



普通高等教育“十一五”国家级规划教材

概率论与数理统计

第四版

□ 浙江大学 盛 骤 谢式千 潘承毅 编



高等教育出版社
HIGHER EDUCATION PRESS

内容提要

本书是普通高等教育“十一五”国家级规划教材,在2001年出版的《概率论与数理统计》(第三版)的基础上增订而成。本次修订新增的内容有:在数理统计中应用 Excel, bootstrap 方法, p 值检验法, 箱线图等;同时吸收了国内外优秀教材的优点对习题的类型和数量进行了调整和充实。

本书主要内容包括概率论、数理统计、随机过程三部分,每章附有习题;同时涵盖了《全国硕士研究生入学统一考试数学考试大纲》的所有知识点。本书可作为高等学校工科、理科(非数学专业)各专业的教材和研究生入学考试的参考书,也可供工程技术人员、科技工作者参考。

图书在版编目(CIP)数据

概率论与数理统计/盛骤,谢式千,潘承毅编. - 4 版.
北京:高等教育出版社,2008.6
ISBN 978-7-04-023896-9

I. 概… II. ①盛…②谢…③潘… III. ①概率论-高等学校-教材②数理统计-高等学校-教材 IV. O21

中国版本图书馆 CIP 数据核字(2008)第 057920 号

策划编辑 李 蕊 责任编辑 蒋 青 封面设计 王凌波 责任绘图 杜晓丹
版式设计 余 杨 责任校对 朱惠芳 责任印制 韩 刚

出版发行	高等教育出版社	购书热线	010-58581118
社 址	北京市西城区德外大街 4 号	免费咨询	800-810-0598
邮政编码	100120	网 址	http://www.hep.edu.cn
总 机	010-58581000		http://www.hep.com.cn
经 销	蓝色畅想图书发行有限公司	网上订购	http://www.landaco.com
印 刷	北京中科印刷有限公司	畅想教育	http://www.widedu.com
开 本	787×960 1/16	版 次	1979 年 3 月第 1 版
印 张	26.75		2008 年 6 月第 4 版
字 数	490 000	印 次	2008 年 6 月第 1 次印刷
		定 价	25.80 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 23896-00

第四版前言

本书自 1979 年 3 月初版至今,已发行近三十年。历经多年教学实践的检验,得到了国内广大院校和任课教师的认可,发行量为国内同类教材中最多的。

第四版是普通高等教育“十一五”国家级规划教材,在第三版的基础上修订编写而成。在编写之前,高等教育出版社在全国有关高校作过相当广泛的调查,本版的编写吸取了相关的意见。

教材应该力求与时俱进。本版新增加了以下内容:

(1) 介绍了 bootstrap 方法的基本思想和方法,介绍了用 bootstrap 方法求参数点估计和区间估计的具体做法。bootstrap 方法是近代统计中的一种用于数据处理的重要的实用方法。

(2) 新增了在数理统计中应用 Excel 软件一章。介绍了 Excel 软件及其在数理统计中的应用,举例介绍了应用 VBA 语言编写“宏”求解具体的数理统计问题。

(3) 新增了假设检验问题的 p 值检验法。新增了箱线图,箱线图能大致描述随机变量分布的一些重要性质,还能检测疑似异常点。

(4) 对第三版原有的例题和习题作了一些调整,增加了有关加强基本概念、基本运算的习题,在例题和习题的选择上扩大了涉及的范围,例如,农业、保险业、医学、商业、管理学、体育、等等。

选用本教材的院校类别较为广泛,专业不一,学生程度不一。我们认为,教材内容要比教学大纲多一些,要比教师在课堂讲授的多一些,这样能照顾到各类学校各个专业的需要,能满足不同程度的学生的学习需要。

我们在目录中打上了一些 * 号,在学时限制下,有 * 号的内容可以不学。这些内容是相对独立的,删去不学不影响全书的讲授。在概率论与数理统计部分中打 * 号的内容有:基于截尾样本的最大似然估计;置信区间与假设检验之间的关系;样本容量的选取;秩和检验。此外还有偏度、峰度检验,以及这一版新增的部分或全部内容。随机过程部分视教学计划中是否有这一门课决定取舍。

本次修订也包括配套辅导书,它们将与教材同时出版。

本书中新增的有关在数理统计中应用 Excel 软件的内容由浙江大学于渤教授编写。

本书由浙江大学范大茵教授审阅,对此我们表示衷心的感谢。

高等教育出版社蒋青、李蕊、兰莹莹同志为本版教材做了很多认真、细致的工作,对此,我们表示诚挚的感谢。

诚恳地希望读者批评、指正。

盛 骤 谢式千 潘承毅

2008年4月



第三版前言

这一版我们对于本书第二版中的一些疏漏和不妥之处作了修改,增加了“基于截尾样本的最大似然估计”和“置信区间与假设检验之间的关系”两小节,对各章的例题和习题作了少量的增减。

为了帮助读者抓住要点,提高学习质量与效率,在各章末增写了“小结”。小结中所包含的内容,有的是用来说明概念的现实背景和含义,对某些概念与方法所基于的概率和统计思想作了进一步的阐述;有的则阐明一章内容的重点和基本要求;有的则指出学习时应注意之点。小结也能起到提纲挈领的作用。

书末还增加了两个参读材料:(一)随机变量样本值的产生,(二)标准正态变量分布函数 $\Phi(x)$ 的数值计算。这些内容在解决实际问题时是常会用到的。

本书这一版承柴根象教授、王静龙教授、谢国瑞教授、范大茵教授审阅,他们提出了许多宝贵意见,对此我们表示衷心的感谢。

盛 骤 谢式千 潘承毅

2000年8月



第二版前言

本书是在1979年出版的第一版的基础上修订的,可作为高等学校工科、理科(非数学专业)概率论与数理统计课程的教材,也可供工程技术人员参考。

本书分三部分。概率论部分(第一章至第五章)作为基础知识,为读者提供了必要的理论基础。数理统计部分(第六章至第九章)主要讲述了参数估计和假设检验,并介绍了方差分析和回归分析。随机过程部分(第十章至第十二章)在讲清基本知识的基础上主要讨论了平稳随机过程,还介绍了马尔可夫过程。数理统计和随机过程这两部分内容是相互独立的,可根据专业的需要选用。

在本书第一版出版后,我们经过进一步的教学实践,积累了不少的经验,并吸收了广大读者的意见,修订稿是在这一基础上写出的。我们修改了第一版中存在的不足之处,并致力于教材质量的提高。我们在选材和叙述上尽量做到联系工科专业的实际,注重应用,力图将概念写得清晰易懂,做到便于教学。我们在例题和习题的选择上作了努力,这些题目既具有启发性,又有广泛的应用性,从题目的广泛性也可看到本门课程涉及生活和技术应用领域的广泛性。读者将会发现,这些例题和习题是饶有趣味的。为适应经济建设的需要,我们加强了数理统计的内容,例如编写了“矩估计法”、“样本容量的选取”和“正态分布的偏度、峰度检验”等,并有意识地加强学习统计计算能力的培养。

书中的一部分内容能直接应用于解决实际课题,另一部分内容为读者今后进一步学习有关课程或在实际应用方面提供一定的基础。

黄纪青同志曾参加过本书第一版编写大纲的讨论,撰写过第一版第一章的初稿。

本书的全部插图是由张礼明同志描绘的。

本书第二版承魏宗舒教授、林少官教授、沈恒范教授、范大茵副教授、樊孝述副教授和汪振鹏副教授审阅,他们提出了很多宝贵意见,对此我们表示衷心的感谢。

书中不足之处,诚恳地希望读者批评指正。

盛 驷 谢式千 潘承毅

1988年1月

目 录

第四版前言	I
第三版前言	I
第二版前言	I
第一章 概率论的基本概念	1
§ 1 随机试验	1
§ 2 样本空间、随机事件	2
§ 3 频率与概率	5
§ 4 等可能概型(古典概型)	9
§ 5 条件概率	14
§ 6 独立性	20
小结	23
习题	24
第二章 随机变量及其分布	30
§ 1 随机变量	30
§ 2 离散型随机变量及其分布律	32
§ 3 随机变量的分布函数	38
§ 4 连续型随机变量及其概率密度	42
§ 5 随机变量的函数的分布	50
小结	54
习题	55
第三章 多维随机变量及其分布	60
§ 1 二维随机变量	60
§ 2 边缘分布	64
§ 3 条件分布	67
§ 4 相互独立的随机变量	72
§ 5 两个随机变量的函数的分布	76
小结	83
习题	84
第四章 随机变量的数字特征	90
§ 1 数学期望	90

§ 2 方差	100
§ 3 协方差及相关系数	106
§ 4 矩、协方差矩阵	110
小结	112
习题	113
第五章 大数定律及中心极限定理	119
§ 1 大数定律	119
§ 2 中心极限定理	121
小结	126
习题	126
第六章 样本及抽样分布	128
§ 1 随机样本	128
§ 2 直方图和箱线图	130
§ 3 抽样分布	135
小结	144
附录	145
习题	147
第七章 参数估计	149
§ 1 点估计	149
§ 2 基于截尾样本的最大似然估计	156
§ 3 估计量的评选标准	158
§ 4 区间估计	161
§ 5 正态总体均值与方差的区间估计	163
§ 6 (0-1)分布参数的区间估计	168
§ 7 单侧置信区间	169
小结	170
习题	173
第八章 假设检验	178
§ 1 假设检验	178
§ 2 正态总体均值的假设检验	183
§ 3 正态总体方差的假设检验	187
§ 4 置信区间与假设检验之间的关系	192
§ 5 样本容量的选取	193
§ 6 分布拟合检验	198
§ 7 秩和检验	208

§ 8 假设检验问题的 p 值检验法	213
小结	217
习题	218
第九章 方差分析及回归分析	224
§ 1 单因素试验的方差分析	224
§ 2 双因素试验的方差分析	233
§ 3 一元线性回归	244
§ 4 多元线性回归	257
小结	261
附录	263
习题	265
第十章 bootstrap 方法	270
§ 1 非参数 bootstrap 方法	270
§ 2 参数 bootstrap 方法	278
小结	281
第十一章 在数理统计中应用 Excel 软件	282
§ 1 概述	282
§ 2 箱线图	284
§ 3 假设检验	285
§ 4 方差分析	287
§ 5 一元线性回归	291
§ 6 bootstrap 方法、宏、VBA	293
本章参考文献	299
第十二章 随机过程及其统计描述	300
§ 1 随机过程的概念	300
§ 2 随机过程的统计描述	303
§ 3 泊松过程及维纳过程	309
小结	316
习题	317
第十三章 马尔可夫链	319
§ 1 马尔可夫过程及其概率分布	319
§ 2 多步转移概率的确定	325
§ 3 遍历性	328
小结	331
习题	333

第十四章 平稳随机过程	335
§ 1 平稳随机过程的概念	335
§ 2 各态历经性	338
§ 3 相关函数的性质	346
§ 4 平稳随机过程的功率谱密度	348
小结	358
习题	360
选做习题	363
参读材料 随机变量样本值的产生	376
附表	379
附表 1 几种常用的概率分布表	379
附表 2 标准正态分布表	382
附表 3 泊松分布表	383
附表 4 t 分布表	385
附表 5 χ^2 分布表	386
附表 6 F 分布表	387
附表 7 均值的 t 检验的样本容量	392
附表 8 均值差的 t 检验的样本容量	394
附表 9 秩和临界值表	396
习题答案	397



第一章 概率论的基本概念

自然界和社会上发生的现象是多种多样的. 有一类现象, 在一定条件下必然发生, 例如, 向上抛一石子必然下落, 同性电荷必相互排斥, 等等. 这类现象称为**确定性现象**. 在自然界和社会上存在着另一类现象, 例如, 在相同条件下抛同一枚硬币, 其结果可能是正面朝上, 也可能是反面朝上, 并且在每次抛掷之前无法肯定抛掷的结果是什么; 用同一门炮向同一目标射击, 各次弹着点不尽相同, 在一次射击之前无法预测弹着点的确切位置. 这类现象, 在一定的条件下, 可能出现这样的结果, 也可能出现那样的结果, 而在试验或观察之前不能预知确切的结果. 但人们经过长期实践并深入研究之后, 发现这类现象在大量重复试验或观察下, 它的结果却呈现出某种规律性. 例如, 多次重复抛一枚硬币得到正面朝上大致有一半, 同一门炮射击同一目标的弹着点按照一定规律分布, 等等. 这种在大量重复试验或观察中所呈现出的固有规律性, 就是我们以后所说的**统计规律性**.

这种在个别试验中其结果呈现出不确定性, 在大量重复试验中其结果又具有统计规律性的现象, 我们称之为**随机现象**. 概率论与数理统计是研究和揭示随机现象统计规律性的一门数学学科.

§ 1 随 机 试 验

我们遇到过各种试验. 在这里, 我们把试验作为一个含义广泛的术语. 它包括各种各样的科学实验, 甚至对某一事物的某一特征的观察也认为是一种试验. 下面举一些试验的例子.

E_1 : 抛一枚硬币, 观察正面 H 、反面 T 出现的情况.

E_2 : 将一枚硬币抛掷三次, 观察正面 H 、反面 T 出现的情况.

E_3 : 将一枚硬币抛掷三次, 观察出现正面的次数.

E_4 : 抛一颗骰子, 观察出现的点数.

E_5 : 记录某城市 120 急救电话台一昼夜接到的呼唤次数.

E_6 : 在一批灯泡中任意抽取一只, 测试它的寿命.

E_7 : 记录某地一昼夜的最高温度和最低温度.

上面举出了七个试验的例子, 它们有着共同的特点. 例如, 试验 E_1 有两种可能结果, 出现 H 或者出现 T , 但在抛掷之前不能确定出现 H 还是出现 T , 这

个试验可以在相同的条件下重复地进行. 又如试验 E_6 , 我们知道灯泡的寿命(以小时计) $t \geq 0$, 但在测试之前不能确定它的寿命有多长. 这一试验也可以在相同的条件下重复地进行. 概括起来, 这些试验具有以下的特点:

1° 可以在相同的条件下重复地进行;

2° 每次试验的可能结果不止一个, 并且能事先明确试验的所有可能结果;

3° 进行一次试验之前不能确定哪一个结果会出现.

在概率论中, 我们将具有上述三个特点的试验称为随机试验.

本书中以后提到的试验都是指随机试验.

我们是通过研究随机试验来研究随机现象的.

§ 2 样本空间、随机事件

(一) 样本空间

对于随机试验, 尽管在每次试验之前不能预知试验的结果, 但试验的所有可能结果组成的集合是已知的. 我们将随机试验 E 的所有可能结果组成的集合称为 E 的样本空间, 记为 S . 样本空间的元素, 即 E 的每个结果, 称为样本点.

下面写出 § 1 中试验 $E_k (k=1, 2, \dots, 7)$ 的样本空间 S_k :

$$S_1: \{H, T\};$$

$$S_2: \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\};$$

$$S_3: \{0, 1, 2, 3\};$$

$$S_4: \{1, 2, 3, 4, 5, 6\};$$

$$S_5: \{0, 1, 2, 3, \dots\};$$

$$S_6: \{t | t \geq 0\};$$

$$S_7: \{(x, y) | T_0 \leq x \leq y \leq T_1\}, \text{ 这里 } x \text{ 表示最低温度 } (^{\circ}\text{C}), y \text{ 表示最高温度 } (^{\circ}\text{C}). \text{ 并设这一地区的温度不会小于 } T_0, \text{ 也不会大于 } T_1.$$

(二) 随机事件

在实际中, 当进行随机试验时, 人们常常关心满足某种条件的那些样本点所组成的集合. 例如, 若规定某种灯泡的寿命(小时)小于 500 为次品, 则在 E_6 中我们关心灯泡的寿命是否有 $t \geq 500$. 满足这一条件的样本点组成 S_6 的一个子集: $A = \{t | t \geq 500\}$. 我们称 A 为试验 E_6 的一个随机事件. 显然, 当且仅当子集 A 中的一个样本点出现时, 有 $t \geq 500$.

一般,我们称试验 E 的样本空间 S 的子集为 E 的随机事件,简称事件^①. 在每次试验中,当且仅当这一子集中的一个样本点出现时,称这一事件发生.

特别,由一个样本点组成的单点集,称为基本事件. 例如,试验 E_1 有两个基本事件 $\{H\}$ 和 $\{T\}$; 试验 E_4 有 6 个基本事件 $\{1\}, \{2\}, \dots, \{6\}$.

样本空间 S 包含所有的样本点,它是 S 自身的子集,在每次试验中它总是发生的, S 称为必然事件. 空集 $\emptyset$ 不包含任何样本点,它也作为样本空间的子集,它在每次试验中都不发生, $\emptyset$ 称为不可能事件.

下面举几个事件的例子.

例 1 在 E_2 中事件 A_1 : “第一次出现的是 H ”, 即

$$A_1 = \{HHH, HHT, HTH, HTT\}.$$

事件 A_2 : “三次出现同一面”, 即

$$A_2 = \{HHH, TTT\}.$$

在 E_6 中,事件 A_3 : “寿命小于 1000 小时”, 即

$$A_3 = \{t | 0 \leq t < 1000\}.$$

在 E_7 中,事件 A_4 : “最高温度与最低温度相差 10 摄氏度”, 即

$$A_4 = \{(x, y) | y - x = 10, T_0 \leq x \leq y \leq T_1\}.$$

□

(三) 事件间的关系与事件的运算

事件是一个集合,因而事件间的关系与事件的运算自然按照集合论中集合之间的关系和集合运算来处理. 下面给出这些关系和运算在概率论中的提法. 并根据“事件发生”的含义,给出它们在概率论中的含义.

设试验 E 的样本空间为 S , 而 $A, B, A_k (k=1, 2, \dots)$ 是 S 的子集.

1° 若 $A \subset B$, 则称事件 B 包含事件 A , 这指的是事件 A 发生必导致事件 B 发生.

若 $A \subset B$ 且 $B \subset A$, 即 $A = B$, 则称事件 A 与事件 B 相等.

2° 事件 $A \cup B = \{x | x \in A \text{ 或 } x \in B\}$ 称为事件 A 与事件 B 的和事件. 当且仅当 A, B 中至少有一个发生时,事件 $A \cup B$ 发生.

类似地,称 $\bigcup_{k=1}^n A_k$ 为 n 个事件 $A_1, A_2, \dots, A_n$ 的和事件; 称 $\bigcup_{k=1}^{\infty} A_k$ 为可列个事件 $A_1, A_2, \dots$ 的和事件.

3° 事件 $A \cap B = \{x | x \in A \text{ 且 } x \in B\}$ 称为事件 A 与事件 B 的积事件. 当且仅

^① 严格地说,事件是指 S 中的满足某些条件的子集. 当 S 是由有限个元素或由可列无限个元素组成时,每个子集都可作为一个事件. 若 S 是由不可列无限个元素组成时,某些子集必须排除在外. 幸而这种不可容许的子集在实际应用中几乎不会遇到. 今后,每当我们讲到一个事件时都是假定它是容许考虑的那种子集. 读者如有兴趣可参考较详细的教材.

当 A, B 同时发生时, 事件 $A \cap B$ 发生. $A \cap B$ 也记作 AB .

类似地, 称 $\bigcap_{k=1}^n A_k$ 为 n 个事件 $A_1, A_2, \dots, A_n$ 的积事件; 称 $\bigcap_{k=1}^{\infty} A_k$ 为可列个事件 $A_1, A_2, \dots$ 的积事件.

4° 事件 $A - B = \{x | x \in A \text{ 且 } x \notin B\}$ 称为事件 A 与事件 B 的差事件. 当且仅当 A 发生、 B 不发生时事件 $A - B$ 发生.

5° 若 $A \cap B = \emptyset$, 则称事件 A 与 B 是互不相容的, 或互斥的. 这指的是事件 A 与事件 B 不能同时发生. 基本事件是两两互不相容的.

6° 若 $A \cup B = S$ 且 $A \cap B = \emptyset$, 则称事件 A 与事件 B 互为逆事件. 又称事件 A 与事件 B 互为对立事件. 这指的是对每次试验而言, 事件 A, B 中必有一个发生, 且仅有一个发生. A 的对立事件记为 $\bar{A}$. $\bar{A} = S - A$.

用图 1-1—图 1-6 可直观地表示以上事件之间的关系与运算. 例如, 在图 1-1 中长方形表示样本空间 S , 圆 A 与圆 B 分别表示事件 A 与事件 B , 事件 B 包含事件 A . 又如在图 1-2 中长方形表示样本空间 S , 圆 A 与圆 B 分别表示事件 A 与事件 B , 而阴影部分表示和事件 $A \cup B$.

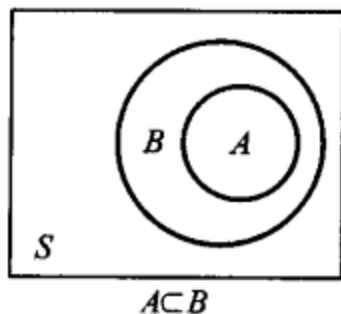


图 1-1

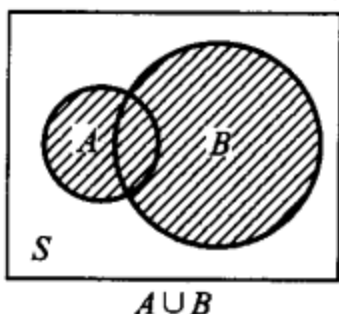


图 1-2

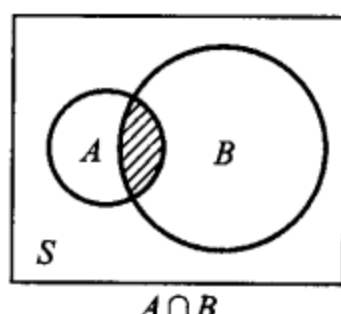


图 1-3

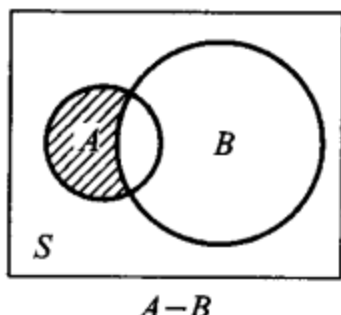


图 1-4

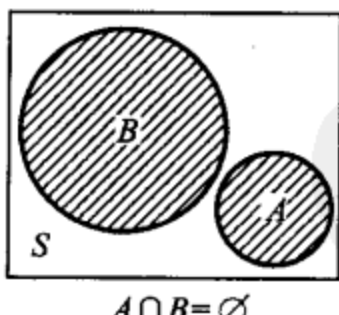


图 1-5

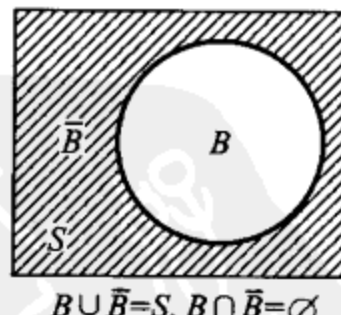


图 1-6

在进行事件运算时, 经常要用到下述定律. 设 A, B, C 为事件, 则有
交换律: $A \cup B = B \cup A$; $A \cap B = B \cap A$.

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

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

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

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

德摩根律: $\overline{A \cup B} = \overline{A} \cap \overline{B}$; $\overline{A \cap B} = \overline{A} \cup \overline{B}$.

例 2 在例 1 中有

$$A_1 \cup A_2 = \{HHH, HHT, HTH, HTT, TTT\},$$

$$A_1 \cap A_2 = \{HHH\},$$

$$A_2 - A_1 = \{TTT\},$$

$$\overline{A_1 \cup A_2} = \{THT, TTH, THH\}.$$

□

例 3 如图 1-7 所示的电路中,以 A 表示“信号灯亮”这一事件,以 B, C, D 分别表示事件:继电器接点 I, II, III 闭合,那么容易知道 $BC \subset A, BD \subset A, BC \cup BD = A$, 而 $\overline{BA} = \emptyset$, 即事件 $\overline{B}$ 与事件 A 互不相容. 又, $\overline{B \cup C} = \overline{B} \cap \overline{C}$. (左边表示事件“I, II 至少有一个闭合”的逆事件,也就是 I, II 都不闭合,即 $\overline{B}, \overline{C}$ 同时发生.)

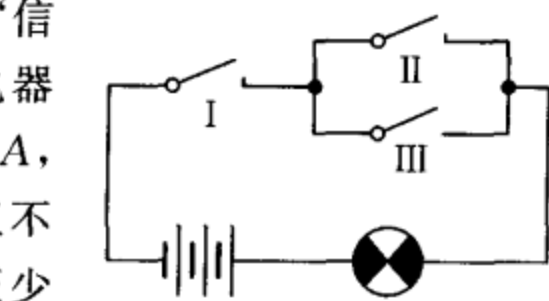


图 1-7

□

§ 3 频率与概率

对于一个事件(除必然事件和不可能事件外)来说,它在一次试验中可能发生,也可能不发生. 我们常常希望知道某些事件在一次试验中发生的可能性究竟有多大. 例如,为了确定水坝的高度,就要知道河流在造水坝地段每年最大洪水达到某一高度这一事件发生的可能性大小. 我们希望找到一个合适的数来表征事件在一次试验中发生的可能性大小. 为此,首先引入频率,它描述了事件发生的频繁程度,进而引出表征事件在一次试验中发生的可能性大小的数——概率.

(一) 频率

定义 在相同的条件下,进行了 n 次试验,在这 n 次试验中,事件 A 发生的次数 n_A 称为事件 A 发生的频数. 比值 n_A/n 称为事件 A 发生的频率,并记成 $f_n(A)$.

由定义,易见频率具有下述基本性质:

$$1^\circ 0 \leq f_n(A) \leq 1;$$

$$2^\circ f_n(S) = 1;$$

3° 若 $A_1, A_2, \dots, A_k$ 是两两互不相容的事件,则

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

由于事件 A 发生的频率是它发生的次数与试验次数之比,其大小表示 A 发

生的频繁程度. 频率大, 事件 A 发生就频繁, 这意味着事件 A 在一次试验中发生的可能性就大. 反之亦然. 因而, 直观的想法是用频率来表示事件 A 在一次试验中发生的可能性的. 但是否可行, 先看下面的例子.

例 1 考虑“抛硬币”这个试验, 我们将一枚硬币抛掷 5 次、50 次、500 次, 各做 10 遍. 得到数据如表 1-1 所示 (其中 n_H 表示 H 发生的频数, $f_n(H)$ 表示 H 发生的频率).

表 1-1

实验序号	$n=5$		$n=50$		$n=500$	
	n_H	$f_n(H)$	n_H	$f_n(H)$	n_H	$f_n(H)$
1	2	0.4	22	0.44	251	0.502
2	3	0.6	25	0.50	249	0.498
3	1	0.2	21	0.42	256	0.512
4	5	1.0	25	0.50	253	0.506
5	1	0.2	24	0.48	251	0.502
6	2	0.4	21	0.42	246	0.492
7	4	0.8	18	0.36	244	0.488
8	2	0.4	24	0.48	258	0.516
9	3	0.6	27	0.54	262	0.524
10	3	0.6	31	0.62	247	0.494

这种试验历史上有人做过, 得到如表 1-2 所示的数据.

表 1-2

实验者	n	n_H	$f_n(H)$
德摩根	2 048	1 061	0.518 1
蒲 丰	4 040	2 048	0.506 9
K·皮尔逊	12 000	6 019	0.501 6
K·皮尔逊	24 000	12 012	0.500 5

从上述数据可以看出: 抛硬币次数 n 较小时, 频率 $f_n(H)$ 在 0 与 1 之间随机波动, 其幅度较大, 但随着 n 增大, 频率 $f_n(H)$ 呈现出稳定性. 即当 n 逐渐增大时 $f_n(H)$ 总是在 0.5 附近摆动, 而逐渐稳定于 0.5. $\square$

例 2 考察英语中特定字母出现的频率. 当观察字母的个数 n (试验的次

数)较小时,频率有较大幅度的随机波动.但当 n 增大时,频率呈现出稳定性.表 1-3 就是一份英文字母频率的统计表^①:

表 1-3

字母	频率	字母	频率	字母	频率
E	0.126 8	L	0.039 4	P	0.018 6
T	0.097 8	D	0.038 9	B	0.015 6
A	0.078 8	U	0.028 0	V	0.010 2
O	0.077 6	C	0.026 8	K	0.006 0
I	0.070 7	F	0.025 6	X	0.001 6
N	0.070 6	M	0.024 4	J	0.001 0
S	0.063 4	W	0.021 4	Q	0.000 9
R	0.059 4	Y	0.020 2	Z	0.000 6
H	0.057 3	G	0.018 7		

□

大量试验证实,当重复试验的次数 n 逐渐增大时,频率 $f_n(A)$ 呈现出稳定性,逐渐稳定于某个常数.这种“频率稳定性”即通常所说的统计规律性.我们让试验重复大量次数,计算频率 $f_n(A)$,以它来表征事件 A 发生可能性的大小是合适的.

但是,在实际中,我们不可能对每一个事件都做大量的试验,然后求得事件的频率,用以表征事件发生可能性的大小.同时,为了理论研究的需要,我们从频率的稳定性和频率的性质得到启发,给出如下表征事件发生可能性大小的概率的定义.

(二) 概率

定义 设 E 是随机试验, S 是它的样本空间.对于 E 的每一事件 A 赋予一个实数,记为 $P(A)$,称为事件 A 的概率,如果集合函数 $P(\cdot)$ 满足下列条件:

- 1° 非负性: 对于每一个事件 A ,有 $P(A) \geq 0$;
- 2° 规范性: 对于必然事件 S ,有 $P(S) = 1$;
- 3° 可列可加性: 设 $A_1, A_2, \dots$ 是两两互不相容的事件,即对于 $A_i A_j = \emptyset$, $i \neq j, i, j = 1, 2, \dots$, 有

^① 这是由 Dewey, G. 统计了约 438 023 个字母得到的. 引自 Relative Frequency of English Spellings (Teachers College Press, Columbia University, New York, 1970).

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots. \quad (3.1)$$

在第五章中将证明, 当 $n \rightarrow \infty$ 时频率 $f_n(A)$ 在一定意义下接近于概率 $P(A)$. 基于这一事实, 我们就有理由将概率 $P(A)$ 用来表征事件 A 在一次试验中发生的可能性的太小.

由概率的定义, 可以推得概率的一些重要性质.

性质 i $P(\emptyset) = 0$.

证 令 $A_n = \emptyset$ ($n=1, 2, \dots$), 则 $\bigcup_{n=1}^{\infty} A_n = \emptyset$, 且 $A_i A_j = \emptyset$, $i \neq j$, $i, j=1, 2, \dots$. 由概率的可列可加性(3.1)得

$$P(\emptyset) = P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} P(\emptyset).$$

由概率的非负性知, $P(\emptyset) \geq 0$, 故由上式知 $P(\emptyset) = 0$. □

性质 ii (有限可加性) 若 $A_1, A_2, \dots, A_n$ 是两两互不相容的事件, 则有

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n). \quad (3.2)$$

(3.2) 式称为概率的有限可加性.

