,34.0))
>samp.m<-apply(samp,2,mean);samp.s<-cov(samp);samp.r<-cor(samp)
```

结果为

```
> samp.m
   x       y
20.5    39.5
> samp.s
         x         y
x   7.5000 -8.8125
y   -8.8125 10.6875
> samp.r
              x              y
x     1.0000000   -0.9843067
y     -0.9843067    1.0000000
```

对于协方差的计算，我们还可以使用函数 var( )，计算结果和 cov( )相同.

如果要对数据进行相关性检验，我们可以用命令

```
>cor.test(x,y)   #表示 Pearson 相关性检验
>cor.test(x,y,method="spearman")   #表示 Spearman 秩检验
>cor.test(x,y,method="kendall")   #表示 Kendall 秩检验
```

例 11.3.2　对例 11.3.1 的两组数据进行相关性检验.

解

```
>cor.test(x,y)
        Pearson's product-moment correlation
data:  x and y
t = -14.758,df = 7,p-value = 1.57e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9968126 -0.9245869
sample estimates:
       cor
 -0.9843067
```

其 p 值为 1.57e-06<0.05，拒绝原假设，即认为两个变量之间线性相关.

注：如果是三组数据，R 也只提供了两两分量的相关性检验.

11.3.3 参数估计

在实际问题中，有时我们可以推断总体分布的类型，但是总体分布中的一些参数却是未知的，因此需要对参数进行估计. 参数估计的形式主要包括参数的点估计和区间估计，下面我们列举几个例子介绍 R 如何进行参数估计.

例 11.3.3 已知保险公司的一个保险品种在一个保单年度内的损失情况如下表所示.

表 11.4 某保险品种的损失情况

损失次数	0	1	2	3	4	5
保单数	1 412	125	121	50	35	4

已知分布类型是泊松分布，求参数 λ 的矩估计.

解 因为泊松分布的参数 λ 的矩估计就是其样本均值. 因此可以用如下命令

```
>num<-c(rep(0:5,c(1412,125,121,50,35,4)))  #用函数 rep( )生成 0,1,2,3,4,5 这 6 个数字,且第二个向量表示每个数字的重复次数
>lambda<-mean(num)
> lambda
[1] 0.3875215
```

即参数 λ 的矩估计值为 0.387 521 5.

例 11.3.4 已知 $X \sim N(\mu, \sigma^2)$，试用极大似然估计法估计参数 μ.

解 因为参数 μ 的极大似然估计就是其样本均值. 因此

```
>x<-rnorm(100,0,1)  #生成 100 个标准正态随机数
>mean(x)
[1] -0.09084007
```

即参数 μ 的极大似然估计值为-0.090 840 07.

事实上，对参数进行点估计，有时会非常困难，因为矩方程组或极大似然方程组的解析解一般不容易得到. 因此，为了得到数值解，需要使用到一些函数，如 uniroot( )，multiroot( )，optimize( )，nlm( )和 nlminb( )等.

例 11.3.5 已知 X 服从 Cauchy 分布，其概率密度为

$$f(x;\ \theta) = \frac{1}{\pi[1+(x-\theta)^2]},\ -\infty < x < \infty,$$

其中 θ 为未知参数. X_1，X_2，…，X_n 是来自总体 X 的样本，求 θ 的极大似然估计.

解 因为 *Cauchy* 的似然函数为

$$L(\theta;\ x) = \frac{1}{\pi^n}\prod_{i=1}^{n}\frac{1}{[1+(x_i-\theta)^2]},$$

取对数得

$$\ln L(\theta;\ x) = -n\ln\pi - \sum_{i=1}^{n} \ln[1 + (x_i - \theta)^2],$$

求导数，并令其为 0，得

$$\sum_{i=1}^{n} \frac{x_i - \theta}{1 + (x_i - \theta)^2} = 0.$$

可以看到，上面这个方程的解析解比较难以得到，故考虑其数值解，可以用如下命令.

```
>x<-rcauchy(1 000,1)  #生成 1 000 个随机数
> f<-function(theta) sum((x-theta)/(1+(x-theta)^2))  #定义似然方程
> out=uniroot(f,c(0,5))  #似然方程在区间(0,5)内的根
> out
 $ root
[1] 1.001189
 $ f.root
[1] -4.846258e-06
 $ iter
[1] 6
 $ estim.prec
[1] 6.103516e-05
```

结果显示：极大似然估计值为 1.001 189. 这里 $ f. root 表示函数 f 在近似值处的函数值，$ iter 表示迭代次数，即用了 6 次迭代，$ estim. prec 表示近似值和准确值之间的误差估计.

例 11.3.6　从一批钉子中随机抽取 8 个，测得其重量(单位：g)为

2.14，2.10，2.13，2.12，2.15，2.14，2.12，2.10.

假设钉子的重量服从正态分布 $N(\mu, \sigma^2)$，且 σ^2 未知，求总体均值 μ 的置信度为 90% 的置信区间.

解　由于总体方差未知时，关于均值的置信度为 $1-\alpha$ 的置信区间为

$$\left(\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1),\ \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1)\right).$$

因此，可以使用如下命令.

```
> x<-c(2.14,2.10,2.13,2.12,2.15,2.14,2.12,2.10)
> xb<-mean(x)
> tmp<-sd(x)/sqrt(8) * qt(1-0.05,7)
> a<-xb-tmp;b<-xb+tmp
> a;b
[1] 2.112597
[1] 2.137403
```

即均值 μ 的置信度为 90%的置信区间为(2.112 597，2.137 403).

事实上，对于总体方差已知或未知两种情况均值μ的区间估计有现成的 R 程序. 下面将其列出(程序名：interval_ estimate1. R).

```
interval_estimate1<-function(x,sigma=-1,alpha=0.05){
  n<-length(x); xb<-mean(x)
  if(sigma>=0){
    tmp<-sigma/sqrt(n)*qnorm(1-alpha/2); df<-n
                }
  else{
      tmp<-sd(x)/sqrt(n)*qt(1-alpha/2,n-1); df<-n-1
      }
  data.frame(mean=xb,df=df,a=xb-tmp,b=xb+tmp)
}
```

于是对于这道例题，可以调用这个程序，命令如下.

```
>source("interval_estimate1.R")   #调用函数 interval_estimate1
> x<-c(2.14,2.10,2.13,2.12,2.15,2.14,2.12,2.10)
>interval_estimate1(x,alpha=0.10)   #alpha 是显著性水平,默认值为 0.05
```

得到

```
   mean  df        a        b
1 2.125   7 2.112597 2.137403
```

例 11.3.7 经验表明：60 日龄的雄鼠体重服从正态分布，且标准差$\sigma=2.1$g，今从 X 射线照射处理过的 60 日龄的雄鼠中随机抽取 16 只测得其体重为

20.3，21.5，22.0，19.8，22.5，23.7，25.4，24.3，
23.2，26.8，18.7，21.9，24.4，22.8，26.2，21.4.

现求其体重均值在置信度为 95%的置信区间.

解 当总体方差已知时，关于均值的置信度为 $1-\alpha$ 的置信区间为

$$\left(\bar{X}-\frac{\sigma}{\sqrt{n}}u_{\alpha/2},\ \bar{X}+\frac{\sigma}{\sqrt{n}}u_{\alpha/2}\right).$$

因此，可以使用如下命令.

```
>source("interval_estimate1.R")   #调用函数 interval_estimate1
> x<-c(20.3,21.5,22.0,19.8,22.5,23.7,25.4,24.3,23.2,26.8,18.7,21.9,24.4,22.8,26.2,21.4)
>interval_estimate1(x,sigma=2.1)   #这里的 sigma 是已知的
```

得到

```
    mean  df        a        b
1 22.80625  16 21.77727 23.83523
```

即体重均值在置信度为 95%的置信区间为(21.777 27，23.835 23).

例 11.3.8　某工厂生产的零件长度被认为服从正态分布，现从该产品中随机抽取 5 个，其长度如下(单位：mm).

15.0，15.2，14.9，15.4，15.0.

试求测量误差(即方差 σ^2)的置信度为 95%的置信区间.

解　由于均值未知时方差的置信度为 $1-\alpha$ 的置信区间为

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right).$$

利用现有的或编制好的总体均值已知或未知两种情况方差 σ^2 的区间估计的 R 程序(程序名：interval_ var1.R).

```
interval_var1<-function(x,mu=Inf,alpha=0.05){
  n<-length(x)
  if(mu<Inf){
    S2 <- sum((x-mu)^2)/n; df <- n
  }
  else{
      S2 <- var(x); df <- n-1
  }
  a<-df*S2/qchisq(1-alpha/2,df)
  b<-df*S2/qchisq(alpha/2,df)
  data.frame(var=S2,df=df,a=a,b=b)
}
```

我们可以进行如下操作.

```
>source("interval_var1.R")   #调用函数 interval_var1
>x=c(15.0,15.2,14.9,15.4,15.0)
>interval_var1(x)
```

得到

```
   var  df         a         b
1 0.04   4 0.01435842 0.3302929
```

即测量误差的置信度为 95%的置信区间为(0.014 358 42，0.330 292 9).

11.3.4　假设检验

在实际问题中，许多随机变量都服从正态分布，因此，下面主要介绍正态参数的假设检验.

1. 对均值的假设检验

对于单个正态总体，我们先讨论对于均值的假设检验. 表 11.5 总结了对均值检验的相

关结果以供参考.

表 11.5 单个正态总体均值μ的假设检验

条件	原假设	备择假设	统计量	临界值	拒绝域
σ^2 已知	$H_0: \mu = \mu_0$	$H_1: \mu \neq \mu_0$	$U = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$u_{\alpha/2}$	$\lvert U \rvert > u_{\alpha/2}$
	$H_0: \mu \geqslant \mu_0$	$H_1: \mu < \mu_0$		$-u_\alpha$	$U < -u_\alpha$
	$H_0: \mu \leqslant \mu_0$	$H_1: \mu > \mu_0$		u_α	$U > u_\alpha$
σ^2 未知	$H_0: \mu = \mu_0$	$H_1: \mu \neq \mu_0$	$T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$	$t_{\alpha/2}(n-1)$	$\lvert T \rvert > t_{\alpha/2}(n-1)$
	$H_0: \mu \geqslant \mu_0$	$H_1: \mu < \mu_0$		$-t_\alpha(n-1)$	$T < -t_\alpha(n-1)$
	$H_0: \mu \leqslant \mu_0$	$H_1: \mu > \mu_0$		$t_\alpha(n-1)$	$T > t_\alpha(n-1)$

例 11.3.9 根据以往的经验可知，某鱼塘单位平均产量服从正态分布 $N(500, \sigma^2)$，先随机抽取 10 口鱼塘，测得各鱼塘产量(单位：kg)为

495，510，505，503，520，498，512，516，505，500，

问该地鱼塘产量是否正常($\alpha=0.05$)？

解 由于方差未知，故采用 t 检验，可用如下命令：

```
> x<-c(495,510,505,503,520,498,512,516,505,500)
> t.test(x,mu=500)
```

显示结果为

```
        One Sample t-test
data:  x
t = 2.5246, df = 9, p-value = 0.03252
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
500.6652 512.1348
sample estimates:
mean of x
    506.4
```

由于计算出的 P 值是 0.032 52(<0.05)，故拒绝原假设，即认为鱼塘的产量不正常. 从显示结果还可以看到 t. test( )提供了均值的区间估计，这是因为区间估计和假设检验是对同一个问题的两种不同角度的回答.

注：在 R 中，t. test 的使用格式如下：

