

大 学 数 学 教 程

概率论 与数理统计

第二版

山东大学数学学院

刘建亚 吴 臻 主编
胡发胜 叶 宏 吕 同 编

 高等教育出版社
HIGHER EDUCATION PRESS

大 学 数 学 教 程

微积分1 第二版

微积分2 第二版

线性代数 第二版

概率论与数理统计 第二版

复变函数与积分变换 第二版

ISBN 978-7-04-032301-6



9 787040 323016 >

定价 20.20 元

大 学 数 学 教 程

概率论 与数理统计

Gailülun yu Shuli Tongji

第二版

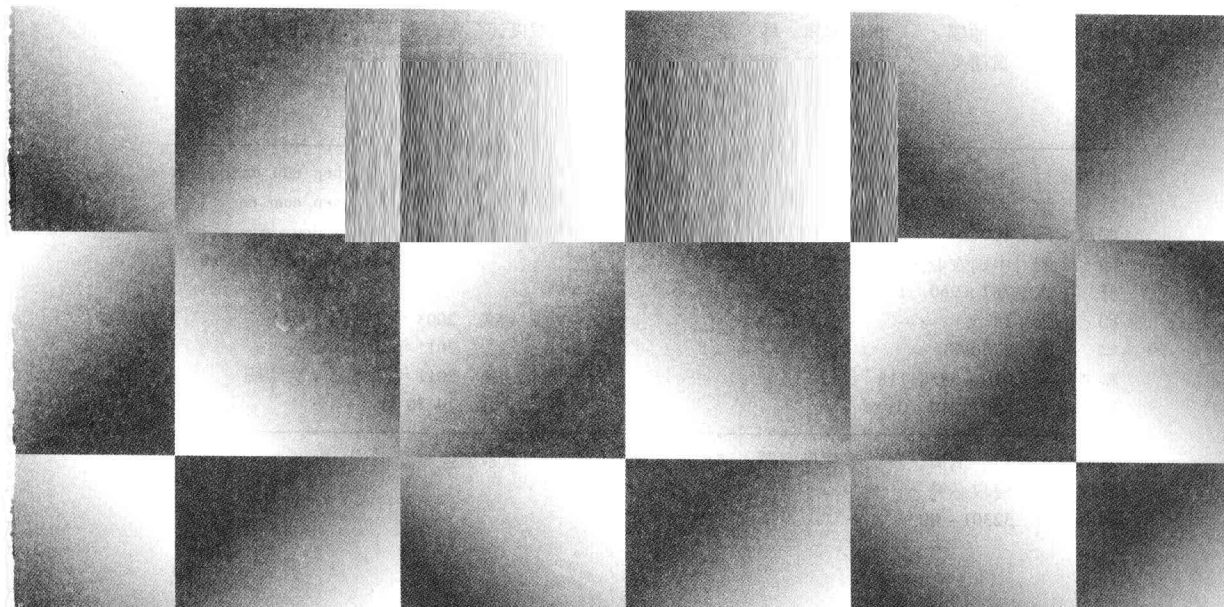
山东大学数学学院

刘建亚 吴 臻 主编

胡发胜 叶 宏 吕 同 编



高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING



内容简介

本书内容少而精,突出数学思想,注重理论与应用相结合,内容与中学知识相衔接,深入浅出,易教易学.本书主要内容包括随机事件及其概率、随机变量及其分布、多维随机变量及其分布、随机变量的数字特征、数理统计的基本知识、参数估计和假设检验、一元线性回归分析和方差分析.

书中每章最后一节均是解决本章主要问题的 MATLAB 程序和例题演示.每章后的习题配置基本练习题和部分综合练习题,书末附有习题答案.附录给出了概率统计发展简介以及概率统计在数学建模中的应用.

本书可供高等学校非数学类专业学生使用,也可供科技工作者学习参考.

图书在版编目(CIP)数据

大学数学教程.概率论与数理统计/刘建亚,吴臻主编;胡发胜,叶宏,吕同编. —2版. —北京:高等教育出版社,2011.7

ISBN 978 - 7 - 04 - 032301 - 6

I. ①大… II. ①刘…②吴…③胡…④叶…⑤吕… III. ①高等数学 - 高等学校 - 教材②概率论 - 高等学校 - 教材③数理统计 - 高等学校 - 教材 IV. ①O13②O21

中国版本图书馆 CIP 数据核字(2011)第 121557 号

策划编辑 于丽娜
插图绘制 黄建英

责任编辑 马丽 李蕊
责任校对 刘春萍

封面设计 张志奇
责任印制 张泽业

版式设计 马敬茹

出版发行 高等教育出版社
社 址 北京市西城区德外大街 4 号
邮政编码 100120
印 刷 中国农业出版社印刷厂
开 本 787 × 960 1/16
印 张 14.5
字 数 270 000
购书热线 010 - 58581118
咨询电话 400 - 810 - 0598

网 址 <http://www.hep.edu.cn>
<http://www.hep.com.cn>
网上订购 <http://www.landaco.com>
<http://www.landaco.com.cn>
版 次 2003 年 7 月第 1 版
2011 年 7 月第 2 版
印 次 2011 年 7 月第 1 次印刷
定 价 20.20 元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换
版权所有 侵权必究
物 料 号 32301 - 00

大学数学教程

编委会

主 编	刘建亚	吴 臻				
编 委	(按姓氏笔画排列)					
	刁在筠	包芳勋	叶 宏	吕 同	许闻天	
	张天德	张光明	郑修才	金 辉	胡发胜	
	秦 静	傅国华	蒋晓芸			

第二版前言

本套教材是由山东大学数学学院具有丰富教学经验的一线教师编写的，第一版是普通高等教育“十五”国家级规划教材，包括《微积分1》、《微积分2》、《线性代数》、《概率论与数理统计》、《复变函数与积分变换》五册。经过多年的教学实践，山东大学数学学院在大学数学课程建设和教学改革方面取得了可喜的成绩。微积分与数学实验、线性代数、复变函数与积分变换分别于2007年、2006年、2010年被评为国家精品课程，以刘建亚教授为带头人的大学数学系列课程教学团队被评为2007年度国家级教学团队。

为更好地将优秀教学改革成果运用并推广，根据当前的教学实际，山东大学数学学院组织中坚力量，对第一版进行了修订。在保持本书第一版优点、特色的前提下，新版教材注重与中学教学内容的衔接，增加了与中学数学接轨的部分内容；增选一些国外教材中的案例、例题和习题，力求题型新颖。为更好地将数学建模思想融入教学，培养学生的建模思想和意识，通过增设有关章节介绍与教学内容相关的建模案例，全方位提升学生的综合素质和创新能力。新版教材力求做到符合大学数学课程教学基本要求，知识结构符合认知规律，同时渗透现代数学思想，加强应用能力培养，便于学生学习和教师教学。

本书为《概率论与数理统计》分册，与第一版相比，新版主要作了如下修改：

1. 根据教学实际，增加了有关基本概念、基本运算的例题，使之更加符合学生的学习和思维习惯。
2. 对原有的习题作了调整，扩大了题目涉及的范围，增加了题型的变化，另外，补充了许多考研经典题目，以满足学生考研复习的需要。
3. 对于各部分内容作了补充与完善，例如，关于多维随机变量函数的分布，增加了离散型随机变量函数的情形；中心极限定理中增加了棣莫弗-拉普拉斯定理等。另外，对于很多定理及方法，补充了证明或者分析，使之更加实用化和通俗化，便于学生理解和使用。
4. 为适应当今社会对数学建模的需要，在附录增加了与概率统计内容对应的数学建模应用实例。
5. 将第一版附录中的著名数学家简介修改为以概率论与数理统计的发展简史为主线、穿插数学家的简介。

本次修订工作主要由吕同(第1章)、叶宏(第2,3,4章)、胡发胜(第5,6,7章)完成. 刘建亚教授、吴臻教授按照丛书总体要求对修改框架提出具体建议, 栾贻会教授对全书进行了审定. 另外, 数学实验内容由傅国华编写, 概率论与数理统计发展简史由包芳勋编写. 在本书修订过程中, 我们得到高等教育出版社、山东大学教务处、山东大学数学学院领导及同事的大力支持, 兄弟院校的同行也对此次修订提出了不少宝贵建议. 在此, 我们表示衷心的感谢.

限于编者水平所限, 新版中难免存在不足, 欢迎广大专家、同行与读者批评指正.

编 者

2011年2月

第一版前言

为了适应新世纪我国高等教育迅速发展的形势和实行学分制的需要,满足新时期高等教育人才培养拓宽口径、增强适应性对数学教育的要求,山东大学数学与系统科学学院从2000年开始按照教育部《高等教育面向21世纪教学内容和课程体系改革计划》的精神和要求,在学院领导的亲自参与下,组织部分教师对非数学类专业大学数学的课程体系进行了认真深入的研究和论证. 针对大学数学是高校非数学类专业所有大学生应当具有的素质,又考虑到不同专业的要求深浅不同、内容多少各异的实际情况,制订了适应这种情况的新课程体系. 新课程体系的主要特点是采取平台加模块的结构,整个大学数学的课程共分三个平台,不同平台反映了不同专业对数学知识的不同层次、级别要求,体现数学知识结构和大学生认知结构的统一. 鉴于人类认识是从感性到理性,由易到难,由浅入深的,因此第一平台(包括微积分(一)、线性代数和概率统计)是体现高等数学的普及和基础,体现所有各专业应当具有的数学素质教育,主要侧重基本概念和基本方法,加强基本运算,努力渗透基本数学思想;第二平台是对第一平台基本概念의加深和认识方法的拓宽,在本平台中还适当体现出数学理论的系统性和严谨性;第三平台(包括数学建模、数值分析、数理方程、复变函数和积分变换、运筹学等)则是为满足某些对数学知识和方法有特殊要求的专业而设置. 各平台的教学内容由浅入深,反映不同专业对数学知识和内容的不同要求;各平台的内容又采取模块组合的方式,模块间相对独立,各专业亦可根据本专业的需要,选用不同的模块组合,这样就使得新的课程体系具有更大的灵活性,能够满足不同层次、不同要求的专业对数学教学的需求. 另外,新课程体系还将利用计算机解决数学问题的数学实验融入其中,做到理论和实践的有机结合.

山东大学教务处对新课程体系给予充分的肯定,并大力支持按新课程体系编写相应的教材. 在我们完成初稿之后,教务处安排几个专业的学生先行试用,并在此基础上加以修改完善. 目前,已完成了前两个平台共计四册的教材编写和修改. 其中,微积分为两册,分属两个平台;线性代数和概率统计各一册. 其中《概率论与数理统计》这本教材还有以下特色:

1. 内容少而精,体现素质教育,突出数学思想. 我们重点介绍概率论与数理统计中的基本概念和基本方法;从培养能力和提高素质为着眼点,有选择地保留了部分定理、性质的证明,对那些用类似的技巧方法,或者读者举一反三

三可以理解或自学的证明部分作省略或简化处理。

2. 扩大了读者的知识面。我们将各专业不同需求的数学内容融进了一套教材中。主要的做法是：用“*”号标明不同层次对数学的要求；从不同的学科例题分析中引进基本概念；习题中也涉及多学科。如在数学要求较低专业学习的读者希望学习更多数学知识（如跨学科考研或工作需要）时，可以从同一本书按“*”号的标示获取。当然，教师在授课时可按本专业的要求有选择地使用。

3. 各节后的习题配置除基本练习外，还有部分综合练习题，以提高读者分析问题、解决问题的能力。综合练习题多置于每节习题后且配以“*”号标示。

4. 紧密联系实际问题，适当反映了统计方法在实际应用中的新进展。增添了利用计算机解决数学问题的内容，在每章后均有解决本章主题问题的MATLAB程序和例题演示。

本书由山东大学数学与系统科学学院组织部分有较高水平和丰富教学经验的教师集体编写，最后聘请有关专家审定。在长达近两年的编写过程中，学院领导给予了极大的关注、支持和具体指导，为此曾多次召开各种类型的会议反复论证，几易手稿。

大学数学教程的主编是刘建亚，概率统计部分由吕同（第1、2、3、4章）、胡发胜（第5、6、7章）编写，由胡发胜完成修改及统稿工作；吴臻、刘锦萼审查定稿；各册的数学实验内容由傅国华编写和制作。

本套教材作为普通高等教育“十五”国家级规划教材正式出版，是教育改革的产物。在此，我们感谢山东大学教务处、山东大学出版基金委、山东大学数学学院领导对改革和教材出版的鼎力支持，感谢仪洪勋、江守礼教授对我们的鼓励和帮助。我们特别感谢高等教育出版社，由于他们的指导和帮助才使本书顺利与读者见面。

新时期大学数学的教学改革是一项非常紧迫，非常重要，也是非常艰巨的工作。限于编者水平，本书肯定会有许多不足和缺点乃至问题，恳请读者批评指正。

编 者

2002年11月

目 录

第 1 章 随机事件及其概率	1
§ 1.1 随机事件及其运算	1
§ 1.2 随机事件的概率	5
§ 1.3 概率的基本运算法则	12
§ 1.4 全概率公式与贝叶斯公式	20
§ 1.5 伯努利概型	23
§ 1.6 随机数的生成与应用	25
习题 1	28
第 2 章 随机变量及其分布	32
§ 2.1 随机变量与分布函数	32
§ 2.2 离散型随机变量及其分布	34
§ 2.3 连续型随机变量及其分布	41
§ 2.4 随机变量函数的分布	49
§ 2.5 用 MATLAB 计算分布函数	53
习题 2	55
第 3 章 多维随机变量及其分布	59
§ 3.1 二维随机变量及其分布	59
§ 3.2 边缘分布	65
§ 3.3 条件分布	68
§ 3.4 随机变量的独立性	72
§ 3.5 二维随机变量函数的分布	74
§ 3.6 用 MATLAB 画二维分布图	80
习题 3	81
第 4 章 随机变量的数字特征	85
§ 4.1 数学期望	85
§ 4.2 方差	93
§ 4.3 协方差与相关系数	96
§ 4.4 大数定律与中心极限定理	100
§ 4.5 用 MATLAB 计算数学期望和方差	104
习题 4	106
第 5 章 数理统计的基本知识	110
§ 5.1 数理统计学	110

§ 5.2 总体与样本	110
§ 5.3 统计量与抽样分布	112
§ 5.4 数据的整理	121
§ 5.5 常用统计量的计算	125
习题 5	130
第 6 章 参数估计和假设检验	132
§ 6.1 参数的点估计	132
§ 6.2 参数的区间估计	141
§ 6.3 假设检验的基本概念	147
§ 6.4 正态总体参数的假设检验	150
§ 6.5 参数估计与假设检验的 MATLAB 计算	158
习题 6	162
第 7 章 一元线性回归分析和方差分析	166
§ 7.1 回归分析的基本概念	166
§ 7.2 一元线性回归	167
§ 7.3 单因素方差分析	173
§ 7.4 用 MATLAB 处理回归与方差分析	181
习题 7	184
附表	186
附表 1 随机数表	186
附表 2 二项分布表	187
附表 3 泊松分布表	189
附表 4 标准正态分布表	190
附表 5 t 分布表	192
附表 6 χ^2 分布表	193
附表 7 F 分布表	196
习题答案	200
附录 I 概率统计在数学建模中的应用	209
附录 II 概率论与数理统计发展简介	212

第 1 章 随机事件及其概率

自然界和人类社会中出现的种种现象，大体上可分为两类：一类现象是在一定条件下必然发生或绝不可能发生的，此类现象称为**确定性现象**。例如，在 1 个标准大气压下把水加热到 $100\text{ }^{\circ}\text{C}$ 时必然会沸腾；电流通过导线时，导线周围必然产生磁场，等等。另一类现象则不然，即在相同的条件下可能发生也可能不发生，或者说，可能出现这个结果也可能出现那个结果，呈现出偶然性，此类现象称为**随机现象**。例如，抛掷一枚硬币究竟是正面（币值面）朝上还是反面朝上，在每次抛掷之前是无法判定的；观察用一种新药治疗某种疾病的疗效，对一个病人来说，可能有效也可能无效。通过大量试验我们知道：大量重复抛掷同一枚硬币时正面朝上的次数约占抛掷总数的一半。这类在个别观察试验中呈现不确定的结果，而在相同条件下，大量重复试验，试验结果呈现出的规律性称之为统计规律性。又如，为评价一种新药的疗效，通过足够多个病例的试用和观察，可以对其效果作出客观的估计。概率论与数理统计就是研究随机现象统计规律性的一门数学学科，它在自然科学、工程技术和社会科学的多领域中有着重要的应用。特别是随着计算机的普及使用，概率统计在经济管理、金融保险、生物医药等方面的应用更加广泛、更加深入。

§ 1.1 随机事件及其运算

1. 随机试验与样本空间

为了确定随机现象的规律性，需要进行多次的试验、实验、调查或观察，我们把这些工作统称为试验。概率论中所说的试验是指随机试验，它具有下列三个特征：

- 1) 试验可在相同的条件下重复进行；
- 2) 试验的结果不止一个；
- 3) 每次试验之前，不能判定哪一个结果将会出现。

例如，掷一颗骰子，观察出现的点数；记录某传呼台在一小时内接收到的呼唤次数；观察日光灯的使用寿命，以及前面提到的掷硬币试验等，都是随机试验，简称为试验，用 E 表示。

试验 E 中的每一个可能结果称为基本事件，或称为样本点，所有基本事件组成的集合称为试验 E 的样本空间，记为 Ω 。

例 1.1.1 在抛掷一枚硬币试验中, 有两个可能的结果: 出现正面, 出现反面. 若分别用“正”、“反”来表示, 即有两个基本事件, 这个试验的样本空间为 $\Omega = \{\text{正}, \text{反}\}$.

例 1.1.2 掷一颗骰子, 观察其出现的点数, 所有可能出现的结果有 6 个: 出现 1 点, 出现 2 点, $\dots$, 出现 6 点. 分别用 $1, 2, \dots, 6$ 表示, 则样本空间为 $\Omega = \{1, 2, \dots, 6\}$.

例 1.1.3 在一批日光灯中任意抽取一只, 测试其寿命, 用 t (单位: 小时) 表示日光灯的使用寿命, 则 t 可取所有非负实数: $t \geq 0$, 对应了试验的所有可能结果, 则样本空间为 $\Omega = \{t | t \geq 0\}$.

由上述例题可以看出, 样本空间应根据随机试验的内容来确定, 这是很重要的.

2. 随机事件

随机试验 E 的样本空间 Ω 中的子集称为试验 E 的随机事件, 简称为事件, 常用大写英文字母 $A, B, C, \dots$ 表示. 在例 1.1.2 中, 基本事件 $\{\text{出现 2 点}\}$ 、 $\{\text{出现 4 点}\}$ 、 $\{\text{出现 6 点}\}$ 以及由它们组成的集合“出现偶数点”, 都是该试验的随机事件. 基本事件是最简单的随机事件, 而一般的随机事件是由若干个基本事件组成的, 称为复合事件.

随机事件中有两个极端的情况: 一是由样本空间 Ω 中的所有元素组成的集合, 称之为必然事件, 用 Ω 来表示, 它在每一次试验中都发生. 例如, 前面所述“抛掷一颗骰子, 出现点数都不大于 6”就是必然事件. 另一种是不含任何元素的空集合, 称之为不可能事件, 用 $\emptyset$ 来表示, 例如“抛掷一颗骰子, 出现点数大于 6”就是不可能事件. 它在每一次试验中都不会发生. 严格来说这两种事件不是随机事件, 但为了今后讨论方便, 我们还是把必然事件与不可能事件作为随机事件的特殊情形来统一处理.

3. 事件间的关系与运算

在同一随机试验中, 事件不止一个. 有些事件简单, 有些事件复杂. 通过研究它们之间的联系. 可以更好地帮助我们理解事件的本质.

设试验 E 的样本空间为 Ω , $A, B, C, A_k (k=1, 2, 3, \dots)$ 是 E 的事件.

(1) 包含 若事件 A 发生必然导致事件 B 发生, 则称事件 B 包含事件 A , 或称事件 A 包含于事件 B 中, 记为 $B \supset A$ 或 $A \subset B$, 如图 1.1.1 所示 (其中, 事件 A 发生和 B 发生分别表示样本点落在圆 A 和圆 B 内, 下同).

显然, 必然事件 Ω 包含任何事件 A , 事件 A 包含不可能事件 $\emptyset$, 即 $\Omega \supset A \supset \emptyset$.

特别地, 若事件 A 包含事件 B , 且事件 B 也包

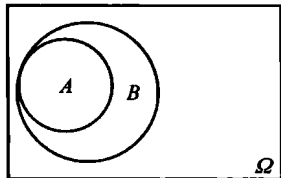


图 1.1.1

含事件 A , 即 $A \supset B$ 且 $A \subset B$, 则称事件 A 与事件 B 相等, 记为 $A = B$.

(2) **事件的并** 若事件 A 与事件 B 至少有一个发生, 这样构成的事件称为事件 A 与事件 B 的并事件(或称为 A 与 B 的和事件), 记为 $A \cup B$. 例如, 10 件产品中有 3 件次品, 从中任取 2 件, 若 A 表示“取到 1 件次品”, B 表示“取到 2 件次品”, 则和事件 $A \cup B$ 表示“至少取到 1 件次品”.

事件 $A \cup B$ 通常包含三个部分: A 发生而 B 不发生, A 不发生而 B 发生, A, B 都发生. 如图 1.1.2 阴影部分所示.

类似地, n 个事件 $A_1, A_2, \dots, A_n$ 的并事件 $A_1 \cup A_2 \cup \dots \cup A_n$ 表示“ $A_1, A_2, \dots, A_n$ 中至少有一个发生”.

(3) **事件的交** 由事件 A 与事件 B 同时发生而构成的事件称为事件 A 与事件 B 的交事件(或称为 A 与 B 的积事件), 记为 AB (或 $A \cap B$). 如图 1.1.3 阴影部分所示.

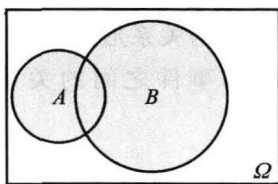


图 1.1.2

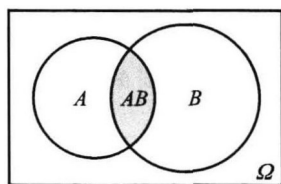


图 1.1.3

在上面“10 件产品中有 3 件次品”的例子中, $A \cap B = \emptyset$; 若用 C 表示“至多取到 2 件次品”, 则 AC 表示“恰好取到 1 件次品”, 即 $AC = A$.

类似地, n 个事件 $A_1, A_2, \dots, A_n$ 的交事件 $A_1 A_2 \dots A_n$ 表示“ $A_1, A_2, \dots, A_n$ 同时发生”.

(4) **事件的差** 事件 A 发生而事件 B 不发生, 这样构成的事件称为事件 A 与事件 B 的差事件, 记为 $A - B$. 如图 1.1.4 阴影部分所示.

(5) **互不相容事件** 在一次试验中, 若事件 A 与事件 B 不能同时发生, 即 $AB = \emptyset$, 则称 A 与 B 为互不相容事件(或称为互斥事件). 例如, 掷一颗骰子, A 表示“出现 3 点”, B 表示“出现 4 点”, 则 A 与 B 为互不相容事件. 如图 1.1.5.

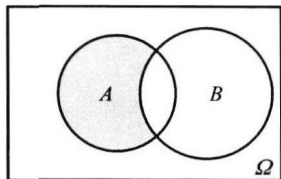


图 1.1.4

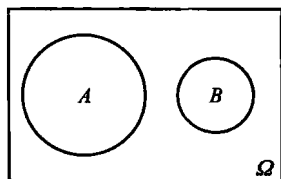


图 1.1.5

一般地, 对于 n 个事件 $A_1, A_2, \dots, A_n$, 若它们之间两两互不相容, 则称这 n 个事件是互不相容的. 通常 $A_1 \cup A_2 \cup \dots \cup A_n$ 记为 $A_1 + A_2 + \dots + A_n$ 或 $\sum_{i=1}^n A_i$.

(6) 对立事件 在一次试验中, 若事件 A 与事件 B 二者必有一个发生且仅有一个发生, 则称 A 与 B 为对立事件(或称为互逆事件), 通常把 A 的对立事件记为 $\bar{A}$ (即 $B = \bar{A}$), $\bar{A}$ 也称为 A 的逆事件. 例如, 掷一枚硬币, 用 A 表示“出现图案面”, 而事件 B 表示“出现币值面”, 则 A 与 B 为对立事件. 如图 1.1.6 所示.

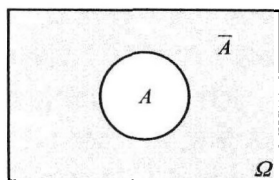


图 1.1.6

显然, $A \cap \bar{A} = \emptyset$, 且 $A \cup \bar{A} = \Omega$.

值得注意的是, 若事件 A 与 B 对立, 则事件 A 与事件 B 互不相容, 反之不然.

概率论中事件间的关系和运算与集合论中集合间的关系形式上是类似的, 利用在中学里学到的集合知识, 可以更好地理解事件之间的关系(见表 1.1.1).

表 1.1.1

符号	概率论中	集合论中
Ω	必然事件	全集
$\emptyset$	不可能事件	空集
A	事件	子集
$\bar{A}$	A 的对立事件	余集
$A \cup B$	事件的并	并集
$A \cap B$	事件的交	交集
$A - B$	事件的差	差集
$A \cap B = \emptyset$	互不相容事件	A 与 B 没有公共元素

集合论中常用的运算律同样适用于概率论的事件的运算.

- (1) 交换律 $A \cup B = B \cup A$,
 $A \cap B = B \cap A$;
- (2) 结合律 $(A \cup B) \cup C = A \cup (B \cup C)$,
 $(A \cap B) \cap C = A \cap (B \cap C)$;
- (3) 分配律 $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$,
 $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;
- (4) 德摩根定律 $\overline{A \cap B} = \bar{A} \cup \bar{B}$,

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

对于 n 个事件 $A_1, A_2, \dots, A_n$ 有

$$\overline{A_1 \cup A_2 \cup \dots \cup A_n} = \overline{A_1} \cap \overline{A_2} \cap \dots \cap \overline{A_n},$$

$$\overline{A_1 \cap A_2 \cap \dots \cap A_n} = \overline{A_1} \cup \overline{A_2} \cup \dots \cup \overline{A_n}.$$

例 1.1.4 设 A, B, C 表示任意三个随机事件, 用 A, B, C 及其运算符号表示下列事件:

- (1) A 发生而 B 与 C 都不发生;
- (2) A 发生且 B 与 C 至少有一个发生;
- (3) A 与 B 发生而 C 不发生;
- (4) A, B, C 三个事件中只有一个发生;
- (5) A, B, C 中至少有两个发生;
- (6) A, B, C 中至多有一个发生.

解 (1) 该事件可表示为 $A \overline{B} \overline{C}$ 或 $A - B - C$ 或 $A - (B \cup C)$;

(2) 因为 $B \cup C$ 表示 B 与 C 至少有一个发生, 故该事件可表示为 $A \cap (B \cup C)$;

(3) A 与 B 发生即 A 和 B 都发生, 就是 AB , C 不发生即 $\overline{C}$ 发生, 故该事件可表示为 $AB \overline{C}$ 或 $AB - C$;

(4) 该事件为 $A \overline{B} \overline{C} \cup \overline{A} B \overline{C} \cup \overline{A} \overline{B} C$;

(5) 该事件为 $AB \cup BC \cup CA$;

(6) 该事件为 $\overline{A} \overline{B} \cup \overline{B} \overline{C} \cup \overline{C} \overline{A}$.

§ 1.2 随机事件的概率

研究随机现象, 我们不仅需要知道它可能出现哪些事件, 更需要知道每个事件出现的可能性的. 所谓事件的概率, 就是刻画事件出现的可能性大小的一种数量指标, 这个数量指标应满足以下两个要求:

(1) 它应是事件本身固有的, 不随人们的意志而改变的一种客观属性量度.

(2) 它必须符合一般常情, 即事件发生可能性大的, 它的值就大; 事件发生可能性小的, 它的值就小.

1. 频率与概率

定义 1.2.1 设在相同的条件下, 进行了 n 次试验, 在这 n 次试验中事件 A 出现了 m 次, 则称

$$f_n(A) = \frac{m}{n} = \frac{\text{事件 } A \text{ 出现的次数}}{\text{试验的总次数}} \quad (1.2.1)$$

为随机事件 A 在 n 次试验中出现的频率, m 称为频数.

由频率的定义, 容易得出它具有下列基本性质:

$$0 \leq f_n(A) \leq 1, f_n(\Omega) = 1 \quad (f_n(\emptyset) = 0).$$

若 $A_1, A_2, \dots, A_n$ 是两两互不相容事件, 则

$$f_n(A_1 \cup A_2 \cup \dots \cup A_n) = f_n(A_1) + f_n(A_2) + \dots + f_n(A_n).$$

事件的频率刻画了事件发生的频繁程度, 由(1.2.1)式容易看出, 频率越大, 事件 A 发生越频繁, 因而在一次试验中发生的可能性越大. 因此, 我们自然会想到用事件 A 的频率表示 A 在一次试验中发生的可能性的. 但频率是随试验的具体结果而定的, 一般不是确定的值.

应该指出, 随机事件的频率与我们进行的试验有关, 而随机事件的概率则是完全客观地存在着的, 随机事件的频率 $f_n(A)$ 可以看作是它的概率的随机表现. 如: 某人进行 n 次试验, A 发生的次数为 m 次. 另一人再进行 n 次相同的试验, 则 A 发生的次数可能仍是 m 次, 也可能不再是 m 次. 因此, 频率总是对特定的 n 次试验而言的. 大量经验表明, 当试验的次数相当大时, 频率总是稳定于某一常数附近, 即它将以某一常数为中心作微小的摆动, 而发生较大偏离的可能性很小. 这一性质称为**频率的稳定性**. 下面的例子说明了这一点.

例 1.2.1 掷一枚质地均匀的硬币, 出现正面与反面的机会是相等的, 即在大量重复试验中, 出现正面的频率应接近于 0.5. 历史上曾有几位数学家做过该试验, 结果见表 1.2.1.

表 1.2.1

试验者	掷硬币次数	正面朝上次数	频 率
蒲 丰	4 040	2 048	0.506 9
皮尔逊	12 000	6 019	0.501 6
皮尔逊	24 000	12 012	0.500 5

例 1.2.2 瑞典 1935 年官方统计资料显示该年各个月份出生的男婴和女婴人数, 以及生女的频率, 见表 1.2.2.

表 1.2.2

月份	总数	男婴	女婴	生女频率
1	7 280	3 743	3 537	0.486
2	6 957	3 550	3 407	0.490
3	7 883	4 017	3 866	0.490
4	7 884	4 173	3 711	0.471
5	7 892	4 117	3 775	0.478
6	7 609	3 944	3 665	0.482

续表

月份	总数	男婴	女婴	生女频率
7	7 585	3 964	3 621	0.477
8	7 393	3 797	3 596	0.486
9	7 203	3 712	3 491	0.485
10	6 903	3 512	3 391	0.491
11	6 552	3 392	3 160	0.482
12	7 132	3 761	3 371	0.473
全年	88 273	45 682	42 591	0.482 5

由表 1.2.2 可以看出瑞典 1935 年各个月份生女频率在 0.482 5 这个数上下摆动.

上述事实表明: 尽管在某一次试验中可能出现这种结果, 也可能出现那种结果, 但在大量试验中, 只要试验条件相同, 观察到的频率就基本上一致. 这说明随机事件在大量试验中存在着一一种客观规律性, 而频率稳定性正是这种规律性的表现. 频率稳定性是由大量统计得来的, 所以也叫“统计规律性”. 它说明随机事件出现的可能性大小是事件本身固有的一种客观属性, 因而, 可以用频率的稳定值作为事件发生可能性大小的度量指标.

定义 1.2.2(概率的统计定义) 在大量重复试验中, 若事件 A 发生的频率稳定地在某一常数 p 的附近摆动, 则称该常数 p 为事件 A 发生的概率, 并记事件 A 的概率为 $P(A)$, 即 $P(A) = p$.

应指出, 频率是变动的, 而概率则为常数. 当试验次数足够多时, 频率相对稳定, 可把频率作为概率的近似值, 即 $f_n(A) \approx P(A)$. 通常生活中, 我们所说的产品的合格率, 彩票的中奖率, 药物治疗某疾病的有效率都是指频率.

2. 古典概型

按照概率的统计定义去确定一个事件的概率, 就得进行大量重复试验. 但是, 在一些情况下, 由于事件本身具有某种“对称性”, 可直接计算其概率. 这类随机试验具有如下特点:

(1) 试验的样本空间中的样本点只有有限个, 即基本事件的总数有限(有限性);

(2) 在每一次试验中, 每个基本事件发生的可能性相等(等可能性).

通常把具有上述特点的随机现象的数学模型称为古典概型, 它在概率论中首先被研究, 古典概型概率的计算在产品质量抽样检查等实际问题中有重要的应用.

定义 1.2.3(概率的古典定义) 设试验结果共有 n 个基本事件 $\omega_1, \omega_2, \dots, \omega_n$, 而且这些事件发生的可能性相等. 事件 A 由其中的 m 个基本事件组成, 则

事件 A 的概率为

$$P(A) = \frac{\text{事件 } A \text{ 包含的基本事件数}(m)}{\text{基本事件总数}(n)}. \quad (1.2.2)$$

例 1.2.3 某城市电话号码升位后为八位数, 且第一位为 6 或 8. 求

(1) 从该城市中随机抽取一个电话号码为不重复的八位数的概率;

(2) 从该城市中随机抽取一个电话号码后三位数都是 8 的概率.

解 先考虑基本事件总数, 注意到除第一位外, 其余位数可取自 0 到 9 这 10 个数字中的任意一个, 因此有 10 种可能结果. 又第一位数只能是 6 和 8, 故只有 2 种可能结果, 由此得, 基本事件总数 $n = 2 \times 10^7$.

(1) 用 A 表示“抽取的一个电话号码为不重复的八位数”, 事件 A 中要求号码不重复, 因此得到事件 A 所包含的基本事件数 $m = 2 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3$, 由概率的古典定义得

$$P(A) = \frac{2 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3}{2 \times 10^7} \approx 0.018.$$

(2) 用 B 表示“抽取的一个电话号码后三位数都是 8”, 则事件 B 所包含的基本事件数 $m = 2 \times 10^4$. 于是求得

$$P(B) = \frac{2 \times 10^4}{2 \times 10^7} = 0.001.$$

例 1.2.4 设有一批产品共计 100 件, 其中有 80 件是正品, 20 件为次品, 现按以下两种方法抽取产品: (a) 每次任意抽取一件, 经观察后放回, 再从中共任取下一件, 这种抽取方法叫做有放回抽样. (b) 每次任取一件, 经观察后不放回, 在剩下的产品中再任取一件, 这种抽取方法叫做不放回抽样. 分别按这两种抽样方法, 求从这 100 件产品中任意抽取 3 件, 其中有 2 件次品的概率.

解 用 A 表示“任意抽取 3 件, 其中有 2 件次品”.

(a) 先考虑有放回的抽样方法: 由于每次抽取后都放回, 故每次抽取产品都是从 100 件中抽取, 则从 100 件中任意抽取 3 件的所有可能的取法共有 100^3 种, 即基本事件总数 $n = 100^3$. 再考虑事件 A 所含的基本事件数. 由于任取 3 件中有 2 件次品的所有可能取法有 C_3^2 种, 而 2 件次品是从 20 件次品中任意取出的, 可能的取法有 20^2 , 另一件是从 80 件正品中任意抽取的, 有 80 种取法. 故事件 A 所含的基本事件数 $m = C_3^2 20^2 \cdot 80$. 由概率的古典定义得

$$P(A) = \frac{C_3^2 20^2 \cdot 80}{100^3} = 0.096.$$

(b) 再按不放回的抽样方法: 由于每抽取一件观察后不放回, 故第一次从 100 件中任意抽取一件, 而第二次是从第一次抽取后剩下的 99 件中任取一件, 再接着第三次是从第二次抽取后剩下的 98 件中任取一件, 从而基本事件

总数 $n = 100 \times 99 \times 98$. 事件 A 所含基本事件数 $m = C_3^2 \times 20 \times 19 \times 80$ 由此得

$$P(A) = \frac{C_3^2 \times 20 \times 19 \times 80}{100 \times 99 \times 98} \approx 0.094.$$

通过本例题可以看出, 采用上述两种抽样方法所求同一事件的概率是不同的. 但当产品的总数较大, 而抽取的数量相对来说较小时, 在这两种抽样方式下所计算的概率值相差不大. 故在实际问题中, 当产品的总数较大时, 可把不放回的抽样视为有放回抽样.

例 1.2.5 在 N 件产品中已知有 M 件次品, 现从中随机抽取 n 件 (取后不放回). 问这 n 件中恰有 k 件次品的概率是多少?

解 设 $A = \{\text{任取 } n \text{ 件中恰有 } k \text{ 件次品}\}$, 基本事件总数是 C_N^n , 抽取的 n 件中的 k 件次品只能取自 M 件次品, 共有 C_M^k 种取法. 而剩下的 $n - k$ 件正品只能在 $N - M$ 件正品中抽取, 共有 C_{N-M}^{n-k} 种取法. 于是事件 A 所包含的基本事件数为 $C_M^k \cdot C_{N-M}^{n-k}$, 由此得

$$P(A) = \frac{C_M^k \cdot C_{N-M}^{n-k}}{C_N^n},$$

其中, $0 \leq n \leq N$, $0 \leq k \leq M$, $0 \leq n - k \leq N - M$.

例 1.2.6 把 n 个不同的球以同样的概率随机放入 $N (N \geq n)$ 个盒子中, 假定每个盒子能容纳的球数没有限制, 试求下列各事件的概率:

- (1) 某指定的 n 个盒子中各有一个球;
- (2) 任意 n 个盒子中各有一个球;
- (3) 有一个盒子中恰有 $m (m \leq n)$ 个球.

解 由于每一个球可以放入 N 个盒子中的任意一个, 所以共有 N 种不同的方法. n 个球放入 N 个盒子相当于从 N 个元素中有放回地选取 n 个进行排列, 应有 N^n 种不同的方法, 故基本事件总数为 N^n .

(1) 记 $A = \{\text{某指定的 } n \text{ 个盒子中各有一球}\}$, 事件 A 包含的基本事件数, 相当于 n 个球在指定的 n 个盒子中进行全排列, 总数为 $n!$, 故

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

(2) 记 $B = \{\text{任意 } n \text{ 个盒子中各有一球}\}$, 那么 B 包含的基本事件应该是任意选出 n 个盒子, 在这 n 个盒子中各放入一个球. 而这 n 个盒子可选自 N 个盒子中的任何 n 个, 有 C_N^n 种选法, 对应于每种选法都有 $n!$ 种放入球的方式, 故 B 中含 $C_N^n \cdot n!$ 个基本事件, 所以

$$P(B) = \frac{C_N^n \cdot n!}{N^n} = \frac{N!}{N^n (N - n)!}.$$

(3) 记 $C = \{\text{有一个盒子中恰好有 } m \text{ 个球}\}$, 这个盒子有 N 种选法, m 个

球可在 n 个球中任意选, 有 C_n^m 种选法, 其余 $n-m$ 个球可任意放在 $N-1$ 个盒子中, 共有 $(N-1)^{n-m}$ 种放法. 因此 C 包含的基本事件数为 $N \cdot C_n^m \cdot (N-1)^{n-m}$. 故

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

这个例子是古典概型中一个很典型的问题, 有许多实际问题的结构形式与它相同. 例如生日问题: 要求参加某次集合的 n 个人中没有两人生日相同的概率. 若把 n 个人看作上面问题中的 n 个球, 而把一年的 365 天作为盒子, 则 $N=365$, 这时, 由(2)便给出所求事件的概率.

3. 几何概率

古典概型的概率定义只适用于试验的可能结果是有限个, 且具有等可能性的情况, 因此有一定的局限性. 在实际问题中, 我们经常会遇到样本空间中的样本点有无穷多的情况, 若等可能的条件依旧成立, 仿照古典概型的概率计算公式, 就得到了几何概率的定义及其计算方法.

定义 1.2.4 如果试验 E 的可能结果可以几何地表示为某区域 Ω 中的一个点(区域 Ω 可以是一维的, 二维的, 三维的, $\dots$), 并且点落在 Ω 中某区域 A 的概率与 A 的测度(长度, 面积, 体积等)成正比, 而与 A 的位置及形状无关, 则随机点落在区域 A 的概率为

$$P(A) = \frac{A \text{ 的测度}}{\Omega \text{ 的测度}} = \frac{mA}{m\Omega}, \quad (1.2.3)$$

称(1.2.3)式定义的概率为几何概率.

例 1.2.7 在线段 $[0,3]$ 上任投一点, 求此点坐标小于 1 的概率.

解 当且仅当所投点落在区间 $[0,1)$ 内时, 此点坐标小于 1, 而必然事件 Ω 就是区间 $[0,3]$. 故按几何概率的定义立即可得所求概率

$$p = \frac{m[0,1)}{m[0,3]} = \frac{1}{3}.$$

例 1.2.8 向区间 $[0,1]$ 上任意投两点, 求此两点间距离小于 $\frac{1}{2}$ 的概率.

解 以 x, y 分别表示两点的坐标, 故

$$0 \leq x \leq 1, \quad 0 \leq y \leq 1.$$

现在把 (x, y) 以平面直角坐标系上的点表示出现, 则一切可能试验结果的集合 Ω 就是单位正方形.

如图 1.2.1 所示. 而“两点间距离小于 $\frac{1}{2}$ ”, 即

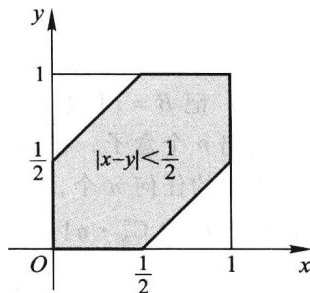


图 1.2.1

$|x-y| < \frac{1}{2}$, 满足这个条件的点 (x, y) 的集合是正方形中两条斜线中间的六边形(阴影部分), 于是所求概率正是投点落在六边形中的概率

$$p = \frac{1^2 - \left(\frac{1}{2}\right)^2}{1^2} = \frac{3}{4}.$$

例 1.2.9 (蒲丰投针问题) 平面上画着一些平行线, 它们之间的距离都等于 d , 向此平面上任投一长度为 $l (l < d)$ 的针, 求此针与任一平行线相交的概率.

解 以 x 表示针的中点到最近的一条平行线的距离, φ 表示针与平行线所成的夹角, 显然有

$$0 \leq x \leq \frac{d}{2}, \quad 0 \leq \varphi \leq \pi.$$

针与平行线的位置关系由图 1.2.2 所示, x 和 φ 完全决定了针的位置, x, φ 的取值范围在平面直角坐标系中确定了一个矩形 $\Omega = \{(x, \varphi) \mid 0 \leq x \leq \frac{d}{2}, 0 \leq \varphi \leq \pi\}$.

为使针与平行线相交, 必须满足关系

$$x \leq \frac{l}{2} \cdot \sin \varphi,$$

这个不等式所确定的区域记为 A , 于是该问题归结为向矩形 Ω 内任投一点, 该点落在区域 A 中的概率即为所求概率.

$$p = \frac{mA}{m\Omega} = \frac{\int_0^\pi \frac{l}{2} \sin \varphi d\varphi}{\frac{d}{2} \pi} = \frac{2l}{\pi d}.$$

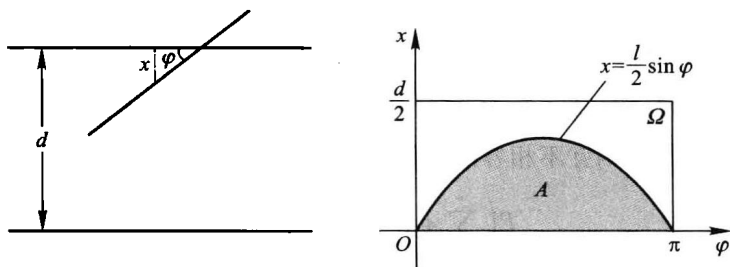


图 1.2.2

4. 概率的公理化定义

上面我们给出了随机事件概率的统计定义、古典定义和几何概率, 并计算了一些事件的概率, 每种定义都有其局限性. 概率的古典定义和几何定义

的假设前提是等可能性,但不总能成立;而对于概率的统计定义,在实际应用中我们无法知道 n 应取多大,也没有认为试验 n 次计算出的频率要比试验 $n-1$ 次计算出的频率更逼近于发生某一事件的概率. 苏联数学家柯尔莫哥洛夫 (A. N. Kolmogorov) 在总结前人的研究基础上于 1933 年给出了概率的公理化定义. 严格地叙述该定义需要涉及较深的数学知识,这里只简述如下:

定义 1.2.5 (概率的公理化定义) 设随机试验 E 的样本空间为 Ω . 若对于 E 的每一个事件 A , 都有一实数 $P(A)$ 与之对应, 并且满足下列三条公理, 那么, 称 $P(A)$ 为事件 A 的概率.

公理 1 非负性: 对于任意一个事件 A , $0 \leq P(A) \leq 1$; (1.2.4)

公理 2 规范性: $P(\Omega) = 1$; (1.2.5)

公理 3 可列可加性: 对于两两互不相容事件 $A_1, A_2, \dots, A_n, \dots$, 有

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \quad (1.2.6)$$

§ 1.3 概率的基本运算法则

1. 概率的加法公式

定理 1.3.1 若事件 A, B 互不相容, 则

$$P(A+B) = P(A) + P(B). \quad (1.3.1)$$

证明 仅就古典概型证明, 设试验的基本事件总数为 n , 其中事件 A 包含的基本事件数为 m_1 , 事件 B 包含的基本事件数为 m_2 , 因为 A 与 B 互不相容, 所以 A 包含的 m_1 个基本事件与 B 包含的 m_2 个基本事件是完全不同的, 故事件 $A+B$ 包含的基本事件数为 $m_1 + m_2$. 由概率的古典定义得

$$P(A+B) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n} = P(A) + P(B).$$

不难把上面结果推广到任意有限个互不相容事件的情形, 即若事件 $A_1, A_2, \dots, A_n$ 是 n 个两两互不相容的事件, 则

$$P\left(\sum_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i). \quad (1.3.2)$$

推论 1 设 $\bar{A}$ 为 A 的对立事件, 则

$$P(\bar{A}) = 1 - P(A). \quad (1.3.3)$$

证明 由于 $A + \bar{A} = \Omega$, 且 A 与 $\bar{A}$ 也是互不相容事件, 由公式 (1.3.1) 得

$$P(A) + P(\bar{A}) = P(A + \bar{A}) = P(\Omega) = 1.$$

因此

$$P(\bar{A}) = 1 - P(A).$$

公式(1.3.3)的形式虽然很简单,但相当有用,它为计算某些较复杂事件的概率提供了很大方便.

例 1.3.1 设有产品 50 件,其中有 3 件次品,其余均为合格品,今从中任意抽取 4 件,求至少有一件次品的概率.

解 设 A 表示“至少取到一件次品”, A_i 表示“恰好取到 i 件次品”($i = 0, 1, 2, 3$), 则 $A = A_1 + A_2 + A_3$. 由于 A_1, A_2, A_3 两两互不相容,利用公式(1.3.2)得

$$\begin{aligned} P(A) &= P(A_1 + A_2 + A_3) = P(A_1) + P(A_2) + P(A_3) \\ &= \frac{C_3^1 \cdot C_{47}^3}{C_{50}^4} + \frac{C_3^2 \cdot C_{47}^2}{C_{50}^4} + \frac{C_3^3 \cdot C_{47}^1}{C_{50}^4} \approx 0.2255. \end{aligned}$$

另解 因为 A 的对立事件 $\bar{A}$ 表示“取到全是合格品”. 于是

$$P(\bar{A}) = \frac{C_{47}^4}{C_{50}^4},$$

由公式(1.3.3)得

$$P(A) = 1 - P(\bar{A}) = 1 - \frac{C_{47}^4}{C_{50}^4} \approx 0.2255.$$

推论 2 设事件 $A \supset B$, 则

$$P(A - B) = P(A) - P(B). \quad (1.3.4)$$

证明 因为 $A \supset B$, 故 $A = (A - B) \cup B$, 并且事件 $(A - B)$ 与 B 为互不相容事件, 由公式(1.3.1)得

$$P(A) = P(A - B) + P(B),$$

因此

$$P(A - B) = P(A) - P(B).$$

定理 1.3.2 若 A, B 为任意两事件, 则

$$P(A \cup B) = P(A) + P(B) - P(AB). \quad (1.3.5)$$

证明 因为 $A \cup B = A + (B - AB)$, 而 A 与 $(B - AB)$ 互不相容, 同时 $B \supset AB$, 所以由公式(1.3.1)和(1.3.4)得

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

类似地, 对于任意三个事件 A, B, C 有

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(AB) - P(BC) \\ &\quad - P(AC) + P(ABC). \end{aligned}$$

例 1.3.2 袋中装有红、白、黑球各一个, 每次从袋中任取一个球, 记录其颜色以后再放回袋中, 这样连取 3 次(有放回地抽取). 求 3 次都没有取到红球或 3 次都没有取到白球的概率.

解 设 $A = \{3 \text{ 次没有取到红球}\}$, $B = \{3 \text{ 次没有取到白球}\}$, 所求事件的概率为 $P(A \cup B)$. 利用公式(1.3.5)

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

其中 $P(A) = P(B) = \frac{2^3}{3^3}$, 而 $AB = \{3 \text{ 次既没有取到红球又没有取到白球}\} = \{3 \text{ 次取到黑球}\}$, 于是 $P(AB) = \frac{1}{3^3}$, 由此

$$P(A \cup B) = \frac{2^3}{3^3} + \frac{2^3}{3^3} - \frac{1}{3^3} = \frac{5}{9}.$$

2. 条件概率与事件的独立性

先考察一个随机试验: 袋中装有 16 个球, 其中 6 个是玻璃球, 另外 10 个是塑料球. 而玻璃球中有 2 个是红色, 4 个是蓝色, 塑料球中有 3 个是红色, 7 个是蓝色. 现从中任取一个, 若令 $A = \{\text{取到蓝色球}\}$, 则 $P(A) = \frac{11}{16}$. 若令 $B = \{\text{取到玻璃球}\}$, 则 $P(B) = \frac{6}{16}$. 因而 $AB = \{\text{取到蓝色的玻璃球}\}$, 且 $P(AB) = \frac{4}{16}$.

在实际问题中, 除了要知道事件 B 的概率 $P(B)$ 外, 有时还需要知道在“事件 A 已发生的”条件下, 事件 B 发生的概率, 这个概率称为条件概率, 记为 $P(B|A)$. 也就是要求在已知取到蓝色球的条件下, 该球又是玻璃球的概率. 将袋中球的分类情况列表如下:

	玻璃	塑料	合计
红	2	3	5
蓝	4	7	11
合计	6	10	16

因为已知取到的是蓝色球, 此时的样本空间仅由 11 个基本事件组成, 而这 11 个蓝色球中有 4 个是玻璃球, 所以 $P(B|A) = \frac{4}{11}$. 即在事件 A 发生的条件下, 原来的样本空间(16 个基本事件)被缩小, 显然 $P(B|A) \neq P(B)$.

值得注意的是

$$P(B|A) = \frac{4}{11} = \frac{\frac{4}{16}}{\frac{11}{16}} = \frac{P(AB)}{P(A)}.$$

$$P(A|B) = \frac{4}{6} = \frac{\frac{4}{16}}{\frac{6}{16}} = \frac{P(AB)}{P(B)}.$$

这两个式子对一般情况也成立. 为此, 我们给出条件概率的定义:

定义 1.3.1 设 A, B 是两事件, 且 $P(A) > 0$, 称

$$P(B|A) = \frac{P(AB)}{P(A)} \quad (1.3.6)$$

为在事件 A 已发生的条件下事件 B 发生的条件概率.

类似地可以定义在事件 B 发生的条件下事件 A 发生的条件概率

$$P(A|B) = \frac{P(AB)}{P(B)} (P(B) > 0). \quad (1.3.7)$$

例 1.3.3 某单位 100 名员工做血压和肝功能指标检查. 结果有 95 人血压正常(事件 A), 94 人肝功能正常(事件 B), 92 人两项检查都正常. 在这个人群中, 如随机抽查一人, 求 $P(A)$, $P(B)$, $P(A|B)$, $P(B|A)$, $P(AB)$.

解 这是古典概型, 由定义

$$P(A) = \frac{95}{100} = 0.95.$$

$$P(B) = \frac{94}{100} = 0.94.$$

求 $P(A|B)$, 实际上是求在事件 B 发生的条件下, A 发生的概率. 即已知某人肝功能正常, 在此条件下做血压检查, 此人血压正常的概率.

$$P(A|B) = \frac{92}{94} \approx 0.98.$$

同理
$$P(B|A) = \frac{92}{95} \approx 0.97.$$

而
$$P(AB) = \frac{92}{100} = 0.92.$$

条件概率 $P(B|A)$, $P(A|B)$ 仍为概率, 只是在缩小的样本空间 Ω_A , Ω_B 中分别计算 B 事件、 A 事件发生的概率, 所以条件概率同样满足概率的公理化定义中的三条公理:

(1) 非负性: $0 \leq P(A|B)$;

(2) 规范性: $P(\Omega|B) = 1$;

(3) 可列可加性: $P\left(\bigcup_{i=1}^{\infty} A_i | B\right) = \sum_{i=1}^{\infty} P(A_i | B)$, 其中 $A_1, A_2, \dots, A_n, \dots$ 为

可列个两两互不相容的事件.

由 (1.3.6) 式和 (1.3.7) 式可得

$$P(AB) = P(A)P(B|A) \quad (P(A) > 0) \quad (1.3.8)$$

$$= P(B)P(A|B) \quad (P(B) > 0), \quad (1.3.9)$$

上述两式称为概率的乘法公式.

例 1.3.4 在 100 件产品中有 3 件次品, 现从中连取两次, 每次取一件, 取后不放回, 求下列事件的概率:

(1) 两次都是正品;

(2) 一件正品, 一件次品.

解 设 $A_1 = \{\text{第一次取到正品}\}$, $A_2 = \{\text{第二次取到正品}\}$.

(1) $A_1 A_2 = \{\text{两件都是正品}\}$.

$$P(A_1 A_2) = P(A_1)P(A_2|A_1) = \frac{97}{100} \times \frac{96}{99} \approx 0.94.$$

(2) “一件正品和一件次品”可用事件 $A_1 \bar{A}_2 + \bar{A}_1 A_2$ 表示, 并且 $A_1 \bar{A}_2, \bar{A}_1 A_2$ 互不相容. 于是

$$\begin{aligned} P(A_1 \bar{A}_2 + \bar{A}_1 A_2) &= P(A_1 \bar{A}_2) + P(\bar{A}_1 A_2) \\ &= P(A_1)P(\bar{A}_2|A_1) + P(\bar{A}_1)P(A_2|\bar{A}_1) \\ &= \frac{97}{100} \times \frac{3}{99} + \frac{3}{100} \times \frac{97}{99} \approx 0.059. \end{aligned}$$

乘法公式还可以推广到有限多个事件的情况, 即

$$P(A_1 A_2 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 A_2) \cdots P(A_n|A_1 A_2 \cdots A_{n-1}).$$

(1.3.10)

例 1.3.5 假设在一个罐中有 b 只黑球及 r 只白球, 随机取出一只, 把原球放回, 并加进去与取出球同色的球 c 只, 再摸第 2 次, 这样下去共摸了 n 次, 问前面 n_1 次出现黑球, 后面 $n_2 = n - n_1$ 次出现白球的概率是多少?

解 设 $A_i = \{\text{第 } i \text{ 次摸出黑球}\} (i = 1, 2, \cdots, n_1)$. $A_j = \{\text{第 } j \text{ 次摸出白球}\} (j = n_1 + 1, n_1 + 2, \cdots, n)$, 则

$$P(A_1) = \frac{b}{b+r}, \quad P(A_2|A_1) = \frac{b+c}{b+r+c},$$

$$P(A_3|A_1 A_2) = \frac{b+2c}{b+r+2c}, \quad \cdots$$

$$P(A_{n_1}|A_1 A_2 \cdots A_{n_1-1}) = \frac{b+(n_1-1)c}{b+r+(n_1-1)c},$$

$$P(A_{n_1+1}|A_1 A_2 \cdots A_{n_1}) = \frac{r}{b+r+n_1 c},$$

$$P(A_{n_1+2}|A_1 A_2 \cdots A_{n_1+1}) = \frac{r+c}{b+r+(n_1+1)c}, \quad \cdots,$$

$$P(A_n | A_1 A_2 \cdots A_{n-1}) = \frac{r + (n_2 - 1)c}{b + r + (n - 1)c},$$

因此

$$\begin{aligned} P(A_1 A_2 \cdots A_n) &= P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 A_2) \cdots P(A_n | A_1 A_2 \cdots A_{n-1}) \\ &= \frac{b}{b+r} \cdot \frac{b+c}{b+r+c} \cdot \frac{b+2c}{b+r+2c} \cdots \frac{b+(n_1-1)c}{b+r+(n_1-1)c} \cdot \\ &\quad \frac{r}{b+r+n_1c} \cdot \frac{r+c}{b+r+(n_1+1)c} \cdots \frac{r+(n_2-1)c}{b+r+(n-1)c}. \end{aligned}$$

注意这个答案只与黑球及白球出现的次数有关,而与出现的顺序无关.该模型常用作描述传染病的数学模型.

在例 1.3.4 中,若将取球的方式改为有放回,则

$$P(A_1 A_2) = \frac{97}{100} \times \frac{97}{100} = P(A_1) P(A_2),$$

$$P(A_1 \bar{A}_2) = \frac{97}{100} \times \frac{3}{100} = P(A_1) P(\bar{A}_2).$$

由此可以看出,尽管是求同一事件的概率,由于取法不同,造成事件的概率也不相等.对于不放回的取法,事件 A_1 的发生与否将影响到事件 A_2 发生的概率,而对于有放回的取法,事件 A_2 发生的概率不因事件 A_1 发生与否而受影响,此时,我们称两个事件相互独立.

定义 1.3.2 设事件 A 与 B , 若有

$$P(AB) = P(A)P(B), \quad (1.3.11)$$

则称事件 A 与 B 相互独立.

若 A, B 相互独立,由条件概率的定义及式(1.3.11)得

$$\begin{aligned} P(B|A) &= \frac{P(AB)}{P(A)} = \frac{P(A)P(B)}{P(A)} \\ &= P(B) \quad (P(A) > 0), \end{aligned} \quad (1.3.12)$$

同理,若 A, B 相互独立, $P(B) > 0$, 则

$$P(A|B) = P(A). \quad (1.3.13)$$

式(1.3.12)与(1.3.13)都与(1.3.11)等价,故它们也可作为两个事件相互独立的定义.

定理 1.3.3 若事件 A, B 相互独立,则事件 $\bar{A}$ 与 B, A 与 $\bar{B}, \bar{A}$ 与 $\bar{B}$ 也相互独立.

证明

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

故 $\bar{A}$ 与 B 也相互独立.

类似地, 也可证明 A 与 $\bar{B}$, $\bar{A}$ 与 $\bar{B}$ 相互独立.

两个事件相互独立的概念可推广到有限多个事件的情形.

定义 1.3.3 设 $A_1, A_2, \dots, A_n$ 为 $n(n \geq 2)$ 个事件, 若对其中任意的 $k(2 \leq k \leq n)$ 个事件 $A_{i_1}, A_{i_2}, \dots, A_{i_k}(1 \leq i_1 < i_2 < \dots < i_k \leq n)$, 满足等式

$$P(A_{i_1} A_{i_2} \cdots A_{i_k}) = P(A_{i_1}) P(A_{i_2}) \cdots P(A_{i_k}), \quad (1.3.14)$$

则称事件 $A_1, A_2, \dots, A_n$ 相互独立.

当事件 $A_1, A_2, \dots, A_n$ 相互独立时, 有

$$P(A_1 A_2 \cdots A_n) = P(A_1) P(A_2) \cdots P(A_n).$$

需要指出: 由定义知 n 个事件相互独立必须有 $2^n - n - 1$ 个式子成立, 而 n 个事件中两两相互独立只需下列 C_n^2 个式子

$$P(A_i A_j) = P(A_i) P(A_j) \quad (i \neq j; i, j = 1, 2, \dots, n)$$

成立. 因此, n 个事件相互独立, 则一定两两独立. 但两两独立, 得不出 n 个事件相互独立.

例如, 对于三个事件 A, B, C 相互独立是指下列等式同时成立

$$P(ABC) = P(A)P(B)P(C),$$

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

$$P(AC) = P(A)P(C),$$

$$P(BC) = P(B)P(C).$$

仅后三个等式成立, 称 A, B, C 两两相互独立. 下面给出三事件两两独立, 但不是相互独立的例子.

例 1.3.6 设有四张卡片, 其中三张分别涂有红色、白色和黑色, 而剩下的一张同时涂有红、白、黑三色, 从中任意抽取一张, 记事件 $A = \{\text{抽取的卡片上有红色}\}$, $B = \{\text{抽取的卡片上有白色}\}$, $C = \{\text{抽取的卡片上有黑色}\}$, 则

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

$$P(AB) = P(AC) = P(BC) = \frac{1}{4},$$

从而 A, B, C 两两相互独立, 但

$$P(ABC) = \frac{1}{4} \neq \frac{1}{8} = P(A)P(B)P(C),$$

故 A, B, C 不是相互独立的.

下面再给出一例子说明有 $P(ABC) = P(A)P(B)P(C)$ 成立, 但 A, B, C 不两两独立.

例 1.3.7 设有八张卡片, 其中一张写有 1, 2, 3 三个数字, 另有两张写

有 1, 2 两个数字, 剩下的五张中有三张写有数字 3, 一张写有数字 2, 一张写有数字 1. 用 A, B, C 分别表示“抽出的一张卡片上有数字 i ” ($i=1, 2, 3$). 则

$$P(A) = P(B) = P(C) = \frac{4}{8} = \frac{1}{2},$$

$$P(AB) = \frac{3}{8}, P(AC) = P(BC) = \frac{1}{8},$$

$$P(ABC) = \frac{1}{8} = P(A)P(B)P(C).$$

但 $P(AB) = \frac{3}{8} \neq P(A)P(B)$, $P(AC) = \frac{1}{8} \neq P(A)P(C)$, $P(BC) = \frac{1}{8} \neq P(B)P(C)$.

由此可见, 当 $P(ABC) = P(A)P(B)P(C)$ 成立时, 一般也不能保证 A, B, C 两两独立成立.

在实际问题中, 判定事件间的独立性, 我们常常不是根据定义来判断, 而是根据实际情况作出判断的. 例如, 有放回的抽取方式两次的抽取结果是相互独立的. 但在具体问题中, 一般都采用不放回的抽取方式, 如产品质量的抽样检查, 此时前后两次抽取结果并不相互独立. 但当产品总数较大时, 前面的抽取对后面的抽取影响很微小, 因而任两次抽取结果亦可以看作是相互独立的.

例 1.3.8 加工某一种零件需要经过三道工序, 设三道工序的次品率分别为 2%, 1%, 5%, 并假设各道工序互不影响, 求加工出来的零件的次品率.

解 设 A_i 表示“第 i 道工序出现次品” ($i=1, 2, 3$), 依题意 A_1, A_2, A_3 相互独立, 且 $P(A_1) = 0.02$, $P(A_2) = 0.01$, $P(A_3) = 0.05$, 又设 A 表示“加工出来的零件是次品”, 则 $A = A_1 \cup A_2 \cup A_3$.

$$\begin{aligned} P(A) &= 1 - P(\bar{A}) = 1 - P(\overline{A_1 \cup A_2 \cup A_3}) \\ &= 1 - P(\bar{A}_1 \bar{A}_2 \bar{A}_3) \\ &= 1 - (1 - 0.02)(1 - 0.01)(1 - 0.05) \\ &\approx 0.0783. \end{aligned}$$

例 1.3.9 设炮兵使用某型号的高射炮, 每一门炮一发击中飞机的概率为 0.2, 问需要多少门炮同时射击, 才能以 90% 的把握一发就击中来犯的敌机?

解 设需要 n 门炮, 并用 $A_i = \{\text{第 } i \text{ 门炮击中敌机}\}$ ($i=1, 2, \dots, n$), $A = \{\text{敌机被击中}\}$, 则

$$A = A_1 \cup A_2 \cup \dots \cup A_n,$$

于是要求 n , 使得 $P(A) = P\left(\sum_{i=1}^n A_i\right) \geq 0.90$.

由于事件 $A_i (i=1, 2, \dots, n)$ 相互独立, 因而事件 $\bar{A}_i (i=1, 2, \dots, n)$ 也是相互独立的. 所以

$$\begin{aligned} P(A) &= 1 - P(\bar{A}) = 1 - P(\bar{A}_1 \bar{A}_2 \cdots \bar{A}_n) \\ &= 1 - P(\bar{A}_1)P(\bar{A}_2) \cdots P(\bar{A}_n) = 1 - (0.8)^n, \end{aligned}$$

从而

$$1 - (0.8)^n \geq 0.90,$$

解得

$$n \geq \lg 0.1 / \lg 0.8 \approx 10.32.$$

故至少需要 11 门高射炮方能以 90% 以上的把握击中来犯的敌机.

例 1.3.10 假若每个人血清中含有肝炎病毒的概率为 0.4%, 混合 100 个人的血清, 求此血清中含有肝炎病毒的概率.

解 以 A_i 表示“第 i 个人的血清中含有肝炎病毒” ($i=1, 2, 3, \dots, 100$), 可认为它们相互独立, 所求概率为

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_{100}) &= 1 - P(\bar{A}_1) \cdot P(\bar{A}_2) \cdots P(\bar{A}_{100}) \\ &= 1 - (0.996)^{100} \approx 0.33. \end{aligned}$$

虽然每个人的血清中有肝炎病毒的概率很小, 但是混合后的概率很大, 在实际问题中, 这种效应值得注意.

§ 1.4 全概率公式与贝叶斯公式

1. 全概率公式

我们通过一实际问题来引入全概率公式. 在体育比赛或实际生活中, 经常要遇到抽签问题. 假设有 7 张足球票, 作 10 张票签, 10 人轮流抽签, 问先抽后抽得到足球票的概率是否相同?

看法一, 10 人轮流抽签, 机会相等.

看法二, 抽签也有先后, 显然第一人抽到的概率是 $\frac{7}{10}$. 若第一人抽到, 第二人抽到的概率仅有 $\frac{6}{9}$. 如果第一人没有抽到, 第二人抽到的概率就是 $\frac{7}{9}$. 抽签未必机会相等.

看法二分析的似乎很有道理, 但实际算出来的两个数 $\frac{6}{9}$ 和 $\frac{7}{9}$, 不是第二人抽到的概率, 而是在第一人抽到或抽不到的条件下, 第二人抽到的条件概率.

实际上, 先抽与后抽却是一样的, 分析如下:

设 $A = \{\text{第一人抽到足球票}\}$, $B = \{\text{第二人抽到足球票}\}$.

相对于 A 而言, B 的发生只有两个可能: $B = AB + \bar{A}B$. 由公式 (1.3.1) 及 (1.3.8),

$$\begin{aligned} P(B) &= P(AB) + P(\bar{A}B) \\ &= P(A)P(B|A) + P(\bar{A})P(B|\bar{A}) \\ &= \frac{7}{10} \cdot \frac{6}{9} + \frac{3}{10} \cdot \frac{7}{9} = \frac{7}{10} = P(A). \end{aligned} \quad (1.4.1)$$

所以第一人和第二人抽到的概率相等. 同理可以算出第三人, 第四人, ... 直到最后一人抽到的概率都等于 $\frac{7}{10}$.

上述分析的实质是将一个复杂事件分解为较简单的几个事件, 然后将概率的加法公式和乘法公式结合起来, 这就产生了所谓的全概率公式.

定理 1.4.1 设事件 $A_1, A_2, \dots, A_n$ 两两互不相容, $P(A_i) > 0 (i = 1, 2, \dots, n)$, 且 $\sum_{i=1}^n A_i = \Omega$, 则对任一事件 B , 有

$$P(B) = \sum_{i=1}^n P(A_i)P(B|A_i). \quad (1.4.2)$$

证明 因为 $A_1, A_2, \dots, A_n$ 两两互不相容, 所以 $A_1B, A_2B, \dots, A_nB$ 也两两互不相容, 且

$$B = B\Omega = A_1B + A_2B + \dots + A_nB.$$

利用互不相容事件的概率加法公式及乘法公式得

$$P(B) = P\left(\sum_{i=1}^n A_iB\right) = \sum_{i=1}^n P(A_iB) = \sum_{i=1}^n P(A_i)P(B|A_i).$$

公式 (1.4.2) 称为全概率公式. 满足全概率公式中的事件组 $A_1, A_2, \dots, A_n$ 叫完备事件组. 实际上, 在上述定理中条件 $\sum_{i=1}^n A_i = \Omega$ 可减弱为 $\sum_{i=1}^n A_i \supset B$, 结论亦然成立.