证 令 $A_{n+1} = A_{n+2} = \cdots = \emptyset$, 即有 $A_i A_j = \emptyset$, $i \neq j$, $i, j=1, 2, \dots$. 由 (3.1) 式得

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) &= P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k) \\ &= \sum_{k=1}^n P(A_k) + 0 = P(A_1) + P(A_2) + \cdots + P(A_n). \end{aligned}$$

(3.2) 式得证. □

性质 iii 设 A, B 是两个事件, 若 $A \subset B$, 则有

$$P(B-A) = P(B) - P(A); \quad (3.3)$$

$$P(B) \geq P(A). \quad (3.4)$$

证 由 $A \subset B$ 知 $B = A \cup (B-A)$ (参见图 1-1), 且 $A(B-A) = \emptyset$, 再由概率的有限可加性(3.2), 得

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

(3.3) 得证; 又由概率的非负性 1° , $P(B-A) \geq 0$ 知

$$P(B) \geq P(A). \quad \square$$

性质 iv 对于任一事件 A ,

$$P(A) \leq 1.$$

证 因 $A \subset S$, 由性质 iii 得

$$P(A) \leq P(S) = 1. \quad \square$$

性质 v(逆事件的概率) 对于任一事件 A , 有

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

证 因 $A \cup \bar{A} = S$, 且 $A\bar{A} = \emptyset$, 由 (3.2) 式, 得

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

性质 v 得证. □

性质 vi(加法公式) 对于任意两事件 A, B 有

$$P(A \cup B) = P(A) + P(B) - P(AB). \quad (3.5)$$

证 因 $A \cup B = A \cup (B - AB)$ (参见图 1-2), 且 $A(B - AB) = \emptyset, AB \subset B$, 故由 (3.2) 式及 (3.3) 式得

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

(3.5) 式还能推广到多个事件的情况. 例如, 设 A_1, A_2, A_3 为任意三个事件, 则有

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) - P(A_1 A_2) \\ &\quad - P(A_1 A_3) - P(A_2 A_3) + P(A_1 A_2 A_3). \end{aligned} \quad (3.6)$$

一般, 对于任意 n 个事件 $A_1, A_2, \dots, A_n$, 可以用归纳法证得

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_n) &= \sum_{i=1}^n P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i A_j) \\ &\quad + \sum_{1 \leq i < j < k \leq n} P(A_i A_j A_k) + \dots + (-1)^{n-1} P(A_1 A_2 \dots A_n). \end{aligned} \quad (3.7)$$

§ 4 等可能概型(古典概型)

§ 1 中所说的试验 E_1, E_4 , 它们具有两个共同的特点:

- 1° 试验的样本空间只包含有限个元素;
- 2° 试验中每个基本事件发生的可能性相同.

具有以上两个特点的试验是大量存在的. 这种试验称为等可能概型. 它在概率论发展初期曾是主要的研究对象, 所以也称为古典概型. 等可能概型的一些概念具有直观、容易理解的特点, 有着广泛的应用.

下面我们来讨论等可能概型中事件概率的计算公式.

设试验的样本空间为 $S = \{e_1, e_2, \dots, e_n\}$. 由于在试验中每个基本事件发生的可能性相同, 即有

$$P(\{e_1\}) = P(\{e_2\}) = \dots = P(\{e_n\}).$$

又由于基本事件是两两互不相容的, 于是

$$\begin{aligned}
 1 &= P(S) = P(\{e_1\} \cup \{e_2\} \cup \cdots \cup \{e_n\}) \\
 &= P(\{e_1\}) + P(\{e_2\}) \\
 &\quad + \cdots + P(\{e_n\}) = nP(\{e_i\}),
 \end{aligned}$$

$$P(\{e_i\}) = \frac{1}{n}, \quad i = 1, 2, \cdots, n.$$

若事件 A 包含 k 个基本事件, 即 $A = \{e_{i_1}\} \cup \{e_{i_2}\} \cup \cdots \cup \{e_{i_k}\}$, 这里 $i_1, i_2, \cdots, i_k$ 是 $1, 2, \cdots, n$ 中某 k 个不同的数. 则有

$$P(A) = \sum_{j=1}^k P(\{e_{i_j}\}) = \frac{k}{n} = \frac{A \text{ 包含的基本事件数}}{S \text{ 中基本事件的总数}}. \quad (4.1)$$

(4.1) 式就是等可能概型中事件 A 的概率的计算公式^①.

例 1 将一枚硬币抛掷三次. (1) 设事件 A_1 为“恰有一次出现正面”, 求 $P(A_1)$; (2) 设事件 A_2 为“至少有一次出现正面”, 求 $P(A_2)$.

解 (1) 我们考虑 § 1 中 E_2 的样本空间:

$$S_2 = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

而 $A_1 = \{HTT, THT, TTH\}$.

S_2 中包含有限个元素, 且由对称性知每个基本事件发生的可能性相同. 故由 (4.1) 式, 得

$$P(A_1) = \frac{3}{8}.$$

(2) 由于 $\bar{A}_2 = \{TTT\}$, 于是

$$P(A_2) = 1 - P(\bar{A}_2) = 1 - \frac{1}{8} = \frac{7}{8}. \quad \square$$

当样本空间的元素较多时, 我们一般不再将 S 中的元素一一列出, 而只需分别求出 S 中与 A 中包含的元素的个数 (即基本事件的个数), 再由 (4.1) 式即可求出 A 的概率.

例 2 一个口袋装有 6 只球, 其中 4 只白球、2 只红球. 从袋中取球两次, 每次随机地取一只. 考虑两种取球方式: (a) 第一次取一只球, 观察其颜色后放回袋中, 搅匀后再取一球. 这种取球方式叫做放回抽样. (b) 第一次取一球不放回袋中, 第二次从剩余的球中再取一球. 这种取球方式叫做不放回抽样. 试分别就上面两种情况求 (1) 取到的两只球都是白球的概率; (2) 取到的两只球颜色相同的概率; (3) 取到的两只球中至少有一只是白球的概率.

解 (a) 放回抽样的情况.

^① 易知由 (4.1) 所确定的概率满足非负性、规范性和有限可加性. 但此时由于 S 中只含有限个子集 (只有 $C_n^0 + C_n^1 + \cdots + C_n^n = 2^n$ 个子集). 因而若在 S 中取可列无限个两两互不相容的事件 $A_1, A_2, \cdots, A_n, \cdots$, 则其中必包含无限多个不可能事件, 即知可列可加性与有限可加性是等价的.

以 A, B, C 分别表示事件“取到的两只球都是白球”, “取到的两只球都是红球”, “取到的两只球中至少有一只是白球”. 易知“取到两只颜色相同的球”这一事件即为 $A \cup B$, 而 $C = \bar{B}$.

在袋中依次取两只球, 每一种取法为一个基本事件, 显然此时样本空间中仅包含有限个元素. 且由对称性知每个基本事件发生的可能性相同, 因而可利用 (4.1) 式来计算事件的概率.

第一次从袋中取球有 6 只球可供抽取, 第二次也有 6 只球可供抽取. 由组合法的乘法原理, 共有 6×6 种取法. 即样本空间中元素总数为 6×6 . 对于事件 A 而言, 由于第一次有 4 只白球可供抽取, 第二次也有 4 只白球可供抽取, 由乘法原理共有 4×4 种取法, 即 A 中包含 4×4 个元素. 同理, B 中包含 2×2 个元素. 于是

$$P(A) = \frac{4 \times 4}{6 \times 6} = \frac{4}{9}.$$

$$P(B) = \frac{2 \times 2}{6 \times 6} = \frac{1}{9}.$$

由于 $AB = \emptyset$, 得

$$P(A \cup B) = P(A) + P(B) = \frac{5}{9}.$$

$$P(C) = P(\bar{B}) = 1 - P(B) = \frac{8}{9}.$$

(b) 不放回抽样的情况.

由读者自己完成. □

例 3 将 n 只球随机地放入 N ($N \geq n$) 个盒子中去, 试求每个盒子至多有一只球的概率 (设盒子的容量不限).

解 将 n 只球放入 N 个盒子中去, 每一种放法是一基本事件. 易知, 这是古典概率问题. 因每一只球都可以放入 N 个盒子中的任一个盒子, 故共有 $N \times N \times \cdots \times N = N^n$ 种不同的放法, 而每个盒子中至多放一只球共有 $N(N-1)\cdots[N-(n-1)]$ 种不同放法. 因而所求的概率为

$$p = \frac{N(N-1)\cdots(N-n+1)}{N^n} = \frac{A_N^n}{N^n}.$$

有许多问题和本例具有相同的数学模型. 例如, 假设每人的生日在一年 365 天中的任一天是等可能的, 即都等于 $1/365$, 那么随机选取 n ($n \leq 365$) 个人, 他们的生日各不相同的概率为

$$\frac{365 \cdot 364 \cdot \cdots \cdot (365 - n + 1)}{365^n}.$$

因而, n 个人中至少有两人生日相同的概率为

$$p = 1 - \frac{365 \cdot 364 \cdot \cdots \cdot (365 - n + 1)}{365^n}.$$

经计算可得下述结果:

n	20	23	30	40	50	64	100
p	0.411	0.507	0.706	0.891	0.970	0.997	0.999 999 7

从上表可看出,在仅有 64 人的班级里,“至少有两人生日相同”这一事件的概率与 1 相差无几,因此,如作调查的话,几乎总是会出现的.读者不妨试一试. $\square$

例 4 设有 N 件产品,其中有 D 件次品,今从中任取 n 件,问其中恰有 k ($k \leq D$) 件次品的概率是多少?

解 在 N 件产品中抽取 n 件(这里是指不放回抽样),所有可能的取法共有 $\binom{N}{n}$ ^① 种,每一种取法为一基本事件,且由于对称性知每个基本事件发生的可能性相同.又因在 D 件次品中取 k 件,所有可能的取法有 $\binom{D}{k}$ 种.在 $N-D$ 件正品中取 $n-k$ 件所有可能的取法有 $\binom{N-D}{n-k}$ 种,由乘法原理知在 N 件产品中取 n 件,其中恰有 k 件次品的取法共有 $\binom{D}{k} \binom{N-D}{n-k}$ 种,于是所求概率为

$$p = \frac{\binom{D}{k} \binom{N-D}{n-k}}{\binom{N}{n}}. \quad (4.2)$$

(4.2) 式即所谓超几何分布的概率公式. $\square$

例 5 袋中有 a 只白球, b 只红球, k 个人依次在袋中取一只球, (1) 作放回抽样; (2) 作不放回抽样, 求第 i ($i=1, 2, \dots, k$) 人取到白球(记为事件 B) 的概率 ($k \leq a+b$).

解 (1) 放回抽样的情况, 显然有

$$P(B) = \frac{a}{a+b}.$$

(2) 不放回抽样的情况. 各人取一只球, 每种取法是一个基本事件. 共有 $(a+b)(a+b-1)\cdots(a+b-k+1) = A_{a+b}^k$ 个基本事件, 且由于对称性知每个基本事

① 对于任意实数 a 以及非负整数 r , 定义 $\binom{a}{r} = \frac{a(a-1)\cdots(a-r+1)}{r!}$, $\binom{a}{0} = 1$. 例如 $\binom{-\pi}{3} = \frac{(-\pi)(-\pi-1)(-\pi-2)}{3!} = -\frac{\pi(\pi+1)(\pi+2)}{3!}$. 特别, 当 a 为正整数, 且 $r \leq a$ 时, $\binom{a}{r}$ 即为组合数, 即 $\binom{a}{r} = C_a^r$.

件发生的可能性相同. 当事件 B 发生时, 第 i 人取的应是白球, 它可以是 a 只白球中的任一只, 有 a 种取法. 其余被取的 $k-1$ 只球可以是其余 $a+b-1$ 只球中的任意 $k-1$ 只, 共有 $(a+b-1)(a+b-2)\cdots[a+b-1-(k-1)+1]=A_{a+b-1}^{k-1}$ 种取法, 于是事件 B 中包含 $a \cdot A_{a+b-1}^{k-1}$ 个基本事件, 故由 (4.1) 式得到

$$P(B) = a \cdot A_{a+b-1}^{k-1} / A_{a+b}^k = \frac{a}{a+b}.$$

值得注意的是 $P(B)$ 与 i 无关, 即 k 个人取球, 尽管取球的先后次序不同, 各人取到白球的概率是一样的, 大家机会相同 (例如在购买福利彩票时, 各人得奖的机会是一样的). 另外还值得注意的是放回抽样的情况与不放回抽样的情况下 $P(B)$ 是一样的. $\square$

例 6 在 $1 \sim 2000$ 的整数中随机地取一个数, 问取到的整数既不能被 6 整除, 又不能被 8 整除的概率是多少?

解 设 A 为事件“取到的数能被 6 整除”, B 为事件“取到的数能被 8 整除”, 则所求概率为

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

由于

$$333 < \frac{2000}{6} < 334,$$

故得

$$P(A) = \frac{333}{2000}.$$

由于

$$\frac{2000}{8} = 250,$$

故得

$$P(B) = \frac{250}{2000}.$$

又由于一个数同时能被 6 与 8 整除, 就相当于能被 24 整除, 因此, 由

$$83 < \frac{2000}{24} < 84,$$

得

$$P(AB) = \frac{83}{2000}.$$

于是所求概率为

$$p = 1 - \left(\frac{333}{2000} + \frac{250}{2000} - \frac{83}{2000} \right) = \frac{3}{4}. \quad \square$$

例 7 将 15 名新生随机地平均分配到三个班级中去, 这 15 名新生中有 3 名是优秀生. 问 (1) 每个班级各分配到一名优秀生的概率是多少? (2) 3 名优秀生分配在同一班级的概率是多少?

解 15 名新生平均分配到三个班级中的分法总数为

$$\binom{15}{5}\binom{10}{5}\binom{5}{5} = \frac{15!}{5!5!5!}.$$

每一种分配法为一基本事件,且由对称性易知每个基本事件发生的可能性相同.

(1) 将 3 名优秀生分配到三个班级使每个班级都有一名优秀生的分法共 3! 种. 对于这每一种分法,其余 12 名新生平均分配到三个班级中的分法共有 $\frac{12!}{4!4!4!}$ 种. 因此,每一班级各分配到一名优秀生的分法共有 $\frac{3!12!}{4!4!4!}$ 种. 于是所求概率为

$$p_1 = \frac{3! \times 12!}{4!4!4!} / \frac{15!}{5!5!5!} = \frac{25}{91}.$$

(2) 将 3 名优秀生分配在同一班级的分法共有 3 种. 对于这每一种分法,其余 12 名新生的分法(一个班级 2 名,另两个班级各 5 名)有 $\frac{12!}{2!5!5!}$ 种. 因此 3 名优秀生分配在同一班级的分法共有 $\frac{3 \times 12!}{2!5!5!}$ 种,于是,所求概率为

$$p_2 = \frac{3 \times 12!}{2!5!5!} / \frac{15!}{5!5!5!} = \frac{6}{91}. \quad \square$$

例 8 某接待站在某一周曾接待过 12 次来访,已知所有这 12 次接待都是在周二和周四进行的,问是否可以推断接待时间是有规定的?

解 假设接待站的接待时间没有规定,而各来访者在一周的任一天中去接待站是等可能的,那么,12 次接待来访者都在周二、周四的概率为

$$\frac{2^{12}}{7^{12}} = 0.000\ 000\ 3.$$

人们在长期的实践中总结得到“概率很小的事件在一次试验中实际上几乎是不发生的”(称之为实际推断原理). 现在概率很小的事件在一次试验中竟然发生了,因此有理由怀疑假设的正确性,从而推断接待站不是每天都接待来访者,即认为其接待时间是有规定的. $\square$

§5 条件概率

(一) 条件概率

条件概率是概率论中的一个重要而实用的概念. 所考虑的是事件 A 已发生的条件下事件 B 发生的概率. 先举一个例子.

例 1 将一枚硬币抛掷两次,观察其出现正反面的情况. 设事件 A 为“至少有一次为 H”,事件 B 为“两次掷出同一面”. 现在来求已知事件 A 已经发生的条件下事件 B 发生的概率.

这里,样本空间为 $S = \{HH, HT, TH, TT\}$, $A = \{HH, HT, TH\}$, $B = \{HH, TT\}$. 易知此属古典概型问题. 已知事件 A 已发生,有了这一信息,知道 TT 不可能发生,即知试验所有可能结果所成的集合就是 A . A 中共有 3 个元素,其中只有 $HH \in B$. 于是,在事件 A 发生的条件下事件 B 发生的概率(记为 $P(B|A)$)为

$$P(B|A) = \frac{1}{3}. \quad \square$$

在这里,我们看到 $P(B) = 2/4 \neq P(B|A)$. 这是很容易理解的,因为在求 $P(B|A)$ 时我们是限制在事件 A 已经发生的条件下考虑事件 B 发生的概率的.

另外,易知

$$P(A) = \frac{3}{4}, P(AB) = \frac{1}{4}, P(B|A) = \frac{1}{3} = \frac{1/4}{3/4},$$

故有

$$P(B|A) = \frac{P(AB)}{P(A)}. \quad (5.1)$$

对于一般古典概型问题,若仍以 $P(B|A)$ 记事件 A 已经发生的条件下事件 B 发生的概率,则关系式(5.1)仍然成立. 事实上,设试验的基本事件总数为 n , A 所包含的基本事件数为 m ($m > 0$), AB 所包含的基本事件数为 k , 即有

$$P(B|A) = \frac{k}{m} = \frac{k/n}{m/n} = \frac{P(AB)}{P(A)}.$$

在一般场合,我们将上述关系式作为条件概率的定义.

定义 设 A, B 是两个事件,且 $P(A) > 0$, 称

$$P(B|A) = \frac{P(AB)}{P(A)} \quad (5.2)$$

为在事件 A 发生的条件下事件 B 发生的条件概率.

不难验证,条件概率 $P(\cdot | A)$ 符合概率定义中的三个条件,即

- 1° 非负性: 对于每一事件 B , 有 $P(B|A) \geq 0$;
- 2° 规范性: 对于必然事件 S , 有 $P(S|A) = 1$;
- 3° 可列可加性: 设 $B_1, B_2, \dots$ 是两两互不相容的事件, 则有

$$P\left(\bigcup_{i=1}^{\infty} B_i \mid A\right) = \sum_{i=1}^{\infty} P(B_i \mid A).$$

既然条件概率符合上述三个条件,故 §3 中对概率所证明的一些重要结果都适用于条件概率. 例如,对于任意事件 B_1, B_2 有

$$P(B_1 \cup B_2 | A) = P(B_1 | A) + P(B_2 | A) - P(B_1 B_2 | A).$$

例 2 一盒子装有 4 只产品,其中有 3 只一等品,1 只二等品. 从中取产品两

次,每次任取一只,作不放回抽样. 设事件 A 为“第一次取到的是一等品”,事件 B 为“第二次取到的是一等品”. 试求条件概率 $P(B|A)$.

解 易知此属古典概型问题. 将产品编号, 1, 2, 3 号为一等品; 4 号为二等品. 以 (i, j) 表示第一次、第二次分别取到第 i 号、第 j 号产品. 试验 E (取产品两次, 记录其号码) 的样本空间为

$$S = \{(1, 2), (1, 3), (1, 4), (2, 1), (2, 3), (2, 4), \dots, (4, 1), (4, 2), (4, 3)\},$$

$$A = \{(1, 2), (1, 3), (1, 4), (2, 1), (2, 3), (2, 4), (3, 1), (3, 2), (3, 4)\},$$

$$AB = \{(1, 2), (1, 3), (2, 1), (2, 3), (3, 1), (3, 2)\}.$$

按(5.2)式, 得条件概率

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{6/12}{9/12} = \frac{2}{3}.$$

也可以直接按条件概率的含义来求 $P(B|A)$. 我们知道, 当事件 A 发生以后, 试验 E 所有可能结果的集合就是 A , A 中有 9 个元素, 其中只有 $(1, 2)$, $(1, 3)$, $(2, 1)$, $(2, 3)$, $(3, 1)$, $(3, 2)$ 属于 B , 故可得

$$P(B|A) = \frac{6}{9} = \frac{2}{3}. \quad \square$$

(二) 乘法定理

由条件概率的定义(5.2), 立即可得下述定理.

乘法定理 设 $P(A) > 0$, 则有

$$P(AB) = P(B|A)P(A). \quad (5.3)$$

(5.3) 式称为乘法公式.

(5.3) 式容易推广到多个事件的积事件的情况. 例如, 设 A, B, C 为事件, 且 $P(AB) > 0$, 则有

$$P(ABC) = P(C|AB)P(B|A)P(A). \quad (5.4)$$

在这里, 注意到由假设 $P(AB) > 0$ 可推得 $P(A) \geq P(AB) > 0$.

一般, 设 $A_1, A_2, \dots, A_n$ 为 n 个事件, $n \geq 2$, 且 $P(A_1 A_2 \cdots A_{n-1}) > 0$, 则有

$$P(A_1 A_2 \cdots A_n) = P(A_n | A_1 A_2 \cdots A_{n-1}) P(A_{n-1} | A_1 A_2 \cdots A_{n-2}) \cdots P(A_2 | A_1) P(A_1). \quad (5.5)$$

例 3 设袋中装有 r 只红球, t 只白球. 每次自袋中任取一只球, 观察其颜色然后放回, 并再放入 a 只与所取出的那只球同色的球. 若在袋中连续取球四次, 试求第一、二次取到红球且第三、四次取到白球的概率.

解 以 $A_i (i=1, 2, 3, 4)$ 表示事件“第 i 次取到红球”, 则 $\bar{A}_3, \bar{A}_4$ 分别表示事件第三、四次取到白球. 所求概率为

$$P(A_1 A_2 \bar{A}_3 \bar{A}_4) = P(\bar{A}_4 | A_1 A_2 \bar{A}_3) P(\bar{A}_3 | A_1 A_2) P(A_2 | A_1) P(A_1)$$

$$= \frac{t+a}{r+t+3a} \cdot \frac{t}{r+t+2a} \cdot \frac{r+a}{r+t+a} \cdot \frac{r}{r+t}.$$

□

例 4 设某光学仪器厂制造的透镜,第一次落下时打破的概率为 $1/2$,若第一次落下未打破,第二次落下打破的概率为 $7/10$,若前两次落下未打破,第三次落下打破的概率为 $9/10$. 试求透镜落下三次而未打破的概率.

解 以 $A_i (i=1,2,3)$ 表示事件“透镜第 i 次落下打破”,以 B 表示事件“透镜落下三次而未打破”. 因为 $B = \bar{A}_1 \bar{A}_2 \bar{A}_3$, 故有

$$\begin{aligned} P(B) &= P(\bar{A}_1 \bar{A}_2 \bar{A}_3) = P(\bar{A}_3 | \bar{A}_1 \bar{A}_2) P(\bar{A}_2 | \bar{A}_1) P(\bar{A}_1) \\ &= \left(1 - \frac{9}{10}\right) \left(1 - \frac{7}{10}\right) \left(1 - \frac{1}{2}\right) = \frac{3}{200}. \end{aligned}$$

另解,按题意

$$\bar{B} = A_1 \cup \bar{A}_1 A_2 \cup \bar{A}_1 \bar{A}_2 A_3.$$

而 $A_1, \bar{A}_1 A_2, \bar{A}_1 \bar{A}_2 A_3$ 是两两互不相容的事件,故有

$$P(\bar{B}) = P(A_1) + P(\bar{A}_1 A_2) + P(\bar{A}_1 \bar{A}_2 A_3).$$

已知 $P(A_1) = \frac{1}{2}, P(A_2 | \bar{A}_1) = \frac{7}{10}, P(A_3 | \bar{A}_1 \bar{A}_2) = \frac{9}{10}$, 即有

$$\begin{aligned} P(\bar{A}_1 A_2) &= P(A_2 | \bar{A}_1) P(\bar{A}_1) = \frac{7}{10} \left(1 - \frac{1}{2}\right) = \frac{7}{20}, \\ P(\bar{A}_1 \bar{A}_2 A_3) &= P(A_3 | \bar{A}_1 \bar{A}_2) P(\bar{A}_2 | \bar{A}_1) P(\bar{A}_1) \\ &= \frac{9}{10} \left(1 - \frac{7}{10}\right) \left(1 - \frac{1}{2}\right) = \frac{27}{200}. \end{aligned}$$

故得

$$P(\bar{B}) = \frac{1}{2} + \frac{7}{20} + \frac{27}{200} = \frac{197}{200},$$

$$P(B) = 1 - \frac{197}{200} = \frac{3}{200}.$$

□

(三) 全概率公式和贝叶斯公式

下面建立两个用来计算概率的重要公式. 先介绍样本空间的划分的定义.

定义 设 S 为试验 E 的样本空间, $B_1, B_2, \dots, B_n$ 为 E 的一组事件. 若

(i) $B_i B_j = \emptyset, i \neq j, i, j = 1, 2, \dots, n$;

(ii) $B_1 \cup B_2 \cup \dots \cup B_n = S$,

则称 $B_1, B_2, \dots, B_n$ 为样本空间 S 的一个划分.

若 $B_1, B_2, \dots, B_n$ 是样本空间的一个划分, 那么, 对每次试验, 事件 $B_1, B_2, \dots, B_n$ 中必有一个且仅有一个发生.

例如, 设试验 E 为“掷一颗骰子观察其点数”. 它的样本空间为 $S = \{1, 2, 3, 4, 5, 6\}$. E 的一组事件 $B_1 = \{1, 2, 3\}, B_2 = \{4, 5\}, B_3 = \{6\}$ 是 S 的一个划分. 而事

件组 $C_1 = \{1, 2, 3\}, C_2 = \{3, 4\}, C_3 = \{5, 6\}$ 不是 S 的划分.

定理 设试验 E 的样本空间为 S , A 为 E 的事件, $B_1, B_2, \dots, B_n$ 为 S 的一个划分, 且 $P(B_i) > 0$ ($i=1, 2, \dots, n$), 则

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n). \quad (5.6)$$

(5.6) 式称为**全概率公式**.

在很多实际问题中 $P(A)$ 不易直接求得, 但却容易找到 S 的一个划分 $B_1, B_2, \dots, B_n$, 且 $P(B_i)$ 和 $P(A|B_i)$ 或为已知, 或容易求得, 那么就可以根据 (5.6) 式求出 $P(A)$.

证 因为

$$A = AS = A(B_1 \cup B_2 \cup \dots \cup B_n) = AB_1 \cup AB_2 \cup \dots \cup AB_n,$$

由假设 $P(B_i) > 0$ ($i=1, 2, \dots, n$), 且 $(AB_i)(AB_j) = \emptyset, i \neq j, i, j=1, 2, \dots, n$ 得到

$$\begin{aligned} P(A) &= P(AB_1) + P(AB_2) + \dots + P(AB_n) \\ &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots \\ &\quad + P(A|B_n)P(B_n). \end{aligned} \quad \square$$

另一个重要公式是下述的贝叶斯公式.

定理 设试验 E 的样本空间为 S . A 为 E 的事件, $B_1, B_2, \dots, B_n$ 为 S 的一个划分, 且 $P(A) > 0, P(B_i) > 0$ ($i=1, 2, \dots, n$), 则

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^n P(A|B_j)P(B_j)}, \quad i=1, 2, \dots, n. \quad (5.7)$$

(5.7) 式称为**贝叶斯(Bayes)公式**①.

证 由条件概率的定义及全概率公式即得

$$P(B_i|A) = \frac{P(B_i A)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^n P(A|B_j)P(B_j)}, \quad i=1, 2, \dots, n. \quad \square$$

特别在 (5.6) 式, (5.7) 式中取 $n=2$, 并将 B_1 记为 B , 此时 B_2 就是 $\bar{B}$, 那么, 全概率公式和贝叶斯公式分别成为

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B}), \quad (5.8)$$

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})}. \quad (5.9)$$

这两个公式是常用的.

① 在全概率公式和贝叶斯公式中, 要求 $B_1, B_2, \dots, B_n$ 是 S 的一个划分, 将这一条件改为“ $B_i B_j = \emptyset, i \neq j, i, j=1, 2, \dots, n$, 且 $P(B_1 \cup B_2 \cup \dots \cup B_n) = 1$ ”两个公式仍然成立.

例 5 某电子设备制造厂所用的元件是由三家元件制造厂提供的. 根据以往的记录有以下的数据:

元件制造厂	次品率	提供元件的份额
1	0.02	0.15
2	0.01	0.80
3	0.03	0.05

设这三家工厂的产品在仓库中是均匀混合的, 且无区别的标志. (1) 在仓库中随机地取一只元件, 求它是次品的概率; (2) 在仓库中随机地取一只元件, 若已知取到的是次品, 为分析此次品出自何厂, 需求出此次品由三家工厂生产的概率分别是多少. 试求这些概率.

解 设 A 表示“取到的是一只次品”, $B_i (i=1, 2, 3)$ 表示“所取到的产品是由第 i 家工厂提供的”. 易知, B_1, B_2, B_3 是样本空间 S 的一个划分, 且有

$$P(B_1)=0.15, P(B_2)=0.80, P(B_3)=0.05,$$

$$P(A|B_1)=0.02, P(A|B_2)=0.01, P(A|B_3)=0.03.$$

(1) 由全概率公式

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\ &= 0.0125. \end{aligned}$$

(2) 由贝叶斯公式

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A)} = \frac{0.02 \times 0.15}{0.0125} = 0.24.$$

$$P(B_2|A) = 0.64, \quad P(B_3|A) = 0.12.$$

以上结果表明, 这只次品来自第 2 家工厂的可能性最大. □

例 6 据美国的一份资料报导, 在美国总的来说患肺癌的概率约为 0.1%, 在人群中 20% 是吸烟者, 他们患肺癌的概率约为 0.4%, 求不吸者患肺癌的概率是多少?

解 以 C 记事件“患肺癌”, 以 A 记事件“吸烟”, 按题意 $P(C)=0.001$, $P(A)=0.20$, $P(C|A)=0.004$. 需要求条件概率 $P(C|\bar{A})$. 由全概率公式有

$$P(C) = P(C|A)P(A) + P(C|\bar{A})P(\bar{A}).$$

将数据代入, 得

$$\begin{aligned} 0.001 &= 0.004 \times 0.20 + P(C|\bar{A})P(\bar{A}) \\ &= 0.004 \times 0.20 + P(C|\bar{A}) \times 0.80, \end{aligned}$$

$$P(C|\bar{A}) = 0.00025. \quad \square$$

例 7 对以往数据分析结果表明, 当机器调整得良好时, 产品的合格率为

98%,而当机器发生某种故障时,其合格率为 55%. 每天早上机器开动时,机器调整良好的概率为 95%. 试求已知某日早上第一件产品是合格品时,机器调整良好的概率是多少?