```
t.test(x,y=NULL,
alternative=c("two.sided","less","greater"),
mu=0,paried=FALSE,var.equal=FALSE,
conf.level=0.95,…)
```

其中 x，y 是由数据构成的向量(如果只提供 x，则作单个正态总体的均值检验；否则作两个总体的均值检验)；alternative 表示备择假设；two. sided(默认)表示双边检验 H_1：$\mu \neq \mu_0$；less 表示单边检验 H_1：$\mu<\mu_0$；greater 表示单边检验 H_1：$\mu>\mu_0$；mu 表示原假设的 μ_0；conf. level 是置信水平，默认值为 0.95.

例 11.3.10　根据以往的经验可知，某鱼塘单位平均产量服从正态分布 $N(500, \sigma^2)$，先随机抽取 10 口鱼塘，测得各鱼塘产量(单位：kg)为

495，510，505，503，520，498，512，516，505，500.

问是否有理由认为鱼塘平均产量大于 500kg($\alpha=0.10$)？

解

```
> x<-c(495,510,505,503,520,498,512,516,505,500)
> t.test(x,alternative="greater",mu=500,conf.level=0.90)
        One Sample t-test
data:  x
t = 2.5246, df = 9, p-value = 0.01626
alternative hypothesis: true mean is greater than 500
90 percent confidence interval:
 502.8939      Inf
sample estimates:
mean of x
    506.4
```

由于计算出的 P 值是 0.016 26(<0.10)，故拒绝原假设，即认为鱼塘平均产量大于 500kg.

例 11.3.11　根据以往的经验可知，某鱼塘单位平均产量服从正态分布 N(500，4)，先随机抽取 10 口鱼塘，测得各鱼塘产量(单位：kg)为

495，510，505，503，520，498，512，516，505，500

问该地鱼塘产量是否正常($\alpha=0.05$)？

解　由于方差已知，故采用 U 检验，于是利用现成的求 P 值的 R 程序(P_ value. R)

```
P_value<-function(cdf,x,paramet=numeric(0),side=0){
  n<-length(paramet)
  P<-switch(n+1,
      cdf(x),
      cdf(x,paramet),
      cdf(x,paramet[1],paramet[2]),
      cdf(x,paramet[1],paramet[2],paramet[3])
  )
  if(side<0)        P
  else if(side>0)   1-P
  else
      if(P<1/2)     2*P
      else          2*(1-P)
}
```

和正态总体均值检验的 R 程序(mean. test1. R)

```
mean.test1<-function(x,mu=0,sigma=-1,side=0){
  source("P_value.R")
  n<-length(x); xb<-mean(x)
   if(sigma>=0){
       z<-(xb-mu)/(sigma/sqrt(n))
      P<-P_value(pnorm,z,side=side)
      data.frame(mean=xb,df=n,Z=z,P_value=P)
  }
  else{
      t<-(xb-mu)/(sd(x)/sqrt(n))
      P<-P_value(pt,t,paramet=n-1,side=side)
      data.frame(mean=xb,df=n-1,T=t,P_value=P)
  }
}
```

我们可以用如下命令.

```
> x<-c(495,510,505,503,520,498,512,516,505,500)
>source("mean.test1. R")
>mean.test1(x,mu=500,sigma=2)
   mean      df        Z         P_value
1 506. 4    10    10. 11929         0
```

由于 P 值为 0<0. 05，所以拒绝原假设，认为鱼塘产量不正常.

2. 对方差的假设检验

对于单个正态总体，表 11. 6 总结了对方差检验的相关结果以供参考.

与均值检验相同，方差检验也需要用到 P 值，根据 P 值的大小来作出接受或拒绝的判断. 而关于 P 值的计算，参考前面的程序(P_ value. R).

根据这张表，可以编制均值已知或未知时方差检验的 R 程序(var. test1. R).

```
var.test1<-function(x,sigma2=1,mu=Inf,side=0){
    source("P_value.R")
    n<-length(x)
    if(mu<Inf){
        S2<-sum((x-mu)^2)/n; df=n
  }
  else{
        S2<-var(x); df=n-1
  }
  chi2<-df*S2/sigma2;
  P<-P_value(pchisq,chi2,paramet=df,side=side)
  data.frame(var=S2,df=df,chisq2=chi2,P_value=P)
}
```

其中 sigma2 表示 σ_0^2. mu 表示均值，Inf 表示无穷，当均值已知时程序采用自由度为 n 的卡方检验；否则默认采用自由度为 $n-1$ 的卡方检验. side = 0(默认)表示作双边检验，side = −1 表示作左边检验(H_1：$\sigma^2<\sigma_0^2$)，side = 1 表示作右边检验(H_1：$\sigma^2>\sigma_0^2$). 最后程序将输出方差(var)、自由度(df)、统计量(chisq2)和 P 值(P_ value).

表 11.6　单个正态总体方差 σ^2 的假设检验

条件	原假设	备择假设	统计量	临界值	拒绝域
μ 已知	H_0：$\sigma^2=\sigma_0^2$	H_1：$\sigma^2\neq\sigma_0^2$	$\chi^2=\dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2}$	$\chi^2_{1-\frac{\alpha}{2}}(n)$ 和 $\chi^2_{\frac{\alpha}{2}}(n)$	$\chi^2<\chi^2_{1-\frac{\alpha}{2}}(n)$ 或 $\chi^2>\chi^2_{1-\frac{\alpha}{2}}(n)$
	H_0：$\sigma^2\geqslant\sigma_0^2$	H_1：$\sigma^2<\sigma_0^2$		$\chi^2_{1-\alpha}(n)$	$\chi^2<\chi^2_{1-\alpha}(n)$
	H_0：$\sigma^2\leqslant\sigma_0^2$	H_1：$\sigma^2>\sigma_0^2$		$\chi^2_{\alpha}(n)$	$\chi^2>\chi^2_{1-\alpha}(n)$
μ 未知	H_0：$\sigma^2=\sigma_0^2$	H_1：$\sigma^2\neq\sigma_0^2$	$\chi^2=\dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma_0^2}$	$\chi^2_{1-\frac{\alpha}{2}}(n-1)$ 和 $\chi^2_{\frac{\alpha}{2}}(n-1)$	$\chi^2<\chi^2_{1-\frac{\alpha}{2}}(n-1)$ 或 $\chi^2>\chi^2_{1-\frac{\alpha}{2}}(n-1)$
	H_0：$\sigma^2\geqslant\sigma_0^2$	H_1：$\sigma^2<\sigma_0^2$		$\chi^2_{1-\alpha}(n-1)$	$\chi^2<\chi^2_{1-\alpha}(n-1)$
	H_0：$\sigma^2\leqslant\sigma_0^2$	H_1：$\sigma^2>\sigma_0^2$		$\chi^2_{\alpha}(n-1)$	$\chi^2>\chi^2_{1-\alpha}(n-1)$

例 11.3.12　已知某花卉株高服从正态分布 N(52，σ^2)，现改变施肥方案，随机抽取新方案下的 7 株花卉，测得株高(单位：cm)为

53.5，54.0，55.5，53.5，50，52.5，53

问在显著性水平 0.05 下，能否认为该花卉株高的方差 $\sigma^2=2^2$?

解　由于均值已知，作方差检验，因此

```
>x<-c(53.5,54.0,55.5,53.5,50,52.5,53)
> source("var.test1.R")
> var.test1(x,sigma2=4,mu=52)
      var  df   chisq2     P_value
1 3.714286   7     6.5   0.9654464
```

结果显示 P 值为 0.965 446 4>0.05，故接受原假设，认为该花卉株高的方差 $\sigma^2=2^2$.

例 11.3.13　已知在一个混杂的小麦品种，其株高服从正态分布 $N(\mu,14^2)$，经提纯后随机抽取 10 株，测得其株高(单位：cm)为

95，100，105，102，104，100，102，93，96，105

问在显著性水平 0.05 下，能否认为提纯后的群体比原群体整齐?

解　由于均值未知，作方差的单边检验，因此

```
>x<-c(95,100,105,102,104,100,102,93,96,105)
> source("var.test1.R")
> var.test1(x,sigma2=196,side=-1)
      var  df    chisq2      P_value
1 18.17778   9 0.8346939 0.0002666677
```

结果显示 P 值为 0.000 266 667 7<0.05，故拒绝原假设，认为提纯之后的群体比原群体整齐.

对于两个正态总体，如果要进行两均值检验，可以使用函数 t. test( )，用法如前所述；如果要进行方差比的检验和相应的区间估计，可以使用函数 var. test( )，其使用格式为

```
var.test(x,y,ratio=1,
        alternative=c("two.sided","less","greater"),
        conf.level=0.95,…)
```

其中 x，y 是来自两个正态总体的数据(向量)；ratio 是方差比的原假设，默认为 1；two. sided 表示双边检验(H_1：$\sigma_1^2/\sigma_2^2 \neq$ ratio)；less 表示单边检验(H_1：$\sigma_1^2/\sigma_2^2 <$ ratio)；greater 表示单边检验(H_1：$\sigma_1^2/\sigma_2^2 >$ ratio).

另外，R 还可以进行非参数检验. 例如：函数 chisq. test( )表示作 Pearson 拟合优度 χ^2 检验，用于检验当理论分布完全已知时，数据是否服从某某分布；函数 ks. test( )则表示当理论分布依赖于若干个未知参数时，数据是否服从某某分布；函数 fisher. test( )表示用 Fisher 精确概率检验来作独立性检验，常用于 2×2 列联表的检验，尤其是数据不满足卡方检验时(单元期望频数小于 4)；函数 cor. test( )表示进行 Spearman 或 Kendall 秩相关检验，用于检验两个变量是否相关；函数 wilcox. test( )表示作 Wilcoxon 符号秩检验，用于检验一个样本是否来自某个总体或两个总体(需要成对样本)是否存在显著差异.

11.3.5 回归分析

为了寻找自变量和因变量之间的定量关系，常用方法之一便是回归分析. 通过建立回归方程并对其进行检验，不仅可以解释各种变量之间的关系，还可以利用回归方程进行预测和控制.

1. 一元线性回归

下面通过例子来说明如何确定一个因变量 Y 与一个自变量 X 之间的定量关系表达式.

例 11.3.14 现有某种大豆的脂肪含量 X 和蛋白质含量 Y 的检测结果如下.

表 11.7 脂肪含量和蛋白质含量数据表

X	16.5	17.5	18.5	19.5	20.5	21.5	22.5	23.5	24.5
Y	43.5	42.6	42.6	40.6	40.3	38.7	37.2	36.0	34.0

试求 Y 与 X 的回归方程，并对相应的回归方程作检验.

解 利用 R 中的函数 lm( )可以非常方便地求出回归方程，并作相应检验，具体操作如下：

```
>x<-c(16.5,17.5,18.5,19.5,20.5,21.5,22.5,23.5,24.5)
>y<-c(43.5,42.6,42.6,40.6,40.3,38.7,37.2,36.0,34.0)
>lm.sol=lm(y~1+x)   #作线性模型,且模型形式为 y=β0+β1x+ε
>summary(lm.sol)   #提取模型结果
Call:
```

```
lm(formula = y ~ 1 + x)
Residuals:
   Min      1Q  Median      3Q     Max
-0.800 -0.425   0.025   0.375   0.800
Coefficients:
             Estimate  Std.Error t value  Pr(>|t|)
(Intercept)  63.58750  1.64510   38.65    2.02e-09 ***
x            -1.17500   0.07962  -14.76    1.57e-06 ***
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6167 on 7 degrees of freedom
Multiple R-squared:  0.9689,  Adjusted R-squared:  0.9644
F-statistic: 217.8 on 1 and 7 DF,p-value: 1.57e-06
```

由计算结果可知回归方程为

$$y=63.587\ 5-1.175x.$$

从 Pr(>|t|)(P 值)可以看出系数是极为显著的.(“***”表示极为显著,“**”表示高度显著,“*”表示显著,“.”表示不太显著,没有记号表示不显著)

另外,从结果中还可以看出残差的标准差(Residual standard error)、相关系数的平方(Multiple R-squared)和 F 统计量(F-statistic).

例 11.3.15　求例 11.3.14 中 $X=25.5$ 时 Y 的概率为 0.95 的预测区间.

解　在回归分析中,函数 predict( )可以非常方便地求出预测值和预测区间.