例 1.4.1 设某厂有甲、乙、丙三个车间生产同一种产品, 已知各车间的产量分别占全厂产量的 25%, 35%, 40%. 并且各车间的次品率依次为 5%, 4%, 2%. 现从该厂这批产品中任取一件, 求这批产品的次品率.

解 设 A_1, A_2, A_3 表示“产品分别由甲、乙、丙车间生产”, B 表示“产品为次品”, 显然 A_1, A_2, A_3 构成完备事件组. 依题意, 有

$$\begin{aligned} P(A_1) &= 25\%, \quad P(A_2) = 35\%, \quad P(A_3) = 40\%, \\ P(B|A_1) &= 5\%, \quad P(B|A_2) = 4\%, \quad P(B|A_3) = 2\%. \end{aligned}$$

所以

$$P(B) = \sum_{i=1}^3 P(A_i)P(B|A_i) = 0.0345.$$

例 1.4.2 甲袋中装有 2 个红球, 6 个白球; 乙袋中装有 5 个红球, 4 个白球. 现从甲袋中任取 3 个放入乙袋, 然后再从乙袋中任取一个球, 求取得红球的概率.

解 设 $A_i = \{\text{从甲袋中任取 3 个球中有 } i \text{ 个红球}\}$, $i=0, 1, 2$,
 $B = \{\text{从乙袋中取得 1 个红球}\}.$

由于 $A_0 \cup A_1 \cup A_2 = \Omega$. 根据全概率公式(1.4.2)得

$$\begin{aligned} P(B) &= \sum_{i=0}^2 P(A_i)P(B|A_i) \\ &= \frac{C_2^0 C_6^3}{C_8^3} \cdot \frac{5}{12} + \frac{C_2^1 C_6^2}{C_8^3} \cdot \frac{6}{12} + \frac{C_2^2 C_6^1}{C_8^3} \cdot \frac{7}{12} \\ &= \frac{5}{14} \cdot \frac{5}{12} + \frac{15}{28} \cdot \frac{6}{12} + \frac{3}{28} \cdot \frac{7}{12} \approx 0.48. \end{aligned}$$

例 1.4.3 甲、乙两人进行射击比赛, 每回射击胜者得 1 分, 并设在每回射击中甲胜的概率为 α , 乙胜的概率为 β , $\alpha + \beta = 1$, 比赛进行到有一人比对方多 2 分为止, 多得 2 分者最终获胜, 求甲最终获胜的概率.

解 记 A 表示“甲最终获胜”, B_1 表示“在第一、二回射击中甲均获胜”, B_2 表示“在第一、二回射击中乙均获胜”, B_3 表示“在第一、二回射击中甲、乙各胜一回”. 由题设得

$$P(B_1) = \alpha^2, P(B_2) = \beta^2, P(B_3) = 2\alpha\beta.$$

若 B_1 发生, 即甲获胜, 则 $P(A|B_1) = 1$; 若 B_2 发生, 乙已获胜, 则 $P(A|B_2) = 0$; 若 B_3 发生, 从第三回开始, 比赛又从头开始, 因此 $P(A|B_3) = P(A)$. 由全概率公式得

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\ &= \alpha^2 + 2\alpha\beta P(A), \end{aligned}$$

由此得

$$P(A) = \frac{\alpha^2}{1 - 2\alpha\beta}.$$

2. 贝叶斯公式

定理 1.4.2 设 $A_1, A_2, \dots, A_n$ 构成完备事件组, 则对任一事件 $B (P(B) > 0)$, 有

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^n P(A_i)P(B|A_i)} \quad (k = 1, 2, \dots, n). \quad (1.4.3)$$

证明 由公式(1.3.7)得

$$P(A_k | B) = \frac{P(A_k B)}{P(B)},$$

又根据公式(1.3.8)及(1.4.2)得

$$P(A_k | B) = \frac{P(A_k)P(B | A_k)}{\sum_{i=1}^n P(A_i)P(B | A_i)}.$$

通常把(1.4.3)式称为贝叶斯公式, 它由数学家贝叶斯(Bayes)于1763年首次提出, 其原理在工程技术、经济分析、投资决策、药理的临床检验及疾病的计量诊断诸多方面有十分重要的实用价值.

例 1.4.4 继续讨论例 1.4.1. 若从该批产品任取一件, 该件是次品, 问该次品由哪个车间生产的概率较大.

解 已知 $P(B) = 0.0345$, 由贝叶斯公式

$$P(A_1 | B) = \frac{P(A_1)P(B | A_1)}{P(B)} = \frac{0.25 \times 0.05}{0.0345} \approx 0.362,$$

$$P(A_2 | B) = \frac{P(A_2)P(B | A_2)}{P(B)} = \frac{0.35 \times 0.04}{0.0345} \approx 0.406,$$

$$P(A_3 | B) = \frac{P(A_3)P(B | A_3)}{P(B)} = \frac{0.4 \times 0.02}{0.0345} \approx 0.232.$$

故该次品由乙车间生产的可能性较大.

例 1.4.5 甲胎蛋白(AFP)免疫检测法被普遍用于肝癌的早期诊断和普查. 已知肝癌患者经 AFP 诊断为肝癌的概率为 95%, 而未患肝癌通过 AFP 被误诊为肝癌的概率为 2%, 在人群中肝癌的发病率一般为 0.04%, 现有一人经 AFP 检验诊断为患肝癌, 求此人的确患肝癌的概率.

解 $A = \{\text{肝癌患者}\}$, $B = \{\text{经 AFP 检验诊断为患肝癌}\}$, 则

$$P(A) = 0.0004, \quad P(\bar{A}) = 0.9996,$$

$$P(B | A) = 0.95, \quad P(B | \bar{A}) = 0.02.$$

由贝叶斯公式, 得

$$\begin{aligned} P(A | B) &= \frac{P(A)P(B | A)}{P(A)P(B | A) + P(\bar{A})P(B | \bar{A})} \\ &= \frac{0.0004 \times 0.95}{0.0004 \times 0.95 + 0.9996 \times 0.02} \approx 0.0187. \end{aligned}$$

可见, 虽然 AFP 的诊断价值很大, 但是一次检验诊断为肝癌的人确患肝癌的可能性并不大. 换言之, 一次检验提供的信息量不足以作出判断.

§ 1.5 伯努利概型

将某试验 E 重复进行 n 次, 且各次试验结果互不影响, 则称这 n 次试验为

n 重独立试验. 比如, 投 n 次硬币或进行 n 次有放回摸球都是重复独立试验的简单例子. 下面我们将研究一类最简单、最重要的重复独立试验——伯努利试验.

若试验 E 只有两个可能的结果: A 及 $\bar{A}$, 记 $P(A) = p$, $P(\bar{A}) = 1 - p = q$, 这种试验称为伯努利 (Bernoulli) 试验. 若将试验 E 重复进行 n 次, 且每次试验结果互不影响 (独立的), 则称为 n 重伯努利试验.

在 n 重伯努利试验中, 事件 A 可能发生 $0, 1, \dots, n$ 次, 如何求事件 A 恰好发生 $k (0 \leq k \leq n)$ 次的概率呢? 先看下述例子.

例 1.5.1 设一袋中有 3 个红球, 7 个白球, 从中任取一球, 有放回地共取 5 次, 试求事件 “4 次取到红球” 的概率.

解 记 $A = \{4 \text{ 次取到红球}\}$,

$A_i = \{\text{第 } i \text{ 次取到红球}\}$,

$\bar{A}_i = \{\text{第 } i \text{ 次取到白球}\}, i = 1, 2, \dots, 5,$

则

$$A = A_1 A_2 A_3 A_4 \bar{A}_5 + A_1 A_2 A_3 \bar{A}_4 A_5 + A_1 A_2 \bar{A}_3 A_4 A_5 + A_1 \bar{A}_2 A_3 A_4 A_5 + \bar{A}_1 A_2 A_3 A_4 A_5,$$

即事件 A 是 5 个 (C_5^4 个) 互不相容事件的和, 其中

$$P(A_i) = \frac{3}{10}, P(\bar{A}_i) = \frac{7}{10}, i = 1, 2, \dots, 5.$$

由于各次试验是相互独立的, 所以有

$$\begin{aligned} P(A_1 A_2 A_3 A_4 \bar{A}_5) &= P(A_1) P(A_2) P(A_3) P(A_4) P(\bar{A}_5) \\ &= \left(\frac{3}{10}\right)^4 \cdot \left(\frac{7}{10}\right)^1. \end{aligned}$$

根据概率加法定理. 所求概率为

$$P(A) = C_5^4 \left(\frac{3}{10}\right)^4 \cdot \left(\frac{7}{10}\right)^1.$$

这种概率模型为伯努利概型, 它是概率论发展初期所研究的概率模型之一, 有着广泛的应用.

定理 1.5.1 (伯努利概型计算公式) 设每次试验中, 事件 A 发生的概率为 $p (0 < p < 1)$, 则在 n 次重复独立试验中, A 发生 k 次的概率为

$$P_n(k) = C_n^k p^k (1-p)^{n-k}, k = 0, 1, 2, \dots, n. \quad (1.5.1)$$

例 1.5.2 对某种药物的疗效进行研究, 设这种药物对某种疾病的有效率为 $p = 0.8$, 现有 10 名患此种疾病的患者同时服用此药, 求其中至少有 6 名患者服药有效的概率.

解 这是伯努利概型, $n = 10$, $p = 0.8$, 记 $A = \{\text{至少有 6 名患者服药有效}\}$, 则

$$\begin{aligned}
 P(A) &= P_{10}(6) + P_{10}(7) + P_{10}(8) + P_{10}(9) + P_{10}(10) \\
 &= \sum_{k=6}^{10} C_{10}^k (0.8)^k (0.2)^{10-k} \approx 0.97.
 \end{aligned}$$

实际中,有些试验的结果不止两个,比如,电子元器件、电器产品的寿命可以是大于零的任何一数值.但根据需要,我们可以把寿命大于 1 000 小时的电子元器件当作合格品,低于该数值的作为不合格品.那么,这类问题就可以归结为伯努利试验.

例 1.5.3 一大批某型号的电子器件,已知其一级品率为 0.3,现从中随机抽查 20 只,问其中有一级品的概率是多少?

解 由于这批电子器件总量较大,而抽取的只有 20 只,相对来说较小,故可将抽查 20 只近似地看作有放回抽样.将“抽取 1 只”作为 1 次试验,则“抽取 20 只”相当于 $n=20$ 的伯努利概型.记 $A = \{\text{抽取 20 只,有一级品}\}$,由伯努利概型公式及逆事件关系得

$$\begin{aligned}
 P(A) &= 1 - P(\bar{A}) = 1 - P_{20}(0) = 1 - C_{20}^0 (0.3)^0 \cdot (0.7)^{20} \\
 &= 1 - (0.7)^{20} \approx 1 - 0.000\,797\,9 \\
 &\approx 0.999\,2
 \end{aligned}$$

在本例中,所抽取的 20 只中不含一级品的概率 $P(\bar{A}) \approx 0.000\,797\,9$.不到万分之八.这种“概率很小的事件在一次试验中几乎不可能发生”,又称为小概率事件的实际不可能原理.它是数理统计中进行统计推断的主要依据.

§ 1.6 随机数的生成与应用

为了能在计算机上方便地处理概率和统计方面的问题, MATLAB 特别提供了统计工具箱(Statistics Toolbox).该工具箱中包括 200 多个 m 文件(函数),能够支持范围广泛的统计计算任务.我们将在今后各章中,结合课程内容逐步进行介绍.如果了解统计工具箱中所有函数的情况,可以使用命令

```
help toolbox \ stats ↓
```

来显示统计工具箱中所有函数的名称及功能.

在统计工具箱中, MATLAB 提供了一组随机数产生函数.使用这些函数,可以产生出满足 20 种概率分布(概率分布的概念在下一章中介绍)的随机数.表 1.6.1 中列出了部分随机数产生函数的名称和调用格式.各函数的详细使用方法请查阅相应使用手册或借助联机帮助命令 help 进行在线查询.

表 1.6.1 随机数产生函数

分布类型	函数名称	调用格式
二项分布	binornd	$R = \text{binornd}(N, P, mm, nn)$
χ^2 分布	chi2rnd	$R = \text{chi2rnd}(N, mm, nn)$
指数分布	exprnd	$R = \text{exprnd}(MU, mm, nn)$
F 分布	frnd	$R = \text{frnd}(M, N, mm, nn)$
正态分布	normrnd	$R = \text{normrnd}(MU, SIGMA, mm, nn)$
泊松分布	poissrnd	$R = \text{poissrnd}(LAMBDA, mm, nn)$
学生 t 分布	trnd	$R = \text{trnd}(N, mm, nn)$
离散均匀分布	unidirnd	$R = \text{unidirnd}(N, mm, nn)$
连续均匀分布	unifrnd	$R = \text{unifrnd}(A, B, mm, nn)$

表 1.6.1 中调用格式栏内的参数 mm 和 nn ，用来指明调用函数时返回的是一个 $mm \times nn$ 阶随机数矩阵。如果只需要返回一个随机数，则调用时参数 mm 和 nn 可以省略。

在研究一些概率问题时，可以用实验的方法来做处理，但这往往需要进行成千上万次的重复操作（比如掷硬币），既花费大量人力，又耗费很多时间。而如果借助统计工具箱中这些随机数产生函数，对所研究问题编写一段模拟程序，便可以轻易地完成实验任务，取得事半功倍的效果。

例 1.6.1 掷两颗骰子时，出现点数之和有多少种可能的结果？其中哪一种结果出现的概率最大，其值是多少？（参见习题 1 第 3 题。）

解 掷骰子所出现的点数服从均匀分布，故可建立如下的一段 MATLAB 程序：

```
clear cs
k=30 000;           %掷骰子的次数
n=2;                %骰子的个数
for m=n:6*n
    cs(m)=0;         %清零计数器
end
for m =1: k
    s2=unidrnd(6,1,n); %模拟一次掷 n 颗骰子
    s=sum(s2);         %算点数和
    cs(s)=cs(s)+1;
```

```

end
disp('点数和  出现次数  概率')
for m = n: 6 * n      %显示实验结果
    disp(sprintf('  %2d  %6d  %6.4f',m,cs(m),cs(m)/k))
end

```

其中函数 `unidrnd(6,1,n)` 的功能是按均匀分布从集合 $\{1,2,3,4,5,6\}$ 中产生一个 $1 \times n$ 阶的随机数矩阵。因此语句 `s2 = unidrnd(6,1,n)` 便可以模拟一次掷 n 颗骰子的实验，并将结果存入变量 `s2` 中。

在 MATLAB 中执行该程序，便可得到一组实验结果：

点数和	出现次数	概率
2	815	0.027 2
3	1 628	0.054 3
4	2 443	0.081 4
5	3 310	0.110 3
6	4 191	0.139 7
7	4 979	0.166 0
8	4 271	0.142 4
9	3 391	0.113 0
10	2 431	0.081 0
11	1 714	0.057 1
12	827	0.027 6

由上述数据可以看出，点数和为 7 的概率最大，且其值应在 0.166 附近。

事实上，该概率值为 $\frac{1}{6} = 1.666\ 6\cdots$ 。

将程序中掷骰子次数 k 的值取得愈大，则程序算出的概率值愈接近其准确值。

顺便指出，如果对程序中的变量 n (骰子个数) 赋值 3, 4, 5, …，则本程序便可以模拟一次掷 3 颗, 4 颗, 5 颗, …骰子时其点数和的实验。

掌握了例 1.6.1 的编程思想后，就可以用这种方法解决各种求概率值的问题。

例 1.6.2 用模拟实验的方法计算例 1.3.4 中的概率值。

解 该问题在例 1.3.4 中已有详细分析，下面仅给出计算所求概率值的 MATLAB 程序及实验结果。程序清单：

```

k = 40 000;      %实验次数

```

```

cs(1)=0;      %清零计数器
cs(2)=0;
cs(3)=0;
for m=1:k
cp(1)=1;      %次品记为 1,正品记为 0
cp(2)=1;
cp(3)=1;
for n=4:100
cp(n)=0;
end
q1=cp(unidrnd(100)); %第一次随机抽样
if q1>0,cp(1)=0;end %如果第一次抽得次品,则次品减少一个
q2=cp(unidrnd(99)); %第二次随机抽样
q=q1+q2+1;        %q 值:1 - 两件正品,2 - 一正一次,3 - 两
                    %件次品

cs(q)=cs(q)+1;
end
disp('两件正品  一正一次  两件次品') %显示实验结果
disp(sprintf('%7.5f  %7.5f  %7.5f',cs(1)/k,cs(2)/k,cs(3)/k))

```

执行该程序若干次,得到如下实验数据:

两件正品	一正一次	两件次品
0.940 45	0.058 98	0.000 57
两件正品	一正一次	两件次品
0.940 15	0.059 28	0.000 57
两件正品	一正一次	两件次品
0.941 63	0.057 75	0.000 63

注意,由于模拟的是随机事件,故每次执行程序的结果是不完全相等的.

习 题 1

1. 设 A, B, C 表示三个不同的随机事件,试将下列事件用 A, B, C 的运算式表示出来:

(1) A, B 都发生,但 C 不发生;

- (2) 三个事件都发生;
- (3) 三个事件都不发生;
- (4) A, B, C 中恰有两个发生;
- (5) 三个事件中至少有一个不发生;
- (6) 三个事件中不多于一个事件发生.

2. 指出下面式子中事件之间的关系:

- (1) $AB = A$; (2) $A \cup B = A$;
- (3) $ABC = A$; (4) $A \cup B \cup C = A$.

3. 掷两颗骰子时, 出现点数之和有多少种可能的结果, 其中哪一种结果出现的概率最大, 其值是多少?

4. 从一副扑克牌的 13 张黑桃中, 一张接一张地有放回地连抽 3 次, 设这 13 张牌中的每一张牌等可能的被抽到, 问抽到的 3 张牌中,

- (1) 没有同号的概率是多少?
- (2) 有同号的概率是多少?

5. 某城市共发行 A, B, C 三种报纸. 在这个城市居民中, 订阅 A 报的占 45%, 订阅 B 报的占 35%, 订阅 C 报的占 30%, 同时订阅 A 及 B 报的占 10%, 同时订阅 A 及 C 报的占 8%, 同时订阅 B 及 C 报的占 5%, 同时订阅 A, B 及 C 报的占 3%, 试求下列事件的概率:

- (1) 只订阅 A 报的;
- (2) 只订阅 A 及 B 报的;
- (3) 只订阅一种报的;
- (4) 恰好订阅两种报的;
- (5) 至少订阅一种报的;
- (6) 不订阅任何报纸的.

6. 一袋中装有 10 个号码球, 分别标有 1 ~ 10 号, 现从袋中任取 3 个球, 记录下其号码, 求:

- (1) 最小号码为 5 的概率;
- (2) 最大号码为 5 的概率;
- (3) 中间的号码恰为 5 的概率.

7. 一部五卷的文集, 按任意次序排放到书架上, 试求下列事件的概率:

- (1) 第一卷出现在两边;
- (2) 第一卷及第五卷出现在两边;
- (3) 第一卷或第五卷出现在两边;
- (4) 第一卷或第五卷不出现在两边.

8. 一根长为 L 的直棍在任意两点截断, 求所得三段能构成三角形的概率.

9. 在某一植物种群中, 有三种基因型: AA, Aa, aa , 每一基因型的数量分别为 200, 600, 50. 设随机地从这个种群中选一个体, 问:

- (1) 其基因型为 AA 的概率是多少?
- (2) 其基因型为 AA 或 aa 的概率是多少?

10. 某种动物由出生活到20岁的概率为0.8, 活到25岁的概率为0.4, 求现在20岁的这种动物能活到25岁的概率.

11. 100件产品中有10件次品, 用不放回的方式从中每次取1件, 连取3次, 求第三次才取得次品的概率.

12. 三人独立地去破译一个密码, 他们单独能译出的概率分别为 $\frac{1}{5}$, $\frac{1}{3}$, $\frac{1}{4}$, 问能将此密码译出的概率是多少?

13. 某药厂灌装一批合格的注射液需经四道工序, 从长期生产经验获知, 各道工序的废品率分别为0.5%, 0.2%, 0.1%, 0.8%. 设各道工序是否合格是相互独立的, 求经四道工序全部合格的概率是多少?

14. 为了提高抗菌素生产的产量和质量, 需要对生产菌种进行诱变处理, 然后从一大批经过处理的变异个体(菌株)中抽取一小部分来培养、测定, 从中找出优良的菌株. 如果某菌种的优良变异率 p 为0.03, 试问从一大批经诱变处理的菌株中, 采取多少只来培养、测定, 才能以95%的把握从中至少可以选到一只优良菌株?

15. 甲袋中装有7个红球, 3个白球, 15个黑球; 乙袋中装有6个红球, 10个白球, 9个黑球. 现从两袋中各取一个球, 求两个球颜色相同的概率.

16. 设一个仓库中共有10箱同样规格的产品, 已知这10箱产品中依次有5, 3, 2箱是甲、乙、丙三个厂生产的, 又知, 甲、乙、丙三厂生产的该产品的次品率依次为 $\frac{1}{10}$, $\frac{1}{15}$, $\frac{1}{20}$. 现从这10箱产品中任取一箱, 再从取得的这箱中任取一件产品, 求取得正品的概率; 假如取到这件产品是正品, 问所取的那箱产品是甲厂生产的概率是多少?

17. 盒中放有12个乒乓球, 其中有9个是新球, 第一次比赛从盒中任取3个来用, 比赛后仍放回盒中, 第二次比赛时又从盒中任取了3个. 求第二次取出的球都是新球的概率. 又若已知第二次取出的球都是新球, 求第一次取到的都是新球的概率.

18. 甲、乙两射手命中的概率分别为0.8和0.9, 如果两射手同时独立地射击一次, 求下列概率:

- (1) 两人都命中;
- (2) 恰有一人命中;
- (3) 至少有一人命中;
- (4) 两人都命中.

19. 某种集成电路使用到2 000 h还能正常工作的概率为0.92, 使用到3 000 h仍能正常工作的概率为0.85, 问已知工作了2 000 h的集成电路, 能继续工作到3 000 h的概率是多少?

20. 某种日光灯使用3 000 h以上的概率为0.8, 求3个日光灯在使用3 000 h以后, 以下事件的概率:

- (1) 都没有坏;
- (2) 坏了一个;
- (3) 最多只有一个坏了.

21. 某单位有 12 台个人计算机, 各计算机是否被使用是相互独立的, 设每台计算机的使用率为 0.7, 求在同一时刻有 9 台或更多计算机被使用的概率.

22. 人的血型为 O, A, B, AB 型的概率分别为 0.46, 0.40, 0.11, 0.03, 今任意挑选五人, 求下列事件的概率:

- (1) 恰有两个人为 O 型;
- (2) 三人为 O 型, 二人为 A 型;
- (3) 没有 AB 型.

23*. 袋中有 a 只黑球, b 只白球. 现从袋中随机地一只一只摸球(不放回). 求第 k 次 ($1 \leq k \leq a+b$) 摸出黑球的概率.

24*. 某人忘记了自己牡丹卡密码的最后一位数字, 因而他随机地按号. 求他按号不超过三次而选择正确的概率. 若已知最后一个数是偶数, 那么此概率是多少?

25*. 设有甲、乙两袋, 甲袋装有 N 个白球、 M 个黑球; 乙袋中装有 n 个白球、 m 个黑球, 今从甲袋中任取一个球放入乙袋中, 再从乙袋中任取一个球, 问取到白球的概率是多少?

26*. 若 $P(A) > 0$, $P(B|A) = P(B|\bar{A})$, 试证: 事件 A 与 B 相互独立.

27*. 甲、乙两人轮流射击, 先命中目标者为胜. 已知他们命中率分别为 p_1 和 p_2 , 甲先射, 求每个人获胜的概率.

28*. 甲、乙两名篮球运动员, 三分远投的命中率分别为 0.7 和 0.6. 现每人远投 3 次, 求两人进球数相等的概率.

第2章 随机变量及其分布

随机事件是对随机试验结果的一种定性描述,为深入研究有关随机试验的问题,需要将随机试验的结果数量化,并且引入现代数学工具.本章将讨论随机变量及其概率分布,它们是概率论研究的中心内容.

§ 2.1 随机变量与分布函数

1. 随机变量

许多随机试验的结果都可以直接用数量来表示,例如,一批产品中的次品数;射击过程中射手击中的环数;某段时间内电话总机接到的呼叫次数等.也有一些随机试验的结果不是用数量表示的,而表现为某种属性,然而可以数量化.例如,在掷硬币试验中,每次出现的结果不是正面就是反面,我们常常用数值“1”和“0”分别表示出现正面和反面.

一般地,对于随机事件 A ,可设

$$X = \begin{cases} 1, & \text{如果 } A \text{ 发生,} \\ 0, & \text{如果 } \bar{A} \text{ 发生.} \end{cases}$$

上面诸例中,试验结果都能用一个数 X 表示,这个数 X 随着试验结果的不同而变化,也就是说它是样本点的函数.由于基本事件(样本点)的出现是随机的,因而 X 的取值也随机,故称 X 为**随机变量**,下面给出其严格定义.

定义 2.1.1 设试验 E 的样本空间为 Ω ,如果对每一个样本点 $\omega \in \Omega$,都有唯一实数 $X(\omega)$ 与之对应,称 $X = X(\omega)$ 为样本空间 Ω 上的**随机变量**.

通常,随机变量用大写字母 X, Y 等表示.

对于随机变量,一般分两类进行讨论.如果随机变量的取值只有有限个或无限可列个数值,则称这种随机变量为**离散型随机变量**.上面列举的例子都属于这一类,其余的统称为**非离散型的随机变量**,在非离散型随机变量中,最重要的一类也是实际工作中遇到最多的是**连续型随机变量**.本书主要讨论离散型与连续型随机变量.

2. 随机变量的分布函数

为了能从数学上统一地对随机变量进行研究,需要构造统一的分布工具.在实际应用中,我们经常会遇到求某一随机变量 X 取值不大于实数 x 的概率 $P(X \leq x)$,显然 $P(X \leq x)$ 是定义在 $(-\infty, +\infty)$ 上的实值函数.

定义 2.1.2 设有随机变量 X , x 是任意一实数, 则函数

$$F(x) = P(X \leq x) \quad (2.1.1)$$

称为随机变量 X 的概率分布函数, 简称为分布函数.

(2.1.1) 式表明, 分布函数 $F(x)$ 的值表示随机变量 X 的取值落在区间 $(-\infty, x]$ 上的概率, 对于任意二实数 $x_1, x_2 (x_1 < x_2)$, 有

$$P(x_1 < X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1),$$

即随机变量 X 的取值落在 $(x_1, x_2]$ 内的概率为 $F(x_2) - F(x_1)$.

由此可见, 分布函数完整地描述了随机变量的统计规律性, 通过它, 我们将能用数学分析的方法来研究随机变量.

分布函数 $F(x)$ 具有以下基本性质:

- (1) $0 \leq F(x) \leq 1 (-\infty < x < +\infty)$;
- (2) $F(x)$ 是非减函数, 且 $F(-\infty) = 0, F(+\infty) = 1$;
- (3) $F(x)$ 在任何点 x 处右连续.

例 2.1.1 设随机变量 X 的分布函数为

$$F(x) = \begin{cases} 0, & x < 0, \\ \frac{x}{3}, & 0 \leq x < 1, \\ \frac{x}{2}, & 1 \leq x < 2, \\ 1, & x \geq 2. \end{cases}$$

求: (1) $P(X \leq \frac{1}{2})$; (2) $P(\frac{1}{2} < X \leq \frac{3}{2})$; (3) $P(X > \frac{3}{2})$.

解 (1) $P(X \leq \frac{1}{2}) = F(\frac{1}{2}) = \frac{1}{6}$;

(2) $P(\frac{1}{2} < X \leq \frac{3}{2}) = F(\frac{3}{2}) - F(\frac{1}{2}) = \frac{3}{4} - \frac{1}{6} = \frac{7}{12}$;

(3) $P(X > \frac{3}{2}) = 1 - F(\frac{3}{2}) = 1 - \frac{3}{4} = \frac{1}{4}$.

例 2.1.2 设随机变量 X 的分布函数为

$$F(x) = \begin{cases} a + be^{-\lambda x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

其中 $\lambda > 0$ 为常数, 求常数 a 与 b 的值.

解 由分布函数的性质 $F(+\infty) = 1$, 可得

$$F(+\infty) = \lim_{x \rightarrow +\infty} F(x) = \lim_{x \rightarrow +\infty} (a + be^{-\lambda x}) = a = 1,$$

即 $a = 1$. 由 $F(x)$ 的右连续性质, 可得

$$\lim_{x \rightarrow 0^+} F(x) = \lim_{x \rightarrow 0^+} (a + be^{-\lambda x}) = a + b = F(0) = 0,$$

则 $b = -1$.

§ 2.2 离散型随机变量及其分布

1. 离散型随机变量的分布律

离散型随机变量的可能取值是可以一一列举的, 因而要掌握它的变化规律, 不仅要知道它可能取哪些值, 而且要知道它取这些值的概率. 为此给出如下定义:

定义 2.2.1 设离散型随机变量 X 所有可能取的值为 $x_k (k=1, 2, \dots)$, 而 X 取值 x_k 的概率为 p_k , 即

$$P(X=x_k) = p_k, \quad k=1, 2, \dots, \quad (2.2.1)$$

称(2.2.1)式为离散型随机变量 X 的概率分布或分布律. 把 X 可能取的值及相应的概率列成表, 如表 2.2.1 所示, 称表 2.2.1 为 X 的概率分布表, 或称为分布列.

表 2.2.1

X	x_1	x_2	$\dots$	x_k	$\dots$
P	p_1	p_2	$\dots$	p_k	$\dots$

随机变量的分布律是指随机变量所有可能取的值与这些值的概率间的一种对应关系, 这种对应关系不仅可用解析式(2.2.1)和概率分布表(表 2.2.1)表示, 还可以用图示法表示(图 2.2.1).

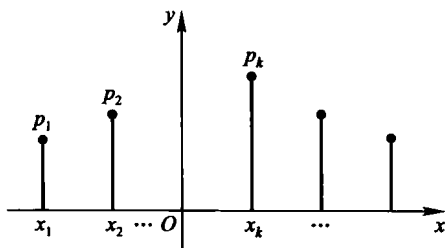


图 2.2.1

对于离散型随机变量, 概率分布中诸 p_k 必须满足下列两个性质:

$$(1) \quad p_k \geq 0, \quad k=1, 2, \dots; \quad (2.2.2)$$

$$(2) \quad \sum_k p_k = 1. \quad (2.2.3)$$

反过来, 满足(2.2.2)和(2.2.3)式的数 p_k 也一定可作为某离散型随机变量的概率分布.

例 2.2.1 设有 10 件产品，其中正品 6 件，次品 4 件，从中任取 3 件产品，用 X 表示从中取出的次品数，求其分布律。

解 X 表示取出 3 件产品中的次品数，则它可能取的值是 0, 1, 2, 3. “ $X=k$ ” ($k=0,1,2,3$) 表示事件“有 k 件次品”. 由概率的古典定义得

$$P(X=0) = \frac{C_6^3}{C_{10}^3} = \frac{1}{6},$$

$$P(X=1) = \frac{C_4^1 C_6^2}{C_{10}^3} = \frac{1}{2},$$

$$P(X=2) = \frac{C_4^2 C_6^1}{C_{10}^3} = \frac{3}{10},$$

$$P(X=3) = \frac{C_4^3}{C_{10}^3} = \frac{1}{30}.$$

其分布列为

X	0	1	2	3
P	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{30}$

一般来说，在总共 N 件产品中，其中有 M 件次品，现从中任取 n 件（不放回地抽取），则这 n 件中所含的次品数 X 是一个离散型随机变量，其概率分布为

$$P(X=k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}, \quad k=0,1,2,\dots,l, \quad (2.2.4)$$

其中 $l = \min(M, n)$. 通常称这个概率分布为超几何分布.

例 2.2.2 从次品率为 p 的一批药品中，有放回地一个一个地抽取，直到抽到次品为止. 设 X 为所需抽取的药品次数，求 X 的概率分布.

解 设 $A_i = \{\text{第 } i \text{ 次抽取次品}\}$, $i=1,2,\dots$, 则

$$P(A_i) = p, \quad P(\bar{A}_i) = 1 - p = q,$$

于是

$$\begin{aligned} P(X=k) &= P(\bar{A}_1 \bar{A}_2 \cdots \bar{A}_{k-1} A_k) \\ &= P(\bar{A}_1) P(\bar{A}_2) \cdots P(\bar{A}_{k-1}) P(A_k) \\ &= q^{k-1} p, \quad k=1,2,\dots \end{aligned} \quad (2.2.5)$$

上式是几何级数的一般项，因此称 (2.2.5) 为几何分布. 显然有

$$\sum_{k=1}^{\infty} P(X=k) = \sum_{k=1}^{\infty} q^{k-1} p = p \sum_{k=1}^{\infty} q^{k-1} = p \frac{1}{1-q} = 1.$$

对于离散型随机变量，只要给出它的分布律，那么它的分布函数以及随机

变量取值落入任一区间 G 内的概率可以随之得到:

$$F(x) = P(X \leq x) = \sum_{x_k \leq x} P(X = x_k) = \sum_{x_k \leq x} p_k. \quad (2.2.6)$$

$$P(X \in G) = \sum_{x_k \in G} P(X = x_k) = \sum_{x_k \in G} p_k. \quad (2.2.7)$$

例 2.2.3 由例 2.2.1 中次品数 X 的分布列, 求:

(1) X 的分布函数; (2) $P(0 < X \leq 2)$ 及 $P(0 \leq X < 2)$.

解 (1) 对于 $x < 0$, $(-\infty, x]$ 内不含 X 的任何可能值, 故 $F(x) = 0$;

对于 $0 \leq x < 1$, $F(x) = P(X \leq x) = P(X = 0) = \frac{1}{6}$;

对于 $1 \leq x < 2$, $F(x) = P(X \leq x) = P(X = 0) + P(X = 1) = \frac{1}{6} + \frac{1}{2} = \frac{2}{3}$;

对于 $2 \leq x < 3$, $F(x) = P(X \leq x) = P(X = 0) + P(X = 1) + P(X = 2)$

$$= \frac{1}{6} + \frac{1}{2} + \frac{3}{10} = \frac{29}{30}.$$

对于 $x \geq 3$, $F(x) = P(X \leq x) = \sum_{k=0}^3 P(X = k) = 1$. 故 X 的分布函数为

$$F(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{6}, & 0 \leq x < 1, \\ \frac{2}{3}, & 1 \leq x < 2, \\ \frac{29}{30}, & 2 \leq x < 3, \\ 1, & x \geq 3. \end{cases}$$

分布函数 $F(x)$ 的图形如图 2.2.2 所示, 由此可以看出, 离散型随机变量的分布函数 $F(x)$ 是一个分段阶梯函数, 在 $x = x_k (k = 1, 2, \dots)$ 处有跳跃, 其跳跃值为 $p_k = P(X = x_k)$.

(2) 利用分布列求概率, 可得

$$P(0 < X \leq 2) = P(X = 1) + P(X = 2)$$

$$= \frac{1}{2} + \frac{3}{10} = \frac{4}{5},$$

$$P(0 \leq X < 2) = P(X = 0) + P(X = 1)$$

$$= \frac{1}{6} + \frac{1}{2} = \frac{2}{3}.$$

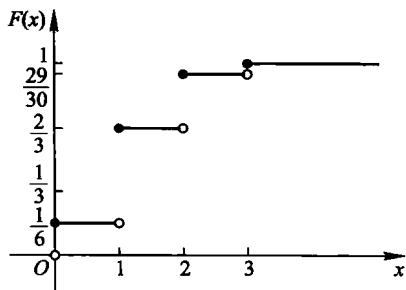


图 2.2.2

若利用分布函数求这些概率值, 则有

$$P(0 < X \leq 2) = F(2) - F(0) = \frac{29}{30} - \frac{1}{6} = \frac{4}{5},$$

$$\begin{aligned} P(0 \leq X < 2) &= P(X < 2) - P(X < 0) = F(2^-) - F(0^-) \\ &= \frac{2}{3} - 0 = \frac{2}{3}. \end{aligned}$$

由此可见, 分布律与分布函数都能完整地描述离散型随机变量的统计规律性, 相比较而言, 对于离散型随机变量, 分布律更为实用与简便.

下面介绍几种常见的离散型随机变量.

2. 两点分布

设在一次伯努利试验中, 只有两个可能结果 A 或 $\bar{A}$, 并且 $P(A) = p$, $P(\bar{A}) = q = 1 - p$. 若以 X 记事件 A 出现的次数, 则 X 的分布列为

X	0	1
P	q	p

或 $P(X=k) = p^k q^{1-k}$, $k=0, 1$, 称 X 服从两点分布或 $(0-1)$ 分布.

两点分布是最简单而又常见的概率分布. 例如掷硬币试验, 检验产品质量是否合格, 药物的毒性试验等都可以用两点分布来描述.

3. 二项分布

在 n 重伯努利试验中, 若事件 A 出现的次数记为 X , 则随机变量 X 可能取的值为 $0, 1, 2, \dots, n$, 相应概率分布为

$$P(X=k) = C_n^k p^k q^{n-k}, \quad k=0, 1, 2, \dots, n. \quad (2.2.8)$$

式中 $p = P(A)$, 且 $0 < p < 1$, $q = 1 - p$. 称 X 服从参数为 n, p 的二项分布, 记作 $X \sim B(n, p)$.

显然, 两点分布是二项分布的特殊情况. 注意到 $C_n^k p^k q^{n-k}$ 恰好是二项式 $(p+q)^n$ 的展开式中的一项, 故不难证明二项分布满足分布律的两个性质 (2.2.2) 式及 (2.2.3) 式. 对于确定的二项分布 $B(n, p)$, 有

$$\begin{aligned} \frac{P(X=k)}{P(X=k-1)} &= \frac{C_n^k p^k q^{n-k}}{C_n^{k-1} p^{k-1} q^{n-k+1}} = \frac{(n-k+1)p}{kq} \\ &= 1 + \frac{(n+1)p-k}{kq}. \end{aligned}$$

所以, 当 $k < (n+1)p$ 时, $P(X=k) > P(X=k-1)$, 即二项分布的概率随 k 值增加而上升; 当 $k > (n+1)p$ 时, $P(X=k) < P(X=k-1)$, 即其概率值随 k 增加而下降. 特别当 $(n+1)p$ 为正整数时, 在 $k = (n+1)p$ 及 $k = (n+1)p - 1$ 处, 两项概率均为最大值, 即 $P(X=k) = P(X=k-1)$. 一般地, 当 $(n+1)p$ 不是整数时, 二项分布在 $k = [(n+1)p]$ 处达到最大值.

上述 $[(n+1)p]$ 称为二项分布 $B(n, p)$ 的最可能出现次数, 又称为最可能值.

例 2.2.4 已知某种大批量产品的一级品率为 0.2, 现从中随机地抽查 20 件, 问 20 件产品中恰有 k 件($k=0, 1, 2, \dots, 20$)为一级品的概率是多少?

解 这里是不放回抽样, 但由于批量很大, 且抽查的件数相对于产品的总数来说又很小, 因而可以当作有放回抽样来处理, 这样做会有误差, 但误差不会大. 我们将抽取的产品是否是一级品看成是一次试验的结果, 检查 20 件产品相当于做 20 重伯努利试验. 设 X 为 20 件产品中一级品的件数, 则 X 服从参数 $n=20, p=0.2$ 的二项分布, 所求的概率为

$$P(X=k) = C_{20}^k (0.2)^k (0.8)^{20-k}, \quad k=0, 1, 2, \dots, 20.$$

对不同的 k 值分别进行计算, 结果见表 2.2.2 和图 2.2.3.

表 2.2.2

$X=k$	0	1	2	3	4	5	6	7	8	9	10	≥ 11
P	0.012	0.058	0.137	0.215	0.218	0.175	0.109	0.055	0.022	0.007	0.002	<0.001

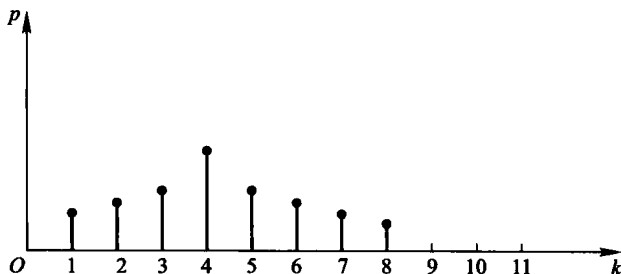


图 2.2.3

例 2.2.5 设每次射击命中目标的概率为 0.01, 现独立地射击 400 次, 求: (1) 最可能命中目标的次数及相应的概率; (2) 至少 3 次命中目标的概率.

解 设在 400 次射击中命中目标的次数为 X , 则由题设知 $X \sim B(n, p)$, 其中 $n=400, p=0.01$.

(1) 最可能命中目标的次数即是使二项分布 $P(X=k)$ 达到最大的 k 值. 由于 $(n+1)p = (400+1) \times 0.01 = 4.01$, 故最可能命中目标的次数为 $[4.01] = 4$, 并且相应的概率为

$$P(X=4) = C_{400}^4 (0.01)^4 (0.99)^{396}.$$

(2) 至少命中 3 次的概率为

$$\begin{aligned}
 P(X \geq 3) &= \sum_{k=3}^{400} P(X=k) = 1 - P(X=0) - P(X=1) - P(X=2) \\
 &= 1 - C_{400}^0 (0.01)^0 (0.99)^{400} - C_{400}^1 (0.01)^1 (0.99)^{399} - \\
 &\quad C_{400}^2 (0.01)^2 (0.99)^{398}.
 \end{aligned}$$

从上面的式子可以看出, 当 n 较大时, 计算服从二项分布的随机变量取值的概率是相当麻烦的. 下面给出一个近似计算公式, 这就是著名的二项分布的泊松逼近.

定理 2.2.1 (泊松定理) 设随机变量 $X_n (n=1, 2, \cdots)$ 服从参数 n, p_n 的二项分布, 若 $\lim_{n \rightarrow \infty} np_n = \lambda$, 则有

$$\lim_{n \rightarrow \infty} P(X_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

证明 记 $\lambda_n = np_n$, 则

$$\begin{aligned}
 P(X_n = k) &= C_n^k p_n^k (1 - p_n)^{n-k} \\
 &= \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda_n}{n}\right)^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \\
 &= \frac{\lambda_n^k}{k!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \left(1 - \frac{\lambda_n}{n}\right)^{n-k}.
 \end{aligned}$$

由于对固定的 k 有

$$\begin{aligned}
 \lim_{n \rightarrow \infty} \lambda_n^k &= \lambda^k, \\
 \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) &= 1,
 \end{aligned}$$

及
$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda_n}{n}\right)^{n-k} = \lim_{n \rightarrow \infty} \left[\left(1 - \frac{\lambda_n}{n}\right)^{-\frac{n}{\lambda_n}} \right]^{-\lambda_n \cdot \frac{n-k}{n}} = e^{-\lambda},$$

故有

$$\lim_{n \rightarrow \infty} P(X_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

根据泊松定理, 当 n 较大而 p 较小时, 有如下近似公式成立:

$$C_n^k p^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}, \quad \lambda = np. \quad (2.2.9)$$

在实际应用中当 $n > 10$, $p < 0.1$ 时就可以利用 (2.2.9) 式近似计算二项分布的概率.

例 2.2.6 利用近似公式 (2.2.9) 计算例 2.2.5 的概率.

解 因为 $\lambda = np = 400 \times 0.01 = 4$, 所以

$$(1) P(X=4) = P(X \geq 4) - P(X \geq 5) \approx \sum_{k=4}^{\infty} \frac{4^k}{k!} e^{-4} - \sum_{k=5}^{\infty} \frac{4^k}{k!} e^{-4}.$$

根据 $\lambda = 4$, 查附表 3 (泊松分布表) 得

$$P(X=4) = 0.566\ 530 - 0.371\ 163 = 0.195\ 367.$$

$$(2) P(X \geq 3) = \sum_{k=3}^{400} P(X=k) \approx \sum_{k=3}^{\infty} \frac{4^k}{k!} e^{-4} = 0.761\ 897.$$

4. 泊松分布

如果随机变量 X 的概率分布为

$$P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k=0, 1, 2, \dots \quad (2.2.10)$$

式中 $\lambda > 0$ 是常数, 则称 X 服从以 λ 为参数的泊松分布, 记作 $X \sim P(\lambda)$.

泊松分布是法国数学家泊松 (Poisson) 研究二项分布在一定条件下的极限分布而发现的.

自然界中有很多稀疏现象, 如生三胞胎、四胞胎; 某段时间内放射性物质放射出的 α 粒子数; 显微镜下在某观察范围内的微生物数; 数字通信中, 传输数字时发生误码的个数; 某商店一天到来的顾客数等都服从或近似服从泊松分布, 所以又称泊松分布为稀疏现象律.

容易验证泊松分布满足分布律性质 (2.2.2) 式及 (2.2.3) 式. 泊松分布图的上升、下降情况与二项分布相仿. 由 (2.2.10) 式, 得

$$\frac{P(X=k)}{P(X=k-1)} = \frac{\lambda}{k},$$

容易看出, 若 λ 不是整数, 则泊松分布的最可能值为 $[\lambda]$, 若 λ 是整数, 则最可能值为 λ 或 $\lambda - 1$.

例 2.2.7 某电话交换台每分钟接到的呼叫数服从参数为 5 的泊松分布, 求在一分钟内呼叫次数不超过 6 次的概率.

解 设每分钟交换台接到的呼叫次数为 X , 则 $X \sim P(5)$, 所以 $P(X=k) = \frac{5^k}{k!} e^{-5}$. 呼叫次数不超过 6 次, 即 $X \leq 6$, 其概率为

$$P(X \leq 6) = \sum_{k=0}^6 P(X=k) = 1 - \sum_{k=7}^{\infty} \frac{5^k}{k!} e^{-5},$$

查泊松分布表 (附表 3) 得

$$\sum_{k=7}^{\infty} \frac{5^k}{k!} e^{-5} = 0.237\ 8,$$

故

$$P(X \leq 6) = 1 - 0.237\ 8 = 0.762\ 2.$$

例 2.2.8 保险公司在开办某项保险业务之前, 要估计该项业务的收益情况. 据调查, 参加某项人身意外保险的人数共有 2 500 人, 而该人群出现意外事故的概率为 0.2%. 如果规定参加该项保险的人 (投保人) 每年交保险金 120 元, 在这一年中若投保人出现意外, 则其家属可从保险公司领取意外赔偿金

20 000 元. 试问: (1) 保险公司在办理该项业务上亏本的概率是多少? (2) 该项业务获利不少于 10 万元的概率有多大?

解 (1) 保险公司在该项业务中共收入 $120 \times 2\,500 = 300\,000$ (元), 设在这一年中有 x 名投保人出现意外事故, 则保险公司要支付赔偿金 $20\,000x$ (元), 只要 $20\,000x > 300\,000$, 即 $x > 15$ (人), 保险公司在该项业务上就亏本. 设出现事故人数为随机变量 X , 则 $X \sim B(2\,500, 0.002)$, 利用近似公式 (2.2.9), $np = \lambda = 5$,

$$\begin{aligned} P(X > 15) &= P(X \geq 16) = \sum_{k=16}^{2\,500} C_{2\,500}^k (0.002)^k (0.998)^{2\,500-k} \\ &\approx \sum_{k=16}^{2\,500} \frac{5^k}{k!} e^{-5} = 6.9 \times 10^{-5}. \end{aligned}$$

(2) 该项业务获利不少于 10 万元, 即 $300\,000 - 20\,000x \geq 100\,000$, $x \leq 10$ (人), 故

$$\begin{aligned} P(X \leq 10) &= 1 - P(X > 10) = 1 - P(X \geq 11) \\ &= 1 - 0.0137 = 0.9863. \end{aligned}$$

由上面的计算可知, 保险公司在办理该项业务亏本的风险很小, 赢利十万元以上的可能性近 99%.

§ 2.3 连续型随机变量及其分布

1. 连续型随机变量及其概率密度

有一些随机变量是在一个区间内连续取值的, 不能像离散型随机变量一一列举, 例如, 将来某一时刻的温度. 对这类随机变量的描述, 也应指出它的取值范围及其对应的取值概率, 其分布函数可用积分的形式来表示.

定义 2.3.1 设随机变量 X 的概率分布函数为 $F(x)$, 如果存在一个函数 $f(x)$, 对于任意的实数 x , 都有

$$F(x) = \int_{-\infty}^x f(t) dt, \quad -\infty < x < +\infty, \quad (2.3.1)$$

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

由 (2.3.1) 式可知, 若 $f(x)$ 连续, 那么连续型随机变量的分布函数的导数就是密度函数, 即 $F'(x) = f(x)$. 同时, 密度函数 $f(x)$ 反映了概率在 x 点附近的“密集程度”, 所以用密度函数描述连续型随机变量的概率分布与离散型用分布列描述在某种意义上有相似之处. 连续型随机变量的密度函数具有如下性质:

$$(1) f(x) \geq 0; \quad (2.3.2)$$

$$(2) \int_{-\infty}^{+\infty} f(x) dx = 1. \quad (2.3.3)$$

(2.3.2)式是因为分布函数 $F(x)$ 为非减函数, 故 $f(x) = F'(x) \geq 0$ (至多有限个点除外). (2.3.3)式是因 $\int_{-\infty}^{+\infty} f(x) dx = F(+\infty) = 1$ 而得. 同时, 满足上述两式的函数 $f(x)$ 也一定可作为某一连续随机变量的概率密度函数.

由分布函数 $F(x)$ 和密度函数 $f(x)$ 的定义可知两者具有下列关系:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt, \quad (2.3.4)$$

进而

$$\begin{aligned} P(a < X \leq b) &= F(b) - F(a) \\ &= \int_a^b f(x) dx. \end{aligned} \quad (2.3.5)$$

由 (2.3.5) 式可以看出, 连续型随机变量 X 落在某一区间内的概率, 在数值上等于密度函数曲线在该区间上的曲边梯形的面积, 如图 (2.3.1) 所示. 并且, 连续型随机变量取一特定值 a 的概率为零, 即

$P(X = a) = 0$. 这是因为, 对任一常数 a , 有

$$0 \leq P(X = a) \leq P(a < X \leq a + \Delta x) = \int_a^{a+\Delta x} f(t) dt.$$

由微积分学中的知识, 有

$$0 \leq P(X = a) \leq \lim_{\Delta x \rightarrow 0} \int_a^{a+\Delta x} f(t) dt = 0,$$

从而 $P(X = a) = 0$.

这一结论告诉我们: 在计算连续型随机变量落在某一区间内的概率时, 不必区分是开区间还是闭区间, 即

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b).$$

还应指出, 一个事件的概率为零, 这个事件并不一定是不可能事件. 同样, 一个事件的概率等于 1, 这事件也不一定就是必然事件.

例 2.3.1 设随机变量 X 的概率密度为

$$f(x) = \begin{cases} a \cos x, & |x| \leq \frac{\pi}{2}, \\ 0, & \text{其它.} \end{cases}$$

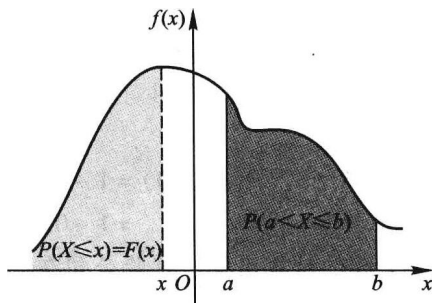


图 2.3.1

求: (1) 常数 a ; (2) X 的分布函数 $F(x)$; (3) $P\left(0 < X < \frac{\pi}{4}\right)$.

解 (1) 由概率密度的性质 $\int_{-\infty}^{+\infty} f(x) dx = 1$, 得

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} a \cos x dx = a \sin x \Big|_{-\frac{\pi}{2}}^{\frac{\pi}{2}} = 2a = 1,$$

故 $a = \frac{1}{2}$, 于是 X 的概率密度为

$$f(x) = \begin{cases} \frac{1}{2} \cos x, & |x| \leq \frac{\pi}{2}, \\ 0, & \text{其它.} \end{cases}$$

(2) X 的分布函数为

$$F(x) = \int_{-\infty}^x f(t) dt,$$

当 $x < -\frac{\pi}{2}$ 时, $F(x) = 0$;

当 $-\frac{\pi}{2} \leq x < \frac{\pi}{2}$ 时, $F(x) = \int_{-\frac{\pi}{2}}^x \frac{1}{2} \cos t dt = \frac{1}{2}(\sin x + 1)$;

当 $x \geq \frac{\pi}{2}$ 时, $F(x) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{2} \cos t dt = 1$,

$$\text{即 } F(x) = \begin{cases} 0, & x < -\frac{\pi}{2}, \\ \frac{1}{2}(\sin x + 1), & -\frac{\pi}{2} \leq x < \frac{\pi}{2}, \\ 1, & x \geq \frac{\pi}{2}. \end{cases}$$

$$(3) \quad P\left(0 < X < \frac{\pi}{4}\right) = F\left(\frac{\pi}{4}\right) - F(0) = \frac{\sqrt{2}}{4}.$$

$$\text{或者 } P\left(0 < X < \frac{\pi}{4}\right) = \int_0^{\frac{\pi}{4}} f(x) dx = \int_0^{\frac{\pi}{4}} \frac{1}{2} \cos x dx = \frac{\sqrt{2}}{4}.$$

下面介绍几种常见的连续型随机变量.

2. 均匀分布

若连续型随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{其它,} \end{cases} \quad (2.3.6)$$

则称 X 在区间 $[a, b]$ 上服从均匀分布, 记作 $X \sim U[a, b]$.

由均匀分布的密度函数 $f(x)$ 可得均匀分布的分布函数为:

$$F(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \leq x \leq b, \\ 1, & x > b. \end{cases} \quad (2.3.7)$$

密度函数 $f(x)$ 与分布函数 $F(x)$ 的图形分别如图 2.3.2 和图 2.3.3 所示.

若 X 在 $[a, b]$ 上服从均匀分布, 对区间内的任一个区间 $[c, d]$, 即 $a \leq c \leq d \leq b$, 有

$$P(c \leq X \leq d) = \int_c^d \frac{1}{b-a} dx = \frac{1}{b-a} (d-c).$$

上式表示, 服从均匀分布的随机变量 X 取值落在 $[a, b]$ 内任一区间 $[c, d]$ 内的概率与该区间的长度成正比, 而与区间的位置无关, 即取值在 $[a, b]$ 上是均匀的.

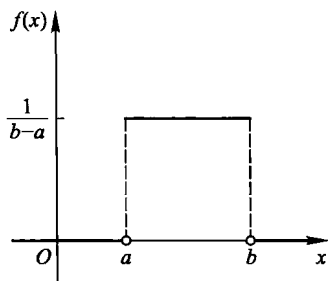


图 2.3.2

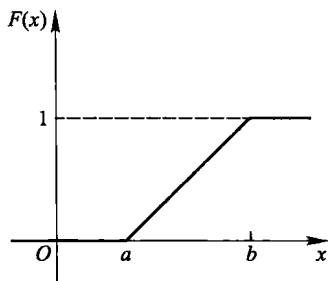


图 2.3.3

例 2.3.2 设随机变量 X 在 $(-1, 1)$ 上服从均匀分布, 求方程 $t^2 - 3Xt + 1 = 0$ 有实根的概率.

解 因为 X 在 $(-1, 1)$ 上服从均匀分布, 故

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

当 $9X^2 - 4 \geq 0$, 即当 $|X| \geq \frac{2}{3}$ 时, 方程有实根. 故方程有实根的概率为

$$P\left(|X| \geq \frac{2}{3}\right) = 1 - \int_{-\frac{2}{3}}^{\frac{2}{3}} \frac{1}{2} dx = \frac{1}{3}.$$

3. 指数分布

若连续型随机变量 X 的密度函数为

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0, \end{cases} \quad (2.3.8)$$

其中 $\lambda > 0$ 为常数, 则称 X 服从参数为 λ 的指数分布, 记作 $X \sim E(\lambda)$. 其分布函数为

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

指数分布在研究“寿命”一类问题中有重要的应用, 如灯泡的寿命、动物的寿命、电话的通话时间等, 都近似服从指数分布. 其中 λ 表示平均寿命的倒数. 指数分布的密度函数与分布函数的图形分别如图 2.3.4、图 2.3.5 所示.

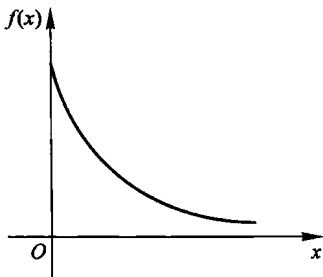


图 2.3.4

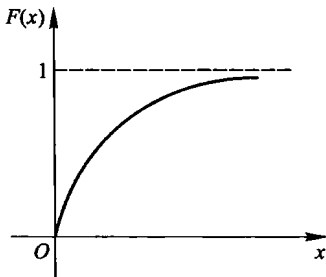


图 2.3.5

指数分布的重要性还表现在它具有“无记忆性”. 设随机变量 X 服从指数分布, 则对于任意 $s > 0, t > 0$, 有

$$\begin{aligned} P(X > s+t | X > s) &= \frac{P(\{X > s+t\} \cap \{X > s\})}{P(X > s)} \\ &= \frac{P(X > s+t)}{P(X > s)} = \frac{1 - P(X \leq s+t)}{1 - P(X \leq s)} \\ &= \frac{1 - (1 - e^{-\lambda(s+t)})}{1 - (1 - e^{-\lambda s})} \\ &= e^{-\lambda t} = 1 - (1 - e^{-\lambda t}) \\ &= 1 - P(X \leq t) = P(X > t). \end{aligned}$$

即

$$P(X > s+t | X > s) = P(X > t). \quad (2.3.9)$$

假如 X 表示为寿命的大小, 则 (2.3.9) 式表示: 如果已知某人的年龄为 s 岁, 则他再存活 t 年的概率与年龄 s 无关. 所以有时又风趣地称指数分布是“永远年轻”的. 可以证明: 指数分布是唯一具有这种性质的连续型分布.

例 2.3.3 设修理某机器所用的时间 X 服从参数为 $\lambda = 0.5$ (h) 的指数分布, 求当机器出现故障时, 在一小时内可以修好的概率.

解 因为 $X \sim E(0.5)$, 故 X 的概率密度为

$$f(x) = \begin{cases} \frac{1}{2}e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

所求概率为

$$P(X \leq 1) = \int_0^1 \frac{1}{2}e^{-\frac{x}{2}} dx = 1 - e^{-\frac{1}{2}}.$$

另外, 也可利用 X 的分布函数 $F(x) = \begin{cases} 1 - e^{-\frac{x}{2}}, & x > 0, \\ 0, & x \leq 0. \end{cases}$ 求出概率

$$P(X \leq 1) = F(1) = 1 - e^{-\frac{1}{2}}.$$

4. 正态分布

正态分布最早是由棣莫弗(A. De Moivre)在1733年研究二项分布的极限分布形式时提出的, 但在当时并没有引起学术界的重视. 之后高斯(C. F. Gauss)在1809年, 拉普拉斯(M. De Laplace)在1812年分别重新提出. 所以人们亦称正态分布为高斯分布或高斯-拉普拉斯分布.

正态分布是自然界中最常见的一种分布, 例如: 人的身高、体重; 各种测量的误差; 某地区居民的年收入等, 都服从或近似服从正态分布, 并且统计分析中的许多分布都以正态分布的理论为基础. 因此, 无论从理论还是实际应用上来说, 正态分布都是十分重要的.

若随机变量 X 的概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty. \quad (2.3.10)$$

其中 μ, σ 为常数, 且 $\sigma > 0$, 则称随机变量 X 服从参数 μ, σ 的正态分布, 记为 $X \sim N(\mu, \sigma^2)$. 服从正态分布的随机变量又称为正态变量.

正态分布密度函数 $f(x)$ 的图形呈钟状, 最大值点在 $x = \mu$ 处; 曲线对称于直线 $x = \mu$; 在 $x = \mu \pm \sigma$ 处有拐点; 当 $x \rightarrow +\infty$ 时, $f(x) \rightarrow 0$, 即曲线以 x 轴为渐近线, 如图 2.3.6 所示. 当 σ 固定而 μ 值改变时, 曲线沿着 x 轴平移, 形状

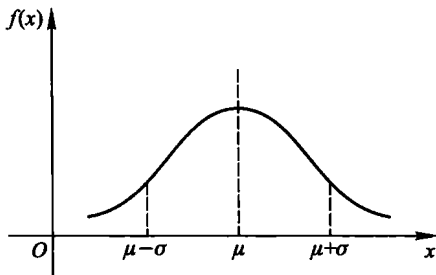


图 2.3.6

不变, 如图 2.3.7. 因此 μ 称为位置参数; 当 μ 固定, 而使 σ 变小时, 曲线变得陡峭, 即分布越集中于 μ 的附近, 反之, σ 增大时, 曲线变得平缓, 即分布越分散, 如图 2.3.8, 因此称 σ 为离散参数.

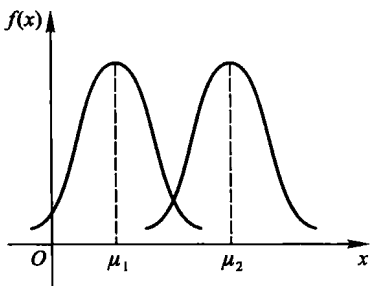


图 2.3.7

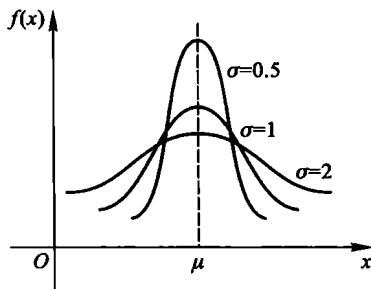


图 2.3.8

正态分布的分布函数为

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt. \quad (2.3.11)$$

当参数 $\mu=0$, $\sigma=1$ 时, 正态分布称为标准正态分布, 记为 $X \sim N(0,1)$. 相应的概率密度函数和分布函数分别特记为 $\varphi(x)$ 及 $\Phi(x)$, 即

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < +\infty, \quad (2.3.12)$$

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt, \quad -\infty < x < +\infty. \quad (2.3.13)$$

如何计算服从正态分布的随机变量落在区间 $(a, b]$ 内的概率 $P(a < X \leq b)$ 呢? 若 $X \sim N(0,1)$,

$$P(a < X \leq b) = \int_a^b \varphi(x) dx = \Phi(b) - \Phi(a).$$

具体计算时, 由于积分 $\Phi(x)$ 无法写成初等函数的表示式, 为此编制了“标准正态分布表”(附表4). $\Phi(x)$ 的数值在图形上为图 2.3.9 中阴影部分, 它的数值可以从附表4中查到. 当遇到负值时, 可利用公式:

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

一般地, 若 $X \sim N(\mu, \sigma^2)$, 则 $\frac{X-\mu}{\sigma} \sim N(0,1)$. 此时 X 的分布函数 $F(x) =$

$\Phi\left(\frac{x-\mu}{\sigma}\right)$, 因此有

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

这样通过查表可求得概率值.

为了便于今后在数理统计中的应用, 对于标准正态随机变量, 我们引入上

α 分位点的定义.

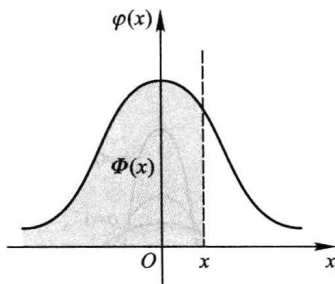


图 2.3.9

图 2.3.10

设 $X \sim N(0,1)$, 若 u_α 满足条件

$$P(X > u_\alpha) = \alpha, \quad 0 < \alpha < 1,$$

则称点 u_α 为标准正态分布的上 α 分位点(数)(如图 2.3.10).

上 α 分位点可通过反查附表 4 得到, 例如 $\alpha = 0.025$ 时, 由附表 4 得, $u_{0.025} = 1.96$.

例 2.3.4 设随机变量 $X \sim N(0,1)$, 求:

(1) $P(X < 1.5)$; (2) $P(X < -1.24)$; (3) $P(|X| < 3)$.

解 (1) $P(X < 1.5) = \Phi(1.5) = 0.9332$.

(2) $P(X < -1.24) = \Phi(-1.24) = 1 - \Phi(1.24) = 1 - 0.8925$
 $= 0.1075$.

(3) $P(|X| < 3) = P(-3 < X < 3) = \Phi(3) - \Phi(-3)$
 $= 2\Phi(3) - 1 = 2 \times 0.9987 - 1 = 0.9974$.

例 2.3.5 设随机变量 $X \sim N(1.5, 4)$, 求:

(1) $P(X < 3.5)$; (2) $P(1.5 < X < 3.5)$; (3) $P(|X| \geq 3)$.

解 $\mu = 1.5$, $\sigma = 2$, 记 $F(x)$ 为 X 的分布函数.

(1) $P(X < 3.5) = F(3.5) = \Phi\left(\frac{3.5 - 1.5}{2}\right) = \Phi(1) = 0.8413$.

(2) $P(1.5 < X < 3.5) = F(3.5) - F(1.5)$
 $= \Phi\left(\frac{3.5 - 1.5}{2}\right) - \Phi\left(\frac{1.5 - 1.5}{2}\right) = \Phi(1) - \Phi(0)$
 $= 0.8413 - 0.5 = 0.3413$.

(3) $P(|X| \geq 3) = 1 - P(|X| < 3) = 1 - P(-3 < X < 3)$
 $= 1 - F(3) + F(-3) = 1 - \Phi\left(\frac{3 - 1.5}{2}\right) + \Phi\left(\frac{-3 - 1.5}{2}\right)$
 $= 1 - \Phi(0.75) + \Phi(-2.25)$
 $= 1 - 0.7734 + (1 - 0.9878) = 0.2388$.

例 2.3.6 设随机变量 $X \sim N(\mu, \sigma^2)$, 求 X 落入区间 $(\mu - \sigma, \mu + \sigma)$ 、 $(\mu - 2\sigma, \mu + 2\sigma)$ 、 $(\mu - 3\sigma, \mu + 3\sigma)$ 内的概率.

解

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &= \Phi\left(\frac{\mu + \sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - \sigma - \mu}{\sigma}\right) \\ &= \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 \\ &= 2 \times 0.8413 - 1 = 0.6826. \end{aligned}$$

类似可得

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 2\Phi(2) - 1 = 2 \times 0.9773 - 1 = 0.9546,$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 2\Phi(3) - 1 = 2 \times 0.9987 - 1 = 0.9974.$$

从而表明, 服从正态分布 $N(\mu, \sigma^2)$ 的随机变量 X 的取值基本上落在区间 $(\mu - 2\sigma, \mu + 2\sigma)$ 内, 几乎全部落在区间 $(\mu - 3\sigma, \mu + 3\sigma)$ 内, 此特征又称为正态分布的“ 3σ ”原则. 如图 2.3.11 所示.

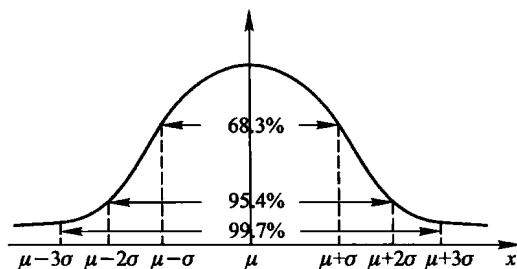


图 2.3.11

§ 2.4 随机变量函数的分布

在实际问题中, 我们经常会遇到随机变量的函数. 这一概念类似于实函数中的 $y = g(x)$. 例如, 在机械加工过程中, 需要测量某圆轴的横截面面积, 而该随机指标往往不能直接得出. 一般我们能测得圆轴截面的直径 X 这个随机变量, 再根据面积公式 $Y = \frac{1}{4}\pi X^2$ 得到我们所关心的随机变量. 下面讨论如何由已知的随机变量 X 的分布求得它的函数 $Y = g(X)$ 的分布.

1. X 是离散型随机变量

设随机变量 X 的分布列为

X	x_1	x_2	$\cdots$	x_k	$\cdots$
P	p_1	p_2	$\cdots$	p_k	$\cdots$

又随机变量 Y 是 X 的函数, $Y = g(X)$. 于是 $Y = g(X)$ 可能取值为 $g(x_1)$, $g(x_2)$, $\dots$, $g(x_k)$, $\dots$. 若 $g(x_k)$ ($k=1, 2, \dots$) 互不相同, 则事件 $\{Y = y_i\}$ 等价于事件 $\{X = x_i\}$, 从而 $Y = g(X)$ 的概率分布为

Y	$g(x_1)$	$g(x_2)$	$\dots$	$g(x_k)$	$\dots$
$P(Y = y_i)$	p_1	p_2	$\dots$	p_k	$\dots$

若某些 $g(x_i)$ 相等, 比如 $g(x_{i1}) = g(x_{i2}) = \dots = g(x_{il}) = y_i$, $i = 1, 2, \dots$, 则事件 $\{Y = y_i\}$ 等价于事件 $\{X = x_{i1}\} \cup \{X = x_{i2}\} \cup \dots \cup \{X = x_{il}\}$, 故有

$$\begin{aligned} P(Y = y_i) &= P(X = x_{i1}) + P(X = x_{i2}) + \dots + P(X = x_{il}) \\ &= \sum_{k=1}^l P(X = x_{ik}). \end{aligned}$$

例 2.4.1 设 X 的分布列为

X	0	1	2	3	4
P	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{3}$

求(1) $Y = 2X - 1$ 的分布列; (2) $Y = (X - 1)^2$ 的分布列.