解 设 A 为事件“产品合格”, B 为事件“机器调整良好”. 已知 $P(A|B)=0.98$, $P(A|\bar{B})=0.55$, $P(B)=0.95$, $P(\bar{B})=0.05$, 所需求的概率为 $P(B|A)$. 由贝叶斯公式

$$\begin{aligned} P(B|A) &= \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|\bar{B})P(\bar{B})} \\ &= \frac{0.98 \times 0.95}{0.98 \times 0.95 + 0.55 \times 0.05} = 0.97. \end{aligned}$$

这就是说,当生产出第一件产品是合格品时,此时机器调整良好的概率为 0.97. 这里,概率 0.95 是由以往的数据分析得到的,叫做先验概率. 而在得到信息(即生产出的第一件产品是合格品)之后再重新加以修正的概率(即 0.97)叫做后验概率. 有了后验概率我们就能对机器的情况有进一步的了解. $\square$

例 8 根据以往的临床记录,某种诊断癌症的试验具有如下的效果:若以 A 表示事件“试验反应为阳性”,以 C 表示事件“被诊断者患有癌症”,则有 $P(A|C)=0.95$, $P(\bar{A}|\bar{C})=0.95$. 现在对自然人群进行普查,设被试验的人患有癌症的概率为 0.005,即 $P(C)=0.005$,试求 $P(C|A)$.

解 已知 $P(A|C)=0.95$, $P(A|\bar{C})=1-P(\bar{A}|\bar{C})=0.05$, $P(C)=0.005$, $P(\bar{C})=0.995$, 由贝叶斯公式

$$P(C|A) = \frac{P(A|C)P(C)}{P(A|C)P(C)+P(A|\bar{C})P(\bar{C})} = 0.087.$$

本题的结果表明,虽然 $P(A|C)=0.95$, $P(\bar{A}|\bar{C})=0.95$, 这两个概率都比较高. 但若将此试验用于普查,则有 $P(C|A)=0.087$, 亦即其正确性只有 8.7% (平均 1000 个具有阳性反应的人中大约只有 87 人确患有癌症). 如果不注意到这一点,将会得出错误的诊断,这也说明,若将 $P(A|C)$ 和 $P(C|A)$ 混淆了会造成不良的后果. $\square$

§6 独立性

设 A, B 是试验 E 的两事件,若 $P(A)>0$, 可以定义 $P(B|A)$. 一般, A 的发生对 B 发生的概率是有影响的,这时 $P(B|A) \neq P(B)$, 只有在这种影响不存在时才会有 $P(B|A)=P(B)$, 这时有

$$P(AB) = P(B|A)P(A) = P(A)P(B).$$

例 1 设试验 E 为“抛甲、乙两枚硬币,观察正反面出现的情况”. 设事件 A

为“甲币出现 H ”，事件 B 为“乙币出现 H ”. E 的样本空间为

$$S = \{HH, HT, TH, TT\}.$$

由 (4.1) 式得

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

$$P(B|A) = \frac{1}{2}, \quad P(AB) = \frac{1}{4}.$$

在这里我们看到 $P(B|A) = P(B)$, 而 $P(AB) = P(A)P(B)$. 事实上, 由题意, 显然甲币是否出现正面与乙币是否出现正面是互不影响的. $\square$

定义 设 A, B 是两事件, 如果满足等式

$$P(AB) = P(A)P(B), \quad (6.1)$$

则称事件 A, B 相互独立, 简称 A, B 独立.

容易知道, 若 $P(A) > 0, P(B) > 0$, 则 A, B 相互独立与 A, B 互不相容不能同时成立.

定理一 设 A, B 是两事件, 且 $P(A) > 0$. 若 A, B 相互独立, 则 $P(B|A) = P(B)$. 反之亦然.

定理的正确性是显然的.

定理二 若事件 A 与 B 相互独立, 则下列各对事件也相互独立:

$$A \text{ 与 } \bar{B}, \quad \bar{A} \text{ 与 } B, \quad \bar{A} \text{ 与 } \bar{B}.$$

证 因为 $A = A(B \cup \bar{B}) = AB \cup A\bar{B}$, 得

$$\begin{aligned} P(A) &= P(AB \cup A\bar{B}) = P(AB) + P(A\bar{B}) \\ &= P(A)P(B) + P(A\bar{B}), \end{aligned}$$

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

因此 A 与 $\bar{B}$ 相互独立. 由此可立即推出 $\bar{A}$ 与 B 相互独立. 再由 $\bar{\bar{B}} = B$, 又推出 $\bar{A}$ 与 B 相互独立. $\square$

下面我们将独立性的概念推广到三个事件的情况.

定义 设 A, B, C 是三个事件, 如果满足等式

$$\left. \begin{aligned} P(AB) &= P(A)P(B), \\ P(BC) &= P(B)P(C), \\ P(AC) &= P(A)P(C), \\ P(ABC) &= P(A)P(B)P(C), \end{aligned} \right\} \quad (6.2)$$

则称事件 A, B, C 相互独立.

一般, 设 $A_1, A_2, \dots, A_n$ 是 n ($n \geq 2$) 个事件, 如果对于其中任意 2 个, 任意 3 个, $\dots$, 任意 n 个事件的积事件的概率, 都等于各事件概率之积, 则称事件 $A_1, A_2, \dots, A_n$ 相互独立.

由定义,可以得到以下两个推论.

1°若事件 $A_1, A_2, \dots, A_n (n \geq 2)$ 相互独立,则其中任意 $k (2 \leq k \leq n)$ 个事件也是相互独立的.

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

1°由独立性定义可直接推出. 2°从直观上看是显然的. 对于 $n=2$ 时,在定理二已作了证明,一般的情况由数学归纳法容易证得,此处略.

两事件相互独立的含义是它们中一个已发生,不影响另一个发生的概率. 在实际应用中,对于事件的独立性常常是根据事件的实际意义去判断. 一般,若由实际情况分析, A, B 两事件之间没有关联或关联很微弱,那就认为它们是相互独立的. 例如 A, B 分别表示甲、乙两人患感冒. 如果甲、乙两人的活动范围相距甚远. 就认为 A, B 相互独立. 若甲乙两人是同住在一个房间里的,那就不能认为 A, B 相互独立了.

例 2 一个元件(或系统)能正常工作的概率称为元件(或系统)的可靠性. 如图 1-8, 设有 4 个独立工作的元件 1, 2, 3, 4 按先串联再并联的方式连接(称为串并联系统). 设第 i 个元件的可靠性为 $p_i (i=1, 2, 3, 4)$, 试求系统的可靠性.

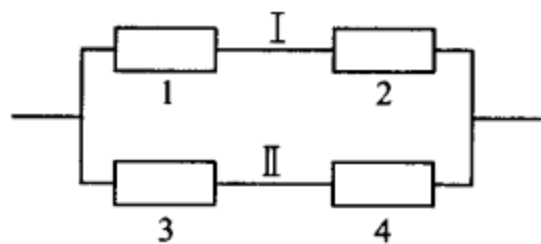


图 1-8

解 以 $A_i (i=1, 2, 3, 4)$ 表示事件“第 i 个元件正常工作”, 以 A 表示事件“系统正常工作”.

系统由两条线路 I 和 II 组成(如图 1-8). 当且仅当至少有一条线路中的两个元件均正常工作时这一系统正常工作, 故有

$$A = A_1 A_2 \cup A_3 A_4.$$

由事件的独立性,得系统的可靠性

$$\begin{aligned} P(A) &= P(A_1 A_2) + P(A_3 A_4) - P(A_1 A_2 A_3 A_4) \\ &= P(A_1)P(A_2) + P(A_3)P(A_4) - P(A_1)P(A_2)P(A_3)P(A_4) \\ &= p_1 p_2 + p_3 p_4 - p_1 p_2 p_3 p_4. \end{aligned}$$

□

例 3 要验收一批 (100 件) 乐器. 验收方案如下: 自该批乐器中随机地取 3 件测试 (设 3 件乐器的测试的结果是相互独立的), 如果 3 件中至少有一件在测试中被认为音色不纯, 则这批乐器就被拒绝接收. 设一件音色不纯的乐器经测试查出其为音色不纯的概率为 0.95; 而一件音色纯的乐器经测试被误认为不纯的概率为 0.01. 如果已知这 100 件乐器中恰有 4 件是音色不纯的. 试问这批乐器被接收的概率是多少?

解 设以 $H_i (i=0, 1, 2, 3)$ 表示事件“随机地取出 3 件乐器, 其中恰有 i 件音色不纯”, H_0, H_1, H_2, H_3 是 S 的一个划分, 以 A 表示事件“这批乐器被接收”.

已知一件音色纯的乐器,经测试被认为音色纯的概率为 0.99,而一件音色不纯的乐器,经测试被误认为音色纯的概率为 0.05,并且 3 件乐器的测试的结果是相互独立的,于是有

$$P(A|H_0)=0.99^3, P(A|H_1)=0.99^2 \times 0.05,$$

$$P(A|H_2)=0.99 \times 0.05^2, P(A|H_3)=0.05^3,$$

而
$$P(H_0)=\frac{\binom{96}{3}}{\binom{100}{3}}, \quad P(H_1)=\frac{\binom{4}{1}\binom{96}{2}}{\binom{100}{3}},$$

$$P(H_2)=\frac{\binom{4}{2}\binom{96}{1}}{\binom{100}{3}}, \quad P(H_3)=\frac{\binom{4}{3}}{\binom{100}{3}}.$$

故
$$P(A)=\sum_{i=0}^3 P(A|H_i)P(H_i)=0.8574+0.0055+0+0$$

$$=0.8629.$$
 □

例 4 甲、乙两人进行乒乓球比赛,每局甲胜的概率为 $p, p \geq 1/2$. 问对甲而言,采用三局二胜制有利,还是采用五局三胜制有利. 设各局胜负相互独立.

解 采用三局二胜制,甲最终获胜,其胜局的情况是:“甲甲”或“乙甲甲”或“甲乙甲”. 而这三种结局互不相容,于是由独立性得甲最终获胜的概率为

$$p_1 = p^2 + 2p^2(1-p).$$

采用五局三胜制,甲最终获胜,至少需比赛 3 局(可能赛 3 局,也可能赛 4 局或 5 局),且最后一局必须是甲胜,而前面甲需胜二局. 例如,共赛 4 局,则甲的胜局情况是:“甲乙甲甲”,“乙甲甲甲”,“甲甲乙甲”,且这三种结局互不相容. 由独立性得在五局三胜制下甲最终获胜的概率为

$$p_2 = p^3 + \binom{3}{2}p^3(1-p) + \binom{4}{2}p^3(1-p)^2,$$

而
$$p_2 - p_1 = p^2(6p^3 - 15p^2 + 12p - 3)$$

$$= 3p^2(p-1)^2(2p-1).$$

当 $p > \frac{1}{2}$ 时 $p_2 > p_1$; 当 $p = \frac{1}{2}$ 时 $p_2 = p_1 = \frac{1}{2}$. 故当 $p > \frac{1}{2}$ 时,对甲来说采用五局三胜制为有利. 当 $p = \frac{1}{2}$ 时两种赛制甲、乙最终获胜的概率是相同的,都是 50%. □

小结

随机试验的全部可能结果组成的集合 S 称为样本空间. 样本空间 S 的子集称为事件,当且仅当这一子集中的一个样本点出现时,称这一事件发生. 事件是一个集合,因而事件间的关系与事件的运算自然按照集合论中集合之间的关系和集合的运算来处理. 集合间的关

系和集合的运算,读者是熟悉的,重要的是要知道它们在概率论中的含义。

在一次试验中,一个事件(除必然事件与不可能事件外)可能发生也可能不发生,其发生的可能性的的大小是客观存在的.事件发生的频率以及它的稳定性,表明能用一个数来表征事件在一次试验中发生的可能性的的大小.我们从频率的稳定性及频率的性质得到启发和抽象,给出了概率的定义.我们定义了一个集合(事件)的函数 $P(\cdot)$,它满足三条基本性质:1°非负性,2°规范性,3°可列可加性.这一函数的函数值 $P(A)$ 就定义为事件 A 的概率.

概率的定义只给出概率必须满足的三条基本性质,并未对事件 A 的概率 $P(A)$ 给定一个具体的数.只在古典概型的情况,对于每个事件 A 给出了概率 $P(A) = k/n$ ((4.1)式).一般,我们可以进行大量的重复试验,得到事件 A 的频率,而以频率作为 $P(A)$ 的近似值.或者根据概率的性质分析,得到 $P(A)$ 的取值.

在古典概型中,我们证明了条件概率的公式

$$P(B|A) = \frac{P(AB)}{P(A)}, \quad P(A) > 0. \quad (5.2)$$

在一般的情况,(5.2)式则作为条件概率的定义.固定 A ,条件概率 $P(\cdot|A)$ 具有概率定义中的三条基本性质,因而条件概率是一种概率.

有两种计算条件概率 $P(B|A)$ 的方法:(1)按条件概率的含义,直接求出 $P(B|A)$.注意到,在求 $P(B|A)$ 时已知事件 A 已发生,样本空间 S 中所有不属于 A 的样本点都被排除,原有的样本空间 S 缩减成为 $S' = A$.在缩减了的样本空间 $S' = A$ 中计算事件 B 的概率就得到 $P(B|A)$.(2)在 S 中计算 $P(AB)$ 及 $P(A)$,再按(5.2)式求得 $P(B|A)$.

将(5.2)式写成

$$P(AB) = P(B|A)P(A), \quad P(A) > 0. \quad (5.3)$$

这就是乘法公式.我们常按上述第一种方法求出条件概率,从而按(5.3)可求得 $P(AB)$.

事件的独立性是概率论中的一个非常重要的概念.概率论与数理统计中的很多很多内容都是在独立的前提下讨论的.应该注意到,在实际应用中,对于事件的独立性,我们往往不是根据定义来验证而是根据实际意义来加以判断的.根据实际背景判断事件的独立性,往往并不困难.

■ 重要术语及主题

下面列出了本章的重要术语及主题,请读者写出它们的定义或内容,然后与教材中的陈述校核,看看你是否写对了.这样做旨在使读者在复习时收到较好的效果.

随机试验 样本空间 随机事件 基本事件 频率 概率 古典概型 A 的对立事件 $\bar{A}$ 及其概率 两个互不相容事件的和事件的概率 概率的加法定理 条件概率 概率的乘法公式 全概率公式 贝叶斯公式 事件的独立性 实际推断原理

习题

1. 写出下列随机试验的样本空间 S :

(1) 记录一个班一次数学考试的平均分数(设以百分制记分).

(2) 生产产品直到有 10 件正品为止,记录生产产品的总件数.

(3) 对某工厂出厂的产品进行检查,合格的记上“正品”,不合格的记上“次品”,如连续查

出了 2 件次品就停止检查,或检查了 4 件产品就停止检查,记录检查的结果.

(4) 在单位圆内任意取一点,记录它的坐标.

2. 设 A, B, C 为三个事件,用 A, B, C 的运算关系表示下列各事件:

(1) A 发生, B 与 C 不发生.

(2) A 与 B 都发生,而 C 不发生.

(3) A, B, C 中至少有一个发生.

(4) A, B, C 都发生.

(5) A, B, C 都不发生.

(6) A, B, C 中不多于一个发生.

(7) A, B, C 中不多于两个发生.

(8) A, B, C 中至少有两个发生.

3. (1) 设 A, B, C 是三个事件,且 $P(A) = P(B) = P(C) = 1/4$, $P(AB) = P(BC) = 0$, $P(AC) = 1/8$,求 A, B, C 至少有一个发生的概率.

(2) 已知 $P(A) = 1/2$, $P(B) = 1/3$, $P(C) = 1/5$, $P(AB) = 1/10$, $P(AC) = 1/15$, $P(BC) = 1/20$, $P(ABC) = 1/30$,求 $A \cup B$, $\bar{A}\bar{B}$, $A \cup B \cup C$, $\bar{A}\bar{B}\bar{C}$, $\bar{A}\bar{B}C$, $\bar{A}\bar{B} \cup C$ 的概率.

(3) 已知 $P(A) = 1/2$, (i) 若 A, B 互不相容,求 $P(\bar{A}\bar{B})$, (ii) 若 $P(AB) = 1/8$,求 $P(\bar{A}\bar{B})$.

4. 设 A, B 是两个事件.

(1) 已知 $\bar{A}\bar{B} = \bar{A}B$,验证 $A = B$.

(2) 验证事件 A 和事件 B 恰有一个发生的概率为 $P(A) + P(B) - 2P(AB)$.

5. 10 片药片中有 5 片是安慰剂.

(1) 从中任意抽取 5 片,求其中至少有 2 片是安慰剂的概率.

(2) 从中每次取一片,作不放回抽样,求前 3 次都取到安慰剂的概率.

6. 在房间里有 10 个人,分别佩戴从 1 号到 10 号的纪念章,任选 3 人记录其纪念章的号码.

(1) 求最小号码为 5 的概率.

(2) 求最大号码为 5 的概率.

7. 某油漆公司发出 17 桶油漆,其中白漆 10 桶、黑漆 4 桶、红漆 3 桶,在搬运中所有标签脱落,交货人随意将这些油漆发给顾客. 问一个订货为 4 桶白漆、3 桶黑漆和 2 桶红漆的顾客,能按所订颜色如数得到订货的概率是多少?

8. 在 1500 件产品中有 400 件次品、1100 件正品. 任取 200 件.

(1) 求恰有 90 件次品的概率.

(2) 求至少有 2 件次品的概率.

9. 从 5 双不同的鞋子中任取 4 只,问这 4 只鞋子中至少有两只配成一双的概率是多少?

10. 在 11 张卡片上分别写上 probability 这 11 个字母,从中任意连抽 7 张,求其排列结果为 ability 的概率.

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

12. 50 只铆钉随机地取来用在 10 个部件上,其中有 3 只铆钉强度太弱. 每个部件用 3 只铆钉. 若将 3 只强度太弱的铆钉都装在一个部件上,则这个部件强度就太弱. 问发生一个部件

强度太弱的概率是多少?

13. 一俱乐部有 5 名一年级学生, 2 名二年级学生, 3 名三年级学生, 2 名四年级学生.

(1) 在其中任选 4 名学生, 求一、二、三、四年级的学生各一名的概率.

(2) 在其中任选 5 名学生, 求一、二、三、四年级的学生均包含在内的概率.

14. (1) 已知 $P(\bar{A})=0.3$, $P(B)=0.4$, $P(A\bar{B})=0.5$, 求条件概率 $P(B|A\cup\bar{B})$.

(2) 已知 $P(A)=1/4$, $P(B|A)=1/3$, $P(A|B)=1/2$, 求 $P(A\cup B)$.

15. 掷两颗骰子, 已知两颗骰子点数之和为 7, 求其中有一颗为 1 点的概率 (用两种方法).

16. 据以往资料表明, 某一 3 口之家, 患某种传染病的概率有以下规律:

$$P\{\text{孩子得病}\}=0.6, P\{\text{母亲得病}|\text{孩子得病}\}=0.5,$$

$$P\{\text{父亲得病}|\text{母亲及孩子得病}\}=0.4,$$

求母亲及孩子得病但父亲未得病的概率.

17. 已知在 10 件产品中有 2 件次品, 在其中取两次, 每次任取一件, 作不放回抽样. 求下列事件的概率:

(1) 两件都是正品.

(2) 两件都是次品.

(3) 一件是正品, 一件是次品.

(4) 第二次取出的是次品.

18. 某人忘记了电话号码的最后一个数字, 因而他随意地拨号. 求他拨号不超过三次而接通所需电话的概率. 若已知最后一个数字是奇数, 那么此概率是多少?

19. (1) 设甲袋中装有 n 只白球、 m 只红球; 乙袋中装有 N 只白球、 M 只红球. 今从甲袋中任意取一只球放入乙袋中, 再从乙袋中任意取一只球. 问取到白球的概率是多少?

(2) 第一只盒子装有 5 只红球, 4 只白球; 第二只盒子装有 4 只红球, 5 只白球. 先从第一盒中任取 2 只球放入第二盒中去, 然后从第二盒中任取一只球. 求取到白球的概率.

20. 某种产品的商标为“MAXAM”, 其中有 2 个字母脱落, 有人捡起随意放回, 求放回后仍为“MAXAM”的概率.

21. 已知男子有 5% 是色盲患者, 女子有 0.25% 是色盲患者. 今从男女人数相等的人群中随机地挑选一人, 恰好是色盲者, 问此人是男性的概率是多少?

22. 一学生接连参加同一课程的两次考试. 第一次及格的概率为 p , 若第一次及格则第二次及格的概率也为 p ; 若第一次不及格则第二次及格的概率为 $p/2$.

(1) 若至少有一次及格则他能取得某种资格, 求他取得该资格的概率.

(2) 若已知他第二次已经及格, 求他第一次及格的概率.

23. 将两信息分别编码为 A 和 B 传送出去, 接收站收到时, A 被误收作 B 的概率为 0.02, 而 B 被误收作 A 的概率为 0.01. 信息 A 与信息 B 传送的频繁程度为 2:1. 若接收站收到的信息是 A , 问原发信息是 A 的概率是多少?

24. 有两箱同种类的零件, 第一箱装 50 只, 其中 10 只一等品; 第二箱装 30 只, 其中 18 只一等品. 今从两箱中任挑出一箱, 然后从该箱中取零件两次, 每次任取一只, 作不放回抽样. 求

(1) 第一次取到的零件是一等品的概率.

(2) 在第一次取到的零件是一等品的条件下,第二次取到的也是一等品的概率.

25. 某人下午 5:00 下班,他所积累的资料表明:

到家时间	5:35~5:39	5:40~5:44	5:45~5:49	5:50~5:54	迟于 5:54
乘地铁的概率	0.10	0.25	0.45	0.15	0.05
乘汽车的概率	0.30	0.35	0.20	0.10	0.05

某日他抛一枚硬币决定乘地铁还是乘汽车,结果他是 5:47 到家的. 试求他是乘地铁回家的概率.

26. 病树的主人外出,委托邻居浇水,设已知如果不浇水,树死去的概率为 0.8. 若浇水则树死去的概率为 0.15. 有 0.9 的把握确定邻居会记得浇水.

(1) 求主人回来树还活着的概率.

(2) 若主人回来树已死去,求邻居忘记浇水的概率.

27. 设本题涉及的事件均有意义. 设 A, B 都是事件.

(1) 已知 $P(A) > 0$, 证明 $P(AB|A) \geq P(AB|A \cup B)$.

(2) 若 $P(A|B) = 1$, 证明 $P(\bar{B}|\bar{A}) = 1$.

(3) 若设 C 也是事件, 且有 $P(A|C) \geq P(B|C)$, $P(A|\bar{C}) \geq P(B|\bar{C})$, 证明 $P(A) \geq P(B)$.

28. 有两种花籽, 发芽率分别为 0.8, 0.9, 从中各取一颗, 设各花籽是否发芽相互独立. 求

(1) 这两颗花籽都能发芽的概率.

(2) 至少有一颗能发芽的概率.

(3) 恰有一颗能发芽的概率.

29. 根据报道美国人血型的分布近似地为: A 型为 37%, O 型为 44%, B 型为 13%, AB 型为 6%. 夫妻拥有的血型是相互独立的.

(1) B 型的人只有输入 B、O 两种血型才安全. 若妻为 B 型, 夫为何种血型未知, 求夫是妻的安全输血者的概率.

(2) 随机地取一对夫妇, 求妻为 B 型夫为 A 型的概率.

(3) 随机地取一对夫妇, 求其中一人为 A 型, 另一人为 B 型的概率.

(4) 随机地取一对夫妇, 求其中至少有一人是 O 型的概率.

30. (1) 给出事件 A, B 的例子, 使得

(i) $P(A|B) < P(A)$, (ii) $P(A|B) = P(A)$, (iii) $P(A|B) > P(A)$.

(2) 设事件 A, B, C 相互独立, 证明 (i) C 与 AB 相互独立. (ii) C 与 $A \cup B$ 相互独立.

(3) 设事件 A 的概率 $P(A) = 0$, 证明对于任意另一事件 B , 有 A, B 相互独立.

(4) 证明事件 A, B 相互独立的充要条件是 $P(A|B) = P(A|\bar{B})$.

31. 设事件 A, B 的概率均大于零, 说明以下的叙述 (1) 必然对. (2) 必然错. (3) 可能对. 并说明理由.

(1) 若 A 与 B 互不相容, 则它们相互独立.

(2) 若 A 与 B 相互独立, 则它们互不相容.

(3) $P(A) = P(B) = 0.6$, 且 A, B 互不相容.

(4) $P(A) = P(B) = 0.6$, 且 A, B 相互独立.

32. 有一种检验艾滋病毒的检验法,其结果有概率 0.005 报导为假阳性(即不带艾滋病毒者,经此检验法有 0.005 的概率被认为带艾滋病毒).今有 140 名不带艾滋病毒的正常人全部接受此种检验,被报道至少有一人带艾滋病毒的概率为多少?

33. 盒中有编号为 1,2,3,4 的 4 只球,随机地自盒中取一只球,事件 A 为“取得的是 1 号或 2 号球”,事件 B 为“取得的是 1 号或 3 号球”,事件 C 为“取得的是 1 号或 4 号球”验证:

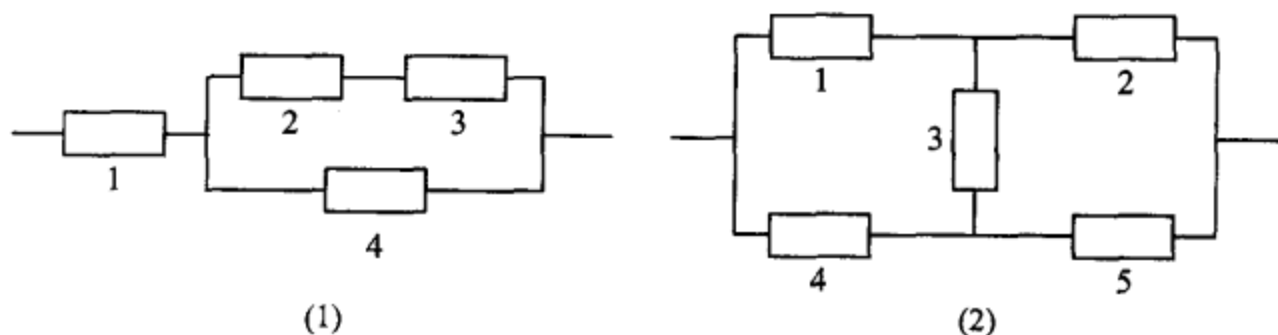
$$P(AB) = P(A)P(B), P(AC) = P(A)P(C), P(BC) = P(B)P(C),$$

但 $P(ABC) \neq P(A)P(B)P(C)$,

即事件 A, B, C 两两独立,但 A, B, C 不是相互独立的.

34. 试分别求以下两个系统的可靠性:

(1) 设有 4 个独立工作的元件 1,2,3,4. 它们的可靠性分别为 p_1, p_2, p_3, p_4 , 将它们按题 34 图(1)的方式连接(称为并串联系统).



题 34 图

(2) 设有 5 个独立工作的元件 1,2,3,4,5. 它们的可靠性均为 p , 将它们按题 34 图(2)的方式连接(称为桥式系统).

35. 如果一危险情况 C 发生时,一电路闭合并发出警报,我们可以借用两个或多个开关并联以改善可靠性. 在 C 发生时这些开关每一个都应闭合,且若至少一个开关闭合了,警报就发出. 如果两个这样的开关并联连接,它们每个具有 0.96 的可靠性(即在情况 C 发生时闭合的概率),问这时系统的可靠性(即电路闭合的概率)是多少? 如果需要有一个可靠性至少为 0.9999 的系统,则至少需要用多少只开关并联? 设各开关闭合与否是相互独立的.

36. 三人独立地去破译一份密码,已知各人能译出的概率分别为 $1/5, 1/3, 1/4$. 问三人中至少有一人能将此密码译出的概率是多少?

37. 设第一只盒子中装有 3 只蓝球,2 只绿球,2 只白球;第二只盒子中装有 2 只蓝球,3 只绿球,4 只白球. 独立地分别在两只盒子中各取一只球.

(1) 求至少有一只蓝球的概率.

(2) 求有一只蓝球一只白球的概率.

(3) 已知至少有一只蓝球,求有一只蓝球一只白球的概率.

38. 袋中装有 m 枚正品硬币、 n 枚次品硬币(次品硬币的两面均印有国徽),在袋中任取一枚,将它投掷 r 次,已知每次都得到国徽. 问这枚硬币是正品的概率为多少?

39. 设根据以往记录的数据分析,某船只运输的某种物品损坏的情况共有三种:损坏 2% (这一事件记为 A_1),损坏 10% (事件 A_2),损坏 90% (事件 A_3),且知 $P(A_1) = 0.8, P(A_2) =$

0.15, $P(A_3) = 0.05$. 现在从已被运输的物品中随机地取 3 件, 发现这 3 件都是好的 (这一事件记为 B). 试求 $P(A_1 | B)$, $P(A_2 | B)$, $P(A_3 | B)$ (这里设物品件数很多, 取出一件后不影响取后一件是否为好品的概率).

40. 将 A、B、C 三个字母之一输入信道, 输出为原字母的概率为 α , 而输出为其他一字母的概率都是 $(1-\alpha)/2$. 今将字母串 AAAA, BBBB, CCCC 之一输入信道, 输入 AAAA, BBBB, CCCC 的概率分别为 p_1, p_2, p_3 ($p_1 + p_2 + p_3 = 1$), 已知输出为 ABCA, 问输入的是 AAAA 的概率是多少? (设信道传输各个字母的工作是相互独立的.)



第二章 随机变量及其分布

§ 1 随机变量

在第一章我们看到一些随机试验,它们的结果可以用数来表示.此时样本空间 S 的元素是一个数,如 S_3, S_5 ;但有些则不然,如 S_1, S_2 . 当样本空间 S 的元素不是一个数时,人们对于 S 就难以描述和研究. 现在来讨论如何引入一个法则,将随机试验的每一个结果,即将 S 的每个元素 e 与实数 x 对应起来. 从而引入了随机变量的概念. 我们从例题开始讨论.

例 1 在第一章 § 4 例 1 中,将一枚硬币抛掷三次,观察出现正面和反面的情况,样本空间是

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

以 X 记三次投掷得到正面 H 的总数,那么,对于样本空间 $S = \{e\}$ ^① 中的每一个样本点 e , X 都有一个数与之对应. X 是定义在样本空间 S 上的一个实值单值函数. 它的定义域是样本空间 S , 值域是实数集合 $\{0, 1, 2, 3\}$. 使用函数记号可将 X 写成

$$X = X(e) = \begin{cases} 3, & e = HHH, \\ 2, & e = HHT, HTH, THH, \\ 1, & e = HTT, THT, TTH, \\ 0, & e = TTT. \end{cases}$$

□

例 2 在一袋中装有编号分别为 1, 2, 3 的 3 只球,在袋中任取一只球,放回,再任取一只球,记录它们的号码,试验的样本空间为 $S = \{e\} = \{(i, j) | i, j = 1, 2, 3\}$, i, j 分别为第 1, 第 2 次取到的球的号码. 以 X 记两球号码之和. 我们看到,对于试验的每一个结果 $e = (i, j) \in S$, X 都有一个指定的值 $i + j$ 与之对应. (如图 2-1). X 是定义在样本空间 S 上的单值实值函数. 它的定义域是样本空间 S . 值域是实数集合 $\{2, 3, 4, 5\}$,

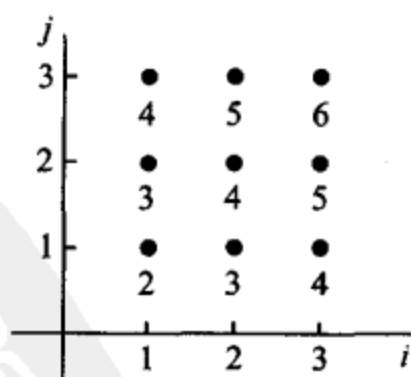


图 2-1

① 我们用 e 代表样本空间的元素,而将样本空间记成 $\{e\}$.