```
>new<-data.frame(x=25.5)   #使用数据框形式
>lm.pred<-predict(lm.sol,new,interval="prediction",level=0.95)   #interval="prediction",level=
0.95(默认)表示给出置信度为 0.95 的置信区间.
>lm.pred
      fit      lwr      upr
1 33.625  31.82245  35.42755
```

结果显示在 $X=25.5$ 时,Y 的概率为 0.95 的预测区间为(31.822 45,35.427 55).

使用如下命令可以得到数据的散点图,并可以将得到的回归直线画在散点图上,如图 11.9 所示.

```
> plot(x,y)
> abline(lm.sol)
```

函数 residuals( )表示计算回归方程的残差.

```
>y.res<-residuals(lm.sol);plot(y.res)   #计算残差并画出残差的散点图
```

结果如图 11.10 所示.

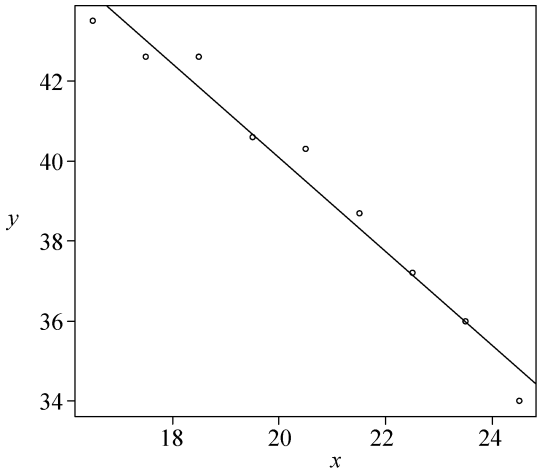


图 11.9 数据的散点图与回归直线

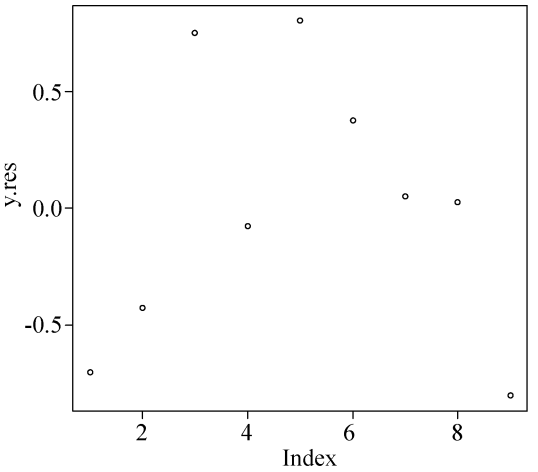


图 11.10 数据残差的散点图

2. 多元线性回归

当影响因变量 Y 的自变量不止一个的时候，我们下面通过例子来说明如何建立简单但是却非常实用的多元线性回归模型.

例 11.3.15 现有 15 个人的血压的收缩压(Y)、体重(X_1)(单位：kg)和年龄(X_2)的数据如下，试建立 Y 关于 X_1 和 X_2 的线性回归方程.

表 11.8 血压、体重、年龄数据表

Y	130	140	125	128	117	125	123	125	155	147	132	123	125	125	155
X_1	75	92	86	83	79	81	60	80	95	93	82	75	80	81	85
X_2	31	35	25	30	21	24	32	25	50	54	36	21	50	23	65

解

```
> bp<-data.frame(x1=c(75,92,86,83,79,81,60,80,95,93,82,75,80,81,85),
                 x2=c(31,35,25,30,21,24,32,25,50,54,36,21,50,23,65),
                 y=c(130,140,125,128,117,125,123,125,155,147,132,123,125,125,155))
> lm.sol<-lm(y~x1+x2,data=bp)
> summary(lm.sol)
Call:
lm(formula = y ~ x1 + x2, data = bp)
Residuals:
     Min       1Q   Median       3Q      Max
-14.6727  -1.6545   0.1112   3.4486   7.4676

Coefficients:
            Estimate  Std.Error  t value  Pr(>|t|)
(Intercept) 68.3161   14.7680    4.626    0.000584 ***
x1          0.5240    0.1955     2.680    0.020042 *
```

```
x2          0.5888      0.1218      4.833      0.000410 ***
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.639 on 12 degrees of freedom
Multiple R-squared:  0.809,    Adjusted R-squared:  0.7771
F-statistic: 25.41 on 2 and 12 DF, p-value: 4.86e-05
```

结果显示回归系数与回归方程的检验都是显著的，因此，Y 关于 X_1 和 X_2 的线性回归方程为

$$\hat{Y}=68.3161+0.5240x_1+0.5888x_2.$$

3. 可线性化的一元回归方程

在实际问题中，有时两个变量之间并非线性关系，但是通过某些变换，可以将其转化为线性关系，下面结合实例说明如何将非线性方程线性化.

例 11.3.16　电容器充电达某电压时为某时间的计算原点，此后电容器串联一电阻放电，测定各时刻的电压 u，测得结果如下：

表 11.9　时间—电压数据表

时间 t(s)	0	1	2	3	4	5	6	7	8	9	10
电压 u(V)	100	75	55	40	30	20	15	10	10	5	5

若 u 与 t 的关系为 $u=u_0e^{-ct}$，其中 u_0 与 c 未知，求 u 对 t 的回归方程.

解　对 $u=u_0e^{-ct}$两端取对数得 $\ln u = \ln u_0 - ct$.

```
>x<-c(0:10)
>y<-c(100,75,55,40,30,20,15,10,10,5,5)
>z<-log(y)
>lm.sol<-lm(z~x)
>summary(lm.sol)
Call:
lm(formula = z ~ 1 + x)
Residuals:
      Min        1Q    Median        3Q       Max
-0.18979 -0.04160  0.01378  0.02917  0.19071
Coefficients:
             Estimate   Std.Error  t value   Pr(>|t|)
(Intercept)  4.61303    0.06195    74.46     7.19e-14 ***
x           -0.31264    0.01047   -29.86     2.59e-10 ***
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1098 on 9 degrees of freedom
Multiple R-squared:  0.99,     Adjusted R-squared:  0.9889
F-statistic: 891.4 on 1 and 9 DF, p-value: 2.592e-10
```

从计算结果中可得 $u_0=e^{4.613\,03}$，$c=0.312\,64$，因此，u 对 t 的回归方程为

$$u=100.7861e^{-0.312\,64t}.$$

除了上述例子外，还有一些非线性方程，也可以通过适当的变量替换将其转化为线性回归方程，从而确定未知参数. 下表给出了一些常用函数的转换公式.

表 11.10 常用曲线的转换公式

函数名称及表达式	转换公式	回归方程
双曲线 $\frac{1}{y}=a+\frac{b}{x}$	$y'=\frac{1}{y}$，$x'=\frac{1}{x}$	$y'=a+bx'$
幂函数曲线 $y=ax^b$	$y'=\ln y$，$x'=\ln x$，$a'=\ln a$	$y'=a+bx'$
负指数函数曲线 $y=ae^{\frac{b}{x}}$	$y'=\ln y$，$x'=\frac{1}{x}$，$a'=\ln a$	$y'=a+bx'$
对数曲线 $y=a+b\ln x$	$x'=\ln x$	$y'=a+bx'$
Logistic 曲线 $y=\frac{K}{1+Ae^{-\lambda x}}$	$y'=\ln\left(\frac{y}{K-y}\right)$，$a=-\ln A$	$y'=a+\lambda x'$

4. 非线性回归

例 11.3.17 已知某产品的广告费 X(百万元)与销售量 Y(百万支)的数据如下：

表 11.11 广告费与销售量数据

X	5.50	5.75	6.00	6.25	6.50	6.75	7.00	7.25	7.50	7.75	8.00	8.25	8.50
Y	7.45	7.55	7.75	8.00	8.30	8.50	9.00	9.8	10.80	11.75	12.6	13.70	14.65

解 先画出数据的散点图，由散点图可以看出，用二次曲线进行拟合较好.

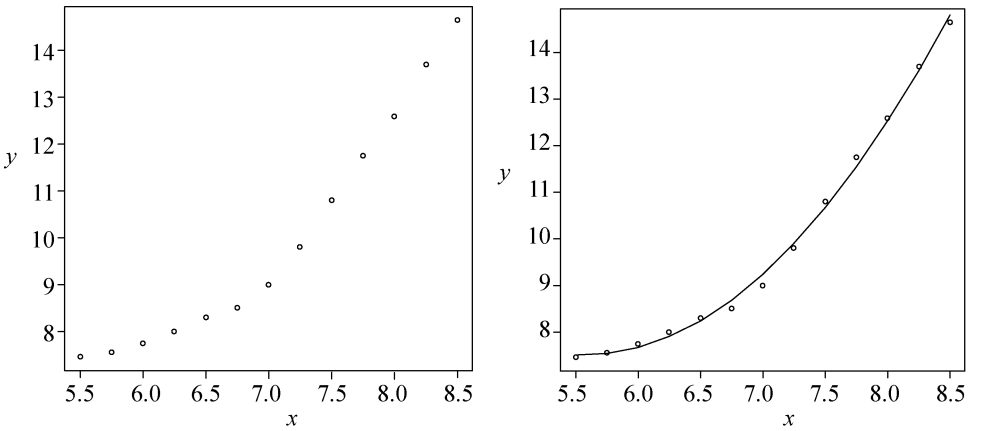


图 11.11 广告费和销售量的散点图和拟合曲线

```
> x<-seq(5.5,8.5,by=0.25)
> y<-c(7.45,7.55,7.75,8.00,8.30,8.5,9.00,9.8,10.80,11.75,12.60,13.70,14.65)
```

```
> plot(x,y)
>lm.sol<-lm(y~x+I(x^2))  #进行二次曲线回归
>z<-predict(lm.sol,data.frame(x))
>lines(x,z)  #绘制回归曲线
>summary(lm.sol)
Call:
lm(formula = y ~ x + I(x^2))
Residuals:
     Min       1Q   Median       3Q      Max
-0.24021 -0.10300  0.06469  0.09201  0.20075
Coefficients:
            Estimate   Std.Error  t value  Pr(>|t|)
(Intercept) 34.08077   2.60681    13.07    1.30e-07 ***
x           -9.53467   0.75411    -12.64   1.78e-07 ***
I(x^2)      0.85514    0.05377    15.90    1.99e-08 ***
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1504 on 10 degrees of freedom
Multiple R-squared:  0.9969,  Adjusted R-squared:  0.9963
F-statistic:  1621 on 2 and 10 DF,p-value: 2.751e-13
```

计算结果通过回归系数检验和回归方程检验，由此得到销售量和广告费之间的关系为

$$y=34.080\ 77-9.534\ 67x+0.855\ 14x^2.$$

11.3.6　方差分析

在实际生产和科学研究中，影响产品质量或产量的因素一般很多，例如：影响水稻产量的因素就有种子品种、肥料品种、施肥量、天气、土壤，等等，然而不同因素的影响大小不等，为了找出影响试验结果的主要因素，就要先做试验，然后对试验结果进行统计分析并进行推断，其中一种重要的统计推断的方法就是方差分析，用来研究一种或多种因素的变化对试验结果的观测值是否有显著影响.

1. 单因素方差分析

例 11.3.18　为了比较 4 个小麦品种对产量的影响，在一片土壤和气候相近的土地上进行了 24 组试验，试验数据见表 11.12，试比较 4 种小麦品种的产量有无显著性差异.

表 11.12　小麦产量

品种	产量					
*A*1	64	72	68	77	56	95
*A*2	91	78	82	97	77	85
*A*3	93	71	78	75	76	63
*A*4	77	55	66	49	70	55

解 先用数据框输入数据，然后调用函数 aov( )进行方差分析，最后用函数 summary( )提取分析结果.