解 (1) 因为 $y = 2x - 1$ 严格单调, 所以 y_i ($i = 1, 2, \dots, 5$) 互不相同, 故 Y 的分布列为

Y	-1	1	3	5	7
$P(Y = y_i)$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{3}$

(2) 由于 $Y = (X - 1)^2$ 的取值分别为 1, 0, 1, 4, 9, 故 Y 的分布列为

Y	0	1	4	9
$P(Y = y_i)$	$\frac{1}{6}$	$\frac{1}{12} + \frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{3}$

2. X 是连续型随机变量

设连续型随机变量 X 的分布函数为 $F_X(x)$, 即

$$F_X(x) = P(X \leq x).$$

如何求随机变量函数 $Y = g(X)$ 的分布呢? 依据分布函数的定义, Y 的分布函数为

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y),$$

由不等式 $g(X) \leq y$ 中解得关于 X 的不等式, 则 $F_Y(y)$ 就可利用已知的分布函

数 $F_X(x)$ 来表示. 下面, 我们从一个具体例子来讨论这个问题.

例 2.4.2 设随机变量 $X \sim N(\mu, \sigma^2)$, 求 $Y = aX + b$ ($a \neq 0, b$ 是常数) 的密度函数.

解 用 $F_X(x)$, $F_Y(y)$ 分别表示 X 与 Y 的分布函数, $f_X(x)$, $f_Y(y)$ 分别表示 X 与 Y 的密度函数, 则

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y).$$

1° 当 $a > 0$ 时

$$F_Y(y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right).$$

将上式两端对 y 求导得

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{a}.$$

2° 当 $a < 0$ 时

$$F_Y(y) = P\left(X \geq \frac{y-b}{a}\right) = 1 - P\left(X \leq \frac{y-b}{a}\right) = 1 - F_X\left(\frac{y-b}{a}\right).$$

上式两端对 y 求导得

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right) \cdot \left(-\frac{1}{a}\right).$$

综合上述 1°, 2° 得 Y 的密度函数为

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right), \quad -\infty < y < +\infty,$$

或者

$$f_Y(y) = \frac{1}{\sqrt{2\pi}|a|\sigma} e^{-\frac{[y-(a\mu+b)]^2}{2(a\sigma)^2}}, \quad -\infty < y < +\infty.$$

由此得到服从正态分布的随机变量一重要性质, 即

若 $X \sim N(\mu, \sigma^2)$, 则 $aX + b \sim N(a\mu + b, a^2\sigma^2)$. 特别, 当 $a = \frac{1}{\sigma}$, $b = -\frac{\mu}{\sigma}$ 时得到 $\frac{X-\mu}{\sigma} \sim N(0, 1)$. 这就是上一节非标准正态分布化成标准正态分布的结论.

一般地, 有如下定理:

定理 2.4.1 设连续型随机变量 X 具有概率密度函数 $f_X(x)$, 又设函数 $y = g(x)$ 严格单调, 其反函数 $g^{-1}(y)$ 有连续导数, 则 $Y = g(X)$ 是连续型随机变量, 其概率密度函数为

$$f_Y(y) = \begin{cases} f_X[g^{-1}(y)] |[g^{-1}(y)]'|, & \alpha < y < \beta, \\ 0, & \text{其它,} \end{cases} \quad (2.4.1)$$

其中 $\alpha = \min[g(-\infty), g(+\infty)]$, $\beta = \max[g(-\infty), g(+\infty)]$.

证明 首先考虑当 $g(x)$ 严格单调增加, 这时其反函数 $g^{-1}(y)$ 也严格单调增加, 则事件 $\{Y \leq y\}$ 与事件 $\{X \leq g^{-1}(y)\}$ 等价, 设 Y 的概率分布函数为 $F_Y(y)$, 从而

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

两端关于 y 求导, 注意到 $[g^{-1}(y)]' > 0$, 于是 Y 的概率密度函数为

$$f_Y(y) = F'_Y(y) = F'_X[g^{-1}(y)] [g^{-1}(y)]' = f_X[g^{-1}(y)] [g^{-1}(y)]',$$

$$\text{或者 } f_Y(y) = \begin{cases} f_X[g^{-1}(y)] [g^{-1}(y)]', & g(-\infty) < y < g(+\infty), \\ 0, & \text{其它.} \end{cases}$$

同理, 当 $g(x)$ 严格单调减小, 其反函数 $g^{-1}(y)$ 也严格单调减小. 事件 $\{Y \leq y\}$ 与事件 $\{X \geq g^{-1}(y)\}$ 等价, 故

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) \\ &= 1 - P(X < g^{-1}(y)) = 1 - P(X \leq g^{-1}(y)) = 1 - F_X(g^{-1}(y)). \end{aligned}$$

此时, $[g^{-1}(y)]' < 0$, 于是 Y 的概率密度函数为

$$f_Y(y) = F'_Y(y) = -F'_X[g^{-1}(y)] [g^{-1}(y)]' = -f_X[g^{-1}(y)] [g^{-1}(y)]',$$

$$\text{或者 } f_Y(y) = \begin{cases} -f_X[g^{-1}(y)] [g^{-1}(y)]', & g(+\infty) < y < g(-\infty), \\ 0, & \text{其它.} \end{cases}$$

综合上述两种情况, 定理成立.

例 2.4.3 若随机变量 X 服从 $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ 上的均匀分布, $Y = \tan X$, 试求 Y 的概率密度函数.

解 令 $y = \tan x$, 则 $x = \arctan y$, $\frac{dx}{dy} = \frac{1}{1+y^2} > 0$, $-\infty < y < +\infty$. 因为 X 的概率密度函数为

$$f_X(x) = \begin{cases} \frac{1}{\pi}, & -\frac{\pi}{2} \leq x \leq \frac{\pi}{2}, \\ 0, & \text{其它,} \end{cases}$$

$\alpha = \lim_{x \rightarrow -\frac{\pi}{2}} \tan x = -\infty$, $\beta = \lim_{x \rightarrow \frac{\pi}{2}} \tan x = +\infty$, 得 $Y = \tan X$ 的概率密度函数为

$$\begin{aligned} f_Y(y) &= f_X(\arctan y) \cdot |(\arctan y)'| \\ &= \frac{1}{\pi} \frac{1}{1+y^2}, \quad -\infty < y < +\infty. \end{aligned} \quad (2.4.2)$$

由 (2.4.2) 式定义的分布称为柯西 (Cauchy) 分布, 它是概率论中重要的分布之一.

例 2.4.4 设随机变量 X 的概率密度函数为 $f_X(x)$, 求 $Y = X^2$ 的概率密度

函数.

解 因为函数 $y = x^2$ 在 $(-\infty, +\infty)$ 上不是单调函数, 当 $y \leq 0$ 时, 事件 $\{X^2 \leq y\}$ 是不可能事件, 所以

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = 0.$$

当 $y > 0$ 时, 因为 $\{X^2 \leq y\} = \{-\sqrt{y} \leq X \leq \sqrt{y}\}$, 故

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P\{-\sqrt{y} \leq X \leq \sqrt{y}\} \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}), \end{aligned}$$

所以, 随机变量 Y 的概率密度函数为

$$f_Y(y) = F'_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}}[f_X(\sqrt{y}) + f_X(-\sqrt{y})], & y > 0, \\ 0, & y \leq 0. \end{cases} \quad (2.4.3)$$

特别地, 若 $X \sim N(0, 1)$, $f_X(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, $-\infty < x < +\infty$,

可得 $Y = X^2$ 的概率密度函数

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}}y^{-\frac{1}{2}}e^{-\frac{y}{2}}, & y > 0, \\ 0, & y \leq 0. \end{cases} \quad (2.4.4)$$

§ 2.5 用 MATLAB 计算分布函数

为了便于研究概率密度函数(pdf)和概率分布函数(cdf), MATLAB 在其统计工具箱中提供了一组 pdf 函数和一组 cdf 函数, 专门用来处理随机变量服从不同概率分布的情况. 表 2.5.1 列出了其中的部分概率密度函数和概率分布函数. 最后三个字母是 pdf 者为概率密度函数, 是 cdf 者为概率分布函数. 当需要计算某种分布的概率密度函数或概率分布函数时, 便可以直接调用表 2.5.1 中的函数来进行处理.

例 2.5.1 若 $X_1, X_2, \dots, X_n$ 为相互独立且服从标准正态分布 $N(0, 1)$ 的随机变量, 则称它们的平方和 $Y = \sum_{i=1}^n X_i^2$ 为服从自由度为 n 的 χ^2 分布(详见第 5 章), 记作 $Y \sim \chi^2(n)$. 试绘制自由度为 5, 10, 20 的 χ^2 分布的密度函数和分布函数的图形.

解 χ^2 分布在 $x < 0$ 时其密度函数等于零, 故取 $x \in [0, 30]$, 则绘制密度函数的命令为:

```
X=0:0.2:30;↓
```

```
P1=chi2pdf(X,5);↓
```

```
P2 = chi2pdf(X,10); ↓
P3 = chi2pdf(X,20); ↓
plot(X,P1,X,P2,X,P3) ↓
```

绘出图形如图 2.5.1.

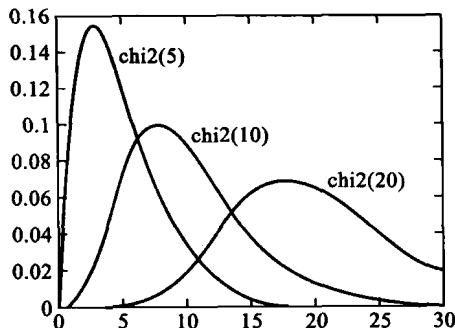


图 2.5.1 χ^2 分布的密度函数

完全类似，绘制分布函数的命令为：

```
X=0:0.2:30; ↓
C1 = chi2cdf(X,5); ↓
C2 = chi2cdf(X,10); ↓
C3 = chi2cdf(X,20); ↓
plot(X,C1,X,C2,X,C3) ↓
```

得到图 2.5.2.

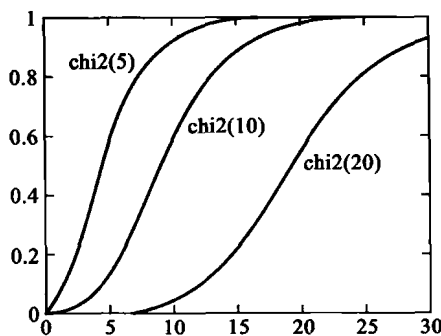


图 2.5.2 χ^2 分布的分布函数

顺便指出，MATLAB 对表 2.5.1 中每一个概率分布函数 (cdf) 都给出了其逆函数，它们可用来计算给定显著概率条件下假设检验的临界值 (详见第 6 章)。这些逆函数的命名规则是：只需将相应概率分布函数名称中最后三个字 cdf 换成 inv 即可。例如，对应 χ^2 分布函数

表 2.5.1 概率密度函数和分布函数

分布类型	函数名称	调用格式	备 注
二项分布	binopdf binocdf	$Y = \text{binopdf}(X, N, P)$ $Y = \text{binocdf}(X, N, P)$	N 为正整数 $P \in (0, 1)$
χ^2 分布	chi2pdf chi2cdf	$Y = \text{chi2pdf}(X, N)$ $Y = \text{chi2cdf}(X, N)$	自由度 N 为正整数
指数分布	exp pdf expcdf	$Y = \text{exp pdf}(X, MU)$ $Y = \text{expcdf}(X, MU)$	$\mu = \frac{1}{\lambda}$ 为正整数 MU 代表 μ
F 分布	f pdf f cdf	$Y = \text{f pdf}(X, M, N)$ $Y = \text{f cdf}(X, M, N)$	M 和 N 都是正整数
正态分布	norm pdf normcdf	$Y = \text{norm pdf}(X, MU, SIGMA)$ $Y = \text{normcdf}(X, MU, SIGMA)$	参数 $SIGMA$ 为正数 $SIGMA$ 代表 σ
泊松分布	poiss pdf poisscdf	$Y = \text{poiss pdf}(X, LAMBDA)$ $Y = \text{poisscdf}(X, LAMBDA)$	参数 $LAMBDA$ 代表 λ $\lambda > 0$, X 为非负整数
学生 t 分布	t pdf t cdf	$Y = \text{t pdf}(X, N)$ $Y = \text{t cdf}(X, N)$	N 为正整数
离散均匀分布	unid pdf unidcdf	$Y = \text{unid pdf}(X, N)$ $Y = \text{unidcdf}(X, N)$	N 为正整数
连续均匀分布	unif pdf unifcdf	$Y = \text{unif pdf}(X, A, B)$ $Y = \text{unifcdf}(X, A, B)$	$A < B$

$$P = \text{chi2cdf}(X, N)$$

的 χ^2 逆分布函数为

$$X = \text{chi2inv}(P, N).$$

习 题 2

1. 下面两表是否可作为离散型随机变量的分布列？为什么？

X_1	-1	0	1
P	-0.5	0.9	0.6

X_2	0	1	2
P	0.6	0.1	0.15

2. 已知袋中有 5 只号码球, 编号为 1, 2, 3, 4, 5. 每次从袋中同时取出 3 只, 记下号码后重新放回, 以 X 表示所取出三只球中的最大号码, 求 X 的分布列.

3. 设在 15 只零件中, 有 3 只是次品, 从中不放回地依次取出 4 只, 以 X 表示取出 4 只中所含次品的只数, 求 X 的概率分布.

4. 设随机变量 X 的分布律分别为

$$(1) P(X=k) = \frac{a}{N}, k=1, 2, \dots, N;$$

$$(2) P(X=k) = \frac{a}{2^k}, k=1, 2, \dots,$$

试确定常数 a .

5. 进行某种试验, 成功的概率为 $\frac{3}{4}$, 失败的概率为 $\frac{1}{4}$. 以 X 表示直到试验首次成功时所需试验的次数, 试写出 X 的概率分布, 并求 X 取偶数的概率.

6. 将一颗骰子连掷两次, 以 X 表示两次所得点数之和, 写出 X 的分布列, 并计算 $P(X \leq 4)$, $P(7 \leq X \leq 10)$.

7. 某批零件的次品率为 0.1, 从这批零件中任取 20 件,

求: (1) 恰有 3 件次品的概率;

(2) 至少有 3 件次品的概率;

(3) 次品数的最可能值.

8. 假定人在一年 365 日中的任一日出生的概率是一样的, 试问: 在一个 730 人的单位中, 至少有两人生于元旦的概率是多少?

9. 在一定条件下, 某种微生物菌落在培养基中出现的概率为 0.2, 在相同的条件下, 分别在 10 个培养基中接种, 求至少在 4 个培养基中出现菌落的概率?

10. 某车间有同类设备 100 台, 每台设备工作互不影响. 如果每台设备发生故障的概率是 0.01, 且一台设备的故障可由一个人来处理. 问至少配备多少维修工, 才能保证设备发生故障但不能及时维修的概率小于 0.01.

11. 某地胃癌的发病率为 0.01%, 现普查 5 万人, 其中没有胃癌患者的概率是多少? 胃癌患者少于 5 人的概率是多少?

12. 一电话交换机每分钟接到的呼唤次数服从参数为 4 的泊松分布. 试求:

(1) 某一分钟内有 8 次呼唤的概率;

(2) 某一分钟内呼唤次数大于 10 次的概率.

13. 设随机变量 X 的密度函数为

$$f(x) = \begin{cases} cx, & 0 \leq x \leq 1, \\ 0, & \text{其它.} \end{cases}$$

求: (1) 常数 c ;

(2) $P(0.3 < X < 0.7)$;

(3) 常数 a , 使 $P(X > a) = P(X < a)$;

(4) X 的分布函数 $F(x)$.

14. 设随机变量 X 的概率密度为

$$f(x) = \begin{cases} \frac{1}{3}, & x \in [0, 1], \\ \frac{2}{9}, & x \in [3, 6], \\ 0, & \text{其它.} \end{cases}$$

若 k 使得 $P(X \geq k) = \frac{2}{3}$, 则 k 的取值范围是多少?

15. 设随机变量 X 的密度函数为

$$f(x) = \begin{cases} \frac{c}{\sqrt{1-x^2}}, & \text{当 } |x| < 1 \text{ 时,} \\ 0, & \text{其它.} \end{cases}$$

求常数 c 及 $P\left(|X| \leq \frac{1}{2}\right)$.

16. 设随机变量 X 的分布函数为

$$F(x) = \begin{cases} 1 - e^{-x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

(1) 求 X 的概率密度函数 $f(x)$;

(2) 计算 $P(X \leq 4)$, $P(X > 1)$.

17. 设随机变量 X 的分布函数为

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

求 $P(X = 1)$.

18. 设随机变量 X 在区间 $(2, 5)$ 上服从均匀分布, 现在对 X 进行 3 次独立观测, 求至少有两次观测值大于 3 的概率.

19. 设顾客在某银行的窗口等待服务的时间 X (单位: min) 服从参数为 $\lambda = \frac{1}{5}$ 的指数分布. 某顾客在窗口等待服务, 且超过 10 min 他就离开. 若他一个月要来银行 5 次, 以 Y 表示一个月内他未等到服务而离开窗口的次数, 写出 Y 的分布律并求至少有一次未等到服务而离开的概率.

20. 设打一次电话所用的时间 X (min) 服从参数 $\lambda = 0.1$ 的指数分布. 如果某人刚好在你前面走进电话间, 求你等待的时间:

(1) 超过 10 min 的概率;

(2) 在 10 min 到 20 min 之间的概率.

21. 某工厂生产螺栓的长度 (cm) 服从参数 $\mu = 10.05$, $\sigma = 0.06$ 的正态分布, 如果规定长度在 10.05 ± 0.12 内为合格品, 求一螺栓为不合格品的概率是多少?

22. 某厂生产的电子管的寿命 $X(h)$ 服从参数 $\mu = 160$, $\sigma = \sigma_0$ 的正态分布, 试问 σ_0 最大为多少时能使 $P(120 \leq X \leq 200) \geq 0.8$.

23. 某单位招聘 155 人, 按考试成绩录用, 共有 526 人报名, 假设报名者的考试成绩 $X \sim N(\mu, \sigma^2)$. 已知 90 分以上的 12 人, 60 分以下的 83 人, 若从高分到低分依次录取, 某人成绩为 78 分, 问此人能否被录取?

24. 设随机变量 $X \sim N(3, 2^2)$.

(1) 求 $P(2 < X \leq 5)$, $P(|X| > 2)$;

(2) 确定 c , 使得 $P(X > c) = P(X \leq c)$;

(3) 设 d 满足 $P(X > d) \geq 0.9$, 问 d 至多为多少?

25. 设测量误差 $X \sim N(0, 10^2)$, 现进行 100 次独立测量, 求误差的绝对值超过 19.6 的次数不小于 3 的概率.

26*. 设随机变量 X 服从几何分布, 试证: $P(X = n + k | X > n) = P(X = k)$, 其中 $n \geq 1$, $k = 1, 2, \dots$.

27*. 若每只母鸡产 k 个蛋的概率服从参数为 λ 的泊松分布, 而每个蛋能孵化成小鸡的概率为 p . 试证: 每只母鸡有 n 只小鸡的概率服从参数为 λp 的泊松分布.

28. 设随机变量 X 的分布列为

X	-2	-1	0	1	2
P	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{5}$

试求: (1) $Y = 2X + 1$ 的分布列;

(2) $Y = X^2$ 的分布列.

29. 设随机变量 X 服从区间 $[0, 1]$ 上的均匀分布, 求随机变量 $Y = -2 \ln X$ 的概率密度函数.

30. 设随机变量 X 服从区间 $[0, 6]$ 上的均匀分布, 求随机变量 $Y = |X - 3|$ 的概率密度函数.

31. 设随机变量 $X \sim N(0, 1)$, 求: (1) $Y = 2X^2 + 1$ 的概率密度函数; (2) $Y = |X|$ 的概率密度函数.

32*. 设随机变量 X 的概率分布为 $P(X = k) = \frac{1}{2^k}$, $k = 1, 2, \dots$. 试求随机变量 $Y = \sin\left(\frac{\pi}{2}X\right)$ 的分布律.

33*. 设随机变量 X 的概率密度为

$$f(x) = \begin{cases} \frac{1}{3\sqrt[3]{x^2}}, & x \in [1, 8], \\ 0, & \text{其它.} \end{cases}$$

$F(x)$ 是 X 的分布函数. 求随机变量 $Y = F(X)$ 的分布函数.

第3章 多维随机变量及其分布

为描述实际问题中的随机现象,有时需要多个随机变量.例如,打靶命中的位置要用两个随机变量——弹着点的横坐标 X 和纵坐标 Y 来描述,衡量儿童的身体发育情况要用多个随机变量——身高 X 、体重 Y 、胸围 Z 等来表述.通常把与同一个随机现象相联系的多个随机变量看作一个整体,称为多维随机变量或随机向量.比如由 n 个随机变量 $X_1, X_2, \dots, X_n$ 构成的 n 维随机变量,记作 $(X_1, X_2, \dots, X_n)$, 其中,第 i 个随机变量 X_i 又称为 $(X_1, X_2, \dots, X_n)$ 的第 i 个分量.

显然,第2章所讨论的随机变量是一维随机变量.为叙述方便,本章主要讨论二维随机变量, n 维随机变量的情形可以类似讨论.

§ 3.1 二维随机变量及其分布

1. 二维随机变量的分布函数

定义 3.1.1 设 (X, Y) 为二维随机变量,对任意实数 x, y , 二元函数

$$F(x, y) = P(X \leq x, Y \leq y) \quad (3.1.1)$$

称为二维随机变量 (X, Y) 的联合分布函数,简称为 (X, Y) 的分布函数.

分布函数 $F(x, y)$ 表示事件 $\{X \leq x\}$ 和事件 $\{Y \leq y\}$ 同时发生的概率.如果把二维随机变量 (X, Y) 看作是平面(二维空间)上的一个随机点,那么它的联合分布函数 $F(x, y)$ 就是随机点落入以 (x, y) 为顶点而位于该点左下方的无穷矩形区域 D 内的概率(图 3.1.1).

由上面的几何解释,容易得到随机点 (X, Y) 落在矩形区域: $x_1 < x \leq x_2, y_1 < y \leq y_2$ 内的概率为(图 3.1.2)

$$\begin{aligned} & P(x_1 < X \leq x_2, y_1 < Y \leq y_2) \\ &= P(X \leq x_2, Y \leq y_2) - P(X \leq x_1, Y \leq y_2) - \\ & \quad P(X \leq x_2, Y \leq y_1) + P(X \leq x_1, Y \leq y_1) \\ &= F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1). \end{aligned} \quad (3.1.2)$$

同一维随机变量类似,二维随机变量联合分布函数 $F(x, y)$ 也具有 3 条基本性质:

(1) $F(x, y)$ 是关于 x 或 y 的非减函数,即对于固定的 y , 当 $x_1 < x_2$ 时, $F(x_1, y) \leq F(x_2, y)$; 对于固定的 x , 当 $y_1 < y_2$ 时, $F(x, y_1) \leq F(x, y_2)$.

(2) $0 \leq F(x, y) \leq 1$, 且

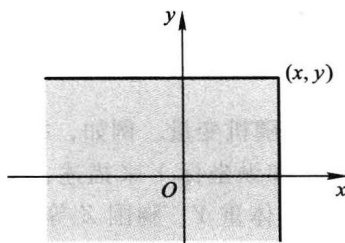


图 3.1.1

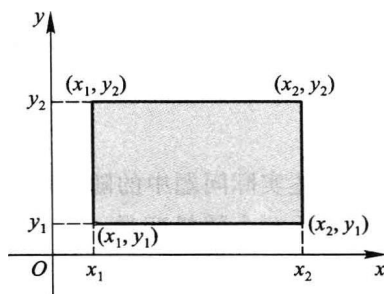


图 3.1.2

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

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

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

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

(3) 关于 x 或 y 都是右连续, 即

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

二维随机变量也有离散型和连续型两种常见的形式, 下面分别讨论.

2. 二维离散型随机变量

定义 3.1.2 若二维随机变量 (X, Y) 只取有限对或可数无限对不同值 (x_i, y_j) , $i, j = 1, 2, \dots$, 则称 (X, Y) 为二维离散型随机变量. 同时, 称

$$P(X = x_i, Y = y_j) = p_{ij}, \quad i, j = 1, 2, \dots \quad (3.1.3)$$

为 (X, Y) 的联合概率分布, 简称为概率分布或分布律.

联合概率分布也可以用下列表格表示.

表 3.1.1

$\begin{matrix} Y \\ X \end{matrix}$	y_1	y_2	$\dots$	y_j	$\dots$
x_1	p_{11}	p_{12}	$\dots$	p_{1j}	$\dots$
x_2	p_{21}	p_{22}	$\dots$	p_{2j}	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	p_{i1}	p_{i2}	$\dots$	p_{ij}	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

自然, p_{ij} 具有下列性质:

(1) $p_{ij} \geq 0$, $i, j = 1, 2, \dots$;

(2) $\sum_i \sum_j p_{ij} = 1$.

对于二维离散型随机变量 (X, Y) , 利用 (X, Y) 的概率分布可以求得 (X, Y) 的分布函数以及 (X, Y) 落入平面区域 G 内的概率:

$$F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p_{ij}. \quad (3.1.4)$$

$$P((X, Y) \in G) = \sum_{(x_i, y_j) \in G} p_{ij}. \quad (3.1.5)$$

例 3.1.1 设随机变量 X 在 1, 2, 3 三个整数中等可能取值, 另一个随机变量 Y 在 $1 \sim X$ 中等可能地取一整数值, 求 (X, Y) 的概率分布.

解 由假设, 随机变量 X 的可能取值为 1, 2, 3. 而 $Y \leq X$, 故 Y 的可能取值范围也为 1, 2, 3. 首先, 当 $j > i$ 时, $\{X=i, Y=j\}$ 为不可能事件, 故

$$P(X=i, Y=j) = 0, \quad j > i.$$

当 $j \leq i$ 时, 根据概率的乘法公式, 有

$$\begin{aligned} P(X=i, Y=j) &= P(X=i) \cdot P(Y=j|X=i) \\ &= \frac{1}{3i}, \quad i=1, 2, 3. \end{aligned}$$

由此得 (X, Y) 的概率分布如下:

表 3.1.2

$X \backslash Y$	1	2	3
1	$\frac{1}{3}$	0	0
2	$\frac{1}{6}$	$\frac{1}{6}$	0
3	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

例 3.1.2 设 (X, Y) 的分布律为

$X \backslash Y$	1	2	3
0	0.1	0.1	0.3
1	0.25	0	0.25

求: (1) $P(X=0)$; (2) $P(Y \leq 2)$; (3) $P(X < 1, Y \leq 2)$;

(4) $P(X+Y=2)$.

解 (1) $P(X=0) = P(X=0, Y=1) + P(X=0, Y=2) + P(X=0, Y=3)$
 $= 0.1 + 0.1 + 0.3 = 0.5.$

$$(2) P(Y \leq 2) = P(X=0, Y=1) + P(X=1, Y=1) + P(X=0, Y=2) + P(X=1, Y=2) \\ = 0.1 + 0.25 + 0.1 + 0 = 0.45.$$

$$(3) P(X < 1, Y \leq 2) = P(X=0, Y=1) + P(X=0, Y=2) \\ = 0.1 + 0.1 = 0.2.$$

$$(4) P(X+Y=2) = P(X=0, Y=2) + P(X=1, Y=1) \\ = 0.1 + 0.25 = 0.35.$$

3. 二维连续型随机变量

定义 3.1.3 设 $F(x, y)$ 是二维随机变量 (X, Y) 的联合分布函数, 如果存在非负二元函数 $f(x, y)$, 对于任意的实数 x, y 都有

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv \quad (3.1.6)$$

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

与一维情形类似, 当 $f(x, y)$ 连续时

$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y), \quad (3.1.7)$$

密度函数 $f(x, y)$ 具有如下性质:

$$(1) f(x, y) \geq 0;$$

$$(2) \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1.$$

可以证明: 若 D 是 Oxy 平面上的任一区域, 则随机点 (X, Y) 落在 D 内的概率为

$$P((X, Y) \in D) = \iint_D f(x, y) dx dy. \quad (3.1.8)$$

在几何上, 二元函数 $f(x, y)$ 在三维空间表示一曲面, (3.1.8) 式表明随机点 (X, Y) 落入区域 D 内的概率恰好等于以 D 为底, 以曲面 $f(x, y)$ 为顶的曲顶柱体的体积.

例 3.1.3 设二维随机变量 (X, Y) 的概率密度为

$$f(x, y) = \begin{cases} kx^2y, & x^2 \leq y \leq 1, \\ 0, & \text{其它.} \end{cases}$$

求: (1) 常数 k ; (2) $P(X \geq Y)$

解 (1) 由概率密度的性质可知

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1,$$

即

$$\int_{-1}^1 dx \int_{x^2}^1 kx^2y dy = \frac{4}{21}k = 1,$$

如图 3.1.3. 所以 $k = \frac{21}{4}$.

$$\begin{aligned}
 (2) \quad P(X \geq Y) &= \iint_{y \leq x} f(x, y) dx dy \\
 &= \int_0^1 dx \int_{x^2}^x \frac{21}{4} x^2 y dy \\
 &= \frac{3}{20},
 \end{aligned}$$

如图 3.1.4.

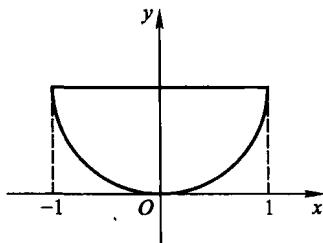


图 3.1.3

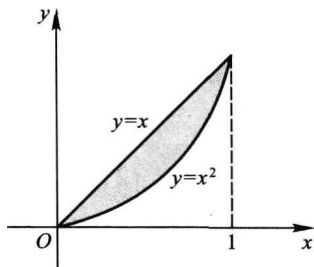


图 3.1.4

例 3.1.4 设二维随机变量 (X, Y) 的概率密度函数为

$$f(x, y) = \frac{12}{\pi^2 (9 + x^2)(16 + y^2)}, \quad -\infty < x, y < +\infty.$$

(1) 求 (X, Y) 的分布函数 $F(x, y)$;

(2) 求 $P(0 < X \leq 3, 0 < Y \leq 4)$.

解 (1) 由 (3.1.6) 式, 得

$$\begin{aligned}
 F(x, y) &= \int_{-\infty}^x \int_{-\infty}^y \frac{12}{\pi^2 (9 + u^2)(16 + v^2)} du dv \\
 &= \frac{12}{\pi^2} \cdot \int_{-\infty}^x \frac{du}{9 + u^2} \cdot \int_{-\infty}^y \frac{dv}{16 + v^2} \\
 &= \frac{1}{\pi^2} \left(\frac{\pi}{2} + \arctan \frac{x}{3} \right) \left(\frac{\pi}{2} + \arctan \frac{y}{4} \right).
 \end{aligned}$$

(2) 利用 (1) 的结果及 (3.1.2) 式,

$$\begin{aligned}
 P(0 < X \leq 3, 0 < Y \leq 4) &= F(3, 4) - F(0, 4) - F(3, 0) + F(0, 0) \\
 &= \frac{9}{16} - \frac{3}{8} - \frac{3}{8} + \frac{1}{4} = \frac{1}{16}.
 \end{aligned}$$

如同一维连续型随机变量一样, 下面介绍两种常用的二维连续型随机变量的分布: 均匀分布与二维正态分布.

定义 3.1.4 设 D 是平面上的有界区域, 其面积为 A , 若二维随机变量

(X, Y) 的概率密度为

$$f(x, y) = \begin{cases} \frac{1}{A}, & \text{当 } (x, y) \in D, \\ 0, & \text{其它,} \end{cases} \quad (3.1.9)$$

则称 (X, Y) 服从区域 D 上的均匀分布.

容易验证, $f(x, y)$ 满足密度函数的性质. 与第2章中服从区间 $[a, b]$ 上的均匀分布类似, 服从区域 D 上的均匀分布 (X, Y) 落在 D 中任一区域 D_1 的概率只与 D_1 的面积成正比, 而与 D_1 的位置和形状无关.

例 3.1.5 设二维随机变量 (X, Y) 服从区域 D 上的均匀分布, 其中 $D: 0 \leq x \leq 1, 0 \leq y \leq x$, 求 $P(X + Y \leq 1)$.

解 如图 3.1.5, D 的面积 $A = \frac{1}{2}$, 所以 (X, Y) 的概率密度为

$$f(x, y) = \begin{cases} 2, & (x, y) \in D, \\ 0, & \text{其它.} \end{cases}$$

$$\begin{aligned} \text{则 } P(X + Y \leq 1) &= \iint_{x+y \leq 1} f(x, y) dx dy \\ &= \iint_{D_1} 2 dx dy = \frac{1}{2}. \end{aligned}$$

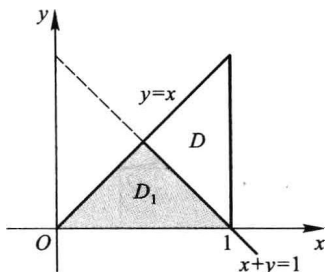


图 3.1.5

定义 3.1.5 若二维随机变量 (X, Y) 的概率密度函数为

$$\begin{aligned} f(x, y) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - \right. \right. \\ &\quad \left. \left. 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\}, \\ &\quad -\infty < x < +\infty, \quad -\infty < y < +\infty. \end{aligned} \quad (3.1.10)$$

其中 $\mu_1, \mu_2, \sigma_1 > 0, \sigma_2 > 0, \rho (|\rho| < 1)$ 均为常数, 则称 (X, Y) 服从参数 $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ 的二维正态分布. 记作 $(X, Y) \sim N(\mu_1, \sigma_1^2; \mu_2, \sigma_2^2; \rho)$.

可以验证函数 (3.1.10) 满足概率密度函数的两个性质, 其函数图像见图 3.1.6.

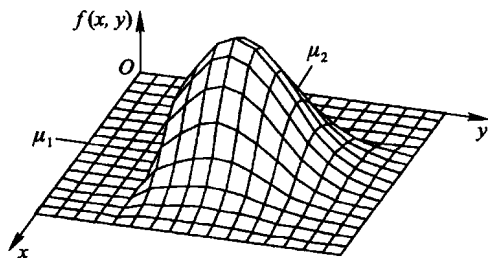


图 3.1.6

§ 3.2 边缘分布

二维随机变量的联合分布是把 (X, Y) 看作一个整体的分布. 其中分量 X 和 Y 都是一维随机变量, 也有各自的分布, 分别称 X 和 Y 的分布为二维随机变量 (X, Y) 关于 X 和 Y 的边缘分布. 本节讨论它们之间的关系.

1. 边缘分布函数

设二维随机变量 (X, Y) 的分布函数为 $F(x, y)$, 分别记关于 X 和 Y 的边缘分布函数为 $F_X(x)$ 和 $F_Y(y)$, 由于

$$F_X(x) = P(X \leq x) = P(X \leq x, Y < +\infty) = F(x, +\infty), \quad (3.2.1)$$

同理, 有

$$F_Y(y) = F(+\infty, y). \quad (3.2.2)$$

由 (3.2.1) 及 (3.2.2) 式可以看出, 边缘分布函数 $F_X(x)$, $F_Y(y)$ 完全由联合分布函数 $F(x, y)$ 来确定.

2. 二维离散型随机变量的边缘分布

设二维离散型随机变量 (X, Y) 的概率分布为

$$P(X = x_i, Y = y_j) = p_{ij}, \quad i = 1, 2, \dots; \quad j = 1, 2, \dots,$$

由 (3.2.1) 式及 (3.1.4) 式可得

$$F_X(x) = F(x, +\infty) = \sum_{x_i \leq x} \sum_{y_j < +\infty} p_{ij}; \quad (3.2.3)$$

另一方面, 有

$$F_X(x) = \sum_{x_i \leq x} P(X = x_i), \quad i = 1, 2, \dots.$$

与 (3.2.3) 式比较, 可得

$$P(X = x_i) = \sum_j p_{ij}, \quad i = 1, 2, \dots. \quad (3.2.4)$$

同理可得

$$P(Y = y_j) = \sum_i p_{ij}, \quad j = 1, 2, \dots. \quad (3.2.5)$$

记

$$p_{i \cdot} = \sum_j p_{ij}, \quad p_{\cdot j} = \sum_i p_{ij}.$$

分别称 (3.2.4) 和 (3.2.5) 为二维离散型随机变量 (X, Y) 的关于 X 和 Y 的边缘概率分布. 它们又可分别表示为

	x_1	x_2	$\dots$	x_i	$\dots$
$P(X = x_i)$	$p_{i \cdot}$	$p_{2 \cdot}$	$\dots$	$p_{i \cdot}$	$\dots$

$Y=y_j$	y_1	y_2	$\cdots$	y_j	$\cdots$
$P(Y=y_j)$	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.j}$	$\cdots$

例 3.2.1 袋中装有 4 只白球和 2 只黑球，分别采取有放回和无放回的取法，每次取一个，共取两次，定义随机变量 X, Y 如下：

$$X = \begin{cases} 0, & \text{第一次摸出黑球,} \\ 1, & \text{第一次摸出白球.} \end{cases} \quad Y = \begin{cases} 0, & \text{第二次摸出黑球,} \\ 1, & \text{第二次摸出白球.} \end{cases}$$

求 (X, Y) 的联合概率分布及其边缘分布。

解 (1) 对于有放回的情形

$$P(X=0, Y=0) = P(X=0)P(Y=0|X=0) = \frac{2}{6} \cdot \frac{2}{6} = \frac{1}{9}.$$

同样可得

$$P(X=0, Y=1) = \frac{2}{9},$$

$$P(X=1, Y=0) = \frac{2}{9},$$

$$P(X=1, Y=1) = \frac{4}{9}.$$

由此， (X, Y) 的联合分布及边缘分布如下：

表 3.2.1

Y	0	1	$P(X=x_i) = p_{i.}$
X			
0	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{3}$
1	$\frac{2}{9}$	$\frac{4}{9}$	$\frac{2}{3}$
$P(Y=y_j) = p_{.j}$	$\frac{1}{3}$	$\frac{2}{3}$	1

(2) 对于无放回的情形， (X, Y) 的联合分布及边缘分布如下：

表 3.2.2

Y	0	1	$P(X=x_i) = p_{i.}$
X			
0	$\frac{1}{15}$	$\frac{4}{15}$	$\frac{1}{3}$
1	$\frac{4}{15}$	$\frac{6}{15}$	$\frac{2}{3}$
$P(Y=y_j) = p_{.j}$	$\frac{1}{3}$	$\frac{2}{3}$	1

由此例我们可以看出, X 和 Y 的边缘分布完全相同, 但是它们的联合分布却不相同. 也就是说, 在一般情况下, 联合分布唯一确定其边缘分布, 反之不然.

3. 二维连续型随机变量的边缘分布

若二维连续型随机变量 (X, Y) 的分布函数与密度函数分别为 $F(x, y)$ 与 $f(x, y)$, 则有

$$\begin{aligned} F_X(x) &= F(x, +\infty) = \int_{-\infty}^x \int_{-\infty}^{+\infty} f(u, y) dy du \\ &= \int_{-\infty}^x \left[\int_{-\infty}^{+\infty} f(u, y) dy \right] du, \end{aligned}$$

$$\begin{aligned} F_Y(y) &= F(+\infty, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^y f(x, v) dx dv \\ &= \int_{-\infty}^y \left[\int_{-\infty}^{+\infty} f(x, v) dx \right] dv, \end{aligned}$$

记

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy, \quad (3.2.6)$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx. \quad (3.2.7)$$

分别称 $f_X(x)$, $f_Y(y)$ 为二维连续型随机变量 (X, Y) 关于 X 和 Y 的边缘概率密度函数, 简称边缘密度函数.

例 3.2.2 设二维随机变量 (X, Y) 服从区域 D 上的均匀分布, 其中 $D = \left\{ (x, y) \mid x \geq 0, y \geq 0, x + \frac{y}{2} \leq 1 \right\}$, 求其边缘密度函数.

解 由于区域 D 为直角三角形, 其面积为 1, 所以 (X, Y) 的密度函数为

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

由图 3.2.1 所示, 当 $0 \leq x \leq 1$ 时,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f(x, y) dy = \int_{-\infty}^0 0 dy + \int_0^{2(1-x)} 1 dy + \int_{2(1-x)}^{+\infty} 0 dy \\ &= 2(1-x), \end{aligned}$$

当 $x < 0$ 或 $x > 1$ 时, 因 $f(x, y) = 0$, 所以 $f_X(x) = 0$. 故

$$f_X(x) = \begin{cases} 2(1-x), & 0 \leq x \leq 1, \\ 0, & \text{其它.} \end{cases}$$

同理

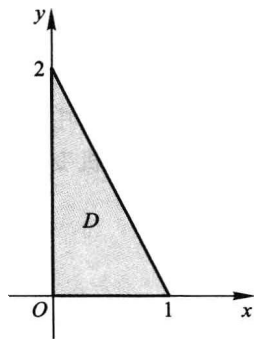


图 3.2.1

$$f_Y(y) = \begin{cases} 1 - \frac{y}{2}, & 0 \leq y \leq 2, \\ 0, & \text{其它.} \end{cases}$$

例 3.2.3 求二维正态随机变量的边缘密度函数.

解 设二维正态随机变量 $(X, Y) \sim N(\mu_1, \sigma_1^2; \mu_2, \sigma_2^2; \rho)$, 关于 X 和 Y 的边缘密度函数分别记为 $f_X(x)$ 和 $f_Y(y)$. 为了计算方便, 令

$$\frac{x - \mu_1}{\sigma_1} = u, \quad \frac{y - \mu_2}{\sigma_2} = v.$$

由 (3.2.6) 及 (3.1.10) 式得

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f(x, y) dy \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2(1-\rho^2)}(u^2 - 2\rho uv + v^2)\right] \sigma_2 dv \\ &= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2(1-\rho^2)}[(u^2 - \rho^2 u^2) + (\rho^2 u^2 - 2\rho uv + v^2)]\right\} dv \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{u^2}{2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{-\frac{(v-\rho u)^2}{2(1-\rho^2)}} dv, \end{aligned}$$

注意到上式中的被积函数恰好是服从正态分布 $N(\rho u, 1 - \rho^2)$ 的随机变量的密度函数, 故

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{u^2}{2}} = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \quad -\infty < x < +\infty. \quad (3.2.8)$$

同理可得

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}, \quad -\infty < y < +\infty. \quad (3.2.9)$$

由此可见: 二维正态分布的两个边缘分布都是一维正态分布, 而且这两个边缘分布都与其中参数 ρ 无关. 这表明仅仅由 X 和 Y 的边缘分布, 一般不能完全确定二维随机变量 (X, Y) 的联合分布.

* § 3.3 条件分布

本节将讨论对于二维随机变量, 当其中某一随机变量的取值确定后, 另一随机变量的分布问题.

1. 离散型随机变量的条件分布

设离散型随机变量 (X, Y) 的联合概率分布为

$$P(X=x_i, Y=y_j) = p_{ij}, \quad i, j=1, 2, \dots,$$

关于 X 和 Y 的边缘分布列为

$$P(X=x_i) = p_{i\cdot}, \quad i=1, 2, \dots,$$

$$P(Y=y_j) = p_{\cdot j}, \quad j=1, 2, \dots,$$

对某一固定的 i , 若 $P(X=x_i) = p_{i\cdot} > 0$, 则称

$$P(Y=y_j|X=x_i) = \frac{P(X=x_i, Y=y_j)}{P(X=x_i)} = \frac{p_{ij}}{p_{i\cdot}}, \quad j=1, 2, \dots. \quad (3.3.1)$$

为在 $X=x_i$ 的条件下, 随机变量 Y 的条件分布律.

同样, 对于某一固定的 j , 若 $P(Y=y_j) > 0$, 则称

$$P(X=x_i|Y=y_j) = \frac{P(X=x_i, Y=y_j)}{P(Y=y_j)} = \frac{p_{ij}}{p_{\cdot j}}, \quad i=1, 2, \dots. \quad (3.3.2)$$

为在 $Y=y_j$ 的条件下, 随机变量 X 的条件分布律.

条件分布律具有下列性质:

$$(1) P(X=x_i|Y=y_j) \geq 0, \quad P(Y=y_j|X=x_i) \geq 0;$$

$$(2) \sum_i P(X=x_i|Y=y_j) = 1, \quad \sum_j P(Y=y_j|X=x_i) = 1.$$

例 3.3.1 在例 3.1.1 中, 求关于 X 和 Y 的边缘分布律及在 $Y=1$ 的条件下, X 的条件分布律.

解 由例 3.1.1 的 (X, Y) 联合概率分布可得关于 X 和 Y 的边缘分布如下:

表 3.3.1

$X \backslash Y$	1	2	3	$P(X=x_i) = p_{i\cdot}$
1	$\frac{1}{3}$	0	0	$\frac{1}{3}$
2	$\frac{1}{6}$	$\frac{1}{6}$	0	$\frac{1}{3}$
3	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$
$P(Y=y_j) = p_{\cdot j}$	$\frac{11}{18}$	$\frac{5}{18}$	$\frac{1}{9}$	1

在 $Y=1$ 的条件下有

$$P(X=1|Y=1) = \frac{\frac{1}{3}}{\frac{11}{18}} = \frac{6}{11},$$

$$P(X=2|Y=1) = \frac{\frac{1}{6}}{\frac{11}{18}} = \frac{3}{11},$$

$$P(X=3|Y=1) = \frac{\frac{1}{9}}{\frac{11}{18}} = \frac{2}{11},$$

故在 $Y=1$ 的条件下, X 的条件分布列为

X	1	2	3
$P(X=x_i Y=1)$	$\frac{6}{11}$	$\frac{3}{11}$	$\frac{2}{11}$

2. 连续型随机变量的条件分布

当 X 和 Y 是连续型随机变量时, 事件 $\{X=x\}$ 和 $\{Y=y\}$ 的概率都是 0, 所以不能直接使用条件概率的公式给出在 $X=x$ 及 $Y=y$ 的条件下的条件分布. 为此考虑随机变量 X 落在 x 的某邻域内的概率, 进而给出连续型随机变量的条件分布.

对于任意 $\varepsilon > 0$, 若 $P(x-\varepsilon < X \leq x+\varepsilon) > 0$, 对任意实数 y , 极限 $\lim_{\varepsilon \rightarrow 0} P(Y \leq y | x-\varepsilon < X \leq x+\varepsilon) = \lim_{\varepsilon \rightarrow 0} \frac{P(x-\varepsilon < X \leq x+\varepsilon, Y \leq y)}{P(x-\varepsilon < X \leq x+\varepsilon)}$ 存在, 则称它为在给定条件 $X=x$ 下, 随机变量 Y 的条件分布函数, 记为 $F_{Y|X}(y|x)$.

若二维随机变量 (X, Y) 的概率分布函数为 $F(x, y)$, 密度函数为 $f(x, y)$, 且 $f(x, y)$ 在点 (x, y) 处连续, 边缘分布函数, 密度函数分别为 $F_X(x)$, $f_X(x)$, 且 $f_X(x) > 0$ 连续, 则有

$$\begin{aligned} F_{Y|X}(y|x) &= \lim_{\varepsilon \rightarrow 0} \frac{P(x-\varepsilon < X \leq x+\varepsilon, Y \leq y)}{P(x-\varepsilon < X \leq x+\varepsilon)} \\ &= \lim_{\varepsilon \rightarrow 0} \frac{F(x+\varepsilon, y) - F(x-\varepsilon, y)}{F_X(x+\varepsilon) - F_X(x-\varepsilon)} \\ &= \lim_{\varepsilon \rightarrow 0} \frac{\frac{F(x+\varepsilon, y) - F(x-\varepsilon, y)}{2\varepsilon}}{\frac{F_X(x+\varepsilon) - F_X(x-\varepsilon)}{2\varepsilon}} \\ &= \frac{\frac{\partial F(x, y)}{\partial x}}{\frac{dF_X(x)}{dx}} = \frac{\int_{-\infty}^y f(x, v) dv}{f_X(x)}, \end{aligned}$$

即

$$F_{Y|X}(y|x) = \int_{-\infty}^y \frac{f(x,v)}{f_X(x)} dv. \quad (3.3.3)$$

若记 $X=x$ 条件下 Y 的条件概率密度函数为 $f_{Y|X}(y|x)$, 则有

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}. \quad (3.3.4)$$

类似地有

$$F_{X|Y}(x|y) = \int_{-\infty}^x \frac{f(u,y)}{f_Y(y)} du, \quad (3.3.5)$$

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}, \quad (3.3.6)$$

其中 $f_Y(y)$ 是 Y 的边缘密度函数, $f_Y(y) > 0$, 且连续.

例 3.3.2 设二维随机变量 $(X,Y) \sim N(\mu_1, \sigma_1^2; \mu_2, \sigma_2^2; \rho)$, 求条件概率密度函数 $f_{Y|X}(y|x)$ 和 $f_{X|Y}(x|y)$.

解 由例 3.2.3 知, $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$. 依 (3.3.4) 式,

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f(x,y)}{f_X(x)} \\ &= \frac{1}{\sqrt{2\pi} \sigma_2} \frac{1}{\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{\rho^2 (x-\mu_1)^2}{\sigma_1^2} - \right. \right. \\ &\quad \left. \left. 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1 \sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\} \\ &= \frac{1}{\sqrt{2\pi} \sigma_2} \frac{1}{\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{\rho(x-\mu_1)}{\sigma_1} - \frac{y-\mu_2}{\sigma_2} \right]^2 \right\} \\ &= \frac{1}{\sqrt{2\pi} \sigma_2} \frac{1}{\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2\sigma_2^2(1-\rho^2)} \left[y - (\mu_2 + \right. \right. \\ &\quad \left. \left. \rho \frac{\sigma_2}{\sigma_1} (x-\mu_1)) \right]^2 \right\}. \end{aligned}$$

即在 $X=x$ 条件下 Y 的条件分布为正态分布

$$N(\mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x-\mu_1), \sigma_2^2(1-\rho^2)).$$

注意到, 这里 $l(x) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x-\mu_1)$ 是 x 的线性函数. 同理可得

$$f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi} \sigma_1} \frac{1}{\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2\sigma_1^2(1-\rho^2)} \left[x - (\mu_1 + \rho \frac{\sigma_1}{\sigma_2} (y-\mu_2)) \right]^2 \right\}.$$

例 3.3.3 设二维随机变量 (X, Y) 的概率密度为

$$f(x, y) = \begin{cases} 3x, & 0 < x < 1, 0 < y < x, \\ 0, & \text{其它.} \end{cases}$$

求条件概率密度 $f_{Y|X}(y|x)$ 以及概率 $P\left(Y \leq \frac{1}{8} \mid X = \frac{1}{4}\right)$.

解 边缘密度 $f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \begin{cases} 3x^2, & 0 < x < 1, \\ 0, & \text{其它.} \end{cases}$

则当 $0 < x < 1$ 时, $f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = \begin{cases} \frac{1}{x}, & 0 < y < x, \\ 0, & \text{其它.} \end{cases}$

于是 $f_{Y|X}\left(y \mid x = \frac{1}{4}\right) = \begin{cases} 4, & 0 < y < \frac{1}{4}, \\ 0, & \text{其它.} \end{cases}$

从而
$$\begin{aligned} P\left(Y \leq \frac{1}{8} \mid X = \frac{1}{4}\right) &= \int_{-\infty}^{\frac{1}{8}} f_{Y|X}\left(y \mid x = \frac{1}{4}\right) dy \\ &= \int_0^{\frac{1}{8}} 4 dy = \frac{1}{2}. \end{aligned}$$

§ 3.4 随机变量的独立性

在第1章中, 我们讨论了随机事件的相互独立性, 下面利用事件相互独立的概念引入随机变量相互独立的概念.

定义 3.4.1 设二维随机变量 (X, Y) 的联合分布函数, 关于 X 和 Y 的边缘分布函数分别为 $F(x, y)$, $F_X(x)$, $F_Y(y)$. 若对任意的实数 x, y 有

$$F(x, y) = F_X(x) F_Y(y), \quad (3.4.1)$$

则称随机变量 X 与 Y 相互独立.

由分布函数的定义, (3.4.1)式可以写成

$$P(X \leq x, Y \leq y) = P(X \leq x) P(Y \leq y). \quad (3.4.2)$$

由此, 随机变量 X 与 Y 的相互独立是指对任意实数 x, y , 随机事件 $\{X \leq x\}$ 与 $\{Y \leq y\}$ 相互独立.

当 (X, Y) 为离散型随机变量时, (3.4.1)式等价于

$$P(X = x_i, Y = y_j) = P(X = x_i) P(Y = y_j), \quad i, j = 1, 2, \dots. \quad (3.4.3)$$

当 (X, Y) 为连续型随机变量时, (3.4.1)式等价于

$$f(x, y) = f_X(x) f_Y(y), \quad (3.4.4)$$

其中 $f(x, y)$, $f_X(x)$, $f_Y(y)$ 分别为 (X, Y) 的联合密度函数与边缘密度函数.

例 3.4.1 讨论例 3.2.1 中随机变量 X 与 Y 的独立性.

解 在有放回抽取的情况下, $P(X=i, Y=j) = P(X=i) \cdot P(Y=j)$, $i, j = 0, 1$, 故 X 与 Y 相互独立; 而在无放回的情况下, 因为 $P(X=0, Y=0) \neq P(X=0)P(Y=0)$, 所以 X 与 Y 不相互独立, 这与直观解释一致.

例 3.4.2 讨论例 3.2.2 中随机变量 X 与 Y 的独立性.

解 因为 (X, Y) 的概率密度为

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

而 X 与 Y 的边缘密度分别为

$$f_X(x) = \begin{cases} 2(1-x), & 0 \leq x \leq 1, \\ 0, & \text{其它.} \end{cases} \quad f_Y(y) = \begin{cases} 1 - \frac{y}{2}, & 0 \leq y \leq 2, \\ 0, & \text{其它.} \end{cases}$$

显然, $f(x, y) \neq f_X(x) \cdot f_Y(y)$, 故 X 与 Y 不相互独立.

例 3.4.3 设二维随机变量 $(X, Y) \sim N(\mu_1, \sigma_1^2; \mu_2, \sigma_2^2; \rho)$, 证明: X 与 Y 相互独立的充分必要条件是 $\rho = 0$.

证明 充分性 设 $\rho = 0$, 由 (3.1.10) 式得到

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{1}{2}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]} = f_X(x)f_Y(y),$$

从而 X 与 Y 相互独立.

必要性 设 X 与 Y 相互独立, 则对任意的 x, y , 有

$$f(x, y) = f_X(x)f_Y(y).$$

特别令 $x = \mu_1, y = \mu_2$, 则有 $f(\mu_1, \mu_2) = f_X(\mu_1)f_Y(\mu_2)$, 即

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} = \frac{1}{\sqrt{2\pi}\sigma_1} \cdot \frac{1}{\sqrt{2\pi}\sigma_2},$$

于是

$$\sqrt{1-\rho^2} = 1,$$

故 $\rho = 0$.

在本章 § 3.2 我们曾经指出, 仅由 X 和 Y 的边缘分布一般不能确定 (X, Y) 的联合分布. 但是, 当 X 与 Y 相互独立时, 边缘概率密度函数 (或边缘分布函数) 的乘积就是 (X, Y) 的联合概率密度函数 (或联合分布函数), 即边缘分布完全确定联合分布.

§ 3.5 二维随机变量函数的分布

类似于一维随机变量,在实际问题中,经常需要考虑二维随机变量的函数及其分布.

设 (X, Y) 为二维随机变量, 则 $Z = g(X, Y)$ 为一维随机变量, 本节讨论如何由已知的 (X, Y) 的分布去寻找 Z 的分布. 由于二维随机变量的函数比较复杂, 所以本节只介绍一般方法和几种常见类型.

1. 离散型随机变量函数的分布

设二维离散型随机变量 (X, Y) 的分布律为 $P(X = x_i, Y = y_j) = p_{ij} (i, j = 1, 2, \dots)$, $Z = g(X, Y)$ 的可能取值为 $z_k (k = 1, 2, \dots)$, 则 Z 的分布律为

$$P(Z = z_k) = P(g(X, Y) = z_k) = \sum_{g(x_i, y_j) = z_k} p_{ij}, \quad k = 1, 2, \dots \quad (3.5.1)$$

例 3.5.1 设二维随机变量 (X, Y) 的分布律为

$X \backslash Y$	-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) $Z = X + Y$ 的分布律; (2) $Z = X^2 + Y^2$ 的分布律.

解 (1) Z 的可能取值为 -2, 0, 1, 3, 4.

$$P(Z = -2) = P(X = -1, Y = -1) = \frac{5}{20},$$

$$P(Z = 0) = P(X = -1, Y = 1) = \frac{2}{20},$$

$$P(Z = 1) = P(X = -1, Y = 2) + P(X = 2, Y = -1) = \frac{9}{20},$$

$$P(Z = 3) = P(X = 2, Y = 1) = \frac{3}{20},$$

$$P(Z = 4) = P(X = 2, Y = 2) = \frac{1}{20},$$

从而得到 $Z = X + Y$ 的分布律为

Z	-2	0	1	3	4
P	$\frac{5}{20}$	$\frac{2}{20}$	$\frac{9}{20}$	$\frac{3}{20}$	$\frac{1}{20}$

(2) Z 的可能取值为 2, 5, 8. 与(1)类似可得 $Z = X^2 + Y^2$ 的分布律:

Z	2	5	8
P	$\frac{7}{20}$	$\frac{12}{20}$	$\frac{1}{20}$

2. 连续型随机变量函数的分布

设二维连续型随机变量 (X, Y) 的概率密度为 $f(x, y)$, 若 $Z = g(X, Y)$ 为连续型变量, 则 Z 的分布函数为

$$F_Z(z) = P(Z \leq z) = P(g(X, Y) \leq z) = \iint_{g(x, y) \leq z} f(x, y) dx dy. \quad (3.5.2)$$

求得 Z 的分布函数后, 经过求导可得到 Z 的密度函数 $f_Z(z) = F'_Z(z)$.

下面介绍几种常见类型.

(1) 和 $(Z = X + Y)$ 的分布

设二维随机变量 (X, Y) 的密度函数为 $f(x, y)$, 记随机变量函数 $Z = X + Y$ 的密度函数与分布函数分别为 $f_Z(z)$, $F_Z(z)$, 则有

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(X + Y \leq z) = \iint_{x+y \leq z} f(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{z-x} f(x, y) dy \right] dx. \end{aligned}$$

积分区域见图 3.5.1 所示, 在积分 $\int_{-\infty}^{z-x} f(x, y) dy$ 中, 作变量代换, 令 $y = u - x$ 或 $u = y + x$, $dy = du$, 则积分 $\int_{-\infty}^{z-x} f(x, y) dy = \int_{-\infty}^z f(x, u - x) du$. 于是得到 Z 的分布函数

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^z f(x, u - x) du \right] dx \\ &= \int_{-\infty}^z \left[\int_{-\infty}^{+\infty} f(x, u - x) dx \right] du. \end{aligned}$$

Z 的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, z - x) dx. \quad (3.5.3)$$

由 X, Y 的对称性, $f_Z(z)$ 又可表示成

$$f_Z(z) = \int_{-\infty}^{+\infty} f(z - y, y) dy. \quad (3.5.4)$$

特别地, 当 X 与 Y 相互独立时, 由于 $f(x, y) = f_X(x)f_Y(y)$, 故有

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x)f_Y(z - x) dx, \quad (3.5.5)$$

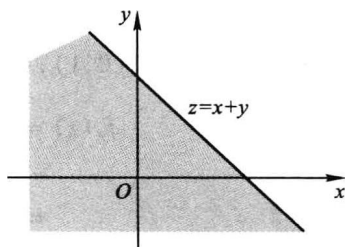


图 3.5.1

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(z-y)f_Y(y)dy. \quad (3.5.6)$$

公式(3.5.5)和(3.5.6)称为概率密度函数的卷(褶)积公式.

例 3.5.2 设二维随机变量 (X, Y) 的概率密度为

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

求 $Z = X + Y$ 的概率密度 $f_Z(z)$.

解 由公式(3.5.3),

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, z-x)dx,$$

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

当 $z \leq 0$ 或 $z \geq 2$ 时, $f_Z(z) = 0$;

当 $0 < z < 1$ 时, $f_Z(z) = \int_0^z (2-z)dx = z(2-z)$;

当 $1 \leq z < 2$ 时, $f_Z(z) = \int_{z-1}^1 (2-z)dx = (2-z)^2$.

即 Z 的概率密度为

$$f_Z(z) = \begin{cases} z(2-z), & 0 < z < 1, \\ (2-z)^2, & 1 \leq z < 2, \\ 0, & \text{其它.} \end{cases}$$

例 3.5.3 设随机变量 X 与 Y 相互独立, 并且都服从 $N(0, 1)$, 求 $Z = X + Y$ 的分布.

解 已知 $X \sim N(0, 1)$, $Y \sim N(0, 1)$, 由(3.5.5)式,

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-x)^2}{2}} dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-\frac{x^2 + (z-x)^2}{2}} dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-(x - \frac{z}{2})^2 - \frac{z^2}{4}} dx \\ &= \frac{e^{-\frac{z^2}{4}}}{2\pi} \int_{-\infty}^{+\infty} e^{-(x - \frac{z}{2})^2} dx. \end{aligned}$$

作变量代换, 令 $x - \frac{z}{2} = \frac{t}{\sqrt{2}}$, $dx = \frac{1}{\sqrt{2}} dt$, 于是

$$\begin{aligned}
 f_Z(z) &= \frac{e^{-\frac{z^2}{4}}}{2\pi} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2}} e^{-\frac{t^2}{2}} dt \\
 &= \frac{1}{2\sqrt{\pi}} e^{-\frac{z^2}{4}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\
 &= \frac{1}{2\sqrt{\pi}} e^{-\frac{z^2}{4}} = \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{z^2}{2(\sqrt{2})^2}},
 \end{aligned}$$

所以 $Z = X + Y \sim N(0, 2)$.

一般地, 若 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, 且 X 与 Y 相互独立, 则

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

(2) 商 $\left(Z = \frac{X}{Y}\right)$ 的分布

设 (X, Y) 的密度函数为 $f(x, y)$, $Z = \frac{X}{Y}$ 的分布函数为

$$\begin{aligned}
 F_Z(z) &= P\left(\frac{X}{Y} \leq z\right) = \iint_{\frac{x}{y} \leq z} f(x, y) dx dy \\
 &= \int_0^{+\infty} \left[\int_{-\infty}^{yz} f(x, y) dx \right] dy + \int_{-\infty}^0 \left[\int_{yz}^{+\infty} f(x, y) dx \right] dy.
 \end{aligned}$$

积分区域见图 3.5.2 所示. 令 $x = yu$, $dx = ydu$, 则

$$\begin{aligned}
 F_Z(z) &= \int_0^{+\infty} \left[\int_{-\infty}^z f(yu, y) y du \right] dy + \int_{-\infty}^0 \left[\int_z^{+\infty} f(yu, y) y du \right] dy \\
 &= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^z f(yu, y) |y| du \right] dy \\
 &= \int_{-\infty}^z \left[\int_{-\infty}^{+\infty} f(yu, y) |y| dy \right] du,
 \end{aligned}$$

从而得 Z 的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} f(yz, y) |y| dy. \quad (3.5.7)$$

特别地, 当 X 与 Y 相互独立时,

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(yz) f_Y(y) |y| dy.$$

(3.5.8)

例 3.5.4 设随机变量 X 与 Y 相互独立, 且服从同一分布, 其概率密度为

$$f(x) = \begin{cases} \frac{1000}{x^2}, & x > 1000, \\ 0, & \text{其它.} \end{cases}$$

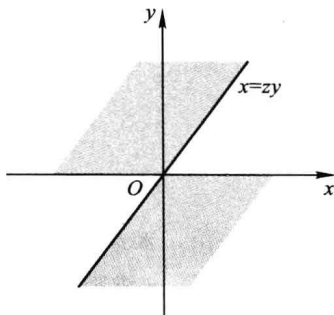


图 3.5.2

求 $Z = \frac{X}{Y}$ 的概率密度.

解 由(3.5.8)式,

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(yz)f_Y(y) |y| dy.$$

当 $z \leq 0$ 时, $f_Z(z) = 0$;

$$\text{当 } 0 < z < 1 \text{ 时, } f_Z(z) = \int_{\frac{1}{1000}}^{+\infty} \frac{1}{z^2 y^3} dy = \frac{1}{2};$$

$$\text{当 } z \geq 1 \text{ 时, } f_Z(z) = \int_{\frac{1}{1000}}^{+\infty} \frac{1}{z^2 y^3} dy = \frac{1}{2z^2}.$$

于是

$$f_Z(z) = \begin{cases} 0, & z \leq 0, \\ \frac{1}{2}, & 0 < z < 1, \\ \frac{1}{2z^2}, & z \geq 1. \end{cases}$$

(3) $M = \max(X, Y)$ 及 $N = \min(X, Y)$ 的分布

设 X, Y 是相互独立的随机变量, 它们的分布函数分别为 $F_X(x)$ 和 $F_Y(y)$, 由于

$$\{\max(X, Y) \leq z\} = \{X \leq z, Y \leq z\},$$

所以

$$\begin{aligned} F_M(z) &= P(M \leq z) = P(\max(X, Y) \leq z) \\ &= P(X \leq z, Y \leq z) = P(X \leq z)P(Y \leq z) \\ &= F_X(z)F_Y(z), \end{aligned}$$

即

$$F_M(z) = F_X(z)F_Y(z). \quad (3.5.9)$$

同理

$$\begin{aligned} F_N(z) &= P(N \leq z) = 1 - P(\min(X, Y) > z) \\ &= 1 - P(X > z, Y > z) \\ &= 1 - P(X > z)P(Y > z) \\ &= 1 - [1 - P(X \leq z)][1 - P(Y \leq z)] \\ &= 1 - [1 - F_X(z)][1 - F_Y(z)]. \end{aligned} \quad (3.5.10)$$

最后指出: 本章中有关二维随机变量的概念及性质, 都可以推广到 n 维随机变量的情形. n 维随机变量 $(X_1, X_2, \dots, X_n)$ 的分布函数为

$$F(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n),$$

其中 $x_1, x_2, \dots, x_n$ 为任意实数.

若存在非负函数 $f(x_1, x_2, \dots, x_n)$, 使得对任意实数 $x_1, x_2, \dots, x_n$ 有

$$F(x_1, x_2, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(u_1, u_2, \dots, u_n) du_1 du_2 \cdots du_n,$$

则称 $(X_1, X_2, \dots, X_n)$ 为 n 维连续型随机变量, 称 $f(x_1, x_2, \dots, x_n)$ 为 $(X_1, X_2, \dots, X_n)$ 的概率密度函数. 同样, 对每个分量 X_i , 其边缘密度函数 $f_{X_i}(x_i) (i = 1, 2, \dots, n)$ 为

$$f_{X_i}(x_i) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n.$$

特别若对任意的实数 $x_1, x_2, \dots, x_n$, 有

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

则称随机变量 $X_1, X_2, \dots, X_n$ 相互独立.

若 $X_i \sim N(\mu_i, \sigma_i^2) (i = 1, 2, \dots, n)$, 并且 $X_1, X_2, \dots, X_n$ 相互独立, 则随机变量

$$Z = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right). \quad (3.5.11)$$

特别地, 当 $X_i \sim N(\mu, \sigma^2) (i = 1, 2, \dots, n)$, 并且 $X_1, X_2, \dots, X_n$ 相互独立, 则

$$Z = \sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2). \quad (3.5.12)$$

若 $X_1, X_2, \dots, X_n$ 相互独立, 设 X_i 的分布函数为 $F_{X_i}(x_i) (i = 1, 2, \dots, n)$, 则 $Z = \max\{X_1, X_2, \dots, X_n\}$ 的分布函数为

$$F_{\max}(z) = F_{X_1}(z) F_{X_2}(z) \cdots F_{X_n}(z), \quad (3.5.13)$$

$Z = \min\{X_1, X_2, \dots, X_n\}$ 的分布函数为

$$F_{\min}(z) = 1 - [1 - F_{X_1}(z)][1 - F_{X_2}(z)] \cdots [1 - F_{X_n}(z)]. \quad (3.5.14)$$

如果 $X_1, X_2, \dots, X_n$ 具有相同的分布函数 $F(x)$, 则有

$$F_{\max}(z) = [F(z)]^n, \quad (3.5.15)$$

$$F_{\min}(z) = 1 - [1 - F(z)]^n. \quad (3.5.16)$$

例 3.5.5 设某种电子器件的寿命(单位:天)近似服从 $N(1\,195, 15^2)$, 现从中随机选取 3 件, 求: (1) 其中没有一件寿命超过 1 210 天的概率; (2) 其中没有一件寿命少于 1 210 天的概率.