6}. X 可写成

$$X = X(e) = X((i, j)) = i + j, \quad i, j = 1, 2, 3.$$

□

一般有以下定义.

定义 设随机试验的样本空间为 $S = \{e\}$. $X = X(e)$ 是定义在样本空间 S 上的实值单值函数. 称 $X = X(e)$ 为随机变量^①.

图 2-2 画出了样本点 e 与实数 $X = X(e)$ 对应的示意图.

有许多随机试验, 它们的结果本身是一个数, 即样本点 e 本身是一个数. 我们令 $X = X(e) = e$, 那么 X 就是一个随机变量. 例如, 用 Y 记某车间一天的缺勤人数, 以 W 记某地区第一季度的降雨量, 以 Z 记某工厂一天的耗电量, 以 N 记某医院某一天的挂号人数. 那么 Y, W, Z, N 都是随机变量.

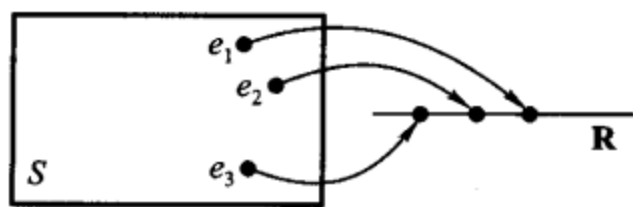


图 2-2

本书中, 我们一般以大写的字母如 $X, Y, Z, W, \dots$ 表示随机变量, 而以小写字母 $x, y, z, w, \dots$ 表示实数.

随机变量的取值随试验的结果而定, 而试验的各个结果出现有一定的概率, 因而随机变量的取值有一定的概率. 例如, 在例 1 中 X 取值为 2, 记成 $\{X=2\}$, 对应于样本点的集合 $A = \{HHT, HTH, THH\}$, 这是一个事件, 当且仅当事件 A 发生时有 $\{X=2\}$. 我们称概率 $P(A) = P\{HHT, HTH, THH\}$ 为 $\{X=2\}$ 的概率, 即 $P\{X=2\} = P(A) = 3/8$. 以后, 还将事件 $A = \{HHT, HTH, THH\}$ 说成是事件 $\{X=2\}$. 类似地有

$$P\{X \leq 1\} = P\{HTT, THT, TTH, TTT\} = \frac{1}{2}.$$

一般, 若 L 是一个实数集合, 将 X 在 L 上取值写成 $\{X \in L\}$. 它表示事件 $B = \{e | X(e) \in L\}$, 即 B 是由 S 中使得 $X(e) \in L$ 的所有样本点 e 所组成的事件, 此时有

$$P\{X \in L\} = P(B) = P\{e | X(e) \in L\}.$$

随机变量的取值随试验的结果而定, 在试验之前不能预知它取什么值, 且它的取值有一定的概率. 这些性质显示了随机变量与普通函数有着本质的差异.

随机变量的引入, 使我们能用随机变量来描述各种随机现象, 并能利用数学分析的方法对随机试验的结果进行深入广泛的研究和讨论.

^① 严格地说“对于任意实数 x , 集合 $\{e | X(e) \leq x\}$ (即: 使得 $X(e) \leq x$ 的所有样本点 e 所组成的集合) 有确定的概率”这一要求应包括在随机变量的定义之中, 一般来说, 不满足这一条件的情况, 在实际应用中是很少遇到的. 因此, 我们在定义中未提及这一要求.

§ 2 离散型随机变量及其分布律

有些随机变量,它全部可能取到的值是有限个或可列无限多个,这种随机变量称为离散型随机变量.例如 § 1 例 1 中的随机变量 X ,它只可能取 $0, 1, 2, 3$ 四个值,它是一个离散型随机变量.又如某城市的 120 急救电话台一昼夜收到的呼唤次数也是离散型随机变量.若以 T 记某元件的寿命,它所可能取的值充满一个区间,是无法按一定次序一一列举出来的,因而它是一个非离散型的随机变量.本节只讨论离散型随机变量.

容易知道,要掌握一个离散型随机变量 X 的统计规律,必须且只需知道 X 的所有可能取值以及取每一个可能值的概率.

设离散型随机变量 X 所有可能取的值为 $x_k (k=1, 2, \dots)$, X 取各个可能值的概率,即事件 $\{X=x_k\}$ 的概率,为

$$P\{X=x_k\}=p_k, k=1, 2, \dots. \quad (2.1)$$

由概率的定义, p_k 满足如下两个条件:

$$1^\circ \quad p_k \geq 0, k=1, 2, \dots; \quad (2.2)$$

$$2^\circ \quad \sum_{k=1}^{\infty} p_k = 1. \quad (2.3)$$

2° 是由于 $\{X=x_1\} \cup \{X=x_2\} \cup \dots$ 是必然事件,且 $\{X=x_j\} \cap \{X=x_k\} = \emptyset, k \neq j$, 故 $1 = P\left[\bigcup_{k=1}^{\infty} \{X=x_k\}\right] = \sum_{k=1}^{\infty} P\{X=x_k\}$, 即 $\sum_{k=1}^{\infty} p_k = 1$.

我们称 (2.1) 式为离散型随机变量 X 的分布律. 分布律也可以用表格的形式来表示

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

(2.4)

(2.4) 直观地表示了随机变量 X 取各个值的概率的规律. X 取各个值各占一些概率,这些概率合起来是 1. 可以想像成: 概率 1 以一定的规律分布在各个可能值上. 这就是 (2.4) 称为分布律的缘故.

例 1 设一汽车在开往目的地的道路上需经过四组信号灯, 每组信号灯以 $1/2$ 的概率允许或禁止汽车通过. 以 X 表示汽车首次停下时, 它已通过的信号灯的组数 (设各组信号灯的工作是相互独立的), 求 X 的分布律.

解 以 p 表示每组信号灯禁止汽车通过的概率, 易知 X 的分布律为

X	0	1	2	3	4
p_k	p	$(1-p)p$	$(1-p)^2 p$	$(1-p)^3 p$	$(1-p)^4$

或写成

$$P\{X=k\}=(1-p)^k p, k=0,1,2,3, P\{X=4\}=(1-p)^4.$$

以 $p=1/2$ 代入得

X	0	1	2	3	4
p_k	0.5	0.25	0.125	0.0625	0.0625

□

下面介绍三种重要的离散型随机变量.

(一) (0-1)分布

设随机变量 X 只可能取 0 与 1 两个值, 它的分布律是

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

则称 X 服从以 p 为参数的 (0-1) 分布或两点分布.

(0-1) 分布的分布律也可写成

X	0	1
p_k	$1-p$	p

对于一个随机试验, 如果它的样本空间只包含两个元素, 即 $S=\{e_1, e_2\}$, 我们总能在 S 上定义一个服从 (0-1) 分布的随机变量

$$X=X(e)=\begin{cases} 0, & \text{当 } e=e_1, \\ 1, & \text{当 } e=e_2 \end{cases}$$

来描述这个随机试验的结果. 例如, 对新生婴儿的性别进行登记, 检查产品的质量是否合格, 某车间的电力消耗是否超过负荷以及前面多次讨论过的“抛硬币”试验等都可以用 (0-1) 分布的随机变量来描述. (0-1) 分布是经常遇到的一种分布.

(二) 伯努利试验、二项分布

设试验 E 只有两个可能结果: A 及 $\bar{A}$, 则称 E 为伯努利 (Bernoulli) 试验. 设 $P(A)=p$ ($0<p<1$), 此时 $P(\bar{A})=1-p$. 将 E 独立重复地进行 n 次, 则称这一串重复的独立试验为 n 重伯努利试验.

这里“重复”是指在每次试验中 $P(A)=p$ 保持不变; “独立”是指各次试验

的结果互不影响,即若以 C_i 记第 i 次试验的结果, C_i 为 A 或 $\bar{A}$, $i=1,2,\dots,n$. “独立”是指

$$P(C_1 C_2 \cdots C_n) = P(C_1)P(C_2) \cdots P(C_n). \quad (2.5)$$

n 重伯努利试验是一种很重要的数学模型,它有广泛的应用,是研究最多的模型之一.

例如, E 是抛一枚硬币观察得到正面或反面. A 表示得正面,这是一个伯努利试验. 如将硬币抛 n 次,就是 n 重伯努利试验. 又如抛一颗骰子,若 A 表示得到“1 点”, $\bar{A}$ 表示得到“非 1 点”. 将骰子抛 n 次,就是 n 重伯努利试验. 再如在袋中装有 a 只白球, b 只黑球. 试验 E 是在袋中任取一只球,观察其颜色. 以 A 表示“取到白球”, $P(A) = a/(a+b)$. 若连续取球 n 次作放回抽样,这就是 n 重伯努利试验. 然而,若作不放回抽样,虽则每次试验都有 $P(A) = a/(a+b)$ (见第一章 §4 例 5),但各次试验不再相互独立^①,因而不再是 n 重伯努利试验了.

以 X 表示 n 重伯努利试验中事件 A 发生的次数, X 是一个随机变量,我们来求它的分布律. X 所有可能取的值为 $0, 1, 2, \dots, n$. 由于各次试验是相互独立的,因此事件 A 在指定的 k ($0 \leq k \leq n$) 次试验中发生,在其他 $n-k$ 次试验中 A 不发生 (例如在前 k 次试验中 A 发生,而后 $n-k$ 次试验中 A 不发生) 的概率为

$$\underbrace{p \cdot p \cdot \cdots \cdot p}_{k \text{ 个}} \cdot \underbrace{(1-p) \cdot (1-p) \cdot \cdots \cdot (1-p)}_{n-k \text{ 个}} = p^k (1-p)^{n-k}.$$

这种指定的方式共有 $\binom{n}{k}$ 种,它们是两两互不相容的,故在 n 次试验中 A 发生 k 次

的概率为 $\binom{n}{k} p^k (1-p)^{n-k}$, 记 $q = 1-p$, 即有

$$P\{X = k\} = \binom{n}{k} p^k q^{n-k}, k = 0, 1, 2, \dots, n. \quad (2.6)$$

显然

$$P\{X = k\} \geq 0, k = 0, 1, 2, \dots, n;$$

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

^① 对于不放回抽样,以 A_1, A_2 分别记第一次、第二次取到白球,则有 $P(A_2 | A_1) = \frac{a-1}{a+b-1}$, 而 $P(A_2) = \frac{a}{a+b}$, $P(A_2 | A_1) \neq P(A_2)$, 故第一次、第二次试验不相互独立. 即知 (2.5) 不成立.

即 $P\{X=k\}$ 满足条件 (2.2), (2.3). 注意到 $\binom{n}{k} p^k q^{n-k}$ 刚好是二项式 $(p+q)^n$ 的展开式中出现 p^k 的那一项, 我们称随机变量 X 服从参数为 n, p 的二项分布, 并记为 $X \sim b(n, p)$.

特别, 当 $n=1$ 时二项分布 (2.6) 化为

$$P\{X=k\} = p^k q^{1-k}, \quad k=0, 1.$$

这就是 (0-1) 分布.

例 2 按规定, 某种型号电子元件的使用寿命超过 1 500 小时的为一级品. 已知某一大批产品的一级品率为 0.2, 现在从中随机地抽查 20 只. 问 20 只元件中恰有 k 只 ($k=0, 1, \dots, 20$) 为一级品的概率是多少?

解 这是不放回抽样. 但由于这批元件的总数很大, 且抽查的元件的数量相对于元件的总数来说又很小, 因而可以当作放回抽样来处理, 这样做会有一些误差, 但误差不大. 我们将检查一只元件看它是否为一级品看成是一次试验, 检查 20 只元件相当于做 20 重伯努利试验. 以 X 记 20 只元件中一级品的只数, 那么, X 是一个随机变量, 且有 $X \sim b(20, 0.2)$. 由 (2.6) 式即得所求概率为

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

将计算结果列表如下:

$P\{X=0\}=0.012$	$P\{X=4\}=0.218$	$P\{X=8\}=0.022$
$P\{X=1\}=0.058$	$P\{X=5\}=0.175$	$P\{X=9\}=0.007$
$P\{X=2\}=0.137$	$P\{X=6\}=0.109$	$P\{X=10\}=0.002$
$P\{X=3\}=0.205$	$P\{X=7\}=0.055$	
当 $k \geq 11$ 时, $P\{X=k\} < 0.001$		

为了对本题的结果有一个直观了解, 我们作出上表的图形, 如图 2-3 所示.

从图 2-3 中看到, 当 k 增加时, 概率 $P\{X=k\}$ 先是随之增加, 直至达到最大值 (本例中当 $k=4$ 时取到最大值), 随后单调减少. 我们指出, 一般, 对于固定的 n 及 p , 二项分布 $b(n, p)$ 都具有这一性质. $\square$

例 3 某人进行射击, 设每次射击的命中率为 0.02, 独立射击 400 次, 试

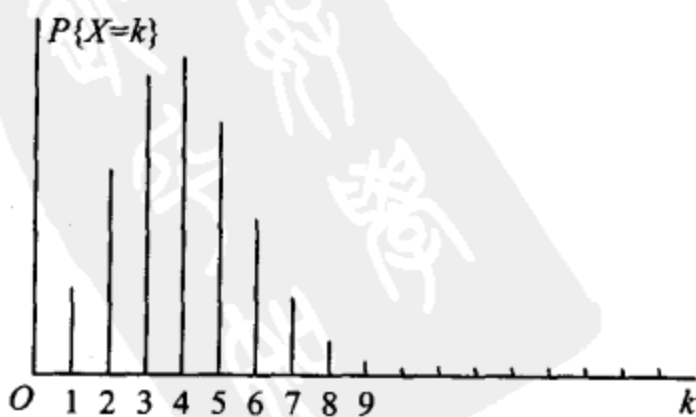


图 2-3

求至少击中两次的概率.

解 将一次射击看成是一次试验. 设击中的次数为 X , 则 $X \sim b(400, 0.02)$. X 的分布律为

$$P\{X = k\} = \binom{400}{k} (0.02)^k (0.98)^{400-k}, k = 0, 1, \dots, 400.$$

于是所求概率为

$$\begin{aligned} P\{X \geq 2\} &= 1 - P\{X = 0\} - P\{X = 1\} \\ &= 1 - (0.98)^{400} - 400(0.02)(0.98)^{399} = 0.9972. \end{aligned}$$

□

这个概率很接近于 1. 我们从两方面来讨论这一结果的实际意义. 其一, 虽然每次射击的命中率很小 (为 0.02), 但如果射击 400 次, 则击中目标至少两次是几乎可以肯定的. 这一事实说明, 一个事件尽管在一次试验中发生的概率很小, 但只要试验次数很多, 而且试验是独立地进行的, 那么这一事件的发生几乎是肯定的. 这也告诉人们绝不能轻视小概率事件. 其二, 如果射手在 400 次射击中, 击中目标的次数竟不到两次, 由于概率 $P\{X < 2\} \approx 0.003$ 很小, 根据实际推断原理, 我们将怀疑“每次射击的命中率为 0.02”这一假设, 即认为该射手射击的命中率达不到 0.02.

例 4 设有 80 台同类型设备, 各台工作是相互独立的, 发生故障的概率都是 0.01, 且一台设备的故障能由一个人处理. 考虑两种配备维修工人的方法, 其一是由 4 人维护, 每人负责 20 台; 其二是由 3 人共同维护 80 台. 试比较这两种方法在设备发生故障时不能及时维修的概率的大小.

解 按第一种方法. 以 X 记“第 1 人维护的 20 台中同一时刻发生故障的台数”, 以 $A_i (i=1, 2, 3, 4)$ 表示事件“第 i 人维护的 20 台中发生故障不能及时维修”, 则知 80 台中发生故障而不能及时维修的概率为

$$P(A_1 \cup A_2 \cup A_3 \cup A_4) \geq P(A_1) = P\{X \geq 2\}.$$

而 $X \sim b(20, 0.01)$, 故有

$$\begin{aligned} P\{X \geq 2\} &= 1 - \sum_{k=0}^1 P\{X = k\} \\ &= 1 - \sum_{k=0}^1 \binom{20}{k} (0.01)^k (0.99)^{20-k} = 0.0169. \end{aligned}$$

即有

$$P(A_1 \cup A_2 \cup A_3 \cup A_4) \geq 0.0169.$$

按第二种方法. 以 Y 记 80 台中同一时刻发生故障的台数. 此时, $Y \sim b(80, 0.01)$, 故 80 台中发生故障而不能及时维修的概率为

$$P\{Y \geq 4\} = 1 - \sum_{k=0}^3 \binom{80}{k} (0.01)^k (0.99)^{80-k} = 0.0087.$$

我们发现, 在后一种情况尽管任务重了 (每人平均维护约 27 台), 但工作效

率不仅没有降低,反而提高了. □

(三) 泊松分布

设随机变量 X 所有可能取的值为 $0, 1, 2, \dots$, 而取各个值的概率为

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

其中 $\lambda > 0$ 是常数. 则称 X 服从参数为 λ 的泊松分布, 记为 $X \sim \pi(\lambda)$.

易知, $P\{X = k\} \geq 0, k = 0, 1, 2, \dots$, 且有

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

即 $P\{X = k\}$ 满足条件 (2.2), (2.3).

参数 λ 的意义将在第四章说明, 有关服从泊松分布的随机变量的数学模型将在第十二章中讨论.

具有泊松分布的随机变量在实际应用中是很多的. 例如, 一本书一页中的印刷错误数、某地区在一天内邮递遗失的信件数、某一医院在一天内的急诊病人数、某一地区一个时间间隔内发生交通事故的次数、在一个时间间隔内某种放射性物质发出的、经过计数器的 α 粒子数等都服从泊松分布. 泊松分布也是概率论中的一种重要分布.

下面介绍一个用泊松分布来逼近二项分布的定理.

泊松定理 设 $\lambda > 0$ 是一个常数, n 是任意正整数, 设 $np_n = \lambda$, 则对于任一固定的非负整数 k , 有

$$\lim_{n \rightarrow \infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}.$$

证 由 $p_n = \frac{\lambda}{n}$, 有

$$\begin{aligned} \binom{n}{k} p_n^k (1-p_n)^{n-k} &= \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \left[1 \cdot \left(1-\frac{1}{n}\right) \cdots \left(1-\frac{k-1}{n}\right)\right] \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k}. \end{aligned}$$

对于任意固定的 k , 当 $n \rightarrow \infty$ 时

$$1 \cdot \left(1-\frac{1}{n}\right) \cdots \left(1-\frac{k-1}{n}\right) \rightarrow 1, \quad \left(1-\frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda}, \quad \left(1-\frac{\lambda}{n}\right)^{-k} \rightarrow 1.$$

故有 $\lim_{n \rightarrow \infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}.$ □

定理的条件 $np_n = \lambda$ (常数) 意味着当 n 很大时 p_n 必定很小, 因此, 上述定理表明当 n 很大, p 很小 ($np = \lambda$) 时有以下近似式

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{\lambda^k e^{-\lambda}}{k!} \quad (\text{其中 } \lambda = np). \quad (2.7)$$

也就是说以 n, p 为参数的二项分布的概率值可以由参数为 $\lambda = np$ 的泊松分布的概率值近似. 上式也能用来作二项分布概率的近似计算.

例 5 计算机硬件公司制造某种特殊型号的微型芯片, 次品率达 0.1%, 各芯片成为次品相互独立. 求在 1000 只产品中至少有 2 只次品的概率. 以 X 记产品中的次品数, $X \sim b(1000, 0.001)$.

解 所求概率为

$$\begin{aligned} P\{X \geq 2\} &= 1 - P\{X=0\} - P\{X=1\} \\ &= 1 - (0.999)^{1000} - \binom{1000}{1} (0.999)^{999} (0.001) \\ &\approx 1 - 0.3676954 - 0.3680635 \approx 0.2642411. \end{aligned}$$

利用(2.7)式来计算得, $\lambda = 1000 \times 0.001 = 1$,

$$\begin{aligned} P\{X \geq 2\} &= 1 - P\{X=0\} - P\{X=1\} \\ &= 1 - e^{-1} - e^{-1} \approx 0.2642411. \end{aligned}$$

□

显然利用(2.7)式的计算来得方便. 一般, 当 $n \geq 20, p \leq 0.05$ 时用 $\frac{\lambda^k e^{-\lambda}}{k!}$ ($\lambda = np$) 作为 $\binom{n}{k} p^k (1-p)^{n-k}$ 的近似值效果颇佳.

§ 3 随机变量的分布函数

对于非离散型随机变量 X , 由于其可能取的值不能一一列举出来, 因而就不能像离散型随机变量那样可以用分布律来描述它. 另外, 我们通常所遇到的非离散型随机变量取任一指定的实数值的概率都等于 0 (这一点在下一节将会讲到). 再者, 在实际中, 对于这样的随机变量, 例如误差 ϵ , 元件的寿命 T 等, 我们并不会对误差 $\epsilon = 0.05 \text{ mm}$, 寿命 $T = 1251.3 \text{ h}$ 的概率感兴趣, 而是考虑误差落在某个区间内的概率, 寿命 T 大于某个数的概率. 因而我们转而去研究随机变量所取的值落在一个区间 $(x_1, x_2]$ 的概率: $P\{x_1 < X \leq x_2\}$. 但由于

$$P\{x_1 < X \leq x_2\} = P\{X \leq x_2\} - P\{X \leq x_1\},$$

所以我们只需知道 $P\{X \leq x_2\}$ 和 $P\{X \leq x_1\}$ 就可以了. 下面引入随机变量的分布函数的概念^①.

^① 虽然对于离散型随机变量, 我们可以用分布律全面地描述它, 但为了从数学上能统一地对随机变量进行研究, 在这里, 我们对离散型随机变量和非离散型随机变量统一地定义了分布函数.

定义 设 X 是一个随机变量, x 是任意实数, 函数

$$F(x) = P\{X \leq x\}, \quad -\infty < x < \infty$$

称为 X 的分布函数.

对于任意实数 $x_1, x_2 (x_1 < x_2)$, 有

$$\begin{aligned} P\{x_1 < X \leq x_2\} &= P\{X \leq x_2\} - P\{X \leq x_1\} \\ &= F(x_2) - F(x_1), \end{aligned} \quad (3.1)$$

因此, 若已知 X 的分布函数, 我们就知道 X 落在任一区间 $(x_1, x_2]$ 上的概率, 从这个意义上说, 分布函数完整地描述了随机变量的统计规律性.

分布函数是一个普通的函数, 正是通过它, 我们将能用数学分析的方法来研究随机变量.

如果将 X 看成是数轴上的随机点的坐标, 那么, 分布函数 $F(x)$ 在 x 处的函数值就表示 X 落在区间 $(-\infty, x]$ 上的概率.

分布函数 $F(x)$ 具有以下的基本性质:

1° $F(x)$ 是一个不减函数.

事实上, 由 (3.1) 式对于任意实数 $x_1, x_2 (x_1 < x_2)$, 有

$$F(x_2) - F(x_1) = P\{x_1 < X \leq x_2\} \geq 0.$$

2° $0 \leq F(x) \leq 1$, 且

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

$$F(\infty) = \lim_{x \rightarrow \infty} F(x) = 1.$$

上面两个式子, 我们只从几何上加以说明. 在图 2-4 中, 将区间端点 x 沿数轴无限向左移动 (即 $x \rightarrow -\infty$), 则

“随机点 X 落在点 x 左边”这一事件趋于不可能事件, 从而其概率趋于 0, 即有 $F(-\infty) = 0$; 又若将点 x 无限右移

(即 $x \rightarrow \infty$), 则“随机点 X 落在点 x 左边”这一事件趋于必然事件, 从而其概率趋于 1, 即有 $F(\infty) = 1$.

3° $F(x+0) = F(x)$, 即 $F(x)$ 是右连续的. (证略)

例 1 设随机变量 X 的分布律为

X	-1	2	3
p_k	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

求 X 的分布函数, 并求 $P\{X \leq \frac{1}{2}\}, P\{\frac{3}{2} < X \leq \frac{5}{2}\}, P\{2 \leq X \leq 3\}$.

解 X 仅在 $x = -1, 2, 3$ 三点处其概率 $\neq 0$, 而 $F(x)$ 的值是 $X \leq x$ 的累积概

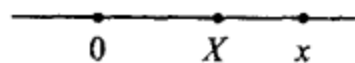


图 2-4

率值,由概率的有限可加性,知它即为小于或等于 x 的那些 x_k 处的概率 p_k 之和,有

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

即

$$F(x) = \begin{cases} 0, & x < -1, \\ \frac{1}{4}, & -1 \leq x < 2, \\ \frac{3}{4}, & 2 \leq x < 3, \\ 1, & x \geq 3. \end{cases}$$

$F(x)$ 的图形如图 2-5 所示,它是一条阶梯形的曲线,在 $x = -1, 2, 3$ 处有跳跃点,跳跃值分别为 $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$. 又

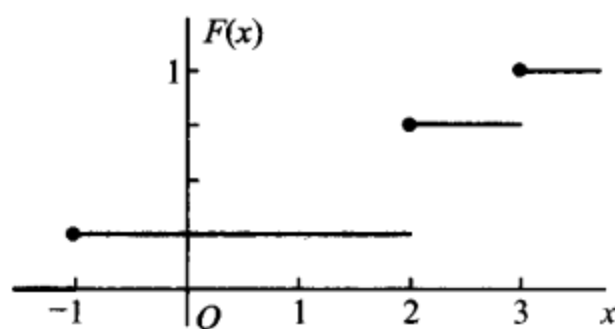


图 2-5

$$P\left\{X \leq \frac{1}{2}\right\} = F\left(\frac{1}{2}\right) = \frac{1}{4},$$

$$P\left\{\frac{3}{2} < X \leq \frac{5}{2}\right\} = F\left(\frac{5}{2}\right) - F\left(\frac{3}{2}\right) = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}.$$

$$\begin{aligned} P\{2 \leq X \leq 3\} &= F(3) - F(2) + P\{X = 2\} \\ &= 1 - \frac{3}{4} + \frac{1}{2} = \frac{3}{4}. \end{aligned}$$

□

一般,设离散型随机变量 X 的分布律为

$$P\{X = x_k\} = p_k, \quad k = 1, 2, \dots.$$

由概率的可列可加性得 X 的分布函数为

$$F(x) = P\{X \leq x\} = \sum_{x_k \leq x} P\{X = x_k\},$$

即

$$F(x) = \sum_{x_k \leq x} p_k, \quad (3.2)$$

这里和式是对于所有满足 $x_k \leq x$ 的 k 求和的. 分布函数 $F(x)$ 在 $x = x_k (k = 1, 2, \dots)$ 处有跳跃, 其跳跃值为 $p_k = P\{X = x_k\}$.

例 2 一个靶子是半径为 2 m 的圆盘, 设击中靶上任一同心圆盘上的点的概率与该圆盘的面积成正比, 并设射击都能中靶, 以 X 表示弹着点与圆心的距离. 试求随机变量 X 的分布函数.

解 若 $x < 0$, 则 $\{X \leq x\}$ 是不可能事件, 于是

$$F(x) = P\{X \leq x\} = 0.$$

若 $0 \leq x \leq 2$, 由题意, $P\{0 \leq X \leq x\} = kx^2$, k 是某一常数, 为了确定 k 的值, 取 $x = 2$, 有 $P\{0 \leq X \leq 2\} = 2^2 k$, 但已知 $P\{0 \leq X \leq 2\} = 1$, 故得 $k = 1/4$, 即

$$P\{0 \leq X \leq x\} = \frac{x^2}{4}.$$

于是

$$F(x) = P\{X \leq x\} = P\{X < 0\} + P\{0 \leq X \leq x\} = \frac{x^2}{4}.$$

若 $x \geq 2$, 由题意 $\{X \leq x\}$ 是必然事件, 于是

$$F(x) = P\{X \leq x\} = 1.$$

综合上述, 即得 X 的分布函数为

$$F(x) = \begin{cases} 0, & x < 0, \\ \frac{x^2}{4}, & 0 \leq x < 2, \\ 1, & x \geq 2. \end{cases}$$

它的图形是一条连续曲线如图 2-6 所示.

另外, 容易看到本例中的分布函数 $F(x)$, 对于任意 x 可以写成形式

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

其中

$$f(t) = \begin{cases} \frac{t}{2}, & 0 < t < 2, \\ 0, & \text{其他.} \end{cases}$$

这就是说, $F(x)$ 恰是非负函数 $f(t)$ 在区间 $(-\infty, x]$ 上的积分, 在这种情况下我们称 X 为连续型随机变量. 下一节我们将给出连续型随机变量的一般定义. $\square$

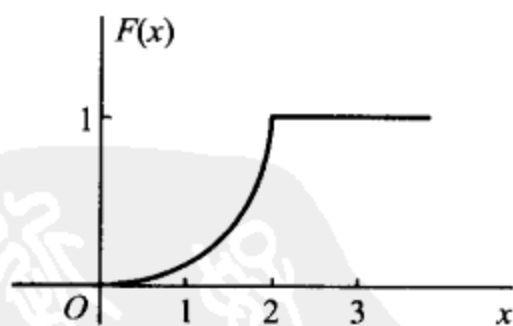


图 2-6

§ 4 连续型随机变量及其概率密度

一般,如上节例 2 中的随机变量那样,如果对于随机变量 X 的分布函数 $F(x)$,存在非负函数 $f(x)$,使对于任意实数 x 有

$$F(x) = \int_{-\infty}^x f(t) dt, \quad (4.1)$$

则称 X 为连续型随机变量,其中函数 $f(x)$ 称为 X 的概率密度函数,简称概率密度①.

由(4.1)式,据数学分析的知识知连续型随机变量的分布函数是连续函数.

在实际应用中遇到的基本上是离散型或连续型随机变量.本书只讨论这两种随机变量.