```
>yield<-data.frame(X=c(64,72,68,77,56,95,91,78,82,97,77,85,93,71,78,75,76,63,77,55,66,49,
70,55),A=factor(rep(1:4,c(6,6,6,6))))
>yield.aov<-aov(X~A,data=yield)
> summary(yield.aov)
            Df  Sum Sq Mean Sq  F value  Pr(>F)
A            3  1636    545.5    4.845  0.0108 *
Residuals   20    2252  112.6
```

其中 Df 表示自由度；Sum Sq 表示平方和；Mean Sq 表示均方和；F value 表示 F 值；Pr(>F)表示 P 值；A 表示因素；Residuals 表示残差(误差).

从计算结果可以看出，P 值小于 0.05，故拒绝原假设，即认为 4 种不同小麦品种的产量有显著的差异.

虽然我们知道 4 个品种之间有显著性差异，但是这并不意味着所有均值间都有差异，有时我们还需要对每一对均值作一对一的比较，即多重比较. 在 R 中，利用函数 pairwise.test( )可以得到多重比较的 P 值.

例 11.3.19 进一步检验

$$H_0:\ \mu_i=\mu_j,\ i=j=1,\ 2,\ 3,\ 4.$$

解

```
>attach(yield)
>mu<-c(mean(X[A==1]),mean(X[A==2]),mean(X[A==3]),mean(X[A==4]));mu
[1] 72 85 76 62
>pairwise.t.test(X,A,p.adjust.method="none")   #计算出的 P 值没有作任何调整,也可选择 p.adjust.
method="holm"或 p.adjust.method="bonferroni"进行调整以此克服多重 t 检验法的缺点(多次重复使用 t 检
验,可能会增大第一类错误的概率,使得结论不一定可靠).
        Pairwise comparisons using t tests with pooled SD
data:  x and A
    1       2       3
2 0.0465  -       -
3 0.5213  0.1574  -
4 0.1183  0.0012  0.0334
P value adjustment method: none
```

从上面计算结果可以看出，μ_1 与 μ_2，μ_2 与 μ_4，μ_3 与 μ_4 均有显著性差异，而 μ_1 与 μ_3，μ_1 与 μ_4，μ_2 与 μ_3 没有显著性差异.

2. 双因素方差分析

例 11.3.20 为了比较 4 种不同的种子和 3 种不同的施肥方法对水稻产量的影响，测得数据如下.

表 11.13　不同施肥方法、不同种子的水稻产量

	*B*1	*B*2	*B*3
*A*1	51	56	45
*A*2	52	57	49
*A*3	52	58	47
*A*4	46	53	35

试对上述数据作双因素方差分析，确定种子与施肥方法对水稻产量有无显著性影响.

解

```
>result<-data.frame(Y=c(51,56,45,52,57,49,52,58,47,46,53,35),A=gl(4,3),B=gl(3,1,12))
>result.aov<-aov(Y~A+B,data=result)
> summary(result.aov)
            Df  Sum Sq  Mean Sq  F value  Pr(>F)
A            3  124.2   41.42    8.148    0.015457 *
B            2  288.2   144.08   28.344   0.000877 ***
Residuals    6  30.5    5.08
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

根据 *P* 值，可以发现不同品种和不同施肥方法对产量均有显著性影响.

事实上，上述例子我们考虑的是无交互作用的方差分析，若两个变量之间存在着交互作用，就要进行有交互作用的方差分析.

例 11.3.21　考察食品喷雾干燥过程中进风速度 *A* 与进料速度 *B* 两个因素对出粉率的影响，因素 *A* 和因素 *B* 各取 4 种水平，在各水平组合上均做两次试验，得到的试验结果见下表，试对试验结果进行方差分析.

表 11.14　食品喷雾干燥试验数据

	*B*1		*B*2		*B*3		*B*4	
*A*1	71	73	72	73	75	73	77	75
*A*2	73	75	76	74	78	77	74	74
*A*3	76	73	79	77	74	75	74	73
*A*4	75	73	73	72	70	71	69	69

解

```
>food<-data.frame(Y=c(71,73,72,73,75,73,77,75,73,75,76,74,78,77,74,74,76,73,79,77,74,75,
74,73,75,73,73,72,70,71,69,69),A=gl(4,8),B=gl(4,2,32))
> food.aov<-aov(Y~A+B+A:B,data=food)
> summary(food.aov)
            Df  Sum Sq  Mean Sq  F value  Pr(>F)
A            3  70.59   23.531   17.512   2.62e-05 ***
```

```
B           3   8.59   2.865   2.132   0.136299
A:B         9   79.53  8.837   6.576   0.000591 ***
Residuals   16  21.50  1.344
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

结果显示因素 A 对试验结果的影响是显著的，而因素 B 对试验结果的影响并不显著，它们之间的交互作用对试验结果的影响是显著的.

3. 正交试验设计

前面介绍了单因素和双因素方差分析，但在实际问题中还有三因素或更多因素的试验，这时会遇到试验次数太多的问题，有时多到让人无法忍受，因此要选择合理的试验设计方案，使得试验次数不多，但是能够得到比较令人满意的结果. 正交试验设计就是其中一种非常重要的试验设计方案，利用一套现成的规格化的表格——正交表来安排多因素试验. 现在正交试验设计方法已经在农业、工业、医药等领域都有了广泛的应用，它已经成为许多科研工作者所需要掌握的必备知识.

例 11.3.22 为提高水稻的产量和质量，提高栽培技术，选用不同品种、密度、施肥量做试验，进行了 3 个因素 3 个水平的正交试验，各因素及其水平见表 11.15.

表 11.15 正交试验的因素与水平表

因素	水平 1	水平 2	水平 3
A：品种	选系 5 号	选系 7 号	广陆矮 4 号
B：密度	25 万株/亩	20 万株/亩	15 万株/亩
C：纯氮用量	5 斤/亩	10 斤/亩	15 斤/亩

如果要进行全面试验，需要进行 27 次试验，但是选用正交表 $L_9(3^4)$，仅须做 9 次试验，试验方案和结果列于下表.

表 11.16 水稻的试验结果

试验号	A：品种	B：密度	C：纯氮用量	试验结果
1	1(选系 5 号)	1(25)	1(5)	815
2	1(选系 5 号)	2(20)	2(10)	908
3	1(选系 5 号)	3(15)	3(15)	932
4	2(选系 7 号)	1(25)	2(10)	883
5	2(选系 7 号)	2(20)	3(15)	980
6	2(选系 7 号)	3(15)	1(5)	790
7	3(广陆矮 4 号)	1(25)	3(15)	1050
8	3(广陆矮 4 号)	2(20)	1(5)	885
9	3(广陆矮 4 号)	3(15)	2(10)	1004

问：较好的方案是什么？并对正交试验进行方差分析.

解

```
>rice<-data.frame(A=gl(3,3),B=gl(3,1,9),C=factor(c(1,2,3,2,3,1,3,1,2)),Y=c(815,908,932,
883,980,790,1050,885,1004))
> K<-matrix(0,nrow=3,ncol=3,dimnames=list(1:3,c("A","B","C")))
> for(j in 1:3)
    for(i in 1:3)
    K[i,j]=mean(rice$Y[rice[j]==i])
> K
      A        B        C
1 885.0000 916.0000 830.0000
2 884.3333 924.3333 931.6667
3 979.6667 908.6667 987.3333
> plot(as.vector(K),axes=F,xlab="Level",ylab="yield")
> xmark=c(NA,"A1","A2","A3","B1","B2","B3","C1","C2","C3",NA)
> axis(1,0:10,labels=xmark)
> axis(2,10*75:105)
> lines(K[,"A"]);lines(4:6,K[,"B"]);lines(7:9,K[,"C"])
```

图形如图 11.12 所示.

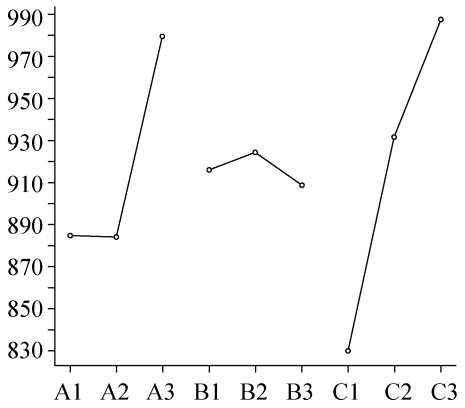


图 11.12　三因素与指标关系

从图中可以看出：广陆矮 4 号的产量最高；密度为 20 万株/亩时产量最高；纯氮用量为 15 斤/亩时产量最高. 综合起来 $A3B2C3$ 可能是较好的工艺条件. 但是，我们发现这个工艺条件不在我们的 9 次试验中，它是否正确还要通过实践来检验. 因此需要对 $A3B2C3$ 再做一次试验，并与最好的试验 $A3B1C3$ 进行比较，从而说明选出的工艺是最好的.

下面对正交试验进行方差分析.

```
>rice.aov<-aov(Y~A+B+C,data=rice)
> summary(rice.aov)
            Df  Sum Sq  Mean Sq  F value Pr(>F)
A            2  18051     9025    11.047  0.083 .
```

```
B             2    369     184    0.226   0.816
C             2  38189   19094   23.371 0.041 *
Residuals     2   1634     817
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

从计算结果可以看到因素 C 对于产量的影响是显著的，所以因素 C 的选取很重要；因素 A 和 B 对产量的影响都不显著.

参考答案

习 题 一

1. (1) 以 n 表示该班的学生人数，样本空间为 $\Omega=\left\{\frac{i}{n}\mid i=0,\ 1,\ 2,\ \cdots,\ 100n\right\}$.

(2) 样本空间为 $\Omega=\{10,\ 11,\ 12,\ \cdots\}$.

(3) 采用 0 表示检查到一个次品，以 1 表示检查到一个正品，样本空间为

$\Omega=\{00,\ 100,\ 0100,\ 0101,\ 0110,\ 1100,\ 1010,\ 1011,\ 0111,\ 1101,\ 1110,\ 1111\}$.