解 设 3 件产品的寿命分别为 $X_i (i = 1, 2, 3)$, 则 $X_i \sim N(1\,195, 15^2)$, 且相互独立. 记

$$M = \max\{X_1, X_2, X_3\}, \quad N = \min\{X_1, X_2, X_3\}.$$

(1) 由 (3.5.15) 式, 所求概率为

$$\begin{aligned}
 P(M \leq 1210) &= F_M(1210) = F_{\max}(1210) = [F(1210)]^3 \\
 &= \left[\Phi\left(\frac{1210 - 1195}{15}\right) \right]^3 = [\Phi(1)]^3 = 0.5955.
 \end{aligned}$$

(2) 所求概率为

$$\begin{aligned}
 P(N \geq 1210) &= 1 - P(N < 1210) = 1 - F_N(1210) \\
 &= 1 - F_{\min}(1210),
 \end{aligned}$$

由(3.5.16)式得到

$$F_{\min}(1210) = 1 - [1 - F(1210)]^3.$$

这里 $F(x)$ 是 $X_i (i=1, 2, 3)$ 的分布函数, 从而

$$F(1210) = \Phi\left(\frac{1210 - 1195}{15}\right) = \Phi(1),$$

于是

$$P(N \geq 1210) = [1 - \Phi(1)]^3 = 0.004.$$

§ 3.6 用 MATLAB 画二维分布图

对于由(3.1.10)式给出的二维正态分布函数, 可以用下面的一段 MATLAB 命令画出其图形.

例 3.6.1 画二维正态分布图.

解 执行下述命令:

```

u1 = 0.5; ↓
u2 = 0.6; ↓
g1 = 8; ↓
g2 = 6; ↓
p = 0.1; ↓
[x,y] = meshgrid(-15:0.6:15); ↓
f = 1/(2 * pi * g1 * g2 * sqrt(1 - p ^ 2)) * ... ↓
    exp(-1/2/(1 - p ^ 2) * ... ↓
        ((x - u1) . ^ 2 / g1 ^ 2 ... ↓
        - 2 * p * (x - u1) . * (y - u2) / g1 / g2 ... ↓
        + (y - u2) . ^ 2 / g2 ^ 2)); ↓
mesh(f) ↓

```

得到分布图形如图 3.6.1 所示.

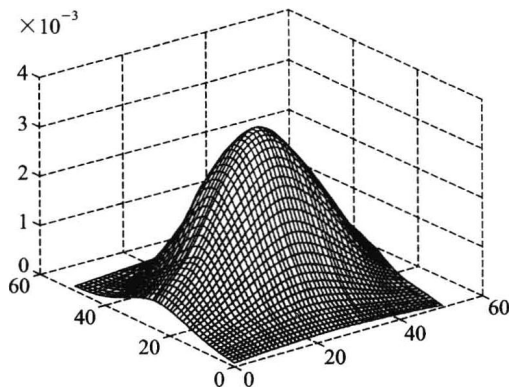


图 3.6.1

习题 3

1. 设二维随机变量 (X, Y) 的分布函数为

$$F(x, y) = \begin{cases} 1 - 2^{-x} - 2^{-y} + 2^{-x-y}, & x \geq 0, y \geq 0, \\ 0, & \text{其它.} \end{cases}$$

求 $P(1 < X \leq 2, 3 < Y \leq 5)$.

2. 设二维随机变量 (X, Y) 所有可能取的值为 $(0, 2)$, $(1.5, 1)$, $(1.5, 2)$, $(2, -1)$, 且取这四个值的概率依次为 $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{8}$, 求 (X, Y) 的概率分布.

3. 从 1, 2, 3, 4 四个整数中随机地取一个数, 并记所取的数为随机变量 X , 再从 1 至 X 中随机地取一个, 记所取的数为随机变量 Y . 求 (X, Y) 的联合概率分布及 X 和 Y 的边缘分布.

4. 袋中装有 3 个黑色球, 2 个白色球, 从中任取 4 个球, 用 X 表示取到的黑色球的个数, 用 Y 表示取到的白色球的个数, 求 (X, Y) 的概率分布.

5. 设二维随机变量 (X, Y) 的概率密度为

$$f(x, y) = \begin{cases} Ae^{-(3x+4y)}, & x > 0, y > 0, \\ 0, & \text{其它.} \end{cases}$$

- (1) 求常数 A ;
- (2) (X, Y) 的联合分布函数;
- (3) $P(0 < X \leq 1, 0 < Y \leq 2)$.

6. 设随机变量 (X, Y) 的分布函数为

$$F(x, y) = A \left(B + \arctan \frac{x}{2} \right) \left(C + \arctan \frac{y}{3} \right),$$

求常数 A, B, C 及 (X, Y) 的联合密度函数.

7. 设随机变量 (X, Y) 的概率密度函数为

$$f(x, y) = \begin{cases} e^{-y}, & 0 < x < y, \\ 0, & \text{其它}, \end{cases}$$

试求 $P(X+Y \leq 1)$.

8. 设随机变量 (X, Y) 的概率密度为

$$f(x, y) = \begin{cases} x^2 + \frac{xy}{3}, & 0 \leq x \leq 1, 0 \leq y \leq 2, \\ 0, & \text{其它}. \end{cases}$$

求 X 和 Y 的边缘密度函数.

9. 设随机变量 (X, Y) 在平面区域 D 上服从均匀分布, 其中 D 是由直线 $y=x$ 和曲线 $y=x^2$ 所围成的平面区域, 求 (X, Y) 的边缘概率密度函数.

10*. 设随机变量 (X, Y) 的概率密度函数为

$$f(x, y) = \begin{cases} 1, & |y| < x, 0 < x < 1, \\ 0, & \text{其它}. \end{cases}$$

求条件概率密度函数 $f_{X|Y}(x|y)$ 和 $f_{Y|X}(y|x)$.

11*. 设随机变量 (X, Y) 的概率密度为

$$f(x, y) = \begin{cases} \frac{21}{4}x^2y, & x^2 \leq y \leq 1, \\ 0, & \text{其它}. \end{cases}$$

求条件密度 $f_{Y|X}(y|x)$ 以及概率 $P\left(Y \geq \frac{3}{4} \mid X = \frac{1}{2}\right)$.

12. 设随机变量 (X, Y) 具有分布函数

$$F(x, y) = \begin{cases} (1 - e^{-\alpha x})y, & x \geq 0, 0 \leq y \leq 1, \\ 1 - e^{-\alpha x}, & x \geq 0, y > 1, (\alpha > 0) \\ 0, & \text{其它}, \end{cases}$$

问 X 与 Y 是否相互独立?

13. 设随机变量 (X, Y) 的概率密度函数为

$$f(x, y) = \begin{cases} Axy^2, & 0 < x < 1, 0 < y < 1, \\ 0, & \text{其它}. \end{cases}$$

求: (1) 常数 A ; (2) 证明 X 与 Y 相互独立.

14. 已知随机变量 (X, Y) 的概率分布为

X \ Y	Y		
	1	2	3
1	$\frac{1}{6}$	$\frac{1}{9}$	$\frac{1}{18}$
2	$\frac{1}{3}$	$\frac{1}{a}$	$\frac{1}{b}$

试问: 当 a, b 为何值时, X 与 Y 相互独立? 并求 $P(X=i|Y=1)$.

15. 设随机变量 X 与 Y 相互独立, 并且它们的密度函数分别为

$$f_1(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & x \leq 0, \end{cases} \quad f_2(y) = \begin{cases} e^{-y}, & y > 0, \\ 0, & y \leq 0. \end{cases}$$

试求: (1) (X, Y) 的概率密度函数; (2) $P(X \leq 1 | Y > 0)$.

16. 已知随机变量 (X, Y) 的分布律为

$\begin{matrix} X \\ Y \end{matrix}$	1	2	3
1	$\frac{1}{5}$	0	$\frac{1}{5}$
2	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

求 $Z_1 = X + Y$, $Z_2 = \max(X, Y)$ 的分布律.

17. 设随机变量 (X, Y) 的概率密度为

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

求: (1) $P(X > 2Y)$; (2) $Z = X + Y$ 的概率密度 $f_Z(z)$.

18. 设随机变量 X 与 Y 相互独立, 其密度函数分别为

$$f_X(x) = \begin{cases} \frac{1}{2}e^{-\frac{x}{2}}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

$$f_Y(y) = \begin{cases} \frac{1}{3}e^{-\frac{y}{3}}, & y \geq 0, \\ 0, & y < 0, \end{cases}$$

求随机变量 $Z = X + Y$ 的概率密度函数.

19. 设随机变量 X, Y 相互独立, 其密度函数分别为

$$f_X(x) = \begin{cases} 1, & 0 \leq x \leq 1, \\ 0, & \text{其它,} \end{cases}$$

$$f_Y(y) = \begin{cases} e^{-y}, & y > 0, \\ 0, & y \leq 0, \end{cases}$$

求随机变量 $Z = X + Y$ 的密度函数.

20. 设随机变量 X 与 Y 相互独立, 分别服从参数为 λ_1 与 λ_2 的指数分布, 求 $Z = \frac{X}{Y}$ 的概率密度.

21. 设系统 L 是由两个相互独立的子系统 L_1, L_2 连接而成, 并且已知 L_1, L_2 的寿命分别为 X, Y , 它们的概率密度函数分别为

$$f_X(x) = \begin{cases} \alpha e^{-\alpha x}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

$$f_Y(y) = \begin{cases} \beta e^{-\beta y}, & y \geq 0, \\ 0, & y < 0, \end{cases}$$

其中参数 $\alpha > 0, \beta > 0$. 求系统 L 在串联、并联方式下寿命 Z 的概率密度函数.

22*. 设随机变量 X 在区间 $(0,1)$ 上服从均匀分布, 在 $X=x(0 < x < 1)$ 的条件下, 随机变量 Y 在区间 $(0,x)$ 上服从均匀分布, 求:

(1) (X,Y) 的联合概率密度 $f(x,y)$;

(2) Y 的概率密度;

(3) 概率 $P(X+Y > 1)$.

23*. 设随机变量 X_1, X_2, X_3, X_4 相互独立且同分布, $P(X_i = 0) = 0.6, P(X_i = 1) = 0.4 (i=1,2,3,4)$, 求行列式 $X = \begin{vmatrix} X_1 & X_2 \\ X_3 & X_4 \end{vmatrix}$ 的概率分布.

24*. 设随机变量 X 与 Y 独立同分布, 且 X 的概率分布为

X	1	2
P	$\frac{2}{3}$	$\frac{1}{3}$

记 $U = \max(X,Y), V = \min(X,Y)$, 求 (U,V) 的概率分布.

25*. 设随机变量 (X,Y) 的概率密度为

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

求: (1) (X,Y) 的边缘概率密度 $f_X(x), f_Y(y)$;

(2) $Z = 2X - Y$ 的概率密度 $f_Z(z)$;

(3) $P\left(Y \leq \frac{1}{2} \mid X \leq \frac{1}{2}\right)$.

26*. 设随机变量 X 与 Y 相互独立, 其中 X 的概率分布为

X	1	2
P	0.3	0.7

而 Y 的概率密度为 $f(y)$, 求随机变量 $U = X + Y$ 的概率密度 $g(u)$.

第4章 随机变量的数字特征

随机变量的概率分布、概率密度函数与分布函数虽能完整地描述随机变量的取值及其对应的取值概率，但在某些实际问题中，有时不容易确定随机变量的分布；有时不需要知道随机变量分布规律的全貌，而只需知道它的某些特征就够了。比如，我们把一个班的每个同学“高等数学”的考试成绩都逐一的罗列起来，虽然已很全面、完整，但这个班级考试成绩的整体特征，如成绩的平均水平，成绩好坏的分散程度，还不够鲜明。就是说，除随机变量的分布律、分布函数外，我们还要对反映随机变量分布状况的数量指标——平均水平和参差程度，给一个定量的表征。刻画随机变量分布状况的某些特征的数量指标称为随机变量的数字特征，常用的数字特征有数学期望、方差等。

§4.1 数 学 期 望

1. 数学期望的定义

对于随机变量 X ，我们希望能找到这样一个数值，它能体现 X 取值的“平均”大小，例如：有一位射手，射击时击中的环数用随机变量 X 表示，假定分布列为：

X	8	9	10
P	0.2	0.2	0.6

求该射手的平均每枪击中的环数。

显然作为 X 可能值的平均数是： $\frac{1}{3}(8+9+10)=9$ ，但是数值 9 并不真正体现 X 的取值的平均，这是视 X 等可能取值“8”、“9”、“10”的结果。实际上，从分布列中可以看出， X 取“10”的可能性远比 X 取“8”、“9”的可能性要大。设想该射手射击 100 次，那么他大约有 60 次击中 10 环，20 次击中 9 环，另 20 次击中 8 环。一般地，让射手射击 n 次，则他击中的环数大约是：

$$8 \times 0.2n + 9 \times 0.2n + 10 \times 0.6n = 9.4n,$$

平均起来该射手每枪击中 9.4 环。虽然 9.4 才是 X 取值的“平均值”。所以在研究随机变量取值的平均值时，不仅要考虑它的各个取值，还应同时考虑到取

各个值相对应的概率.

定义 4.1.1 设离散型随机变量 X 的概率分布为

$$P(X = x_k) = p_k, \quad k = 1, 2, \dots.$$

若级数 $\sum_{k=1}^{+\infty} x_k p_k$ 绝对收敛, 则称级数 $\sum_{k=1}^{+\infty} x_k p_k$ 的和为离散型随机变量的数学期望, 或均值, 记为 $E(X)$. 即

$$E(X) = \sum_{k=1}^{+\infty} x_k p_k. \quad (4.1.1)$$

对于连续型随机变量, 若它的密度函数为 $f(x)$, 注意到 $f(x)$ 的作用与离散型随机变量 p_k 相当, 有以下定义.

定义 4.1.2 设连续型随机变量 X 的概率密度函数为 $f(x)$, 若积分 $\int_{-\infty}^{+\infty} x f(x) dx$ 绝对收敛, 则称该积分值为 X 的数学期望, 或均值, 并记为

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx. \quad (4.1.2)$$

例 4.1.1 求服从两点分布的随机变量 X 的数学期望.

解 设两点分布的分布列为

X	0	1
P	q	p

由定义 4.1.1 得 $E(X) = 0 \times q + 1 \times p = p$.

例 4.1.2 为了普查某种疾病, 需要对 N 个人抽血检验. 检验方式可以有两种: (1) 对每个人的血样进行检验, 逐一断定是否呈阳性反应, 需要检验 N 次; (2) 把 k 个人分为一组, 同一组的 k 个人的血样混在一起进行检验. 如果混合血样呈阴性反应, 则表明这 k 个人的血都呈阴性, 此时这 k 个人平均每人只需检验 $\frac{1}{k}$ 次; 如果混合血样呈阳性, 再对这 k 个人的血样逐个检验, 共需检验 $k+1$ 次, 平均每人需检验 $\frac{k+1}{k}$ 次. 假定对每个人来说化验呈阳性反应的概率都是 p (p 较小), 而且各人呈阳性反应是相互独立的. 试说明第二种检验办法能减少化验的次数.

解 按第二种检验办法, 每个人的血样需检验的次数 X 是随机变量, 其可能取值为 $\frac{1}{k}$ 或 $\frac{k+1}{k}$. 若记 $q = 1 - p$, 表示每个人血样呈阴性反应的概率, 故 k 个人混合成的血呈阴性的概率是 q^k , 呈阳性的概率是 $1 - q^k$. 于是, X 的分布列为

X	$\frac{1}{k}$	$\frac{k+1}{k}$
P	q^k	$1-q^k$

X 的数学期望(每个人平均需化验的次数)为

$$E(X) = \frac{1}{k} \times q^k + \left(1 + \frac{1}{k}\right)(1 - q^k) = 1 - q^k + \frac{1}{k}.$$

由于 p 一般很小, 从而 q 接近于 1. 当 $k \geq 2$ 时, $q^k > \frac{1}{k}$, 故 $E(X) < 1$, 当 p 已知时, 取适当的 k , 使 $E(X)$ 达到最小值. 例如, 当 $N = 1\,000$, $p = 0.1$, $q = 0.9$ 时, 取 $k = 4$, $E(X)$ 达到最小, 此时平均化验的次数为

$$1\,000 \times \left(1 - 0.9^4 + \frac{1}{4}\right) \approx 594 (\text{次}),$$

减少约 40% 的工作量.

例 4.1.3 设随机变量 $X \sim B(n, p)$, 求 $E(X)$.

解 因为

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

由离散型随机变量数学期望的定义, 则

$$\begin{aligned} E(X) &= \sum_{k=0}^n k C_n^k p^k (1-p)^{n-k} \\ &= \sum_{k=1}^n \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^n \frac{(n-1)!}{(k-1)![(n-1)-(k-1)]!} p^{k-1} (1-p)^{(n-1)-(k-1)}, \end{aligned}$$

令 $i = k - 1$, 则

$$\begin{aligned} &\sum_{k=1}^n \frac{(n-1)!}{(k-1)![(n-1)-(k-1)]!} p^{k-1} (1-p)^{(n-1)-(k-1)} \\ &= \sum_{i=0}^{n-1} C_{n-1}^i p^i (1-p)^{(n-1)-i} = [p + (1-p)]^{n-1} = 1, \end{aligned}$$

从而

$$E(X) = np.$$

例 4.1.4 设随机变量 $X \sim P(\lambda)$, 求 $E(X)$.

解 因为 $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$, $k = 0, 1, 2, \dots$, 所以

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\ &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda. \end{aligned}$$

例 4.1.5 设随机变量 X 服从 $[a, b]$ 上的均匀分布, 求 $E(X)$.

解 由于均匀分布的密度函数为

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

则

$$E(X) = \int_a^b \frac{x}{b-a} dx = \frac{1}{2(b-a)} x^2 \Big|_a^b = \frac{a+b}{2}.$$

可见, 服从均匀分布的随机变量 X 的数学期望就是该区间的中点.

例 4.1.6 设随机变量 X 服从参数为 λ 的指数分布, 试求 $E(X)$.

解 已知 X 的密度函数为

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{当 } x \geq 0 \text{ 时}, \\ 0, & \text{当 } x < 0 \text{ 时}. \end{cases}$$

故

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} x f(x) dx = \lambda \int_0^{+\infty} x e^{-\lambda x} dx \\ &= -x e^{-\lambda x} \Big|_0^{+\infty} + \int_0^{+\infty} e^{-\lambda x} dx = \frac{1}{\lambda}. \end{aligned}$$

例 4.1.7 设 $X \sim N(\mu, \sigma^2)$, 求 $E(X)$.

解 因为随机变量 X 的密度函数为

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

所以

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \stackrel{\frac{x-\mu}{\sigma} = t}{=} \int_{-\infty}^{+\infty} (\mu + \sigma t) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2}} \sigma dt \\ &= \mu \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt + \sigma \int_{-\infty}^{+\infty} \frac{t}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt, \end{aligned}$$

故

$$E(X) = \mu.$$

2. 随机变量函数的数学期望.

设 X 为一维随机变量, 下面讨论 X 的函数 $Y = g(X)$ 的数学期望.

定理 4.1.1 设 $y = g(x)$ 是连续函数, Y 是随机变量 X 的函数, $Y = g(X)$.

(1) 若 X 是离散型随机变量, 其概率分布为

$$P(X = x_k) = p_k, \quad k = 1, 2, \dots,$$

且级数 $\sum_k g(x_k) p_k$ 绝对收敛, 则有

$$E(Y) = E[g(X)] = \sum_k g(x_k) p_k. \quad (4.1.3)$$

(2) 若 X 是连续型随机变量, 其概率密度函数为 $f(x)$, 且积分 $\int_{-\infty}^{+\infty} g(x)f(x)dx$ 绝对收敛, 则有

$$E(Y) = E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx. \quad (4.1.4)$$

定理 4.1.1 提供了一种直接由随机变量 X 的分布, 求其函数 $g(X)$ 的数学期望的方法, 而不需先求 $Y=g(X)$ 的分布. 同时, 定理 4.1.1 的结论可以推广到 n 维随机变量的情形. 例如, 设 (X,Y) 为二维随机变量, $Z=g(X,Y)$, 对离散型情况, 设 (X,Y) 的联合分布律为

$$P(X=x_i, Y=y_j) = p_{ij}, \quad i=1, 2, \dots; j=1, 2, \dots.$$

则

$$E(Z) = E[g(X,Y)] = \sum_i \sum_j g(x_i, y_j) p_{ij}. \quad (4.1.5)$$

对连续型情况, 设 (X,Y) 的联合密度函数为 $f(x,y)$, 则

$$E(Z) = E[g(X,Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y)f(x,y)dx dy. \quad (4.1.6)$$

例 4.1.8 设随机变量 $X \sim B(n, p)$, 求 $E(X^2)$.

解 因为

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

由 (4.1.3) 式, 则有

$$\begin{aligned} E(X^2) &= \sum_{k=0}^n k^2 C_n^k p^k (1-p)^{n-k} \\ &= \sum_{k=0}^n [k(k-1) + k] C_n^k p^k (1-p)^{n-k} \\ &= \sum_{k=0}^n k(k-1) \frac{n!}{k! (n-k)!} p^k (1-p)^{n-k} + \sum_{k=0}^n k C_n^k p^k (1-p)^{n-k} \\ &= \sum_{k=2}^n \frac{n!}{(k-2)! (n-k)!} p^k (1-p)^{n-k} + E(X) \\ &= n(n-1)p^2 \sum_{k=2}^n \frac{(n-2)!}{(k-2)! [(n-2)-(k-2)]!} p^{k-2} (1-p)^{(n-2)-(k-2)} + np \\ &= n(n-1)p^2 [p + (1-p)]^{n-2} + np \\ &= n(n-1)p^2 + np. \end{aligned}$$

例 4.1.9 设随机变量 X 服从 $[0, 2\pi]$ 上的均匀分布, $Y = \sin X$, 求 $E(Y)$.

解 随机变量 X 的密度函数为

$$f(x) = \begin{cases} \frac{1}{2\pi}, & 0 \leq x \leq 2\pi, \\ 0, & \text{其它}, \end{cases}$$

由(4.1.4)式得到

$$E(Y) = \int_{-\infty}^{+\infty} \sin x f(x) dx = \int_0^{2\pi} \frac{1}{2\pi} \sin x dx = 0.$$

例 4.1.10 已知随机变量 (X, Y) 的分布律为

$X \backslash Y$	0	1
0	$\frac{1}{3}$	0
1	$\frac{1}{2}$	$\frac{1}{6}$

求 $E(2X+3Y)$, $E(XY)$.

解 由(4.1.5)式得

$$\begin{aligned} E(2X+3Y) &= \sum_i \sum_j (2x_i + 3y_j) p_{ij} \\ &= (2 \times 0 + 3 \times 0) \times \frac{1}{3} + (2 \times 0 + 3 \times 1) \times 0 + \\ &\quad (2 \times 1 + 3 \times 0) \times \frac{1}{2} + (2 \times 1 + 3 \times 1) \times \frac{1}{6} \\ &= \frac{11}{6}, \end{aligned}$$

$$\begin{aligned} E(XY) &= \sum_i \sum_j (x_i y_j) p_{ij} \\ &= (0 \times 0) \times \frac{1}{3} + (0 \times 1) \times 0 + (1 \times 0) \times \frac{1}{2} + (1 \times 1) \times \frac{1}{6} \\ &= \frac{1}{6}. \end{aligned}$$

例 4.1.11 设随机变量 (X, Y) 的概率密度为

$$f(x, y) = \begin{cases} 8xy, & 0 \leq x \leq 1, 0 \leq y \leq x, \\ 0, & \text{其它.} \end{cases}$$

求 $E(X)$, $E(Y)$, $E(XY)$.

解 由(4.1.6)式得

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xf(x, y) dx dy \\ &= \int_0^1 dx \int_0^x x \cdot 8xy dy = \frac{4}{5}, \end{aligned}$$

$$\begin{aligned} E(Y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} yf(x, y) dx dy \\ &= \int_0^1 dx \int_0^x y \cdot 8xy dy = \frac{8}{15}, \end{aligned}$$

$$\begin{aligned}
 E(XY) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xyf(x, y) dx dy \\
 &= \int_0^1 dx \int_0^x xy \cdot 8xy dy = \frac{4}{9}.
 \end{aligned}$$

例 4.1.12 若市场对某种产品的需求量为随机变量 X , 且 X 服从 $U[200, 400]$, 设商场每销售一单位的产品, 可赚三百元, 但若销不出去, 则每一单位产品需付保管费一百元. 问应组织多少货源, 该商场才能使收益的期望值最大?

解 设商场组织 n 单位货源, 收益为 Y (百元), 由题意得

$$Y = g(X) = \begin{cases} 3n, & X \geq n, \\ 3X - (n - X), & X < n, \end{cases}$$

即

$$g(x) = \begin{cases} 3n, & x \geq n, \\ 4x - n, & x < n, \end{cases}$$

而 X 的密度函数为

$$f(x) = \begin{cases} \frac{1}{200}, & 200 \leq x \leq 400, \\ 0, & \text{其它}, \end{cases}$$

从而 Y 的数学期望为

$$\begin{aligned}
 E(Y) &= \int_{-\infty}^{+\infty} g(x)f(x) dx = \frac{1}{200} \left[\int_{200}^n (4x - n) dx + \int_n^{400} 3n dx \right] \\
 &= \frac{1}{200} (-2n^2 + 1400n - 2 \cdot 200^2),
 \end{aligned}$$

令

$$L(n) = E(Y) = \frac{1}{100} (-n^2 + 700n - 200^2),$$

由 $L'(n) = 0$, 解得 $n = 350$. 故该商场应组织 350 单位货源才能使收益的期望值最大.

3. 数学期望的性质

随机变量的数学期望有下列性质:

- (1) 常量 C 的数学期望等于常量 C , 即 $E(C) = C$.
- (2) 如果 C 是常量, X 是任一随机变量, 则 $E(CX) = CE(X)$.
- (3) 设 X, Y 是任意两个随机变量, 则

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

这一性质可推广到有限个的情形. 即

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n).$$

(4) 若随机变量 X 与 Y 相互独立, 则有

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

这一性质也可推广到有限个的情形. 即 $E(X_1 X_2 \cdots X_n) = E(X_1)E(X_2) \cdots E(X_n)$, 其中 $X_1, X_2, \cdots, X_n$ 为相互独立随机变量.

(5) 设 X, Y 为两个随机变量, 若 $E(X^2), E(Y^2)$ 存在, 则

$$[E(XY)]^2 \leq E(X^2) \cdot E(Y^2).$$

证明 (1) 常数 C 作为随机变量 X 而言, 是一个离散型的, 它只可能取到 C 值, 故 $P(X=C)=1$, 于是 $E(C)=C \times 1=C$.

(2) 仅以连续型随机变量为例, 设 X 的密度函数为 $f(x)$, 则

$$E(CX) = \int_{-\infty}^{+\infty} Cxf(x)dx = C \int_{-\infty}^{+\infty} xf(x)dx = CE(X).$$

(3) 设二维随机变量 (X, Y) 的联合密度函数为 $f(x, y)$, 关于 X 和 Y 的边缘密度函数分别为 $f_X(x), f_Y(y)$, 由 (4.1.6) 式则有

$$\begin{aligned} E(X+Y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x+y)f(x,y)dx dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xf(x,y)dx dy + \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} yf(x,y)dx dy \\ &= \int_{-\infty}^{+\infty} xdx \int_{-\infty}^{+\infty} f(x,y)dy + \int_{-\infty}^{+\infty} ydy \int_{-\infty}^{+\infty} f(x,y)dx \\ &= \int_{-\infty}^{+\infty} xf_X(x)dx + \int_{-\infty}^{+\infty} yf_Y(y)dy \\ &= E(X) + E(Y). \end{aligned}$$

对离散型随机变量类似可证.

(4) 同样仅以连续型为例, 由于 X 与 Y 相互独立, 所以

$$f(x,y) = f_X(x)f_Y(y).$$

于是

$$\begin{aligned} E(XY) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xyf(x,y)dx dy \\ &= \int_{-\infty}^{+\infty} xf_X(x)dx \int_{-\infty}^{+\infty} yf_Y(y)dy \\ &= E(X)E(Y). \end{aligned}$$

(5) 考虑实变量 t 的二次函数

$$g(t) = E(tX - Y)^2 = t^2 E(X^2) - 2tE(XY) + E(Y^2)$$

对一切 t , 容易看出 $g(t) \geq 0$, 故二次方程 $g(t) = 0$ 的判别式

$$\Delta = 4[E(XY)]^2 - 4E(X^2) \cdot E(Y^2) \leq 0,$$

即

$$[E(XY)]^2 \leq E(X^2) \cdot E(Y^2).$$

该不等式称为柯西 - 施瓦茨 (Cauchy-Schwarz) 不等式.

例 4.1.13 已知随机变量 $X \sim N(5, 10^2)$, $Y = 3X + 5$, 求 $E(Y)$.

解 因为 $X \sim N(5, 10^2)$, 故 $E(X) = 5$, 则由数学期望的性质可得

$$E(Y) = E(3X + 5) = 3E(X) + 5 = 3 \times 5 + 5 = 20.$$

例 4.1.14 设 X 服从超几何分布, 其分布列为

$$P(X = k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}, \quad k = 0, 1, \dots, n, \quad k \leq M,$$

求其数学期望 $E(X)$.

解 超几何分布描述的是不放回抽样情形, 以 X_i 表示第 i 次抽样的结果, 令

$$X_i = \begin{cases} 1, & \text{第 } i \text{ 次抽得正品,} \\ 0, & \text{第 } i \text{ 次抽得次品,} \end{cases}$$

则

$$P(X_i = 1) = \frac{M}{N}, \quad P(X_i = 0) = 1 - \frac{M}{N}, \quad E(X_i) = \frac{M}{N},$$

而 $X = X_1 + X_2 + \dots + X_n$, 利用性质(3)得到

$$E(X) = E(X_1) + E(X_2) + \dots + E(X_n) = \frac{nM}{N}.$$

§ 4.2 方 差

1. 方差的定义

数学期望体现了随机变量取值的平均水平, 它是随机变量的重要数字特征. 但仅知道均值是不够的, 还需要知道随机变量取值的波动程度, 即随机变量所取的值与它的数学期望的偏离程度. 例如, 检查一批日光灯的寿命, 如果波动大, 说明生产不够稳定; 生物的某种特征(血压、白细胞数)波动大, 表示该生物处于病态. 所以, 研究随机变量与其平均值的偏离程度也十分重要. 设随机变量 X 的数学期望为 $E(X)$. 偏离量 $X - E(X)$ 也是随机的, 由于 $E(X - E(X)) = 0$, 无法刻画随机变量的偏离程度, 通常用 $[X - E(X)]^2$ 的数学期望来描述 X 与 $E(X)$ 的偏离程度.

定义 4.2.1 设 X 为随机变量, 如果 $E[X - E(X)]^2$ 存在, 则称之为 X 的方差, 记为 $D(X)$, 即

$$D(X) = E[X - E(X)]^2. \quad (4.2.1)$$

对于离散型随机变量, 若 $P(X = x_k) = p_k$, 则

$$D(X) = \sum_k [x_k - E(X)]^2 p_k. \quad (4.2.2)$$

对于连续型随机变量, 若 X 的密度函数为 $f(x)$, 则

$$D(X) = \int_{-\infty}^{+\infty} [X - E(X)]^2 f(x) dx. \quad (4.2.3)$$

称 $\sqrt{D(X)}$ 为标准差.

在计算方差时, 常用如下重要公式:

$$D(X) = E(X^2) - E^2(X). \quad (4.2.4)$$

这是因为

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

例 4.2.1 分别求服从两点分布、二项分布、泊松分布的随机变量的方差.

解 由本章例 4.1.1 知两点分布的数学期望为 $E(X) = p$, 则

$$E(X^2) = 0^2 \times q + 1^2 \times p = p.$$

由(4.2.4)式得两点分布的方差为

$$D(X) = p - p^2 = pq,$$

由例 4.1.3 及例 4.1.8 知二项分布的数学期望为

$$E(X) = np, \quad E(X^2) = n(n-1)p^2 + np.$$

由(4.2.4)式得二项分布的方差为

$$D(X) = n(n-1)p^2 + np - (np)^2 = np - np^2 = npq.$$

由例 4.1.4 知泊松分布的数学期望为 $E(X) = \lambda$, 而

$$\begin{aligned} E(X^2) &= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=1}^{\infty} [k(k-1) + k] \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \lambda^2 + \lambda. \end{aligned}$$

由(4.2.4)式得泊松分布的方差为

$$D(X) = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

例 4.2.2 设 X 在 $[a, b]$ 上服从均匀分布, 求方差 $D(X)$.

解 由本章例 4.1.5 可知 $E(X) = \frac{1}{2}(a+b)$, 而

$$\begin{aligned} E(X^2) &= \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{3(b-a)} x^3 \Big|_a^b \\ &= \frac{1}{3}(b^2 + ba + a^2), \end{aligned}$$

于是

$$D(X) = \frac{1}{3}(b^2 + ba + a^2) - \frac{1}{4}(a+b)^2 = \frac{1}{12}(b-a)^2.$$

例 4.2.3 设 X 服从参数为 λ 的指数分布, 求 $D(X)$.

解 由本章例 4.1.6 知, $E(X) = \frac{1}{\lambda}$, 而

$$\begin{aligned} E(X^2) &= \lambda \int_0^{+\infty} x^2 e^{-\lambda x} dx \\ &= -x^2 e^{-\lambda x} \Big|_0^{+\infty} + 2 \int_0^{+\infty} x e^{-\lambda x} dx = \frac{2}{\lambda^2}, \end{aligned}$$

故

$$D(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

例 4.2.4 设 $X \sim N(\mu, \sigma^2)$, 求方差 $D(X)$.

解 由定义

$$\begin{aligned} D(X) &= \int_{-\infty}^{+\infty} (x - \mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &\stackrel{t = \frac{x-\mu}{\sigma}}{=} \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} t^2 e^{-\frac{t^2}{2}} dt \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \left[-te^{-\frac{t^2}{2}} \Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt \right] \\ &= \sigma^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt = \sigma^2. \end{aligned}$$

至此, 我们清楚了正态分布 $N(\mu, \sigma^2)$ 的两个参数的意义, 即 μ 是数学期望, σ^2 是方差, σ 为其标准差.

2. 方差的性质

方差具有如下性质:

(1) 常量的方差等于零, 即 $D(C) = 0$.

上述事实直观上也容易理解. 因为方差刻画了随机变量取值围绕其均值的波动情况, 作为常数的随机变量, 其波动为零. 故常数的方差等于零.

(2) 设 C 为常量, $D(CX) = C^2 D(X)$.

(3) 设 X, Y 是相互独立的随机变量, 则

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

这一性质可以推广: 有限多个相互独立的随机变量代数和的方差, 等于它们方差之和.

例 4.2.5 设随机变量 X 和 Y 相互独立, 且 X 服从二项分布 $B(16, 0.5)$, Y 服从泊松分布 $P(9)$, 求 $D(X - 2Y + 1)$.

解 由例 4.2.1 可知

$$D(X) = npq = 16 \times 0.5 \times 0.5 = 4, \quad D(Y) = \lambda = 9.$$

根据方差的性质,

$$\begin{aligned} D(X - 2Y + 1) &= D(X) + D(-2Y) + D(1) \\ &= D(X) + 4D(Y) \\ &= 4 + 4 \times 9 = 40. \end{aligned}$$

§ 4.3 协方差与相关系数

1. 协方差与相关系数的概念

在本章前二节中, 我们主要讨论了一维随机变量的数字特征. 对于二维随机变量 (X, Y) , 除了研究 X, Y 它们各自的数学期望和方差以外, 还需要研究描述这两个随机变量之间相互关系的数字特征. 协方差和相关系数就是用来描述 X 与 Y 之间相互关系的数字特征.

定义 4.3.1 设二维随机变量 (X, Y) , 若

$$E\{[X - E(X)][Y - E(Y)]\}$$

存在, 则称它为随机变量 X 与 Y 的协方差, 记作 $\text{Cov}(X, Y)$, 或 σ_{XY} , 即

$$\text{Cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\}. \quad (4.3.1)$$

又若 $D(X) \neq 0, D(Y) \neq 0$, 则称

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)} \sqrt{D(Y)}} \quad (4.3.2)$$

为随机变量 X 与 Y 的相关系数.

如果记 $\sigma_X = \sqrt{D(X)}, \sigma_Y = \sqrt{D(Y)}$, 则相关系数又可写成

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

当 (X, Y) 分别是离散型和连续型随机变量时, 协方差的计算公式分别为:

$$\text{Cov}(X, Y) = \sum_i \sum_j [x_i - E(X)][y_j - E(Y)] p_{ij}, \quad (4.3.3)$$

$$\text{Cov}(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [x - E(X)][y - E(Y)] f(x, y) dx dy. \quad (4.3.4)$$

计算协方差时, 还常用公式

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y). \quad (4.3.5)$$

例 4.3.1 设随机变量 (X, Y) 的分布律为

$\begin{array}{c} Y \\ \diagdown \\ X \end{array}$	0	1
0	$\frac{1}{4}$	0
1	$\frac{1}{4}$	$\frac{1}{2}$

求 $\text{Cov}(X, Y)$, ρ_{XY} .

解

$$E(X) = 0 \times \frac{1}{4} + 1 \times \frac{3}{4} = \frac{3}{4},$$

$$E(X^2) = 0^2 \times \frac{1}{4} + 1^2 \times \frac{3}{4} = \frac{3}{4},$$

则
$$D(X) = E(X^2) - [E(X)]^2 = \frac{3}{4} - \left(\frac{3}{4}\right)^2 = \frac{3}{16}.$$

同理 $E(Y) = \frac{1}{2}$, $D(Y) = \frac{1}{4}$. 又

$$E(XY) = 1 \times 1 \times \frac{1}{2} = \frac{1}{2},$$

故

$$\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y) = \frac{1}{2} - \frac{3}{4} \times \frac{1}{2} = \frac{1}{8}.$$

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)} \sqrt{D(Y)}} = \frac{\frac{1}{8}}{\sqrt{\frac{3}{4}} \times \frac{1}{2}} = \frac{\sqrt{3}}{3}.$$

例 4.3.2 求例 4.1.11 中随机变量 (X, Y) 的 $\text{Cov}(X, Y)$, ρ_{XY} .

解 由例 4.1.11 可知

$$E(X) = \frac{4}{5}, \quad E(Y) = \frac{8}{15}, \quad E(XY) = \frac{4}{9}.$$

故

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{4}{9} - \frac{4}{5} \times \frac{8}{15} = \frac{4}{225}.$$

又

$$E(X^2) = \int_0^1 dx \int_0^x x^2 \cdot 8xy dy = \frac{2}{3},$$

则

$$D(X) = E(X^2) - [E(X)]^2 = \frac{2}{3} - \left(\frac{4}{5}\right)^2 = \frac{2}{75}.$$

同理 $D(Y) = \frac{11}{225}$. 所以

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)} \sqrt{D(Y)}} = \frac{\frac{4}{225}}{\sqrt{\frac{2}{75}} \cdot \sqrt{\frac{11}{225}}} = \frac{2}{33} \sqrt{\frac{66}{33}}.$$

例 4.3.3 设 (X, Y) 服从二维正态分布 $N(\mu_1, \sigma_1^2; \mu_2, \sigma_2^2; \rho)$, 求 ρ_{XY} .

解 由例 3.2.3 知, $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, 故 $E(X) = \mu_1$, $D(X) = \sigma_1^2$, $E(Y) = \mu_2$, $D(Y) = \sigma_2^2$. 令

$$u = \frac{x - \mu_1}{\sigma_1}, \quad v = \frac{y - \mu_2}{\sigma_2},$$

则由二维正态分布的概率密度函数及 (4.3.4) 式有

$$\begin{aligned} \text{Cov}(X, Y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_1)(y - \mu_2) f(x, y) dx dy \\ &= \frac{\sigma_1 \sigma_2}{2\pi \sqrt{1 - \rho^2}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} uv \exp \left[-\frac{1}{2(1 - \rho^2)} (u^2 - 2\rho uv + v^2) \right] du dv \\ &= \frac{\sigma_1 \sigma_2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} v dv \frac{1}{\sqrt{2\pi} \sqrt{1 - \rho^2}} \int_{-\infty}^{+\infty} u \exp \cdot \left\{ -\frac{1}{2(1 - \rho^2)} \right. \\ &\quad \left. [(u - \rho v)^2 + (1 - \rho^2)v^2] \right\} du \\ &= \frac{\sigma_1 \sigma_2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} v e^{-\frac{v^2}{2}} dv \left\{ \frac{1}{\sqrt{2\pi} \sqrt{1 - \rho^2}} \int_{-\infty}^{+\infty} u \exp \left[-\frac{1}{2(1 - \rho^2)} (u - \rho v)^2 \right] du \right\}. \end{aligned}$$

上式大括号中的积分恰好是服从正态分布 $N(\rho v, 1 - \rho^2)$ 的随机变量的数学期望, 故

$$\text{Cov}(X, Y) = \frac{\sigma_1 \sigma_2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \rho v^2 e^{-\frac{v^2}{2}} dv = \rho \sigma_1 \sigma_2.$$

于是

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)} \sqrt{D(Y)}} = \frac{\rho \sigma_1 \sigma_2}{\sigma_1 \sigma_2} = \rho.$$

因而, 二维正态分布 (X, Y) 的分布中所含的参数 ρ 就是 X 和 Y 的相关系数.

2. 协方差与相关系数的性质

(1) $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.

(2) $\text{Cov}(a_1 X + b_1, a_2 Y + b_2) = a_1 a_2 \text{Cov}(X, Y)$, 其中 a_1, a_2, b_1, b_2 为常数.

(3) $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$.

(4) $D(X \pm Y) = D(X) + D(Y) \pm 2\text{Cov}(X, Y)$, 并且当 X 与 Y 相互独立时, $\text{Cov}(X, Y) = 0$.

$$(5) |\rho_{XY}| \leq 1.$$

(6) $|\rho_{XY}| = 1$ 的充分必要条件是 $P(Y = aX + b) = 1$. 其中 a, b 为常数, 且 $a \neq 0$.

性质(1)~(4)容易依据协方差的定义推得, 性质(5)可以利用柯西-施瓦茨不等式推得. 性质(5), (6)刻画了 X 与 Y 之间的线性关系的程度. 一般地, 当 $|\rho_{XY}|$ 较大时, X 与 Y 的线性关系程度较好; 反之, X 与 Y 的线性关系程度较差.

当 $\rho_{XY} = 0$ 时, 称 X 与 Y 不相关.

若 X 与 Y 相互独立, 则 $\text{Cov}(X, Y) = 0$, 于是 $\rho_{XY} = 0$, 即 X 与 Y 不相关. 反之, 当 X 与 Y 不相关时, 并不能保证 X 与 Y 相互独立.

例 4.3.4 设二维随机变量 (X, Y) , 若 X 服从 $[-1, 1]$ 上的均匀分布, 而 $Y = X^2$, 问 X, Y 是否不相关? 是否相互独立?

解 由均匀分布的数学期望知 $E(X) = 0$. 由(4.3.5)式知 X 与 Y 的协方差为

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = E(X^3),$$

而

$$E(X^3) = \int_{-1}^1 \frac{1}{2} x^3 dx = 0,$$

所以

$$\text{Cov}(X, Y) = 0.$$

注意到 $D(X) \neq 0$, $D(Y) \neq 0$, 故 $\rho_{XY} = 0$, 即 X 与 Y 不相关; 但由于 $Y = X^2$, 显然 X 与 Y 不独立.

事实上, $\rho_{XY} = 0$ 只说明了 X 与 Y 之间没有线性关系, 但这时 X 与 Y 之间可能存在某种非线性关系. 特别地, 对二维正态随机变量 (X, Y) 而言, X 与 Y 相互独立和 X 与 Y 不相关是等价的.

例 4.3.5 已知 $D(X) = 4$, $D(Y) = 1$, $\rho_{XY} = \frac{1}{2}$, 求 $D(3X - 2Y)$.

解 由方差及协方差的性质:

$$\begin{aligned} D(3X - 2Y) &= D(3X) + D(2Y) - 2\text{Cov}(3X, 2Y) \\ &= 9D(X) + 4D(Y) - 12\text{Cov}(X, Y) \\ &= 9D(X) + 4D(Y) - 12\rho_{XY} \cdot \sqrt{DX} \cdot \sqrt{DY} \\ &= 9 \times 4 + 4 \times 1 - 12 \times \frac{1}{2} \times 2 \times 1 = 28. \end{aligned}$$

3. 矩和协方差矩阵

定义 4.3.2 设 X 和 Y 是随机变量, 若

$$E(X^k), k = 1, 2, \dots$$

存在, 则称它为 X 的 k 阶原点矩.

若

$$E\{[X - E(X)]^k\}, k = 1, 2, \dots$$

存在, 则称它为 X 的 k 阶中心矩.

若

$$E(X^k Y^l), k, l = 1, 2, \dots$$

存在, 则称它为 X 和 Y 的 $k+l$ 阶混合矩.

若

$$E\{[X - E(X)]^k [Y - E(Y)]^l\}, k, l = 1, 2, \dots$$

存在, 则称它为 X 和 Y 的 $k+l$ 阶中心混合矩.

显然, 数学期望 $E(X)$, $E(Y)$ 是随机变量 X, Y 的一阶原点矩, 方差 $D(X)$, $D(Y)$ 是二阶中心矩, 协方差 $\text{Cov}(X, Y)$ 是 X 和 Y 的二阶混合中心矩.

定义 4.3.3 设 $(X_1, X_2, \dots, X_n)$ 为 n 维随机变量, 记 $c_{ij} = \text{Cov}(X_i, X_j)$, $i, j = 1, 2, \dots, n$, 若它们都存在, 则称矩阵

$$\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}$$

为 n 维随机变量 $(X_1, X_2, \dots, X_n)$ 的协方差矩阵.

易证协方差矩阵是非负定对称矩阵.

例如, 二维随机变量 $(X, Y) \sim N(\mu_1, \sigma_1^2; \mu_2, \sigma_2^2; \rho)$, 则其协方差矩阵为

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

§ 4.4 大数定律与中心极限定理

概率论与数理统计的研究内容是随机现象的统计规律性, 而随机现象的规律是通过大量的重复试验呈现出来的. 为了精确地描述这种规律性, 本节将引入极限定理, 其中最主要的是大数定律与中心极限定理. 它们在概率论与数理统计的理论研究和实际应用中具有重要的意义.

1. 切比雪夫不等式

定理 4.4.1 (切比雪夫不等式) 设随机变量 X 的数学期望 $E(X)$ 与方差 $D(X)$ 存在, 则对任意的 $\varepsilon > 0$, 有

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{D(X)}{\varepsilon^2}. \quad (4.4.1)$$

证明 这里仅对 X 是连续型随机变量证明. 设 X 的概率密度函数为 $f(x)$, 对于任意 $\varepsilon > 0$, 有

$$\begin{aligned} P(|X - E(X)| \geq \varepsilon) &= \int_{|X - E(X)| \geq \varepsilon} f(x) dx \\ &\leq \int_{|X - E(X)| \geq \varepsilon} \frac{[x - E(X)]^2}{\varepsilon^2} f(x) dx \\ &\leq \frac{1}{\varepsilon^2} \int_{-\infty}^{+\infty} [x - E(X)]^2 f(x) dx \\ &= \frac{D(X)}{\varepsilon^2}. \end{aligned}$$

因为 $\{|X - E(X)| \geq \varepsilon\}$ 与 $\{|X - E(X)| < \varepsilon\}$ 为对立事件, 所以切比雪夫不等式也可以写成如下形式:

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{D(X)}{\varepsilon^2}. \quad (4.4.2)$$

由切比雪夫不等式可看出: 当误差 ε 取定时, 随着方差 $D(X)$ 减小, X 围绕 $E(X)$ 取值的概率增大. 反之, 随着方差 $D(X)$ 的增大, X 围绕 $E(X)$ 取值的概率减少. 因而进一步说明方差 $D(X)$ 能描述 X 对其均值 $E(X)$ 的偏离程度.

切比雪夫不等式对理论研究和实际应用都具有重要的价值, 它在理论上是证明大数定律的工具, 在实际应用中可以对不便计算的概率加以估计.

例 4.4.1 某车间生产一种电子器件, 月平均产量为 9 500 只, 方差为 10 000, 试估计车间月产量为 9 000 至 10 000 只之间的概率.

解 设该车间的月生产量为 X , 则 $E(X) = 9\,500$, $D(X) = 10\,000$, 取 $\varepsilon = 500$, 由切比雪夫不等式得

$$\begin{aligned} P(9\,000 < X < 10\,000) &= P(|X - 9\,500| < 500) \\ &\geq 1 - \frac{(100)^2}{(500)^2} = 0.96. \end{aligned}$$

即该车间月生产量为 9 000 至 10 000 之间的概率不小于 0.96.

2. 大数定律

我们已经知道在一定条件下, 多次重复进行某一试验, 随机事件发生的频率随着次数的增多逐渐稳定在某一个常数附近, 这一数值也就是随机事件的概率. 但到目前为止, 我们还没有在理论上对这种稳定性给以说明. 另外, 直观的经验表明, 大量观测值的算术平均值也具有稳定性, 即在相同条件下随着观测次数的增多, 观测值的算术平均值逐渐稳定于某一常数附近, 这一数值就是观测值(看作随机变量)的数学期望. 概率论中用来阐明大量随机现象平均结果的稳定性的定理统称为大数定律.

定理 4.4.2 (伯努利大数定律) 设 k_n 是 n 次独立重复试验中事件 A 发生

的次数, 而 p 是事件 A 在每次试验中发生的概率, 对于任意给的正数 ε , 都有

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{k_n}{n} - p\right| < \varepsilon\right) = 1. \quad (4.4.3)$$

伯努利大数定律, 从理论上说明任一随机事件的频率具有稳定性, 这为概率的统计定义提供了理论依据. 于是, 实际问题中, 当试验次数很大时, 可以用事件 A 发生的频率作为概率 p 的近似值.

定理 4.4.3 (平均数法则) 设随机变量 $X_1, X_2, \dots, X_n, \dots$ 相互独立, 且具有相同的数学期望和方差, 即

$$E(X_k) = \mu, D(X_k) = \sigma^2, k = 1, 2, \dots,$$

则对于任意的 $\varepsilon > 0$, 都有

$$\lim_{n \rightarrow +\infty} P\left(\left|\frac{1}{n} \sum_{k=1}^n X_k - \mu\right| < \varepsilon\right) = 1 \quad (4.4.4)$$

或

$$\lim_{n \rightarrow +\infty} P\left(\left|\frac{1}{n} \sum_{k=1}^n X_k - \mu\right| \geq \varepsilon\right) = 0.$$

即在试验次数无限增多的情况下, 算术平均值与 μ 有较大偏差的可能性很小. 利用切比雪夫不等式可以证明上述两个定理.

3. 中心极限定理

在实际问题中, 有许多随机现象可以看作是由大量相互独立的因素综合影响的结果, 而每一个因素对该现象的影响都很微小, 那么作为因素总和的随机变量, 往往服从或近似服从正态分布. 概率论中有关阐述独立随机变量和的极限分布是正态分布的定理称为中心极限定理. 这里只介绍其中之一.

定理 4.4.4 设 $X_1, X_2, \dots, X_n, \dots$ 是具有相同分布且相互独立的一列随机变量,

$$E(X_k) = \mu, D(X_k) = \sigma^2, k = 1, 2, \dots,$$

则对于任意 x 都有

$$\lim_{n \rightarrow +\infty} P\left(\frac{\sum_{k=1}^n X_k - n\mu}{\sigma\sqrt{n}} \leq x\right) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(x). \quad (4.4.5)$$

定理 4.4.4 表明以下结论:

当 n 充分大时, 独立同分布的随机变量之和 $\sum_{k=1}^n X_k$ 近似服从正态分布 $N(n\mu, n\sigma^2)$, 即

$$\frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\sigma} \underset{\text{近似地}}{\sim} N(0, 1).$$

若将上式左端改写成

$$\frac{\frac{1}{n} \sum_{k=1}^n X_k - \mu}{\sigma / \sqrt{n}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}},$$

则上述结论可写成: 当 n 充分大时,

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \underset{\text{近似地}}{\sim} N(0, 1) \text{ 或 } \bar{X} \underset{\text{近似地}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

这是中心极限定理的另一表达形式, 这一结论在数理统计中有重要应用.

如果在定理 4.4.4 中令 X_i 服从 $(0-1)$ 分布, 则 $Y_n = \sum_{k=1}^n X_k$ 的极限分布是正态分布, 由此可得到定理 4.4.4 的特例如下:

定理 4.4.5 (棣莫弗-拉普拉斯定理) 设 $Y_n (n=1, 2, \dots)$ 服从二项分布 $B(n, p)$, 则对于任意 x , 有

$$\lim_{n \rightarrow \infty} P\left(\frac{Y_n - np}{\sqrt{np(1-p)}} \leq x\right) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(x). \quad (4.4.6)$$

该定理表明, 正态分布是二项分布的极限分布, 当 n 充分大时, Y_n 近似服从 $N(np, np(1-p))$.

例 4.4.2 一生产线生产的产品成箱包装, 每箱的质量是随机的. 假设每箱平均重 50 kg, 标准差为 5 kg. 若用最大载重量为 5 t 的汽车承运, 问每辆车最多可以装多少箱, 才能保障不超载的概率大于 0.977?

解 设 $X_k (k=1, 2, \dots, n)$ 是装运的第 k 箱产品的质量 (单位: kg), n 是所求箱数. 由已知条件知 $X_1, X_2, \dots, X_n$ 独立同分布, $E(X_k) = 50$, $D(X_k) = 5^2$.

根据定理 4.4.4, n 箱的总质量 $Y_n = \sum_{k=1}^n X_k$ 近似服从正态分布 $N(50n, 25n)$, 则

$$\begin{aligned} P(Y_n \leq 5000) &= P\left(\frac{Y_n - 50n}{5\sqrt{n}} \leq \frac{5000 - 50n}{5\sqrt{n}}\right) \\ &\approx \Phi\left(\frac{5000 - 50n}{5\sqrt{n}}\right) > 0.977, \end{aligned}$$

查标准正态分布表可得 $\Phi(2) = 0.977$, 故

$$\frac{5000 - 50n}{5\sqrt{n}} > 2,$$

即 $n < 98.02$. 可见每辆车最多可以装 98 箱, 才能使不超载的概率大于 0.977.

例 4.4.3 某单位有一台电话总机, 装有 200 台分机. 设每台分机要求使用外线通话的概率为 0.2, 且是相互独立的, 试问要装置多少外线才能以 95%

的概率保证每台分机对外线的使用?

解 在任一时刻, 分机使用外线的个数为一随机变量 X , 则 $X \sim B(200, 0.2)$. 由题意, 需求出一个最小的正整数 r , 使

$$P(0 \leq X \leq r) = \sum_{k=0}^r C_{200}^k (0.2)^k (0.8)^{200-k} \geq 0.95.$$

直接求 r 有困难, 因为 n 较大, 由定理 4.4.5, X 近似服从正态分布 $N(40, 32)$. 所以

$$\begin{aligned} P(0 \leq X \leq r) &= P\left(-\frac{40}{\sqrt{32}} \leq \frac{X-40}{\sqrt{32}} \leq \frac{r-40}{\sqrt{32}}\right) \\ &\approx \Phi\left(\frac{r-40}{\sqrt{32}}\right) - \Phi\left(-\frac{40}{\sqrt{32}}\right). \end{aligned}$$

于是, 问题归纳为求满足下式的 r :

$$\Phi\left(\frac{r-40}{\sqrt{32}}\right) - \Phi\left(-\frac{40}{\sqrt{32}}\right) = 0.95.$$

查标准正态分布表可知

$$\Phi\left(-\frac{40}{\sqrt{32}}\right) \approx 0.$$

而

$$\Phi(1.65) \approx 0.95.$$

故可令

$$\frac{r-40}{\sqrt{32}} = 1.65,$$

得

$$r \approx 50.$$

也就是说该单位内同时有 50 台以上分机使用外线的概率只有 0.05, 所以只需装置 50 条外线就能以 95% 的概率保证每个分机使用外线.

§ 4.5 用 MATLAB 计算数学期望和方差

当随机变量服从已知的概率分布时, 可直接使用表 4.5.1 中的函数来计算其数学期望和方差. 表 2.4.1 中函数各输入参数的意义与表 4.5.1 中的相同, 而返回值中 E 是期望值、 D 是方差值. 例如下面命令:

```
[E,D]=normstat(0.5,2) ↓
```

```
E =
```

```
0.5
```

D =

4

表 4.5.1 分布的数学期望和方差计算函数

分布类型	调用格式
二项分布	$[E, D] = \text{binostat}(N, P)$
χ^2 分布	$[E, D] = \text{chi2stat}(N)$
指数分布	$[E, D] = \text{expstat}(MU)$
F 分布	$[E, D] = \text{fstat}(M, N)$
正态分布	$[E, D] = \text{normstat}(MU, SIGMA)$
泊松分布	$[E, D] = \text{poisstat}(LAMBDA)$
学生 t 分布	$[E, D] = \text{tstat}(N)$
离散均匀分布	$[E, D] = \text{unidstat}(N)$
连续均匀分布	$[E, D] = \text{unifstat}(A, B)$

计算出正态分布 $N(0.5, 2^2)$ 的期望为 0.5, 方差为 4.

而当随机变量的概率分布是用数据表的形式给出时, 利用 MATLAB 的矩阵运算功能, 可以很容易地按定义来计算其期望和方差. 特别当数据量较大时, 用 MATLAB 作计算工具是非常有效的.

例 4.5.1 在相同条件下, 用甲、乙两种方法测定某一样品的含量. 测定结果分别用 X_1, X_2 表示. 由以往大量测定结果获知, X_1, X_2 的分布如下表. 试比较哪一种测定方法的精度较高(习题 4 第 2 题).

X_1, X_2 的可能取值	48	49	50	51	52
$P(X_1)$	0.1	0.1	0.6	0.1	0.1
$P(X_2)$	0.2	0.2	0.2	0.2	0.2

解 首先输入随机变量 X 及两种测定法之概率 $P1, P2$ 的数据:

$X = [48 \ 49 \ 50 \ 51 \ 52]; \downarrow$

$P1 = [0.1 \ 0.1 \ 0.6 \ 0.1 \ 0.1]; \downarrow$

$P2 = [0.2 \ 0.2 \ 0.2 \ 0.2 \ 0.2]; \downarrow$

然后按定义计算数学期望

$E1 = P1 * X' \downarrow$

$E1 =$

50.0000

$$E2 = P2 * X' \downarrow$$

$$E2 =$$

$$50$$

可见两种测定法的期望值相同. 再按定义计算方差:

$$E = E1 * \text{ones}(1,5) \downarrow$$

$$E =$$

$$50 \quad 50 \quad 50 \quad 50 \quad 50$$

$$D1 = (X - E).^2 * P1' \downarrow$$

$$D1 =$$

$$1$$

$$D2 = (X - E).^2 * P2' \downarrow$$

$$D2 =$$

$$2$$

显然 $D1 < D2$, 于是知甲测定方法的精度高于乙测定方法.

习 题 4

1. 某项动物试验难度颇高, 稍有疏忽便需另换动物重做. 由长期的经验获知, 学生们做此试验需要一、二、三、四、五只动物的概率分别为 0.25, 0.4, 0.2, 0.1, 0.05. 试问做该项试验, 平均每组需要多少只动物?

2. 在相同条件下, 用甲、乙两种方法测定某一样品的含量, 测定结果分别用 X_1, X_2 表示, 由以往大量测定结果获知, X_1, X_2 的分布如下. 试比较哪一种测定方法的精度较高.

X_1, X_2 的可能取值	48	49	50	51	52
$P(X_1)$	0.1	0.1	0.6	0.1	0.1
$P(X_2)$	0.2	0.2	0.2	0.2	0.2

3. 一批零件中有 9 个合格品及 3 个废品, 从这批零件中任取一个, 如果每次取出的废品不再放回, 求在取得合格品以前已取出的废品数的数学期望, 方差及标准差.

4. 设随机变量 X 的数学期望为 $E(X)$, 方差为 $D(X) > 0$, 若令

$$X^* = \frac{X - E(X)}{\sqrt{D(X)}},$$

验证: $E(X^*) = 0, D(X^*) = 1$.

5. 设随机变量 X 的概率密度为

$$f(x) = \begin{cases} \frac{1}{\pi \sqrt{1-x^2}}, & |x| < 1, \\ 0, & \text{其它.} \end{cases}$$

求 $E(X)$, $D(X)$.

6. 设随机变量 X 服从拉普拉斯分布. 其概率密度为

$$f(x) = \frac{1}{2}e^{-|x|}, \quad -\infty < x < +\infty,$$

求 X 的数学期望及方差.

7. 设随机变量 X 的概率分布为

$$P(X=k) = \frac{1}{1+a} \left(\frac{a}{1+a} \right)^k, \quad k=0, 1, 2, \dots,$$

其中 $a>0$ 为常数, 求 $E(X)$, $D(X)$.

8. 证明对任意随机变量, 若 $E(X)$, $D(X)$ 存在, 有

$$(1) E(aX+b) = aE(X) + b;$$

$$(2) D(cX) = c^2 D(X).$$

9. 设 X 为随机变量, C 为常数, 试证: $D(X) \leq E(X-C)^2$.

10. 调查某城市 11 岁男孩的身高, 已知其数学期望是 143.10 cm, 标准差是 5.67 cm, 假设身高服从正态分布, 试指出该城市 11 岁男孩身高的 95% 正常值范围.

11. 证明: 若随机变量 X 的密度函数 $f(x)$ 满足

$$f(c+x) = f(c-x), \quad x>0,$$

其中 c 为常数, 又 $E(X)$ 存在, 则 $E(X) = c$.

12. 已知随机变量 X 的概率密度函数为

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

求: (1) $Y=2X$ 的数学期望; (2) $Y=e^{-2X}$ 的数学期望.

13. 设随机变量 (X, Y) 的分布律为

$X \backslash Y$	1	2
1	0.25	0.32
2	0.08	0.35

求 $E(X^2 + Y)$.

14. 设二维随机变量 (X, Y) 的概率密度为

$$f(x, y) = \begin{cases} x+y, & 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0, & \text{其它,} \end{cases}$$

求: $E(X)$, $E(Y)$, $E(XY)$.

15. 设 X, Y 相互独立, 其密度函数分别为

$$f_X(x) = \begin{cases} 2x, & 0 \leq x \leq 1, \\ 0, & \text{其它,} \end{cases}$$

$$f_Y(y) = \begin{cases} e^{-(y-5)}, & y>5, \\ 0, & \text{其它,} \end{cases}$$

求 $E(XY)$.

16. 设随机变量 X 服从几何分布,

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

其中 $0 < p < 1$ 是常数, 求 $E(X)$ 和 $D(X)$.

17. 设随机变量 X 服从瑞利分布, 其密度函数为

$$f(x) = \begin{cases} \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

其中 $\sigma > 0$ 是常数. 求 $E(X)$ 和 $D(X)$.

18. 设 $X_1, X_2, \dots, X_n$ 是 n 个独立同分布的随机变量, 数学期望 $E(X_i) = \mu$, 方差 $D(X_i) = \sigma^2$, $i = 1, 2, \dots, n$. 设 $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$, 求 $E(\bar{X})$ 和 $D(\bar{X})$.

19. 设某种商品每周的需求量 X 为随机变量, $X \sim U[10, 30]$, 而经销商进货数量为区间 $[10, 30]$ 中的某一整数, 经销商每销售一单位的商品可获利 500 元, 若供大于求则必须削价处理, 每处理一单位商品要亏损 100 元; 若供不应求, 则可以从其他商店调剂供货, 但此时每销售一单位商品仅获利 300 元. 为使商店所获利润的期望值不少于 9 280 元, 试确定最少进货量.

20. 假设一部机器在一天内发生故障的概率为 0.2, 且一旦发生故障全天停止工作. 一周 5 个工作日里无故障, 可获利润 10 万元; 只发生 1 次故障, 可获利润 5 万元; 发生 2 次故障, 获利润 0 元; 若发生 3 次或 3 次以上故障, 要亏损 2 万元. 求一周内利润的期望值.

21. 设随机变量 (X, Y) 的分布律为

$X \backslash Y$	-1	0	1
0	0.07	0.18	0.15
1	0.08	0.32	0.20

求 ρ_{XY} , $\text{Cov}(X^2, Y^2)$.

22. 证明: (1) $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$;

(2) $D(X \pm Y) = D(X) + D(Y) \pm 2\text{Cov}(X, Y)$.

23. 设 (X, Y) 在三角形区域 $D = \{(x, y) \mid 0 \leq x \leq 1, 0 \leq y \leq 1 - x\}$ 上服从均匀分布, 求 $\text{Cov}(X, Y)$ 和 ρ_{XY} .

24. 设随机变量 (X, Y) 具有密度函数

$$f(x, y) = \begin{cases} \frac{1}{8}(x+y), & 0 \leq x \leq 2, 0 \leq y \leq 2, \\ 0, & \text{其它,} \end{cases}$$

求 $E(X)$, $E(Y)$, ρ_{XY} ,

25. 设 $X \sim N(0, 1)$, 而 $Y = X^n$ (n 为正整数), 求 ρ_{XY} .

26. 设随机变量 X 和 Y 分别服从正态分布 $N(1, 3^2)$ 和 $N(0, 4^2)$, 且 X 与 Y 的相关系数 $\rho_{XY} = -\frac{1}{2}$. 设 $Z = \frac{X}{3} + \frac{Y}{2}$, 求: (1) $E(Z)$, $D(Z)$; (2) X 与 Z 的相关系数 ρ_{XZ} .

27. 设 $Y = aX + b$, 其中 a, b 为常数, 并且 $a > 0$, 证明 $\rho_{XY} = 1$.

28. 设 $X_1 \sim N(\mu, \sigma^2)$, $X_2 \sim N(\mu, \sigma^2)$, 且 X_1 与 X_2 相互独立, 试求 $Y_1 = aX_1 + bX_2$, $Y_2 = aX_1 - bX_2$ 的相关系数 (其中 a, b 为不为零的常数).

29. 已知正常男性成人血液中, 每一毫升白细胞数平均是 7 300, 标准差为 700, 利用

切比雪夫不等式估计每毫升血液中含白细胞数在 5 200 ~ 9 400 之间的概率.

30. 某市有 50 个无线寻呼台, 每个寻呼台在每分钟内收到的电话呼叫次数服从参数为 $\lambda = 0.05$ 的泊松分布. 试求该市某一分钟内的呼叫次数的总和大于 3 次的概率.

31. 若有一万人参加某保险公司开办的人身意外保险, 按规定, 每人付 18 元的保险费. 若出现意外投保人可向保险公司领取 2 500 元. 假设该人群出现意外的概率为 0.006, 问保险公司亏本的概率多大?

32*. 设随机变量 X 与 Y 相互独立, 且都服从期望为 0, 方差为 $\frac{1}{2}$ 的正态分布, 求随机变量 $|X - Y|$ 的期望与方差.

33*. 已知三个随机变量 X, Y, Z 中, $E(X) = E(Y) = 1, E(Z) = -1, D(X) = D(Y) = D(Z) = 1, \rho_{XY} = 0, \rho_{XZ} = \frac{1}{2}, \rho_{YZ} = -\frac{1}{2}$. 求 $E(X + Y + Z), D(X + Y + Z)$.

34*. 设随机变量 $X_1, X_2, \dots, X_n (n > 1)$ 独立同分布, 且其方差为 $\sigma^2 > 0$. 令 $Y = \frac{1}{n} \sum_{i=1}^n X_i$, 求 $\text{Cov}(X_1, Y), D(X_1 - Y)$.

35*. 设 A, B 是随机事件, 随机变量

$$X = \begin{cases} 1, & \text{若 } A \text{ 发生,} \\ -1, & \text{若 } A \text{ 不发生.} \end{cases} \quad Y = \begin{cases} 1, & \text{若 } B \text{ 发生,} \\ -1, & \text{若 } B \text{ 不发生.} \end{cases}$$

试证明: 随机变量 X 和 Y 不相关的充分必要条件是 A 与 B 相互独立.

36*. 某种汽车的碳排放量的数学期望为 0.9 g/km, 标准差为 1.9 g/km, 某公司有这种汽车 100 辆, 以 $\bar{X}$ 表示这些车辆的碳排放量的算术平均, 问当 L 为何值时, $\bar{X}$ 大于 L 的概率不超过 0.01?