由定义知道,概率密度 $f(x)$ 具有以下性质:

$$1^\circ f(x) \geq 0;$$

$$2^\circ \int_{-\infty}^{\infty} f(x) dx = 1;$$

$$3^\circ \text{ 对于任意实数 } x_1, x_2 (x_1 \leq x_2),$$

$$P\{x_1 < X \leq x_2\} = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x) dx;$$

$$4^\circ \text{ 若 } f(x) \text{ 在点 } x \text{ 处连续, 则有 } F'(x) = f(x).$$

由性质 2° 知道介于曲线 $y=f(x)$ 与 Ox 轴之间的面积等于 1 (图 2-7). 由 3° 知道 X 落在区间 $(x_1, x_2]$ 的概率 $P\{x_1 < X \leq x_2\}$ 等于区间 $(x_1, x_2]$ 上曲线 $y=f(x)$ 之下的曲边梯形的面积(图 2-8). 由性质 4° 在 $f(x)$ 的连续点 x 处有

$$\begin{aligned} f(x) &= \lim_{\Delta x \rightarrow 0^+} \frac{F(x+\Delta x) - F(x)}{\Delta x} \\ &= \lim_{\Delta x \rightarrow 0^+} \frac{P\{x < X \leq x+\Delta x\}}{\Delta x}. \end{aligned} \quad (4.2)$$

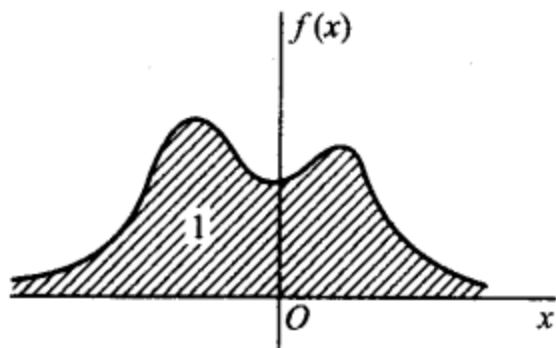


图 2-7

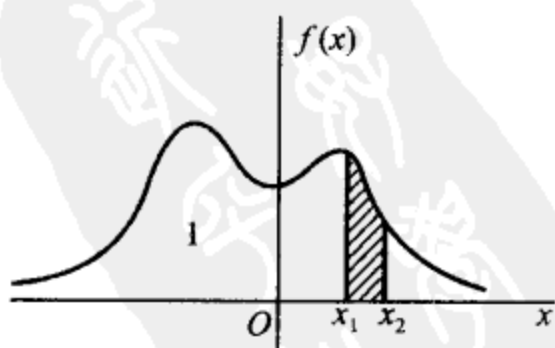


图 2-8

① 由定义知道,改变概率密度 $f(x)$ 在个别点的函数值不影响分布函数 $F(x)$ 的取值. 因此,并不在乎改变概率密度在个别点上的值.

从这里我们看到概率密度的定义与物理学中的线密度的定义相类似,这就是为什么称 $f(x)$ 为概率密度的缘故.

由(4.2)式知道,若不计高阶无穷小,有

$$P\{x < X \leq x + \Delta x\} \approx f(x)\Delta x. \quad (4.3)$$

这表示 X 落在小区间 $(x, x + \Delta x]$ 上的概率近似地等于 $f(x)\Delta x$.

例 1 设随机变量 X 具有概率密度

$$f(x) = \begin{cases} kx, & 0 \leq x < 3, \\ 2 - \frac{x}{2}, & 3 \leq x \leq 4, \\ 0, & \text{其他.} \end{cases}$$

(1) 确定常数 k ; (2) 求 X 的分布函数 $F(x)$; (3) 求 $P\{1 < X \leq \frac{7}{2}\}$.

解 (1) 由 $\int_{-\infty}^{\infty} f(x)dx = 1$, 得

$$\int_0^3 kx dx + \int_3^4 (2 - \frac{x}{2}) dx = 1,$$

解得 $k = \frac{1}{6}$, 于是 X 的概率密度为

$$f(x) = \begin{cases} \frac{x}{6}, & 0 \leq x < 3, \\ 2 - \frac{x}{2}, & 3 \leq x < 4, \\ 0, & \text{其他.} \end{cases}$$

(2) X 的分布函数为

$$F(x) = \begin{cases} 0, & x < 0, \\ \int_0^x \frac{x}{6} dx, & 0 \leq x < 3, \\ \int_0^3 \frac{x}{6} dx + \int_3^x (2 - \frac{x}{2}) dx, & 3 \leq x < 4, \\ 1, & x \geq 4. \end{cases}$$

$$F(x) = \begin{cases} 0, & x < 0, \\ \frac{x^2}{12}, & 0 \leq x < 3, \\ -3 + 2x - \frac{x^2}{4}, & 3 \leq x < 4, \\ 1, & x \geq 4. \end{cases}$$

即

$$(3) \quad P\{1 < X \leq \frac{7}{2}\} = F\left(\frac{7}{2}\right) - F(1) = \frac{41}{48}.$$

□

需要指出的是,对于连续型随机变量 X 来说,它取任一指定实数值 a 的概率均为 0,即 $P\{X=a\}=0$. 事实上,设 X 的分布函数为 $F(x)$, $\Delta x > 0$,则由 $\{X=a\} \subset \{a-\Delta x < X \leq a\}$ 得

$$0 \leq P\{X=a\} \leq P\{a-\Delta x < X \leq a\} = F(a) - F(a-\Delta x).$$

在上述不等式中令 $\Delta x \rightarrow 0$, 并注意到 X 为连续型随机变量,其分布函数 $F(x)$ 是连续的. 即得

$$P\{X=a\}=0. \quad (4.4)$$

据此,在计算连续型随机变量落在某一区间的概率时,可以不必区分该区间是开区间或闭区间或半闭区间. 例如有

$$P\{a < X \leq b\} = P\{a \leq X \leq b\} = P\{a < X < b\}.$$

在这里,事件 $\{X=a\}$ 并非不可能事件,但有 $P\{X=a\}=0$. 这就是说,若 A 是不可能事件,则有 $P(A)=0$; 反之,若 $P(A)=0$, 并不一定意味着 A 是不可能事件.

以后当我们提到一个随机变量 X 的“概率分布”时,指的是它的分布函数; 或者,当 X 是连续型随机变量时,指的是它的概率密度,当 X 是离散型随机变量时,指的是它的分布律.

下面介绍三种重要的连续型随机变量.

(一) 均匀分布

若连续型随机变量 X 具有概率密度

$$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b, \\ 0, & \text{其他,} \end{cases} \quad (4.5)$$

则称 X 在区间 (a, b) 上服从均匀分布. 记为 $X \sim U(a, b)$.

易知 $f(x) \geq 0$, 且 $\int_{-\infty}^{\infty} f(x) dx = 1$.

在区间 (a, b) 上服从均匀分布的随机变量 X , 具有下述意义的等可能性, 即它落在区间 (a, b) 中任意等长度的子区间内的可能性是相同的. 或者说它落在 (a, b) 的子区间内的概率只依赖于子区间的长度而与子区间的位置无关. 事实上, 对于任一长度 l 的子区间 $(c, c+l)$, $a \leq c < c+l \leq b$, 有

$$P\{c < X \leq c+l\} = \int_c^{c+l} f(x) dx = \int_c^{c+l} \frac{1}{b-a} dx = \frac{l}{b-a}.$$

由(4.1)式得 X 的分布函数为

$$F(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \leq x < b, \\ 1, & x \geq b. \end{cases} \quad (4.6)$$

$f(x)$ 及 $F(x)$ 的图形分别如图 2-9, 图 2-10 所示.

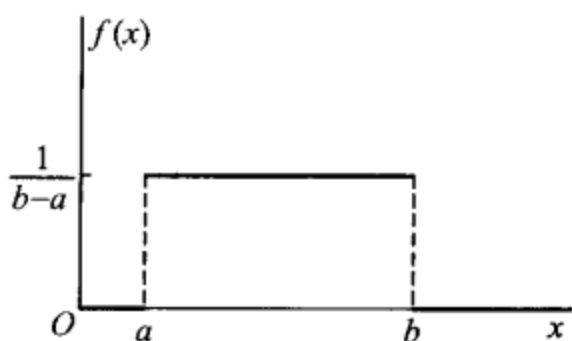


图 2-9

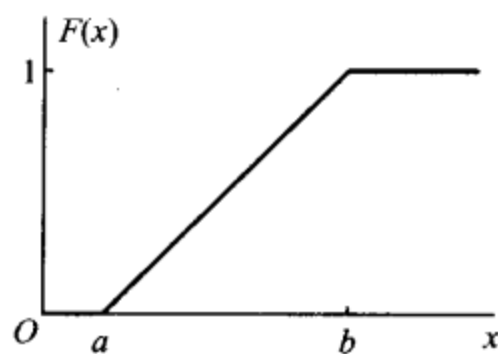


图 2-10

例 2 设电阻值 R 是一个随机变量, 均匀分布在 $900\ \Omega \sim 1100\ \Omega$. 求 R 的概率密度及 R 落在 $950\ \Omega \sim 1050\ \Omega$ 的概率.

解 按题意, R 的概率密度为

$$f(r) = \begin{cases} \frac{1}{1100-900}, & 900 < r < 1100, \\ 0, & \text{其他.} \end{cases}$$

故有
$$P\{950 < R \leq 1050\} = \int_{950}^{1050} \frac{1}{200} dr = 0.5. \quad \square$$

(二) 指数分布

若连续型随机变量 X 的概率密度为

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta}, & x > 0, \\ 0, & \text{其他,} \end{cases} \quad (4.7)$$

其中 $\theta > 0$ 为常数, 则称 X 服从参数为 θ 的指数分布.

易知 $f(x) \geq 0$, 且 $\int_{-\infty}^{\infty} f(x) dx = 1$. 图 2-11 中画出了 $\theta = 1/3, \theta = 1, \theta = 2$ 时 $f(x)$ 的图形.

由 (4.7) 式容易得到随机变量 X 的分布函数为

$$F(x) = \begin{cases} 1 - e^{-x/\theta}, & x > 0, \\ 0, & \text{其他.} \end{cases} \quad (4.8)$$

服从指数分布的随机变量 X 具有以下有趣的性质:

对于任意 $s, t > 0$, 有

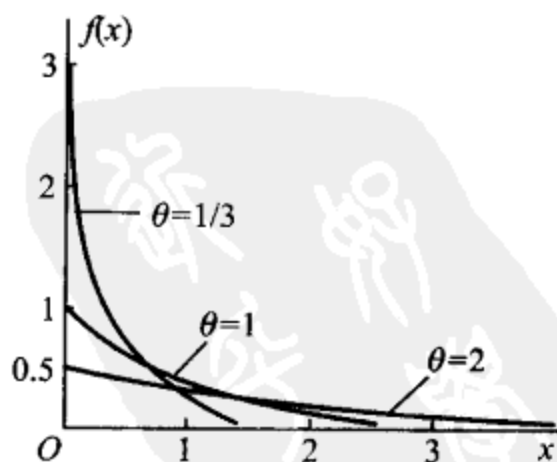


图 2-11

$$P\{X > s+t \mid X > s\} = P\{X > t\}. \quad (4.9)$$

事实上

$$\begin{aligned} P\{X > s+t \mid X > s\} &= \frac{P\{(X > s+t) \cap (X > s)\}}{P\{X > s\}} \\ &= \frac{P\{X > s+t\}}{P\{X > s\}} = \frac{1-F(s+t)}{1-F(s)} \\ &= \frac{e^{-(s+t)/\theta}}{e^{-s/\theta}} = e^{-t/\theta} \\ &= P\{X > t\}. \end{aligned}$$

性质(4.9)称为无记忆性. 如果 X 是某一元件的寿命, 那么(4.9)式表明: 已知元件已使用了 s 小时, 它总共能使用至少 $s+t$ 小时的条件概率, 与从开始使用时算起它至少能使用 t 小时的概率相等. 这就是说, 元件对它已使用过 s 小时没有记忆. 具有这一性质是指数分布有广泛应用的重要原因.

指数分布在可靠性理论与排队论中有广泛的应用.

(三) 正态分布

若连续型随机变量 X 的概率密度为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty, \quad (4.10)$$

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

显然 $f(x) \geq 0$, 下面来证明 $\int_{-\infty}^{\infty} f(x) dx = 1$. 令 $(x-\mu)/\sigma = t$, 得到

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

记 $I = \int_{-\infty}^{\infty} e^{-t^2/2} dt$, 则有 $I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(t^2+u^2)/2} dt du$, 利用极坐标将它化成累次积分, 得到

$$I^2 = \int_0^{2\pi} \int_0^{\infty} r e^{-r^2/2} dr d\theta = 2\pi.$$

而 $I > 0$, 故有 $I = \sqrt{2\pi}$, 即有

$$\int_{-\infty}^{\infty} e^{-t^2/2} dt = \sqrt{2\pi}, \quad (4.11)$$

于是

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

参数 μ, σ 的意义将在第四章中说明. $f(x)$ 的图形如图 2-12 所示, 它具有以下的性质.

1° 曲线关于 $x=\mu$ 对称. 这表明对于任意 $h>0$ 有(图 2-12)

$$P\{\mu-h < X \leq \mu\} = P\{\mu < X \leq \mu+h\}.$$

2° 当 $x=\mu$ 时取到最大值

$$\max f(\mu) = \frac{1}{\sqrt{2\pi}\sigma}.$$

x 离 μ 越远, $f(x)$ 的值越小. 这表明对于同样长度的区间, 当区间离 μ 越远, X 落在这个区间上的概率越小.

在 $x=\mu\pm\sigma$ 处曲线有拐点. 曲线以 Ox 轴为渐近线.

另外, 如果固定 σ , 改变 μ 的值, 则图形沿着 Ox 轴平移, 而不改变其形状(如图 2-12), 可见正态分布的概率密度曲线 $y=f(x)$ 的位置完全由参数 μ 所确定. μ 称为位置参数.

如果固定 μ , 改变 σ , 由于最大值 $f(\mu) = \frac{1}{\sqrt{2\pi}\sigma}$, 可知当 σ 越小时图形变得越尖(如图 2-13), 因而 X 落在 μ 附近的概率越大.

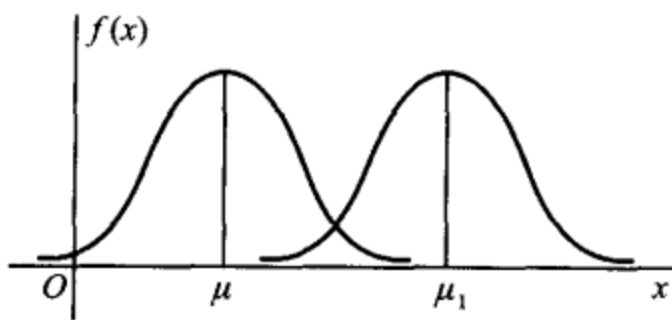


图 2-12

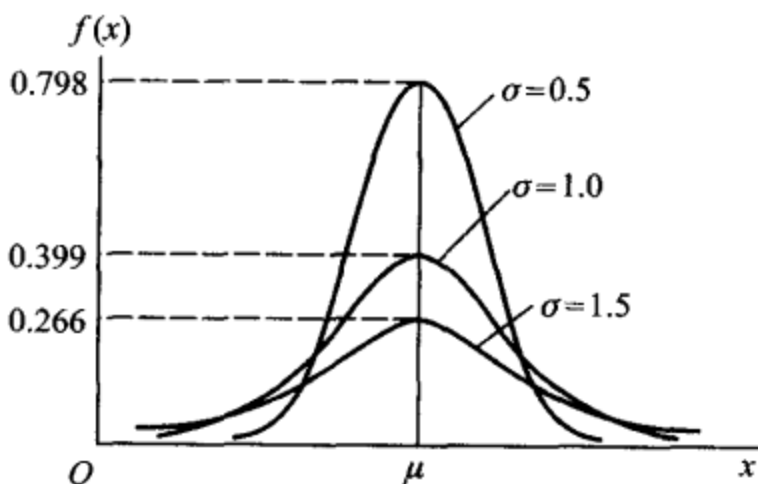


图 2-13

由(4.10)式得 X 的分布函数为(如图 2-14)

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt, \quad (4.12)$$

特别, 当 $\mu=0, \sigma=1$ 时称随机变量 X 服从标准正态分布. 其概率密度和分布函数分别用 $\varphi(x), \Phi(x)$ 表示, 即有

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad (4.13)$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt. \quad (4.14)$$

易知

$$\Phi(-x) = 1 - \Phi(x) \quad (4.15)$$

(参见图 2-15).

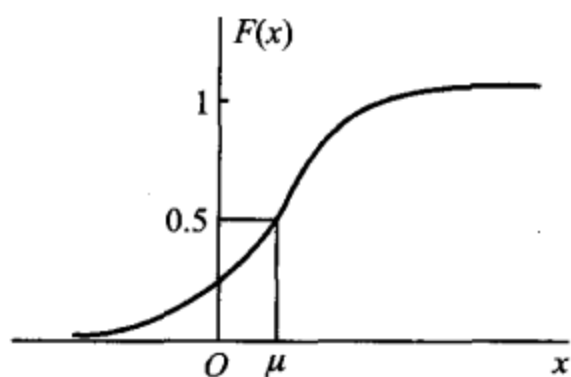


图 2-14

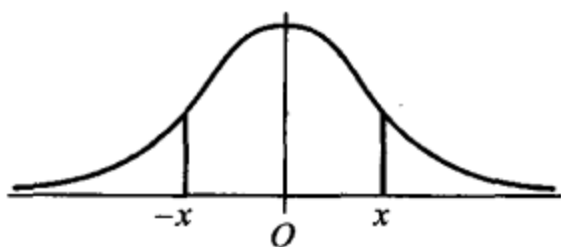


图 2-15

人们已编制了 $\Phi(x)$ 的函数表, 可供查用 (见附表 2).

一般, 若 $X \sim N(\mu, \sigma^2)$, 我们只要通过一个线性变换就能将它化成标准正态分布.

引理 若 $X \sim N(\mu, \sigma^2)$, 则 $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$.

证 $Z = \frac{X - \mu}{\sigma}$ 的分布函数为

$$\begin{aligned} P\{Z \leq x\} &= P\left\{\frac{X - \mu}{\sigma} \leq x\right\} = P\{X \leq \mu + \sigma x\} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\mu + \sigma x} e^{-\frac{(t - \mu)^2}{2\sigma^2}} dt, \end{aligned}$$

令 $\frac{t - \mu}{\sigma} = u$, 得

$$P\{Z \leq x\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2} du = \Phi(x),$$

由此知 $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$. □

于是, 若 $X \sim N(\mu, \sigma^2)$, 则它的分布函数 $F(x)$ 可写成

$$F(x) = P\{X \leq x\} = P\left\{\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right\} = \Phi\left(\frac{x - \mu}{\sigma}\right). \quad (4.16)$$

对于任意区间 $(x_1, x_2]$, 有

$$\begin{aligned} P\{x_1 < X \leq x_2\} &= P\left\{\frac{x_1 - \mu}{\sigma} < \frac{X - \mu}{\sigma} \leq \frac{x_2 - \mu}{\sigma}\right\} \\ &= \Phi\left(\frac{x_2 - \mu}{\sigma}\right) - \Phi\left(\frac{x_1 - \mu}{\sigma}\right). \end{aligned} \quad (4.17)$$

例如, 设 $X \sim N(1, 4)$, 查表得

$$P\{0 < X \leq 1.6\} = \Phi\left(\frac{1.6 - 1}{2}\right) - \Phi\left(\frac{0 - 1}{2}\right)$$

$$\begin{aligned}
 &= \Phi(0.3) - \Phi(-0.5) \\
 &= 0.6179 - [1 - \Phi(0.5)] \\
 &= 0.6179 - 1 + 0.6915 = 0.3094.
 \end{aligned}$$

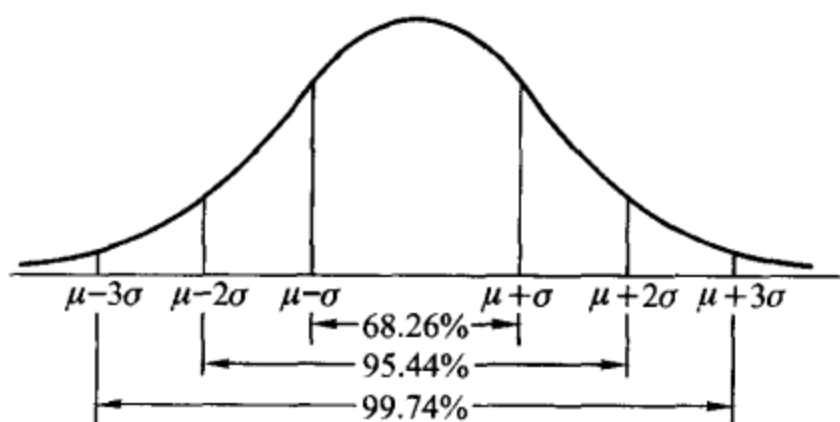


图 2-16

设 $X \sim N(\mu, \sigma^2)$, 由 $\Phi(x)$ 的函数表还能得到(图 2-16):

$$\begin{aligned}
 P\{\mu - \sigma < X < \mu + \sigma\} &= \Phi(1) - \Phi(-1) \\
 &= 2\Phi(1) - 1 = 68.26\%,
 \end{aligned}$$

$$P\{\mu - 2\sigma < X < \mu + 2\sigma\} = \Phi(2) - \Phi(-2) = 95.44\%,$$

$$P\{\mu - 3\sigma < X < \mu + 3\sigma\} = \Phi(3) - \Phi(-3) = 99.74\%.$$

我们看到, 尽管正态变量的取值范围是 $(-\infty, \infty)$, 但它的值落在 $(\mu - 3\sigma, \mu + 3\sigma)$ 内几乎是肯定的事. 这就是人们所说的“3σ”法则.

例 3 将一温度调节器放置在贮存着某种液体的容器内. 调节器整定在 $d^\circ\text{C}$, 液体的温度 X (以 $^\circ\text{C}$ 计) 是一个随机变量, 且 $X \sim N(d, 0.5^2)$. (1) 若 $d = 90^\circ\text{C}$, 求 X 小于 89°C 的概率. (2) 若要求保持液体的温度至少为 80°C 的概率不低于 0.99, 问 d 至少为多少?

解 (1) 所求概率为

$$\begin{aligned}
 P\{X < 89\} &= P\left\{\frac{X - 90}{0.5} < \frac{89 - 90}{0.5}\right\} \\
 &= \Phi\left(\frac{89 - 90}{0.5}\right) = \Phi(-2) \\
 &= 1 - \Phi(2) = 1 - 0.9772 = 0.0228.
 \end{aligned}$$

(2) 按题意需求 d 满足

$$\begin{aligned}
 0.99 &\leq P\{X \geq 80\} = P\left\{\frac{X - d}{0.5} \geq \frac{80 - d}{0.5}\right\} \\
 &= 1 - P\left\{\frac{X - d}{0.5} < \frac{80 - d}{0.5}\right\} = 1 - \Phi\left(\frac{80 - d}{0.5}\right).
 \end{aligned}$$

即

$$\Phi\left(\frac{d - 80}{0.5}\right) \geq 0.99 = \Phi(2.327),$$

亦即

$$\frac{d-80}{0.5} \geq 2.327.$$

故需

$$d > 81.1635.$$

□

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.

设 $X \sim N(0,1)$, 若 z_α 满足条件

$$P\{X > z_\alpha\} = \alpha, \quad 0 < \alpha < 1, \quad (4.18)$$

则称点 z_α 为标准正态分布的上 α 分位点(如图 2-17). 下面列出了几个常用的 z_α 的值.

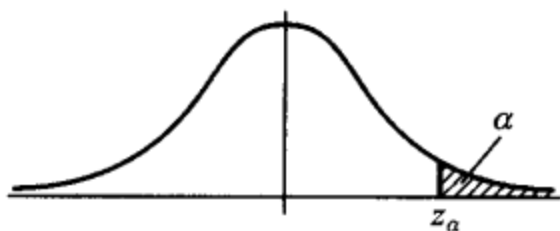


图 2-17

α	0.001	0.005	0.01	0.025	0.05	0.10
z_α	3.090	2.576	2.326	1.960	1.645	1.282

另外,由 $\varphi(x)$ 图形的对称性知道 $z_{1-\alpha} = -z_\alpha$.

在自然现象和社会现象中,大量随机变量都服从或近似服从正态分布.例如,一个地区的男性成年人的身高;测量某零件长度的误差,海洋波浪的高度,半导体器件中的热噪声电流或电压等,都服从正态分布.在概率论与数理统计的理论研究和实际应用中正态随机变量起着特别重要的作用.在第五章我们将进一步说明正态随机变量的重要性.

§ 5 随机变量的函数的分布

在实际中,我们常对某些随机变量的函数更感兴趣.例如,在一些试验中,所关心的随机变量往往不能由直接测量得到,而它却是某个能直接测量的随机变量的函数.比如我们能测量圆轴截面的直径 d ,而关心的却是截面面积 $A = \frac{1}{4}\pi d^2$. 这里,随机变量 A 是随机变量 d 的函数.在这一节中,我们将讨论如何由已知的随机变量 X 的概率分布去求得它的函数 $Y = g(X)$ ($g(\cdot)$ 是已知的连续函数)的概率分布.这里 Y 是这样的随机变量,当 X 取值 x 时, Y 取值 $g(x)$.

例 1 设随机变量 X 具有以下的分布律,试求 $Y = (X-1)^2$ 的分布律.

X	-1	0	1	2
p_k	0.2	0.3	0.1	0.4

解 Y 所有可能取的值为 0, 1, 4. 由

$$P\{Y=0\} = P\{(X-1)^2=0\} = P\{X=1\} = 0.1,$$

$$P\{Y=1\} = P\{X=0\} + P\{X=2\} = 0.7,$$

$$P\{Y=4\} = P\{X=-1\} = 0.2,$$

即得 Y 的分布律为

Y	0	1	4
p_k	0.1	0.7	0.2

□

例 2 设随机变量 X 具有概率密度

$$f_X(x) = \begin{cases} \frac{x}{8}, & 0 < x < 4, \\ 0, & \text{其他.} \end{cases}$$

求随机变量 $Y=2X+8$ 的概率密度.

解 分别记 X, Y 的分布函数为 $F_X(x), F_Y(y)$. 下面先来求 $F_Y(y)$.

$$\begin{aligned} F_Y(y) &= P\{Y \leq y\} = P\{2X+8 \leq y\} \\ &= P\left\{X \leq \frac{y-8}{2}\right\} = F_X\left(\frac{y-8}{2}\right). \end{aligned}$$

将 $F_Y(y)$ 关于 y 求导数, 得 $Y=2X+8$ 的概率密度为

$$\begin{aligned} f_Y(y) &= f_X\left(\frac{y-8}{2}\right) \left(\frac{y-8}{2}\right)' \\ &= \begin{cases} \frac{1}{8} \left(\frac{y-8}{2}\right) \cdot \frac{1}{2}, & 0 < \frac{y-8}{2} < 4, \\ 0, & \text{其他} \end{cases} \\ &= \begin{cases} \frac{y-8}{32}, & 8 < y < 16, \\ 0, & \text{其他.} \end{cases} \end{aligned}$$

□

例 3 设随机变量 X 具有概率密度 $f_X(x)$, $-\infty < x < \infty$, 求 $Y=X^2$ 的概率密度.

解 分别记 X, Y 的分布函数为 $F_X(x), F_Y(y)$. 先来求 Y 的分布函数 $F_Y(y)$. 由于 $Y=X^2 \geq 0$, 故当 $y \leq 0$ 时 $F_Y(y)=0$. 当 $y > 0$ 时有

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

将 $F_Y(y)$ 关于 y 求导数, 即得 Y 的概率密度为

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}} [f_X(\sqrt{y}) + f_X(-\sqrt{y})], & y > 0, \\ 0, & y \leq 0. \end{cases} \quad (5.1) \square$$

例如, 设 $X \sim N(0, 1)$, 其概率密度为

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

由(5.1)得 $Y=X^2$ 的概率密度为

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

此时称 Y 服从自由度为 1 的 χ^2 分布.

上述两个例子解法的关键一步是在“ $Y \leq y$ ”中,即在“ $g(X) \leq y$ ”中解出 X ,从而得到一个与“ $g(X) \leq y$ ”等价的 X 的不等式,并以后者代替“ $g(X) \leq y$ ”.例如,在例 2 中以“ $X \leq \frac{y-8}{2}$ ”代替“ $2X+8 \leq y$ ”;在例 3 中,当 $y > 0$ 时以“ $-\sqrt{y} \leq X \leq \sqrt{y}$ ”代替“ $X^2 \leq y$ ”.一般来说,可以用这样的方法^①求连续型随机变量的函数的分布函数或概率密度.下面我们仅对 $Y=g(X)$,其中 $g(\cdot)$ 是严格单调函数的情况,写出一般的结果.

定理 设随机变量 X 具有概率密度 $f_X(x)$, $-\infty < x < \infty$, 又设函数 $g(x)$ 处处可导且恒有 $g'(x) > 0$ (或恒有 $g'(x) < 0$), 则 $Y=g(X)$ 是连续型随机变量, 其概率密度为

$$f_Y(y) = \begin{cases} f_X[h(y)] |h'(y)|, & \alpha < y < \beta, \\ 0, & \text{其他,} \end{cases} \quad (5.2)$$

其中 $\alpha = \min\{g(-\infty), g(\infty)\}$, $\beta = \max\{g(-\infty), g(\infty)\}$, $h(y)$ 是 $g(x)$ 的反函数.

我们只证 $g'(x) > 0$ 的情况. 此时 $g(x)$ 在 $(-\infty, \infty)$ 严格单调增加, 它的反函数 $h(y)$ 存在, 且在 (α, β) 严格单调增加, 可导. 分别记 X, Y 的分布函数为 $F_X(x), F_Y(y)$. 现在先来求 Y 的分布函数 $F_Y(y)$.

因为 $Y=g(X)$ 在 (α, β) 取值, 故当 $y \leq \alpha$ 时 $F_Y(y) = P\{Y \leq y\} = 0$; 当 $y \geq \beta$ 时, $F_Y(y) = P\{Y \leq y\} = 1$.

当 $\alpha < y < \beta$ 时,

$$\begin{aligned} F_Y(y) &= P\{Y \leq y\} = P\{g(X) \leq y\} \\ &= P\{X \leq h(y)\} = F_X[h(y)]. \end{aligned}$$

将 $F_Y(Y)$ 关于 y 求导数, 即得 Y 的概率密度

$$f_Y(y) = \begin{cases} f_X[h(y)] h'(y), & \alpha < y < \beta, \\ 0, & \text{其他.} \end{cases} \quad (5.3)$$

对于 $g'(x) < 0$ 的情况可以同样地证明, 此时有

$$f_Y(y) = \begin{cases} f_X[h(y)] [-h'(y)], & \alpha < y < \beta, \\ 0, & \text{其他.} \end{cases} \quad (5.4)$$

^① 连续型随机变量 X 的函数 $Y=g(X)$ 不一定是连续型的随机变量.

合并(5.3)与(5.4)两式,(5.2)得证. $\square$

若 $f(x)$ 在有限区间 $[a, b]$ 以外等于零, 则只需假设在 $[a, b]$ 上恒有 $g'(x) > 0$ (或恒有 $g'(x) < 0$), 此时

$$\alpha = \min\{g(a), g(b)\}, \beta = \max\{g(a), g(b)\}.$$

例 4 设随机变量 $X \sim N(\mu, \sigma^2)$. 试证明 X 的线性函数 $Y = aX + b$ ($a \neq 0$) 也服从正态分布.

证 X 的概率密度为

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

现在 $y = g(x) = ax + b$, 由这一式子解得

$$x = h(y) = \frac{y-b}{a}, \text{ 且有 } h'(y) = \frac{1}{a}.$$

由(5.2)式得 $Y = aX + b$ 的概率密度为

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

即

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

即有 $Y = aX + b \sim N(a\mu + b, (a\sigma)^2)$.