(4) 取直角坐标系，则有 $\Omega=\{(x,\ y)\mid x^2+y^2<1\}$，若取极坐标系，则有

$$\Omega=\{(\rho,\ \theta)\mid 0\leqslant\rho<1,\ 0\leqslant\theta<2\pi\}.$$

2. (1) $\overline{A}\overline{B}\overline{C}$ 或 $A-B-C$ 或 $A-(B\cup C)$；

(2) $\overline{A}\overline{B}\overline{C}\cup\overline{A}\overline{B}C\cup\overline{A}B\overline{C}$；

(3) $A\cup B\cup C$ 或 $\overline{A}\overline{B}\overline{C}\cup\overline{A}\overline{B}C\cup\overline{A}B\overline{C}\cup\overline{A}BC\cup A\overline{B}\overline{C}\cup A\overline{B}C\cup AB\overline{C}\cup ABC$；

(4) $\overline{A}\overline{B}C\cup\overline{A}B\overline{C}\cup A\overline{B}C$；

(5) $AB\cup AC\cup BC$ 或 $\overline{A}BC\cup A\overline{B}C\cup AB\overline{C}\cup ABC$；(6) $\overline{A}\overline{B}\overline{C}\cup\overline{A}\overline{B}C\cup\overline{A}B\overline{C}\cup A\overline{B}\overline{C}$.

3. (1) $AB=\{x\mid 0.8<x\leqslant 1\}$；

(2) $A-B=\{x\mid 0.5\leqslant x\leqslant 0.8\}$；

(3) $\overline{A-B}=\{x\mid 0\leqslant x<0.5$ 或 $0.8<x\leqslant 2\}$；

(4) $\overline{A\cup B}=\{x\mid 0\leqslant x<0.5$ 或 $1.6<x\leqslant 2\}$.

4. $p=-3+\sqrt{11}$.

5. (1) 0；(2) 0.3；(3) 0.2.

6. $P(B)=1-p$.

7. $P(A\cup B\cup C)=0.9$.

8. 以 A_i，$i=1$，2，3 表示事件“杯子中球的最大个数为 i”，则

$$P(A_1)=\frac{6}{16},\ P(A_2)=\frac{9}{16},\ P(A_3)=\frac{1}{16}.$$

9. $\frac{41}{90}$.

10. (1) $\frac{2}{5}$；(2) $\frac{1}{10}$；(3) $\frac{7}{10}$；(4) $\frac{3}{10}$；(5) $\frac{1}{5}$.

11. $\frac{2}{5}$.

12. $\frac{A_9^7}{9^7}$.

13. $\frac{1}{6}$.

14. $\frac{15}{64}$.

15. (1)0.988；(2)0.8285.

16. $P(A\cup B)=0.7$.

17. 设 A_i 表示事件“第 i 次取得合格品”，则 $P(\bar{A}_1\bar{A}_2A_3)\approx 0.00835$.

18. 设从第一个袋子摸出黑球 A，从第二个袋中摸出黑球为 B，则

$$P(B)=P(B\mid A)P(A)+P(B\mid \bar{A})P(\bar{A})=\frac{a}{a+b}$$

19. 设 C 表示“机床停机”，则 $P(C)=0.367$.

20. 设甲、乙、丙分别抽到难签的事件为 A，B，C，则 $P(A)=P(B)=P(C)=\frac{4}{10}$.

21. 设 A 表示“零件由甲生产”，B 表示“零件是次品”，则 $P(A\mid B)=0.2$.

22. $\frac{1}{2}$

23. 0.684

24. 0.6

25. 0.09

26. 0.458

28. 0.089.

29. 0.8629.

30. KL 通达的概率为 $p^3(3-p-2p^2+p^3)$；KR 通达的概率为 $p^2(2+2p-5p^2+2p^3)$.

习　题　二

1. 随机变量 X 的所有可能取值为 1，2，3，4，5，6，分布律为

X	1	2	3	4	5	6
P_k	$\frac{11}{36}$	$\frac{9}{36}$	$\frac{7}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{1}{36}$

2. (1) $\frac{1}{3}$；(2) $\frac{1}{4}$.

3. X 的分布律为

X	0	1	2
$\frac{22}{35}$	$\frac{22}{35}$	$\frac{12}{35}$	$\frac{1}{35}$

的分布函数为 $F(x)=\begin{cases}0, & x<0,\\ \frac{22}{35}, & 0\leqslant x<1,\\ \frac{34}{35}, & 1\leqslant x<2,\\ 1, & x\geqslant 2.\end{cases}$

4. $e-1$.

5. (1)0.072 9；(2)0.008 56；(3)0.999 54；(4)0.409 51.

6. (1)0.321；(2)0.243.

7. (1) $\frac{1}{70}$；(2) 猜对 3 次的概率约为 3×10^{-4}，这个概率很小，根据实际推断原理，可以认为他确有区分能力.

8. (1) $e^{-\frac{3}{2}}$；(2) $1-e^{-\frac{5}{2}}$.

9. (1) 至少配备 4 人；(2) 约为 0.017 5；(3) 约为 0.014 4.

10. $\frac{20}{27}$.

11. 0.2.

12. (1) $\ln2\approx0.693\ 15$，1，$\ln1.25\approx0.223\ 14$；(2) $f(x)=\begin{cases}x^{-1},1<x<e,\\0,\quad \text{其他}.\end{cases}$

13. (1) $a=1$，$b=-1$；(2)$f(x)=\begin{cases}xe^{-\frac{x^2}{2}},x\geqslant0,\\0,\quad x<0.\end{cases}$ (3) 0.25.

14. (1) $F(x)=\begin{cases}0, & x<1,\\ 2x+\frac{2}{x}-4, & 1\leqslant x<2,\\ 1, & x\geqslant2;\end{cases}$ (2) $F(x)=\begin{cases}0, & x<0,\\ \frac{x^2}{2}, & 0\leqslant x<1,\\ -\frac{x^2}{2}+2x-1, & 1\leqslant x<2,\\ 1, & x\geqslant2.\end{cases}$

15. $F_T(t)=\begin{cases}1-e^{-t/241},t\geqslant0,\\0,\quad \text{其他};\end{cases}$ $P\{50<T<100\}=e^{-\frac{50}{241}}-e^{-\frac{100}{241}}$.

16. 0.954 7.

17. $Y\sim B(5,\ e^{-2})$，即 $P\{Y=k\}=C_5^k e^{-2k}(1-e^{-2})^{5-k}$，$k=0$，1，2，3，4，5；$P\{Y\geqslant1\}\approx0.516\ 7$.

18. (1) 0.532 8，0.999 6，0.697 7，0.5；(2) $c=3$；(3) $d\leqslant0.42$.

19. 应允许 σ 最大为 31.25.

20. 129.8.

21. 0.682.

22. 184 厘米.

23. (1)

Y	0	π^2	$4\pi^2$
q_i	0.2	0.7	0.1

(2)

Y	-1	1
p_i	0.7	0.3

24.（1）

Y	-1	1	2
q_i	0.3	0.5	0.2

（2）

Y	1	2
p_i	0.8	0.2

25.（1）$f_Y(y)=\dfrac{1}{2\sqrt{2\pi}}e^{-(y+1)^2/8}$，$-\infty<y<+\infty$；

（2）$f_Y(y)=\begin{cases}\dfrac{1}{\sqrt{2\pi}y}e^{-(\ln y)^2/2}, & y>0,\\ 0, & y\leqslant 0;\end{cases}$

（3）$f_Y(y)=\begin{cases}\dfrac{1}{\sqrt{2\pi}y}e^{-y/2}, & y>0,\\ 0, & y\leqslant 0.\end{cases}$

26.（1）$f_Y(y)=\begin{cases}\dfrac{1}{2\pi}e^{y/2}, & -\infty<y\leqslant 2\ln\pi,\\ 0, & 2\ln\pi<y<+\infty;\end{cases}$

（2）$f_Y(y)=\begin{cases}\dfrac{1}{\pi\sqrt{1-y^2}}, & -1<y<1,\\ 0, & \text{其他};\end{cases}$

（3）$f_Y(y)=\begin{cases}\dfrac{2}{\pi\sqrt{1-y^2}}, & 0<y<1,\\ 0, & \text{其他}.\end{cases}$

习　题　三

1. $\dfrac{3}{128}$.

2.（1）有放回摸取时的分布律为

X \ Y	0	1
0	$\frac{9}{25}$	$\frac{6}{25}$
1	$\frac{6}{25}$	$\frac{4}{25}$

（2）无放回摸取时的分布律为

X \ Y	0	1
0	$\frac{3}{10}$	$\frac{3}{10}$
1	$\frac{3}{10}$	$\frac{1}{10}$

3.（1）有放回摸取时，(X, Y)的边缘分布律为

X \ Y	0	1	$p_{i\cdot}$
0	$\frac{9}{25}$	$\frac{6}{25}$	$\frac{3}{5}$
1	$\frac{6}{25}$	$\frac{4}{25}$	$\frac{2}{5}$
$p_{\cdot j}$	$\frac{3}{5}$	$\frac{2}{5}$	

（2）无放回摸取时，(X, Y)的边缘分布律为

X \ Y	0	1	$p_{i\cdot}$
0	$\frac{3}{10}$	$\frac{3}{10}$	$\frac{3}{5}$
1	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{2}{5}$
$p_{\cdot j}$	$\frac{3}{5}$	$\frac{2}{5}$	

此结果说明不同的联合分布律可以确定相同的边缘分布律，因此边缘分布不能唯一确定联合分布.

4.（1）(X, Y)的联合分布律为

X \ Y	0	1
−1	$\frac{1}{2}$	0
0	$\frac{1}{3}$	$\frac{1}{6}$

(2)$(X,\ Y)$的分布函数为

$$F(x,\ y)=\begin{cases}0,\ x<-1\text{ 或 }y<0,\\ \dfrac{1}{2},-1\leqslant x<0,\ y\geqslant 0,\\ \dfrac{5}{6},x\geqslant 0,\ 0\leqslant y<1,\\ 1,\ x\geqslant 0,\ y\geqslant 1.\end{cases}$$

5. $(X,\ Y)$的联合分布律为

X \ Y	$-\frac{1}{2}$	1	3
-2	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$
-1	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$
0	$\frac{1}{24}$	$\frac{1}{48}$	$\frac{1}{48}$
$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$

6. (1)X 的分布函数为

X	1	2	3	4
P	0	1	0	0

(2) Y 的分布函数为

Y	1	2	3	4
P	0	$\frac{1}{2}$	0	$\frac{1}{2}$

7. (1)$\dfrac{1}{9}$; (2)$\dfrac{5}{12}$; (3)$\dfrac{8}{27}$.

8. (1) $F(x,\ y)=\begin{cases}(1-\mathrm{e}^{-2x})(1-\mathrm{e}^{-Y}),x>0,\ y>0,\\ 0,\qquad\qquad\text{其他;}\end{cases}$ (2) $\dfrac{1}{3}$.

9. $\dfrac{a^2}{1+a^2}$.