第5章 数理统计的基本知识

§ 5.1 数理统计学

数理统计学是研究大量随机现象规律性的一门数学学科。它从实际观测数据出发，以概率论为基础，研究如何用有效的方法去收集和利用数据资料，对资料进行整理、分析和推断。

在概率论中，我们把对随机现象的观测叫**随机试验**，引进了**样本空间**及**随机变量**等概念，研究了事件发生的概率以及随机变量的分布函数（概率分布或密度函数）和数字特征等。但在实际问题中，人们一般事前并不知道某随机事件发生的概率，对随机变量的分布也往往了解不多。请看下面的例子。

例 5.1.1 研究某厂生产的一批灯泡的质量状况，回答下列问题：

(1) 估计这一批灯泡的平均使用寿命；

(2) 若规定使用寿命低于 800 h 的灯泡为次品，当次品率超过 5% 时，则不接受这批产品，试对这批产品作出接受与否的判断。

要回答上述问题，就必须对灯泡的使用寿命这一随机变量进行研究。

随机性的普遍存在，促进了数理统计的发展，也为数理统计的应用提供了一个广阔的用武之地。数理统计的方差分析、显著性检验及回归分析等方法已广泛应用于工农业生产、医学、质量控制等许多领域中。

§ 5.2 总体与样本

1. 总体与个体

在数理统计中将研究对象的全体称为**总体**，构成总体的每一成员称为**个体**。在实际问题中，人们关心的往往是每一成员的某种**数量指标**，而不是每个成员的种种具体特性。人们关心的这种数量指标随着成员的不同而不同。如在例 5.1.1 中，我们关心的是灯泡的使用寿命这一数量指标。又如，我们要研究北京市中学生的身高和体重的分布情况，把北京市全体中学生看作总体。但这不能抓住问题的本质，我们真正的研究对象是该市中学生的身高和体重这两个数量指标。因而我们通常把个体所具有的数量指标的全体作为**总体**，而个体即为每一成员的数量指标。

在例 5.1.1 中, 灯泡的使用寿命在观测之前只知道其一切可能取值, 并不能确定其具体的取值. 因而我们可以用随机变量 X 去描述总体, 即把要研究的随机变量 X 定义为总体, 简称总体 X . 所谓总体的分布也就是随机变量 X 的分布.

如果被研究对象的数量指标不止一个, 不妨设为 $k(k \geq 2)$ 个, 则它们构成 k 维向量, 称为 k 维总体. 例如上面提到的北京市中学生的身高和体重是二维总体. 又如, 考察企业的经济效益状况时, 有劳动生产率、资金利润率等多个数量指标, 是一个多维总体问题.

根据总体所含个体的数量多少可分为有限总体和无限总体. 若总体含有有限个个体, 称为有限总体; 若总体含有无限或多得无法计数的个体, 则称为无限总体. 在实际应用中, 有时一个总体尽管所含个体有限, 但个体数目足够大时, 也可视为无限总体.

2. 样本

为了研究总体, 就必须对其个体进行试验与观测. 但在绝大多数情况下, 对个体的观测往往要付出一定的人力、物力和财力, 有些试验或观测具有破坏性(例如, 观测灯泡使用寿命), 因此, 通常人们只是从总体中抽取若干个体, 通过测定这些个体的值对总体进行推断.

从总体中抽取的待测个体组成的集合称为样本, 样本所含的个体数目称为样本容量. 从总体 X 中抽取的容量为 n 的样本常记为 $X_1, X_2, \dots, X_n$, 这里每个 X_i 都被看作是一个随机变量, 这是因为第 i 个被抽到的个体具有随机性, 其取值在观测前是无法知道的.

样本 $X_1, X_2, \dots, X_n$ 所有可能取值的全体(是 $\mathbf{R}^n$ 或 $\mathbf{R}^n$ 的一子集)称为样本空间. 一组样本观测值 $x_1, x_2, \dots, x_n$ 就是样本空间的一个点(元素). 为了有效地利用样本来推断总体, 一般对样本作如下要求:

(1) 独立性: 要求 $X_1, X_2, \dots, X_n$ 是相互独立的随机变量, 即每个观测结果既不影响其它观测结果, 也不受其它观测结果的影响.

(2) 代表性: 要求 $X_1, X_2, \dots, X_n$ 能代表总体的特性, 即要求 $X_1, X_2, \dots, X_n$ 与总体 X 具有相同的分布.

凡满足上述两条性质的样本称为简单随机样本. 本书中, 凡提到样本, 都是指简单随机样本.

在实践中如何才能得到简单随机样本呢?

对于有限总体, 若采用有放回的抽样观察, 则所得样本即为简单随机样本. 若采用不放回的抽样观察, 当样本容量相对于总体所含个体数目很小时, 也可近似视为简单随机样本. 对于无限总体, 有放回和不放回的抽样所得样本都是简单随机样本.

对于简单随机样本, 有下列性质:

定理 5.2.1 设 $X_1, X_2, \dots, X_n$ 为来自总体 X 的简单随机样本.

(1) 如果总体 X 的分布函数为 $F(x)$, 则样本 $X_1, X_2, \dots, X_n$ 的联合分布函数为

$$F(x_1, x_2, \dots, x_n) = \prod_{k=1}^n F(x_k).$$

(2) 如果总体 X 的密度函数为 $f(x)$, 则样本 $X_1, X_2, \dots, X_n$ 的联合密度函数为

$$f(x_1, x_2, \dots, x_n) = \prod_{k=1}^n f(x_k).$$

(3) 如果总体 X 的均值和方差分别为 μ, σ^2 , 则

$$E(X_k) = \mu, D(X_k) = \sigma^2, k = 1, 2, \dots, n.$$

§ 5.3 统计量与抽样分布

1. 统计量

样本取自于总体, 它含有总体的信息. 为了对总体进行推断, 需要对样本的 n 个观测值进行加工处理, 把样本中所包含的有关信息集中起来. 具体地说, 就是针对不同问题构造样本的不同函数, 再利用这些函数的取值对总体进行推断. 这类函数就是统计量.

定义 5.3.1 设 $X_1, X_2, \dots, X_n$ 是来自总体 X 的一组样本. $T(X_1, X_2, \dots, X_n)$ 为一实值函数, 且不含任何未知参数, 则称 $T = T(X_1, X_2, \dots, X_n)$ 为一统计量.

从上述定义不难看出, 统计量仍然是一个随机变量, 当样本的观测值 $x_1, x_2, \dots, x_n$ 给定后, 统计量的取值就完全确定了. $T(x_1, x_2, \dots, x_n)$ 就是统计量 $T(X_1, X_2, \dots, X_n)$ 的观测值.

例如, 若 $X_1, X_2, \dots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的一组样本, 其中 $\mu = \mu_0$ 已知, 而 σ^2 未知. 则 $\frac{1}{n} \sum_{i=1}^n (X_i - \mu), X_1^2 + X_2^2 + 1, X_1 + X_2$ 都是统计量. 但 $\frac{X_1^2}{\sigma^2}, \frac{X_1 - \mu}{\sigma}$ 都不是统计量, 因为它们含有未知参数 σ .

2. 常用统计量

设 $X_1, X_2, \dots, X_n$ 是来自总体 X 的一组样本, $x_1, x_2, \dots, x_n$ 是样本观测值. 理论上, 可以构造无穷多的统计量, 但实际应用中用到的统计量并不多, 常用统计量有

(1) 样本均值

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i;$$

(2) 样本方差

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right);$$

(3) 样本标准差

$$S = \sqrt{S^2};$$

(4) 样本 k 阶原点矩

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k, \quad k = 1, 2, \dots;$$

(5) 样本 k 阶中心矩

$$B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k, \quad k = 1, 2, \dots.$$

它们的观测值分别为

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

$$s = \sqrt{s^2},$$

$$a_k = \frac{1}{n} \sum_{i=1}^n x_i^k, \quad k = 1, 2, \dots,$$

$$b_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k, \quad k = 1, 2, \dots.$$

另外, 还有顺序统计量、样本偏度及样本峰度等统计量.

设 $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ 是来自二维总体 (X, Y) 的一组样本, 则常用统计量有

(1) 样本协方差

$$S_{XY}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y});$$

(2) 样本相关系数

$$\rho_{XY} = \frac{S_{XY}^2}{S_X \cdot S_Y},$$

其中

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

3. 抽样分布

构造统计量的目的, 是对总体进行某种推断, 这就需要了解统计量的特性. 由于统计量是随机变量, 因而我们必须了解它的分布.

统计量的分布称为**抽样分布**.

下面是数理统计中常用的三种分布:

(1) χ^2 分布

定义 5.3.2 设随机变量 $X_1, X_2, \dots, X_n$ 相互独立, 并且 $X_i \sim N(0, 1)$, $1 \leq i \leq n$, 则称随机变量

$$\chi^2 = X_1^2 + X_2^2 + \dots + X_n^2$$

为服从自由度为 n 的 χ^2 分布, 记为 $\chi^2(n)$, 其密度函数为

$$f_{\chi^2}(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & x > 0, \\ 0, & x \leq 0. \end{cases}$$

χ^2 分布的数学期望和方差分别为

$$E[\chi^2(n)] = n, \quad D[\chi^2(n)] = 2n.$$

利用独立随机变量和的密度函数公式可以证明 χ^2 分布的**可加性**:

如果随机变量 X 和 Y 相互独立, 并且分别服从自由度为 n_1, n_2 的 χ^2 分布, 则

$$X + Y \sim \chi^2(n_1 + n_2).$$

(2) t 分布 (“学生”分布)

定义 5.3.3 设随机变量 X 和 Y 相互独立, 并且 $X \sim N(0, 1)$, $Y \sim \chi^2(n)$. 则称随机变量

$$t = \frac{X}{\sqrt{Y/n}}$$

为服从自由度为 n 的 t 分布, 记为 $t(n)$, 其密度函数为

$$f_t(x) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi} \Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

t 分布的数学期望和方差分别为

$$E(t(n)) = 0, \quad n > 1,$$

$$D(t(n)) = \frac{n}{n-2}, \quad n > 2.$$

t 分布的密度函数是偶函数. 可以证明: 当自由度 n 趋于无穷时, t 分布将趋近于标准正态分布 $N(0, 1)$. 一般当 $n \geq 30$ 时, t 分布可近似看作标准正态分布.

(3) F 分布

定义 5.3.4 设随机变量 X 和 Y 相互独立, 并且 $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, 则称随机变量

$$F = \frac{X/m}{Y/n}$$

为服从自由度为 (m, n) 的 F 分布, 记为 $F(m, n)$, 其中 m, n 分别称为第一、二自由度. F 分布的密度函数为

$$f_F(x) = \begin{cases} \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{x^{\frac{n}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

F 分布的数学期望和方差分别为

$$E[F(m, n)] = \frac{n}{n-2}, \quad n > 2,$$

$$D[F(m, n)] = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, \quad n > 4.$$

由定义可知, 若 $F \sim F(m, n)$, 则

$$\frac{1}{F} \sim F(n, m).$$

(4) 分位数(点)

定义 5.3.5 设随机变量 X 的分布函数为 $F(x)$, 满足等式

$$P(X > x_\alpha) = 1 - F(x_\alpha) = \alpha, \quad 0 < \alpha < 1$$

的实数 x_α 称为 X 的上 α 分位数(点).

标准正态分布, χ^2 分布, t 分布, F 分布的上 α 分位数示意图如图 5.3.1.

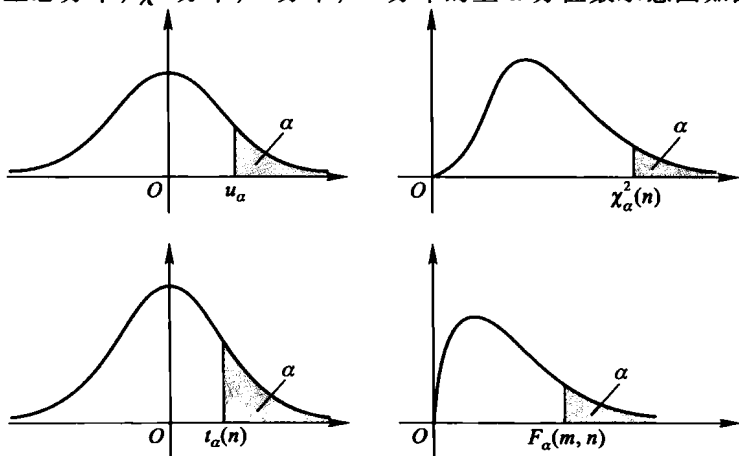


图 5.3.1

例如, $u_{0.05} = 1.645$, $\chi_{0.05}^2(8) = 15.507$,
 $t_{0.01}(10) = 2.7638$, $F_{0.01}(8, 10) = 5.06$.

上述几种分布的分位数有如下性质:

(1) 当 n 充分大时 ($n > 40$ 即可), 近似地有

$$\chi_{\alpha}^2(n) \approx \frac{1}{2}(u_{\alpha} + \sqrt{2n-1})^2;$$

(2) 由 t 分布密度函数图形的对称性及上 α 分位数的定义有

$$t_{1-\alpha}(n) = -t_{\alpha}(n).$$

此外, $n \geq 30$ 时有

$$t_{\alpha}(n) \approx u_{\alpha};$$

(3) 利用上 α 分位数的定义不难得到

$$F_{1-\alpha}(m, n) = \frac{1}{F_{\alpha}(n, m)}.$$

类似可定义下 α 分位数(点).

4. 正态总体的有关抽样分布

从理论上讲, 当总体的分布函数已知时, 统计量的分布就可通过随机变量函数的分布而求出. 但在实际中, 由于涉及复杂的求和或积分运算, 往往得不到统计量分布的明显表达式, 这就限制了统计量的应用. 但在正态总体的条件下, 某些常用统计量有比较简单的结果. 下面给出正态总体条件下某些常用统计量的分布.

定理 5.3.1 设 $X_1, X_2, \dots, X_n$ 是来自正态总体 $X \sim N(\mu, \sigma^2)$ 的一组样本, $\bar{X}$, S^2 分别是样本均值和样本方差, 则

$$(1) \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{或} \quad U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1);$$

$$(2) \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1);$$

$$(3) \bar{X} \text{ 与 } S^2 \text{ 相互独立};$$

$$(4) \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

证明 (1) 由于 $\sum_{i=1}^n X_i$ 是相互独立的正态分布的和, 因而 $\sum_{i=1}^n X_i$ 也服从正态分布. 由于

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i) = n\mu,$$

$$D\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n D(X_i) = n\sigma^2,$$

从而

$$\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2),$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

将 $\bar{X}$ 标准化得

$$U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

(2) 令 $Z_i = \frac{X_i - \mu}{\sigma} (i = 1, 2, \dots, n)$, 则 $Z_1, Z_2, \dots, Z_n$, 相互独立, 且都服从 $N(0, 1)$ 分布, 从而

$$\begin{aligned}\bar{Z} &= \frac{1}{n} \sum_{i=1}^n Z_i = \frac{\bar{X} - \mu}{\sigma}; \\ \frac{(n-1)S^2}{\sigma^2} &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^n \left[\frac{(X_i - \mu) - (\bar{X} - \mu)}{\sigma} \right]^2 \\ &= \sum_{i=1}^n (Z_i - \bar{Z})^2 = \sum_{i=1}^n Z_i^2 - n\bar{Z}^2.\end{aligned}$$

取一 n 阶正交矩阵 $A = (a_{ij})_{n \times n}$, 其中第 n 行元素均为 $\frac{1}{\sqrt{n}}$, 作正交变换

$$Y = AZ,$$

其中 $Y = (Y_1, Y_2, \dots, Y_n)^T$, $Z = (Z_1, Z_2, \dots, Z_n)^T$. 由于 $Y_i = \sum_{j=1}^n a_{ij} Z_j (i = 1, 2, \dots, n)$, 故 $Y_1, Y_2, \dots, Y_n$ 也服从正态分布. 由 $Z \sim N_n(\mathbf{0}, E_n)$ 知,

$$E(Y) = E(AZ) = AE(Z) = \mathbf{0},$$

$$D(Y) = \text{Cov}(Y, Y) = \text{Cov}(AZ, AZ) = A \text{Cov}(Z, Z) A^T = AE_n A^T = E_n \textcircled{1},$$

故 $Y_1, Y_2, \dots, Y_n$ 两两不相关, 而且 $Y_i \sim N(0, 1) (i = 1, 2, \dots, n)$. 又由于 $Y = (Y_1, Y_2, \dots, Y_n)^T$ 是由 $Z = (Z_1, Z_2, \dots, Z_n)^T$ 经过线性变换而得到的, 因而 $Y_1, Y_2, \dots, Y_n$ 相互独立. 而

$$Y_n = \sum_{j=1}^n a_{nj} Z_j = \sum_{j=1}^n \frac{1}{\sqrt{n}} Z_j = \sqrt{n} \bar{Z};$$

$$\sum_{i=1}^n Y_i^2 = Y^T Y = (AZ)^T (AZ) = Z^T (A^T A) Z = Z^T E_n Z = Z^T Z = \sum_{i=1}^n Z_i^2,$$

① $\text{Cov}(AX, BY) = A \text{Cov}(X, Y) B^T$.

于是

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n Z_i^2 - n\bar{Z}^2 = \sum_{i=1}^n Y_i^2 - Y_n^2 = \sum_{i=1}^{n-1} Y_i^2.$$

由于 $Y_1, Y_2, \dots, Y_{n-1}$ 相互独立, 且 $Y_i \sim N(0, 1) (i = 1, 2, \dots, n-1)$, 知

$$\sum_{i=1}^{n-1} Y_i^2 \sim \chi^2(n-1). \text{ 故}$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

(3) 又 $\bar{X} = \sigma\bar{Z} + \mu = \frac{\sigma}{\sqrt{n}}Y_n + \mu$ 只依赖于 Y_n , 而 $S^2 = \frac{\sigma^2}{n-1} \sum_{i=1}^{n-1} Y_i^2$ 仅依赖于 $Y_1, Y_2, \dots, Y_{n-1}$, 从而 $\bar{X}$ 与 S^2 相互独立.

(4) 由于 $U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, 而且两者相互独立, 根据 t 分布的定义得

$$\frac{U}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

定理 5.3.2 设 $X_1, X_2, \dots, X_m$ 是来自正态总体 $X \sim N(\mu_1, \sigma_1^2)$ 的一组样本, $Y_1, Y_2, \dots, Y_n$ 是来自正态总体 $Y \sim N(\mu_2, \sigma_2^2)$ 的一组样本, $\bar{X}, S_x^2$ 及 $\bar{Y}, S_y^2$ 分别表示两样本的样本均值和样本方差, 假定两样本相互独立, 则

$$(1) \frac{S_x^2/\sigma_1^2}{S_y^2/\sigma_2^2} \sim F(m-1, n-1);$$

$$(2) \text{ 当 } \sigma_1^2 = \sigma_2^2 = \sigma^2 \text{ 时,}$$

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2),$$

其中

$$S_w^2 = \frac{(m-1)S_x^2 + (n-1)S_y^2}{m+n-2};$$

$$(3) \text{ 当 } m=n \text{ 时,}$$

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S/\sqrt{n}} \sim t(n-1),$$

其中

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n [(X_i - Y_i) - (\bar{X} - \bar{Y})]^2 = S_x^2 + S_y^2 - 2S_{xy}.$$

证明 (1) 由定理 5.3.1 的结论(2)知,

$$\frac{(m-1)S_x^2}{\sigma_1^2} \sim \chi^2(m-1), \quad \frac{(n-1)S_y^2}{\sigma_2^2} \sim \chi^2(n-1),$$

而 S_X^2 只依赖于 $X_1, X_2, \dots, X_m$, S_Y^2 只依赖于 $Y_1, Y_2, \dots, Y_n$, 从而 S_X^2 与 S_Y^2 相互独立, 由 F 分布的定义知,

$$\frac{\frac{(m-1)S_X^2}{\sigma_1^2} / (m-1)}{\frac{(n-1)S_Y^2}{\sigma_2^2} / (n-1)} = \frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} \sim F(m-1, n-1).$$

(2) 根据定理 5.3.1 的结论(1)知,

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma^2}{m}\right), \quad \bar{Y} \sim N\left(\mu_2, \frac{\sigma^2}{n}\right),$$

又因为 $X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n$ 相互独立, 从而

$$\begin{aligned} \bar{X} - \bar{Y} &\sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{m} + \frac{\sigma^2}{n}\right), \\ U &= \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0, 1). \end{aligned}$$

又根据定理 5.3.1 的结论(2)及 χ^2 分布的可加性知,

$$V = \frac{(m-1)S_X^2}{\sigma^2} + \frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2(m+n-2).$$

因为 $\bar{X}$ 与 S_X^2 及 S_Y^2 相互独立, $\bar{Y}$ 与 S_X^2 及 S_Y^2 相互独立, 因而 U 与 V 相互独立. 根据 t 分布的定义知,

$$\frac{U}{\sqrt{V/(m+n-2)}} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2),$$

其中

$$S_w^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}.$$

(3) 当 $m=n$ 时, 令

$$Z_i = X_i - Y_i, \quad i=1, 2, \dots, n.$$

由于两组样本相互独立, 因而 $Z_1, Z_2, \dots, Z_n$ 相互独立且均服从正态分布 $N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$, 故它们可看作是来自正态总体 $Z \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$ 的一组样本, 根据定理 5.3.1 的结论(4)得

$$\frac{\bar{Z} - (\mu_1 - \mu_2)}{S_Z/\sqrt{n}} \sim t(n-1),$$

由于 $\bar{Z} = \bar{X} - \bar{Y}$, $S_Z^2 = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2 = S^2$, 所以

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S/\sqrt{n}} \sim t(n-1),$$

其中
$$S^2 = \frac{1}{n-1} \sum_{i=1}^n [(X_i - Y_i) - (\bar{X} - \bar{Y})]^2.$$

例 5.3.1 设 X_1, X_2, X_3, X_4 是来自正态总体 $N(0, 9)$ 的样本, $Y = a(X_1 - X_2)^2 + b(3X_3 + 4X_4)^2$, 问当 a, b 各为多少时, 统计量 Y 服从 $\chi^2(2)$ 分布?

解 由已知 $X_i \sim N(0, 3^2) (i = 1, 2, 3, 4)$, 且 X_1, X_2, X_3, X_4 相互独立, 故

$$\begin{aligned} X_1 - X_2 &\sim N(0, 2 \times 3^2), \quad \frac{X_1 - X_2}{3\sqrt{2}} \sim N(0, 1), \\ 3X_3 + 4X_4 &\sim N(0, 25 \times 3^2), \quad \frac{3X_3 + 4X_4}{15} \sim N(0, 1), \end{aligned}$$

因而

$$\begin{aligned} Y &= \left(\frac{X_1 - X_2}{3\sqrt{2}} \right)^2 + \left(\frac{3X_3 + 4X_4}{15} \right)^2 \\ &= \frac{1}{18} (X_1 - X_2)^2 + \frac{1}{225} (3X_3 + 4X_4)^2 \sim \chi^2(2), \end{aligned}$$

故取 $a = \frac{1}{18}, b = \frac{1}{225}$ 时, Y 服从 $\chi^2(2)$ 分布.

例 5.3.2 设随机变量 X 和 Y 相互独立且都服从正态分布 $N(0, 3^2)$, 而 $X_1, X_2, \dots, X_9$ 和 $Y_1, Y_2, \dots, Y_9$ 分别是来自总体 X 和 Y 的简单随机样本, 则统计量

$$U = \frac{X_1 + X_2 + \dots + X_9}{\sqrt{Y_1^2 + Y_2^2 + \dots + Y_9^2}}$$

服从什么分布?

解 令 $W_i = \frac{X_i}{3}, S_i = \frac{Y_i}{3}, i = 1, 2, \dots, 9$, 则

$$\begin{aligned} W_i &\sim N(0, 1), \quad S_i \sim N(0, 1). \\ W &= W_1 + W_2 + \dots + W_9 \sim N(0, 3^2), \\ S &= S_1^2 + S_2^2 + \dots + S_9^2 \sim \chi^2(9), \end{aligned}$$

因此

$$U = \frac{X_1 + X_2 + \dots + X_9}{\sqrt{Y_1^2 + Y_2^2 + \dots + Y_9^2}} = \frac{W_1 + W_2 + \dots + W_9}{\sqrt{S_1^2 + S_2^2 + \dots + S_9^2}} = \frac{W}{\sqrt{S}} = \frac{W/3}{\sqrt{S/9}},$$

由于 $W/3 \sim N(0, 1), S \sim \chi^2(9)$, 故 $U \sim t(9)$.

例 5.3.3 设总体 $X \sim N(0, 2^2)$, $X_1, X_2, \dots, X_{15}$ 是来自总体 X 的样本, 则统计量

$$Y = \frac{2(X_1^2 + X_2^2 + \dots + X_5^2)}{X_6^2 + X_7^2 + \dots + X_{15}^2}$$

服从什么分布?

解 由于 $X_i \sim N(0, 2^2)$, $i = 1, 2, \dots, 15$, 且相互独立, 故

$$S_1^2 = \left(\frac{X_1}{2}\right)^2 + \left(\frac{X_2}{2}\right)^2 + \dots + \left(\frac{X_5}{2}\right)^2 \sim \chi^2(5),$$

$$S_2^2 = \left(\frac{X_6}{2}\right)^2 + \left(\frac{X_7}{2}\right)^2 + \dots + \left(\frac{X_{15}}{2}\right)^2 \sim \chi^2(10),$$

于是

$$Y = \frac{2(X_1^2 + X_2^2 + \dots + X_5^2)}{X_6^2 + X_7^2 + \dots + X_{15}^2} = \frac{2S_1^2}{S_2^2} = \frac{S_1^2/5}{S_2^2/10} \sim F(5, 10).$$

§ 5.4 数据的整理

在实际统计工作中, 人们对总体的分布类型往往不甚了解. 为此需要对样本观测数据进行整理, 以便获取有关总体分布类型的信息. 数据的整理主要有列表、作图等方法.

1. 频数、频率及累积频率分布表及其图示

如果总体 X 是离散型随机变量, 其所有可能取值为 $a_1, a_2, \dots, a_k$ (也可以为可列个). 假定样本的 n 个观测值为 $x_1, x_2, \dots, x_n$. 由于样本与总体具有相同的分布, 因此样本的任一观测值 $x_j (1 \leq j \leq n)$ 必是某个 a_i . 记 n_i 表示 $x_1, x_2, \dots, x_n$ 中取值为 a_i 的个数, 称 n_i 为频数, $f_i = n_i/n$ 为频率, $(n_1 + n_2 + \dots + n_i)/n$ 为累积频率. 将它们列成一张表, 分别称为频数、频率和累积频率分布表(示例见表 5.4.1).

例 5.4.1 某高校调查了 100 名二年级同学的英语四级成绩, 设成绩分为五类: 优、良、中、及格、不及格, 分别用 0, 1, 2, 3, 4 表示.

表 5.4.1 100 名学生的英语四级成绩频数、频率及累积频率分布表

成绩	频数 n_i	频率 f_i	累积频率 $f_1 + f_2 + \dots + f_i$
0	10	0.10	0.10
1	25	0.25	0.35
2	35	0.35	0.70
3	24	0.24	0.94
4	6	0.06	1
合计	$n = 100$	1	

从表可看出成绩为良好的同学有 25 人, 占 25%, 而及格率为 94%.

为直观起见,可用条形图来表示频数(或频率),图 5.4.1 及图 5.4.2 是例 5.4.1 的频数、频率及累积频率条形图,其中横坐标表示成绩.

根据大数定理知,随着样本容量的增大,频率分布将趋于总体 X 的概率分布,而累积频率将趋于总体 X 的分布函数.

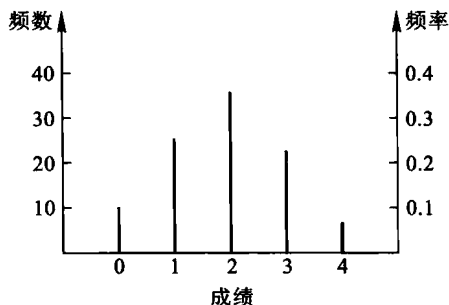


图 5.4.1

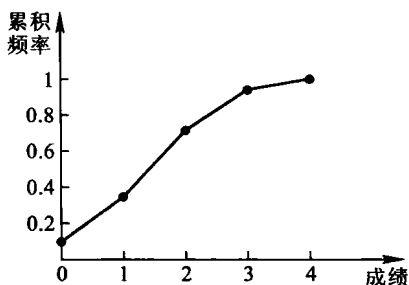


图 5.4.2

如果总体 X 是连续型随机变量,其所有可能取值是某一区间,如灯泡的使用寿命为区间 $(0, M)$, 其中 M 是常数(可能取值较大). 这时可将总体 X 的取值范围分成 k 个小区间 $[a_0, a_1)$, $[a_1, a_2)$, $\dots$, $[a_{k-1}, a_k)$, 然后统计样本观测值 $x_1, x_2, \dots, x_n$ 落在每一小区间中的频数,并计算其频率和累积频率.

例 5.4.2 某工厂生产了一批灯泡,为了考察它们的使用寿命(单位:h)大小,今从中抽取 50 个灯泡进行测量,测得它们的寿命如下:

1 450 1 480 1 640 1 510 1 680 1 670 1 700 1 530 1 640 1 550
 1 590 1 480 1 550 1 590 1 570 1 680 1 730 1 800 1 770 1 420
 1 370 1 350 1 610 1 510 1 630 1 600 1 490 1 460 1 810 1 570
 1 430 1 660 1 570 1 560 1 630 1 740 1 650 1 720 1 540 1 550
 1 730 1 540 1 670 1 430 1 490 1 570 1 580 1 640 1 660 1 390

下面利用分组的方法整理上述数据,具体步骤如下:

- (1) 找出 $x_1, x_2, \dots, x_n$ 的最小值 $x_{(1)}$ 与最大值 $x_{(n)}$, 并计算极差

$$R_n = x_{(n)} - x_{(1)}.$$

本例中 $n=50$, $x_{(1)}=1\ 350$, $x_{(50)}=1\ 810$, $R_{50}=460$.

- (2) 根据样本容量 n 确定分组数 k . 一般 n 越大, 组数 k 取的相应大. 但 k 不宜过大或过小. 本例中取 $k=7$.

- (3) 确定各组端点 $a_0 < a_1 < \dots < a_k$. 通常 $a_0 < x_{(1)}$, $a_k > x_{(n)}$. 在某些情况下, 可取 $a_0 = -\infty$, $a_k = +\infty$. 分组可分等组距与不等组距两种, 一般取等组距分组, 组距 $d \approx R_n/k$.

在本例中取 $a_0 = 1\,335$, $d = 70$, 而 $a_i = a_0 + 70i$, $i = 1, 2, \dots, 7$.

(4) 统计落在每一区间 $[a_{i-1}, a_i)$ 中的频数 n_i , 并计算频率 $f_i = n_i/n$ 及累积频率 $f_1 + f_2 + \dots + f_i$.

本例中的频数、频率及累积频率分布表如表 5.4.2:

表 5.4.2

组号 i	分组区间 $[a_{i-1}, a_i)$	频数 n_i	频率 f_i	累积频率 $f_1 + f_2 + \dots + f_i$
1	[1 335, 1 405)	3	0.06	0.06
2	[1 405, 1 475)	5	0.10	0.16
3	[1 475, 1 545)	9	0.18	0.34
4	[1 545, 1 615)	13	0.26	0.60
5	[1 615, 1 685)	12	0.24	0.84
6	[1 685, 1 755)	5	0.10	0.94
7	[1 755, 1 825)	3	0.06	1
合 计		$n = 50$	1	

为直观起见, 可用下列直方图表示.

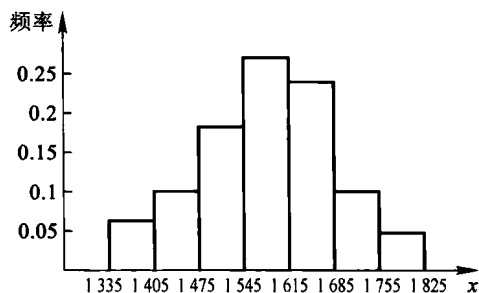


图 5.4.3

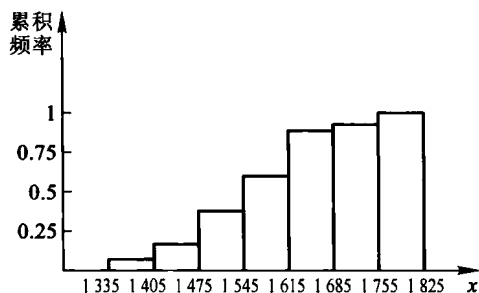


图 5.4.4

当总体 X 是连续型随机变量时, 若将频率直方图的纵坐标刻度变为频率/组距, 这样每一小矩形的面积即为样本观测值落在该区间的频率, 所有小矩形面积之和为 1. 随着样本容量 n 越来越大, 分组可越来越多, 此时组距将越来越小, 变化刻度后的频率直方图顶部折线将趋于总体 X 的密度函数曲线, 而累积频率直方图顶部折线将趋于总体 X 的分布函数曲线.

从图 5.4.3 我们可初步推断总体 X ——灯泡使用寿命服从正态分布, 这就是从数据处理得出的总体分布的有关信息.

2. 经验分布函数

利用频率直方图或累积频率直方图去推断总体的分布类型要受到分组的影响, 因为不同的分组将导出不同的直方图. 下面给出另一种推断总体分布的方法.

定义 5.4.1 设总体 X 的分布函数为 $F(x)$ (未知), 其样本观测值为 $x_1, x_2, \dots, x_n$, 将它们从小到大排列成 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. 设其中互不相同的共 l 个, 分别为 $x_{(1)}^* < x_{(2)}^* < \dots < x_{(l)}^*$, 其个数分别为 $n_1, n_2, \dots, n_l$, $\sum_{i=1}^l n_i = n$. 令

$$F_n(x) = \begin{cases} 0, & x < x_{(1)}^*, \\ \frac{n_1 + n_2 + \dots + n_k}{n}, & x_{(k)}^* \leq x < x_{(k+1)}^*, \quad k = 1, 2, \dots, l-1, \\ 1, & x \geq x_{(l)}^*, \end{cases}$$

则称 $F_n(x)$ 为该样本的经验分布函数.

例 5.4.3 从某个总体 X 中抽取 10 个样本, 测得样本观测值为 $-4, 0, 1, -2, 5, 3, 3, 5, 6, 5$. 试求该样本的经验分布函数.

解 将样本观测值从小到大的顺序排列为

$-4 \quad -2 \quad 0 \quad 1 \quad 3 \quad 3 \quad 5 \quad 5 \quad 5 \quad 6$

从而其经验分布函数为

$$F_n(x) = \begin{cases} 0, & x < -4, \\ 0.1, & -4 \leq x < -2, \\ 0.2, & -2 \leq x < 0, \\ 0.3, & 0 \leq x < 1, \\ 0.4, & 1 \leq x < 3, \\ 0.6, & 3 \leq x < 5, \\ 0.9, & 5 \leq x < 6, \\ 1, & x \geq 6. \end{cases}$$

经验分布函数完全由样本观测值确定, 容易验证它具有分布函数的性质. 对于固定的 x , $F_n(x)$ 表示事件 $\{X \leq x\}$ 在 n 次独立试验中出现的频率. 根据第 4 章的大数定律知道, 在满足一定条件下, 事件发生的频率依概率收敛于这个事件发生的概率. 人们自然要问: 当试验次数 n 增大时, 样本的经验分布函数 $F_n(x)$ 是否接近于总体的分布函数呢? 格利文科于 1933 年作了肯定的回答.

定理 5.4.1 (格利文科定理) 设总体 ξ 的分布函数为 $F(x)$, 经验分布函数为 $F_n(x)$, 对于任何实数 x , 记

$$D_n = \sup_{-\infty < x < +\infty} |F_n(x) - F(x)|,$$

则有

$$P\left\{\lim_{n \rightarrow \infty} D_n = 0\right\} = 1.$$

格利文科定理证明了统计量 D_n 以概率为 1 地收敛于 0. 通俗地说, 就是当 n 足够大时, 对于所有的 x 值, $F_n(x)$ 与 $F(x)$ 之差的绝对值都很小这个事件发生的概率接近于 1.

§ 5.5 常用统计量的计算

在实际应用中, 样本数据少则几十个, 多则成千上万, 如果仅靠人工对常用统计量进行计算的话, 工作量将是十分巨大的. 因此用计算机对这些数据进行统计处理是非常必要的. 特别在当今信息社会里, 大量的样本数据通过计算机及网络来储存和传送, 这更为各种统计量的计算提供了很好的应用环境.

MATLAB 在其统计工具箱中提供了可以计算各种统计量的工具函数, 与本章所介绍内容相关的部分函数列于表 5.5.1 中, 下面介绍这些函数的使用方法.

表 5.5.1 常用统计量的计算函数

函数名	功 能	函数名	功 能
mean	均值	max	最大值
var	方差	min	最小值
std	标准差	range	极差 = 最大值 - 最小值
moment	中心矩	median	样本中值(中位数)
cov	协方差	prtile	分位数
corrcoef	相关系数		

以下设 $X = \{x_1, x_2, \dots, x_n\}$, $Y = \{y_1, y_2, \dots, y_n\}$ 都是由 n 个样本数据组成的向量.

1. 均值

函数 mean 用于计算样本均值

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i,$$

其调用格式为

$$x = \text{mean}(X).$$

2. 方差

函数 var 用于计算样本方差

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$$

其调用格式为

$$S2 = \text{var}(X).$$

3. 标准差

函数 `std` 用于计算样本标准差

$$S = \left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right]^{\frac{1}{2}},$$

其调用格式为

$$S = \text{std}(X).$$

4. 中心矩

函数 `moment` 用于计算 k 阶中心矩

$$B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k, \quad k = 1, 2, 3, \dots,$$

其调用格式为

$$Bk = \text{moment}(X, k).$$

在统计工具箱中没有给出计算原点矩的函数. 但对于 k 阶原点矩 $A_k =$

$\frac{1}{n} \sum_{i=1}^n X_i^k$, 用 MATLAB 的数组运算功能可以很容易地算得

$$Ak = \text{sum}(X.^k)/n.$$

5. 协方差

函数 `cov` 用于计算样本协方差

$$S_{XY}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}),$$

其调用格式为

$$S = \text{cov}(X, Y).$$

而返回值 S 是一个 2×2 矩阵, 其各元素的意义为

$$S = \begin{bmatrix} S_X^2 & S_{YX}^2 \\ S_{XY}^2 & S_Y^2 \end{bmatrix},$$

其中

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

6. 相关系数

函数 `corrcoef` 用于计算样本相关系数

$$\rho_{XY} = \frac{S_{XY}^2}{S_X \cdot S_Y},$$

其调用格式为

$$P = \text{corrcoef}(X, Y).$$

返回值 P 是一个 2×2 阶矩阵, 其各元素的意义为

$$P = \begin{bmatrix} \rho_{XX} & \rho_{YX} \\ \rho_{XY} & \rho_{YY} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{XY} \\ \rho_{XY} & 1 \end{bmatrix}.$$

7. 分位数

函数 `prctile` 的调用格式为

$$Y = \text{prctile}(X, p).$$

返回结果 Y 是 X 中大于 $p\%$ 的值, p 的取值区间是 $[0, 100]$.

函数 `max`, `min`, `range`, `median` 的调用格式为

$$Y = \text{max}(X),$$

$$Y = \text{min}(X),$$

$$Y = \text{range}(X),$$

$$Y = \text{median}(X).$$

例 5.5.1 设给定两组样本数据:

$$X = [24 \ 11 \ 6 \ 24 \ 28 \ 6 \ 7 \ 18 \ 10 \ 3 \ 23 \ 15]; \downarrow$$

$$Y = [21 \ 9 \ 17 \ 17 \ 23 \ 25 \ 2 \ 16 \ 25 \ 10 \ 12 \ 10]; \downarrow$$

用 MATLAB 计算其各统计量的操作命令及显示结果如下:

计算样本 X 的最大值、最小值、极差、中位数、均值:

`max(X)` ↓

`ans =`

28

`min(X)` ↓

`ans =`

3

`range(X)` ↓

`ans =`

25

`median(X)` ↓

`ans =`

13

`mean(X)` ↓

```
ans =
```

```
14.5833
```

计算 X 中大于 30% 的分位数:

```
prctile(X,30) ↓
```

```
ans =
```

```
7.3000
```

计算 X 的方差和标准差:

```
S2 = var(X) ↓
```

```
S2 =
```

```
73.9015
```

```
std(X) ↓
```

```
ans =
```

```
8.5966
```

计算 X 的 3 阶中心矩:

```
moment(X,3) ↓
```

```
ans =
```

```
110.3345
```

计算 X 的 4 阶原点矩:

```
A4 = sum(X.^4)/12 ↓
```

```
A4 =
```

```
1.4528e+005
```

计算 X 与 Y 的协方差矩阵 S 和相关系数矩阵 P :

```
S = cov(X,Y) ↓
```

```
S =
```

```
73.9015 17.0833
```

```
17.0833 51.7197
```

```
P = corrcoef(X,Y) ↓
```

```
P =
```

```
1.0000 0.2763
```

```
0.2763 1.0000
```

为了观察样本数据的分布状况,经常需要绘制样本数据的统计频数直方图. MATLAB 中的 `hist` 函数就是为此而设计的,其调用格式为

```
hist(X,m),
```

```
[n,a] = hist(X,m),
```

其中 X 是样本数据向量或矩阵. 该函数将区间 $[\min(X), \max(X)]$ 等分为 m

份(缺省时 m 设定为 10), 用直方图表示在这 m 个小区间上算得的 X 的频数函数. 如果 X 是矩阵, 则 `hist` 函数按照列的方向进行运算. 参数 m 也可以是向量, 此时 `hist` 绘制在以 m 的各个分量为中心的区间上计算得到的 X 的频数函数的直方图.

第二种调用格式 `[n,a] = hist(X,m)` 只计算不绘图, 式中的返回值 n 和 a 都是向量, n 返回 m 个小区间上的频数, a 返回 m 个小区间的中点.

例 5.5.2 用 `hist` 函数绘制本章例 5.4.2 中灯泡寿命样本数据的频数直方图.

解 首先输入数据:

```
X = [1450,1480,1640,1510,1680,1670,1700,1530,1640,1550,... ↵
1590,1480,1550,1590,1570,1680,1730,1800,1770,1420,... ↵
1370,1350,1610,1510,1630,1600,1490,1460,1810,1570,... ↵
1430,1660,1570,1560,1630,1740,1650,1720,1540,1550,... ↵
1730,1540,1670,1430,1490,1570,1580,1640,1660,1390]; ↵
```

计算最大值、最小值、极差:

```
Xmax = max(X), Xmin = min(X), R50 = range(X) ↵
```

```
Xmax =
```

```
1810
```

```
Xmin =
```

```
1350
```

```
R50 =
```

```
460
```

计算以 1 370, 1 440, 1 510, 1 580, 1 650, 1 720, 1 790 为中心的 7 个小区间上的频数 n_i :

```
m = [1370,1440,1510,1580,1650,1720,1790]; ↵
```

```
ni = hist(X,m) ↵
```

```
ni =
```

```
3 5 9 13 12 5 3
```

绘制 n_i 的直方图:

```
hist(X,m) ↵
```

得图 5.5.1.

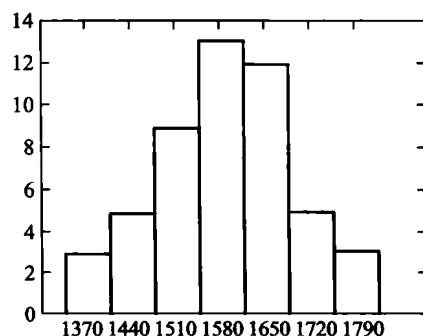


图 5.5.1 灯泡寿命的直方图

习 题 5

1. 设总体 $X \sim N(\mu, \sigma^2)$, 其中 μ 和 σ^2 都是未知参数, 随机变量 $X_1, X_2, \dots, X_n$ 是来自总体的样本.

(1) 写出样本 $(X_1, X_2, \dots, X_n)$ 的样本空间和联合密度函数;

(2) 指出下列样本的函数中哪些是统计量, 哪些不是统计量, 并说明原因:

$$T_1 = \frac{1}{n-1} \sum_{i=1}^n X_i, \quad T_2 = X_n - E(X_1), \quad T_3 = 2X_2 + X_3,$$

$$T_4 = \max(X_1, X_2, \dots, X_n), \quad T_5 = \frac{X_1 - \mu}{\sigma}, \quad T_6 = \sum_{i=1}^n \left(\frac{X_i}{\sigma} \right)^2.$$

2. 设 $X_1, X_2, \dots, X_n$ 是来自泊松分布 $P(\lambda)$ 的样本, 样本均值和方差分别为 $\bar{X}$ 和 S^2 , 试求 $E(\bar{X})$, $D(\bar{X})$ 和 $E(S^2)$.

3. 从总体 $X \sim N(10, 2^2)$ 中随机抽取容量为 5 的样本 $X_1, X_2, \dots, X_5$, 求:

(1) $P(|\bar{X} - 10| > 2)$;

(2) $P(\max(X_1, X_2, X_3, X_4, X_5) > 12)$;

(3) $P(\min(X_1, X_2, X_3, X_4, X_5) > 8)$.

4. 设 $X_1, X_2, \dots, X_8$ 是来自总体 $N(0, 0.2^2)$ 的样本, 试求 a , 使 $P\left(\sum_{i=1}^8 X_i^2 < a\right) = 0.95$.

5. 某厂生产的灯泡使用寿命(单位:h) $X \sim N(2500, 250^2)$. 现进行质量检查, 方法如下: 任意挑选若干个灯泡, 如果这些灯泡的平均寿命超过 2450, 就认为该厂生产的灯泡质量合格, 若要使检查能通过的概率超过 99%, 至少应检查多少个灯泡?

6. 设电子元件的寿命(单位:h) X 服从参数 $\lambda = 0.0015$ 的指数分布, 即密度函数 $f(x) = 0.0015e^{-0.0015x} (x > 0)$, 今独立测试 6 个元件, 并记录它们各自失效的时间, 试问:

(1) 至 800 h 时没有一个元件失效的概率是多少?

(2) 至 3000 h 时所有元件都失效的概率是多少?

7. 查表求下列分位数:

$$\chi_{0.9}^2(16), \chi_{0.95}^2(16), \chi_{0.10}^2(16), \chi_{0.05}^2(16), t_{0.95}(20), t_{0.05}(20),$$

$$F_{0.95}(4, 6), F_{0.05}(4, 6), u_{0.025}, u_{0.05}.$$

8. 证明 F 分布具有下列性质.

(1) 若 $F \sim F(m, n)$, 则 $1/F \sim F(n, m)$;

(2) $F_{1-\alpha}(m, n) = 1/F_{\alpha}(n, m)$.

9. 设 X_1, X_2, X_3, X_4 是来自总体 $N(0, 2^2)$ 的样本.

(1) 试求常数 C , 使 $Y = C[(X_1 - X_2)^2 + (X_3 + X_4)^2]$ 服从 χ^2 分布, 并指出其自由度是多少?

(2) 证明 $Z = \frac{(X_1 - X_2)^2}{(X_1 + X_2)^2}$ 服从 $F(1, 1)$.

10. 设 $X_1, X_2, \dots, X_n$ 是来自总体 $N(\mu, \sigma^2)$ 的样本, 又设 $X_{n+1} \sim N(\mu, \sigma^2)$ 且与 $X_1,$

$X_2, \dots, X_n$ 独立. 试求

$$Y = \frac{X_{n+1} - \bar{X}}{S} \sqrt{\frac{n}{n+1}}$$

的分布.

11. 设 $X_1, X_2, \dots, X_9$ 是来自正态总体 $X \sim N(\mu, \sigma^2)$ 的简单随机样本,

$$Y_1 = \frac{1}{6}(X_1 + \dots + X_6), \quad Y_2 = \frac{1}{3}(X_7 + X_8 + X_9),$$

$$S^2 = \frac{1}{2} \sum_{i=7}^9 (X_i - Y_2)^2, \quad Z = \frac{\sqrt{2}(Y_1 - Y_2)}{S}.$$

证明统计量 Z 服从自由度为 2 的 t 分布.

12. 设总体 X 服从正态分布 $N(\mu, \sigma^2)$, 从该总体中抽取简单随机样本 $X_1, X_2, \dots, X_{2n}$ ($n \geq 2$), 其样本均值为 $\bar{X} = \frac{1}{2n} \sum_{i=1}^{2n} X_i$, 求统计量 $Y = \sum_{i=1}^n (X_i + X_{n+i} - 2\bar{X})^2$ 的数学期望 $E(Y)$.

13. 设 $X_1, X_2, \dots, X_n$ 是来自总体 X 的样本, 给定任意常数 a 及任意非零常数 b , 证明:

(1) 对变换 $Y_i = X_i - a$, 有

$$\bar{X} = \bar{Y} + a, \quad S_X^2 = S_Y^2;$$

(2) 对变换 $Y_i = \frac{X_i - a}{b}$, 有

$$\bar{X} = b\bar{Y} + a, \quad S_X^2 = b^2 S_Y^2.$$

其中 $\bar{X}, \bar{Y}, S_X^2, S_Y^2$ 分别为 $X_1, X_2, \dots, X_n$ 及 $Y_1, Y_2, \dots, Y_n$ 的样本均值和样本方差.

14. 设 $-1, 2, 1, 3, 2, 1, 4, -2, 0, 4, 2.5, -2, 0.5, 1.5, 3, -0.5, 1.5, 1, 2, 3$ 是某总体 X 的 20 个样本观测值,

(1) 将数据分成 5 组, 画出频数、频率及累积频率直方图;

(2) 写出经验分布函数;

(3) 比较由累积频率直方图顶部折线形成的曲线与经验分布函数曲线是否一致, 并考虑何时二者一致?

15. 设 $X_1, X_2, \dots, X_m$ 和 $Y_1, Y_2, \dots, Y_n$ 分别是来自总体 $N(\mu_1, \sigma^2)$ 和 $N(\mu_2, \sigma^2)$ 的样本, 且它们相互独立, α 和 β 是两个已知常数. 试求

$$\frac{\alpha(\bar{X} - \mu_1) + \beta(\bar{Y} - \mu_2)}{\sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}} \cdot \sqrt{\frac{\alpha^2}{m} + \frac{\beta^2}{n}}}$$

的分布.

16*. 设 $X_1, X_2, \dots, X_n$ 是来自总体 X 的样本, 总体 X 的分布函数为 $F(x)$, 密度函数为 $f(x)$, 记 $Y = \max(X_1, X_2, \dots, X_n)$, $Z = \min(X_1, X_2, \dots, X_n)$. 试求 Y, Z 的密度函数及 (Y, Z) 的联合密度函数.

17*. 设 $X_1, X_2, \dots, X_n$ 是来自总体 $X \sim N(\mu, \sigma^2)$ 的样本, 令 $Y = \frac{1}{n} \sum_{i=1}^n |X_i - \mu|$, 试证:

$$E(Y) = \sqrt{\frac{2}{\pi}} \sigma, \quad D(Y) = \left(1 - \frac{2}{\pi}\right) \frac{\sigma^2}{n}.$$

第6章 参数估计和假设检验

参数估计和假设检验是统计推断的核心内容, 它们具有深刻的统计思想内涵, 是各种数理统计方法的基础.

§ 6.1 参数的点估计

1. 参数估计问题

数理统计的基本任务是以样本推断总体. 在很多情况下, 人们根据经验对总体的分布形式是可以事先确定的, 但对其中的某些参数(比如总体均值、方差等)并不清楚. 当这些未知参数确定后, 我们就对总体有了全面的认识. 参数估计就是根据样本对总体未知参数的取值进行估计的一种统计推断方法.

例 6.1.1 某砖厂生产的砖的抗断强度 X 服从正态分布 $N(\mu, \sigma^2)$, 其中 μ, σ^2 未知. 从中抽取 n 块砖进行检测, 得到它们的抗断强度分别为 $X_1, X_2, \dots, X_n$. 试估计 μ, σ^2 .

类似的例子非常多.

本节中, 假定 θ (可以是向量) 是总体的待估参数, Θ 表示参数空间—— θ 所有可能的取值构成的集合. $X_1, X_2, \dots, X_n$ 表示来自总体 X 的样本容量为 n 的样本, $x_1, x_2, \dots, x_n$ 是样本的一组观测值.

参数估计的形式有两种. 一种是点估计, 就是根据样本构造一个统计量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$, 然后用 $\hat{\theta}$ 去估计 θ . 称 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 为 θ 的估计量, 称 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 为 θ 的估计值. 另一种是区间估计, 就是根据样本构造两个统计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$, 且 $\hat{\theta}_1 < \hat{\theta}_2$, 然后用区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 估计 θ , 并对事件“区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 包含 θ ”发生的概率给予一定的保证.

本节将介绍获得点估计的两种常用方法——矩估计法和最大似然估计法, 最后给出评价估计量优劣的几个标准.

2. 矩估计

矩估计法是用样本的 k 阶原点矩作为总体的 k 阶原点矩的估计, 进而求解未知参数的一种估计方法. 其步骤如下:

设总体 X 的未知参数 $\theta = (\theta_1, \theta_2, \dots, \theta_m)$. 总体 X 的 k 阶原点矩 EX^k ($k \leq m$) 存在, 它们显然是 θ 的函数, 不妨记 $\tilde{A}_k(\theta_1, \theta_2, \dots, \theta_m) = EX^k$. 设样本 $X_1, X_2, \dots, X_n$ 的 k 阶原点矩为 A_k . 根据大数定律, EX^k 的一个合理的估计是样本

的 k 阶原点矩 A_k , 故有方程组

$$\begin{cases} \tilde{A}_1(\theta_1, \theta_2, \dots, \theta_m) = A_1, \\ \tilde{A}_2(\theta_1, \theta_2, \dots, \theta_m) = A_2, \\ \dots\dots\dots \\ \tilde{A}_m(\theta_1, \theta_2, \dots, \theta_m) = A_m. \end{cases}$$

若上述方程组有唯一的解

$$\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m).$$

这样得到的估计量 $(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m)$ 称为未知参数 $(\theta_1, \theta_2, \dots, \theta_m)$ 的矩估计量.

例 6.1.2 设总体 X 的均值 μ 和方差 σ^2 都存在, 它们是未知参数. $X_1, X_2, \dots, X_n$ 是总体的一个样本, 试求 μ 和 σ^2 的矩估计量.

解 由矩估计法得方程组

$$\begin{cases} \mu = EX = A_1 = \bar{X}, \\ \mu^2 + \sigma^2 = EX^2 = A_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 = \frac{n-1}{n} S^2 + \bar{X}^2. \end{cases}$$

求解方程组得

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{n-1}{n} S^2,$$

此即为 μ, σ^2 的矩估计量.

上述结果表明: 不论总体服从什么分布, 其均值的矩估计量就是样本均值、方差的矩估计量与样本方差相差一个因子 $\frac{n-1}{n}$.

例 6.1.3 设总体 X 的概率密度为

$$f(x) = \begin{cases} (\theta+1)x^\theta, & 0 < x < 1, \\ 0, & \text{其它,} \end{cases}$$

其中 $\theta > -1$ 是未知参数. $X_1, X_2, \dots, X_n$ 是总体 X 的一组样本, 试求 θ 的矩估计量.

解 $EX = \int_{-\infty}^{+\infty} xf(x) dx = \int_0^1 x(\theta+1)x^\theta dx = \frac{\theta+1}{\theta+2},$

$$A_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$

由矩估计法得方程

$$\frac{\theta+1}{\theta+2} = \bar{X},$$

解得 θ 的矩估计量为

$$\hat{\theta} = \frac{2\bar{X} - 1}{1 - \bar{X}}.$$

例 6.1.4 设样本 $X_1, X_2, \dots, X_n$ 来自服从几何分布的总体 X , 其概率分布为

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

其中 p 未知, $0 < p < 1$, 试求 p 的矩估计量.

解 首先求出总体 X 的期望

$$\begin{aligned} E(X) &= \sum_{k=1}^{\infty} kp(1-p)^{k-1} = -p \sum_{k=1}^{\infty} [(1-p)^k]' \\ &= -p \left[\sum_{k=1}^{\infty} (1-p)^k \right]' \\ &= -p \left(\frac{1-p}{p} \right)' = \frac{1}{p}, \end{aligned}$$

从而 $\frac{1}{\hat{p}} = \bar{X}$, 所以 p 的矩估计量为 $\hat{p} = \frac{1}{\bar{X}}$.

矩估计法是一种古老的估计方法. 这种方法只要求知道总体的适当阶矩存在, 并不要求知道总体的分布类型, 因而它有广泛的应用. 但是矩估计法有两点不足: 一是求解所要求的总体阶矩可能不存在, 二是求解方程组可能很困难.

3. 最大似然估计

最大似然估计法是参数估计的另一种常用方法, 下面通过一个例子来说明这一估计法思想.

例 6.1.5 设盒内黑球和白球个数之比为 9:1, 但不知何种球多. 今有放回地抽取三次, 每次抽一球, 发现第一、三次抽到白球, 第二次抽到黑球. 试判断何种球多.

不妨设 θ 表示盒中白球所占比例, 则 θ 要么是 0.9, 要么是 0.1. 要判断何种球多, 就是判断 $\theta=0.9$ 还是 $\theta=0.1$. 上述问题可重新描述为: 令 $X=0$ 表示抽到黑球, $X=1$ 表示抽到白球. 一次抽球的结果 X 服从两点分布, 其分布列为

X	0	1
P	$1-\theta$	θ

其中 θ 表示白球所占比例, $\theta=0.9$ 或者 0.1. 有放回地抽三次球, 可视为从总体 X 中抽取了容量为 3 的样本, 其观测值分别为 $X_1=1, X_2=0, X_3=1$, 即观测的结果是事件 $A=\{X_1=1, X_2=0, X_3=1\}$ 发生了, 要判断 $\theta=0.9$ 还是 0.1.

显然, 不论 $\theta=0.9$ 还是 0.1, 事件 A 都可能发生. 下面考察 θ 取不同值时, 事件 A 发生的概率. 当 $\theta=0.9$ 时, $P(A)=0.9^2 \times 0.1=0.081$; 当 $\theta=0.1$

时, $P(A) = 0.1^2 \times 0.9 = 0.009$, 而 $0.081 > 0.009$. 这说明 θ 取不同的值, 事件 A (观测结果) 发生的概率是不一样的. 现在事件 A 已经发生, 因而试验的条件应当有利于 A 发生 (其概率相对比较大), 故应判断 $\theta = 0.9$, 也就是盒中白球多.

上述解决问题的思路可描述为: 设总体 X 含有待估参数 $\theta = (\theta_1, \theta_2, \dots, \theta_m)$, 参数空间为 Θ . 我们要在 Θ 中选取一个 $\hat{\theta}$, 使得 $\theta = \hat{\theta}$ 时, 样本观测结果即事件 $\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}$ 出现的概率 $L(\theta)$ 达到最大值, 我们称 $\hat{\theta}$ 为 θ 的最大似然估计. 这就是最大似然估计法的基本思想.

求最大似然估计关键有两步: 一是 $L(\theta)$ 的表达式, 二是 $L(\theta)$ 的极值解. 下面分别进行讨论.

(1) 关于 $L(\theta)$ 的表达式

若总体 X 是离散型的, 其分布概率为 $P(X = a_i) = p(a_i; \theta)$, $i = 1, 2, \dots$, $\theta \in \Theta$, 则

$$\begin{aligned} L(\theta) &= P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= \prod_{i=1}^n P(X_i = x_i) = \prod_{i=1}^n p(x_i; \theta). \end{aligned}$$

若总体 X 是连续型的, 其概率密度函数为 $f(x; \theta)$. 由于 $P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = 0$. 故考察样本 $X_1, X_2, \dots, X_n$ 落在观测值 $x_1, x_2, \dots, x_n$ 某邻域内的概率, 即事件 $\{x_1 \leq X_1 \leq x_1 + dx_1, \dots, x_n \leq X_n \leq x_n + dx_n\}$ 发生的概率. 此概率近似等于 $\prod_{i=1}^n f(x_i; \theta) dx_i$, 我们的目的是选取 $\theta \in \Theta$ 使这一概率最大. 这等价于选取 θ 使 $\prod_{i=1}^n f(x_i; \theta)$ 最大, 故

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta).$$

称 $L(\theta)$ 为样本的似然函数.

(2) 关于 $L(\theta)$ 的极值解

求 $L(\theta)$ 的最大值常需要求导. 为求导方便, 一般对似然函数求对数 $\ln L(\theta)$ ——称之为对数似然函数. 由于 $\ln x$ 是增函数, 故 $L(\theta)$ 和 $\ln L(\theta)$ 在同一点 $\theta = \hat{\theta}$ 上达到最大值. 由微积分知, θ 的最大似然估计 $\hat{\theta}$ 可由似然方程 (组)

$$\frac{\partial \ln L(\theta)}{\partial \theta_i} = 0, \quad i = 1, 2, \dots, m$$

注: 求解似然方程 (组) 得到 $\hat{\theta}$ 后, 还须验证 $L(\theta)$ 在 $\hat{\theta}$ 处是否取最大值. 另外, 有时 $L(\theta)$ 不是 θ 的连续可导函数, 有时参数空间是有界区域, 此时不

能用求解似然方程(组)的方法,一般利用定义进行判断分析求解.

例 6.1.6 试求例 6.1.3 中未知参数 θ 的最大似然估计.

解 似然函数

$$L(\theta) = \prod_{i=1}^n [(\theta+1)x_i^\theta] = (\theta+1)^n \left(\prod_{i=1}^n x_i \right)^\theta,$$

两边取对数得

$$\ln L(\theta) = n \ln(\theta+1) + \theta \sum_{i=1}^n \ln x_i,$$

故似然方程为

$$\frac{d \ln L(\theta)}{d\theta} = \frac{n}{\theta+1} + \sum_{i=1}^n \ln x_i = 0,$$

解得

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^n \ln x_i} - 1.$$

由于

$$\left. \frac{d^2 \ln L(\theta)}{d\theta^2} \right|_{\theta=\hat{\theta}} = -\frac{\left(\sum_{i=1}^n \ln x_i \right)^2}{n} < 0,$$

所以 $\ln L(\theta)$ 在 $\hat{\theta}$ 处取最大值, 因而 $\hat{\theta}$ 是 θ 的最大似然估计.

这个例子说明矩估计与最大似然估计有时并不一致.

例 6.1.7 设总体 X 服从正态分布

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

其中 μ, σ^2 是未知参数, $x_1, \dots, x_n$ 为样本的一组观测值, 试求 μ 和 σ^2 的最大似然估计值.

解 似然函数

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2}.$$

两边取对数得

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2,$$

所以似然方程组为

$$\begin{cases} \frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} = 0, \\ \frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^4} = 0, \end{cases}$$

解得

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n-1}{n} s^2.$$

经验证 $L(\mu, \sigma^2)$ 在 $(\hat{\mu}, \hat{\sigma}^2)$ 处达到最大值, 故 $\hat{\mu}$, $\hat{\sigma}^2$ 分别是未知参数 μ 和 σ^2 的最大似然估计值.

例 6.1.8 设某种元件的使用寿命 X 的概率密度函数为

$$f(x; \theta) = \begin{cases} 2e^{-2(x-\theta)}, & x \geq \theta, \\ 0, & x < \theta, \end{cases}$$

其中 $\theta > 0$ 为未知参数. 又设 $x_1, x_2, \dots, x_n$ 是 X 的一组样本观测值, 求参数 θ 的最大似然估计值.

解 似然函数为

$$L(\theta) = L(x_1, \dots, x_n; \theta) = \begin{cases} 2^n \exp\left[-2 \sum_{i=1}^n (x_i - \theta)\right], & x_i \geq \theta, \\ 0, & \text{其它.} \end{cases}$$

当 $x_i \geq \theta (i=1, 2, \dots, n)$ 时, $L(\theta) > 0$, 取对数得

$$\ln L(\theta) = n \ln 2 - 2 \sum_{i=1}^n (x_i - \theta),$$

因为 $\frac{d \ln L(\theta)}{d\theta} = 2n > 0$, 所以 $L(\theta)$ 单调增加. 由于 θ 必须满足 $\theta \leq x_i (i=1, 2, \dots, n)$,

因此, 当 θ 取 $x_1, x_2, \dots, x_n$ 中的最小值时, $L(\theta)$ 取最大值, 所以 θ 的最大似然估计值为

$$\hat{\theta} = \min(x_1, x_2, \dots, x_n).$$

例 6.1.9 设 $X_1, X_2, \dots, X_n$ 是来自参数为 λ 的泊松分布总体的一个样本, 试求 λ 的最大似然估计量.

解 设 $x_1, x_2, \dots, x_n$ 是相应于样本 $X_1, X_2, \dots, X_n$ 的一组观测值, 则似然函数为

$$L = \prod_{i=1}^n \left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right) = e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i} / \prod_{i=1}^n x_i!,$$

$$\ln L = -n\lambda + \left(\sum_{i=1}^n x_i \right) \ln \lambda - \ln \prod_{i=1}^n x_i!.$$

令

$$\frac{d}{d\lambda} \ln L = -n + \left(\sum_{i=1}^n x_i \right) / \lambda = 0,$$

得 λ 的最大似然估计值为 $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$, 最大似然估计量为 $\hat{\lambda} = \bar{X}$.

4. 估计量优劣的评价标准

参数的点估计是构造一个统计量作为参数取值的估计, 矩估计法和最大似然估计法只是构造参数统计量的两种常用方法. 例 6.1.3 中 θ 的矩估计量是

$$\hat{\theta}_1 = \frac{2\bar{X} - 1}{1 - \bar{X}}, \text{ 而其最大似然估计量是 } \hat{\theta}_2 = -\frac{n}{\sum_{i=1}^n \ln X_i} - 1, \text{ 自然还可构造出其}$$

它统计量去估计 θ . 那么在这些估计量中, 哪一个更好? 为回答这一问题, 必须给出评价估计量好坏的标准. 也就是说, 估计量具有什么“特点”我们认为它“好”. 下面给出几个常用的评价标准.

(1) 无偏性

未知参数 θ 的估计量 $\hat{\theta}$ 是一个随机变量, $\hat{\theta}$ 的取值由样本观测值确定, 具有随机性, 因此, 用 $\hat{\theta}$ 去估计 θ 一般会有偏差. 我们自然希望 $\hat{\theta}$ 的取值总是在 θ 附近左右摆动, 不应该总是偏大或总是偏小. 也就是从平均意义上讲, 希望 $E\hat{\theta}$ 与 θ 越接近越好. 当其差为 0 时, 便有了无偏性的概念.

定义 6.1.1 设 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 是未知参数 θ 的估计量. 若 $E(\hat{\theta}) = \theta$, 则称 $\hat{\theta}$ 是 θ 的无偏估计, 否则称为有偏估计. 若 $\lim_{n \rightarrow \infty} E(\hat{\theta}) = \theta$, 则称 $\hat{\theta}$ 是 θ 的渐近无偏估计.

在例 6.1.2 中, 总体 X 的均值 μ 和方差 σ^2 的矩估计量分别为 $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = \frac{n-1}{n}S^2$. 由于

$$\begin{aligned} E(\hat{\mu}) &= E(\bar{X}) = \mu, \\ E(\hat{\sigma}^2) &= E\left(\frac{n-1}{n}S^2\right) = E\left[\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right] \\ &= \frac{1}{n}\sum_{i=1}^n E(X_i^2) - E(\bar{X}^2) \\ &= \frac{1}{n}\sum_{i=1}^n (\mu^2 + \sigma^2) - \left(\mu^2 + \frac{\sigma^2}{n}\right). \\ &= \frac{n-1}{n}\sigma^2, \end{aligned}$$

所以 $\hat{\mu} = \bar{X}$ 是 μ 的无偏估计, $\hat{\sigma}^2 = \frac{n-1}{n}S^2$ 不是 σ^2 的无偏估计. 由于

$$E(S^2) = E\left(\frac{n}{n-1}\hat{\sigma}^2\right) = \frac{n}{n-1} \cdot E(\hat{\sigma}^2) = \sigma^2,$$

故 S^2 是 σ^2 的无偏估计. 又因为

$$\lim_{n \rightarrow \infty} E(\hat{\sigma}^2) = \lim_{n \rightarrow \infty} E\left(\frac{n-1}{n}S^2\right) = \lim_{n \rightarrow \infty} \frac{n-1}{n}\sigma^2 = \sigma^2,$$

因而 $\hat{\sigma}^2$ 是 σ^2 的渐近无偏估计.

上述结论与总体的分布类型没有关系. 只要总体均值存在, 样本均值总是它的无偏估计; 只要总体方差存在, 样本方差总是其无偏估计.

例 6.1.10 设随机变量 $X_1, X_2, \dots, X_n$ 是来自正态分布 $N(\mu, \sigma^2)$ 的一个样本, 适当选择常数 c , 使 $c \sum_{i=1}^{n-1} (X_{i+1} - X_i)^2$ 为 σ^2 的无偏估计.