特别, 在上例中取 $a = \frac{1}{\sigma}, b = -\frac{\mu}{\sigma}$ 得

$$Y = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

这就是上一节引理的结果. $\square$

例 5 设电压 $V = A \sin \Theta$, 其中 A 是一个已知的正常数, 相角 Θ 是一个随机变量, 且有 $\Theta \sim U\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, 试求电压 V 的概率密度.

解 现在 $v = g(\theta) = A \sin \theta$ 在 $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ 上恒有 $g'(\theta) = A \cos \theta > 0$, 且有反函数

$$\theta = h(v) = \arcsin \frac{v}{A}, \quad h'(v) = \frac{1}{\sqrt{A^2 - v^2}}.$$

又, Θ 的概率密度为

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

由(5.2)式得 $V = A \sin \Theta$ 的概率密度为

$$\phi(v) = \begin{cases} \frac{1}{\pi} \cdot \frac{1}{\sqrt{A^2 - v^2}}, & -A < v < A, \\ 0, & \text{其他.} \end{cases}$$

□

若在上题中 $\Theta \sim U(0, \pi)$, 因为此时 $v = g(\theta) = A \sin \theta$ 在 $(0, \pi)$ 上不是单调函数, 上述定理失效, 应仍按例 3 的方法来做. 请读者自行求出其结果.

小结

随机变量 $X = X(e)$ 是定义在样本空间 $S = \{e\}$ 上的实值单值函数. 也就是说, 它是随机试验结果的函数. 它的取值随试验的结果而定, 是不能预先确定的, 它的取值有一定的概率. 随机变量的引入, 使概率论的研究由个别随机事件扩大为随机变量所表征的随机现象的研究. 今后, 我们主要研究随机变量和它的分布.

一个随机变量, 如果它所有可能的值是有限个或可列无限个, 这种随机变量称为离散型随机变量, 不是这种情况则称为非离散型的. 不论是离散型的或非离散型的随机变量 X , 都可以借助分布函数

$$F(x) = P\{X \leq x\}, \quad -\infty < x < \infty$$

来描述. 若已知随机变量 X 的分布函数, 就能知道 X 落在任一区间 $(x_1, x_2]$ 上的概率

$$P\{x_1 < X \leq x_2\} = F(x_2) - F(x_1), \quad x_1 < x_2.$$

这样, 分布函数就能完整地描述随机变量取值的统计规律性.

对于离散型随机变量, 我们需要掌握的是它可能取哪些值, 以及它以怎样的概率取这些值, 这就是离散型随机变量取值的统计规律性. 因而, 对于离散型随机变量, 用分布律

$$P\{X = x_k\} = p_k, \quad k = 1, 2, \dots$$

或写成

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

(这里, $\sum_{k=1}^{\infty} p_k = 1$) 来描述它的取值的统计规律性较为直观和简洁. 分布律与分布函数有以下关系

$$F(x) = P\{X \leq x\} = \sum_{x_k \leq x} P\{X = x_k\},$$

它们是一一对应的.

设随机变量 X 的分布函数为 $F(x)$, 如果存在非负可积函数 $f(x)$, 使得对于任意 x , 有

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

则称 X 是连续型随机变量, 其中 $f(x) \geq 0$ 称为 X 的概率密度.

给定 X 的概率密度 $f(x)$ 就能确定 $F(x)$, 由于 $f(x)$ 位于积分号之内, 故改变 $f(x)$ 在个别点上的函数值并不改变 $F(x)$ 的值. 因此, 改变 $f(x)$ 在个别点上的值, 是无关紧要的.

连续型随机变量 X 的分布函数是连续的, 连续型随机变量取任一指定实数值 a 的概率

为 0, 即 $P\{X=a\}=0$. 这两点性质离散型随机变量是不具备的.

我们将随机变量分成为

$$\text{随机变量} \begin{cases} \text{离散型} \\ \text{非离散型} \begin{cases} \text{连续型} \\ \text{其他} \end{cases} \end{cases}$$

读者不要误以为, 一个随机变量, 如果它不是离散型的那一定是连续型的. 但本书只讨论两类重要的随机变量: 离散型和连续型随机变量.

读者应掌握分布函数、分布律、概率密度的性质. 本章引入了几种重要的随机变量的分布: (0-1) 分布, 二项分布, 泊松分布, 指数分布, 均匀分布和正态分布. 读者必须熟知这几种随机变量的分布律或概率密度.

随机变量 X 的函数 $Y=g(X)$ 也是一个随机变量, 要掌握如何由已知的 X 的分布 (X 的分布律或概率密度) 去求得 $Y=g(X)$ 的分布 (Y 的分布律或概率密度).

■ 重要术语及主题

随机变量 分布函数 离散型随机变量及其分布律 连续型随机变量及其概率密度
伯努利试验 (0-1) 分布 n 重伯努利试验 二项分布 泊松分布 指数分布 均匀分布
正态分布 随机变量函数的分布

习题

1. 考虑为期一年的一张保险单, 若投保人在投保后一年内因意外死亡, 则公司赔付 20 万元, 若投保人因其他原因死亡, 则公司赔付 5 万元, 若投保人在投保期末生存, 则公司无需付给任何费用. 若投保人在一年内因意外死亡的概率为 0.0002, 因其他原因死亡的概率为 0.0010, 求公司赔付金额的分布律.

2. (1) 一袋中装有 5 只球, 编号为 1, 2, 3, 4, 5. 在袋中同时取 3 只, 以 X 表示取出的 3 只球中的最大号码, 写出随机变量 X 的分布律.

(2) 将一颗骰子抛掷两次, 以 X 表示两次中得到的小的点数, 试求 X 的分布律.

3. 设在 15 只同类型的零件中有 2 只是次品, 在其中取 3 次, 每次任取 1 只, 作不放回抽样. 以 X 表示取出的次品的只数.

(1) 求 X 的分布律.

(2) 画出分布律的图形.

4. 进行重复独立试验, 设每次试验的成功概率为 p , 失败概率为 $q=1-p$ ($0 < p < 1$).

(1) 将试验进行到出现一次成功为止, 以 X 表示所需的试验次数, 求 X 的分布律. (此时称 X 服从以 p 为参数的几何分布.)

(2) 将试验进行到出现 r 次成功为止, 以 Y 表示所需的试验次数, 求 Y 的分布律. (此时称 Y 服从以 r, p 为参数的巴斯卡分布或负二项分布.)

(3) 一篮球运动员的投篮命中率为 45%. 以 X 表示他首次投中时累计已投篮的次数, 写出 X 的分布律, 并计算 X 取偶数的概率.

5. 一房间有 3 扇同样大小的窗子, 其中只有一扇是打开的. 有一只鸟自开着的窗子飞入了房间, 它只能从开着的窗子飞出去. 鸟在房子里飞来飞去, 试图飞出房间. 假定鸟是没有记

忆的,它飞向各扇窗子是随机的.

(1) 以 X 表示鸟为了飞出房间试飞的次数,求 X 的分布律.

(2) 户主声称,他养的一只鸟是有记忆的,它飞向任一窗子的尝试不多于一次.以 Y 表示这只聪明的鸟为了飞出房间试飞的次数.如户主所说是确实的,试求 Y 的分布律.

(3) 求试飞次数 X 小于 Y 的概率和试飞次数 Y 小于 X 的概率.

6. 一大楼装有 5 台同类型的供水设备.设各台设备是否被使用相互独立.调查表明在任一时刻 t 每台设备被使用的概率为 0.1,问在同一时刻,

(1) 恰有 2 台设备被使用的概率是多少?

(2) 至少有 3 台设备被使用的概率是多少?

(3) 至多有 3 台设备被使用的概率是多少?

(4) 至少有 1 台设备被使用的概率是多少?

7. 设事件 A 在每次试验发生的概率为 0.3. A 发生不少于 3 次时,指示灯发出信号.

(1) 进行了 5 次重复独立试验,求指示灯发出信号的概率.

(2) 进行了 7 次重复独立试验,求指示灯发出信号的概率.

8. 甲、乙两人投篮,投中的概率分别为 0.6, 0.7. 今各投 3 次. 求

(1) 两人投中次数相等的概率;

(2) 甲比乙投中次数多的概率.

9. 有一大批产品,其验收方案如下,先作第一次检验:从中任取 10 件,经检验无次品接受这批产品,次品数大于 2 拒收;否则作第二次检验,其做法是从中再任取 5 件,仅当 5 件中无次品时接受这批产品.若产品的次品率为 10%,求

(1) 这批产品经第一次检验就能接受的概率.

(2) 需作第二次检验的概率.

(3) 这批产品按第二次检验的标准被接受的概率.

(4) 这批产品在第一次检验未能作决定且第二次检验时被通过的概率.

(5) 这批产品被接受的概率.

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

(1) 某人随机地去猜,问他试验成功一次的概率是多少?

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

11. 尽管在几何教科书中已经讲过仅用圆规和直尺三等分一个任意角是不可能的,但每一年总是有一些“发明者”撰写关于仅用圆规和直尺将角三等分的文章. 设某地区每年撰写此类文章的篇数 X 服从参数为 6 的泊松分布. 求明年没有此类文章的概率.

12. 一电话总机每分钟收到呼唤的次数服从参数为 4 的泊松分布. 求

(1) 某一分钟恰有 8 次呼唤的概率;

(2) 某一分钟的呼唤次数大于 3 的概率.

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

(1) 求某一天中午 12 时至下午 3 时未收到紧急呼救的概率.

(2) 求某一天中午 12 时至下午 5 时至少收到 1 次紧急呼救的概率.

14. 某人家中在时间间隔 t (小时) 内接到电话的次数 X 服从参数为 $2t$ 的泊松分布.

(1) 若他外出计划用时 10 分钟, 问其间有电话铃响一次的概率是多少?

(2) 若他希望外出时没有电话的概率至少为 0.5, 问他外出应控制最长时间是多少?

15. 保险公司在一天内承保了 5000 张相同年龄, 为期一年的寿险保单, 每人一份. 在合同有效期内若投保人死亡, 则公司需赔付 3 万元. 设在一年内, 该年龄段的死亡率为 0.0015, 且各投保人是否死亡相互独立. 求该公司对于这批投保人的赔付总额不超过 30 万元的概率 (利用泊松定理计算).

16. 有一繁忙的汽车站, 每天有大量汽车通过, 设一辆汽车在一天的某段时间内出事故的概率为 0.0001. 在某天的该时间段内有 1000 辆汽车通过. 问出事故的车辆数不小于 2 的概率是多少? (利用泊松定理计算)

17. (1) 设 X 服从 $(0-1)$ 分布, 其分布律为 $P\{X=k\} = p^k(1-p)^{1-k}, k=0,1$, 求 X 的分布函数, 并作出其图形.

(2) 求第 2 题(1)中的随机变量的分布函数.

18. 在区间 $[0, a]$ 上任意投掷一个质点, 以 X 表示这个质点的坐标. 设这个质点落在 $[0, a]$ 中任意小区间内的概率与这个小区间的长度成正比例. 试求 X 的分布函数.

19. 以 X 表示某商店从早晨开始营业起直到第一个顾客到达的等待时间 (以分计), X 的分布函数是

$$F_X(x) = \begin{cases} 1 - e^{-0.4x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

求下述概率:

(1) $P\{\text{至多 3 分钟}\}$.

(2) $P\{\text{至少 4 分钟}\}$.

(3) $P\{3 \text{ 分钟至 } 4 \text{ 分钟之间}\}$.

(4) $P\{\text{至多 3 分钟或至少 4 分钟}\}$.

(5) $P\{\text{恰好 2.5 分钟}\}$.

20. 设随机变量 X 的分布函数为

$$F_X(x) = \begin{cases} 0, & x < 1, \\ \ln x, & 1 \leq x < e, \\ 1, & x \geq e. \end{cases}$$

(1) 求 $P\{X < 2\}, P\{0 < X \leq 3\}, P\{2 < X < 5/2\}$.

(2) 求概率密度 $f_X(x)$.

21. 设随机变量 X 的概率密度为

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

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

求 X 的分布函数 $F(x)$, 并画出(2)中的 $f(x)$ 及 $F(x)$ 的图形.

22. (1) 分子运动速度的绝对值 X 服从麦克斯韦(Maxwell)分布, 其概率密度为

$$f(x) = \begin{cases} Ax^2 e^{-x^2/b}, & x > 0, \\ 0, & \text{其他}, \end{cases}$$

其中 $b = m/(2kT)$, k 为玻耳兹曼(Boltzmann)常数, T 为绝对温度, m 是分子的质量, 试确定常数 A .

(2) 研究了英格兰在 1875 年~1951 年期间, 在矿山发生导致不少于 10 人死亡的事实的频繁程度, 得知相继两次事故之间的时间 T (日)服从指数分布, 其概率密度为

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

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

23. 某种型号器件的寿命 X (以小时计)具有概率密度

$$f(x) = \begin{cases} \frac{1000}{x^2}, & x > 1000, \\ 0, & \text{其他}. \end{cases}$$

现有一大批此种器件(设各器件损坏与否相互独立), 任取 5 只, 问其中至少有 2 只寿命大于 1500 小时的概率是多少?

24. 设顾客在某银行的窗口等待服务的时间 X (min)服从指数分布, 其概率密度为

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

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

25. 设 K 在 $(0, 5)$ 服从均匀分布, 求 x 的方程

$$4x^2 + 4Kx + K + 2 = 0$$

有实根的概率.

26. 设 $X \sim N(3, 2^2)$.

(1) 求 $P\{2 < X \leq 5\}$, $P\{-4 < X \leq 10\}$, $P\{|X| > 2\}$, $P\{X > 3\}$.

(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?

27. 某地区 18 岁的女青年的血压(收缩压, 以 mmHg 计)服从 $N(110, 12^2)$ 分布. 在该地区任选一 18 岁的女青年, 测量她的血压 X . 求

(1) $P\{X \leq 105\}$, $P\{100 < X \leq 120\}$;

(2) 确定最小的 x , 使 $P\{X > x\} \leq 0.05$.

28. 由某机器生产的螺栓的长度(cm)服从参数 $\mu = 10.05$, $\sigma = 0.06$ 的正态分布. 规定长度在范围 10.05 ± 0.12 内为合格品, 求一螺栓为不合格品的概率.

29. 一工厂生产的某种元件的寿命 X (以小时计)服从参数为 $\mu = 160$, σ ($\sigma > 0$) 的正态分布. 若要求 $P\{120 < X \leq 200\} \geq 0.80$, 允许 σ 最大为多少?

30. 设在一电路中,电阻两端的电压(V)服从 $N(120, 2^2)$, 今独立测量了 5 次, 试确定有 2 次测定值落在区间 $[118, 122]$ 之外的概率.

31. 某人上班, 自家里去办公楼要经过一交通指示灯, 这一指示灯有 80% 时间亮红灯, 此时他在指示灯旁等待直至绿灯亮. 等待时间在区间 $[0, 30]$ (以秒计) 服从均匀分布. 以 X 表示他的等待时间, 求 X 的分布函数 $F(x)$. 画出 $F(x)$ 的图形, 并问 X 是否为连续型随机变量, 是否为离散型的? (要说明理由)

32. 设 $f(x), g(x)$ 都是概率密度函数, 求证

$$h(x) = \alpha f(x) + (1 - \alpha)g(x), \quad 0 \leq \alpha \leq 1$$

也是一个概率密度函数.

33. 设随机变量 X 的分布律为

X	-2	-1	0	1	3
p_k	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{15}$	$\frac{11}{30}$

求 $Y = X^2$ 的分布律.

34. 设随机变量 X 在区间 $(0, 1)$ 服从均匀分布.

(1) 求 $Y = e^X$ 的概率密度.

(2) 求 $Y = -2 \ln X$ 的概率密度.

35. 设 $X \sim N(0, 1)$.

(1) 求 $Y = e^X$ 的概率密度.

(2) 求 $Y = 2X^2 + 1$ 的概率密度.

(3) 求 $Y = |X|$ 的概率密度.

36. (1) 设随机变量 X 的概率密度为 $f(x)$, $-\infty < x < \infty$. 求 $Y = X^3$ 的概率密度.

(2) 设随机变量 X 的概率密度为

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

求 $Y = X^2$ 的概率密度.

37. 设随机变量 X 的概率密度为

$$f(x) = \begin{cases} \frac{2x}{\pi^2}, & 0 < x < \pi, \\ 0, & \text{其他.} \end{cases}$$

求 $Y = \sin X$ 的概率密度.

38. 设电流 I 是一个随机变量, 它均匀分布在 $9 \text{ A} \sim 11 \text{ A}$ 之间. 若此电流通过 2Ω 的电阻, 在其上消耗的功率 $W = 2I^2$. 求 W 的概率密度.

39. 某物体的温度 $T(^{\circ}\text{F})$ 是随机变量, 且有 $T \sim N(98.6, 2)$, 已知 $\Theta = \frac{5}{9}(T - 32)$, 试求 $\Theta(^{\circ}\text{C})$ 的概率密度.

第三章 多维随机变量及其分布

§ 1 二维随机变量

以上我们只限于讨论一个随机变量的情况,但在实际问题中,对于某些随机试验的结果需要同时用两个或两个以上的随机变量来描述.例如,为了研究某一地区学龄前儿童的发育情况,对这一地区的儿童进行抽查.对于每个儿童都能观察到他的身高 H 和体重 W . 在这里,样本空间 $S = \{e\} = \{\text{某地区的全部学龄前儿童}\}$,而 $H(e)$ 和 $W(e)$ 是定义在 S 上的两个随机变量.又如炮弹弹着点的位置需要由它的横坐标和纵坐标来确定,而横坐标和纵坐标是定义在同一个样本空间的两个随机变量.

一般,设 E 是一个随机试验,它的样本空间是 $S = \{e\}$,设 $X = X(e)$ 和 $Y = Y(e)$ 是定义在 S 上的随机变量,由它们构成的一个向量 (X, Y) ,叫做 二维随机向量 或 二维随机变量 (如图 3-1). 第二章讨论的随机变量也叫一维随机变量.

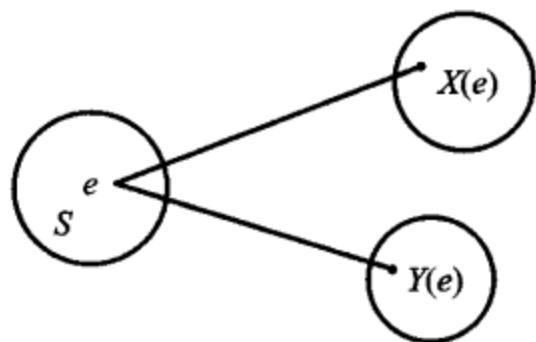


图 3-1

二维随机变量 (X, Y) 的性质不仅与 X 及 Y 有关,而且还依赖于这两个随机变量的相互关系.因此,逐个地来研究 X 或 Y 的性质是不够的,还需将 (X, Y) 作为一个整体来进行研究.

和一维的情况类似,我们也借助“分布函数”来研究二维随机变量.

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

$$F(x, y) = P\{(X \leq x) \cap (Y \leq y)\} \stackrel{\text{记成}}{=} P\{X \leq x, Y \leq y\}$$

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

如果将二维随机变量 (X, Y) 看成是平面上随机点的坐标,那么,分布函数 $F(x, y)$ 在 (x, y) 处的函数值就是随机点 (X, Y) 落在如图 3-2 所示的,以点 (x, y) 为顶点而位于该点左下方的无穷矩形域内的概率.

依照上述解释,借助于图 3-3 容易算出随机点 (X, Y) 落在矩形域 $\{(x, y) | x_1 < x \leq x_2, y_1 < y \leq y_2\}$ 的概率为

$$\begin{aligned}
 &P\{x_1 < X \leq x_2, y_1 < Y \leq y_2\} \\
 &= F(x_2, y_2) - F(x_2, y_1) + F(x_1, y_1) - F(x_1, y_2).
 \end{aligned} \tag{1.1}$$

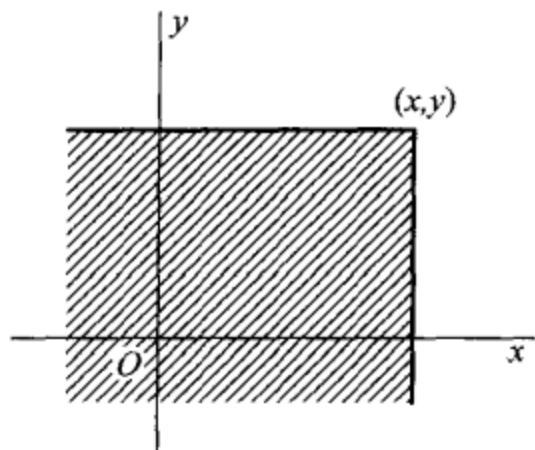


图 3-2

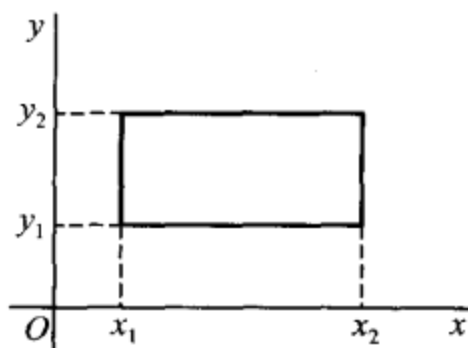


图 3-3

分布函数 $F(x, y)$ 具有以下的基本性质:

1° $F(x, y)$ 是变量 x 和 y 的不减函数, 即对于任意固定的 y , 当 $x_2 > x_1$ 时 $F(x_2, y) \geq F(x_1, y)$; 对于任意固定的 x , 当 $y_2 > y_1$ 时 $F(x, y_2) \geq F(x, y_1)$.

2° $0 \leq F(x, y) \leq 1$, 且

对于任意固定的 $y, F(-\infty, y) = 0$,

对于任意固定的 $x, F(x, -\infty) = 0$,

$F(-\infty, -\infty) = 0, F(\infty, \infty) = 1$.

上面四个式子可以从几何上加以说明. 例如, 在图 3-2 中将无穷矩形的右面边界向左无限平移 (即 $x \rightarrow -\infty$), 则“随机点 (X, Y) 落在这个矩形内”这一事件趋于不可能事件, 故其概率趋于 0, 即有 $F(-\infty, y) = 0$; 又如当 $x \rightarrow \infty, y \rightarrow \infty$ 时图 3-2 中的无穷矩形扩展到全平面, 随机点 (X, Y) 落在其中这一事件趋于必然事件, 故其概率趋于 1, 即 $F(\infty, \infty) = 1$.

3° $F(x+0, y) = F(x, y), F(x, y+0) = F(x, y)$, 即 $F(x, y)$ 关于 x 右连续, 关于 y 也右连续.

4° 对于任意 $(x_1, y_1), (x_2, y_2), x_1 < x_2, y_1 < y_2$, 下述不等式成立:

$$F(x_2, y_2) - F(x_2, y_1) + F(x_1, y_1) - F(x_1, y_2) \geq 0.$$

这一性质由 (1.1) 式及概率的非负性即可得.

如果二维随机变量 (X, Y) 全部可能取到的值是有限对或可列无限多对, 则称 (X, Y) 是离散型的随机变量.

设二维离散型随机变量 (X, Y) 所有可能取的值为 $(x_i, y_j), i, j = 1, 2, \dots$, 记 $P\{X = x_i, Y = y_j\} = p_{ij}, i, j = 1, 2, \dots$, 则由概率的定义有

$$p_{ij} \geq 0, \quad \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} p_{ij} = 1.$$

我们称 $P\{X=x_i, Y=y_j\}=p_{ij}, i, j=1, 2, \dots$ 为二维离散型随机变量 (X, Y) 的分布律, 或随机变量 X 和 Y 的联合分布律.

我们也能用表格来表示 X 和 Y 的联合分布律, 如下表所示.

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

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

解 由乘法公式容易求得 (X, Y) 的分布律. 易知 $\{X=i, Y=j\}$ 的取值情况是: $i=1, 2, 3, 4, j$ 取不大于 i 的正整数, 且

$$P\{X=i, Y=j\} = P\{Y=j | X=i\} P\{X=i\} = \frac{1}{i} \cdot \frac{1}{4},$$

$$i=1, 2, 3, 4, j \leq i.$$

于是 (X, Y) 的分布律为

$Y \backslash X$	1	2	3	4
1	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{16}$
2	0	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{16}$
3	0	0	$\frac{1}{12}$	$\frac{1}{16}$
4	0	0	0	$\frac{1}{16}$

□

将 (X, Y) 看成一个随机点的坐标, 由图 3-2 知道离散型随机变量 X 和 Y 的联合分布函数为

$$F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p_{ij}, \quad (1.2)$$

其中和式是对一切满足 $x_i \leq x, y_j \leq y$ 的 i, j 来求和的.

与一维随机变量相似, 对于二维随机变量 (X, Y) 的分布函数 $F(x, y)$, 如果存在非负的函数 $f(x, y)$ 使对于任意 x, y 有

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

则称 (X, Y) 是连续型的二维随机变量, 函数 $f(x, y)$ 称为二维随机变量 (X, Y) 的概率密度, 或称为随机变量 X 和 Y 的联合概率密度.

按定义, 概率密度 $f(x, y)$ 具有以下性质:

$$1^\circ f(x, y) \geq 0.$$

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

3° 设 G 是 xOy 平面上的区域, 点 (X, Y) 落在 G 内的概率为

$$P\{(X, Y) \in G\} = \iint_G f(x, y) dx dy. \quad (1.3)$$

4° 若 $f(x, y)$ 在点 (x, y) 连续, 则有

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

由性质 4°, 在 $f(x, y)$ 的连续点处有

$$\begin{aligned} & \lim_{\substack{\Delta x \rightarrow 0+ \\ \Delta y \rightarrow 0+}} \frac{P\{x < X \leq x + \Delta x, y < Y \leq y + \Delta y\}}{\Delta x \Delta y} \\ & \stackrel{\text{由(1.1)}}{=} \lim_{\substack{\Delta x \rightarrow 0+ \\ \Delta y \rightarrow 0+}} \frac{1}{\Delta x \Delta y} [F(x + \Delta x, y + \Delta y) - F(x + \Delta x, y) \\ & \quad - F(x, y + \Delta y) + F(x, y)] \\ & = \frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y). \end{aligned}$$

这表示若 $f(x, y)$ 在点 (x, y) 处连续, 则当 $\Delta x, \Delta y$ 很小时

$$P\{x < X \leq x + \Delta x, y < Y \leq y + \Delta y\} \approx f(x, y) \Delta x \Delta y,$$

也就是点 (X, Y) 落在小长方形 $(x, x + \Delta x] \times (y, y + \Delta y]$ 内的概率近似地等于 $f(x, y) \Delta x \Delta y$.

在几何上 $z = f(x, y)$ 表示空间的一个曲面. 由性质 2° 知, 介于它和 xOy 平面的空间区域的体积为 1. 由性质 3°, $P\{(X, Y) \in G\}$ 的值等于以 G 为底, 以曲面 $z = f(x, y)$ 为顶面的柱体体积.

例 2 设二维随机变量 (X, Y) 具有概率密度

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

(1) 求分布函数 $F(x, y)$; (2) 求概率 $P\{Y \leq X\}$.

$$\begin{aligned} \text{解} \quad (1) \quad F(x, y) &= \int_{-\infty}^y \int_{-\infty}^x f(x, y) dx dy \\ &= \begin{cases} \int_0^y \int_0^x 2e^{-(2x+y)} dx dy, & x > 0, y > 0, \\ 0, & \text{其他.} \end{cases} \end{aligned}$$

即有
$$F(x, y) = \begin{cases} (1 - e^{-2x})(1 - e^{-y}), & x > 0, y > 0, \\ 0, & \text{其他.} \end{cases}$$

(2) 将 (X, Y) 看作是平面上随机点的坐标. 即有

$$\{Y \leq X\} = \{(X, Y) \in G\},$$

其中 G 为 xOy 平面上直线 $y = x$ 及其下方的部分, 如图 3-4. 于是

$$\begin{aligned} P\{Y \leq X\} &= P\{(X, Y) \in G\} = \iint_G f(x, y) dx dy \\ &= \int_0^\infty \int_y^\infty 2e^{-(2x+y)} dx dy = \frac{1}{3}. \quad \square \end{aligned}$$

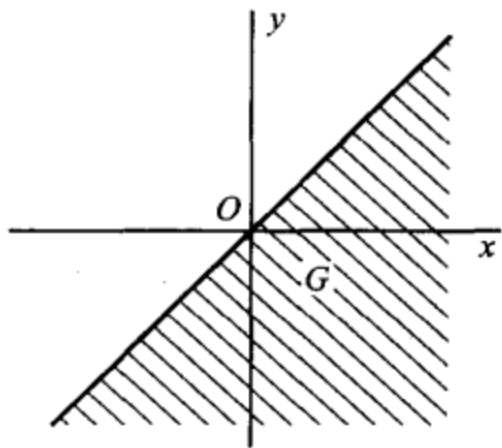


图 3-4

以上关于二维随机变量的讨论, 不难推广到 n ($n > 2$) 维随机变量的情况. 一般, 设 E 是一个随机试验, 它的样本空间是 $S = \{e\}$, 设 $X_1 = X_1(e), X_2 = X_2(e), \dots, X_n = X_n(e)$ 是定义在 S 上的随机变量, 由它们构成的一个 n 维向量 $(X_1, X_2, \dots, X_n)$ 叫做 n 维随机向量或 n 维随机变量.

对于任意 n 个实数 $x_1, x_2, \dots, x_n$, n 元函数

$$F(x_1, x_2, \dots, x_n) = P\{X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n\}$$

称为 n 维随机变量 $(X_1, X_2, \dots, X_n)$ 的分布函数或随机变量 $X_1, X_2, \dots, X_n$ 的联合分布函数. 它具有类似于二维随机变量的分布函数的性质.

§2 边缘分布

二维随机变量 (X, Y) 作为一个整体, 具有分布函数 $F(x, y)$. 而 X 和 Y 都是随机变量, 各自也有分布函数, 将它们分别记为 $F_X(x), F_Y(y)$, 依次称为二维随机变量 (X, Y) 关于 X 和关于 Y 的边缘分布函数. 边缘分布函数可以由 (X, Y) 的分布函数 $F(x, y)$ 所确定, 事实上,

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

$$\text{即} \quad F_X(x) = F(x, \infty). \quad (2.1)$$

就是说, 只要在函数 $F(x, y)$ 中令 $y \rightarrow \infty$ 就能得到 $F_X(x)$. 同理

$$F_Y(y) = F(\infty, y). \quad (2.2)$$

对于离散型随机变量, 由 (1.2)、(2.1) 式可得

$$F_X(x) = F(x, \infty) = \sum_{x_i \leq x} \sum_{j=1}^{\infty} p_{ij}.$$

与第二章 (3.2) 式比较, 知道 X 的分布律为

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

同样, Y 的分布律为

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

记

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

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

分别称 $p_{i\cdot}$ ($i=1, 2, \dots$) 和 $p_{\cdot j}$ ($j=1, 2, \dots$) 为 (X, Y) 关于 X 和关于 Y 的边缘分布律 (注意, 记号 $p_{i\cdot}$ 中的“ $\cdot$ ”表示 $p_{i\cdot}$ 是由 p_{ij} 关于 j 求和后得到的; 同样, $p_{\cdot j}$ 是由 p_{ij} 关于 i 求和后得到的).