10. (1)$f(x,\ y)=\begin{cases}4,(x,\ y)\in B,\\ 0,\text{其他;}\end{cases}$

(2) $F(x,y)=\begin{cases}0, & x<-\frac{1}{2}\text{或 } y<0,\\ y(4x+2-y), & -\frac{1}{2}\leqslant x<0,\ 0\leqslant y<2x+1,\\ y(2-y), & x\geqslant 0,\ 0\leqslant y<1,\\ (2x+1)^2, & -\frac{1}{2}\leqslant x<0,\ y\geqslant 2x+1,\\ 1, & x\geqslant 0,\ y\geqslant 1.\end{cases}$

11. $f_X(x)=\begin{cases}4(2x+1), & -\frac{1}{2}\leqslant x<0,\\ 0, & \text{其他};\end{cases}$ $f_Y(y)=\begin{cases}2(1-y), & 0\leqslant y<1,\\ 0, & \text{其他}.\end{cases}$

12. $f_X(x)=\begin{cases}\frac{x}{2}, & 0\leqslant x\leqslant 2,\\ 0, & \text{其他};\end{cases}$ $f_Y(y)=\begin{cases}3y^2, & 0\leqslant y\leqslant 1,\\ 0, & \text{其他}.\end{cases}$

13. $f_X(x)=\begin{cases}2.4x^2(2-x), & 0\leqslant x\leqslant 1,\\ 0, & \text{其他}.\end{cases}$ $f_Y(y)=\begin{cases}2.4y(3-4y+y^2), & 0\leqslant y\leqslant 1,\\ 0, & \text{其他}.\end{cases}$

14. $f_{Y|X}(y\mid x)=\begin{cases}\frac{1}{2(1-x)}, & 0\leqslant y<2(1-x),\\ 0, & \text{其他}.\end{cases}$

15. $f_{X|Y}(x\mid y)=\frac{6x^2+2xy}{2+y}$, $f_{Y|X}(y\mid x)=\frac{3x+y}{6x+2}$, $0\leqslant x\leqslant 1$, $0\leqslant y\leqslant 2$.

$$P\{Y<\frac{1}{2}\mid X=\frac{1}{2}\}=\frac{7}{40}.$$

16. (1) X 和 Y 相互独立；(2) X 和 Y 不相互独立.

17. $a=\frac{2}{9}$；$b=\frac{1}{9}$.

18. 习题 12 中的 X 和 Y 相互独立；习题 13 中的 X 和 Y 不相互独立.

19. $\frac{13}{24}\approx 0.541\ 7$.

20. 相互独立.

21. $F_X(x)=F(x,\infty)=\begin{cases}1-e^{-x}, & x\geqslant 0,\\ 0, & x<0;\end{cases}$

$F_Y(y)=F(\infty,y)=\begin{cases}1-e^{-y}, & y\geqslant 0,\\ 0, & y<0.\end{cases}$

因为 $F(x,y)=F_X(x)F_Y(y)$，所以 X 与 Y 相互独立.

22. $f_Z(z)=\begin{cases}1-e^{-z}, & 0\leqslant y\leqslant 0,\\ (e-1)e^{-z}, & z>1.\\ 0, & \text{其他}.\end{cases}$

23. $f_Z(z)=\begin{cases}\dfrac{1}{2\sigma^2}e^{\frac{z}{2\sigma^2}}, z>0,\\ 0, \quad z\leqslant 1.\end{cases}$

24. $f_Z(z)=\begin{cases}4ze^{-2z}, z>0,\\ 0, \quad z\leqslant 0.\end{cases}$

25. $f_R(r)=\begin{cases}\dfrac{1}{15\ 000}(600r-60r^2+r^3), 0\leqslant r<10,\\ \dfrac{1}{15\ 000}(20-r)^3, \quad \text{当 } 0\leqslant r<20,\\ 0, \quad \text{其他.}\end{cases}$

习　题　四

1. $E(X)=1$；$E(X^2+2)=3.5$；$D(X)=0.5$.

2. $E(X)=\dfrac{81}{64}$；$D(X)=\dfrac{1695}{64^2}$.

3. $E(X)=\dfrac{1}{3}$；$D(X)=\dfrac{1}{18}$.

4. $E(X)=0$；$D(X)=\dfrac{1}{6}$.

5. $E(X^2)=16.24$.

6. $E(3X-2)=4$.

7. 5.208 96.

8. $\dfrac{\pi}{12}(b^2+ab+a^2)$.

9. (1) $E(Y)=E(2X)=2$；(2) $E(Y)=E(e^{-2X})=\dfrac{1}{3}$.

10. $E(X)=\dfrac{7}{6}$；$E(Y)=\dfrac{7}{6}$；$E(XY)=\dfrac{4}{3}$；$E(X^2+Y^2)=\dfrac{10}{3}$.

11. (1) $E(X+Y)=\dfrac{3}{4}$，$E(2X-3Y^2)=\dfrac{5}{8}$；(2) $E(XY)=\dfrac{1}{8}$，$D(X+Y)=\dfrac{5}{16}$.

12. 随机变量 $Z=2X-Y+3$ 的概率密度为

$$f(z)=\frac{1}{\sqrt{2\pi}\cdot 3}e^{-\frac{(z-5)^2}{2\cdot 9}}=\frac{1}{3\sqrt{2\pi}}e^{-\frac{(z-5)^2}{18}}.$$

13. 所求期望值为 $10\left(1-\dfrac{p}{10}\right)^{10}$.

14. 所求期望值为 35.

15. $E(X)=0.2$，$E(Y)=0.6$，$Cov(X, Y)=0$.

16. $E(X)=\dfrac{2}{3}$，$E(Y)=0$，$Cov(X, Y)=0$.

17. (1) $E(X)=0$，$D(X)=2$；(2) $Cov(X, |X|)=0$，X 与 $|X|$ 不相关；(3) X 与 $|X|$ 不

相互独立.

18. $n=6$，$p=0.4$.

19. $E(X)=\dfrac{1}{p}$；$D(X)=\dfrac{1-p}{p^2}$.

20. $E(Y)=0$，$D(Y)=1$.

21. $E(Y^2)=5$.

22. (1)(X_1, X_2)的所有可能取值为(0，0)，(0，1)，(1，0)，(1，1)，且

$$P(X_1=0, X_2=0)=1-e^{-1},\ P(X_1=0, X_2=1)=0,$$
$$P(X_1=1, X_2=0)=e^{-1}-e^{-2},\ P(X_1=1, X_2=1)=e^{-2}.$$

(2) $E(X_1+X_2)=e^{-1}+e^{-2}$.

23. $E[\,|X-Y|\,]=\dfrac{2}{\sqrt{2\pi}}$.

24. (1)求 $E(Z)=\dfrac{1}{3}$，$D(Z)=3$；(2)X 和 Z 的相关系数为 0.

25. (1)(X, Y)的分布律为

X \ Y	0	1
0	$\frac{2}{3}$	$\frac{1}{12}$
1	$\frac{1}{6}$	$\frac{1}{12}$

(2) X 和 Y 的相关系数为 $\rho_{XY}=\dfrac{1}{\sqrt{15}}$.

26. X 和 Y 的相关系数为 $\rho_{XY}=-1$.

习　题　五

1. $P\{\,|X-\mu|<3\sigma\}\geqslant\dfrac{8}{9}$.

2. 设 X 为晚上开着的路灯数，则 $X\sim B(20\ 000, 0.6)$，由切比雪夫不等式有

$$P\{11\ 000<X<13\ 000\}=P\{\,|X-12\ 000|<1\ 000\}\geqslant 0.995\ 2.$$

3. $n\geqslant 18\ 750$.

4. 假设 X 表示任取 10 000 件产品中合格品的数量，由中心极限定理，

$$P\{5\ 980<X<6\ 020\}\approx 2\Phi\left(\frac{20}{\sqrt{2\ 400}}\right)-1=0.318\ 2.$$

5. (1)由中心极限定理，保险公司一年获利不少于 240 000 元的概率近似等于 0.958；
(2) 同理，保险公司亏本的概率近似等于 0.

6. (1) 根据中心极限定理，所求概率近似等于 0.180 2；

（2）最多可有 $n=443$ 个数相加使得误差总和的绝对值小于 10 的概率不小于 0.90.

7. 根据中心极限定理，所求概率近似等于 0.876 4.

8. 根据中心极限定理，所求概率近似等于 0.006 21.

9. 由切比雪夫不等式，需要掷 250 次；由棣莫弗-拉普拉斯定理，仅须掷 68 次.

10.（1）根据中心极限定理 $P\{\overline{X}<2\}\approx 0.151\ 5$；

（2）同理可得所求概率近似等于 0.076 4.

11. 治愈率为 80%的概率近似等于 0.008 9，因此假定不可靠.

12. 根据中心极限定理，需要的车位数大约为 $n=254$.

13. 根据棣莫弗-拉普拉斯定理，每个戏院大约应设 $n=537$ 个座位.

习　题　六

1. 样本均值 98.44，样本方差 2.993.

2. 略.

3.（1）1；（2）0.089.

4.（1）$\chi^2(1)$；（2）$F(1,\ n-1)$.

5. 提示：概率密度函数为奇函数.

6.（1）0.1；（2）26.10.

7. 41.

习　题　七

1.（156.9，43.7），（10.1，8.0）.

2. $\hat{\lambda}=\dfrac{1}{\overline{X}}$ 或 $\hat{\lambda}=\sqrt{\dfrac{n}{\sum\limits_{i=1}^{n}(X_i-\overline{X})^2}}$.

3. 矩估计 $\hat{p}=\dfrac{1}{\overline{X}}$，最大似然估计 $\hat{p}=\dfrac{1}{\overline{X}}$.

4. 矩估计 $\hat{p}=\dfrac{\overline{X}}{N}$，最大似然估计 $\hat{p}=\dfrac{\overline{X}}{N}$.

5. 矩估计 $\hat{\theta}=\dfrac{\overline{X}}{1-\overline{X}}$，最大似然估计 $\hat{\theta}=-\dfrac{n}{\sum\limits_{i=1}^{n}\ln X_i}$.

6. 最大似然估计 $\hat{\theta}=\overline{X}$.

7.（1）$\hat{\mu}_1$，$\hat{\mu}_2$ 为无偏估计量；（2）无偏估计量 $\hat{\mu}_2$ 方差最小.

8.［487.99，515.35］.

9.（1）［1 790.14，2 109.86］；（2）［221.61，464.31］.

10. 62，106.

11. 385.

12.［202.84，2 363.71］.

13. [667.983 8，4 708.580 4].

14. [−0.898 6，0.018 56].

15. [0.215 2，4.571 4].

习 题 八

1. 拒绝原假设，即认为包装机工作不正常.

2. 接受原假设，即认为这批农药的含磷均值为 3.25.

3. 拒绝原假设，即认为男生身高有明显变化.

4. 拒绝原假设，即认为该批原木平均直径低于 12cm.

5. 接受原假设，即认为该葡萄的方差为发生变化.

6. 接受原假设，即认为该水稻亩产量方差未变.

7. 拒绝原假设，即认为新工艺的铁水含碳量方差不为 0.108^2.

8. 在抽取的 100 个样本中次品为 8，时，u 值未落入拒绝域，故不能拒绝原假设 H_0，因此在 $\alpha=0.05$ 水平上不认为次品率大于 5%，可出厂.

9. 拒绝原假设，即认为改良前后石榴籽含水率发生变化.

10. 接受原假设，即认为两总体方差相同.

习 题 九

1. 方差分析表为

方差来源	自由度	平方和	均方	F value	Pr(>F)
因素 A	4	282.27	70.57	16.54	0.000 209
残差	10	42.67	4.267		
总和	14	328.54			

由检验结果，可认为温度对得率有显著影响.