解 令 $Y_i = X_{i+1} - X_i$, 则 $Y_i \sim N(0, 2\sigma^2)$, 由

$$\begin{aligned} E\left(c \sum_{i=1}^{n-1} Y_i^2\right) &= c \sum_{i=1}^{n-1} E(Y_i^2) = c(n-1) \{D(Y_i) + [E(Y_i)]^2\} \\ &= c(n-1)2\sigma^2 = \sigma^2, \end{aligned}$$

因此

$$c = \frac{1}{2(n-1)}.$$

在实际应用中, 要求估计量具有无偏性是有实际意义的. 例如在大批商品的买卖中, 买卖双方通常通过抽样去估计产品次品率. 若估计值高于实际值, 将给卖者带来损失. 反之, 若估计值低于实际值, 就会给买者带来损失. 但只要采用的估计量是无偏估计, 而双方的买卖是长期的, 则总的来说是互不吃亏.

(2) 有效性

具有无偏性只是对“好”估计的基本要求, 对同一待估参数往往有很多无偏估计. 例如对于任意常数 $a_1, a_2, \dots, a_n$, 只要 $\sum_{i=1}^n a_i = 1$, 则 $\sum_{i=1}^n a_i X_i$ 都是总体均值 μ 的无偏估计. 因此, 必须给出另外的标准以便在众多的无偏估计中选出“最优”者.

对于未知参数 θ 的两个无偏估计 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 而言, 显然与 θ 偏离程度小的估计要更“好”, 而无偏估计 $\hat{\theta}$ 的方差恰好反映了估计 $\hat{\theta}$ 与 θ 的偏离程度, 由此便有了有效性的概念.

定义 6.1.2 设 $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \dots, X_n)$, $\hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \dots, X_n)$ 都是参数 θ 的无偏估计, 如果 $D(\hat{\theta}_1) < D(\hat{\theta}_2)$, 则称估计量 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效.

进一步, 若对于 θ 的任一无偏估计 $\hat{\theta}$, 有 $D(\hat{\theta}_1) \leq D(\hat{\theta})$, 则称 $\hat{\theta}_1$ 是 θ 的最小方差无偏估计.

例如, 设 X_1, X_2, X_3 是来自总体 X 容量为 3 的样本, 总体均值 μ 未知. 易验证 $\hat{\theta}_1 = \frac{1}{3}(X_1 + X_2 + X_3)$ 和 $\hat{\theta}_2 = \frac{1}{2}X_1 + \frac{1}{3}X_2 + \frac{1}{6}X_3$ 都是 μ 的无偏估计. 但 $D(\hat{\theta}_1) = \frac{1}{3}D(X) < D(\hat{\theta}_2) = \frac{7}{18}D(X)$, 故 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效. 可以证明(习题 6 第 8

题): μ 的所有形如 $\sum_{i=1}^n c_i X_i$ (其中 $\sum_{i=1}^n c_i = 1$) 的无偏估计中, $\bar{X}$ 的方差最小.

例 6.1.11 设 Y_1 和 Y_2 是 θ 的两个相互独立的无偏估计, 若 $D(Y_1) = 2D(Y_2)$, 求常数 a, b , 使得 $aY_1 + bY_2$ 为此种线性组合中有最小方差的无偏估计.

解 因为 $E(Y_1) = \theta, E(Y_2) = \theta$, 所以

$$E(aY_1 + bY_2) = (a+b)\theta = \theta,$$

可知 $a+b=1$. 又

$$\begin{aligned} D(aY_1 + bY_2) &= a^2 D(Y_1) + b^2 D(Y_2) \quad (\text{因为 } Y_1, Y_2 \text{ 相互独立}) \\ &= 2a^2 D(Y_2) + b^2 D(Y_2) \\ &= [2a^2 + (1-a)^2] D(Y_2) \\ &= (3a^2 - 2a + 1) D(Y_2), \end{aligned}$$

令 $\frac{dD}{da} = 0$, 即 $6a - 2 = 0$, 则 $a = \frac{1}{3}, b = \frac{2}{3}$. 因为 $\frac{d^2 D}{da^2} = 6 > 0$. 所以 $a = \frac{1}{3}, b = \frac{2}{3}$ 时, 即 $\frac{1}{3}Y_1 + \frac{2}{3}Y_2$ 为 θ 的最小方差无偏估计.

(3)* 相合性(一致性)

无偏性和有效性都是在样本容量 n 固定的前提下提出的, 随着试验次数 n 的不断增加, 样本所包含的信息量也逐步增加. 因此, 一个“好”的估计在样本容量 n 增加时, 必须越来越接近于参数的真值, 由此引出了相合性的概念.

定义 6.1.3 设对每一个自然数 n , 统计量 $\hat{\theta}_n(X_1, X_2, \dots, X_n)$ 都是未知参数 θ 的一个估计量, 若对任意的 $\varepsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0,$$

则称 $\hat{\theta}_n$ 是 θ 的相合(一致)估计.

定理 6.1.1 设 $\hat{\theta}_n$ 是 θ 的一个估计量, 若 $\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta, \lim_{n \rightarrow \infty} D(\hat{\theta}_n) = 0$, 则 $\hat{\theta}_n$ 是 θ 的相合估计.

证 对任给的 $\varepsilon > 0$, 有

$$P(|\hat{\theta}_n - \theta| \geq \varepsilon) = P((\hat{\theta}_n - \theta)^2 \geq \varepsilon^2) \leq \frac{E(\hat{\theta}_n - \theta)^2}{\varepsilon^2} \quad (\text{易证}).$$

又知 $D(\hat{\theta}_n) = D(\hat{\theta}_n - \theta) = E(\hat{\theta}_n - \theta)^2 - [E(\hat{\theta}_n - \theta)]^2$, 得

$$P(|\hat{\theta}_n - \theta| \geq \varepsilon) \leq \frac{D(\hat{\theta}_n) + [E(\hat{\theta}_n - \theta)]^2}{\varepsilon^2} = \frac{D(\hat{\theta}_n) + [E(\hat{\theta}_n) - \theta]^2}{\varepsilon^2},$$

而当 $n \rightarrow \infty$ 时, $E(\hat{\theta}_n) \rightarrow \theta, D(\hat{\theta}_n) \rightarrow 0$, 故

$$P(|\hat{\theta}_n - \theta| \geq \varepsilon) \rightarrow 0,$$

因而 $\hat{\theta}_n$ 是 θ 的相合估计.

由上述定理知, $\bar{X}$ 和 S^2 分别是总体均值 μ 和方差 σ^2 的相合估计.

§ 6.2 参数的区间估计

当样本观测值给定以后, 点估计能给出未知参数 θ 一个确定的数值. 但估计值只是 θ 的一个近似值, 究竟它与 θ 的真值有没有误差, 误差是多少并不知道. 而在实际问题中, 这种误差的大小往往是人们比较关心的. 例如通过产品抽样对废品率进行估计, 估计误差达到 1% 就可能对交易的某一方带来重大损失. 为此, 引入了估计的另一种形式——区间估计.

定义 6.2.1 设 $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \dots, X_n)$, $\hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \dots, X_n)$ 是两个统计量, 且 $\hat{\theta}_1 < \hat{\theta}_2$, θ 是总体 X 的未知参数, 若对给定的 $\alpha \in (0, 1)$, 有

$$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha,$$

则称随机区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 是 θ 的一个区间估计或置信区间. $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 分别称为置信下限、置信上限, $1 - \alpha$ 称为置信水平或置信度.

这里需要指出的是: 区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 是随机区间, 不同的样本观测值就得到不同的区间 $[\hat{\theta}_1, \hat{\theta}_2]$. 当样本观测值给定后, 区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 可能包含 θ 的真值, 也可能不包含 θ 的真值, 而 $1 - \alpha$ 给出了随机区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 包含真值 θ 的可信程度. 例如若 $\alpha = 5\%$, 则置信度为 95%, 这表明若重复抽样 100 次, 将得到 100 个不同的区间, 其中约有 95 个区间包含了真值 θ , 不包含真值 θ 的区间约有 5 个.

例 6.2.1 设 $X_1, X_2, \dots, X_n$ 是总体 $X \sim N(\mu, \sigma^2)$ 的样本. 试讨论下列四类问题:

- (1) 若 σ^2 已知, μ 未知, 试求参数 μ 的置信水平为 $1 - \alpha$ 的置信区间;
- (2) 若 σ^2 未知, μ 未知, 试求参数 μ 的置信水平为 $1 - \alpha$ 的置信区间;
- (3) 若 μ 已知, σ^2 未知, 试求参数 σ^2 的置信水平为 $1 - \alpha$ 的置信区间;
- (4) 若 μ 未知, σ^2 未知, 试求参数 σ^2 的置信水平为 $1 - \alpha$ 的置信区间.

上述问题是单个正态总体的参数(均值、方差)区间估计问题, 今以情况(1)为例说明如何求区间估计.

解 μ 的一个点估计是 $\bar{X}$, 由定理 5.3.1 知,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

根据分位数的定义有

$$P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq u_{\alpha/2}\right) = 1 - \alpha,$$

所以

$$P\left(\bar{X} - u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

从而参数 μ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left[\bar{X} - u_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right].$$

由这个例子可给出求置信区间的一般方法, 总结为以下几条:

(1) 找一个与要估计的参数 θ 有关的统计量 T , 一般是 θ 的一良好点估计 $\hat{\theta}$ (此例中 T 为样本均值).

(2) 设法找出 T 和 θ 的某一函数 $H(T, \theta)$, 要求 $H(T, \theta)$ 的分布已知且与 θ 无关 (此例中 $H(T, \theta) = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$, 其分布为 $N(0, 1)$), H 称为枢轴变量.

(3) 寻找适当的常数 c, d , 使

$$P(c \leq H(T, \theta) \leq d) = 1 - \alpha.$$

(4) 将不等式 $c \leq H(T, \theta) \leq d$ 等价变形为

$$\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2,$$

所以

$$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha.$$

故随机区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 即为参数 θ 的置信水平为 $1 - \alpha$ 的置信区间.

例 6.2.2 设总体 $X \sim N(\mu, \sigma^2)$, $X_1, X_2, \dots, X_n$ 为 X 的简单随机样本, μ, σ^2 均未知, 求 σ^2 的置信区间.

解 方差 σ^2 的一个良好点估计为 S^2 , 因

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

于是对于给定的 α , 由

$$P\left(\chi_{1-\alpha/2}^2(n-1) \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\alpha/2}^2(n-1)\right) = 1 - \alpha,$$

即

$$P\left[\frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)}\right] = 1 - \alpha,$$

于是方差 σ^2 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left[\frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)}\right].$$

求置信区间的关键是选取枢轴变量 $H(T, \theta)$. 利用上述方法可讨论例 6.2.1 的其它置信区间问题, 也可讨论两个正态总体均值差、方差比的置信区

间问题. 它们用到的枢轴变量 $H(T, \theta)$ 分别见定理 5.3.1 和定理 5.3.2, 列表如下.

表 6.2.1 单个正态总体的参数区间估计表 (置信水平为 $1 - \alpha$)

待估参数	条件	枢轴变量及其分布	置 信 区 间
均值 μ	方差 σ^2 已知	$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$	$\left[\bar{X} - u_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$
	方差 σ^2 未知	$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$	$\left[\bar{X} - t_{\alpha/2}(n-1) \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1) \frac{S}{\sqrt{n}} \right]$
方差 σ^2	均值 μ 已知	$\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \sim \chi^2(n)$	$\left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{\alpha/2}^2(n)}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{1-\alpha/2}^2(n)} \right]$
	均值 μ 未知	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left[\frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)} \right]$

表 6.2.2 两个正态总体均值差、方差比的区间估计表 (置信水平为 $1 - \alpha$)

待估参数	条件	枢轴变量及其分布	置 信 区 间
$\mu_1 - \mu_2$	σ_1^2, σ_2^2 已知	$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0, 1)$	$\left[\bar{X} - \bar{Y} - u_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}, \bar{X} - \bar{Y} + u_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \right]$
	σ_1^2, σ_2^2 未知, 但 $\sigma_1^2 = \sigma_2^2$	$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2)$	$\left[\bar{X} - \bar{Y} - t_{\alpha/2}(m+n-2) S_w \sqrt{\frac{1}{m} + \frac{1}{n}}, \bar{X} - \bar{Y} + t_{\alpha/2}(m+n-2) S_w \sqrt{\frac{1}{m} + \frac{1}{n}} \right]$
$\frac{\sigma_1^2}{\sigma_2^2}$	μ_1, μ_2 已知	$\frac{\frac{1}{m} \sum_{i=1}^m (X_i - \mu_1)^2 / \sigma_1^2}{\frac{1}{n} \sum_{i=1}^n (Y_i - \mu_2)^2 / \sigma_2^2} \sim F(m, n)$	$\left[\frac{A}{F_{\alpha/2}(m, n)}, \frac{A}{F_{1-\alpha/2}(m, n)} \right]$
	μ_1, μ_2 未知	$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F(m-1, n-1)$	$\left[\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{\alpha/2}(m-1, n-1)}, \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{1-\alpha/2}(m-1, n-1)} \right]$

注: 在表 6.2.2 中, 符号 $X_1, X_2, \dots, X_m$ 表示来自总体 $X \sim N(\mu_1, \sigma_1^2)$ 的样本, $Y_1, Y_2, \dots, Y_n$ 表示来自总体 $Y \sim N(\mu_2, \sigma_2^2)$ 的样本, 两样本假定相互独立. $\bar{X}$, S_1^2 及 $\bar{Y}$, S_2^2 分别表示两样本的均值和方差,

$$S_w^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}, \quad A = \frac{\frac{1}{m} \sum_{i=1}^m (X_i - \mu_1)^2}{\frac{1}{n} \sum_{i=1}^n (Y_i - \mu_2)^2}.$$

最后指出, 置信区间的长度描述了估计的“精确性”. 置信区间的长度越小, 则估计越精确. 置信区间的长度越大, 则估计越不精确. 例如, 若给出灯泡平均寿命的置信水平为 99% 的置信区间是 (100, 5 000), 这种区间估计没有多大的实际意义. 因为寿命比较低的 (比如低于 2 000) 是不合格品, 寿命比较高的是合格品, 而区间估计 (100, 5 000) 并没有提供多少有关灯泡质量的信息. 反之, 若给出的置信区间为 (2 100, 2 500), 则我们对灯泡的质量状况就比较了解了, 表明这是一批质量相当高的灯泡. 一般说来, 当样本容量 n 固定后, 置信水平要求越高, 则精确性 (置信区间长度) 就越差. 处理的方法是: 固定置信水平, 精确性尽可能地高. 当然, 可通过增加样本容量 n , 使得置信水平和精确性都能达到满意的要求.

例 6.2.3 某厂家生产的灯泡寿命近似服从正态分布, 标准差 $\sigma = 40$, 若 25 个灯泡的样本平均寿命为 1 180 h,

- (1) 求此厂家生产的所有灯泡总体均值 μ 的 0.95 置信区间;
- (2) 若希望误差 $e = |\bar{x} - \mu| < 10$ h, 需要多大的样本?

解 (1) 因为标准差 σ 已知, 所以求 μ 的置信区间可用 $U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ 作为枢轴变量, $U \sim N(0, 1)$, 即对于所给的置信水平 $1 - \alpha = 0.95$, $\alpha = 0.05$, 由

$$P\left(-u_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq u_{\alpha/2}\right) = 1 - \alpha$$

得 μ 的置信区间为

$$\left[\bar{X} - u_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + u_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right].$$

查正态分布表得 $u_{\alpha/2} = u_{0.025} = 1.96$, 代入数据得置信区间为

$$\left[1\,180 - 1.96 \times \frac{40}{\sqrt{25}}, 1\,180 + 1.96 \times \frac{40}{\sqrt{25}}\right],$$

即

$$[1\,164.32, 1\,195.68].$$

(2) 若希望误差 $e = |\bar{X} - \mu| < 10$ h, 只要 $u_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < 10$, 即

$$n > \left(\sigma \cdot \frac{u_{\alpha/2}}{10} \right)^2 = \left(40 \times \frac{1.96}{10} \right)^2 = 61.4656,$$

所以若以 0.95 的置信水平使误差 $e < 10$ h, 需样本大小为 62 个灯泡.

例 6.2.4 已知某种木材的横纹抗压力 (单位: kg/cm²) 服从 $N(\mu, \sigma^2)$, 现对 10 个试件作横纹抗压力试验, 得数据如下:

482 493 457 471 510 446 435 418 394 469

(1) 求 μ 的置信水平为 0.95 的置信区间;

(2) 求 σ 的置信水平为 0.90 的置信区间.

解 (1) 根据样本计算可得 $\bar{x} = 457.5$, $s^2 = 1240.28$. 则 $s = 35.22$, 由于 σ 未知, 因此利用 $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ 作为枢轴变量, $T \sim t(n-1)$, 由表 6.2.1 知, 均值 μ 的置信区间为

$$\left[\bar{X} - t_{\alpha/2}(n-1) \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1) \frac{S}{\sqrt{n}} \right].$$

代入数据得置信区间为

$$\left[457.5 - 2.2622 \times \frac{35.22}{\sqrt{10}}, 457.5 + 2.2622 \times \frac{35.22}{\sqrt{10}} \right],$$

即

$$[432.3, 482.7].$$

(2) 由于均值 μ 未知, 因此利用 $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ 作为枢轴变量, $\chi^2 \sim \chi^2(n-1)$,

由表 6.2.1 知, 标准差 σ 的置信区间为

$$\left[\sqrt{\frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)}}, \sqrt{\frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)}} \right].$$

查表得 $\chi_{0.05}^2(9) = 16.919$, $\chi_{0.95}^2(9) = 3.325$, 代入数据得置信区间为

$$[25.69, 57.94].$$

例 6.2.5 两台机床生产同一型号的滚珠, 从甲机床生产的滚珠中抽取 8 个, 从乙机床生产的滚珠中抽取 9 个, 测得这些滚珠的直径 (单位: mm) 如下:

甲机床: 15.0 14.8 15.2 15.4 14.9 15.1 15.2 14.8

乙机床: 15.2 15.0 14.8 15.1 15.0 14.6 14.8 15.1 14.5

设两台机床生产的滚珠直径服从正态分布, 如果

(1) 已知两台机床生产的滚珠直径的标准差分别是 $\sigma_1 = 0.18$, $\sigma_2 = 0.24$;

(2) 未知 σ_1 及 σ_2 , 但假设 $\sigma_1 = \sigma_2$,

求这两台机床生产的滚珠直径差 $\mu_1 - \mu_2$ 的置信水平为 0.90 的置信区间.

解 (1) 根据样本计算可得,

$$m=8, \bar{x}=15.05, s_1^2=0.0457;$$

$$n=9, \bar{y}=14.9, s_2^2=0.0575.$$

由已知 $\sigma_1=0.18, \sigma_2=0.24$, 因此利用 $U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$ 作为枢轴变

量, $U \sim N(0, 1)$, 由表 6.2.2 知, $\mu_1 - \mu_2$ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left[\bar{X} - \bar{Y} - u_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}, \bar{X} - \bar{Y} + u_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \right].$$

查表得 $u_{\alpha/2} = u_{0.05} = 1.645$, 代入数据得置信区间为

$$[-0.018, 0.318]$$

(2) 由于 σ_1, σ_2 未知, 且 $\sigma_1 = \sigma_2$, 因此利用 $T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{m} + \frac{1}{n}}}$ 作

为枢轴变量, $T \sim t(m+n-2)$, 由表 6.2.2 知; $\mu_1 - \mu_2$ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left[\bar{X} - \bar{Y} - t_{\alpha/2}(m+n-2) S_w \sqrt{\frac{1}{m} + \frac{1}{n}}, \bar{X} - \bar{Y} + t_{\alpha/2}(m+n-2) S_w \sqrt{\frac{1}{m} + \frac{1}{n}} \right].$$

查表得 $t_{0.05}(15) = 1.7531$, 代入数据得置信区间为

$$[-0.044, 0.344]$$

例 6.2.6 为了考察温度对某物体断裂强力的影响, 在 70 °C 和 80 °C 下分别重复做了 8 次试验, 测得断裂强力(单位:N)的数据如下:

70 °C: 20.5 18.8 19.8 20.9 21.5 19.5 21.0 21.2

80 °C: 17.5 20.3 20.0 18.8 19.0 20.1 20.2 19.1

假定 70 °C 和 80 °C 下断裂强力分别服从 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 分布, μ_1 和 μ_2

及 σ_1^2 和 σ_2^2 均未知, 求 $\frac{\sigma_1^2}{\sigma_2^2}$ 的置信水平为 0.90 的置信区间.

解 根据样本计算得,

$$m=8, \bar{x}=20.4, s_1^2=0.8857;$$

$$n=8, \bar{y}=19.4, s_2^2=0.9314,$$

其中 X 表示 70 °C 下断裂强力, Y 表示 80 °C 下断裂强力. 利用 $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ 作为

枢轴变量, $F \sim F(m-1, n-1)$, 由表 6.2.2 知, $\frac{\sigma_1^2}{\sigma_2^2}$ 的置信区间为

$$\left[\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{\alpha/2}(m-1, n-1)}, \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{1-\alpha/2}(m-1, n-1)} \right].$$

查表得 $F_{0.05}(7,7) = 3.79$, $F_{0.95}(7,7) = \frac{1}{F_{0.05}(7,7)} = 0.2639$, 代入数据得置信区间为

$$[0.2509, 3.6034].$$

§ 6.3 假设检验的基本概念

1. 问题的提出

前面讲的参数估计是通过样本给未知参数一个合适的估计值, 进而达到“认识”总体的目的. 在实际中, 还经常遇到另一类统计推断问题. 先看下面几个例子.

例 6.3.1 某工厂在正常情况下, 生产的灯泡使用寿命(单位:h) X 服从正态分布 $N(1800, 100^2)$. 今从该厂生产的一批灯泡中随机地取 25 个进行检测, 测得其平均寿命 $\bar{x} = 1730$. 如果标准差不变, 能否认为该厂生产的这批灯泡的寿命均值为 1800 h.

例 6.3.2 在公路上 50 min 内记录每 15 s 通过汽车的车辆数, 得如下数据:

辆数	0	1	2	3	4	5
频数	92	68	28	11	1	0

问能否认为每 15 s 通过的车辆数服从泊松分布?

例 6.3.3 某制药厂推出一种治疗感冒的新药, 为检验其疗效, 将 180 名患者均分为两组, 分别服药或不服药. 三天后, 发现服药的 90 人中有 50 人痊愈, 而不服药的 90 人中有 41 人痊愈, 问这种新药疗效是否显著?

例 6.3.1 的问题可描述为: 已知总体(灯泡寿命)的分布形式(正态分布), 要推断“总体均值为 1800 h”这一假设是否成立. 此类问题可归结为: 已知总体的分布形式, 只对总体的某些未知参数取值作出假设, 通过抽样来判断假设是否成立, 这种检验称为参数检验. 如果不知道总体分布的具体类型, 只对未知分布函数的类型或者它的某些特性提出假设, 然后对这种假设进行检验, 这种检验称为非参数检验. 参数检验和非参数检验统称为假设检验.

例 6.3.2 推断的是“总体服从泊松分布”这一假设是否成立. 在例 6.3.3 中, 假定 X 表示是否服药, Y 表示是否痊愈, 则要推断的是“ X 与 Y 是否存在密切关系”. 也就是说, 要推断的是“二维总体 (X, Y) 的两个分量 X 和 Y 是否独立”. 这两种检验都属于非参数检验问题, 本章只介绍参数检验的内容.

2. 假设检验的基本思想和步骤

下面利用例 6.3.1 来说明假设检验的基本原理.

从直观上看, 如果灯泡的寿命均值 $\mu = 1\,800$ 成立, 由于样本均值的观测值 $\bar{x} = 1\,730$, 故它与均值的偏差为 $1\,730 - 1\,800 = -70$. 我们知道, $\bar{X}$ 是 μ 的一个“好”的估计, 一般情况下, $\bar{X}$ 应与 μ 偏离不大. 我们观测到的偏离 -70 算不算大呢? 为回答这一问题, 我们考察一下 $\bar{X} - \mu$ 的分布情况, 由定理 5.3.1 知, $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$. 因此 $\bar{X} - \mu$ 落入 $(-3\sigma/\sqrt{n}, 3\sigma/\sqrt{n}) = (-60, 60)$

内的概率约为 99.73%, 落入此区间之外的概率仅为 0.27%. 现在的情况是 $\bar{x} - \mu = -70$ 落在区间之外, 一个概率很小的事件发生了. 这违背了“小概率事件在一次试验中不可能发生”这一原理. 偏差 -70 落在区间之外是由最初的假设 $\mu = 1\,800$ 所造成的, 因此这一假设的正确性值得怀疑, 应当予以否定, 即认为总体寿命均值不是 $1\,800$ h.

上述思路包含了反证法的思想, 但它不同于一般的反证法. 一般的反证法是假定某个假设成立, 若推出了与事实相矛盾的结论, 则否定原假设. 因此, “否定原假设”是绝对正确的. 而假设检验中的反证法是假定某个假设成立. 若推出小概率事件发生, 则否定原假设, 这种反证法带有概率的性质. 我们知道小概率事件并不是绝对不可能发生, 只是它发生的可能性很小而已. 由此可知, 假设检验有时要犯错误, 即所得结论与事实可能不符. 这一点后面还要讨论.

下面结合前面的分析, 以例 6.3.1 为例, 给出假设检验的步骤:

(1) 建立假设

$$H_0: \mu = 1\,800, \quad H_1: \mu \neq 1\,800.$$

通常把假设 H_0 叫原假设, 而把与 H_0 对立的假设 H_1 叫做备择假设. 检验的目的就是要在 H_0 和 H_1 二者之中选择其一.

(2) 选取检验的统计量, 并在 H_0 成立的条件下确定该统计量的分布

选取 $U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ 为检验的统计量, 当 H_0 成立时,

$$U = \frac{\bar{X} - 1\,800}{\sigma/\sqrt{n}} \sim N(0, 1).$$

(3) 选取检验的显著性水平 α 与临界值

显著性水平 α 就是所谓的小概率的具体数值, 它一般事先给定. 通常取 $\alpha = 0.05, 0.01$. 临界值表示使小概率事件发生的统计量的数值界限, 也就是统计量分布的相应分位数.

对例 6.3.1, 假定 $\alpha = 0.05$. 查标准正态分布表, 得临界值 $u_{0.025} = 1.96$, 它满足

$$P(|U| \geq u_{0.025}) = 0.05.$$

(4) 作判断

计算统计量的观测值, 并与临界值比较, 即考察小概率事件是否发生. 若发生, 则否定 H_0 , 接受 H_1 ; 反之, 则接受 H_0 .

对例 6.3.1, 统计量 U 的观测值为

$$|U| = \left| \frac{1\,730 - 1\,800}{20} \right| = 3.5 > u_{0.025},$$

故小概率事件 $\{|U| \geq u_{0.025}\}$ 发生了, 从而否定 H_0 , 即认为寿命均值不是 1 800 h.

注: 设 Ω 为样本空间, 则 Ω 中使小概率事件发生的元素构成的集合称为拒绝域 V , 而 $\Omega - V$ 称为接受域. 显然, 若 $(x_1, x_2, \dots, x_n) \in V$, 则应拒绝 H_0 . 若 $(x_1, x_2, \dots, x_n) \in \Omega - V$, 则应接受 H_0 .

3. 假设检验的两类错误

前面已经指出: 假设检验的推理方法是根据“小概率事件在一次试验中不可能发生”进行判断的一种反证法. 但是, 当原假设正确时, 小概率事件无论其概率多么小, 还是可能发生的. 当小概率事件发生时, 我们将作出错误的判断, 因而检验可能犯错误. 所谓犯错误, 就是检验得到的结论与实际情况不符. 有两种情况: 一是原假设 H_0 确实成立, 而检验的结果是拒绝了 H_0 , 称这种错误为第一类错误或“弃真”错误. 二是原假设 H_0 确定不成立, 而检验的结果是接受了 H_0 , 称这种错误为第二类错误或“存伪”错误.

不难推知: 犯第一类错误的概率就是显著性水平 α , 而犯第二类错误的概率 β 的计算通常比较复杂.

例 6.3.4 设总体 $X \sim N(\mu, \sigma_0^2)$, 其中 σ_0^2 已知, $X_1, X_2, \dots, X_n$ 是来自总体的简单随机样本. 对于假设检验

$$H_0: \mu = \mu_0, \quad H_1: \mu = \mu_1,$$

其中 μ_0 和 μ_1 均为已知常数, 且 $\mu_0 < \mu_1$. 设拒绝域取为 $V = \{U \geq u_\alpha\}$, 其中 $U = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}$, 试求犯第二类错误的概率.

解 当 H_1 成立时, 统计量 $U \sim N\left(\frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}}, 1\right)$. 根据第二类错误的定义得犯

第二类错误的概率为

$$\beta = P(\bar{V} | H_1) = P(U < u_\alpha | \mu = \mu_1)$$

$$\begin{aligned}
 &= P\left(U - \frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}} < u_\alpha - \frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}} \mid \mu = \mu_1\right) \\
 &= \Phi\left(u_\alpha - \frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}}\right).
 \end{aligned}$$

从上例可知, 当 H_0, H_1 给定时, 若样本容量 n 固定, α 取值越小, u_α 越大; 因而犯第二类错误的概率 β 越大; 而当 α 固定, 样本容量 n 增大时, β 的取值将变小. 所以要使 α, β 都达到很小, 一般通过适当增加样本容量 n 来实现.

§ 6.4 正态总体参数的假设检验

正态总体 $N(\mu, \sigma^2)$ 是最常见的分布, 其参数 μ 和 σ^2 的有关检验在实际中经常遇到, 下面分别讨论它们的假设检验.

1. 关于均值 μ 的检验

设 $X_1, X_2, \dots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的一个样本, 样本均值和方差分别记为 $\bar{X}$ 和 S^2 , 现要考虑均值 μ 的检验

$$H_0: \mu = \mu_0, \quad H_1: \mu \neq \mu_0.$$

下面根据参数 σ^2 是否已知分别进行讨论.

(1) σ^2 已知

上一节的例 6.3.1 就是 σ^2 已知条件下检验 μ 的问题. 我们知道, 样本均值是总体均值的一个“好”的估计, 它具有无偏性等特点, 因此, 在一般情况下, 样本均值 $\bar{X}$ 与总体均值 μ 应相差不大. 故当 H_0 成立时, 样本均值 $\bar{X}$ 与总体均值 μ_0 应相差不大. 若二者相差偏大, 应拒绝 H_0 . 为此, 选择

$$U = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

为检验统计量. 当 H_0 成立时, 统计量 $U \sim N(0, 1)$. 统计量 U 的绝对值偏大, 说明 $\bar{X}$ 与 μ_0 相差偏大, 此时应拒绝 H_0 . 故对于给定的显著性水平 α , 选取临界值 $u_{\alpha/2}$ 满足

$$P(|U| \geq u_{\alpha/2}) = \alpha,$$

从而得拒绝域为

$$V = \{(x_1, x_2, \dots, x_n) : |U(x_1, x_2, \dots, x_n)| \geq u_{\alpha/2}\}.$$

这种以 U 作为检验统计量的检验称为 U 检验.

(2) σ^2 未知

同(1)的分析一样知, 当 H_0 成立时, 样本均值 $\bar{X}$ 与总体均值 μ_0 应相差不

大. 若二者相差过大, 应拒绝 H_0 . 但不能选择 U 作为检验统计量, 因为 σ^2 未知; 随机变量 U 不再是统计量, 从而无法给出拒绝域. 根据定理 5.3.1, 我们选择

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

作为检验统计量. 当 H_0 成立时, 统计量 $T \sim t(n-1)$. 统计量 T 的绝对值偏大, 说明 $\bar{X}$ 与 μ_0 相差偏大, 此时应拒绝 H_0 . 故对于给定的显著性水平 α , 选取临界值 $t_{\alpha/2}(n-1)$ 满足

$$P(|T| \geq t_{\alpha/2}(n-1)) = \alpha,$$

从而得拒绝域为

$$V = \{(x_1, x_2, \dots, x_n) : |T(x_1, x_2, \dots, x_n)| \geq t_{\alpha/2}(n-1)\}.$$

这种以 t 分布作为检验统计量的检验称为 t 检验.

2. 关于方差 σ^2 的检验

设 $X_1, X_2, \dots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的一个样本, 样本均值和方差分别记为 $\bar{X}$ 和 S^2 , 现要考虑方差 σ^2 的检验

$$H_0: \sigma^2 = \sigma_0^2, \quad H_1: \sigma^2 \neq \sigma_0^2.$$

下面讨论均值 μ 未知(实际往往如此)时方差 σ^2 的检验.

由 § 6.1 知, 样本方差 S^2 是总体方差 σ^2 的良好估计. 因此, 当 H_0 成立时, 样本方差 S^2 与总体方差 σ_0^2 应相差不大. 若二者相差偏大, 应拒绝 H_0 . 也就是说, 样本方差 S^2 与总体方差 σ_0^2 的比值 S^2/σ_0^2 偏大或偏小时, 应拒绝 H_0 . 根据定理 5.3.1, 可选择

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

作为检验统计量. 当 H_0 成立时, $\chi^2 \sim \chi^2(n-1)$. 统计量 χ^2 的取值偏大或偏小时, 说明样本方差 S^2 与 σ_0^2 相差偏大, 此时应拒绝 H_0 . 故对于给定的显著性水平 α , 选取临界值 $\chi_{\alpha/2}^2(n-1)$, $\chi_{1-\alpha/2}^2(n-1)$ 分别满足

$$P(\chi^2 \geq \chi_{\alpha/2}^2(n-1)) = \frac{\alpha}{2},$$

$$P(\chi^2 \leq \chi_{1-\alpha/2}^2(n-1)) = \frac{\alpha}{2},$$

从而得拒绝域为

$$V = \{(x_1, x_2, \dots, x_n) : \chi^2 \geq \chi_{\alpha/2}^2(n-1) \text{ 或 } \chi^2 \leq \chi_{1-\alpha/2}^2(n-1)\}.$$

这种以 χ^2 分布作为检验统计量的检验称为 χ^2 检验.

3. 均值 μ 和方差 σ^2 检验的几个例子

例 6.4.1 某化肥厂用自动包装机将化肥装袋, 规定每袋的标准质量(单

位:500 g)为100. 设每袋的化肥重量 X 服从正态分布. 由以往经验知 X 的标准差 $\sigma = 0.9$ 并保持不变. 某天开工后, 为了检验包装机的工作是否正常, 随机抽取该机所装的9袋, 称得其质量为

99.3 98.7 101.2 100.5 98.3 99.7 102.6 100.5 105.1

问该日此包装机工作是否正常, 即总体均值是否为 $\mu = 100$? ($\alpha = 0.05$)

解 这是方差已知的正态总体均值的检验问题, 应当用 U 检验法, 检验步骤如下:

(1) 待检验假设为

$$H_0: \mu = 100, \quad H_1: \mu \neq 100.$$

(2) 选取

$$U = \frac{\bar{X} - 100}{\sigma / \sqrt{n}}$$

作为检验统计量. 当 H_0 成立时, 统计量 $U \sim N(0, 1)$.

(3) 对于显著性水平 $\alpha = 0.05$, 查表得临界值 $u_{0.025} = 1.96$, 故得拒绝域为

$$V = \{(x_1, x_2, \dots, x_9) : |U| \geq 1.96\}.$$

(4) 据样本观测值得 $\bar{x} = 100.66$, 从而

$$|U| = \frac{100.66 - 100}{0.9/\sqrt{9}} = 2.2 > 1.96,$$

样本落在拒绝区域内, 故应拒绝 H_0 , 即认为包装机工作不正常.

例 6.4.2 已知金属锰的熔点服从正态分布, 对锰进行5次试验, 测得其熔点(单位: $^{\circ}\text{C}$)如下:

1 267 1 271 1 256 1 265 1 254

问是否可以认为锰的熔点为 $1\,260^{\circ}\text{C}$? ($\alpha = 0.05$)

解 由题意知. 需检验假设

$$H_0: \mu = 1\,260, \quad H_1: \mu \neq 1\,260.$$

因为总体服从正态分布且方差 σ^2 未知, 所以要用 t 检验法. 由所给数据可算出

$$\bar{x} = 1\,262.6, \quad s^2 = 53.3,$$

从而

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1\,262.6 - 1\,260}{7.30/\sqrt{5}} = 0.796.$$

对给定的显著性水平 $\alpha = 0.05$, 查表得临界值 $t_{0.025}(n-1) = t_{0.025}(4) = 2.776\,4$. 因为 $|T| = 0.796 < 2.776\,4 = t_{0.025}(4)$, 所以样本没有落在拒绝区域内, 故可以认为锰的熔点为 $1\,260^{\circ}\text{C}$.

例 6.4.3 设维尼纶纤度 X 在正常情况下服从正态分布 $N(\mu, \sigma^2)$, 按规定加工的精度为 $\sigma_0^2 = 0.048^2$, 今抽出 5 根纤维进行测试, 测得其纤度为

1.44 1.36 1.40 1.55 1.32

试问该产品的精度有无显著性变化? ($\alpha = 0.05$)

解 由题意需检验假设

$$H_0: \sigma^2 = \sigma_0^2 = 0.048^2, \quad H_1: \sigma^2 \neq \sigma_0^2 = 0.048^2.$$

这是正态总体的方差检验问题, 故用 χ^2 检验法. 对给定的水平 $\alpha = 0.05$, 查 χ^2 分布表 (自由度为 $n - 1 = 4$) 得临界值 $\chi_{0.025}^2(4) = 11.143$, $\chi_{0.975}^2(4) = 0.484$. 由题中数据可算出 $(n - 1)s^2 = 0.03112$, 从而

$$\chi^2 = \frac{(n - 1)s^2}{\sigma_0^2} = \frac{0.03112}{0.048^2} = 13.51 > \chi_{0.025}^2(4),$$

故应拒绝 H_0 , 即可认为生产精度已有了显著性的变化.

4. 双侧检验与单侧检验

假设检验的原假设 H_0 和备择假设 H_1 的具体内容是由问题的侧重点确定的. 例如, 例 6.4.2 所关心的是“锰的熔点是否为 1260°C ”, 因而检验的假设为 $H_0: \mu = 1260$, $H_1: \mu \neq 1260$. 如果关心的是“锰的熔点是否高于 1260°C ”, 则检验的假设应为 $H_0: \mu \leq 1260$, $H_1: \mu > 1260$. 反之, 如果关心的问题是“锰的熔点是否低于 1260°C ”, 则检验的假设应为 $H_0: \mu \geq 1260$, $H_1: \mu < 1260$. 由此可见, 对于同一问题, 不同的侧重点就会有不同形式 (内容) 的原假设 H_0 和备择假设 H_1 .

尽管原假设 H_0 和备择假设 H_1 的内容随问题的侧重点不同而不同, 但它们本质上都是将参数空间 Θ 分解为两个互不相容的子集 (不妨设为 Θ_0 及 $\Theta - \Theta_0$), 然后检验参数 θ 属于哪一个子集. 即检验的假设都可写成 $H_0: \theta \in \Theta_0, H_1: \theta \in \Theta - \Theta_0$. 如果在数轴上, 集合 $\Theta - \Theta_0$ 位于集合 Θ_0 的两侧, 这种类型的参数检验称为**双侧检验**; 反之, 如果在数轴上, 集合 $\Theta - \Theta_0$ 位于集合 Θ_0 的一侧, 这种类型的参数检验称为**单侧检验**.

前面讨论的参数检验 $H_0: \mu = \mu_0, H_1: \mu \neq \mu_0$ 及 $H_0: \sigma^2 = \sigma_0^2, H_1: \sigma^2 \neq \sigma_0^2$ 都是双侧检验. 而参数检验 $H_0: \mu \leq \mu_0, H_1: \mu > \mu_0$ 及 $H_0: \mu \geq \mu_0, H_1: \mu < \mu_0$ 则属于单侧检验. 单侧检验的假设 H_0 和 H_1 还有其它许多形式 (内容). 同样对方差 σ^2 也有类似的单侧检验. 这里需要指出的是, 对双侧检验, 总是按 $\alpha/2$ 查表; 对单侧检验, 总是按 α 查表.

下面讨论方差已知时, 正态总体均值 μ 的单侧检验

$$H_0: \mu \leq \mu_0, \quad H_1: \mu > \mu_0,$$

这里 μ_0 是一个已知常数.

由于 $\bar{X}$ 是 μ 的一个良好估计, 因此 $\bar{X}$ 与 μ 一般应相差不大. 当 H_0 成立

时, 有不等式 $\bar{X} - \mu_0 \leq \bar{X} - \mu$, 所以此时 $\bar{X} - \mu_0$ 有偏小的趋势; 而 H_1 成立时, 有不等式 $\bar{X} - \mu_0 > \bar{X} - \mu$, 此时 $\bar{X} - \mu_0$ 有偏大的趋势. 因此, 当 H_0 成立时, 若 $\bar{X} - \mu_0$ 偏大, 应拒绝 H_0 . 根据定理 5.3.1, 选择

$$U = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

作为检验统计量. 当 H_0 成立时,

$$U = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1).$$

统计量 U 的值偏大等价于 $\bar{X} - \mu_0$ 的值偏大, 故 U 的观测值偏大时, 应拒绝 H_0 . 对于给定的显著性水平 α , 有

$$P(U \geq u_\alpha) \leq P\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \geq u_\alpha\right) = \alpha,$$

因此拒绝域为

$$V = \{(x_1, x_2, \dots, x_n) : U \geq u_\alpha\}.$$

对另一种类型的单侧检验

$$H_0: \mu \geq \mu_0, \quad H_1: \mu < \mu_0,$$

仍可取 $U = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$ 为检验统计量, 对 U 的取值进行类似上面的分析知, 当观测

值 U 偏小时应拒绝 H_0 , 故拒绝域应为

$$V = \{(x_1, x_2, \dots, x_n) : U \leq -u_\alpha\}.$$

类似可讨论方差未知时, 正态总体均值 μ 的单侧检验问题, 以及正态总体方差 σ^2 的单侧检验问题.

例 6.4.4 某公司生产的一种灯泡, 其寿命 X 服从正态分布 $N(\mu, 300^2)$. 很长时间以来, 灯泡的平均寿命 μ 一直没有超过 2 000 h. 现在采用新工艺后, 从所生产的灯泡中抽取 16 只, 测得平均寿命为 2 168 h. 问采用新工艺后, 灯泡的寿命是否有显著提高? ($\alpha = 0.05$)

解 这是正态总体均值的单侧检验问题, 需检验假设

$$H_0: \mu \leq 2\,000, \quad H_1: \mu > 2\,000.$$

选取

$$U = \frac{\bar{X} - 2\,000}{300 / \sqrt{n}}$$

作为检验统计量. 对显著性水平 $\alpha = 0.05$, 查表得 $u_{0.05} = 1.645$, 故得拒绝域为

$$V = \{(x_1, \dots, x_{16}) : U \geq 1.645\}.$$

据样本知

$$U = \frac{2\,168 - 2\,000}{300/\sqrt{16}} = 2.24 > 1.645,$$

因而拒绝 $H_0: \mu \leq 2\,000$, 即可以认为新灯泡的寿命有显著提高.

5. 两个正态总体的假设检验

在实际工作中还常常遇到两个正态总体的比较问题, 有以下两种情况: 均值的比较和方差的比较. 设 $X_1, X_2, \dots, X_m$ 是来自总体 $X \sim N(\mu_1, \sigma_1^2)$ 的样本, $Y_1, Y_2, \dots, Y_n$ 是来自总体 $Y \sim N(\mu_2, \sigma_2^2)$ 的样本, 且两样本独立, $\bar{X}, \bar{Y}$ 为两样本的均值, S_1^2, S_2^2 为两样本的方差. 根据已知条件的不同, 均值及方差的比较可进一步细分成若干情形(见表 6.4.2), 这里仅介绍两种最简单的情形:

(1) 未知 μ_1, μ_2 , 检验假设

$$H_0: \sigma_1^2 = \sigma_2^2, \quad H_1: \sigma_1^2 \neq \sigma_2^2.$$

(2) 未知 σ_1^2, σ_2^2 , 但 $\sigma_1^2 = \sigma_2^2$, 检验假设

$$H_0: \mu_1 = \mu_2, \quad H_1: \mu_1 \neq \mu_2.$$

例 6.4.5 机床厂某日从两台机器所加工的同一种零件中, 分别抽取 11 个和 9 个样品测量, 其尺寸(单位: cm)分别为

第一台机器: 6.2 5.7 6.0 6.3 6.5 6.0 5.7 5.8 6.0 5.8 6.0

第二台机器: 5.6 5.7 5.9 5.5 5.6 6.0 5.8 5.5 5.7

已知零件尺寸服从正态分布, 问在显著性水平 $\alpha = 0.05$ 条件下, 加工精度(方差)是否有显著性差异?

记 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ 分别表示第一、第二台机器生产的零件尺寸, 所给的两组数据分别是来自于 X 和 Y 的样本观测值. 此问题就是在未知 μ_1, μ_2 的条件下, 对假设 $H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 \neq \sigma_2^2$ 进行检验.

由于样本方差 S^2 是总体方差 σ^2 的良好估计(据 § 6.1), 一般情况下二者相差不大. 因而当 H_0 成立时, 比值 $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ 的值不应过大, 也不应过小, 否则就应拒绝 H_0 . 根据定理 5.3.2, 可选择

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

作为检验统计量. 当 H_0 成立时, $F = \frac{S_1^2}{S_2^2} \sim F(m-1, n-1)$. 若 F 的取值偏大或偏小, 应拒绝 H_0 . 故对于给定的显著性水平 α , 选取临界值 $F_{\alpha/2}(m-1, n-1)$, $F_{1-\alpha/2}(m-1, n-1)$, 分别满足

$$P(F \geq F_{\alpha/2}(m-1, n-1)) = \frac{\alpha}{2},$$

$$P(F \leq F_{1-\alpha/2}(m-1, n-1)) = \frac{\alpha}{2},$$

从而得拒绝域为

$$V = \{(x_1, x_2, \dots, x_n) : F \geq F_{\alpha/2}(m-1, n-1) \text{ 或 } F \leq F_{1-\alpha/2}(m-1, n-1)\}.$$

下面给出例 6.4.5 的求解过程.

解 由题设可知

$$m = 11, \bar{x} = 6, s_1^2 = 0.064;$$

$$n = 9, \bar{y} = 5.7, s_2^2 = 0.030.$$

要检验的假设为

$$H_0: \sigma_1^2 = \sigma_2^2, \quad H_1: \sigma_1^2 \neq \sigma_2^2,$$

选择检验统计量

$$F = \frac{S_1^2}{S_2^2} \stackrel{H_0 \text{ 成立}}{\sim} F(m-1, n-1)$$

由于 $F = \frac{0.064}{0.030} = 2.13$, 查表得 $F_{0.025}(10, 8) = 4.30$, $F_{0.975}(10, 8) = 1/F_{0.025}(8, 10) =$

0.26, 故

$$F_{0.975}(10, 8) < F < F_{0.025}(10, 8).$$

因而在显著性水平 0.05 下, 接受原假设 H_0 , 即认为两台机器加工精度(方差)无显著性差异.

例 6.4.6 某种物品在处理前与处理后分别抽取样本分析其含脂率如下:

处理前: 0.19 0.18 0.21 0.30 0.41 0.12 0.27

处理后: 0.15 0.13 0.07 0.24 0.19 0.06 0.08 0.12

假定处理前后的含脂率都服从正态分布, 且方差不变, 问在显著性水平 $\alpha = 0.05$ 条件下, 处理前后含脂率的均值是否有显著性差异?

解 由题设可知

$$\bar{x} = 0.24, s_1^2 = 0.0091, m = 7;$$

$$\bar{y} = 0.13, s_2^2 = 0.0039, n = 8.$$

$$\sigma_1^2 = \sigma_2^2.$$

要检验的假设为

$$H_0: \mu_1 = \mu_2, \quad H_1: \mu_1 \neq \mu_2.$$

根据定理 5.3.2 选择检验统计量

$$T = \frac{\bar{X} - \bar{Y}}{S_w \sqrt{\frac{1}{m} + \frac{1}{n}}} \stackrel{H_0 \text{ 成立}}{\sim} t(m+n-2),$$

其中

$$S_w^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}.$$

代入数据计算得 $T = 2.68$, 查表得 $t_{0.025}(13) = 2.16$, 故

$$|T| > t_{0.025}(13),$$

所以拒绝 H_0 , 即认为处理前后含脂率的均值有显著性差异.

6. 正态总体的假设检验表

下面列表总结正态总体的假设检验问题.

表 6.4.1 单正态总体的假设检验表(显著性水平为 α)

条 件	原假设 H_0	备择假设 H_1	检验统计量	拒绝域
$\sigma^2 = \sigma_0^2$ 已知	$\mu = \mu_0$	$\mu \neq \mu_0$	$U = \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}}$	$ U \geq u_{\alpha/2}$
	$\mu \leq \mu_0$	$\mu > \mu_0$		$U \geq u_\alpha$
	$\mu \geq \mu_0$	$\mu < \mu_0$		$U \leq -u_\alpha$
σ^2 未知	$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$	$ T \geq t_{\alpha/2}(n-1)$
	$\mu \leq \mu_0$	$\mu > \mu_0$		$T \geq t_\alpha(n-1)$
	$\mu \geq \mu_0$	$\mu < \mu_0$		$T \leq -t_\alpha(n-1)$
$\mu = \mu_0$ 已知	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2 = \frac{1}{\sigma_0^2} \sum_{i=1}^n (X_i - \mu_0)^2$	$\chi^2 \geq \chi_{\alpha/2}^2(n)$ 或 $\chi^2 \leq \chi_{1-\alpha/2}^2(n)$
	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$		$\chi^2 \geq \chi_\alpha^2(n)$
	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$		$\chi^2 \leq \chi_{1-\alpha}^2(n)$
μ 未知	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$	$\chi^2 \geq \chi_{\alpha/2}^2(n-1)$ 或 $\chi^2 \leq \chi_{1-\alpha/2}^2(n-1)$
	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$		$\chi^2 \geq \chi_\alpha^2(n-1)$
	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$		$\chi^2 \leq \chi_{1-\alpha}^2(n-1)$

表 6.4.2 两个正态总体的假设检验表(显著性水平为 α)

条 件	原假设 H_0	备择假设 H_1	检验统计量	拒绝域
σ_1^2, σ_2^2 已知	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$U = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$	$ U \geq u_{\alpha/2}$
	$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$		$U \geq u_\alpha$
	$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$		$U \leq -u_\alpha$

续表

条 件	原假设 H_0	备择假设 H_1	检验统计量	拒绝域
σ_1^2, σ_2^2 未知 但 $\sigma_1^2 = \sigma_2^2$	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$T = \frac{\bar{X} - \bar{Y}}{S_w \sqrt{\frac{1}{m} + \frac{1}{n}}}$	$ T \geq t_{\alpha/2}(m+n-2)$
	$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$		$T \geq t_{\alpha}(m+n-2)$
	$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$		$T \leq -t_{\alpha}(m+n-2)$
μ_1, μ_2 已知	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$F = \frac{\sum_{i=1}^m (X_i - \mu_1)^2 / m}{\sum_{j=1}^n (Y_j - \mu_2)^2 / n}$	$F \geq F_{\alpha/2}(m, n)$ 或 $F \leq F_{1-\alpha/2}(m, n)$
	$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$		$F \geq F_{\alpha}(m, n)$
	$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$		$F \leq F_{1-\alpha}(m, n)$
μ_1, μ_2 未知	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$F \geq F_{\alpha/2}(m-1, n-1)$ 或 $F \leq F_{1-\alpha/2}(m-1, n-1)$
	$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$		$F \geq F_{\alpha}(m-1, n-1)$
	$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$		$F \leq F_{1-\alpha}(m-1, n-1)$

§ 6.5 参数估计与假设检验的 MATLAB 计算

1. 参数估计

MATLAB 统计工具箱中提供了一组参数估计函数, 这些函数采用最大似然法给出了常用概率分布模型的参数的点估计和区间估计值. 表 6.5.1 中列出与本书相关的部分函数, 其中 X 为样本数据 (向量或矩阵), α 为显著性水平 α , α 的缺省值为 0.05.

表 6.5.1 参数估计函数

分布	函数名	调用格式	说明
二项分布	binofit	[p,pci] = binofit(X,n,alpha)	p: 试验成功的概率 p 的估计值. pci: p 的置信区间. n: 总试验次数.
指数分布	expfit	[mu,muci] = expfit(X,alpha)	mu: 参数 $\mu = \frac{1}{\lambda}$ 的估计值. muci: μ 的置信区间.

续表

分布	函数名	调用格式	说明
正态分布	normfit	[mu,sigma,muci,sigmaci] =normfit(X,alpha)	mu, sigma: 参数 μ, σ 的估计值. muci, sigmaci: 参数 μ, σ 的置信区间.
泊松分布	poissfit	[lambda,lambdaci] =poissfit(X,alpha)	lambda: 参数 λ 的估计值. lambdaci: 参数 λ 的置信区间.
均匀分布	unifit	[a,b,aci,bci]=unifit(X,alpha)	a, b: 参数 a, b 的估计值. aci, bci: 参数 a, b 的置信区间.

例 6.5.1 据媒体报道某地区的大米零售价格为 1.15 元/500 g, 为检验该报道是否真实, 某记者在该地区随机采访了一些大米零售点, 得到最近两个月的大米价格(单位:元/500 g)为:

第一个月: 1.19 1.17 1.15 1.16 1.12 1.21 1.15 1.22 1.16 1.18
1.09 1.12 1.19 1.12 1.17 1.13 1.14 1.09 1.09 1.18

第二个月: 1.18 1.19 1.15 1.22 1.18 1.21 1.20 1.22 1.28 1.16
1.20 1.23 1.21 1.19 1.17 1.19 1.28 1.26 1.18 1.25

根据这些数据, 试问媒体的报道是否可信? 并给出大米价格的置信区间.
($\alpha=0.05$)

解 输入第一个月的样本数据:

```
X1=[1.19,1.17,1.15,1.16,1.12,1.21,1.15,1.22,1.16,1.18,...√  
1.09,1.12,1.19,1.12,1.17,1.13,1.14,1.09,1.09,1.18];√
```

估计其均值:

```
[mu,sigma,muci,sigmaci]=normfit(X1,0.05);√
```

```
mu, muci=muci'√
```

```
mu =
```

```
1.1515
```

```
muci =
```

```
1.1334 1.1696
```

由 $\mu=1.1515$ 可见, 第一个月时大米价格可以认为是 1.15 元/500 g, 其置信区间为 [1.1334, 1.1696].

再考察第二个月的样本数据:

```
X2=[1.18,1.19,1.15,1.22,1.18,1.21,1.20,1.22,1.28,1.16,...√  
1.20,1.23,1.21,1.19,1.17,1.19,1.28,1.26,1.18,1.25];√
```

估计其均值:

```
[mu,sigma,muci,sigmaci] = normfit(X2,0.05); √
mu, muci = muci' √
mu =
    1.2075
muci =
    1.1901    1.2249
```

显然该月的大米价格不能认为是 1.15 元/500 g, 其置信区间为 [1.190 1, 1.224 9].

2. 假设检验

假定总体服从正态分布 $N(\mu, \sigma^2)$, 下面介绍用 MATLAB 对其参数 μ 和 σ^2 进行检验的方法.

(1) 对均值的检验

总体方差 σ^2 已知时, 对均值的假设检验用 U 检验, 调用格式为

```
[h,p,ci] = ztest(X,mu,sigma,alpha,tail).
```

总体方差 σ^2 未知时, 对均值的假设检验用 t 检验, 调用格式为

```
[h,p,ci] = ttest(X,mu,alpha,tail).
```

以上两式中, 输入参数 X 是样本数据; μ 是 H_0 中的 μ_0 ; σ 是总体标准差 σ ; α 是显著性水平 α (α 缺省时设定为 0.05); tail 是对备选假设 H_1 的选择: H_1 为 $\mu \neq \mu_0$ 时用 $\text{tail}=0$ (缺省值); H_1 为 $\mu > \mu_0$ 时用 $\text{tail}=1$; H_1 为 $\mu < \mu_0$ 时用 $\text{tail}=-1$; 输出参数 $h=0$ 表示接受 H_0 , $h=1$ 表示拒绝 H_0 ; p 表示在假设 H_0 下样本均值出现的概率, p 越小 H_0 越值得怀疑; ci 是 μ_0 的置信区间.

例 6.5.2 考虑计算本章例 6.4.1 的问题, 此时已知方差为 $\sigma^2 = 0.9^2$, $\alpha = 0.05$, 待检假设为

$$H_0: \mu = 100, \quad H_1: \mu \neq 100.$$

解 由于方差已知, 故采用 U 检验法.

```
X=[99.3 98.7 101.2 100.5 98.3 99.7 102.6 100.5 105.1]; √
[h,p,ci] = ztest(X,100,0.9) √
h =
    1
p =
    0.0289
ci =
    100.0676    101.2435
```

计算结果为 $h=1$, $p=0.0289$, μ_0 的置信区间为 [100.067 6, 101.243 5], 显然应拒绝 H_0 假设.

例 6.5.3 考虑计算本章例 6.4.2 的问题, 此时方差未知, $\alpha = 0.05$, 待检假设为

$$H_0: \mu = 1\,260, H_1: \mu \neq 1\,260.$$

解 因方差未知, 故采用 t 检验法.

$X = [1267, 1271, 1256, 1265, 1254]; \downarrow$

$[h, p, ci] = ttest(X, 1260) \downarrow$

$h =$

0

$p =$

0.4704

$ci =$

1253.5 1271.7

计算结果 $h = 0$, $p = 0.4704$, 置信区间 $[1\,253.5, 1\,271.7]$ 包含 $1\,260$, 由此可见应接受 H_0 假设.

例 6.5.4 计算本章例 6.4.4 的问题, 已知方差 $\sigma^2 = 300^2$, 待检假设为

$$H_0: \mu \leq 2\,000, H_1: \mu > 2\,000.$$

解 这是一个单侧检验问题, 因方差已知, 故使用 `ztest` 函数并取其输入参数中的 `tail = 1`.

$X = 2168 * \text{ones}(1, 16); \downarrow$

$[h, p, ci] = \text{ztest}(X, 2000, 300, 0.05, 1) \downarrow$

$h =$

1

$p =$

0.0125

$ci =$

2045.6 Inf

由计算结果知, 应拒绝 H_0 假设.

(2) 对方差的检验

对方差 σ^2 的检验可采用 χ^2 检验法. 下面仅通过示例进行说明.

例 6.5.5 考虑计算本章例 6.4.3 的问题. 此时有 $\sigma_0^2 = 0.048^2$, $\alpha = 0.05$, 待检假设为

$$H_0: \sigma^2 = \sigma_0^2, H_1: \sigma^2 \neq \sigma_0^2.$$

解 首先输入基本数据:

$X = [1.44, 1.36, 1.40, 1.55, 1.32]; \downarrow$

$\text{sigma} = 0.048; \downarrow$

alpha = 0.05; ↓

n = length(X); ↓

然后计算 $\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$:

X2 = (n - 1) * var(X) / sigma ^ 2 ↓

X2 =

13.5069

再使用逆 χ^2 分布函数 chi2inv 计算 $\chi_{\alpha/2}^2$ 和 $\chi_{1-\alpha/2}^2$:

X1a = chi2inv(alpha/2, n - 1) ↓

X1a =

0.4844

Xa = chi2inv(1 - alpha/2, n - 1) ↓

Xa =

11.1433

因本教材采用的是上 α 分位数(参见定义 5.3.5), 故得 $\chi_{1-\alpha/2}^2 = 0.4844$, $\chi_{\alpha/2}^2 =$

11.1433. 如果 $\chi_{1-\alpha/2}^2 \leq \chi^2 \leq \chi_{\alpha/2}^2$, 则接受 H_0 假设, 否则拒绝 H_0 假设:

if X2 >= X1a & X2 <= Xa, h = 0, else h = 1, end ↓

h =

1

即计算结果为拒绝 H_0 假设.

如果将上述命令存入一个命令文件(m 文件)中, 便可得到一个对方差 σ^2 进行假设检验的通用程序.

习 题 6

1. 某中学从二年级的男生中随机抽取 9 人, 测得其身高(单位:cm)和体重(单位:kg)值如下(括号内第一个数字为身高,第二个数字为体重):

(155,45) (160,52) (158,49) (162,55) (157,50)

(163,53) (155,47) (159,51) (153,46)

试求身高 X 和体重 Y 的均值 μ 和方差 σ^2 的矩估计.

2. 已知总体 X 在 $[a, b]$ 上服从均匀分布

$$p(x; a, b) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{其它}, \end{cases}$$

其中 $a < b$, 设 $X_1, X_2, \dots, X_n$ 是来自总体 X 的样本. 试求:

(1) 参数 a 与 b 的矩估计量;

(2) 参数 a 与 b 的最大似然估计量.

3. 设总体 X 服从正态分布, 并且已知总体均值 μ , 如果样本观测值为 $x_1, x_2, \dots, x_n$, 求方差 σ^2 的最大似然估计值.

4. 设总体 X 服从几何分布

$$P(X=k) = p(1-p)^{k-1}, \quad k=1, 2, \dots.$$

又 $x_1, x_2, \dots, x_n$ 是来自总体 X 的 n 个样本观测值, 试求 p 与 $E(X)$ 的最大似然估计.

5. 已知总体 X 的概率密度为

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

其中未知参数 $\theta > 0$, 设 $X_1, X_2, \dots, X_n$ 为取自总体 X 的简单随机样本,

(1) 求 θ 的最大似然估计量;

(2) 试问该估计量是否为无偏估计量? 说明理由.

6. 设 $X_1, X_2, \dots, X_n$ 是取自正态总体 $N(\mu, \sigma^2)$ 的一个样本, 试选择适当的 C , 使 $S^2 = C \sum_{i=1}^{n-1} (X_{i+1} - X_i)^2$ 为 σ^2 的无偏估计.

7. 设样本 $X_1, X_2, \dots, X_n$ 来自于参数为 λ 的泊松分布, 试证明样本均值 $\bar{X}$ 与 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 都是 λ 的无偏估计, 且对任一 $a (0 \leq a \leq 1)$ 值, 统计量 $a\bar{X} + (1-a)S^2$ 也是 λ 的无偏估计.

8. 设样本为 $X_1, X_2, \dots, X_n$, 又 $C_1, C_2, \dots, C_n$ 为常数, 且 $\sum_{i=1}^n C_i = 1$. 证明:

(1) $\hat{\mu} = \sum_{i=1}^n C_i X_i$ 是总体均值 μ 的无偏估计;

(2) 在所有这些无偏估计中, 样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 的方差最小.

9*. 设 $X_1, X_2, \dots, X_n$ 是取自正态总体 $N(\mu, \sigma^2)$ 的一个样本, $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 是否为 σ^2 相合估计? 为什么?

10*. 设总体 X 的均值为 μ , 统计量 $\hat{\mu}_1$ 和 $\hat{\mu}_2$ 是参数 μ 的两个无偏估计量, 它们的方差分别为 σ_1^2, σ_2^2 , 相关系数为 ρ . 试确定常数 $c_1 > 0, c_2 > 0, c_1 + c_2 = 1$, 使得 $c_1 \hat{\mu}_1 + c_2 \hat{\mu}_2$ 有最小方差.

11. 设总体 X 的密度函数为

$$f(x) = (\alpha+1)x^\alpha, \quad 0 < x < 1, \quad \alpha > -1,$$

又 $X_1, X_2, \dots, X_n$ 为其样本, 试求参数 α 的矩估计及最大似然估计. 今有样本观测值为 0.3, 0.5, 0.7, 0.4, 0.23, 0.9 及 0.65. 求参数 α 的估计值.

12*. 设有 k 个事件 $A_1, A_2, \dots, A_k$ 两两互不相容, 其概率 $p_1, p_2, \dots, p_k$ 之和等于 1. 做 n 次重复独立试验, 事件 $A_1, A_2, \dots, A_k$ 发生的次数分别为 $n_1, n_2, \dots, n_k$. 试证明事件 $A_i (i=1, 2, \dots, k)$ 发生的频率 n_i/n 为其概率 p_i 的最大似然估计值.

13. 设总体 $X \sim N(\mu, \sigma^2)$, 数据 $x_1, x_2, \dots, x_{15}$ 是其一组样本观测值, 已知 $\sum_{i=1}^{15} x_i = 8.7, \sum_{i=1}^{15} x_i^2 = 25.05$, 试分别求 μ 和 σ^2 的置信水平为 0.95 的置信区间.

14. 某工厂生产滚球, 从某日生产的产品中随机抽取 9 个, 测得直径(单位:mm)如下:

14.8 14.7 15.1 14.9 14.8 15.2 14.9 15.0 14.6

设滚球直径服从正态分布, 如果(1)已知标准差为 0.15 mm; (2)未知标准差, 求直径 X 的均值的置信水平为 0.95 的置信区间.

15*. 设总体 $X \sim N(\mu, \sigma_0^2)$, 其中 σ_0 为已知数. 需要抽取容量 n 为多大的样本, 才能使总体均值 μ 的置信水平为 $1 - \alpha$ 的置信区间的长度不大于常数 d ?

16*. 随机抽取某种炮弹 9 发做试验, 得炮口速度(单位:m/s)的样本标准差为 10.5, 设炮口速度服从正态分布. 求这种炮弹的炮口速度的标准差 σ 的置信水平为 95% 的置信区间.

17. 今从两个正态总体 $N(\mu_1, \sigma_1^2)$ 及 $N(\mu_2, \sigma_2^2)$ 中分别独立地取样本, 它们的样本容量分别为 $n_1 = 10, n_2 = 8$. 样本方差分别为 $S_1^2 = 3.4, S_2^2 = 2.5$. 试求二总体方差比 $\frac{\sigma_1^2}{\sigma_2^2}$ 的置信水平为 0.95 的置信区间.

18. 已知在正常生产条件下, 某种汽车零件的质量服从正态分布 $N(50, 0.8^2)$. 某日从生产的零件中任取 11 件, 测得样本(质量)均值 $\bar{x} = 50.45$. 如果方差不变, 问该日生产的零件的平均质量有无显著差异? ($\alpha = 0.05$)

19. 设某次英语四级考试, 考生的成绩服从正态分布. 从中随机地抽取 40 名考生的成绩, 算得平均成绩为 68 分, 标准差为 17 分. 在显著性水平 $\alpha = 0.05$ 下, 是否可以认为这次英语考试全体考生的平均成绩为 72 分?

20. 根据过去几年农产品产量的调查资料, 某县小麦单位产量服从方差为 56.25 的正态分布. 今年随机抽取了 10 块地. 测得小麦单位产量(单位:500 g)分别为

969 695 743 836 748 558 675 631 654 685

据上述数据, 能否认为该县小麦单位产量的方差没有发生变化? ($\alpha = 0.05$)

21*. 某城市的物价部门对当前市场的鸡蛋价格情况进行调查, 共调查了 30 个集市上的鸡蛋售价测得它们的平均价格为 2.21 元/500 g, 已知以往鸡蛋平均售价一直稳定在 2 元/500 g. 如果该城市鸡蛋售价服从正态分布 $N(\mu, 0.18)$, 假定方差不变, 能否据上述数据认为该城市当前的鸡蛋售价明显高于往年? ($\alpha = 0.05$)

22*. 测定某种溶液中的水分. 它的 10 个测定值给出 $\bar{x} = 0.637\%, s^2 = 0.044\%$. 设该溶液中的水分服从正态分布, 在显著性水平 $\alpha = 0.05$ 下, 试检验假设

$$H_0: \sigma^2 = 0.045\%, H_1: \sigma^2 < 0.045\%.$$

23*. 设总体 $X \sim N(\mu, \sigma^2)$, 方差 σ^2 已知, 又 $X_1, X_2, \dots, X_n$ 为来自总体 X 的样本. 对假设检验

$$H_0: \mu = \mu_0, H_1: \mu = \mu_1 \quad (\mu_1 > \mu_0),$$

已知检验该假设的否定域为

$$V = \{(x_1, \dots, x_n): \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > 1.64\}, \quad (\alpha = 0.05).$$

试计算该检验法犯第二类错误的概率 β , 要求计算结果用标准正态分布的分布函数 $\Phi(x)$ 表示, 并讨论 β 与 n 的关系.

24*. 设有甲、乙两种安眠药, 考虑它们的治疗效果, 以 X 及 Y 分别表示失眠病人服甲药或服乙药后睡眠时间延长的时数. 现在独立观察 20 个失眠病人, 其中 10 人服甲药, 另 10 人服乙药, 其结果如下:

时数 病人 药物	1	2	3	4	5	6	7	8	9	10
X	1.9	0.8	1.1	0.1	0.1	5.4	5.5	1.6	5.6	3.4
Y	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0

假定 $X \sim N(\mu_1, \sigma^2)$, $Y \sim N(\mu_2, \sigma^2)$, 其中 σ^2 相同且未知, 试问这两种药物的疗效有无显著性差异? ($\alpha = 0.05$)

25*. 一辆货车从甲地到乙地有 A 和 B 两条行车路线, 行车时间分别服从 $N(\mu_i, \sigma_i^2)$ ($i=1, 2$). 现让一名驾驶员在每条路上跑 50 次, 记录其行车时间(单位: min). 在线路 A 上有 $\bar{x} = 95$, $s_x = 20$, 在线路 B 上有 $\bar{y} = 76$, $s_y = 15$. 试在 $\alpha = 0.05$ 的水平上检验两者方差是否一样, 两者均值是否一样?