对于连续型随机变量 (X, Y) , 设它的概率密度为 $f(x, y)$, 由于

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

由第二章(4.1)式知道, X 是一个连续型随机变量, 且其概率密度为

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy. \quad (2.3)$$

同样, Y 也是一个连续型随机变量, 其概率密度为

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx. \quad (2.4)$$

分别称 $f_X(x)$, $f_Y(y)$ 为 (X, Y) 关于 X 和关于 Y 的边缘概率密度.

例1 一整数 N 等可能地在 $1, 2, 3, \dots, 10$ 十个值中取一个值. 设 $D = D(N)$ 是能整除 N 的正整数的个数, $F = F(N)$ 是能整除 N 的素数的个数 (注意 1 不是素数). 试写出 D 和 F 的联合分布律. 并求边缘分布律.

解 先将试验的样本空间及 D, F 取值的情况列出如下:

样本点	1	2	3	4	5	6	7	8	9	10
D	1	2	2	3	2	4	2	4	3	4
F	0	1	1	1	1	2	1	1	1	2

D 所有可能取的值为 $1, 2, 3, 4$; F 所有可能取的值为 $0, 1, 2$. 容易得到 (D, F) 取 (i, j) , $i=1, 2, 3, 4$, $j=0, 1, 2$ 的概率, 例如

$$P\{D = 1, F = 0\} = \frac{1}{10}, \quad P\{D = 2, F = 1\} = \frac{4}{10},$$

可得 D 和 F 的联合分布律及边缘分布律如下表所示:

$F \backslash D$	1	2	3	4	$P\{F=j\}$
0	$\frac{1}{10}$	0	0	0	$\frac{1}{10}$
1	0	$\frac{4}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{7}{10}$
2	0	0	0	$\frac{2}{10}$	$\frac{2}{10}$
$P\{D=i\}$	$\frac{1}{10}$	$\frac{4}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	1

即有边缘分布律

D	1	2	3	4
p_k	$\frac{1}{10}$	$\frac{4}{10}$	$\frac{2}{10}$	$\frac{3}{10}$

F	0	1	2
p_k	$\frac{1}{10}$	$\frac{7}{10}$	$\frac{2}{10}$

□

我们常常将边缘分布律写在联合分布律表格的边缘上如上表所示. 这就是“边缘分布律”这个名词的来源.

例 2 设随机变量 X 和 Y 具有联合概率密度 (图 3-5)

$$f(x, y) = \begin{cases} 6, & x^2 \leq y \leq x, \\ 0, & \text{其他.} \end{cases}$$

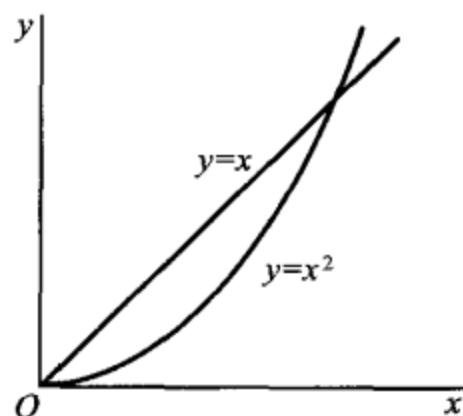


图 3-5

求边缘概率密度 $f_X(x)$, $f_Y(y)$.

解

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \begin{cases} \int_{x^2}^x 6 dy = 6(x - x^2), & 0 \leq x \leq 1, \\ 0, & \text{其他.} \end{cases}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \begin{cases} \int_y^{\sqrt{y}} 6 dx = 6(\sqrt{y} - y), & 0 \leq y \leq 1, \\ 0, & \text{其他.} \end{cases}$$

□

例 3 设二维随机变量 (X, Y) 的概率密度为

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

其中 $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ 都是常数, 且 $\sigma_1 > 0, \sigma_2 > 0, -1 < \rho < 1$. 我们称 (X, Y) 为服从参数为 $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ 的二维正态分布(这五个参数的意义将在下一章说明), 记为 $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. 试求二维正态随机变量的边缘概率密度.

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

由于
$$\begin{aligned} & \frac{(y - \mu_2)^2}{\sigma_2^2} - 2\rho \frac{(x - \mu_1)(y - \mu_2)}{\sigma_1 \sigma_2} \\ &= \left(\frac{y - \mu_2}{\sigma_2} - \rho \frac{x - \mu_1}{\sigma_1} \right)^2 - \rho^2 \frac{(x - \mu_1)^2}{\sigma_1^2}, \end{aligned}$$

于是

$$f_X(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{y-\mu_2}{\sigma_2} - \rho\frac{x-\mu_1}{\sigma_1}\right)^2} dy.$$

令 $t = \frac{1}{\sqrt{1-\rho^2}} \left(\frac{y-\mu_2}{\sigma_2} - \rho \frac{x-\mu_1}{\sigma_1} \right)$, 则有

$$f_X(x) = \frac{1}{2\pi\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \int_{-\infty}^{\infty} e^{-t^2/2} dt,$$

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

同理

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}, \quad -\infty < y < \infty. \quad \square$$

我们看到二维正态分布的两个边缘分布都是一维正态分布, 并且都不依赖于参数 ρ , 亦即对于给定的 $\mu_1, \mu_2, \sigma_1, \sigma_2$, 不同的 ρ 对应不同的二维正态分布, 它们的边缘分布却都是一样的. 这一事实表明, 单由关于 X 和关于 Y 的边缘分布, 一般来说是不能确定随机变量 X 和 Y 的联合分布的.

§3 条件分布

我们由条件概率很自然地引出条件概率分布的概念.

设 (X, Y) 是二维离散型随机变量, 其分布律为

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

(X, Y) 关于 X 和关于 Y 的边缘分布律分别为

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

$$P\{Y=y_j\} = p_{\cdot j} = \sum_{i=1}^{\infty} p_{ij}, \quad j=1, 2, \dots.$$

设 $p_{\cdot j} > 0$, 我们来考虑在事件 $\{Y = y_j\}$ 已发生的条件下事件 $\{X = x_i\}$ 发生的概率, 也就是来求事件

$$\{X = x_i | Y = y_j\}, \quad i = 1, 2, \dots$$

的概率. 由条件概率公式, 可得

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

易知上述条件概率具有分布律的性质:

$$1^\circ P\{X = x_i | Y = y_j\} \geq 0;$$

$$2^\circ \sum_{i=1}^{\infty} P\{X = x_i | Y = y_j\} = \sum_{i=1}^{\infty} \frac{p_{ij}}{p_{\cdot j}} = \frac{1}{p_{\cdot j}} \sum_{i=1}^{\infty} p_{ij} = \frac{p_{\cdot j}}{p_{\cdot j}} = 1.$$

于是我们引入以下的定义.

定义 设 (X, Y) 是二维离散型随机变量, 对于固定的 j , 若 $P\{Y = y_j\} > 0$, 则称

$$P\{X = x_i | Y = y_j\} = \frac{P\{X = x_i, Y = y_j\}}{P\{Y = y_j\}} = \frac{p_{ij}}{p_{\cdot j}}, \quad i = 1, 2, \dots \quad (3.1)$$

为在 $Y = y_j$ 条件下随机变量 X 的条件分布律.

同样, 对于固定的 i , 若 $P\{X = x_i\} > 0$, 则称

$$P\{Y = y_j | X = x_i\} = \frac{P\{X = x_i, Y = y_j\}}{P\{X = x_i\}} = \frac{p_{ij}}{p_{i\cdot}}, \quad j = 1, 2, \dots \quad (3.2)$$

为在 $X = x_i$ 条件下随机变量 Y 的条件分布律.

例 1 在一汽车工厂中, 一辆汽车有两道工序是由机器人完成的. 其一是紧固 3 只螺栓, 其二是焊接 2 处焊点. 以 X 表示由机器人紧固的螺栓紧固得不良的数目, 以 Y 表示由机器人焊接的不良焊点的数目. 据积累的资料知 (X, Y) 具有分布律:

$\begin{matrix} X \\ Y \end{matrix}$	0	1	2	3	$P\{Y = j\}$
0	0.840	0.030	0.020	0.010	0.900
1	0.060	0.010	0.008	0.002	0.080
2	0.010	0.005	0.004	0.001	0.020
$P\{X = i\}$	0.910	0.045	0.032	0.013	1.000

(1) 求在 $X = 1$ 的条件下, Y 的条件分布律; (2) 求在 $Y = 0$ 的条件下, X 的条件分布律.

解 边缘分布律已经求出列在上表中. 在 $X = 1$ 的条件下, Y 的条件分布律为

$$P\{Y = 0 | X = 1\} = \frac{P\{X = 1, Y = 0\}}{P\{X = 1\}} = \frac{0.030}{0.045},$$

$$P\{Y=1|X=1\} = \frac{P\{X=1, Y=1\}}{P\{X=1\}} = \frac{0.010}{0.045},$$

$$P\{Y=2|X=1\} = \frac{P\{X=1, Y=2\}}{P\{X=1\}} = \frac{0.005}{0.045},$$

或写成

$Y=k$	0	1	2
$P\{Y=k X=1\}$	$\frac{6}{9}$	$\frac{2}{9}$	$\frac{1}{9}$

同样可得在 $Y=0$ 的条件下 X 的条件分布律为

$X=k$	0	1	2	3
$P\{X=k Y=0\}$	$\frac{84}{90}$	$\frac{3}{90}$	$\frac{2}{90}$	$\frac{1}{90}$

□

例 2 一射手进行射击, 击中目标的概率为 p ($0 < p < 1$), 射击直至击中目标两次为止. 设以 X 表示首次击中目标所进行的射击次数, 以 Y 表示总共进行的射击次数, 试求 X 和 Y 的联合分布律及条件分布律.

解 按题意 $Y=n$ 就表示在第 n 次射击时击中目标, 且在第 1 次, 第 2 次, ……第 $n-1$ 次射击中恰有一次击中目标. 已知各次射击是相互独立的, 于是不管 m ($m < n$) 是多少, 概率 $P\{X=m, Y=n\}$ 都应等于

$$p \cdot p \cdot \underbrace{q \cdot q \cdots q}_{n-2 \uparrow} = p^2 q^{n-2} \quad (\text{这里 } q=1-p).$$

即得 X 和 Y 的联合分布律为

$$P\{X=m, Y=n\} = p^2 q^{n-2}, n=2, 3, \cdots; m=1, 2, \cdots, n-1.$$

$$\begin{aligned} \text{又} \quad P\{X=m\} &= \sum_{n=m+1}^{\infty} P\{X=m, Y=n\} = \sum_{n=m+1}^{\infty} p^2 q^{n-2} \\ &= p^2 \sum_{n=m+1}^{\infty} q^{n-2} = \frac{p^2 q^{m-1}}{1-q} = pq^{m-1}, m=1, 2, \cdots, \end{aligned}$$

$$\begin{aligned} P\{Y=n\} &= \sum_{m=1}^{n-1} P\{X=m, Y=n\} \\ &= \sum_{m=1}^{n-1} p^2 q^{n-2} = (n-1)p^2 q^{n-2}, \quad n=2, 3, \cdots. \end{aligned}$$

于是由 (3.1), (3.2) 式得到所求的条件分布律为

当 $n=2, 3, \cdots$ 时,

$$P\{X=m|Y=n\} = \frac{p^2 q^{n-2}}{(n-1)p^2 q^{n-2}} = \frac{1}{n-1}, m=1, 2, \dots, n-1;$$

当 $m=1, 2, \dots$ 时,

$$P\{Y=n|X=m\} = \frac{p^2 q^{n-2}}{p q^{m-1}} = p q^{n-m-1}, n=m+1, m+2, \dots.$$

$$\text{例如, } P\{X=m|Y=3\} = \frac{1}{2}, \quad m=1, 2;$$

$$P\{Y=n|X=3\} = p q^{n-4}, \quad n=4, 5, \dots. \quad \square$$

现设 (X, Y) 是二维连续型随机变量, 这时由于对任意 x, y 有 $P\{X=x\}=0$, $P\{Y=y\}=0$, 因此就不能直接用条件概率公式引入“条件分布函数”了.

设 (X, Y) 的概率密度为 $f(x, y)$, (X, Y) 关于 Y 的边缘概率密度为 $f_Y(y)$. 给定 y , 对于任意固定的 $\epsilon > 0$, 对于任意 x , 考虑条件概率

$$P\{X \leq x | y < Y \leq y + \epsilon\},$$

设 $P\{y < Y \leq y + \epsilon\} > 0$, 则有

$$\begin{aligned} P\{X \leq x | y < Y \leq y + \epsilon\} &= \frac{P\{X \leq x, y < Y \leq y + \epsilon\}}{P\{y < Y \leq y + \epsilon\}} \\ &= \frac{\int_{-\infty}^x \left[\int_y^{y+\epsilon} f(x, y) dy \right] dx}{\int_y^{y+\epsilon} f_Y(y) dy}. \end{aligned}$$

在某些条件下, 当 ϵ 很小时, 上式右端分子、分母分别近似于 $\epsilon \int_{-\infty}^x f(x, y) dx$ 和 $\epsilon f_Y(y)$, 于是当 ϵ 很小时, 有

$$P\{X \leq x | y < Y \leq y + \epsilon\} \approx \frac{\epsilon \int_{-\infty}^x f(x, y) dx}{\epsilon f_Y(y)} = \int_{-\infty}^x \frac{f(x, y)}{f_Y(y)} dx. \quad (3.3)$$

与一维随机变量概率密度的定义式第二章(4.1)式比较. 我们给出以下的定义.

定义 设二维随机变量 (X, Y) 的概率密度为 $f(x, y)$, (X, Y) 关于 Y 的边缘概率密度为 $f_Y(y)$. 若对于固定的 y , $f_Y(y) > 0$, 则称 $\frac{f(x, y)}{f_Y(y)}$ 为在 $Y=y$ 的条件下 X 的条件概率密度, 记为①

$$f_{X|Y}(x | y) = \frac{f(x, y)}{f_Y(y)}. \quad (3.4)$$

① 条件概率密度满足条件: $f_{X|Y}(x | y) = \frac{f(x, y)}{f_Y(y)} \geq 0$;

$$\int_{-\infty}^{\infty} f_{X|Y}(x | y) dx = \int_{-\infty}^{\infty} \frac{f(x, y)}{f_Y(y)} dx = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f(x, y) dx = 1.$$

称 $\int_{-\infty}^x f_{X|Y}(x|y)dx = \int_{-\infty}^x \frac{f(x,y)}{f_Y(y)}dx$ 为在 $Y=y$ 的条件下 X 的条件分布函数, 记为 $P\{X \leq x | Y=y\}$ 或 $F_{X|Y}(x|y)$, 即

$$F_{X|Y}(x|y) = P\{X \leq x | Y=y\} = \int_{-\infty}^x \frac{f(x,y)}{f_Y(y)}dx. \quad (3.5)$$

类似地, 可以定义 $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$ 和 $F_{Y|X}(y|x) = \int_{-\infty}^y \frac{f(x,y)}{f_X(x)}dy$.

由(3.3)知道, 当 ϵ 很小时, 有

$$P\{X \leq x | y < Y \leq y + \epsilon\} \approx \int_{-\infty}^x f_{X|Y}(x|y)dx = F_{X|Y}(x|y),$$

上式说明了条件密度和条件分布函数的含义.

例3 设 G 是平面上的有界区域, 其面积为 A . 若二维随机变量 (X, Y) 具有概率密度

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

则称 (X, Y) 在 G 上服从均匀分布. 现设二维随机变量 (X, Y) 在圆域 $x^2 + y^2 \leq 1$ 上服从均匀分布, 求条件概率密度 $f_{X|Y}(x|y)$.

解 由假设随机变量 (X, Y) 具有概率密度

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

且有边缘概率密度

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)dx \\ &= \begin{cases} \frac{1}{\pi} \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} dx = \frac{2}{\pi} \sqrt{1-y^2}, & -1 \leq y \leq 1, \\ 0, & \text{其他}. \end{cases} \end{aligned}$$

于是当 $-1 < y < 1$ 时有

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

当 $y=0$ 和 $y=\frac{1}{2}$ 时 $f_{X|Y}(x|y)$ 的图形分别如图 3-6, 图 3-7 所示.

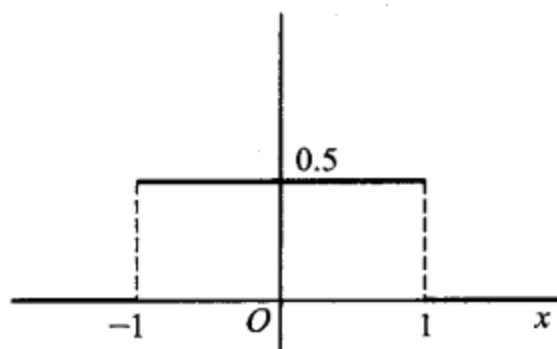


图 3-6

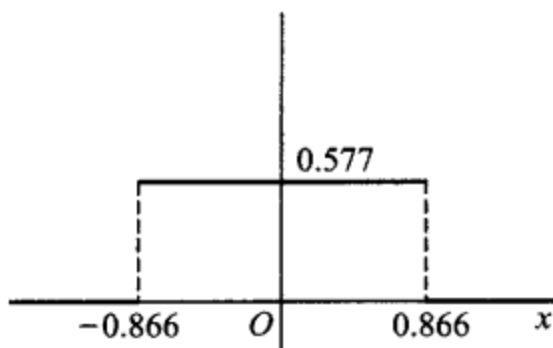


图 3-7

例 4 设数 X 在区间 $(0, 1)$ 上随机地取值, 当观察到 $X=x$ ($0 < x < 1$) 时, 数 Y 在区间 $(x, 1)$ 上随机地取值. 求 Y 的概率密度 $f_Y(y)$.

解 按题意 X 具有概率密度

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

对于任意给定的值 x ($0 < x < 1$), 在 $X=x$ 的条件下 Y 的条件概率密度为

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{1-x}, & x < y < 1, \\ 0, & \text{其他.} \end{cases}$$

由 (3.4) 式得 X 和 Y 的联合概率密度为

$$f(x, y) = f_{Y|X}(y|x)f_X(x) = \begin{cases} \frac{1}{1-x}, & 0 < x < y < 1, \\ 0, & \text{其他.} \end{cases}$$

于是得关于 Y 的边缘概率密度为

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x, y) dx \\ &= \begin{cases} \int_0^y \frac{1}{1-x} dx = -\ln(1-y), & 0 < y < 1, \\ 0, & \text{其他.} \end{cases} \end{aligned}$$

§ 4 相互独立的随机变量

本节我们将利用两个事件相互独立的概念引出两个随机变量相互独立的概念, 这是一个十分重要的概念.

定义 设 $F(x, y)$ 及 $F_X(x), F_Y(y)$ 分别是二维随机变量 (X, Y) 的分布函数及边缘分布函数. 若对于所有 x, y 有

$$P\{X \leq x, Y \leq y\} = P\{X \leq x\}P\{Y \leq y\}, \quad (4.1)$$

$$\text{即} \quad F(x, y) = F_X(x)F_Y(y), \quad (4.2)$$

则称随机变量 X 和 Y 是相互独立的.

设 (X, Y) 是连续型随机变量, $f(x, y), f_X(x), f_Y(y)$ 分别为 (X, Y) 的概率密度和边缘概率密度, 则 X 和 Y 相互独立的条件(4.2)等价于: 等式

$$f(x, y) = f_X(x)f_Y(y) \quad (4.3)$$

在平面上几乎处处^①成立.

当 (X, Y) 是离散型随机变量时, X 和 Y 相互独立的条件(4.2)式等价于: 对于 (X, Y) 的所有可能取的值 (x_i, y_j) 有

$$P\{X = x_i, Y = y_j\} = P\{X = x_i\}P\{Y = y_j\}. \quad (4.4)$$

在实际中使用(4.3)式或(4.4)式要比使用(4.2)式方便.

例如 § 1 例 2 中的随机变量 X 和 Y , 由于

$$f_X(x) = \begin{cases} 2e^{-2x}, & x > 0, \\ 0, & \text{其他}, \end{cases} \quad f_Y(y) = \begin{cases} e^{-y}, & y > 0, \\ 0, & \text{其他}, \end{cases}$$

故有 $f(x, y) = f_X(x)f_Y(y)$, 因而 X, Y 是相互独立的.

又如, 若 X, Y 具有联合分布律

Y \ X	X		$P\{Y=j\}$
	0	1	
1	1/6	2/6	1/2
2	1/6	2/6	1/2
$P\{x=i\}$	1/3	2/3	1

则有 $P\{X=0, Y=1\} = 1/6 = P\{X=0\}P\{Y=1\},$

$$P\{X=0, Y=2\} = 1/6 = P\{X=0\}P\{Y=2\},$$

$$P\{X=1, Y=1\} = 2/6 = P\{X=1\}P\{Y=1\},$$

$$P\{X=1, Y=2\} = 2/6 = P\{X=1\}P\{Y=2\},$$

因而 X, Y 是相互独立的.

再如 § 2 例 1 中的随机变量 F 和 D , 由于 $P\{D=1, F=0\} = 1/10 \neq P\{D=1\} \times P\{F=0\}$, 因而 F 和 D 不是相互独立的.

下面考察二维正态随机变量 (X, Y) . 它的概率密度为

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

① 此处“几乎处处成立”的含义是: 在平面上除去“面积”为零的集合以外, 处处成立.

由 § 2 中例 3 知道, 其边缘概率密度 $f_X(x), f_Y(y)$ 的乘积为

$$f_X(x)f_Y(y) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\}.$$

因此, 如果 $\rho=0$, 则对于所有 x, y 有 $f(x, y) = f_X(x)f_Y(y)$, 即 X 和 Y 相互独立. 反之, 如果 X 和 Y 相互独立, 由于 $f(x, y), f_X(x), f_Y(y)$ 都是连续函数, 故对于所有的 x, y 有 $f(x, y) = f_X(x)f_Y(y)$. 特别, 令 $x=\mu_1, y=\mu_2$, 自这一等式得到

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} = \frac{1}{2\pi\sigma_1\sigma_2},$$

从而 $\rho=0$. 综上所述, 得到以下的结论:

对于二维正态随机变量 (X, Y) , X 和 Y 相互独立的充要条件是参数 $\rho=0$.

例 一负责人到达办公室的时间均匀分布在 8~12 时, 他的秘书到达办公室的时间均匀分布在 7~9 时, 设他们两人到达的时间相互独立, 求他们到达办公室的时间相差不超过 5 分钟(1/12 小时)的概率.

解 设 X 和 Y 分别是负责人和他的秘书到达办公室的时间, 由假设 X 和 Y 的概率密度分别为

$$f_X(x) = \begin{cases} \frac{1}{4}, & 8 < x < 12, \\ 0, & \text{其他}, \end{cases} \quad f_Y(y) = \begin{cases} \frac{1}{2}, & 7 < y < 9, \\ 0, & \text{其他}, \end{cases}$$

因为 X, Y 相互独立, 故 (X, Y) 的概率密度为

$$f(x, y) = f_X(x)f_Y(y) = \begin{cases} \frac{1}{8}, & 8 < x < 12, 7 < y < 9, \\ 0, & \text{其他}. \end{cases}$$

按题意需要求概率 $P\{|X-Y| \leq 1/12\}$. 画出区域: $|x-y| \leq 1/12$, 以及长方形 $[8 < x < 12; 7 < y < 9]$, 它们的公共部分是四边形 $BCC'B'$, 记为 G (如图 3-8). 显然仅当 (X, Y) 取值于 G 内, 他们两人到达的时间相差不超过 1/12 小时. 因此, 所求的概率为

$$\begin{aligned} P\{|X-Y| \leq \frac{1}{12}\} &= \iint_G f(x, y) dx dy \\ &= \frac{1}{8} \times (G \text{ 的面积}). \end{aligned}$$

而 G 的面积 = 三角形 ABC 的面积 - 三角形 $AB'C'$ 的面积

$$= \frac{1}{2} \left(\frac{13}{12}\right)^2 - \frac{1}{2} \left(\frac{11}{12}\right)^2 = \frac{1}{6}.$$

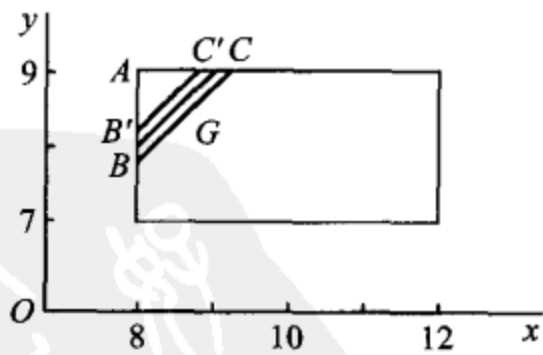


图 3-8

于是
$$P\{|X-Y|\leq \frac{1}{12}\} = \frac{1}{48}.$$

即负责人和他的秘书到达办公室的时间相差不超过 5 分钟的概率为 $1/48$. $\square$

以上所述关于二维随机变量的一些概念,容易推广到 n 维随机变量的情况.

上面说过, n 维随机变量 $(X_1, X_2, \dots, X_n)$ 的分布函数定义为

$$F(x_1, x_2, \dots, x_n) = P\{X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n\},$$

其中 $x_1, x_2, \dots, x_n$ 为任意实数.

若存在非负函数 $f(x_1, x_2, \dots, x_n)$, 使对于任意实数 $x_1, x_2, \dots, x_n$ 有

$$\begin{aligned} F(x_1, x_2, \dots, x_n) \\ = \int_{-\infty}^{x_n} \int_{-\infty}^{x_{n-1}} \cdots \int_{-\infty}^{x_1} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n, \end{aligned}$$

则称 $f(x_1, x_2, \dots, x_n)$ 为 $(X_1, X_2, \dots, X_n)$ 的概率密度函数.

设 $(X_1, X_2, \dots, X_n)$ 的分布函数 $F(x_1, x_2, \dots, x_n)$ 为已知, 则 $(X_1, X_2, \dots, X_n)$ 的 k ($1 \leq k < n$) 维边缘分布函数就随之确定. 例如 $(X_1, X_2, \dots, X_n)$ 关于 X_1 、关于 (X_1, X_2) 的边缘分布函数分别为

$$F_{X_1}(x_1) = F(x_1, \infty, \infty, \dots, \infty),$$

$$F_{X_1, X_2}(x_1, x_2) = F(x_1, x_2, \infty, \infty, \dots, \infty).$$

又若 $f(x_1, x_2, \dots, x_n)$ 是 $(X_1, X_2, \dots, X_n)$ 的概率密度, 则 $(X_1, X_2, \dots, X_n)$ 关于 X_1 、关于 (X_1, X_2) 的边缘概率密度分别为

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_2 dx_3 \cdots dx_n,$$

$$f_{X_1, X_2}(x_1, x_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_3 dx_4 \cdots dx_n.$$

若对于所有的 $x_1, x_2, \dots, x_n$ 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 $X_1, X_2, \dots, X_n$ 是相互独立的.

若对于所有的 $x_1, x_2, \dots, x_m; y_1, y_2, \dots, y_n$ 有

$$F(x_1, x_2, \dots, x_m, y_1, y_2, \dots, y_n) = F_1(x_1, x_2, \dots, x_m) F_2(y_1, y_2, \dots, y_n),$$

其中 F_1, F_2, F 依次为随机变量 $(X_1, X_2, \dots, X_m)$, $(Y_1, Y_2, \dots, Y_n)$ 和 $(X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n)$ 的分布函数, 则称随机变量 $(X_1, X_2, \dots, X_m)$ 和 $(Y_1, Y_2, \dots, Y_n)$ 是相互独立的.

我们有以下的定理, 它在数理统计中是很有用的.

定理 设 $(X_1, X_2, \dots, X_m)$ 和 $(Y_1, Y_2, \dots, Y_n)$ 相互独立, 则 X_i ($i=1, 2, \dots, m$) 和 Y_j ($j=1, 2, \dots, n$) 相互独立. 又若 h, g 是连续函数, 则 $h(X_1, X_2, \dots, X_m)$ 和 $g(Y_1, Y_2, \dots, Y_n)$ 相互独立.

(证明略)

§5 两个随机变量的函数的分布

上一章 §5 中已经讨论过一个随机变量的函数的分布, 本节讨论两个随机变量的函数的分布. 我们只就下面几个具体的函数来讨论.

(一) $Z=X+Y$ 的分布

设 (X, Y) 是二维连续型随机变量, 它具有概率密度 $f(x, y)$. 则 $Z=X+Y$ 仍为连续型随机变量, 其概率密度为

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f(z-y, y) dy, \quad (5.1)$$

或

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f(x, z-x) dx. \quad (5.2)$$

又若 X 和 Y 相互独立, 设 (X, Y) 关于 X, Y 的边缘密度分别为 $f_X(x)$, $f_Y(y)$, 则 (5.1), (5.2) 分别化为

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy \quad (5.3)$$

和

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx. \quad (5.4)$$

这两个公式称为 f_X 和 f_Y 的卷积公式, 记为 $f_X * f_Y$, 即

$$f_X * f_Y = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx.$$

证 先来求 $Z=X+Y$ 的分布函数 $F_Z(z)$, 即有

$$F_Z(z) = P\{Z \leq z\} = \iint_{x+y \leq z} f(x, y) dx dy,$$

这里积分区域 $G: x+y \leq z$ 是直线 $x+y=z$ 及其左下方的半平面 (如图 3-9). 将二重积分化成累次积分, 得

$$F_Z(z) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z-y} f(x, y) dx \right] dy.$$

固定 z 和 y 对积分 $\int_{-\infty}^{z-y} f(x, y) dx$ 作变量变换, 令 $x = u - y$, 得

$$\int_{-\infty}^{z-y} f(x, y) dx = \int_{-\infty}^z f(u-y, y) du.$$

于是

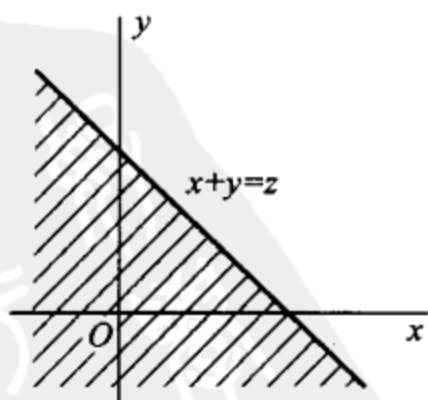


图 3-9

$$F_Z(z) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^z f(u-y, y) du \right] dy = \int_{-\infty}^z \left[\int_{-\infty}^{\infty} f(u-y, y) dy \right] du.$$

由概率密度的定义即得(5.1)式. 类似可证得(5.2)式. $\square$

例1 设 X 和 Y 是两个相互独立的随机变量. 它们都服从 $N(0, 1)$ 分布, 其概率密度为

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

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

求 $Z = X + Y$ 的概率密度.