2. 方差分析表为

方差来源	自由度	平方和	均方	F value	Pr(>F)
因素 A	3	346.0	115.33	14.66	0.000 002 79
残差	20	157.3	7.87		
总和	23	503.3			

（1）由检验结果，可认为 4 个厂生产的产品的变化率有显著差异. 且各水平的均值分别为

17.500 00　24.500 00　23.333 33

(2)水平间两两均值相等性检验的 p 值为

	1	2	3
2	0.000 50		
3	0.001 98	0.479 58	
4	0.102 97	0.000 39	0.000 85

3. 方差分析表为

方差来源	自由度	平方和	均方	F value	Pr(>F)
因素 A	3	27 387	9 129	1.315	0.295
残差	22	152 725	6 942		
总和	25	180 112			

由检验结果，可认为 4 种不同配方下元件的使用寿命无显著差异.

4. 方差分析表为

方差来源	自由度	平方和	均方	F value	Pr(>F)
因素 A	3	3 308	1 102.7	1.378	0.277
残差	21	16 804	800.2		
总和	24	20 112			

由检验结果，可认为各食谱的营养效果无显著差异.

5. 方差分析表为

方差来源	自由度	平方和	均方	F value	Pr(>F)
因素 A	2	73.12	36.56	9.104	0.001 42
残差	21	84.33	4.02		
总和	23	157.45			

由检验结果，可认为不同饲料的小鼠肝中的铁含量有显著差异.

6. 方差分析表为

方差来源	自由度	平方和	均方	F value	Pr(>F)
因素 A	5	1 435.11	287.022	17.802 9	0.000 108
因素 B	2	141.44	70.722	4.386 6	0.042 885
残差	10	161.22	16.122		

由检验结果，可认为不同的管理方法和不同的地块对草莓产量均有显著差异.

7. 方差分析表为

方差来源	自由度	平方和	均方	F value	Pr(>F)
因素 A	2	352.53	176.267	8.958 9	0.000 494
因素 B	3	87.52	29.172	1.482 7	0.231 077
交互 AB	6	71.73	11.956	0.607 7	0.722 890
残差	48	944.40	19.675		

由检验结果，可认为因素 A 有显著影响，因素 B 和交互 AB 均无显著影响.

习 题 十

1.（1）略；（2）$\hat{y} = -1 + 7x$；（3）显著；（4）28.4，（6.06，50.7）.

2. $\hat{y} = -0.005\ 71 + 0.023\ 4x$，$y(12) = 0.275\ 43$，预测区间(0.252 54，0.298 32).

3. $\hat{y} = \dfrac{2.827}{1 + 19.961\ 4\mathrm{e}^{-0.519\ 97x}}$，$\alpha = 0.01$ 时，回归方程显著.

4. $\hat{y} = 1\ 051.423\ 2\mathrm{e}^{-0.247\ 3x}$，$\alpha = 0.05$ 时，回归方程显著.

5.（1）$\hat{y} = 1\ 113 + 356x_1 + 6.29x_2 - 9.16x_3 + 40.0x_4 - 557x_5$；

（2）$\alpha = 0.05$ 时，回归方程显著；

（3）$\alpha = 0.05$ 时，回归系数 $\hat{\beta}_2$ 显著，其余系数不显著；

（4）2 558，（1 515，3 601）.

附　录

表 1　泊松分布表

设 $X \sim P(\lambda)$，表中给出概率

$$P(X \geqslant x) = \sum_{r=x}^{\infty} \frac{e^{-\lambda}\lambda^r}{r!}$$

x		$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$
0	1.000 000 0	1.000 000 0	1.000 000 0	1.000 000 0	1.000 000 0
1	0.181 269 2	0.259 181 8	0.329 680 0	0.323 469	0.451 188
2	0.017 523 1	0.036 936 3	0.061 551 9	0.090 204	0.121 901
3	0.001 148 5	0.003 599 5	0.007 926 3	0.014 388	0.023 115
4	0.000 056 8	0.000 265 8	0.000 776 3	0.001 752	0.003 358
5	0.000 002 3	0.000 015 8	0.000 061 2	0.000 172	0.000 394
6	0.000 000 1	0.000 000 8	0.000 004 0	0.000 014	0.000 039
7			0.000 000 2	0.000 001	0.000 003
x	$\lambda = 0.7$	$\lambda = 0.8$	$\lambda = 0.9$	$\lambda = 1.0$	$\lambda = 1.2$
0	1.000 000	1.000 000	1.000 000	1.000 000	1.000 000
1	0.503 415	0.550 671	0.593 430	0.632 121	0.698 806
2	0.155 805	0.191 208	0.227 518	0.264 241	0.337 373
3	0.034 142	0.047 423	0.062 857	0.080 301	0.120 513
4	0.005 753	0.009 080	0.013 459	0.018 988	0.033 769
5	0.000 786	0.001 411	0.002 344	0.003 660	0.007 746
6	0.000 090	0.000 184	0.000 343	0.000 594	0.001 500
7	0.000 009	0.000 021	0.000 043	0.000 083	0.000 251
8	0.000 001	0.000 002	0.000 005	0.000 010	0.000 037
9					0.000 005
10				0.000 001	0.000 001
x	$\lambda = 1.4$	$\lambda = 1.6$	$\lambda = 1.8$		
0	1.000 000	1.000 000	1.000 000		
1	0.753 403	0.798 103	0.834 701		
2	0.408 167	0.475 069	0.537 163		
3	0.166 502	0.216 642	0.269 379		
4	0.053 725	0.078 813	0.108 708		
5	0.014 253	0.023 682	0.036 407		
6	0.003 201	0.006 040	0.010 378		
7	0.000 622	0.001 336	0.002 569		
8	0.000 107	0.000 260	0.000 562		
9	0.000 016	0.000 045	0.000 110		
10	0.000 002	0.000 007	0.000 019		
11		0.000 001	0.000 003		

表 2　标准正态分布函数 $\Phi(x)$ 数值表

本表列出标准正态分布函数 $\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt$ 的值.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500 0	0.504 0	0.508 0	0.512 0	0.516 0	0.519 9	0.523 9	0.527 9	0.531 9	0.535 9
0.1	0.539 8	0.543 8	0.547 8	0.551 7	0.555 7	0.559 6	0.563 6	0.567 5	0.571 4	0.575 3
0.2	0.579 3	0.583 2	0.587 1	0.591 0	0.594 8	0.598 7	0.602 6	0.606 4	0.610 3	0.614 1
0.3	0.617 9	0.621 7	0.625 5	0.629 3	0.633 1	0.636 8	0.640 6	0.644 3	0.648 0	0.651 7
0.4	0.655 4	0.659 1	0.662 8	0.666 4	0.670 0	0.673 6	0.677 2	0.680 8	0.684 4	0.687 9
0.5	0.691 5	0.695 0	0.698 5	0.701 9	0.705 4	0.708 8	0.712 3	0.715 7	0.719 0	0.722 4
0.6	0.725 7	0.729 1	0.732 4	0.735 7	0.738 9	0.742 2	0.745 4	0.748 6	0.751 7	0.754 9
0.7	0.758 0	0.761 1	0.764 2	0.767 3	0.770 4	0.773 4	0.776 4	0.779 4	0.782 3	0.785 2
0.8	0.788 1	0.791 0	0.793 9	0.796 7	0.799 5	0.802 3	0.805 1	0.807 8	0.810 6	0.813 3
0.9	0.815 9	0.818 6	0.821 2	0.823 8	0.826 4	0.828 9	0.831 5	0.834 0	0.836 5	0.838 9
1.0	0.841 3	0.843 8	0.846 1	0.848 5	0.850 8	0.853 1	0.855 4	0.857 7	0.859 9	0.862 1
1.1	0.864 3	0.866 5	0.868 6	0.870 8	0.872 9	0.874 9	0.877 0	0.879 0	0.881 0	0.883 0
1.2	0.884 9	0.886 9	0.888 8	0.890 7	0.892 5	0.894 4	0.896 2	0.898 0	0.899 7	0.901 5
1.3	0.903 2	0.904 9	0.906 6	0.908 2	0.909 9	0.911 5	0.913 1	0.914 7	0.916 2	0.917 7
1.4	0.919 2	0.920 7	0.922 2	0.923 6	0.925 1	0.926 5	0.927 9	0.929 2	0.930 6	0.931 9
1.5	0.933 2	0.934 5	0.935 7	0.937 0	0.938 2	0.939 4	0.940 6	0.941 8	0.942 9	0.944 1
1.6	0.945 2	0.946 3	0.947 4	0.948 4	0.949 5	0.950 5	0.951 5	0.952 5	0.953 5	0.954 5
1.7	0.955 4	0.956 4	0.957 3	0.958 2	0.959 1	0.959 9	0.960 8	0.961 6	0.962 5	0.963 3
1.8	0.964 1	0.964 9	0.965 6	0.966 4	0.967 1	0.967 8	0.968 6	0.969 3	0.969 9	0.970 6
1.9	0.971 3	0.971 9	0.972 6	0.973 2	0.973 8	0.974 4	0.975 0	0.975 6	0.976 1	0.976 7
2.0	0.977 2	0.977 8	0.978 3	0.978 8	0.979 3	0.979 8	0.980 3	0.980 8	0.981 2	0.981 7
2.1	0.982 1	0.982 6	0.983 0	0.983 4	0.983 8	0.984 2	0.984 6	0.985 0	0.985 4	0.985 7
2.2	0.986 1	0.986 4	0.986 8	0.987 1	0.987 5	0.987 8	0.988 1	0.988 4	0.988 7	0.989 0
2.3	0.989 3	0.989 6	0.989 8	0.990 1	0.990 4	0.990 6	0.990 9	0.991 1	0.991 3	0.991 6
2.4	0.991 8	0.992 0	0.992 2	0.992 5	0.992 7	0.992 9	0.993 1	0.993 2	0.993 4	0.993 6
2.5	0.993 8	0.994 0	0.994 1	0.994 3	0.994 5	0.994 6	0.994 8	0.994 9	0.995 1	0.995 2
2.6	0.995 3	0.995 5	0.995 6	0.995 7	0.995 9	0.996 0	0.996 1	0.996 2	0.996 3	0.996 4
2.7	0.996 5	0.996 6	0.996 7	0.996 8	0.996 9	0.997 0	0.997 1	0.997 2	0.997 3	0.997 4
2.8	0.997 4	0.997 5	0.997 6	0.997 7	0.997 7	0.997 8	0.997 9	0.997 9	0.998 0	0.998 1
2.9	0.998 1	0.998 2	0.998 2	0.998 3	0.998 4	0.998 4	0.998 5	0.998 5	0.998 6	0.998 6

x	3.0	3.2	3.5	4.0	5.0
$\Phi(x)$	0.998 650	0.999 313	0.999 767	0.999 968 31	0.999 999 71

Excel 函数：$\Phi(x)$ = NORMSDIST(x)；若 $y=\Phi(x)$，则 x = NORMSINV(Vy).

表 3　t 分布上侧分位数表

设 T 服从自由度为 n 的 t 分布，本表列出使得 $P(T > t_n(\alpha)) = \alpha$ 的 $t_n(\alpha)$.

n \ α	0.10	0.05	0.025	0.01	0.005	0.002 5	0.001	0.000 5
1	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	4.773	5.894	6.869
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.689
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.660
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
100	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
∞	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.290

表 4　χ^2 分布上侧分位数表

设χ^2 服从自由度为 n 的χ^2 分布，本表列出使得 $P(\chi^2 > \chi_n^2(\alpha)) = \alpha$ 的$\chi_n^2(\alpha)$.

n \ α	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672

表 5　F 分布上侧分位数表

设 F 服从自由度为 m，n 的 F 分布，本表列出使得 $P(F > F_{m,n}(\alpha)) = \alpha$ 的 $F_{m,n}(\alpha)$.