26. 某物质在化学处理前后的含脂率如下:

处理前: 0.19 0.18 0.21 0.30 0.66 0.42 0.08 0.12 0.30 0.27

处理后: 0.15 0.13 0.00 0.07 0.24 0.24 0.19 0.04 0.08 0.20 0.12

假定处理前后含脂率分别服从正态分布, 问处理后是否降低了含脂率? ($\alpha = 0.05$)

第7章 一元线性回归分析和方差分析

§ 7.1 回归分析的基本概念

参数估计和假设检验都是通过样本去推断总体 Y . 在实际应用中, 人们有时关心另一类问题: 研究 Y 与对其影响比较显著 (或存在密切关系) 的变量之间的关系, 并将这种关系用某种数学式子表示出来. 例如, 人们的消费数量 Y 与其可支配收入 X 有密切的关系. 一般说来, 当 X 比较大时, Y 也倾向于大, 但 X 不能完全决定 Y . 研究 X 和 Y 之间的关系, 就能为制定合理的收入分配政策提供决策依据. 又如, 纱的质量与原棉的质量具有密切的关系, 而原棉的质量往往与其品种、产地有关. 因此为了预测和控制纱的质量, 就必须研究棉花的质量指标 $(X_1, X_2, \dots, X_p)$ (如棉花成熟度、韧度、强力、长度均匀度等) 与纺出的纱的品质指标 Y (也可是向量) 之间的关系, 从而通过对各种原棉的合理搭配, 以保证所纺出的纱合乎要求. 类似的例子非常多.

在上面的例子中, 随机变量 Y 与变量 X (或向量) 关系非常密切, 但不是确定的函数关系, 变量之间的这种关系称为相关关系. 回归分析就是研究相关关系的一种数理统计方法.

回归一词是英国著名人类学家和气象学家高尔顿 (Francis Galton) 首次提出的. 在其论文《身高遗传中的平庸回归》中, 他发现: 通常父代身材高大的, 其子代身材也高大; 父代身材矮小的, 其子代身材亦矮小, 但子代的身高有向其平均值靠拢的趋势. 后来, 英国著名统计学家皮尔逊 (K. Pearson) 通过对上千个家庭成员的身高 (单位: in^①) 数据进行分析, 得出儿子的身高 Y 与父亲的身高 X 大致有以下关系: $Y = 0.516X + 33.73$.

上述例子中的随机变量 Y 称为因变量或响应变量, X (或 $(X_1, X_2, \dots, X_p)$) 称为自变量或预报变量. 我们可以设想, Y 的值由两部分组成: 一部分是由 X 决定的, 它是 X 的函数, 记为 $f(X)$. 另一部分是由其它众多未考虑的因素 (随机或非随机) 所决定的, 它可视为随机误差, 记为 ε . 于是得到模型

$$Y = f(X) + \varepsilon.$$

① 1 in = 25.4 mm.

为了尽可能把 X 对 Y 的影响 $f(X)$ 确定出来, 一般要求 $E(\varepsilon) = 0$, $D(\varepsilon) = \sigma^2$ (未知), 上述模型称为一元回归模型.

同样, 当自变量为 $X_1, X_2, \dots, X_p$ (共 p 个) 时, 有模型

$$Y = f(X_1, X_2, \dots, X_p) + \varepsilon,$$

其中 ε 是随机误差, $E(\varepsilon) = 0$, $D(\varepsilon) = \sigma^2$ (未知), 称此模型为多元回归模型. 本章中, 我们只讨论一元回归模型.

§ 7.2 一元线性回归

1. 散点图与回归直线

研究一元回归模型 $Y = f(X) + \varepsilon$ 的主要任务之一就是给出 $f(X)$ 的具体表达式. 确定 $f(X)$ 的具体表达式包含两方面的内容: 一是选择函数类型 (如线性函数、非线性函数等); 二是确定函数中的未知参数, 其中函数类型的选择是最关键的. 那么究竟选择何种类型的函数作为 $f(X)$ 呢? 一般情况下, 我们只能从样本出发去选择 $f(X)$ 的函数类型.

设 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 是 n 个观测值, 为了考察 $f(x)$ 的函数类型, 我们将这些观测值描在直角坐标系中, 得到下列类型的图:

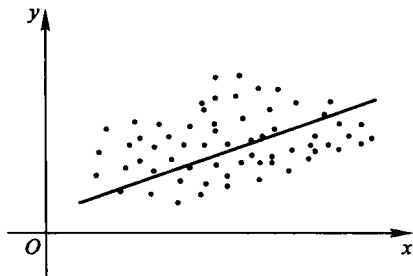


图 7.2.1

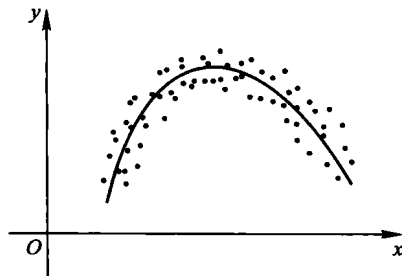


图 7.2.2

这种图称为散点图. 如果 n 个散点密集在一条直线周围, 呈现出直线趋势 (如图 7.2.1), 则称变量 X 和 Y 为直线相关 (或线性相关); 若 n 个散点呈现曲线趋势 (如图 7.2.2), 则称变量 X 和 Y 为曲线相关.

通过对散点图的观察, 可大致了解 X 和 Y 之间的关系类型. 如果认为 X 和 Y 是线性相关, 自然应取 $f(X) = \beta_0 + \beta_1 X$; 如果认为 X 和 Y 是曲线相关, 则 $f(X)$ 应取相应的函数曲线 (如抛物线、椭圆等).

如果认为 X 和 Y 线性相关, 则所要确定的直线 $\hat{Y} = f(X) = \beta_0 + \beta_1 X$ 称为回归直线.

本节只研究线性相关问题, 即研究下列模型:

$$\begin{cases} Y = \beta_0 + \beta_1 X + \varepsilon, \\ E(\varepsilon) = 0, D(\varepsilon) = \sigma^2 \text{ 未知.} \end{cases} \quad (7.2.1)$$

模型(7.2.1)称为一元线性回归模型, 未知参数 β_1 称为回归系数.

在实际问题中, 自变量 X 的值一般是可以控制或可以精确测量的, 因此为了讨论简便起见, 我们认为自变量 X 是非随机变量.

2. 回归直线的确定(估计)

设 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 是一组样本观测值, 由散点图知变量 X 和 Y 线性相关, 那么如何根据样本观测值来确定回归直线 $\hat{Y} = \beta_0 + \beta_1 X$ 呢? 显然直线 $\hat{Y} = \beta_0 + \beta_1 X$ 与 n 个观测点 $(x_i, y_i) (i=1, 2, \dots, n)$ 愈接近就愈能反映 Y 与 X 之间的内在联系.

点 (x_i, y_i) 与直线 $\hat{Y} = \beta_0 + \beta_1 X$ 的偏差为 $|y_i - (\beta_0 + \beta_1 x_i)|$ (见图 7.2.3).

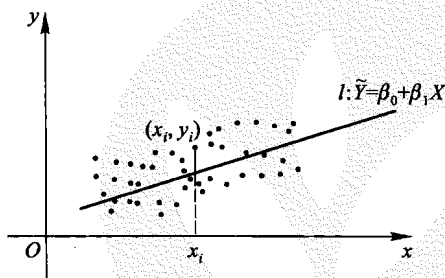


图 7.2.3

令

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2,$$

称 $Q(\beta_0, \beta_1)$ 为偏差平方和. 偏差平方和 Q 反映了 n 个观测值与直线 l 的偏离程度, Q 越小直线就越能反映 Y 与 X 的内在联系. 当样本观测值给定后, Q 的取值就完全依赖于 β_0 和 β_1 . 因此, 我们应当选择适当的 β_0 和 β_1 , 使偏差平方和 $Q(\beta_0, \beta_1)$ 达到最小.

由微积分知, $Q(\beta_0, \beta_1)$ 的最小值可通过令 Q 对 β_0, β_1 的偏导数等于 0 来求, 于是有

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0, \\ \frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0. \end{cases}$$

经整理有

$$\begin{cases} n\beta_0 + n\bar{x}\beta_1 = n\bar{y}, \\ n\bar{x}\beta_0 + \beta_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i. \end{cases}$$

称上述方程组为正规方程组. 不难求得

$$\hat{\beta}_1 = \frac{L_{xy}}{L_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

其中

$$L_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2,$$

$$L_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}.$$

可验证 $(\hat{\beta}_0, \hat{\beta}_1)$ 使 $Q(\beta_0, \beta_1)$ 达到最小值, 称 $\hat{\beta}_0, \hat{\beta}_1$ 为参数 β_0 和 β_1 的最小二乘估计, 这种求解参数估计的方法称为最小二乘估计法, 称直线方程 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 为 Y 关于 X 的一元线性回归方程.

最后指出, 可以证明 $E[Q(\hat{\beta}_0, \hat{\beta}_1)] = (n-2)\sigma^2$, 即

$$E\left[\frac{Q(\hat{\beta}_0, \hat{\beta}_1)}{n-2}\right] = \sigma^2,$$

所以

$$\hat{\sigma}^2 = \frac{Q(\hat{\beta}_0, \hat{\beta}_1)}{n-2}$$

是 σ^2 的无偏估计, 称为剩余方差.

例 7.2.1 今调查了某种产品的月产量 Y 与设备投资额 X 之间一组 (共 10 个) 统计数据如下表:

X /万元	17.1	16.3	18.4	19.0	15.8	16.0	17.5	17.8	18	16.4
Y /t	65	63	70	75	60	55	64	69	65	58

试求 Y 与 X 的一元线性回归方程.

解 由散点图 (略) 知, Y 与 X 线性相关, 故采用一元线性回归模型. 由表中数据可算得

$$\bar{x} = 17.23, \quad L_{xx} = 10.62,$$

$$\bar{y} = 64.4, \quad L_{xy} = 52.68,$$

进而得到 β_0, β_1 的最小二乘估计值为

$$\hat{\beta}_0 = -21.06, \quad \hat{\beta}_1 = 4.96.$$

所以 Y 关于 X 的一元线性回归方程为

$$\hat{y} = -21.06 + 4.96x.$$

3. 回归方程的显著性检验

在实际工作中, 我们事先往往不能确定 Y 和 X 是否线性相关. 根据散点图作出线性相关的判断只是一种直观认识, 这种认识必须经过检验后, 才能确定其是否正确. 为了检验的需要, 这里假定随机误差 $\varepsilon \sim N(0, \sigma^2)$.

由 $Y = \beta_0 + \beta_1 X + \varepsilon$ 可知, 当 $|\beta_1|$ 越大时, Y 随 X 变化的趋势就越明显; 反之, $|\beta_1|$ 越小时, Y 随 X 变化就越不明显. 特别当 $\beta_1 = 0$ 时, 则认为 X 与 Y 不存在线性关系(当然可能存在其它非线性关系); 当 $\beta_1 \neq 0$ 时, 则认为 Y 与 X 存在线性关系. 因此, 检验 Y 与 X 是否线性相关, 等价于对假设

$$H_0: \beta_1 = 0, \quad H_1: \beta_1 \neq 0$$

进行检验. 若拒绝 H_0 , 则认为 Y 与 X 存在线性关系, 所求的线性回归方程有意义; 若接受 H_0 , 则认为 Y 与 X 不存在线性关系, 所求的回归方程无意义.

可以证明(略)有下列等式

$$L_{yy} = Q + U,$$

其中

$$L_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 \text{ 称为离差平方和,}$$

$$Q = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \text{ 称为残差平方和,}$$

$$U = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 L_{xx} \text{ 称为回归平方和.}$$

上式表明, Y 的 n 个观测值的离散程度(L_{yy})可分解成两部分: 前一部分是由于实际观测值没有落在回归直线上引起的, 后一部分则是由回归系数引起的. 因此, 若 U 越大, Q 就越小, 表示 Y 与 X 的线性关系就越明显; 反之, U 越小, Q 就越大, 表示 Y 与 X 的线性关系就越不明显. 故我们找到了判断 Y 与 X 线性相关程度的(定性)准则: 如果 U/Q 较大, 则 Y 与 X 的线性关系显著, 此时应拒绝 H_0 ; 反之; 如果 U/Q 较小, 则 Y 与 X 的线性关系不显著, 此时应接受 H_0 ; 也就是说, 当 H_0 不成立时, U/Q 的值应较大, 因而 $\frac{U}{Q/(n-2)}$ 的值也较大. 可以证明, 当 H_0 成立时,

$$F = \frac{U}{Q/(n-2)} \sim F(1, n-2),$$

故对给定的显著性水平 α , 检验的(定量)准则为:

当 $F > F_{\alpha}(1, n-2)$ 时, 拒绝 H_0 , 否则就接受 H_0 .

例 7.2.2 对例 7.2.1 所求的线性回归方程进行显著性检验, $\alpha = 0.05$.

解 利用表中数据经计算得

$$\begin{aligned} L_{yy} &= 316.4, \quad U = \hat{\beta}_1^2 L_{xx} = 261.3, \\ Q &= L_{yy} - U = 55.1, \end{aligned}$$

故

$$F = \frac{U}{Q/(n-2)} = 37.9 > 5.32 = F_{0.05}(1, 8),$$

所以拒绝 H_0 , 即认为 Y 与 X 的线性关系显著, 所求的线性回归方程是有效的.

对回归方程的显著性检验还有 t 检验法和 r 检验法, 它们与上面讲的 F 检验法是等价的. 对它们就不再介绍了.

4. 预测

若经过检验拒绝了 $H_0: \beta_1 = 0$, 则表明 Y 与 X 的线性关系显著, 也就是说所得的线性回归方程能反映变量 Y 和 X 之间的内在联系. 因此对于给定的 $x = x_0$, 我们可利用回归方程 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 去预测相应的 $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$, 其中 $\varepsilon_0 \sim N(0, \sigma^2)$.

根据回归方程的意义, 我们可用 Y_0 的回归值 $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 作为 Y_0 的点预测值. 由于 $E(\hat{Y}_0) = E(Y_0)$, 故 $\hat{Y}_0$ 是 $E(Y_0)$ 的无偏估计 (它不是 Y_0 的无偏估计, 因为 Y_0 是随机变量不是未知参数). 另外可以证明

$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma} \sqrt{1 + 1/n + (x_0 - \bar{x})^2 / L_{xx}}} \sim t(n-2),$$

故给定显著性水平 α , Y_0 的置信水平为 $1 - \alpha$ 的预测区间为

$$[\hat{Y}_0 - \delta(x_0), \hat{Y}_0 + \delta(x_0)],$$

其中 $\delta(x_0) = \hat{\sigma} t_{\alpha/2}(n-2) \sqrt{1 + 1/n + (x_0 - \bar{x})^2 / L_{xx}}$, $\hat{\sigma} = \sqrt{Q/(n-2)}$.

例如, 对例 7.2.1, 给定 $x_0 = 17.5$, $\alpha = 0.05$, 可求得

$$\hat{Y}_0 = 65.7, \quad \delta(x_0) = 6.37,$$

故 Y_0 的置信水平为 95% 的预测区间为

$$[59.33, 72.07].$$

例 7.2.3 为确定广告费用 X 和销售额 Y 的关系, 现作一统计, 得资料如下表:

$X/\text{万元}$	40	25	20	30	40	40	25	20	50	20	50	50
$Y/\text{万元}$	490	395	420	475	385	525	480	400	560	365	510	540

- (1) 求销售额 Y 对广告费 X 的线性回归方程;
 (2) 对于显著性水平 $\alpha = 0.05$, 检验假设 $H_0: b = 0$, 其中 b 是回归直线的斜率;

(3) 当广告费 $x_0 = 35$ 时, 求销售额 Y 的点预测值和预测区间.

解 假设 Y 和 X 有如下线性相关关系:

$$Y = a + bX + \varepsilon, \text{ 其中 } \varepsilon \sim N(0, \sigma^2), \sigma^2 \text{ 未知.}$$

(1) 用最小二乘法求 a 和 b 的估计值 $\hat{a}$ 和 $\hat{b}$. 经计算

$$\bar{x} = 34.1667, \bar{y} = 462.0833,$$

$$\sum_{i=1}^{12} x_i^2 = 15\,650, \sum_{i=1}^{12} x_i y_i = 196\,325, \sum_{i=1}^{12} y_i^2 = 2\,610\,925;$$

$$L_{xx} = \sum_{i=1}^{12} (x_i - \bar{x})^2 = \sum_{i=1}^{12} x_i^2 - 12\bar{x}^2 = 1\,641.639,$$

$$L_{xy} = \sum_{i=1}^{12} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{12} x_i y_i - 12\bar{x}\bar{y} = 6\,870.662;$$

$$\therefore \hat{b} = \frac{L_{xy}}{L_{xx}} = 4.1853, \hat{a} = \bar{y} - \hat{b}\bar{x} = 319.0854.$$

于是得到 Y 对 X 的线性回归方程为

$$\hat{y} = 319.0854 + 4.1853x.$$

(2) 利用表中数据经计算得

$$L_{yy} = \sum_{i=1}^{12} y_i^2 - 12\bar{y}^2 = 48\,673.286,$$

$$U = \hat{b}^2 L_{xx} = 28\,756.157, Q = L_{yy} - U = 19\,917.129,$$

为检验假设 $H_0: b = 0$, 利用 F 检验, 得统计量

$$F = \frac{U}{Q/(12-2)} = 14.44 > 4.96 = F_{0.05}(1, 10),$$

所以拒绝 H_0 , 即认为 Y 与 X 的线性关系显著.

(3) 将 $x_0 = 35$ 代入回归方程, 当广告费为 35 万元时, 销售额的预测值 $\hat{Y}_0 = 465.5709$, 给定显著性水平 $\alpha = 0.05$, Y_0 的置信水平为 0.95 的预测区间为

$$[\hat{Y}_0 - \delta(x_0), \hat{Y}_0 + \delta(x_0)],$$

其中 $\delta(x_0) = \hat{\sigma} t_{\alpha/2}(n-2) \sqrt{1 + 1/n + (x_0 - \bar{x})^2/L_{xx}}$, $\hat{\sigma} = \sqrt{Q/(n-2)}$. 经计算得 $\delta(35) = 103.5175$. 最后得广告费为 35 万元时销售额的置信水平为 0.95 的预测区间为 $[362.0534, 569.0884]$.

§ 7.3 单因素方差分析

在科学实验和生产实践中,经常要考虑不同的实验条件是否对实验结果有显著的影响.我们知道,影响实验结果的因素往往很多.例如,在化工生产中,有原料成分、原料剂量、催化剂、反应温度、反应压力、反应时间、环境状况等因素,其中每一因素的变化都可能影响产品的数量和质量.这些因素可分成两类:一类是人们可控制的因素,如原料成分、反应温度等;另一类是人们不能控制的因素,如环境状况.一种因素是属于可控因素还是不可控因素往往取决于科技发展的水平和企业的技术水平.例如,在粮食生产过程中,影响产量的施水量、施肥量等因素人们可加以控制,但目前还不能控制气温因素.

一般说来,随着科学技术的发展,人们可控制的因素会越来越多,但是不可控因素不会完全消除.

在可控因素中,有些因素对生产和实验影响较大,有些影响较小.为了使生产过程得以稳定,保证优质高产,有必要找出对产品数量和质量有显著影响的那些因素,以便找到最优的生产或实验条件.本章将介绍推断某个因素的影响是否显著的方法——方差分析.方差分析是英国统计学家费希尔(R. A. Fisher)在20世纪20年代创立的,它最早源于对农业生产试验结果的分析,尔后被成功地应用于工业、气象预报、医学、生物学等许多领域.

1. 问题与基本概念

为了研究各种可控因素对实验结果的影响程度,首先要进行试验,以获取有关的数据.我们所关心的试验结果称为**试验指标**,如单位产量;试验中需要考察的、可以控制的条件称为**因素(因子)**,如施肥量等,常记作 $A, B, C, \dots$.而因素的每一状态(或等级)称为因素的一个**水平**,因素 A 的 r 个水平记为 $A_1, A_2, \dots, A_r$.

试验指标是所有因素综合作用的结果,为了考察某一种因素对试验指标的影响,往往把影响试验指标的其它可控因素固定,而把要考察的那个因素控制在几个不同状态或等级上进行试验,这样的试验称为**单因素试验**.相应的方差分析称为**单因素方差分析**.

例 7.3.1 (灯丝的配料方案优选)某灯泡厂用四种不同配料方案制成的灯丝,生产了四批灯泡.在每批灯泡中随机地抽取若干个灯泡测其使用寿命(单位:h),得到下列数据(表 7.3.1):

表 7.3.1

使用寿命 灯丝	灯泡样本数	1	2	3	4	5	6	7	8
甲		1 600	1 610	1 650	1 680	1 700	1 700	1 780	
乙		1 500	1 640	1 400	1 700	1 750			
丙		1 640	1 550	1 600	1 620	1 640	1 600	1 740	1 800
丁		1 510	1 520	1 530	1 570	1 640	1 680		

试问用这四种灯丝生产的灯泡的使用寿命有无显著性差异？

这里灯泡的使用寿命为试验指标，灯丝的配料为因素，它有四个水平，这是一个单因素四水平试验。

显然，每个灯泡的使用寿命都是由灯丝因素和其它不可控因素（随机因素）共同决定的。就甲种灯丝生产的全部灯泡而言，它们的灯丝是相同的，其使用寿命的差异完全是由随机因素造成的。这些随机因素种类很多，可以假定任何一种随机因素对灯泡使用寿命都不起决定性作用。因此，由甲种灯丝生产的全部灯泡可视为一正态总体 $X_1 \sim N(\mu_1, \sigma^2)$ ，而调查的 7 个数据是来自于该总体的 7 个样本，分别记为 $X_{11}, X_{12}, \dots, X_{17}$ 。同样由乙、丙、丁这三种灯丝生产的全部灯泡可分别视为总体 $X_2 \sim N(\mu_2, \sigma^2)$ ， $X_3 \sim N(\mu_3, \sigma^2)$ 及 $X_4 \sim N(\mu_4, \sigma^2)$ ，而调查的相应数据是来自于它们的样本。这里，四个总体均值不一致是由于生产它们的灯丝是不同的，而相同的方差是因为影响它们的随机因素是相同的。检验四种灯泡的使用寿命有无显著性差异就是检验四总体的均值 μ_i 是否相同。

基于上述分析，我们把问题重新叙述为：

设 X_1, X_2, X_3 及 X_4 分别表示四种灯泡的使用寿命，假定 $X_i \sim N(\mu_i, \sigma^2)$ ($i=1, 2, 3, 4$)。现从总体 X_i 中抽取容量为 n_i 的样本 $X_{i1}, X_{i2}, \dots, X_{in_i}$ ($i=1, 2, 3, 4, n_1=7, n_2=5, n_3=8, n_4=6$)，并且假定所有样本 X_{ij} 之间是相互独立的，欲对假设

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4, H_1: \mu_1, \mu_2, \mu_3, \mu_4 \text{ 不全相等}$$

作显著性检验。

由此可见，方差分析就是检验具有相同方差的多个正态总体的均值是否相同的一种统计方法。

2. 数学模型

设因素 A 有 r 个水平 $A_1, A_2, \dots, A_r$ ，水平 A_i 条件下的试验结果全体记为总体 X_i ，假定 $X_i \sim N(\mu_i, \sigma^2)$ ($i=1, 2, \dots, r$)。在水平 A_i 条件下进行 n_i ($n_i \geq 2$) 次独立试验，相当于从总体 X_i 抽取了容量为 n_i 的样本 $X_{i1}, X_{i2}, \dots, X_{in_i}$ ($i=1,$

2, ..., r), 假定全体 X_{ij} 之间相互独立, 即

$$\begin{cases} X_{ij} \sim N(\mu_i, \sigma^2), \\ X_{ij} \text{ 相互独立, } i=1, 2, \dots, r, j=1, 2, \dots, n_i, \\ \mu_i, \sigma^2 \text{ 未知.} \end{cases} \quad (7.3.1)$$

要检验统计假设

$$H_0: \mu_1 = \mu_2 = \dots = \mu_r, \quad H_1: \mu_1, \mu_2, \dots, \mu_r \text{ 不全相等.}$$

这就是单因素方差分析的数学模型.

在水平 A_i 下, 试验结果 $X_{ij} \sim N(\mu_i, \sigma^2)$, μ_i 表示水平 A_i 对试验结果的影响, 而 $\varepsilon_{ij} \triangleq X_{ij} - \mu_i$ 表示随机因素对试验结果的影响, 称 ε_{ij} 为随机误差, 从而有 $X_{ij} = \mu_i + \varepsilon_{ij}$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, 并且彼此相互独立.

模型(7.3.1)等价于下列模型:

$$\begin{cases} X_{ij} = \mu_i + \varepsilon_{ij}, \\ \varepsilon_{ij} \sim N(0, \sigma^2) \text{ 且相互独立,} \\ \mu_i, \sigma^2 \text{ 未知, } i=1, 2, \dots, r, j=1, 2, \dots, n_i. \end{cases} \quad (7.3.2)$$

要检验假设

$$H_0: \mu_1 = \mu_2 = \dots = \mu_r, \quad H_1: \mu_1, \mu_2, \dots, \mu_r \text{ 不全相等.}$$

为了便于讨论, 我们引入因素各水平效应的概念, 并对模型作等价变形, 记

$$n = \sum_{i=1}^r n_i, \quad \mu = \frac{1}{n} \sum_{i=1}^r n_i \mu_i,$$

$$\alpha_i = \mu_i - \mu, \quad i=1, 2, \dots, r.$$

显然, 符号 n 表示样本的总个数, 即样本总容量; 而 μ 表示 n 个样本均值的平均数, 称为理论总均值; 由 α_i 的定义知, α_i 表示总体 X_i 的均值与理论总均值的差异, 它反映了水平 A_i 对试验结果的影响大小(相对于总均值而言). 称 α_i 为水平 A_i 的效应. 易得

$$\mu_i = \mu + \alpha_i, \quad \sum_{i=1}^r n_i \alpha_i = 0,$$

从而 $H_0: \mu_1 = \mu_2 = \dots = \mu_r$ 等价于

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_r = 0.$$

于是模型(7.3.2)等价于下列模型:

$$\begin{cases} X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \\ \sum_{i=1}^r n_i \alpha_i = 0, \\ \varepsilon_{ij} \sim N(0, \sigma^2) \text{ 且相互独立,} \\ \mu, \alpha_i, \sigma^2 \text{ 未知, } i=1, 2, \dots, r, j=1, 2, \dots, n_i. \end{cases} \quad (7.3.3)$$

要检验假设

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0, \quad H_1: \alpha_1, \alpha_2, \cdots, \alpha_r \text{ 不全为 } 0.$$

3. 统计分析

由模型(7.3.3)知, 样本 X_{ij} 与 X_{kl} 的差异为 $(\alpha_i - \alpha_k) + (\varepsilon_{ij} - \varepsilon_{kl})$, 前一项表示因素取不同水平而引起的差异, 后一项表示由随机因素而引起的差异. 这启发我们, 全体 X_{ij} 之间的差异是否有类似的分解? 若有, 则当因素取不同水平对试验结果有显著性影响时, 前者应比后者显著偏大; 反之, 当影响不显著时, 前者就不会比后者偏大. 下面的分析过程就是基于这一条思路.

用 $\bar{X}_i$ 和 S_i^2 分别表示第 i 组样本的样本均值和样本方差, $\bar{X}$ 和 S^2 分别表示全体样本的样本均值和样本方差, 称为样本总均值和样本总方差. 引入总偏差平方和

$$S_T = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2,$$

S_T 反映了全体 X_{ij} 之间差异程度的大小. 由于

$$\begin{aligned} S_T &= \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 \\ &= \sum_{i=1}^r \sum_{j=1}^{n_i} [(X_{ij} - \bar{X}_i) + (\bar{X}_i - \bar{X})]^2 \\ &= \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 + 2 \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(\bar{X}_i - \bar{X}), \\ &\quad \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(\bar{X}_i - \bar{X}) = \sum_{i=1}^r (\bar{X}_i - \bar{X}) \left[\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i) \right] = 0, \end{aligned}$$

我们得到 S_T 的平方和分解式

$$S_T = S_e + S_A,$$

式中

$$\begin{aligned} S_e &= \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^r (n_i - 1) S_i^2, \\ S_A &= \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 = \sum_{i=1}^r n_i (\bar{X}_i - \bar{X})^2. \end{aligned}$$

显然 $\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$ 反映了第 i 组样本 $X_{i1}, X_{i2}, \cdots, X_{in_i}$ 之间的差异程度, 又因它们来自同一正态总体, 彼此的差异完全由随机因素所决定, 因而 S_e 反映了所有样本组内部的差异程度, 其值也完全由随机因素所决定, 称 S_e 为误差平方和或组内平方和. 下面的推导进一步验证了 S_e 的上述性质. 事实上, 由于 $X_{ij} \sim N(\mu_i, \sigma^2)$, 且 $X_{i1}, X_{i2}, \cdots, X_{in_i}$ 相互独立, 所以

$$X_{ij} - \bar{X}_i \sim N\left(0, \frac{n_i - 1}{n_i} \sigma^2\right),$$

$$E(X_{ij} - \bar{X}_i)^2 = 0^2 + \frac{n_i - 1}{n_i} \sigma^2 = \frac{n_i - 1}{n_i} \sigma^2,$$

$$\begin{aligned} E(S_e) &= E\left[\sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2\right] = \sum_{i=1}^r \sum_{j=1}^{n_i} E(X_{ij} - \bar{X}_i)^2 \\ &= \sum_{i=1}^r \sum_{j=1}^{n_i} \frac{n_i - 1}{n_i} \sigma^2 = \sum_{i=1}^r (n_i - 1) \sigma^2 = (n - r) \sigma^2. \end{aligned}$$

同样 $\sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 = n_i (\bar{X}_i - \bar{X})^2$ 反映了第 i 组样本均值和样本总均值的差异大小. 当 H_0 成立时, 全体样本可视为来自于同一正态总体, 此时, 该差异完全由随机因素所引起; 当 H_0 不成立时, 该差异除受到随机因素影响之外, 还受到因素水平 A_i 的效应 α_i 的影响. 因而 S_A 除反映随机因素引起的波动之外, 还反映了因素不同水平效应的差异所引起的波动, 称 S_A 为因素平方和或组间平方和. 下面的分析再次验证 S_A 的这种性质. 事实上, 由于 $\bar{X}_i \sim N\left(\mu + \alpha_i, \frac{\sigma^2}{n_i}\right)$, 且 $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_r$ 相互独立, 所以

$$\bar{X} = \frac{1}{n} \sum_{i=1}^r n_i \bar{X}_i \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

$$\bar{X}_i - \bar{X} \sim N\left(\alpha_i, \frac{n - n_i}{nn_i} \sigma^2\right),$$

$$E(\bar{X}_i - \bar{X})^2 = \alpha_i^2 + \frac{n - n_i}{nn_i} \sigma^2,$$

$$\begin{aligned} E(S_A) &= E\left[\sum_{i=1}^r n_i (\bar{X}_i - \bar{X})^2\right] \\ &= \sum_{i=1}^r n_i \left(\alpha_i^2 + \frac{n - n_i}{nn_i} \sigma^2\right) \\ &= (r - 1) \sigma^2 + \sum_{i=1}^r n_i \alpha_i^2. \end{aligned}$$

下面给出假设 H_0 的检验过程. 由于

$$E\left(\frac{S_e}{n - r}\right) = \sigma^2, \quad E\left(\frac{S_A}{r - 1}\right) = \sigma^2 + \frac{1}{r - 1} \sum_{i=1}^r n_i \alpha_i^2.$$

因此, 当 H_0 成立时, 统计量 $S_e/(n - r)$ 与 $S_A/(r - 1)$ 都是 σ^2 的无偏估计, 从而比值 $\frac{S_A/(r - 1)}{S_e/(n - r)}$ 应接近于 1; 反之 H_0 不成立时, 有 $E\left(\frac{S_A}{r - 1}\right) > \sigma^2$, 因而比值

$\frac{S_A/(r-1)}{S_e/(n-r)}$ 应比 1 明显的偏大, 故统计量 $F = \frac{S_A/(r-1)}{S_e/(n-r)}$ 可作为 H_0 的检验统计量. 当 H_0 成立时,

$$F = \frac{S_A/(r-1)}{S_e/(n-r)} \sim F(r-1, n-r).$$

根据第 6 章的内容知, 对给定的显著性水平 α , 检验准则为:

当 $F > F_{\alpha}(r-1, n-r)$ 时, 拒绝 H_0 , 即认为因素取不同水平对试验结果有显著影响, 否则接受 H_0 . 由上述步骤可得如下的单因素方差分析表.

表 7.3.2 单因素方差分析表

方差来源	平方和	自由度	方差	F 值
组间(因素 A)	S_A	$r-1$	$S_A/(r-1)$	$\frac{S_A/(r-1)}{S_e/(n-r)}$
组内(误差)	S_e	$n-r$	$S_e/(n-r)$	
总和	S_T	$n-1$		

计算 S_T , S_e , S_A 时常用下列公式:

$$S_T = \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij}^2 - \frac{1}{n} \left(\sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij} \right)^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij}^2 - n \bar{X}^2,$$

$$S_A = \sum_{i=1}^r \frac{1}{n_i} \left(\sum_{j=1}^{n_i} X_{ij} \right)^2 - \frac{1}{n} \left(\sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij} \right)^2 = \sum_{i=1}^r n_i \bar{X}_i^2 - n \bar{X}^2,$$

$$S_e = S_T - S_A.$$

有时对数据 X_{ij} 施以线性变换 $Y_{ij} = b(X_{ij} - a)$, 其中 a, b 为适当常数且 $b \neq 0$, 使得 Y_{ij} 较为简单. 易证根据 Y_{ij} 算出的 F 值不变. 特别 $b=1$ 时, S_e 和 S_A 的值也不变.

最后指出: (1) 利用方差分析检验因素取不同水平对试验结果的影响是否显著, 是相对意义下的“显著”, 是相对随机因素而言的. 如果随机因素的影响很小, 即使判断出因素不同水平的影响具有“显著性”, 其绝对意义下的影响也可能较小.

(2) 当 $\alpha=0.01$ 时否定 H_0 , 称因素 A 的影响高度显著, 用符号 “**” 表示. 当 $\alpha=0.01$ 时接受 H_0 , 而 $\alpha=0.05$ 时否定 H_0 , 称因素 A 的影响显著. 用符号 “*” 表示. 当 $\alpha=0.05$ 时, 接受 H_0 , 而 $\alpha=0.1$ 时否定 H_0 , 称因素 A 有一定影响. 反之, 称因素 A 无显著性影响.

(3) 可以估计模型中的未知参数 μ , α_i 及 $\sigma^2 (i=1, 2, \dots, r)$, 它们的无偏估计(点估计)分别为

$$\hat{\mu} = \bar{X}, \quad \hat{\alpha}_i = \bar{X}_i - \bar{X}, \quad \hat{\sigma}^2 = S_e / (n - r).$$

亦可给出它们的区间估计(略).

4. 例子

例 7.3.1 的计算结果如表 7.3.3. 其中 $r=4$, $n_1=7$, $n_2=5$, $n_3=8$, $n_4=6$, $n=26$.

表 7.3.3

方差来源	平方和	自由度	方差	F 值
因素	39 776.4	3	13 258.8	1.638
误差	178 089	22	8 095	
总和	217 865.4	25		

对于给定的 $\alpha=0.1$, $F_{0.1}(3,22)=2.35$, 因 $1.638 < 2.35$, 故接受 H_0 , 即认为四种灯丝生产的灯泡其平均使用寿命无显著性差异.

例 7.3.2 某工厂用五台机器生产同一种产品, 今对每种机器观察三天的日产量, 测得数据如表 7.3.4, 这里假定每台机器的日产量服从方差相等的正态分布. 试分析五台机器的日产量是否存在显著的差异.

表 7.3.4

日产量 机器	时期	1	2	3
A		90	92	88
B		91	93	92
C		96	93	93
D		84	83	88
E		84	86	82

解 这是一个单因素方差分析问题, 其中 $r=5$, $n_1=n_2=n_3=n_4=n_5=3$, $n=15$.

(1) 作变换 $Y_{ij} = X_{ij} - 90$ 得到关于 Y_{ij} 的表(略).

(2) 利用 Y_{ij} 计算得

$$S_T = \sum_{i=1}^5 \sum_{j=1}^3 Y_{ij}^2 - \frac{1}{15} \left(\sum_{i=1}^5 \sum_{j=1}^3 Y_{ij} \right)^2 = 281 - \frac{(-15)^2}{15} = 266,$$

$$S_A = \sum_{i=1}^5 \frac{1}{3} \left(\sum_{j=1}^3 Y_{ij} \right)^2 - \frac{1}{15} \left(\sum_{i=1}^5 \sum_{j=1}^3 Y_{ij} \right)^2 = 243 - \frac{(-15)^2}{15} = 228,$$

$$S_e = S_T - S_A = 38.$$

(3) 方差分析表如表 7.3.5,

表 7.3.5

方差来源	平方和	自由度	方差	F 值
因 素	228	4	57	15
误 差	38	10	3.8	
总 和	266	14		

由于 $F_{0.01}(4, 10) = 5.99$, 故 $F > F_{0.01}(4, 10)$, 因而五台机器的日产量存在显著的差异.

例 7.3.3 某公司欲研究 3 种内容的广告宣传对某种大型机械销售量的影响, 第 1 种广告强调运输方便性, 第 2 种广告强调节省燃料, 第 3 种广告强调噪音低. 在经过广告广泛宣传后, 抽取了一部分订单, 按订单上的订购量计算的销售量如下表:

表 7.3.6

	第 1 季 度	第 2 季 度	第 3 季 度	第 4 季 度
广 告 1	163	176	170	185
广 告 2	184	198	179	190
广 告 3	206	191	218	224

试根据以上资料判断广告内容不同对销售量是否有显著影响($\alpha = 0.05$).

解 在本题中广告内容是因素, 每一种广告表示一个水平, 共 3 个水平, 分别以 μ_1, μ_2, μ_3 记 3 种不同内容广告下订购量的平均值, 要检验 $H_0: \mu_1 = \mu_2 = \mu_3$. 现在 $r = 3, n_1 = n_2 = n_3 = 4, n = 12$, 计算结果列入下表:

表 7.3.7

水 平	$\sum_{j=1}^4 x_{ij}$	$\left(\sum_{j=1}^4 x_{ij}\right)^2$	$\sum_{j=1}^4 x_{ij}^2$
A_1	694	481 636	120 670
A_2	751	564 001	141 201
A_3	839	703 921	176 617
Σ	2 284	1 749 558	438 488

$$S_T = 438\,488 - \frac{1}{12} \times 2\,284^2 = 3\,766.67,$$

$$S_A = \frac{1}{4} \times 1\,749\,558 - \frac{1}{12} \times 2\,284^2 = 2\,668.17,$$

$$S_e = S_T - S_A = 1\,098.5.$$

方差分析表如下:

表 7.3.8

方差来源	平方和	自由度	方差	F 值	临界值
因素	2 668. 17	2	1 334. 085	10. 93	$F_{0.05}(2,9) = 4.26$
误差	1 098. 5	9	122. 056		
总和	3 766. 67	11			

由于 $F > F_{\alpha}(r-1, n-r)$, 故拒绝 $H_0: \mu_1 = \mu_2 = \mu_3$, 认为广告内容对销售量有显著影响.

§ 7.4 用 MATLAB 处理回归与方差分析

MATLAB 在其统计工具箱中提供了一组专门用于处理一元、多元、线性、非线性等回归问题和单因素、双因素方差分析的函数, 本节仅介绍其中用于多项式回归的函数 `polyfit`, `polyconf` 和用于单因素方差分析的函数 `anova1`. 多项式回归是指回归函数是回归变量的多项式的回归. 由于实际应用中一般函数都可用多项式逼近, 因此多项式回归有广泛的应用.

1. `polyfit` 函数

该函数的调用格式为

$$P = \text{polyfit}(x, y, n),$$

$$[P, S] = \text{polyfit}(x, y, n).$$

第一种格式的功能是在最小二乘意义下, 将观测数据 $x = \{x_1, x_2, \dots, x_m\}$, $y = \{y_1, y_2, \dots, y_m\}$ 拟合成次数为 n 的多项式

$$P(x) = P_1 x^n + P_2 x^{n-1} + \dots + P_n x + P_{n+1}.$$

返回结果 P 为此多项式按降幂次排列的系数向量, 即

$$P = \{P_1, P_2, \dots, P_n, P_{n+1}\}.$$

输入参数 n 用来指定所选多项式的次数. 当取 $n = 1$ 时, `polyfit` 计算的正是一元线性回归.

第二种格式中除返回多项式系数向量 P 外, 还返回矩阵 S . 该矩阵主要用于 `polyconf` 函数中, 用来进行多项式预测和置信区间评估. 因此如果无意进行预测和误差估计的话, 则只需使用第一种调用格式.

2. `polyconf` 函数

该函数的调用格式为:

$$[Y, \text{delta}] = \text{polyconf}(P, x, S, \alpha).$$

该函数的功能是进行多项式预测和置信区间评估. 输入参数中的 P 和 S 是多项式回归函数 `polyfit` 的输出结果. P 是向量, 其元素为按降次幂排列的多项式的系数. x 为矩阵或向量. `polyconf` 计算在 x 的每个点处多项式 P 的预测值 Y .

假设 `polyfit` 输入数据的误差服从方差为常值的独立正态分布, 则 `DELTA` 给出 Y 的置信度为 $100(1 - \alpha)\%$ 的预测区间

$$[Y - \text{delta}, Y + \text{delta}].$$

α 的缺省值为 0.05, 此时 $[Y - \text{delta}, Y + \text{delta}]$ 给出的是 95% 的置信区间.

3. anova 函数

该函数的调用格式为:

$$P = \text{anova}(X),$$

$$P = \text{anova}(x, \text{grnp}).$$

第一种格式对样本矩阵 X 中的各列数据进行均衡的单因素方差分析, 以比较各列的均值. 函数返回 H_0 (即 X 中各列的均值相同) 假设成立的概率值.

第二种格式对样本向量 x 中由向量 `grnp` 索引的两组或多组数据进行单因素方差分析. 输入参数 `grnp` 标明向量 x 中相应元素的组别. `grnp` 元素的值为整数, 最大值为需要比较的不同组的数量, 最小值为 1. 每组中至少有一个元素, 但并不要求各组中元素个数相同, 因此适用于数据不均衡的情况. 此时函数同样返回 H_0 假设成立的概率值.

`anova` 函数同时显示出样本数据的单因素方差分析表(表 7.3.2)和 box 图.

4. 计算举例

例 7.4.1 现有一组统计数据(参见例 7.2.1):

X	17.1	16.3	18.4	19.0	15.8	16.0	17.5	17.8	18.0	16.4
Y	65	63	70	75	60	55	64	69	65	58

试用 MATLAB 求 Y 与 X 的一元线性回归方程, 并求 $x_0 = 17.5$ 时, Y 的预测值 $\hat{Y}_0$ 和 Y_0 的置信水平为 95% 的置信区间.

解 首先输入统计数据:

$$X = [17.1 \ 16.3 \ 18.4 \ 19.0 \ 15.8 \ 16.0 \ 17.5 \ 17.8 \ 18.0 \ 16.4]; \downarrow$$

$$Y = [65 \ 63 \ 70 \ 75 \ 60 \ 55 \ 64 \ 69 \ 65 \ 58]; \downarrow$$

求线性回归方程:

$$[P, S] = \text{polyfit}(X, Y, 1); \downarrow$$

$$\text{disp}(P) \downarrow$$

$$4.9600 \quad -21.0605$$

得一元线性回归方程为

$$y = 4.96x - 21.0605.$$

将统计数据的散点图和回归方程的直线图画在一幅图形中, 其中统计数据点用“+”号标示:

```
X1 = 15:0.5:20; ↓
Y1 = polyconf(P,X1); ↓
plot(X,Y,'b +') ↓
hold on ↓
plot(X1,Y1) ↓
```

得图形如图 7.4.1.

对 $x_0 = 17.5, \alpha = 0.05$, 计算 Y 的预测值 $\hat{Y}_0$:

```
X0 = 17.5; ↓
alpha = 0.05; ↓
[Y0,delta] = polyconf(P,X0,S,alpha) ↓
Y0 =
```

65.7392

delta =

6.3675

计算 Y_0 的 95% 的置信区间:

```
confidence_interval = [Y0 - delta, Y0 + delta] ↓
confidence_interval =
59.3717    72.1067
```

例 7.4.2 用 `anova` 函数求解例 7.3.1 中的问题.

解 首先将表 7.3.1 中的样本数据存入向量 X 中:

```
X = [1600 1610 1650 1680 1700 1700 1780 ... ↓
     1500 1640 1400 1700 1750 ... ↓
     1640 1550 1600 1620 1640 1600 1740 1800 ... ↓
     1510 1520 1530 1570 1640 1680]; ↓
```

然后输入索引向量 g :

```
g = [1 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 3 3 3
     4 4 4 4 4 4 ]; ↓
```

其中的 1, 2, 3, 4 分别对应于例 7.3.1 中的甲、乙、丙、丁. 用单因素方差分析函数 `anova` 进行计算:

```
P = anova(X,g) ↓
```

P =

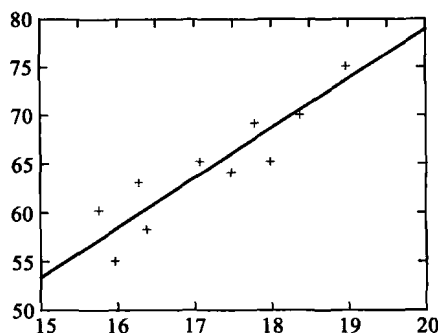


图 7.4.1 统计数据点及其回归直线

0.2094

于是知 H_0 假设成立的概率为 0.2094. 同时得到单因素方差分析表与表 7.3.3 相同(该表及其 box 图省略).

习 题 7

1. 证明一元线性回归模型

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

的参数 β_0 和 β_1 的最小二乘估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是无偏估计, 其中模型的随机误差 ε 满足 $E(\varepsilon) = 0$.

2. 假定 X 是可控变量(非随机变量), Y 是服从正态分布的随机变量, 变量 X 和变量 Y 存在相关关系. 今测得下列 10 组数据:

X	2	3	4	5	6	8	10	12	14	16
Y	15	20	25	30	35	45	60	80	82	105

(1) 作散点图, 能否认为 Y 与 X 间存在线性相关关系?

(2) 写出 Y 关于 X 的一元线性回归模型;

(3) 求出 Y 关于 X 的一元线性回归方程;

(4) 用 F 检验法对回归方程的显著性进行检验; ($\alpha = 0.05$)

(5) 当 $x_0 = 7$ 时, 求 Y_0 的置信度水平为 0.95 的置信区间.

3. 假设儿子的体重 Y 和父亲的体重 X 适合一元线性回归模型, 且随机误差 $\varepsilon \sim N(0, \sigma^2)$. 观测了 11 对父子的体重(单位: kg)如下表:

X	60	62	64	65	66	67	68	70	72	73	69
Y	63.6	65.2	66	65.5	67	67.5	67.2	63.5	71.5	72.6	70

(1) 求 Y 与 X 之间的回归方程;

(2) 对回归方程进行显著性检验; ($\alpha = 0.05$)

(3) 给定 $x_0 = 71$ 时, 求 Y_0 的置信水平为 95% 的置信区间.

4. 证明回归方程 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 过点 $(\bar{x}, \bar{y})$.

5. 证明 $Q = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n y_i^2 - \hat{\beta}_0 \sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i y_i$. (用此结果计算残差平方和 Q 十分方便.)

6. 假定某企业的某种产品产量与单位成本资料如下:

月 份	1	2	3	4	5	6
产量 x /千件	2	3	4	3	4	5
单位成本 y /元	73	72	71	73	69	68

- (1) 试确定回归直线;
 (2) 指出产量每增加 1 000 件时, 单位成本下降多少?
 (3) 假定产量为 6 000 件时, 单位成本是多少? 单位成本为 70 元时, 产量应为多少件?

7. 某长途运输公司在同一类型的卡车中, 对行驶公里数 y 与行驶天数 x 进行统计, 得数据如下:

$x/\text{天}$	3.5	1.0	4.0	2.0	1.0	3.0	4.5	1.5	3.0	5.0
y/km	825	215	1 070	550	480	920	1 350	325	670	1 215

过原点的回归模型为

$$\begin{cases} y_i = \beta x_i + \varepsilon_i, & i = 1, 2, \dots, n, \\ \text{各 } \varepsilon_i \text{ 相互独立, 且均服从 } N(0, \sigma^2). \end{cases}$$

试求 β 的最小二乘估计, 用这一方法建立本题 y 关于 x 的过原点的回归方程.

8. 粮食加工厂试验 5 种贮藏粮食的方法, 检验它们对粮食的含水率是否有不同的影响. 在贮藏前这些粮食的含水率几乎相同, 贮藏后其含水率如下表. 假定每种贮藏方法的粮食含水率服从方差相等的正态分布, 问不同贮藏方法对含水率的影响是否有明显的差异? ($\alpha = 0.05$)

含 水 率		试 验 批 号				
		1	2	3	4	5
因素 A(贮藏方法)	A_1	7.3	8.3	7.6	8.4	8.3
	A_2	5.4	7.4	7.1		
	A_3	8.1	6.4			
	A_4	7.9	9.5	10.0		
	A_5	7.2	7.8	9.0		

9. 某中学一年级有三个班, 他们进行了一次英语测验, 现从三个班中分别随机抽取 10, 12, 11 名同学记录其成绩如下:

I: 65 81 92 83 74 50 69 77 85 66

II: 84 75 63 95 47 86 67 93 50 77 80 68

III: 76 87 60 94 55 71 74 65 83 71 62

设各班成绩服从方差相等的正态分布, 试问各班的平均分数有无显著差异? ($\alpha = 0.05$)

10. 一试验用来比较 4 种不同药品解除外科手术后疼痛和延续时间(单位:h), 结果如下表:

药 品	时间长度/h			
A	8	6	4	2
B	6	6	4	4
C	8	10	10	10
D	4	4	2	

试在显著性水平 $\alpha = 0.05$ 下检验各种药品对解除疼痛的延续时间有无显著差异.

附表

附表 1 随机数表

03 47 43 73 86	36 96 47 36 61	46 98 63 71 62	33 26 16 80 45	60 11 14 10
97 74 24 67 62	42 81 14 57 20	42 53 32 37 32	27 07 36 07 51	24 51 79 89
16 76 62 27 66	56 50 26 71 07	32 90 79 78 53	13 55 38 58 59	88 97 54 14
12 56 85 99 26	96 96 68 27 31	05 03 72 93 15	57 12 10 14 21	88 26 49 81
55 59 56 35 64	38 54 82 46 22	31 62 43 09 90	06 18 44 32 53	23 83 01 30
16 22 77 94 39	49 54 43 54 82	17 37 93 23 78	87 35 20 96 43	84 26 34 91
84 42 17 53 31	57 24 55 06 88	77 04 74 47 67	21 76 33 50 25	83 92 12 06
63 01 63 78 59	16 95 55 67 19	98 10 50 71 75	12 86 73 58 07	44 39 52 38
33 21 12 34 29	78 64 56 07 82	52 42 07 44 38	15 51 00 13 42	99 66 02 79
57 60 86 32 44	09 47 27 96 54	49 17 46 09 62	90 52 84 77 27	08 02 73 43
18 18 07 92 46	44 17 16 58 09	79 83 86 19 62	06 76 50 03 10	55 23 64 05
26 62 38 97 75	84 16 07 44 99	83 11 46 32 24	20 14 85 88 45	10 93 72 88
23 42 40 64 74	82 97 77 77 81	07 45 32 14 08	32 98 94 07 72	93 85 79 10
52 36 28 19 95	50 92 26 11 97	00 56 76 31 38	80 22 02 53 53	86 60 42 04
37 85 94 35 12	83 39 50 08 30	42 34 07 96 88	54 22 06 87 98	35 85 29 48
70 29 17 12 13	40 33 20 38 26	13 89 51 03 74	17 76 37 13 04	07 74 21 19
56 62 18 37 35	96 83 50 87 75	97 12 25 93 47	70 33 24 03 54	97 77 46 44
99 49 57 22 77	88 42 95 45 72	16 64 36 16 00	04 43 18 66 79	94 77 24 11
16 08 15 04 72	33 27 14 34 09	45 59 34 68 49	12 72 07 34 45	99 27 72 95
31 16 93 32 43	50 27 89 87 19	20 15 37 00 49	52 85 66 60 44	38 68 88 11
68 34 30 13 70	55 74 30 77 40	44 22 78 84 26	04 33 46 09 52	68 07 97 06
74 57 25 65 76	59 29 97 68 60	71 91 38 67 54	13 58 18 24 76	15 54 55 95
27 42 37 86 53	48 55 90 65 72	96 57 69 36 10	96 46 92 42 45	97 60 49 04
00 39 68 29 61	66 37 32 20 30	77 84 57 03 29	10 45 65 04 26	11 04 96 67
29 94 98 94 24	68 49 69 10 82	53 75 91 93 30	34 25 20 57 27	40 48 73 51

附表 2 二项分布表

本表给出了累积二项分布的概率

$$\sum_{k=0}^r C_n^k p^k (1-p)^{n-k}, \quad r=0, 1, \dots, n-1.$$

		p									
n	r	0.01	0.05	0.10	0.20	0.25	0.30	0.333	0.40	0.50	
2	0	0.980 1	0.902 5	0.810 0	0.640 0	0.562 5	0.490 0	0.444 4	0.360 0	0.250 0	
	1	0.999 9	0.997 5	0.990 0	0.960 0	0.937 5	0.910 0	0.888 8	0.840 0	0.500 0	
3	0	0.970 3	0.857 4	0.729 0	0.512 0	0.421 9	0.343 0	0.296 3	0.216 0	0.125 0	
	1	0.999 7	0.992 8	0.972 0	0.896 0	0.843 8	0.784 0	0.740 7	0.648 0	0.500 0	
	2	1.000 0	0.999 9	0.999 0	0.992 0	0.984 4	0.973 0	0.962 9	0.936 0	0.875 0	
4	0	0.960 6	0.814 5	0.656 1	0.409 6	0.316 4	0.240 1	0.197 5	0.129 6	0.062 5	
	1	0.999 4	0.986 0	0.947 7	0.819 2	0.738 3	0.651 7	0.592 6	0.474 2	0.312 5	
	2	1.000 0	0.999 5	0.996 3	0.972 8	0.949 2	0.916 3	0.888 9	0.819 8	0.687 5	
	3		1.000 0	0.999 9	0.998 4	0.996 1	0.991 9	0.987 7	0.973 4	0.937 5	
5	0	0.951 0	0.773 8	0.590 5	0.327 7	0.237 3	0.168 1	0.131 7	0.077 8	0.031 2	
	1	0.999 0	0.977 4	0.918 5	0.737 3	0.623 8	0.528 3	0.460 9	0.337 0	0.187 4	
	2	1.000 0	0.998 8	0.991 4	0.942 1	0.896 5	0.837 0	0.790 1	0.682 6	0.499 9	
	3		0.999 9	0.999 5	0.993 3	0.984 4	0.969 3	0.954 7	0.913 0	0.812 4	
	4		1.000 0	1.000 0	0.999 7	0.999 0	0.997 7	0.995 9	0.989 8	0.968 6	
6	0	0.941 5	0.735 1	0.531 4	0.262 1	0.178 0	0.117 6	0.087 8	0.046 7	0.015 6	
	1	0.998 6	0.967 2	0.885 7	0.655 3	0.534 0	0.420 1	0.351 2	0.233 3	0.109 4	
	2	1.000 0	0.997 7	0.984 1	0.901 1	0.830 6	0.744 2	0.680 4	0.544 3	0.343 8	
	3		0.999 8	0.998 7	0.983 0	0.962 4	0.929 4	0.899 9	0.820 8	0.656 3	
	4		0.999 9	0.999 9	0.998 4	0.995 4	0.988 9	0.982 2	0.959 0	0.890 7	
	5		1.000 0	1.000 0	0.999 9	0.999 8	0.999 1	0.998 7	0.995 9	0.984 5	
7	0	0.932 1	0.698 3	0.478 3	0.209 7	0.133 5	0.082 4	0.058 5	0.028 0	0.007 8	
	1	0.998 0	0.955 6	0.655 4	0.576 7	0.445 0	0.329 4	0.263 3	0.158 6	0.062 5	
	2	1.000 0	0.996 2	0.850 3	0.825 0	0.756 5	0.647 1	0.570 6	0.419 9	0.226 6	
	3		0.999 8	0.974 3	0.966 7	0.929 5	0.874 0	0.826 7	0.710 2	0.500 0	
	4		1.000 0	0.997 3	0.995 3	0.987 2	0.971 2	0.954 7	0.903 7	0.773 4	
	5			0.999 8	0.999 6	0.998 7	0.996 2	0.998 1	0.981 2	0.937 5	
	6			1.000 0	1.000 0	0.999 9	0.999 8	0.999 5	0.998 4	0.992 2	

续表

		<i>p</i>								
<i>n</i>	<i>r</i>	0.01	0.05	0.10	0.20	0.25	0.30	0.333	0.40	0.50
8	0	0.922 7	0.663 4	0.430 5	0.167 8	0.100 1	0.057 6	0.039 0	0.016 8	0.003 9
	1	0.997 3	0.942 7	0.813 1	0.503 3	0.367 1	0.255 3	0.195 1	0.106 4	0.035 2
	2	0.999 9	0.994 2	0.961 9	0.796 9	0.678 6	0.551 8	0.468 2	0.315 4	0.144 5
	3	1.000 0	0.999 6	0.995 0	0.943 7	0.886 2	0.805 9	0.741 3	0.594 1	0.363 3
	4		1.000 0	0.999 6	0.989 6	0.972 7	0.942 0	0.912 0	0.826 3	0.636 7
	5			1.000 0	0.998 8	0.995 8	0.988 7	0.980 3	0.950 2	0.855 5
	6				1.000 0	0.999 6	0.998 7	0.997 4	0.991 5	0.964 8
	7					1.000 0	0.999 9	0.999 8	0.999 3	0.996 1
9	0	0.913 5	0.630 2	0.387 4	0.134 2	0.075 1	0.040 4	0.026 0	0.010 1	0.002 0
	1	0.996 5	0.928 7	0.774 8	0.436 2	0.300 4	0.196 0	0.143 1	0.070 6	0.019 6
	2	0.999 9	0.991 6	0.947 0	0.738 2	0.600 7	0.462 8	0.377 2	0.231 8	0.089 9
	3	1.000 0	0.999 3	0.991 6	0.914 4	0.834 3	0.729 6	0.650 3	0.482 6	0.254 0
	4		0.999 9	0.999 0	0.980 5	0.951 1	0.901 1	0.855 1	0.733 4	0.500 1
	5		1.000 0	0.999 8	0.997 0	0.990 0	0.974 6	0.957 5	0.900 6	0.746 2
	6			0.999 9	0.999 8	0.998 7	0.995 6	0.991 6	0.974 9	0.910 3
	7			1.000 0	1.000 0	0.999 9	0.999 5	0.998 9	0.996 1	0.980 6
	8					1.000 0	0.999 9	0.999 8	0.999 6	0.998 2
10	0	0.904 4	0.598 7	0.348 7	0.107 4	0.056 3	0.028 2	0.017 3	0.006 0	0.001 0
	1	0.995 8	0.913 8	0.736 1	0.375 8	0.244 0	0.149 3	0.104 0	0.046 3	0.010 8
	2	1.000 0	0.988 4	0.929 8	0.677 8	0.525 6	0.382 8	0.299 1	0.167 2	0.054 7
	3		0.998 9	0.987 2	0.879 1	0.775 9	0.649 6	0.559 2	0.381 2	0.171 9
	4		0.999 9	0.998 4	0.967 2	0.921 9	0.849 7	0.786 8	0.632 0	0.377 0
	5		1.000 0	0.999 9	0.993 6	0.980 3	0.952 6	0.923 4	0.832 7	0.623 1
	6			1.000 0	0.999 1	0.996 5	0.989 4	0.986 3	0.944 2	0.828 2
	7				0.999 9	0.999 6	0.998 4	0.996 6	0.986 7	0.945 4
	8				1.000 0	1.000 0	0.999 8	0.999 6	0.997 3	0.989 3
	9						1.000 0	0.999 9	0.999 9	0.999 1

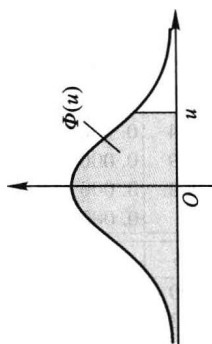
附表 3 泊松分布表

表中列出了 $\sum_{k=x}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!}$ 的值

x	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$
0	1.000 000 0	1.000 000 0	1.000 000 0	1.000 000 0	1.000 000 0	1.000 000 0	1.000 000 0
1	0.095 162 6	0.181 269 2	0.259 181 8	0.329 680 0	0.393 469	0.451 188	0.503 415
2	0.004 678 8	0.017 523 1	0.036 936 3	0.061 551 9	0.090 204	0.121 901	0.155 805
3	0.000 154 7	0.001 148 5	0.003 569 5	0.007 926 3	0.014 388	0.023 115	0.034 142
4	0.000 003 8	0.000 056 8	0.000 265 8	0.000 776 3	0.001 752	0.003 358	0.005 753
5		0.000 002 3	0.000 015 8	0.000 061 2	0.000 172	0.000 394	0.000 786
6		0.000 000 1	0.000 000 8	0.000 004 0	0.000 014	0.000 039	0.000 090
7				0.000 000 2	0.000 001	0.000 003	0.000 009
8							0.000 001

x	$\lambda = 0.8$	$\lambda = 0.9$	$\lambda = 1.0$	$\lambda = 1.2$	$\lambda = 1.4$	$\lambda = 1.6$	$\lambda = 1.8$
0	1.000 000 0	1.000 000 0	1.000 000 0	1.000 000 0	1.000 000 0	1.000 000 0	1.000 000 0
1	0.550 671	0.593 430	0.632 121	0.698 806	0.753 403	0.798 103	0.834 701
2	0.191 208	0.227 518	0.264 241	0.337 373	0.408 167	0.475 069	0.537 163
3	0.047 423	0.062 857	0.080 301	0.120 513	0.166 502	0.216 642	0.269 379
4	0.009 080	0.013 459	0.018 988	0.033 769	0.053 725	0.078 813	0.108 708
5	0.001 411	0.002 344	0.003 660	0.007 746	0.014 253	0.023 682	0.036 407
6	0.000 184	0.000 343	0.000 594	0.001 500	0.003 201	0.006 040	0.010 378
7	0.000 021	0.000 043	0.000 083	0.000 251	0.000 622	0.001 336	0.002 569
8	0.000 002	0.000 005	0.000 010	0.000 037	0.000 107	0.000 260	0.000 562
9			0.000 001	0.000 005	0.000 016	0.000 045	0.000 110
10				0.000 001	0.000 002	0.000 007	0.000 019
11						0.000 001	0.000 003

x	$\lambda = 2.5$	$\lambda = 3.0$	$\lambda = 3.5$	$\lambda = 4.0$	$\lambda = 4.5$	$\lambda = 5.0$
0	1.000 000	1.000 000	1.000 000	1.000 000	1.000 000	1.000 000
1	0.917 915	0.950 213	0.969 803	0.981 684	0.988 891	0.993 262
2	0.712 703	0.800 852	0.864 112	0.908 422	0.938 901	0.959 572
3	0.456 187	0.576 810	0.679 153	0.761 897	0.826 422	0.875 348
4	0.242 424	0.352 768	0.463 367	0.566 530	0.657 704	0.734 974
5	0.108 822	0.184 737	0.274 555	0.371 163	0.467 896	0.559 507
6	0.042 021	0.083 918	0.142 386	0.214 870	0.297 070	0.384 039
7	0.014 187	0.033 509	0.065 288	0.110 674	0.168 949	0.237 817
8	0.004 247	0.011 905	0.026 739	0.051 134	0.086 586	0.133 372
9	0.001 140	0.003 803	0.009 874	0.021 363	0.040 257	0.068 094
10	0.000 277	0.001 102	0.003 315	0.008 132	0.017 093	0.031 828
11	0.000 062	0.000 292	0.001 019	0.002 840	0.006 669	0.013 695
12	0.000 013	0.000 071	0.000 289	0.000 915	0.002 404	0.005 453
13	0.000 002	0.000 016	0.000 076	0.000 274	0.000 805	0.002 019
14		0.000 003	0.000 019	0.000 076	0.000 252	0.000 698
15		0.000 001	0.000 004	0.000 020	0.000 074	0.000 226
16			0.000 001	0.000 005	0.000 020	0.000 069
17				0.000 001	0.000 005	0.000 020
18					0.000 001	0.000 005
19						0.000 001



附表 4 标准正态分布表

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{1}{2}x^2} dx (u \geq 0)$$

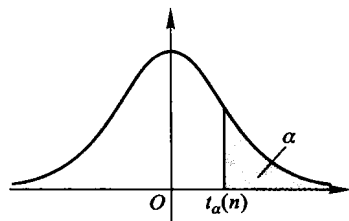
u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	u
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	0.0
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	0.1
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	0.2
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517	0.3
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	0.4
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224	0.5
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	0.6
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852	0.7
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	0.8
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	0.9
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	1.0
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	1.1
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9014	1.2
1.3	0.9032	0.9049	0.9065	0.9082	0.9098	0.9114	0.9130	0.9146	0.9162	0.9177	1.3
1.4	0.9192	0.9207	0.9222	0.9236	0.9250	0.9264	0.9278	0.9292	0.9306	0.9319	1.4
1.5	0.9332	0.9345	0.9357	0.9369	0.9382	0.9394	0.9406	0.9417	0.9429	0.9440	1.5
1.6	0.9452	0.9463	0.9474	0.9485	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	1.6
1.7	0.9555	0.9565	0.9575	0.9585	0.9595	0.9605	0.9615	0.9625	0.9635	0.9645	1.7
1.8	0.9655	0.9665	0.9675	0.9685	0.9695	0.9705	0.9715	0.9725	0.9735	0.9745	1.8
1.9	0.9755	0.9765	0.9775	0.9785	0.9795	0.9805	0.9815	0.9825	0.9835	0.9845	1.9
2.0	0.9775	0.9778	0.9781	0.9782	0.9783	0.9784	0.9785	0.9786	0.9787	0.9788	2.0
2.1	0.9821	0.9825	0.9828	0.9831	0.9834	0.9837	0.9840	0.9843	0.9846	0.9849	2.1

续表

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	u
2.2	0.986 10	0.986 45	0.986 79	0.987 13	0.987 45	0.987 78	0.988 09	0.988 40	0.988 70	0.988 99	2.2
2.3	0.989 28	0.989 56	0.989 83	0.990 07	0.990 35	0.990 63	0.990 93	0.991 16	0.991 44	0.991 76	2.3
2.4	0.991 82	0.992 04	0.992 24	0.992 41	0.992 56	0.992 78	0.993 03	0.993 24	0.993 41	0.993 63	2.4
2.5	0.993 90	0.993 63	0.993 32	0.992 97	0.992 44	0.991 84	0.991 10	0.990 26	0.989 26	0.988 01	2.5
2.6	0.995 39	0.994 73	0.993 94	0.992 97	0.991 55	0.989 55	0.986 93	0.983 67	0.979 82	0.975 39	2.6
2.7	0.996 33	0.995 36	0.993 94	0.991 84	0.988 82	0.984 82	0.979 26	0.972 06	0.963 32	0.953 01	2.7
2.8	0.997 45	0.996 23	0.994 29	0.991 55	0.987 44	0.981 84	0.974 62	0.965 88	0.955 59	0.944 84	2.8
2.9	0.998 13	0.996 59	0.994 32	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	2.9
3.0	0.998 50	0.996 94	0.994 44	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	3.0
3.1	0.998 84	0.997 16	0.994 54	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	3.1
3.2	0.999 13	0.997 36	0.994 64	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	3.2
3.3	0.999 38	0.997 54	0.994 79	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	3.3
3.4	0.999 59	0.997 74	0.994 94	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	3.4
3.5	0.999 76	0.997 89	0.995 04	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	3.5
3.6	0.999 91	0.998 04	0.995 19	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	3.6
3.7	0.999 99	0.998 16	0.995 31	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	3.7
3.8	0.999 99	0.998 16	0.995 31	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	3.8
3.9	0.999 99	0.998 16	0.995 31	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	3.9
4.0	0.999 99	0.998 16	0.995 31	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	4.0
4.1	0.999 99	0.998 16	0.995 31	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	4.1
4.2	0.999 99	0.998 16	0.995 31	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	4.2
4.3	0.999 99	0.998 16	0.995 31	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	4.3
4.4	0.999 99	0.998 16	0.995 31	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	4.4
4.5	0.999 99	0.998 16	0.995 31	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	4.5
4.6	0.999 99	0.998 16	0.995 31	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	4.6
4.7	0.999 99	0.998 16	0.995 31	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	4.7
4.8	0.999 99	0.998 16	0.995 31	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	4.8
4.9	0.999 99	0.998 16	0.995 31	0.990 82	0.985 82	0.979 26	0.971 06	0.961 32	0.950 06	0.938 26	4.9