($\alpha = 0.1$)

n \ m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	39.864	49.500	53.593	55.833	57.240	58.204	58.906	59.439	59.857	60.195	60.473	60.705	60.902	61.073	61.220	61.350
2	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392	9.401	9.408	9.415	9.420	9.425	9.429
3	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230	5.222	5.216	5.210	5.205	5.200	5.196
4	4.545	4.325	4.191	4.107	4.051	4.010	3.979	3.955	3.936	3.920	3.907	3.896	3.886	3.878	3.870	3.864
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297	3.282	3.268	3.257	3.247	3.238	3.230
6	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937	2.920	2.905	2.892	2.881	2.871	2.863
7	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703	2.684	2.668	2.654	2.643	2.632	2.623
8	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538	2.519	2.502	2.488	2.475	2.464	2.454
9	3.360	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.440	2.416	2.396	2.379	2.364	2.351	2.340	2.330
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323	2.302	2.284	2.269	2.255	2.244	2.233
11	3.225	2.860	2.660	2.536	2.451	2.389	2.342	2.304	2.274	2.248	2.227	2.209	2.193	2.179	2.167	2.156
12	3.177	2.807	2.606	2.480	2.394	2.331	2.283	2.245	2.214	2.188	2.166	2.147	2.131	2.117	2.105	2.094
13	3.136	2.763	2.560	2.434	2.347	2.283	2.234	2.195	2.164	2.138	2.116	2.097	2.080	2.066	2.053	2.042
14	3.102	2.726	2.522	2.395	2.307	2.243	2.193	2.154	2.122	2.095	2.073	2.054	2.037	2.022	2.010	1.998
15	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059	2.037	2.017	2.000	1.985	1.972	1.961
16	3.048	2.668	2.462	2.333	2.244	2.178	2.128	2.088	2.055	2.028	2.005	1.985	1.968	1.953	1.940	1.928

($\alpha = 0.05$)

n \ m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	161.446	199.499	215.707	224.583	230.160	233.988	236.767	238.884	240.543	241.882	242.981	243.905	244.690	245.363	245.949	246.466
2	18.513	19.000	19.164	19.247	19.296	19.329	19.353	19.371	19.385	19.396	19.405	19.412	19.419	19.424	19.429	19.433
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.785	8.763	8.745	8.729	8.715	8.703	8.692
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.936	5.912	5.891	5.873	5.858	5.844
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.704	4.678	4.655	4.636	4.619	4.604
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.027	4.000	3.976	3.956	3.938	3.922
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.603	3.575	3.550	3.529	3.511	3.494
8	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.347	3.313	3.284	3.259	3.237	3.218	3.202
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.102	3.073	3.048	3.025	3.006	2.989
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.943	2.913	2.887	2.865	2.845	2.828
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.818	2.788	2.761	2.739	2.719	2.701
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.717	2.687	2.660	2.637	2.617	2.599
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.635	2.604	2.577	2.554	2.533	2.515
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.565	2.534	2.507	2.484	2.463	2.445
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.507	2.475	2.448	2.424	2.403	2.385
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.456	2.425	2.397	2.373	2.352	2.333

($\alpha=0.025$)

n \ m	1.000	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	647.793	799.482	864.151	899.599	921.835	937.114	948.203	956.643	963.279	968.634	973.028	976.725	979.839	982.545	984.874	986.911
2	38.506	39.000	39.166	39.248	39.298	39.331	39.356	39.373	39.387	39.398	39.407	39.415	39.421	39.427	39.431	39.436
3	17.443	16.044	15.439	15.101	14.885	14.735	14.624	14.540	14.473	14.419	14.374	14.337	14.305	14.277	14.253	14.232
4	12.218	10.649	9.979	9.604	9.364	9.197	9.074	8.980	8.905	8.844	8.794	8.751	8.715	8.684	8.657	8.633
5	10.007	8.434	7.764	7.388	7.146	6.978	6.853	6.757	6.681	6.619	6.568	6.525	6.488	6.456	6.428	6.403
6	8.813	7.260	6.599	6.227	5.988	5.820	5.695	5.600	5.523	5.461	5.410	5.366	5.329	5.297	5.269	5.244
7	8.073	6.542	5.890	5.523	5.285	5.119	4.995	4.899	4.823	4.761	4.709	4.666	4.628	4.596	4.568	4.543
8	7.571	6.059	5.416	5.053	4.817	4.652	4.529	4.433	4.357	4.295	4.243	4.200	4.162	4.130	4.101	4.076
9	7.209	5.715	5.078	4.718	4.484	4.320	4.197	4.102	4.026	3.964	3.912	3.868	3.831	3.798	3.769	3.744
10	6.937	5.456	4.826	4.468	4.236	4.072	3.950	3.855	3.779	3.717	3.665	3.621	3.583	3.550	3.522	3.496
11	6.724	5.256	4.630	4.275	4.044	3.881	3.759	3.664	3.588	3.526	3.474	3.430	3.392	3.359	3.330	3.304
12	6.554	5.096	4.474	4.121	3.891	3.728	3.607	3.512	3.436	3.374	3.321	3.277	3.239	3.206	3.177	3.152
13	6.414	4.965	4.347	3.996	3.767	3.604	3.483	3.388	3.312	3.250	3.197	3.153	3.115	3.082	3.053	3.027
14	6.298	4.857	4.242	3.892	3.663	3.501	3.380	3.285	3.209	3.147	3.095	3.050	3.012	2.979	2.949	2.923
15	6.200	4.765	4.153	3.804	3.576	3.415	3.293	3.199	3.123	3.060	3.008	2.963	2.925	2.891	2.862	2.836
16	6.115	4.687	4.077	3.729	3.502	3.341	3.219	3.125	3.049	2.986	2.934	2.889	2.851	2.817	2.788	2.761

($\alpha=0.001$)

n \ m	1.000	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	4052.185	4999.340	5403.534	5624.257	5763.955	5858.950	5928.334	5980.954	6022.397	6055.925	6083.399	6106.682	6125.774	6143.004	6156.974	6170.012
2	98.502	99.000	99.164	99.251	99.302	99.331	99.357	99.375	99.390	99.397	99.408	99.419	99.422	99.426	99.433	99.437
3	34.116	30.816	29.457	28.710	28.237	27.911	27.671	27.489	27.345	27.228	27.132	27.052	26.983	26.924	26.872	26.826
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.452	14.374	14.306	14.249	14.198	14.154
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.963	9.888	9.825	9.770	9.722	9.680
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.790	7.718	7.657	7.605	7.559	7.519
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.538	6.469	6.410	6.359	6.314	6.275
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.734	5.667	5.609	5.559	5.515	5.477
9	10.562	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.178	5.111	5.055	5.005	4.962	4.924
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.772	4.706	4.650	4.601	4.558	4.520
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.462	4.397	4.342	4.293	4.251	4.213
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.220	4.155	4.100	4.052	4.010	3.972
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	4.025	3.960	3.905	3.857	3.815	3.778
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.864	3.800	3.745	3.698	3.656	3.619
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.730	3.666	3.612	3.564	3.522	3.485
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.616	3.553	3.498	3.451	3.409	3.372

Excel 函数：　$F_{m,n}(\alpha)$= FINV(α, m, n).

（$\alpha=0.005$）

n \ m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	16212.463	19997.358	21614.134	22500.753	23055.822	23439.527	23715.198	23923.814	24091.452	24221.838	24333.596	24426.728	24504.960	24572.015	24631.619	24683.774
2	198.503	199.012	199.158	199.245	199.303	199.332	199.361	199.376	199.390	199.390	199.419	199.419	199.419	199.419	199.434	199.449
3	55.552	49.800	47.468	46.195	45.391	44.838	44.434	44.125	43.881	43.685	43.525	43.387	43.270	43.172	43.085	43.008
4	31.332	26.284	24.260	23.154	22.456	21.975	21.622	21.352	21.138	20.967	20.824	20.705	20.603	20.515	20.438	20.371
5	22.785	18.314	16.530	15.556	14.939	14.513	14.200	13.961	13.772	13.618	13.491	13.385	13.293	13.215	13.146	13.086
6	18.635	14.544	12.917	12.028	11.464	11.073	10.786	10.566	10.391	10.250	10.133	10.034	9.950	9.878	9.814	9.758
7	16.235	12.404	10.883	10.050	9.522	9.155	8.885	8.678	8.514	8.380	8.270	8.176	8.097	8.028	7.968	7.915
8	14.688	11.043	9.597	8.805	8.302	7.952	7.694	7.496	7.339	7.211	7.105	7.015	6.938	6.872	6.814	6.763
9	13.614	10.107	8.717	7.956	7.471	7.134	6.885	6.693	6.541	6.417	6.314	6.227	6.153	6.089	6.032	5.983
10	12.827	9.427	8.081	7.343	6.872	6.545	6.303	6.116	5.968	5.847	5.746	5.661	5.589	5.526	5.471	5.422
11	12.226	8.912	7.600	6.881	6.422	6.102	5.865	5.682	5.537	5.418	5.320	5.236	5.165	5.103	5.049	5.001
12	11.754	8.510	7.226	6.521	6.071	5.757	5.524	5.345	5.202	5.085	4.988	4.906	4.836	4.775	4.721	4.674
13	11.374	8.186	6.926	6.233	5.791	5.482	5.253	5.076	4.935	4.820	4.724	4.643	4.573	4.513	4.460	4.413
14	11.060	7.922	6.680	5.998	5.562	5.257	5.031	4.857	4.717	4.603	4.508	4.428	4.359	4.299	4.247	4.201
15	10.798	7.701	6.476	5.803	5.372	5.071	4.847	4.674	4.536	4.424	4.329	4.250	4.181	4.122	4.070	4.024
16	10.576	7.514	6.303	5.638	5.212	4.913	4.692	4.521	4.384	4.272	4.179	4.099	4.031	3.972	3.920	3.875

参考文献

1. 王松桂，张忠占，程维虎，高旅端. 概率论与数理统计[M]. 北京：科学出版社，2006.
2. 茆诗松，程依明，濮晓龙. 概率论与数理统计教程[M]. 北京：高等教育出版社，2012.
3. 张国权，刘金山. 概率论与数理统计[M]. 北京：中国农业出版社，2015.
4. 刘金山. 概率论[M]. 北京：中国农业出版社，2014.
5. 吴坚，刘金山. 概率论[M]. 北京：中国农业出版社，2007.
6. 张国权. 应用概率统计[M]. 北京：科学出版社，2005.
7. 孙荣恒. 应用概率论[M]. 北京：科学出版社，2006.
8. 盛骤，谢式千，潘承毅. 概率论与数理统计[M]. 北京：高等教育出版社，2005.
9. 梁之舜，邓集贤，杨维权，司徒荣. 概率论与数理统计[M]. 北京：高等教育出版社，2005.
10. 陈希儒. 概率论与数理统计[M]. 北京：科学出版社，2002.
11. 刘金山. 概率论与数理统计教程[M]. 北京：科学出版社，2016.