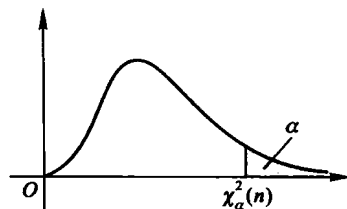
附表 5 t 分布表

$$P(t(n) \geq t_{\alpha}(n)) = \alpha$$



$\alpha \backslash n$	0.25	0.10	0.05	0.025	0.01	0.005
1	1.000 0	3.077 7	6.313 8	12.706 2	31.820 7	63.657 4
2	0.816 5	1.885 6	2.920 0	4.302 7	6.964 6	9.924 8
3	0.764 9	1.637 7	2.353 4	3.182 4	4.540 7	5.840 9
4	0.740 7	1.533 2	2.131 8	2.776 4	3.746 9	4.604 1
5	0.726 7	1.475 9	2.015 0	2.570 6	3.364 9	4.032 2
6	0.717 6	1.439 8	1.943 2	2.446 9	3.142 7	3.707 4
7	0.711 1	1.414 9	1.894 6	2.364 6	2.998 0	3.499 5
8	0.706 4	1.396 8	1.859 5	2.306 0	2.896 5	3.355 4
9	0.702 7	1.383 0	1.833 1	2.262 2	2.821 4	3.249 8
10	0.699 8	1.372 2	1.812 5	2.228 1	2.763 8	3.169 3
11	0.697 4	1.363 4	1.795 9	2.201 0	2.718 1	3.108 5
12	0.695 5	1.356 2	1.782 3	2.178 8	2.681 0	3.054 5
13	0.693 8	1.350 2	1.770 9	2.160 4	2.650 3	3.012 3
14	0.692 4	1.345 0	1.761 3	2.144 8	2.624 5	2.976 8
15	0.691 2	1.340 6	1.753 1	2.131 5	2.602 5	2.946 7
16	0.690 1	1.336 8	1.745 9	2.119 9	2.583 5	2.920 8
17	0.689 2	1.333 4	1.739 6	2.109 8	2.566 9	2.898 2
18	0.688 4	1.330 4	1.734 1	2.100 9	2.552 4	2.878 4
19	0.687 6	1.327 7	1.729 1	2.093 0	2.539 5	2.860 9
20	0.687 0	1.325 3	1.724 7	2.086 0	2.528 0	2.845 3
21	0.686 4	1.323 2	1.720 7	2.079 6	2.517 7	2.831 4
22	0.685 8	1.321 2	1.717 1	2.073 9	2.508 3	2.818 8
23	0.685 3	1.319 5	1.713 9	2.068 7	2.499 9	2.807 3
24	0.684 8	1.317 8	1.710 9	2.063 9	2.492 2	2.796 9
25	0.684 4	1.316 3	1.708 1	2.059 5	2.485 1	2.787 4
26	0.684 0	1.315 0	1.705 6	2.055 5	2.478 6	2.778 7
27	0.683 7	1.313 7	1.703 3	2.051 8	2.472 7	2.770 7
28	0.683 4	1.312 5	1.701 1	2.048 4	2.467 1	2.763 3
29	0.683 0	1.311 4	1.699 1	2.045 2	2.462 0	2.756 4
30	0.682 8	1.310 4	1.697 3	2.042 3	2.457 3	2.750 0
31	0.682 5	1.309 5	1.695 5	2.039 5	2.452 8	2.744 0
32	0.682 2	1.308 6	1.693 9	2.036 9	2.448 7	2.738 5
33	0.682 0	1.307 7	1.692 4	2.034 5	2.444 8	2.733 3
34	0.681 8	1.307 0	1.690 9	2.032 2	2.441 1	2.728 4
35	0.681 6	1.306 2	1.689 6	2.030 1	2.437 7	2.723 8
36	0.681 4	1.305 5	1.688 3	2.028 1	2.434 5	2.719 5
37	0.681 2	1.304 9	1.687 1	2.026 2	2.431 4	2.715 4
38	0.681 0	1.304 2	1.686 0	2.024 4	2.428 6	2.711 6
39	0.680 8	1.303 6	1.684 9	2.022 7	2.425 8	2.707 9
40	0.680 7	1.303 1	1.683 9	2.021 1	2.423 3	2.704 5
41	0.680 5	1.302 5	1.682 9	2.019 5	2.420 8	2.701 2
42	0.680 4	1.302 0	1.682 0	2.018 1	2.418 5	2.698 1
43	0.680 2	1.301 6	1.681 1	2.016 7	2.416 3	2.695 1
44	0.680 1	1.301 1	1.680 2	2.015 4	2.414 1	2.692 3
45	0.680 0	1.300 6	1.679 4	2.014 1	2.412 1	2.689 6

附表 6 χ^2 分布表
 $P(\chi^2(n) \geq \chi_\alpha^2(n)) = \alpha$



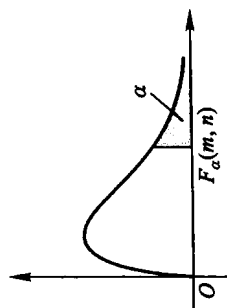
α n	0.995	0.99	0.975	0.95	0.90	0.75
1	—	—	0.001	0.004	0.016	0.102
2	0.010	0.020	0.051	0.103	0.211	0.575
3	0.072	0.115	0.216	0.352	0.584	1.213
4	0.207	0.297	0.484	0.711	1.064	1.923
5	0.412	0.554	0.831	1.145	1.610	2.675
6	0.676	0.872	1.237	1.635	2.204	3.455
7	0.989	1.239	1.690	2.167	2.833	4.255
8	1.344	1.646	2.180	2.733	3.490	5.071
9	1.735	2.088	2.700	3.325	4.168	5.899
10	2.156	2.558	3.247	3.940	4.865	6.737
11	2.603	3.053	3.816	4.575	5.578	7.584
12	3.074	3.571	4.404	5.226	6.304	8.438
13	3.565	4.107	5.009	5.892	7.042	9.299
14	4.075	4.660	5.629	6.571	7.790	10.165
15	4.601	5.229	6.262	7.261	8.547	11.037
16	5.142	5.812	6.908	7.962	9.312	11.912
17	5.697	6.408	7.564	8.672	10.085	12.792
18	6.265	7.015	8.231	9.390	10.865	13.675
19	6.844	7.633	8.907	10.117	11.651	14.562
20	7.434	8.260	9.591	10.851	12.443	15.452
21	8.034	8.897	10.283	11.591	13.240	16.344
22	8.643	9.542	10.982	12.338	14.042	17.240
23	9.260	10.196	11.689	13.091	14.848	18.137
24	9.886	10.856	12.401	13.848	15.659	19.037
25	10.520	11.524	13.120	14.611	16.473	19.939
26	11.160	12.198	13.844	15.379	17.292	20.843
27	11.808	12.879	14.573	16.151	18.114	21.779
28	12.461	13.565	15.308	16.928	18.939	22.657
29	13.121	14.257	16.047	17.708	19.768	23.567
30	13.787	14.954	16.791	18.493	20.599	24.478

续表

$n \backslash \alpha$	0.995	0.99	0.975	0.95	0.90	0.75
31	14.458	15.655	17.539	19.281	21.434	25.390
32	15.134	16.362	18.291	20.072	22.271	26.304
33	15.815	17.074	19.047	20.867	23.110	27.219
34	16.501	17.789	19.806	21.664	23.952	28.136
35	17.192	18.509	20.569	22.465	24.797	29.054
36	17.887	19.233	21.336	23.269	25.643	29.973
37	18.586	19.960	22.106	24.075	26.492	30.893
38	19.289	20.691	22.878	24.884	27.343	31.815
39	19.996	21.426	23.654	25.695	28.196	32.737
40	20.707	22.164	24.433	26.509	29.051	33.660
41	21.421	22.906	25.215	27.326	29.907	34.585
42	22.138	23.650	25.999	28.144	30.765	35.510
43	22.859	24.398	26.785	28.965	31.625	36.436
44	23.584	25.148	27.575	29.787	32.487	37.363
45	24.311	25.901	28.366	30.612	33.350	38.291
$n \backslash \alpha$	0.25	0.10	0.05	0.025	0.01	0.005
1	1.323	2.706	3.841	5.024	6.635	7.879
2	2.773	4.605	5.991	7.378	9.210	10.597
3	4.108	6.251	7.815	9.348	11.345	12.838
4	5.385	7.779	9.488	11.143	13.277	14.860
5	6.626	9.236	11.071	12.833	15.086	16.750
6	7.841	10.645	12.592	14.449	16.812	18.548
7	9.037	12.017	14.067	16.013	18.475	20.278
8	10.219	13.362	15.507	17.535	20.090	21.955
9	11.389	14.684	16.919	19.023	21.666	23.589
10	12.549	15.987	18.307	20.483	23.209	25.188
11	13.701	17.275	19.675	21.920	24.725	26.757
12	14.845	18.549	21.026	23.337	26.217	28.299
13	15.984	19.812	22.362	24.736	27.688	29.819
14	17.117	21.064	23.685	26.119	29.141	31.319
15	18.245	22.307	24.996	27.488	30.578	32.801

续表

$n \backslash \alpha$	0.25	0.10	0.05	0.025	0.01	0.005
16	19.369	23.542	26.296	28.845	32.000	34.267
17	20.489	24.769	27.587	30.191	33.409	35.718
18	21.605	25.989	28.869	31.526	34.805	37.156
19	22.718	27.204	30.144	32.852	36.191	38.582
20	23.828	28.412	31.410	34.170	37.566	39.997
21	24.935	29.615	32.671	35.479	38.932	41.401
22	26.039	30.813	33.924	36.781	40.289	42.796
23	27.141	32.007	35.172	38.076	41.638	44.181
24	28.241	33.196	36.415	39.364	42.980	45.559
25	29.339	34.382	37.652	40.646	44.314	46.928
26	30.435	35.563	38.885	41.923	45.642	48.290
27	31.528	36.741	40.113	43.194	46.963	49.645
28	32.620	37.916	41.337	44.461	48.278	50.993
29	33.711	39.087	42.557	45.722	49.588	52.336
30	34.800	40.256	43.773	46.979	50.892	53.672
31	35.887	41.422	44.985	48.232	52.191	55.003
32	36.973	42.585	46.194	49.480	53.486	56.328
33	38.058	43.745	47.400	50.725	54.776	57.648
34	39.141	44.903	48.602	51.966	56.061	58.964
35	40.223	46.059	49.802	53.203	57.342	60.275
36	41.304	47.212	50.998	54.437	58.619	61.581
37	42.383	48.363	52.192	55.668	59.892	62.883
38	43.462	49.513	53.384	56.896	61.162	64.181
39	44.539	50.660	54.572	58.120	62.428	65.476
40	45.616	51.805	55.758	59.342	63.691	66.766
41	46.692	52.949	56.942	60.561	64.950	68.053
42	47.766	54.090	58.124	61.777	66.206	69.336
43	48.840	55.230	59.304	62.990	67.459	70.616
44	49.913	56.369	60.481	64.201	68.710	71.893
45	50.985	57.505	61.656	65.410	69.957	73.166



附表 7 F 分布表

$$P(F(m, n) \geq F_{\alpha}(m, n)) = \alpha$$

α	n	m																
		1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
0.10	1	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.7	61.2	61.7	62.3	62.8	63.1	63.3
0.05		161	200	216	225	230	234	237	239	241	242	244	246	248	250	252	253	254
0.025		648	800	864	900	922	937	948	957	963	969	977	985	993	1 000	1 010	1 010	1 020
0.01		4 050	5 000	5 400	5 620	5 760	5 860	5 930	5 980	6 020	6 060	6 110	6 160	6 210	6 260	6 310	6 340	6 370
0.005		16 200	20 000	21 600	22 500	23 100	23 400	23 700	23 900	24 100	24 200	24 400	24 600	24 800	25 000	25 200	25 400	25 500
0.10	2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.46	9.47	9.48	9.49
0.05		18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5
0.025		38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5	39.5
0.01		98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5
0.005		199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199
0.10	3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.17	5.15	5.14	5.13
0.05		10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.57	8.55	8.53
0.025		17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.3	14.2	14.1	14.0	13.9	13.9
0.01		34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.5	26.3	26.2	26.1
0.005		55.6	49.8	47.5	46.2	45.4	44.8	44.4	44.1	43.9	43.7	43.4	43.1	42.8	42.5	42.1	42.0	41.8

续表

α	n	m																
		1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
0.10	4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.93	3.92	3.90	3.87	3.84	3.82	3.79	3.78	3.76
0.05		7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.69	5.66	5.63
0.025		12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.46	8.36	8.31	8.26
0.01		21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.8	13.7	13.6	13.5
0.005		31.3	26.3	24.3	23.2	22.5	22.0	21.6	21.4	21.1	21.0	20.7	20.4	20.2	19.9	19.6	19.5	19.3
0.10	5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.17	3.14	3.12	3.11
0.05		6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.43	4.40	4.37
0.025		10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.23	6.12	6.07	6.02
0.01		16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.38	9.20	9.11	9.02
0.005		22.8	18.3	16.5	15.6	14.9	14.5	14.2	14.0	13.8	13.6	13.4	13.1	12.9	12.7	12.4	12.3	12.1
0.10	6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.80	2.76	2.74	2.72
0.05		5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.74	3.70	3.67
0.025		8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.07	4.96	4.90	4.85
0.01		13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23	7.06	6.97	6.88
0.005		18.6	14.5	12.9	12.0	11.5	11.1	10.8	10.6	10.4	10.2	10.0	9.81	9.59	9.36	9.12	9.00	8.88
0.10	7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.56	2.51	2.49	2.47
0.05		5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.30	3.27	3.23
0.025		8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.36	4.25	4.20	4.14
0.01		12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99	5.82	5.74	5.65
0.005		16.2	12.4	10.9	10.1	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.53	7.31	7.19	7.08

续表

α	n	m																
		1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
0.10	8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.38	2.34	2.31	2.29
0.05		5.32	4.46	4.07	3.84	3.60	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.45	3.08	3.01	2.97	2.93
0.025		7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.89	3.78	3.73	3.67
0.01		11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20	5.03	4.95	4.86
0.005		14.7	11.0	9.60	8.81	8.30	7.95	7.60	7.50	7.34	7.21	7.01	6.81	6.61	6.40	6.18	6.06	5.95
0.10	9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.25	2.21	2.18	2.16
0.05		5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.79	2.75	2.71
0.025		7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.56	3.45	3.39	3.33
0.01		10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65	4.48	4.40	4.31
0.005		13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.62	5.41	5.30	5.19
0.10	10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.15	2.11	2.08	2.06
0.05		4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.84	2.77	2.70	2.62	2.58	2.54
0.025		6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.31	3.20	3.14	3.08
0.01		10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25	4.08	4.00	3.91
0.005		12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.07	4.86	4.75	4.64
0.10	12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.01	1.96	1.93	1.90
0.05		4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.60	2.62	2.54	2.47	2.38	2.34	2.30
0.025		6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	2.96	2.85	2.79	2.72
0.01		9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.70	3.54	3.45	3.36
0.005		11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.33	4.12	4.01	3.90
0.10	15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.87	1.82	1.79	1.76
0.05		4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.16	2.11	2.07
0.025		6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.64	2.52	2.46	2.40
0.01		8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.21	3.05	2.96	2.87
0.005		10.8	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.69	3.48	3.37	3.26

续表

α	n	m																	∞
		1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120		
0.10	20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.74	1.68	1.64	1.61	
0.05		4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.95	1.90	1.84	
0.025		5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.35	2.22	2.16	2.09	
0.01		8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78	2.61	2.52	2.42	
0.005		9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	3.12	2.92	2.81	2.69	
0.10	30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.61	1.54	1.50	1.46	
0.05		4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.74	1.68	1.62	
0.025		5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.07	1.94	1.87	1.79	
0.01		7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39	2.21	2.11	2.01	
0.005		9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.63	2.42	2.30	2.18	
0.10	60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.43	1.40	1.35	1.29	
0.05		4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.53	1.47	1.39	
0.025		5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.82	1.67	1.53	1.48	
0.01		7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03	1.84	1.73	1.60	
0.005		8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.19	1.96	1.83	1.69	
0.10	120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.54	1.48	1.41	1.32	1.26	1.19	
0.05		3.92	3.07	2.63	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.55	1.43	1.35	1.25	
0.025		5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.60	1.53	1.43	1.31	
0.01		6.85	4.79	3.95	3.43	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.86	1.66	1.53	1.38	
0.005		8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	1.98	1.75	1.61	1.43	
0.10	∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.34	1.24	1.17	1.00	
0.05		3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.32	1.22	1.00	
0.025		5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.57	1.39	1.27	1.00	
0.01		6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.70	1.47	1.32	1.00	
0.005		7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.79	1.53	1.36	1.00	

习 题 答 案

习题 1

1. (1) $AB\bar{C}$; (2) ABC ; (3) $\bar{A}\bar{B}\bar{C}$; (4) $AB\bar{C} \cup A\bar{B}C \cup \bar{A}BC$;
(5) $\bar{A} \cup \bar{B} \cup \bar{C}$; (6) $A\bar{B}\bar{C} \cup \bar{A}B\bar{C} \cup \bar{A}\bar{B}C \cup A\bar{B}C$.
2. (1) $A \subset B$; (2) $B \subset A$; (3) $A \subset BC$; (4) $B \cup C \subset A$.
3. 11 种 (2~12), 7, $\frac{1}{6}$.
4. (1) $\frac{132}{169}$; (2) $\frac{37}{169}$.
5. (1) 0.3; (2) 0.07; (3) 0.73; (4) 0.14; (5) 0.9; (6) 0.1.
6. (1) $\frac{1}{12}$; (2) $\frac{1}{20}$; (3) $\frac{1}{6}$.
7. (1) $\frac{2}{5}$; (2) $\frac{1}{10}$; (3) $\frac{7}{10}$; (4) $\frac{9}{10}$.
8. $\frac{1}{4}$.
9. (1) $\frac{4}{17}$; (2) $\frac{5}{17}$.
10. 0.5.
11. 0.082 6.
12. 0.6.
13. 0.984.
14. 99.
15. 0.331.
16. 0.92, $\frac{45}{92}$.
17. (1) 0.146; (2) 0.238.
18. (1) 0.72; (2) 0.26; (3) 0.98; (4) 0.02.
19. 0.923 9.
20. (1) 0.512; (2) 0.384; (3) 0.896.
21. 0.492.
22. (1) 0.333; (2) 0.156; (3) 0.859.
- 23*. $\frac{a}{a+b}$.
- 24*. $\frac{3}{10}, \frac{3}{5}$.

$$25^* . \frac{Mn + Nn + N}{(m + n + 1)(M + N)}.$$

26* . 略.

$$27^* . \frac{P_1}{P_1 + P_2 - P_1 P_2}; \quad \frac{P_2 - P_1 P_2}{P_1 + P_2 - P_1 P_2}.$$

28* . 0.321.

习题 2

1. 略.

X	3	4	5
P	0.1	0.3	0.6

$$3. P(X=k) = \frac{C_3^k C_{12}^{4-k}}{C_{15}^4}, \quad k=0, 1, 2, 3.$$

4. (1) $a=1$; (2) $a=1$.

$$5. P(X=k) = \frac{3}{4} \cdot \left(\frac{1}{4}\right)^{k-1}, \quad k=1, 2, \dots; \quad P(X=\text{偶数})=0.2.$$

$$6. \frac{1}{6}, \quad \frac{1}{2}.$$

7. (1) 0.19; (2) 0.323 1; (3) 2.

8. 0.595.

9. 0.120 9.

10. 4.

11. 0.006 738, 0.265 0.

12. (1) 0.029 77; (2) 0.440 5.

13. (1) $c=2$; (2) 0.4; (3) $\frac{1}{\sqrt{2}}$;

$$(4) F(x) = \begin{cases} 0, & x < 0, \\ x^2, & 0 \leq x < 1, \\ 1, & x \geq 1. \end{cases}$$

14. $1 \leq k \leq 3$.

$$15. c = \frac{1}{\pi}, \quad \frac{1}{3}.$$

$$16. f(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & x \leq 0; \end{cases} \quad 1 - e^{-4}, \quad e^{-1}.$$

$$17. \frac{1}{2} - e^{-1}.$$

$$18. \frac{20}{27}.$$

$$19. P(Y=k) = C_5^k (e^{-2})^k (1 - e^{-2})^{5-k}, \quad k=0, 1, \dots, 5; \quad 0.516 7.$$

20. (1) 0.368; (2) 0.233.

21. 0.045 5.

22. 31. 25.

23. 能被录取.

24. (1) 0.532 8, 0.697 7; (2) $c=3$; (3) $d \leq 0.436$.

25. 0.87.

26*. 略.

27*. 略.

28.	$Y=2X+1$	-3	-1	1	3	5
	P	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{5}$
	$Y=X^2$	0	1	4		
	P	$\frac{1}{4}$	$\frac{9}{20}$	$\frac{3}{10}$		

$$29. f(y) = \begin{cases} \frac{1}{2} e^{-\frac{y}{2}}, & y > 0, \\ 0, & y \leq 0. \end{cases}$$

$$30. f(y) = \begin{cases} \frac{1}{3}, & 0 \leq y \leq 3, \\ 0, & \text{其它.} \end{cases}$$

$$31. (1) f(y) = \begin{cases} \frac{1}{2\sqrt{\pi(y-1)}} e^{-\frac{y-1}{4}}, & y > 1, \\ 0, & \text{其它.} \end{cases}$$

$$(2) f(y) = \begin{cases} \sqrt{\frac{2}{\pi}} e^{-\frac{y^2}{2}}, & y > 0, \\ 0, & y \leq 0. \end{cases}$$

32*.	Y	-1	0	1
	P	$\frac{2}{15}$	$\frac{1}{3}$	$\frac{8}{15}$

$$33*. F_Y(y) = \begin{cases} 0, & y < 0, \\ y, & 0 \leq y < 1, \\ 1, & y \geq 1. \end{cases}$$

习题 3

1. $\frac{3}{128}$.

2.	Y	-1	1	2
	X			
	0	0	0	$\frac{1}{2}$
	1.5	0	$\frac{1}{4}$	$\frac{1}{8}$
	2	$\frac{1}{8}$	0	0

3.	$\begin{array}{c cccc} & Y & & & \\ X & 1 & 2 & 3 & 4 \\ \hline 1 & \frac{1}{4} & 0 & 0 & 0 \\ 2 & \frac{1}{8} & \frac{1}{8} & 0 & 0 \\ 3 & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & 0 \\ 4 & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} \\ \hline P_{\cdot j} & \frac{25}{48} & \frac{13}{48} & \frac{7}{48} & \frac{3}{48} \end{array}$	$\begin{array}{c} p_{i \cdot} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ 1 \end{array}$
----	--	---

4.	$\begin{array}{c cc} & Y & \\ X & 1 & 2 \\ \hline 2 & 0 & \frac{3}{5} \\ 3 & \frac{2}{5} & 0 \end{array}$
----	---

$$5. (1) 12; \quad (2) F(x, y) = \begin{cases} (1 - e^{-3x})(1 - e^{-4y}), & x > 0, y > 0, \\ 0, & \text{其它;} \end{cases}$$

$$(3) 0.9499.$$

$$6. \frac{1}{\pi^2}, \frac{\pi}{2}, \frac{\pi}{2}, f(x, y) = \frac{6}{\pi^2(4+x^2)(9+y^2)}.$$

$$7. 1 + e^{-1} - 2e^{-\frac{1}{2}}.$$

$$8. f_x(x) = \begin{cases} 2x^2 + \frac{2}{3}x, & 0 < x < 1, \\ 0, & \text{其它,} \end{cases} \quad f_y(y) = \begin{cases} \frac{y}{6} + \frac{1}{3}, & 0 < y < 2, \\ 0, & \text{其它.} \end{cases}$$

$$9. f_x(x) = \begin{cases} 6(x - x^2), & 0 \leq x \leq 1, \\ 0, & \text{其它,} \end{cases} \quad f_y(y) = \begin{cases} 6(\sqrt{y} - y), & 0 \leq y \leq 1, \\ 0, & \text{其它.} \end{cases}$$

$$10^*. f_{X|Y}(x|y) = \begin{cases} \frac{1}{1-|y|}, & |y| < x < 1, \\ 0, & \text{其它,} \end{cases} \quad f_{Y|X}(y|x) = \begin{cases} \frac{1}{2x}, & |y| < x < 1, \\ 0, & \text{其它.} \end{cases}$$

$$11^*. f_{Y|X}(y|x) = \begin{cases} \frac{2y}{1-x^4}, & -1 < x < 1, x^2 < y < 1, \\ 0, & \text{其它;} \end{cases} \quad \frac{7}{15}.$$

12. 相互独立.

13. 6.

$$14. a = \frac{9}{2}, b = 9, \quad \begin{array}{c|cc} X & 1 & 2 \\ \hline P & \frac{1}{3} & \frac{2}{3} \end{array}$$

$$15. (1) f(x, y) = \begin{cases} e^{-(x+y)}, & x > 0, y > 0 \\ 0, & \text{其它;} \end{cases}$$

$$(2) P(X \leq 1 | Y > 0) = 1 - e^{-1}.$$

$$16. \begin{array}{c|cccc} Z_1 & 2 & 3 & 4 & 5 \\ \hline P & \frac{1}{5} & \frac{1}{5} & \frac{2}{5} & \frac{1}{5} \\ \hline Z_2 & 1 & 2 & 3 \\ \hline P & \frac{1}{5} & \frac{2}{5} & \frac{2}{5} \end{array}$$

$$17. (1) \frac{7}{24}; \quad (2) f_z(z) = \begin{cases} z(2-z), & 0 < z < 1, \\ (2-z)^2, & 1 \leq z < 2, \\ 0, & \text{其它.} \end{cases}$$

$$18. f_z(z) = \begin{cases} e^{-\frac{z}{3}}(1 - e^{-\frac{z}{6}}), & z \geq 0, \\ 0, & z < 0. \end{cases}$$

$$19. f_z(z) = \begin{cases} 0, & z < 0, \\ 1 - e^{-z}, & 0 \leq z < 1, \\ (e-1)e^{-z}, & z \geq 1. \end{cases}$$

$$20. f_z(z) = \begin{cases} \frac{\lambda_1 \lambda_2}{(\lambda_2 + \lambda_1 z)^2}, & z > 0, \\ 0, & z \leq 0. \end{cases}$$

$$21. f_{\min}(z) = \begin{cases} (\alpha + \beta)e^{-(\alpha+\beta)z}, & z \geq 0, \\ 0, & z < 0, \end{cases}$$

$$f_{\max}(z) = \begin{cases} \alpha e^{-\alpha z} + \beta e^{-\beta z} - (\alpha + \beta)e^{-(\alpha+\beta)z}, & z \geq 0, \\ 0, & z < 0. \end{cases}$$

$$22^*. (1) f(x, y) = \begin{cases} \frac{1}{x}, & 0 < y < x < 1, \\ 0, & \text{其它.} \end{cases}$$

$$(2) f_Y(y) = \begin{cases} -\ln y, & 0 < y < 1, \\ 0, & \text{其它.} \end{cases}$$

$$(3) 1 - \ln 2.$$

$$23^*. \begin{array}{c|ccc} X & -1 & 0 & 1 \\ \hline P & 0.1344 & 0.7312 & 0.1344 \end{array}$$

$$24^*. \begin{array}{c|cc} & V & \\ \hline U & & \\ \hline 1 & \frac{4}{9} & 0 \\ \hline 2 & \frac{4}{9} & \frac{1}{9} \end{array}$$

$$25^* . (1) f_x(x) = \begin{cases} 2x, & 0 < x < 1, \\ 0, & \text{其它}; \end{cases} \quad f_y(y) = \begin{cases} 1 - \frac{y}{2}, & 0 < y < 2, \\ 0, & \text{其它}. \end{cases}$$

$$(2) f_z(z) = \begin{cases} 1 - \frac{z}{2}, & 0 < z < 2, \\ 0, & \text{其它}. \end{cases}$$

$$(3) \frac{3}{4}.$$

$$26^* . 0.3f(u-1) + 0.7f(u-2).$$

习题 4

1. 2. 3.

2. 略.

3. 0.301, 0.322, 0.567.

4. 略.

5. 0, $\frac{1}{2}$.

6. 0, 2.

7. $a, a + a^2$.

8. 略.

9. 略.

10. (131.99, 154.21).

11. 略.

12. (1) 2; (2) $\frac{1}{3}$.

13. 3.96.

14. $\frac{7}{12}, \frac{7}{12}, \frac{1}{3}$.

15. 4.

16. $\frac{1}{p}, \frac{1-p}{p^2}$.17. $\sqrt{\frac{\pi}{2}}\sigma, \frac{4-\pi}{2}\sigma^2$.18. $\mu, \frac{\sigma^2}{n}$.

19. 21.

20. 5.21.

21. 0, -0.02.

22. 略

23. $-\frac{1}{36}, -\frac{1}{2}$.

$$24. \frac{7}{6}, -\frac{1}{11}.$$

$$25. \rho_{xy} = \begin{cases} \frac{n!!}{\sqrt{(2n-1)!!}}, & n \text{ 为奇}, \\ 0, & n \text{ 为偶}. \end{cases}$$

$$26. (1) \frac{1}{3}, 3; \quad (2) 0.$$

27. 略.

$$28. \frac{a^2 - b^2}{a^2 + b^2}.$$

$$29. p \geq \frac{8}{9}.$$

$$30. 0.374 \ 5.$$

$$31. 0.060 \ 6.$$

$$32^*. \sqrt{\frac{2}{\pi}}, 1 - \frac{2}{\pi}.$$

$$33^*. 1, 3.$$

$$34^*. \frac{\sigma^2}{n}, \frac{n-1}{n}\sigma^2.$$

35*. 略.

$$36^*. 1.342 \ 7.$$

习题 5

$$1. (1) \mathbf{R}^n, (\sqrt{2\pi}\sigma)^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right);$$

(2) T_1, T_3, T_4 是统计量, 其余不是.

$$2. E(\bar{X}) = \lambda, \quad D(\bar{X}) = \frac{\lambda}{n}, \quad E(S^2) = \lambda.$$

$$3. (1) 0.026; \quad (2) 0.578 \ 5; \quad (3) 0.421 \ 5.$$

$$4. 0.62.$$

$$5. 137.$$

$$6. (1) 0.000 \ 75; \quad (2) 0.935.$$

7. 及 8. 略.

$$9. (1) \frac{1}{8}; \quad (2) \text{略}.$$

$$10. t(n-1).$$

11. 略.

$$12. 2(n-1)\sigma^2.$$

13. 及 14. 略.

$$15. t(m+n-2).$$

$$16^*. f(y, z) = \begin{cases} n(n-1)f(y)f(z)[F(y) - F(z)]^{n-2}, & y \geq z, \\ 0, & y < z. \end{cases}$$

17*. 略.

习题 6

1. $\hat{\mu}_x = 158$, $\hat{\sigma}_x^2 = 10$, $\hat{\mu}_y = 49.78$, $\hat{\sigma}_y^2 = 9.95$.

2. (1) $\hat{a} = \bar{X} - \sqrt{\frac{3}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$, $\hat{b} = \bar{X} + \sqrt{\frac{3}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$.

(2) $\hat{a} = \min\{X_1, X_2, \dots, X_n\}$, $\hat{b} = \max\{X_1, X_2, \dots, X_n\}$.

3. $\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$.

4. $\hat{p} = \frac{1}{X}$, $\hat{E}(X) = \bar{X}$.

5. (1) $\bar{X}$; (2) 是.

6. $\frac{1}{2(n-1)}$.

7. 及 8. 略.

9*. 是.

10*. 当 $\rho \leq \min\left\{\frac{\sigma_2}{\sigma_1}, \frac{\sigma_1}{\sigma_2}\right\}$ 时, $c_1 = \frac{\sigma_2(\sigma_2 - \rho\sigma_1)}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2}$,

$$c_2 = \frac{\sigma_1(\sigma_1 - \rho\sigma_2)}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2}.$$

当 $\rho > \min\left\{\frac{\sigma_2}{\sigma_1}, \frac{\sigma_1}{\sigma_2}\right\}$ 时, $c_1 = 1$, $c_2 = 0$ (若 $\sigma_1 < \sigma_2$)

或 $c_1 = 0$, $c_2 = 1$ (若 $\sigma_1 > \sigma_2$).

11. 矩估计量: $\frac{2\bar{X}-1}{1-\bar{X}}$, 最大似然估计量: $-\frac{n}{\sum_{i=1}^n \ln X_i} - 1$.

12*. 略.

13. $[-0.082, 1.242]$, $[0.766, 3.554]$.

14. (1) $[14.791, 14.987]$; (2) $[14.743, 15.035]$.

15*. $[4u_{\alpha/2}^2 \sigma_0^2 / d^2] + 1$.

16*. $[20.74, 21.33]$.

17. $[0.274, 5.553]$.

18. 无显著性差异.

19. 可以.

20. 发生了变化.

21*. 可以.

22*. 接受 H_0 .

23*. $\beta = \Phi\left(1.64 - \frac{\mu_1 - \mu_0}{\sigma} \sqrt{n}\right)$, β 随 n 的增加而减少.

24*. 疗效无显著性差异.

25*. 方差一样, 均值不一样.

26. 处理后降低了含脂率.

习题 7

1. 略.

2. (1) 略; (2) $Y = \beta_0 + \beta_1 X + \varepsilon$; (3) $\hat{y} = -0.66 + 6.295x$;

(4) 略; (5) $[33.77, 53.04]$.

3. (1) $\hat{y} = 29.766 + 0.56x$; (2) 略; (3) $[64.379, 74.723]$.

4. 及 5. 略.

6. (1) $\hat{y} = 77.363 - 1.818x$; (2) 1.818;

(3) $\hat{y} = 66.455$, $x = 4\ 050$ (件).

7. (1) β 的最小二乘估计为 $\hat{\beta} = \frac{\sum_{\alpha=1}^n x_{\alpha} y_{\alpha}}{\sum_{\alpha=1}^n x_{\alpha}^2}$

(2) $\hat{y} = 264.36x$.

8. 无明显差异.

9. 无明显差异.

10. 药品效果的差异是显著的.

附录 I 概率统计在数学建模中的应用

航空公司的预订票策略^①

在激烈的市场竞争中，航空公司为争取更多的客源而开展的一个优质服务项目是预订票业务。

开展预订票业务时，对于一次航班，若公司限制预订票的数量恰好等于飞机的容量，那么由于总会有一些订了机票的乘客不按时前来登机，致使飞机因不满员飞行而利润降低，甚至亏本。而如果不限制预订票数量，那么当持票按时前来登机的乘客超过飞机容量时，必然会引起那些不能飞走的乘客的抱怨，公司不管以什么方式补救，都会导致声誉受损和一定的经济损失，如客源减少、挤掉以后班机的乘客、公司无偿供应食宿、付给一定的赔偿金等。所以航空公司需要综合考虑经济利益和社会声誉，确定预订票数量的最佳限额。

问题分析

公司的经济利益可以用机票收入扣除飞行费用和赔偿金后的利润来衡量，社会声誉可以用持票按时前来登机、但因满员不能飞走的乘客（以下称被挤掉者）限制在一定数量为标准，注意到这个问题的关键因素——预订票的乘客是否按时前来登机——是随机的，所以经济利益和社会声誉两个指标都应该在平均意义下衡量，这是个两目标的优化问题，决策变量是预订票数量的限额。

模型假设

1. 飞机容量为常数 n ，机票价格为常数 g ，飞行费用为常数 r ， r 与乘客数量无关（实际上关系很小），机票价格按照 $g = r/(\lambda n)$ 来制订，其中 $\lambda (< 1)$ 是利润调节因子，如 $\lambda = 0.6$ 表示飞机 60% 满员率就不亏本。
2. 预订票数量的限额为常数 $m (> n)$ ，每位乘客不按时前来登机的概率为 p ，各位乘客是否按时前来登机是相互独立的，这适合于单独行动的商人、游客等。
3. 每位被挤掉者获得的赔偿金为常数 b 。

模型建立

1. 公司的经济利益可以用平均利润 S 来衡量，每次航班的利润 s 为从机票收入中减去飞行费用和可能发生的赔偿金。当 m 位乘客中有 k 位不按时前来

^① 本文引自姜启源，谢金星，叶俊．数学模型．4 版．北京：高等教育出版社，2011．

登机时

$$s = \begin{cases} (m-k)g-r, & m-k \leq n, \\ ng-r-(m-k-n)b, & m-k > n, \end{cases} \quad (1)$$

由假设 2, 不按时前来登机的乘客数 K 服从二项分布, 于是概率

$$p_k = P(K=k) = C_m^k p^k q^{m-k}, \quad q = 1-p. \quad (2)$$

平均利润 S (即 s 的期望) 为

$$S(m) = \sum_{k=0}^{m-n-1} [ng-r-(m-k-n)b]p_k + \sum_{k=m-n}^m [(m-k)g-r]p_k, \quad (3)$$

化简(3)式, 并注意到 $\sum_{k=0}^m kp_k = mp$, 可得

$$S(m) = qmg - r - (g+b) \sum_{k=0}^{m-n-1} (m-k-n)p_k. \quad (4)$$

当 n, g, r, p 给定后可以求 m 使 $S(m)$ 最大.

2. 公司从社会声誉和经济利益两方面考虑, 应该要求被挤掉的乘客不要太多, 而由于被挤掉者的数量是随机的, 可以用被挤掉的乘客数超过若干人的概率作为度量指标. 记被挤掉的乘客数超过 j 人的概率为 $P_j(m)$, 因为被挤掉的乘客数超过 j 人, 等价于 m 位预订票的乘客中不按时前来登机的不超过 $m-n-j-1$ 人, 所以

$$P_j(m) = \sum_{k=0}^{m-n-j-1} p_k \quad (5)$$

对于给定的 n, j , 显然当 $m=n+j$ 时被挤掉的乘客不会超过 j 人, 即 $P_j(m)=0$. 而当 m 变大时 $P_j(m)$ 单调增加.

综上, $S(m)$ 和 $P_j(m)$ 虽然是这个优化问题的两个目标, 但是可以将 $P_j(m)$ 不超过某给定值作为约束条件, 以 $S(m)$ 为单目标函数来求解.

模型求解

为了减少 $S(m)$ 中的参数, 取 $S(m)$ 除以飞行费用 r 为新的目标函数 $J(m)$, 其含义是单位费用获得的平均利润, 注意到假设 1 中有 $g=r/\lambda n$, 由(4)式可得

$$J(m) = \frac{S(m)}{r} = \frac{1}{\lambda n} \left[qm - \left(1 + \frac{b}{g} \right) \sum_{k=0}^{m-n-1} (m-k-n)p_k \right] - 1, \quad (6)$$

其中 b/g 是赔偿金占机票价格的比例. 问题化为给定 $\lambda, n, p, b/g$, 求 m 使 $J(m)$ 最大, 而约束条件为

$$P_j(m) = \sum_{k=0}^{m-n-j-1} p_k \leq \alpha, \quad (7)$$

其中 α 是小于 1 的正数.

模型(6), (7)无法解析地求解, 我们设定几组数据, 用 MATLAB 软件作

数值计算, 结果如下.

设 $n = 150$, $\lambda = 0.6$, $p = 0.05$ 和 0.1 , $b/g = 0.2$ 和 0.4 , 计算 $J(m)$, $P_5(m)$, $P_{10}(m)$, 得表 1.

表 1 $n = 150$ 时的计算结果

m	$p = 0.05$				$p = 0.1$			
	J		P_5	P_{10}	J		P_5	P_{10}
	$b/g = 0.2$	$b/g = 0.4$			$b/g = 0.2$	$b/g = 0.4$		
150	0.583 3	0.583 3	0	0	0.500 0	0.500 0	0	0
152	0.604 4	0.604 4	0	0	0.520 0	0.520 0	0	0
154	0.624 5	0.624 4	0	0	0.540 0	0.540 0	0	0
156	0.640 8	0.639 9	0.000 5	0	0.560 0	0.560 0	0.000 0	0
158	0.650 0	0.647 0	0.018 2	0	0.579 7	0.579 7	0.000 0	0
160	0.651 9	0.645 7	0.125 6	0	0.598 5	0.598 2	0.000 5	0
162	0.649 0	0.638 9	0.372 9	0.004 1	0.614 8	0.613 9	0.005 6	0.000 0
164	0.644 3	0.629 8	0.663 8	0.054 8	0.626 7	0.624 5	0.030 7	0.000 1
166	0.638 9	0.620 0	0.867 8	0.234 4	0.633 0	0.628 5	0.106 0	0.001 9
168	0.633 3	0.610 0	0.961 5	0.522 8	0.634 0	0.626 3	0.254 5	0.014 0
170	0.627 8	0.600 0	0.991 5	0.780 9	0.631 1	0.619 6	0.460 2	0.060 0
172					0.625 9	0.610 3	0.669 4	0.170 9
174					0.619 8	0.599 8	0.830 8	0.353 0
176					0.613 3	0.588 8	0.927 9	0.568 2

结果分析

1. 对于所取的各个 n , p , b/g , 平均利润 $J(m)$ 随着 m 的变大都是先增加再减少, 但是在最大值附近变化很小, 而被挤掉的乘客数超过 5 人和 10 人的概率 $P_5(m)$ 和 $P_{10}(m)$ 增加得相当快, 所以应该参考 $J(m)$ 的最大值, 给定约束条件(7)式中可以接受的 α , 确定合适的 m .

2. 对于一定的 n , p , 当 b/g 由 0.2 增加到 0.4 时 $J(m)$ 的减少不超过 2%, 所以不妨付给被挤掉的乘客以较多的赔偿金, 赢得社会声誉.

3. 综合考虑经济效益和社会声誉, 可给定 $P_5(m) < 0.2$, $P_{10}(m) < 0.05$, 由表 1 知, 对于 $n = 150$, 若估计 $p = 0.05$, 取 $m = 160$; 若估计 $p = 0.1$, 取 $m = 166$. (表 1 中黑体数字)

附录 II 概率论与数理统计发展简介

概率论发展史

一、概率论的创立

概率思想的萌芽 概率论起源于博弈问题。在概率思想产生和发展的早期，意大利数学家卡尔达诺 (G. Cardano, 1501—1576)、法国数学家帕斯卡 (B. Pascal, 1623—1662) 和费马 (P. de Fermat, 1601—1665) 等做出了重要贡献。

卡尔达诺是文艺复兴时期意大利数学家、物理学家、哲学家、诗人兼医生和占星术士，据称是一位私生子，一生命运坎坷，但崇尚个人奋斗。早年从医，热衷赌博、占卜、算命，且因赌债和给耶稣基督算命的异端罪而两次入狱。其间曾因不光彩的身世和行为，或被拒绝进入米兰医学院，或被迫辞去数学教授职务。为了成全对自己寿命的预言，卡尔达诺最终自杀身亡。1526年，卡尔达诺完成了他的著作《论机会游戏》(或译为《游戏机遇的学说》、《论赌博》、《赌博的游戏》)，但该书直到卡尔达诺死后近一个世纪才得以出版(1663年)。

卡尔达诺本人是一个狂热的赌徒。他在《论机会游戏》中讲述了与赌博有关的各种各样的问题，有关机会问题只是其中的一小部分。正是在这部有关论赌博的著作中，卡尔达诺首次从理性的角度，认真地思考了赌博中某些于己有利事件发生的可能性大小的问题。尤其重要的是，卡尔达诺给出了等可能性事件发生的概率的一个粗略定义：一个特殊结果发生的概率等于得到这一结果的各种可能方式的数目除以“全范围”。

值得一提的是在卡尔达诺之前，意大利数学家帕乔利 (L. Pacioli, 约 1445—1517) 在其 1494 年出版的教科书中，曾讨论过“点数问题”，即在赌博提前中断的情况下，如何在两赌徒之间分配赌金的问题。卡尔达诺重新讨论了这一问题，他已经认识到需要分析的不是已经赌过的次数，而是剩下的次数，但未能提供正确的解答。虽然如此，卡尔达诺的想法已向“点数问题”的解决迈出一大步。这个问题的真正解决是由费马和帕斯卡来完成的。

帕斯卡和费马在概率论方面的主要工作是赌博中赌金的“公平”分配和计算。1654年，帕斯卡和费马通信讨论一些赌博中提出的问题，这些信件被

看作是数学史上最早的概率文献。其中一个最有名的问题就是“点数问题”，这是帕斯卡的朋友梅雷提出来要求帕斯卡帮助解决的一个问题。帕斯卡又将此问题写信给费马，从此两人开始了就这一问题研究的通信。这一问题是古典概率的经典题例并对后来古典概率思想的发展产生了一定影响，此问题是说：两赌徒各出 32 枚金币作为赌金，以先得 3 分为赢。第一人现得 2 分，第二人仅得 1 分，设赌局因故中断，问怎样分配赌金才算公平？

当时有人按 3:2 和 3:1 的比例分配 64 枚赌金。帕斯卡和费马否定了这种算法，他们从不同的角度给出了“点数问题”的正确答案。尤其是费马，他的解法正是现在通行的概率教科书中的解法。费马虽没有用“概率”一词，但他已经认识到第一个人赢的概率是 $\frac{3}{4}$ ，也就是有利情形数与所有情形数的比。在费马的解答中，自然要假定所有的情形都是等可能的，即机会均等。

概率概念的要旨，在于对未发生事件的估计和评价。这种思想在帕斯卡和费马的通信讨论中被体现出来，因而他们被认为是概率论的启蒙者。

概率论的诞生 1657 年，荷兰数学家惠更斯 (C. Huygens, 1629—1695) 发表了《论赌博中的计算》，这是最早的概率论著作，也是概率论诞生的标志性著作。惠更斯出生于海牙的一个富豪之家。其父知识渊博，擅长数学研究，同时又是一杰出的诗人和外交家。惠更斯从小受到了父亲的熏陶，喜欢学习和钻研科学问题。16 岁进入莱顿大学学习，后转到布雷达大学学习法律和数学。26 岁获得法学博士学位。

惠更斯在数学方面的最大贡献，就是以《论赌博中的计算》一书奠基了概率论的基础。像帕斯卡和费马一样，惠更斯本身并不热衷赌博，正如他在该书中所言：“在许多情况下，我认为如果读者仔细研究对象，当可注意到你所处理的不仅是赌博问题而已，实际上其中包含着很有趣、很深刻的理论基础”。该书仅包含 14 个命题，并以 5 个留给读者的练习题结束。他先从关于公平赌博值的一条公理出发，推导出有关数学期望的三个基本定理，利用这些定理和递推公式，解决了点子问题及其它一些博弈问题。惠更斯是第一个提出“数学期望”概念的数学家。对于命题 14，惠更斯通过列代数方程来求解，这是其独创，该方法后被瑞士数学家雅各布称之为“惠更斯分析法”。

《论赌博中的计算》第一次把概率论建立在公理、命题和问题上而构成一个较完整的理论体系，第一次对以前概率论知识系统化、公式化和一般化。书末所提出的问题在随后的几十年里被许多数学家讨论。该书为概率论的进一步发展奠定了坚实的基础。

二、古典概率思想的成熟发展

概率论作为一门独立的数学分支，真正的奠基人是瑞士数学家雅各布·伯

努利(Jacob Bernoulli, 1654—1705). 雅各布出生于瑞士巴塞尔的一个商人世家. 他毕业于巴塞尔大学, 1671 年获艺术硕士学位, 后来遵照父亲的意愿又取得神学硕士学位, 但他却不顾父亲的反对, 自学了数学和天文学. 雅各布在 1678 年和 1681 年的两次学习旅行, 使他接触了许多数学家和科学家, 丰富了他的知识, 拓宽了他的兴趣. 1687 年, 雅各布成为巴塞尔大学的数学教授, 直到 1705 年去世.

雅各布是 17—18 世纪期间, 欧洲大陆在数学方面做过特殊贡献的伯努利家族(共产生过 11 位数学家)的重要成员之一, 他在数学上的贡献涉及微积分、微分方程、解析几何、概率论以及变分法等领域.

雅各布·伯努利一生最有创造力的著作就是 1713 年出版的《猜度术》, 该书共分为四个部分: 第一部分相当于惠更斯著作《论赌博中的计算》的再版, 只是附加了雅各布·伯努利的一个评论; 第二部分集中讨论了排列组合理论; 第三部分包含了各种机会游戏问题的解答; 第四部分建议把概率理论应用于解决伦理道德和经济社会等问题. 这部著作是对概率论建立在稳固的数学基础上的首次认真的尝试, 在这部著作中雅各布提出了概率论中的“伯努利定理”以及在数论中很有影响的“伯努利数”和“伯努利多项式”等基本内容.“伯努利定理”是“大数定律”的最早形式, 雅各布对大数定律的陈述与现代的标准概率论著作十分一致. 由于伯努利兄弟在科学问题上的过于激烈的争论, 致使双方的家庭也被卷入, 以至于雅各布死后, 他的《猜度术》手稿被他的遗孀和儿子在外藏匿多年, 直到 1713 年才得以出版, 几乎使这部经典著作的价值受到损害.

虽然雅各布提出了将概率思想应用于社会伦理、政治、经济学等, 但他并没有给出实例. 不过, 雅各布指明了一条路, 据此向这些问题发起冲击的是一位比他稍晚的同时代的年轻人——法(英)国数学家棣莫弗(Abraham De Moivre 1667—1754). 棣莫弗出生于法国维特里勒弗朗索瓦的一个乡村医生之家, 后因教派斗争遭到监禁并移居英国. 1697 年棣莫弗当选为皇家学会会员, 1735 年、1754 年又分别被接纳为柏林科学院和巴黎科学院院士. 棣莫弗终生未婚, 在英国他靠讲授数学和回答与机会和年度奖金的问题维持生计, 一直过着贫困潦倒的生活, 直到 1754 年在伦敦去世. 棣莫弗在 87 岁时患上了嗜眠症, 每天睡觉长达 20 小时, 关于棣莫弗的死有一个颇具数学色彩的神奇传说: 在临终前若干天, 棣莫弗发现, 他每天需要比前一天多睡 $1/4$ 小时, 那么各天睡眠时间将构成一个算术级数, 当此算术级数达到 24 小时时, 棣莫弗就长眠不醒了.

1711 年, 棣莫弗撰写了题为《抽签的计量》的论文, 并于 1718 年扩充为《机会的学说》一书. 书中首次定义了独立事件的乘法定理, 给出了二项分布公式, 讨论了掷骰子和其它赌博的许多问题. 二项分布作为正态分布或高斯分

布的原始形式,成为其后两个世纪概率论和统计中有效的发现工具;1730年,他又在另一部著作《关于级数和求积的综合分析》中,讨论了排列和组合理论,书中最早使用了概率积分 $\int_0^{\infty} e^{-x^2} dx = \frac{\sqrt{\pi}}{2}$, 得到 n 阶乘的级数表达式,并指出对于很大的 n , $n! \sim \left(\frac{n}{e}\right)^n \sqrt{2n\pi}$. 1733年他又用阶乘的近似公式导出正态曲线,并以此作为二项分布的近似. 后来拉普拉斯将此结果推广到一般情形,得到了现在所谓的“棣莫弗-拉普拉斯极限定理”. 定理揭示了大量服从 0-1 分布的随机变量的和的分布以正态分布为极限,是历史上最早的中心极限定理. 在此后的两个世纪里,中心极限定理的研究始终占据着概率论研究的中心地位.

另外,棣莫弗还是 18 世纪力主将概率论应用于人文、社会科学研究的重要人物之一. 1725 年,棣莫弗出版的《年金论》一书不仅改进了以往的关于人口统计的方法,而且在假定死亡率所遵循的规律以及银行利息不变的情况下,推导出了计算年金的公式,从而为保险业提供了合理处理有关问题的依据.

对古典概率论做出重要贡献的另一位学者就是法国数学家蒲丰 (G. L. L. Buffon, 1707—1788). 蒲丰出生于法国的蒙巴尔,10 岁时就读耶稣会学院,16 岁主修法学,21 岁后转向数学、自然科学特别是植物学,1733 年当选为法国科学院院士,并于 1739 年任巴黎植物园园长,卒于法国大革命前夕. 蒲丰是几何概率的开创者,他在概率论方面的工作以设计“蒲丰投针问题”而著称. 由于投针问题的结果可以用来计算圆周率 π 的近似值,故受到特别重视,成为几何概率的经典范例. 几何概率保留等可能这一特性,突破了古典概型所要求的样本点有限的限制.

三、分析概率论的发展

继棣莫弗之后,数学家拉普拉斯、高斯、泊松等人相继对概率论做出了进一步的奠基性贡献. 尤其是拉普拉斯 1812 年出版的《概率的分析理论》,实现了概率论研究中由组合技巧向分析方法的过渡,开创了概率论发展的新阶段.

拉普拉斯 (P. S. Laplace, 1749—1827) 是法国著名的天文学家和数学家,出生于法国诺曼底的一个农场主家庭,他从青年时期就显示出卓越的数学才能,18 岁时离家赴巴黎,决定从事数学工作. 拉普拉斯生活的时代恰逢政治和社会动荡的法国大革命时期,大革命前夕他主要从事教学工作,并积极从事政治活动. 1795 年任巴黎综合工科学学校教授,后又在高等师范学校任教授. 1816 年被选为法兰西科学院院士,1817 年任该学院院长. 1827 年 3 月 5 日卒于巴黎. 拉普拉斯一生发表的天文学、数学和物理学的论文有 270 多篇,专著合计

有 4 000 多页, 其中最有代表性的专著有《天体力学》、《宇宙体系论》和《概率的解析理论》, 他被誉为法国的牛顿和天体力学之父。

拉普拉斯也是古典概率理论和思想的集大成者, 他对概率论的研究始于 1772 年, 1812 年著成《概率的解析理论》, 1814 年再版时增加的一个长达 150 页的绪论。该书是拉普拉斯对前人及他自己研究成果的全面总结, 运用 17、18 世纪发展起来的强有力的分析工具处理概率论的基本内容, 使以往零散的结果系统化。在这部著作中, 拉普拉斯给出了著名的概率古典定义, 导出了原始形式的中心极限定理(后称棣莫弗-拉普拉斯极限定理), 并将概率论广泛应用于观测误差估计、人口统计、保险等科学和社会问题。诸如随机变量、数字特征、特征函数、拉普拉斯变换和拉普拉斯中心极限定律等都可以说是拉普拉斯引入或者经他改进的。

德国著名数学家、物理学家、天文学家高斯(C. F. Gauss 1777—1855)奠定了最小二乘法和误差理论的基础。18 岁时他发现了素数分布定理和最小二乘法, 通过对足够多的测量数据的处理后, 可以得到一个新的、概率性质的测量结果。在此基础上, 高斯成功得到正态分布曲线, 其函数被命名为标准正态分布或高斯分布, 在概率计算中大量使用。而法国数学家泊松(S. D. Poisson 1781—1840)则推广了伯努利大数定律, 引入了十分重要的“泊松分布”。

19 世纪后期, 极限理论的发展成为概率论研究的中心课题, 俄国数学家切比雪夫(P. L. Chebyshev, 1821—1894)对此做出了重要贡献。切比雪夫出身于贵族家庭, 他的左脚生来有残疾, 因而童年时代的他经常独坐家中, 养成了在孤寂中思索的习惯。1832 年, 切比雪夫全家迁往莫斯科, 1837 年, 切比雪夫进入莫斯科大学物理数学系学习, 大学毕业后, 他留在莫斯科大学当助教并同时攻读硕士学位, 1846 年以题为《试论概率论的基础分析》的论文获硕士学位。1847 年起在圣彼得堡大学任教, 1850 年晋升为副教授, 1860 年晋升为教授, 1859 年当选为圣彼得堡科学院院士。他还先后当选为法兰西科学院、柏林皇家科学院、意大利皇家科学院、瑞典皇家科学院的外籍院士和伦敦皇家学会会员。切比雪夫终身未娶, 把一生献给了科学教育事业。他不仅是土生土长的学者, 而且以他自己的卓越才能和独特的魅力吸引了一批年轻的俄国数学家, 形成了一个具有鲜明风格的数学学派, 从而使俄罗斯数学摆脱了落后境地而开始走向世界前列。他是圣彼得堡数学学派的奠基人和当之无愧的领袖。在概率论、解析数论和函数逼近论等领域, 切比雪夫都做出了开创性工作。

在概率论方面, 切比雪夫 1866 年建立了关于独立随机变量序列的大数定理, 使伯努利大数定律和泊松大数定理成为其特例。特别是他给出的切比雪夫不等式以及建立的证明极限定理的新方法——矩方法, 是证明概率论中大数定理的强有力的工具。他还研究了服从正态规律的独立随机变量和函数的收敛条

件, 将棣莫弗-拉普拉斯极限定理推广至更为一般的中心极限定理. 他引出的一系列概念和研究题材与方法, 后又被他的学生马尔可夫 (A. A. Markov, 1856—1922) 等发扬光大, 影响了 20 世纪概率论的发展进程, 实际上概率论的中心从这时起已由欧洲转移到苏联.

四、概率论的公理化

19 世纪末, 一方面概率论在统计、物理等领域的应用提出了对概率论基本概念与原理进行解释的需要. 另外, 1899 年由法国学者贝特朗 (J. L. F. Bertrand) 提出的所谓“贝特朗悖论”更揭示了古典概率在基本概念方面存在的含混之处. 总之, 无论是从概率论实际应用的需要还是其自身发展的需要, 都要求对概率论的逻辑基础做出更加严格的考察.

20 世纪初由数学家希尔伯特 (D. Hilbert) 倡导的公理化运动在概率论这门学科上结出了丰硕的果实, 形成了公理化概率论, 使这门学科在理论和应用两个方向都迈进了一个崭新的发展阶段.

最早对概率论严格化做出尝试的是苏联数学家伯恩斯坦 (S. N. Bernstein, 1880—1968) 和奥地利数学家冯·米泽斯 (R. von Mises, 1883—1953). 他们都提出了一些公理来作为概率论的前提, 促进了概率论公理化的建立, 但他们的公理理论并不完善;

伯恩斯坦是苏联数学家, 出生于敖德萨, 卒于莫斯科. 1893 年伯恩斯坦毕业于法国巴黎大学, 1901 年又毕业于巴黎综合工科学学校, 1904 年在巴黎获数学博士学位, 1907 年成为教授. 先后在哈尔科夫大学、列宁格勒综合技术学院、列宁格勒大学和苏联科学院数学研究所工作. 1925 年当选为乌克兰科学院院士, 1929 年当选为苏联科学院院士, 他还是巴黎科学院的外国院士. 伯恩斯坦曾获得许多国家的荣誉称号和奖励.

伯恩斯坦对偏微分方程、函数构造论、多项式逼近理论以及概率论都做出了重要贡献. 在概率论方面, 伯恩斯坦与莱维共同开创了相关随机变量之和的收敛问题的研究, 并于 1917 年得到了相当于独立随机变量之和的中心极限定理, 其特点是把独立性换为渐近独立性; 从 1922 年起, 他又着手研究一些应用的实例, 诸如马尔可夫单链成果的推广等; 与莱维在研究一维布朗扩散运动时, 曾尝试用概率论方式研究所谓的随机微分方程, 并可将它推广到多维扩散过程的研究.

事实上, 通过大批数学家的努力和对概率论基本概念的分析, 数学家们认识到: 真正严格的公理化概率论只能建立在测度论与实变函数理论之上. 在这方面的先行者是法国数学家博雷尔.

博雷尔 (E. Borel, 1871—1956) 出生于圣阿弗里克, 1893 年毕业于巴黎高

等师范学校, 1894 年获博士学位. 自 1893 年毕业后, 先后在里尔大学、巴黎高等师范学校、巴黎大学理学院等任教近 50 年之久. 1921 年被选为法国科学院院士. 1928 年组建庞加莱研究所, 并任所长直至去世. 博雷尔不仅是一位杰出的数学家, 而且还是一位著名的社会活动家. 他曾任法国国民会议员、海军部长, 还当过市长. 第二次世界大战期间参加抵抗运动, 并荣获抵抗运动奖章. 他还曾被授予大十字军功章和国家科学研究中心颁发的第一枚金质奖章. 1920 年, 博雷尔曾到中国进行过学术访问, 并逗留了五个月.

博雷尔是 20 世纪一流的、多产的数学家, 他对数学分析、函数论、数论、代数、几何、数学物理、概率论等诸多分支都有杰出的贡献. 博雷尔是测度论的奠基人, 对概率论也有深入的研究, 早在 1905 年他就认识到概率论理论应建立在测度论的基础上, 并首先将测度论方法引入到概率论重要问题的研究中; 他把概率论同测度论相结合, 引进了可数事件集的概率, 填补了古典有限概率和几何概率之间的空白.

博雷尔的工作引起了数学家们沿着这一思路的一系列探索, 尤其是苏联数学家柯尔莫哥洛夫的工作最为突出. 其实, 现在通行的概率论教科书中关于概率的公理化定义就是柯尔莫哥洛夫于 1933 年提出的.

柯尔莫哥洛夫 (A. N. Kolmogorov, 1903—1987) 是苏联杰出的数学家. 他的父亲是一位农学家, 十月革命后曾担任农业部某部门的领导, 1919 年在战争中牺牲. 他的母亲因为难产, 在他出生后 10 天去世. 柯尔莫哥洛夫的童年是在外祖父家度过的, 姨妈把他抚养成人, 也同时培养了他对书本和大自然的兴趣和好奇心. 1920 年, 柯尔莫哥洛夫高中毕业, 进入莫斯科大学, 先学习冶金, 后来转学数学, 并决心以数学为终身职业. 1925 年柯尔莫哥洛夫大学毕业, 1931 年起他担任莫斯科大学教授. 1933 年担任莫斯科大学数学力学研究所所长, 1939 年当选为苏联科学院院士、主席团委员和数学研究所所长, 1954 年担任莫斯科大学数学力学系主任, 1966 年当选为苏联教育科学院院士. 柯尔莫哥洛夫非常重视中学数学教育, 从 20 世纪 30 年代起就指导全苏中学生数学奥林匹克竞赛活动, 编写辅导书籍并亲自给学生讲课, 培养了大批优秀中学生. 柯尔莫哥洛夫一生直接指导的研究生近 70 人, 他们大多成为世界级数学家, 其中 14 人成为苏联科学院院士.

柯尔莫哥洛夫一生发表学术论文 488 篇 (包含合作文章)、科普文章 57 篇. 作为科学巨人, 他所涉及的研究领域极其广泛, 在概率论、实变函数论、数理统计、拓扑空间、动力系统理论、微分方程、泛函分析、射影几何、流体力学、自动控制、数学史等许多数学分支, 他的工作都带有开创性或奠基性. 他还被认为创建了一些新的数学分支, 如信息算法论、概率算法论和语言统计学等. 柯尔莫哥洛夫的成就得到了世界的公认, 他被聘为美、英、法等 12 个

国家的科学院名誉院士，并多次获得国际大奖，但他淡泊名利，1980 年他荣获沃尔夫奖，却没有领取奖金。

在随机数学(概率论,随机过程论和数理统计)方面,柯尔莫哥洛夫在 1924 年念大学四年级时,就和当时的苏联数学家辛钦一起建立了关于独立随机变量的三级数定理;1928 年得到了随机变量序列服从大数定律的充要条件;1929 年得到了独立同分布随机变量序列的重对数律;1930 年得到了强大数定律的非常一般的充分条件;1931 年发表了《概率论的解析方法》一文,奠定了马尔可夫过程论的基础,马尔可夫过程的研究至今仍然是当今世界数学研究的热点和重点之一。

柯尔莫哥洛夫从 20 世纪 20 年代中期开始沿测度论途径研究整个概率论理论的严格表述,其结果发表在他在 1933 年以德文出版的经典著作《概率论基础》。柯尔莫哥洛夫是莫斯科函数论学派领袖卢津(N. N. Luzin, 1883—1950)的学生,对实变函数理论掌握得可谓炉火纯青。他在这部著作中建立起集合测度与事件概率的类比、积分与数学期望的类比、函数正交性与随机变量独立性的类比,等等。这种广泛的类比终于使概率论具有了演绎数学的特征。柯尔莫哥洛夫提出了 6 条公理,整个概率论大厦可以从这 6 条公理出发建立起来。柯尔莫哥洛夫公理化概率定义的实质是把概率定义为随机试验样本空间上的集合函数。柯尔莫哥洛夫的公理化体系逐渐得到了数学家们的普遍承认。由于公理化,概率论成为一门严格的演绎科学,具有了与其它数学分支同等的地位,并通过集合论与其它数学分支密切的联系着。

在公理化基础上,现代概率论取得了一系列理论上的重大突破。一是随机过程的研究由马尔可夫过程推进到对一类特殊的马尔可夫过程——布朗运动的研究;二是对具有重要意义的随机过程“鞅”的研究;三是由日本数学家伊藤清(Itô Kiyosi)于 1942 年引入的随机积分与随机微分方程的研究。

尽管概率论的“出身不好”,但它像任何一门公理化的数学分支一样,一旦完成公理化,就允许各种具体的解释。概率论的公理化是将概率概念从频率解释中抽象出来的,同时又可以从公理系统回到现实世界,因而概率论的应用范围被大大拓宽了。

数理统计发展简介

在 18、19 世纪出现了统计推断思想的萌芽并有了一定发展,但以概率论为基础,以统计推断为主要内容的现代意义上的数理统计学,则到 20 世纪才渐近成熟。

1763 年,自学成材的英国数学家贝叶斯(T. Bayes, 1702—1761)给出的

“贝叶斯定理”(贝叶斯公式)可以看作是一种最早的统计推断程序,它在现代概率论和数理统计中仍有重要作用.拉普拉斯和高斯等人利用贝叶斯公式进行参数估计,高斯由于计算行星轨道的需要而建立了以“最小二乘法”为基础的误差分析,这些都促使统计学摆脱对观测数据的单纯描述而向强调推断的阶段过渡.

英国统计学家 K·皮尔逊(K. Pearson, 1857—1936)对现代数理统计的建立起了重要作用.他在 19 世纪末、20 世纪初发展了他老师高尔顿(F. Galton)首先提出的“相关”与“回归”的理论,成功地建立了生物统计学.皮尔逊明确指出统计学不是研究样本本身,而是要根据样本对总体进行推断,并据此提出了“拟合优度检验”.皮尔逊的工作是所谓“大样本统计”的前驱,他的学生戈塞特(W. S. Gosset)1908 年以笔名“学生”发表的“学生分布”则开创了小样本统计理论,从而使统计学研究对象从群体现象转变为随机现象.

作为一门独立学科的现代数理统计学的奠基人是英国统计学家费希尔(R. A. Fisher, 1890—1962).20 世纪 20—30 年代,费希尔提出了许多重要的统计方法,开辟了一系列统计学的分支领域.他发展了正态总体下的各种统计量的抽样分布,将已有的相关、回归理论建造为系统的相关分析和回归分析.1923 年,费希尔提出了方差分析这一重要的数据分析方法.1925 年,他与耶茨(F. Yates)合作创立了试验设计这一重要的统计分支,他还是假设检验和多元统计分析等重要统计分支的先驱.费希尔做过中学教员,曾长期在农业试验站工作,并致力于数理统计在农业科学和遗传学中的应用.在 20 世纪 20—50 年代,费希尔是数理统计学研究的中心人物.

1928 年,威沙特(J. Wishart)将费希尔的狭义的多元分析发展为统计学中的一个独立分支.中国数学家许宝騄和美国数学家霍特林(H. Hotelling)也是多元统计分析的奠基人.

1946 年,瑞典数学家克拉默(H. Cramér)的著作《统计学的数学方法》,用测度论系统总结了数理统计的发展,标志着现代数理统计学的成熟.

第二次世界大战期间,数理统计学的研究出现了一些重要的动向,这些新的动向在很大程度上决定了战后数理统计学的发展方向.其中最有影响的是瓦尔德(A. Wald, 1902—1950)提出的序贯分析和统计决策理论.

序贯分析的主旨是以“序贯抽样方案”代替统计推断中的传统的固定抽样方案.为了解决二战中军方提出的实际问题,瓦尔德提出序贯分析这一崭新的统计方法.1947 年,瓦尔德发表了《序贯分析》专著,使序贯分析在战后发展为数理统计中的一个重要分支.

1950 年,瓦尔德出版了著作《统计决策函数》.他的统计决策理论用博弈的观点看待数理统计问题,对于推断所获得的论断会产生什么后果,应采取何

种对策或行动等这些不属于经典统计的内容，统计决策理论也将其纳入统计的范畴。瓦尔德的思想方法对 20 世纪下半叶整个数理统计学的发展有着重要影响。

数理统计在近些年来有所发展，但理论上突破不大，最引人注目的是它的普及和广泛的应用。它几乎渗透到一切学科之中，哪里有试验，哪里有数据，哪里就少不了数理统计，它已成为现代科学研究最基本的工具之一，没有数理统计就无法应付大量的数据和信息。数理统计还将为社会的进步做出更大贡献。