解 由(5.4)式

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} \cdot e^{-\frac{(z-x)^2}{2}} dx = \frac{1}{2\pi} e^{-\frac{z^2}{4}} \int_{-\infty}^{\infty} e^{-(x-\frac{z}{2})^2} dx, \end{aligned}$$

令 $t = x - \frac{z}{2}$, 得

$$f_Z(z) = \frac{1}{2\pi} e^{-\frac{z^2}{4}} \int_{-\infty}^{\infty} e^{-t^2} dt = \frac{1}{2\pi} e^{-\frac{z^2}{4}} \sqrt{\pi} = \frac{1}{2\sqrt{\pi}} e^{-\frac{z^2}{4}}.$$

即 Z 服从 $N(0, 2)$ 分布.

一般, 设 X, Y 相互独立且 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$. 由(5.4)式经过计算知 $Z = X + Y$ 仍然服从正态分布, 且有 $Z \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. 这个结论还能推广到 n 个独立正态随机变量之和的情况. 即若 $X_i \sim N(\mu_i, \sigma_i^2) (i=1, 2, \dots, n)$, 且它们相互独立, 则它们的和 $Z = X_1 + X_2 + \dots + X_n$ 仍然服从正态分布, 且有 $Z \sim N(\mu_1 + \mu_2 + \dots + \mu_n, \sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2)$.

更一般地, 可以证明有限个相互独立的正态随机变量的线性组合仍然服从正态分布. $\square$

例2 在一简单电路中, 两电阻 R_1 和 R_2 串联连接, 设 R_1, R_2 相互独立, 它们的概率密度均为

$$f(x) = \begin{cases} \frac{10-x}{50}, & 0 \leq x \leq 10, \\ 0, & \text{其他.} \end{cases}$$

求总电阻 $R = R_1 + R_2$ 的概率密度.

解 由(5.4)式, R 的概率密度为

$$f_R(z) = \int_{-\infty}^{\infty} f(x) f(z-x) dx.$$

易知仅当

$$\begin{cases} 0 < x < 10, \\ 0 < z-x < 10, \end{cases} \quad \text{即} \quad \begin{cases} 0 < x < 10, \\ z-10 < x < z \end{cases}$$

时上述积分的被积函数不等于零. 参考图 3-10, 即得

$$f_R(z) = \begin{cases} \int_0^z f(x)f(z-x)dx, & 0 \leq z < 10, \\ \int_{z-10}^{10} f(x)f(z-x)dx, & 10 \leq z \leq 20, \\ 0, & \text{其他.} \end{cases}$$

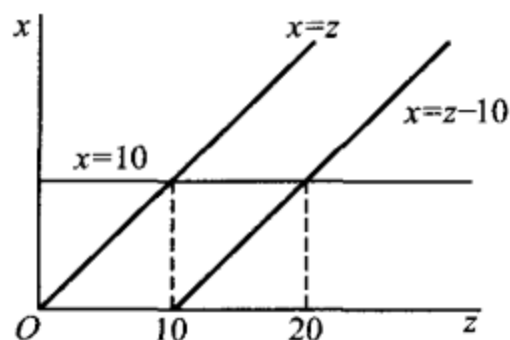


图 3-10

将 $f(x)$ 的表达式代入上式得

$$f_R(z) = \begin{cases} \frac{1}{15\,000}(600z - 60z^2 + z^3), & 0 \leq z < 10, \\ \frac{1}{15\,000}(20-z)^3, & 10 \leq z < 20, \\ 0, & \text{其他.} \end{cases}$$

□

例 3 设随机变量 X, Y 相互独立, 且分别服从参数为 $\alpha, \theta; \beta, \theta$ 的 Γ 分布 (分别记成 $X \sim \Gamma(\alpha, \theta), Y \sim \Gamma(\beta, \theta)$). X, Y 的概率密度分别为

$$f_X(x) = \begin{cases} \frac{1}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\theta}, & x > 0, \\ 0, & \text{其他,} \end{cases} \quad \alpha > 0, \theta > 0.$$

$$f_Y(y) = \begin{cases} \frac{1}{\theta^\beta \Gamma(\beta)} y^{\beta-1} e^{-y/\theta}, & y > 0, \\ 0, & \text{其他,} \end{cases} \quad \beta > 0, \theta > 0.$$

试证明 $Z = X + Y$ 服从参数为 $\alpha + \beta, \theta$ 的 Γ 分布, 即 $X + Y \sim \Gamma(\alpha + \beta, \theta)$.

证 由 (5.4) 式 $Z = X + Y$ 的概率密度为

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx.$$

易知仅当

$$\begin{cases} x > 0, \\ z-x > 0, \end{cases} \quad \text{亦即} \quad \begin{cases} x > 0, \\ x < z. \end{cases}$$

时上述积分的被积函数不等于零, 于是 (参见图 3-11) 知

当 $z < 0$ 时 $f_Z(z) = 0$, 而当 $z > 0$ 时有

$$f_Z(z) = \int_0^z \frac{1}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\theta} \frac{1}{\theta^\beta \Gamma(\beta)} (z-x)^{\beta-1} e^{-(z-x)/\theta} dx$$

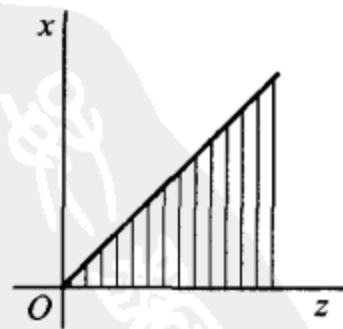


图 3-11

$$\begin{aligned}
 &= \frac{e^{-z/\theta}}{\theta^{\alpha+\beta}\Gamma(\alpha)\Gamma(\beta)} \int_0^z x^{\alpha-1}(z-x)^{\beta-1} dx \quad (\text{令 } x=zt) \\
 &= \frac{z^{\alpha+\beta-1}e^{-z/\theta}}{\theta^{\alpha+\beta}\Gamma(\alpha)\Gamma(\beta)} \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt \stackrel{\text{记成}}{=} Az^{\alpha+\beta-1}e^{-z/\theta},
 \end{aligned}$$

其中 $A = \frac{1}{\theta^{\alpha+\beta}\Gamma(\alpha)\Gamma(\beta)} \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt.$ (5.5)①

现在来计算 A . 由概率密度的性质得到:

$$\begin{aligned}
 1 &= \int_{-\infty}^{\infty} f_Z(z) dz = \int_0^{\infty} Az^{\alpha+\beta-1}e^{-z/\theta} dz \\
 &= A\theta^{\alpha+\beta} \int_0^{\infty} (z/\theta)^{\alpha+\beta-1} e^{-z/\theta} d(z/\theta) \\
 &= A\theta^{\alpha+\beta}\Gamma(\alpha+\beta),
 \end{aligned}$$

即有 $A = \frac{1}{\theta^{\alpha+\beta}\Gamma(\alpha+\beta)}.$ (5.6)

于是 $f_Z(z) = \begin{cases} \frac{1}{\theta^{\alpha+\beta}\Gamma(\alpha+\beta)} z^{\alpha+\beta-1} e^{-z/\theta}, & z > 0, \\ 0, & \text{其他.} \end{cases}$

即 $X+Y \sim \Gamma(\alpha+\beta, \theta).$ □

上述结论还能推广到 n 个相互独立的 Γ 分布变量之和的情况. 即若 $X_1, X_2, \dots, X_n$ 相互独立, 且 X_i 服从参数为 α_i, β ($i=1, 2, \dots, n$) 的 Γ 分布, 则 $\sum_{i=1}^n X_i$ 服从参数为 $\sum_{i=1}^n \alpha_i, \beta$ 的 Γ 分布. 这一性质称为 Γ 分布的可加性.

(二) $Z = \frac{Y}{X}$ 的分布、 $Z = XY$ 的分布

设 (X, Y) 是二维连续型随机变量, 它具有概率密度 $f(x, y)$, 则 $Z = \frac{Y}{X}$ 、 $Z = XY$ 仍为连续型随机变量, 其概率密度分别为

$$f_{Y/X}(z) = \int_{-\infty}^{\infty} |x| f(x, xz) dx, \quad (5.7)$$

$$f_{XY}(z) = \int_{-\infty}^{\infty} \frac{1}{|x|} f(x, \frac{z}{x}) dx. \quad (5.8)$$

① (5.5) 式中的积分

$$\int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt \stackrel{\text{记成}}{=} B(\alpha, \beta), \quad \alpha, \beta > 0,$$

称为 Beta 函数. 由 (5.5), (5.6) 式知 Beta 函数与 Γ 函数有如下关系

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

又若 X 和 Y 相互独立. 设 (X, Y) 关于 X, Y 的边缘密度分别为 $f_X(x)$, $f_Y(y)$, 则(5.7)式化为

$$f_{Y/X}(z) = \int_{-\infty}^{\infty} |x| f_X(x) f_Y(xz) dx. \quad (5.9)$$

而(5.8)式化为

$$f_{XY}(z) = \int_{-\infty}^{\infty} \frac{1}{|x|} f_X(x) f_Y\left(\frac{z}{x}\right) dx. \quad (5.10)$$

证 $Z=Y/X$ 的分布函数为(如图 3-12)

$$\begin{aligned} F_{Y/X}(z) &= P\{Y/X \leq z\} = \iint_{G_1 \cup G_2} f(x, y) dx dy \\ &= \iint_{y/x \leq z, x < 0} f(x, y) dy dx + \iint_{y/x \leq z, x > 0} f(x, y) dy dx \\ &= \int_{-\infty}^0 \left[\int_{zx}^{\infty} f(x, y) dy \right] dx + \int_0^{\infty} \left[\int_{-\infty}^{zx} f(x, y) dy \right] dx \\ &\stackrel{\text{令 } y=xu}{=} \int_{-\infty}^0 \left[\int_z^{\infty} x f(x, xu) du \right] dx + \int_0^{\infty} \left[\int_{-\infty}^z x f(x, xu) du \right] dx \\ &= \int_{-\infty}^0 \left[\int_{-\infty}^z (-x) f(x, xu) du \right] dx + \int_0^{\infty} \left[\int_{-\infty}^z x f(x, xu) du \right] dx \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^z |x| f(x, xu) du \right] dx \\ &= \int_{-\infty}^z \left[\int_{-\infty}^{\infty} |x| f(x, xu) dx \right] du, \end{aligned}$$

由概率密度的定义即得(5.7)式. □

类似地, 可求出 $f_{XY}(z)$ 的概率密度为(5.8)式.

例 4 某公司提供一种地震保险, 保险费 Y 的概率密度为

$$f(y) = \begin{cases} \frac{y}{25} e^{-y/5}, & y > 0, \\ 0, & \text{其他.} \end{cases}$$

保险赔付 X 的概率密度为

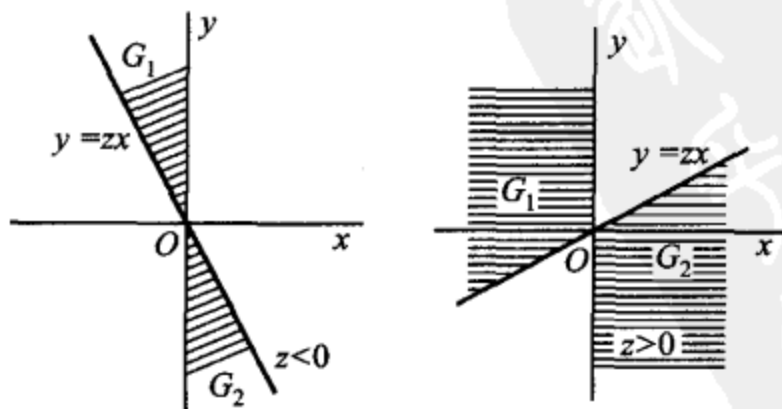


图 3-12

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

设 X 与 Y 相互独立, 求 $Z=Y/X$ 的概率密度.

解 由(5.7)式知, 当 $z < 0$ 时, $f_z(z) = 0$. 当 $z > 0$ 时, Z 的概率密度为

$$\begin{aligned} f_z(z) &= \int_0^{\infty} x \cdot \frac{1}{5}e^{-x/5} \cdot \frac{xz}{25}e^{-xz/5} dx = \frac{z}{125} \int_0^{\infty} x^2 e^{-x(\frac{1+z}{5})} dx \\ &= \frac{z}{125} \frac{\Gamma(3)}{[(1+z)/5]^3} = \frac{2z}{(1+z)^3}. \end{aligned} \quad \square$$

(三) $M=\max\{X, Y\}$ 及 $N=\min\{X, Y\}$ 的分布

设 X, Y 是两个相互独立的随机变量, 它们的分布函数分别为 $F_X(x)$ 和 $F_Y(y)$. 现在来求 $M=\max\{X, Y\}$ 及 $N=\min\{X, Y\}$ 的分布函数.

由于 $M=\max\{X, Y\}$ 不大于 z 等价于 X 和 Y 都不大于 z , 故有

$$P\{M \leq z\} = P\{X \leq z, Y \leq z\}.$$

又由于 X 和 Y 相互独立, 得到 $M=\max\{X, Y\}$ 的分布函数为

$$F_{\max}(z) = P\{M \leq z\} = P\{X \leq z, Y \leq z\} = P\{X \leq z\}P\{Y \leq z\}.$$

$$\text{即有 } F_{\max}(z) = F_X(z)F_Y(z). \quad (5.11)$$

类似地, 可得 $N=\min\{X, Y\}$ 的分布函数为

$$\begin{aligned} F_{\min}(z) &= P\{N \leq z\} = 1 - P\{N > z\} \\ &= 1 - P\{X > z, Y > z\} = 1 - P\{X > z\} \cdot P\{Y > z\}. \end{aligned}$$

$$\text{即 } F_{\min}(z) = 1 - [1 - F_X(z)][1 - F_Y(z)]. \quad (5.12)$$

以上结果容易推广到 n 个相互独立的随机变量的情况. 设 $X_1, X_2, \dots, X_n$ 是 n 个相互独立的随机变量. 它们的分布函数分别为 $F_{X_i}(x_i)$ ($i=1, 2, \dots, n$), 则 $M=\max\{X_1, X_2, \dots, X_n\}$ 及 $N=\min\{X_1, X_2, \dots, X_n\}$ 的分布函数分别为

$$F_{\max}(z) = F_{X_1}(z)F_{X_2}(z)\cdots F_{X_n}(z), \quad (5.13)$$

$$F_{\min}(z) = 1 - [1 - F_{X_1}(z)][1 - F_{X_2}(z)]\cdots[1 - F_{X_n}(z)]. \quad (5.14)$$

特别, 当 $X_1, X_2, \dots, X_n$ 相互独立且具有相同分布函数 $F(x)$ 时有

$$F_{\max}(z) = [F(z)]^n, \quad (5.15)$$

$$F_{\min}(z) = 1 - [1 - F(z)]^n. \quad (5.16)$$

例 5 设系统 L 由两个相互独立的子系统 L_1, L_2 连接而成, 连接的方式分

别为(i)串联,(ii)并联,(iii)备用(当系统 L_1 损坏时,系统 L_2 开始工作),如图 3-13 所示. 设 L_1 , L_2 的寿命分别为 X, Y , 已知它们的概率密度分别为

$$f_X(x) = \begin{cases} \alpha e^{-\alpha x}, & x > 0, \\ 0, & x \leq 0, \end{cases} \quad (5.17)$$

$$f_Y(y) = \begin{cases} \beta e^{-\beta y}, & y > 0, \\ 0, & y \leq 0, \end{cases} \quad (5.18)$$

其中 $\alpha > 0, \beta > 0$ 且 $\alpha \neq \beta$. 试分别就以上三种连接方式写出 L 的寿命 Z 的概率密度.

解 (i) 串联的情况.

由于当 L_1, L_2 中有一个损坏时, 系统 L 就停止工作, 所以这时 L 的寿命为

$$Z = \min\{X, Y\}.$$

由(5.17), (5.18)式 X, Y 的分布函数分别为

$$F_X(x) = \begin{cases} 1 - e^{-\alpha x}, & x > 0, \\ 0, & x \leq 0, \end{cases} \quad F_Y(y) = \begin{cases} 1 - e^{-\beta y}, & y > 0, \\ 0, & y \leq 0. \end{cases}$$

由(5.12)式得 $Z = \min\{X, Y\}$ 的分布函数为

$$F_{\min}(z) = \begin{cases} 1 - e^{-(\alpha+\beta)z}, & z > 0, \\ 0, & z \leq 0. \end{cases}$$

于是 $Z = \min\{X, Y\}$ 的概率密度为

$$f_{\min}(z) = \begin{cases} (\alpha + \beta)e^{-(\alpha+\beta)z}, & z > 0, \\ 0, & z \leq 0. \end{cases}$$

(ii) 并联的情况.

由于当且仅当 L_1, L_2 都损坏时, 系统 L 才停止工作, 所以这时 L 的寿命 Z 为

$$Z = \max\{X, Y\}.$$

按(5.11)式得 $Z = \max\{X, Y\}$ 的分布函数为

$$F_{\max}(z) = F_X(z)F_Y(z) = \begin{cases} (1 - e^{-\alpha z})(1 - e^{-\beta z}), & z > 0, \\ 0, & z \leq 0. \end{cases}$$

于是 $Z = \max\{X, Y\}$ 的概率密度为

$$f_{\max}(z) = \begin{cases} \alpha e^{-\alpha z} + \beta e^{-\beta z} - (\alpha + \beta)e^{-(\alpha+\beta)z}, & z > 0, \\ 0, & z \leq 0. \end{cases}$$

(iii) 备用的情况.

由于这时当系统 L_1 损坏时系统 L_2 才开始工作, 因此整个系统 L 的寿命 Z

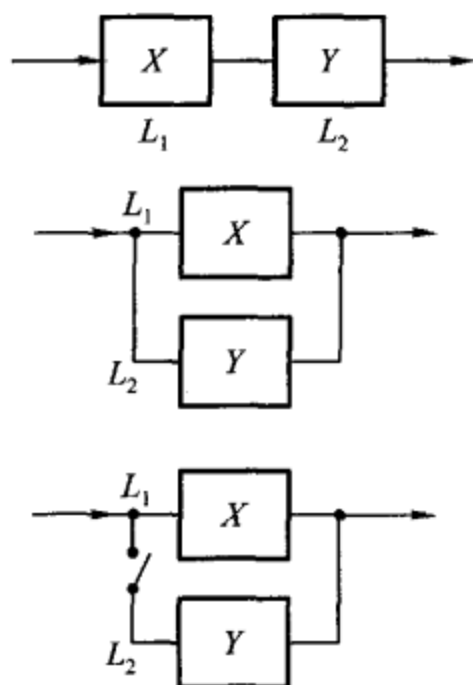


图 3-13

是 L_1, L_2 两者寿命之和, 即

$$Z = X + Y.$$

按(5.3)式, 当 $z > 0$ 时 $Z = X + Y$ 的概率密度为

$$\begin{aligned} f(z) &= \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy = \int_0^z \alpha e^{-\alpha(z-y)} \beta e^{-\beta y} dy \\ &= \alpha \beta e^{-\alpha z} \int_0^z e^{-(\beta-\alpha)y} dy \\ &= \frac{\alpha \beta}{\beta - \alpha} (e^{-\alpha z} - e^{-\beta z}). \end{aligned}$$

当 $z \leq 0$ 时, $f(z) = 0$, 于是 $Z = X + Y$ 的概率密度为

$$f(z) = \begin{cases} \frac{\alpha \beta}{\beta - \alpha} (e^{-\alpha z} - e^{-\beta z}), & z > 0, \\ 0, & z \leq 0. \end{cases}$$

□

小结

将一维随机变量的概念加以扩充, 就得到多维随机变量. 我们着重讨论了二维随机变量. 和一维随机变量一样, 我们定义二维随机变量 (X, Y) 的分布函数

$$F(x, y) = P\{X \leq x, Y \leq y\}, \quad -\infty < x < \infty, -\infty < y < \infty.$$

对于离散型随机变量 (X, Y) 定义了分布律

$$P\{X = x_i, Y = y_j\} = p_{ij}, \quad i = 1, 2, \dots, j = 1, 2, \dots, \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} p_{ij} = 1.$$

对于连续型随机变量 (X, Y) 定义了概率密度 $f(x, y)$ ($f(x, y) \geq 0$)

$$F(x, y) = \int_{-\infty}^y \int_{-\infty}^x f(x, y) dx dy, \quad \text{对于任意 } x, y.$$

二维随机变量的分布律与概率密度的性质与一维的类似. 特别, 对于二维连续型随机变量, 有公式

$$P\{(X, Y) \in G\} = \iint_G f(x, y) dx dy,$$

其中, G 是平面上的某区域 (它是一维连续型变量的公式 $P\{a < X \leq b\} = \int_a^b f(x) dx$ 的扩充). 这一公式常用来求随机变量的不等式成立的概率, 例如

$$P\{Y \leq X\} = P\{(X, Y) \in G\} = \iint_G f(x, y) dx dy,$$

其中, G 为半平面 $y \leq x$.

在研究二维随机变量 (X, Y) 时, 除了讨论上述与一维随机变量类似的内容外, 还要讨论以下的新内容: 边缘分布、条件分布、随机变量的独立性等.

注意到, 对于 (X, Y) 而言, 由 (X, Y) 的分布可以确定关于 X 、关于 Y 的边缘分布. 反之, 由关于 X 和关于 Y 的边缘分布一般是不能确定 (X, Y) 的分布的. 只有当 X, Y 相互独立时, 由两边缘分布能确定 (X, Y) 的分布.

随机变量的独立性是随机事件独立性的扩充. 我们也常利用问题的实际意义去判断两个随机变量的独立性. 例如, 若 X, Y 分别表示两个工厂生产的显像管的寿命, 我们可以认为 X, Y 是相互独立的.

我们还讨论了 $Z = X + Y, Z = Y/X, Z = XY, M = \max\{X, Y\}, N = \min\{X, Y\}$ 的分布的求法 (设 (X, Y) 的分布已知).

本章在进行各种问题的计算时, 要用到二重积分或用到二元函数固定其中一个变量对另一个变量的积分. 此时千万要搞清楚积分变量的变化范围. 题目做错, 往往是由于在进行积分运算时, 将有关的积分区间或积分区域搞错了. 在做题时, 画出有关函数的定义域的图形, 对于正确确定积分上下限肯定是有帮助的. 另外, 所求得的边缘密度、条件密度或 $Z = X + Y$ 的密度等, 往往是分段函数, 正确写出分段函数的表达式当然是必须的.

■重要术语及主题

二维随机变量 (X, Y) (X, Y) 的分布函数 离散型随机变量 (X, Y) 的分布律 连续型随机变量 (X, Y) 的概率密度 离散型随机变量 (X, Y) 的边缘分布律 连续型随机变量 (X, Y) 的边缘概率密度 条件分布函数 条件分布律 条件概率密度 两个随机变量 X, Y 的独立性 $Z = X + Y, Z = Y/X, Z = XY$ 的概率密度 $M = \max\{X, Y\}, N = \min\{X, Y\}$ 的概率密度

习题

1. 在一箱子中装有 12 只开关, 其中 2 只是次品, 在其中取两次, 每次任取一只, 考虑两种试验: (1) 放回抽样; (2) 不放回抽样. 我们定义随机变量 X, Y 如下:

$$X = \begin{cases} 0, & \text{若第一次取出的是正品,} \\ 1, & \text{若第一次取出的是次品;} \end{cases}$$

$$Y = \begin{cases} 0, & \text{若第二次取出的是正品,} \\ 1, & \text{若第二次取出的是次品.} \end{cases}$$

试分别就 (1)、(2) 两种情况, 写出 X 和 Y 的联合分布律.

2. (1) 盒子里装有 3 只黑球、2 只红球、2 只白球, 在其中任取 4 只球. 以 X 表示取到黑球的只数, 以 Y 表示取到红球的只数. 求 X 和 Y 的联合分布律.

(2) 在 (1) 中求 $P\{X > Y\}, P\{Y = 2X\}, P\{X + Y = 3\}, P\{X < 3 - Y\}$.

3. 设随机变量 (X, Y) 的概率密度为

$$f(x, y) = \begin{cases} k(6 - x - y), & 0 < x < 2, 2 < y < 4, \\ 0, & \text{其他.} \end{cases}$$

(1) 确定常数 k .

(2) 求 $P\{X < 1, Y < 3\}$.

(3) 求 $P\{X < 1.5\}$.

(4) 求 $P\{X + Y \leq 4\}$.

4. 设 X, Y 都是非负连续型随机变量, 它们相互独立.

(1) 证明 $P\{X < Y\} = \int_0^{\infty} F_X(x) f_Y(x) dx$,

其中 $F_X(x)$ 是 X 的分布函数, $f_Y(y)$ 是 Y 的概率密度.

(2) 设 X, Y 相互独立, 其概率密度分别为

$$f_X(x) = \begin{cases} \lambda_1 e^{-\lambda_1 x}, & x > 0, \\ 0, & \text{其他}, \end{cases} \quad f_Y(y) = \begin{cases} \lambda_2 e^{-\lambda_2 y}, & y > 0, \\ 0, & \text{其他}, \end{cases}$$

求 $P\{X < Y\}$.

5. 设随机变量 (X, Y) 具有分布函数

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

求边缘分布函数.

6. 将一枚硬币掷 3 次, 以 X 表示前 2 次中出现 H 的次数, 以 Y 表示 3 次中出现 H 的次数. 求 X, Y 的联合分布律以及 (X, Y) 的边缘分布律.

7. 设二维随机变量 (X, Y) 的概率密度为

$$f(x, y) = \begin{cases} 4.8y(2-x), & 0 \leq x \leq 1, 0 \leq y \leq x, \\ 0, & \text{其他}. \end{cases}$$

求边缘概率密度.

8. 设二维随机变量 (X, Y) 的概率密度为

$$f(x, y) = \begin{cases} e^{-y}, & 0 < x < y, \\ 0, & \text{其他}. \end{cases}$$

求边缘概率密度.

9. 设二维随机变量 (X, Y) 的概率密度为

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

(1) 确定常数 c .

(2) 求边缘概率密度.

10. 将某一医药公司 8 月份和 9 月份收到的青霉素针剂的订货单数分别记为 X 和 Y . 据以往积累的资料知 X 和 Y 的联合分布律为

$\begin{matrix} X \\ Y \end{matrix}$	51	52	53	54	55
51	0.06	0.05	0.05	0.01	0.01
52	0.07	0.05	0.01	0.01	0.01
53	0.05	0.10	0.10	0.05	0.05
54	0.05	0.02	0.01	0.01	0.03
55	0.05	0.06	0.05	0.01	0.03

(1) 求边缘分布律.

(2) 求 8 月份的订单数为 51 时, 9 月份订单数的条件分布律.

11. 以 X 记某医院一天出生的婴儿的个数, Y 记其中男婴的个数, 设 X 和 Y 的联合分布律为

$$P\{X = n, Y = m\} = \frac{e^{-14} (7.14)^m (6.86)^{n-m}}{m!(n-m)!},$$

$$m = 0, 1, 2, \dots, n; \quad n = 0, 1, 2, \dots.$$

(1) 求边缘分布律.

(2) 求条件分布律.

(3) 特别, 写出当 $X=20$ 时, Y 的条件分布律.

12. 求 §1 例 1 中的条件分布律: $P\{Y=k|X=i\}$.

13. 在第 9 题中

(1) 求条件概率密度 $f_{X|Y}(x|y)$, 特别, 写出当 $Y=\frac{1}{2}$ 时 X 的条件概率密度.

(2) 求条件概率密度 $f_{Y|X}(y|x)$, 特别, 分别写出当 $X=\frac{1}{3}$, $X=\frac{1}{2}$ 时 Y 的条件概率密度.

(3) 求条件概率

$$P\left\{Y \geq \frac{1}{4} \mid X = \frac{1}{2}\right\}, \quad P\left\{Y \geq \frac{3}{4} \mid X = \frac{1}{2}\right\}.$$

14. 设随机变量 (X, Y) 的概率密度为

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

求条件概率密度 $f_{Y|X}(y|x)$, $f_{X|Y}(x|y)$.

15. 设随机变量 $X \sim U(0, 1)$, 当给定 $X=x$ 时, 随机变量 Y 的条件概率密度为

$$f_{Y|X}(y|x) = \begin{cases} x, & 0 < y < \frac{1}{x}, \\ 0, & \text{其他.} \end{cases}$$

(1) 求 X 和 Y 的联合概率密度 $f(x, y)$.

(2) 求边缘密度 $f_Y(y)$, 并画出它的图形.

(3) 求 $P\{X > Y\}$.

16. (1) 问第 1 题中的随机变量 X 和 Y 是否相互独立?

(2) 问第 14 题中的随机变量 X 和 Y 是否相互独立(需说明理由)?

17. (1) 设随机变量 (X, Y) 具有分布函数

$$F(x, y) = \begin{cases} (1 - e^{-ax})y, & x \geq 0, 0 \leq y \leq 1, \\ 1 - e^{-ax}, & x \geq 0, y > 1, \\ 0, & \text{其他.} \end{cases} \quad a > 0,$$

证明 X, Y 相互独立.

(2) 设随机变量 (X, Y) 具有分布律

$$P\{X=x, Y=y\} = p^2(1-p)^{x+y-2}, \quad 0 < p < 1, x, y \text{ 均为正整数},$$

问 X, Y 是否相互独立.

18. 设 X 和 Y 是两个相互独立的随机变量, X 在区间 $(0, 1)$ 上服从均匀分布, Y 的概率密度为

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

(1) 求 X 和 Y 的联合概率密度.

(2) 设含有 a 的二次方程为 $a^2 + 2Xa + Y = 0$, 试求 a 有实根的概率.

19. 进行打靶, 设弹着点 $A(X, Y)$ 的坐标 X 和 Y 相互独立, 且都服从 $N(0, 1)$ 分布, 规定

点 A 落在区域 $D_1 = \{(x, y) | x^2 + y^2 \leq 1\}$ 得 2 分;

点 A 落在 $D_2 = \{(x, y) | 1 < x^2 + y^2 \leq 4\}$ 得 1 分;

点 A 落在 $D_3 = \{(x, y) | x^2 + y^2 > 4\}$ 得 0 分.

以 Z 记打靶的得分. 写出 X, Y 的联合概率密度, 并求 Z 的分布律.

20. 设 X 和 Y 是相互独立的随机变量, 其概率密度分别为

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0, \end{cases} \quad f_Y(y) = \begin{cases} \mu e^{-\mu y}, & y > 0, \\ 0, & y \leq 0, \end{cases}$$

其中 $\lambda > 0, \mu > 0$ 是常数. 引入随机变量

$$Z = \begin{cases} 1, & \text{当 } X \leq Y, \\ 0, & \text{当 } X > Y. \end{cases}$$

(1) 求条件概率密度 $f_{X|Y}(x|y)$.

(2) 求 Z 的分布律和分布函数.

21. 设随机变量 (X, Y) 的概率密度为

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

分别求 (1) $Z = X + Y$, (2) $Z = XY$ 的概率密度.

22. 设 X 和 Y 是两个相互独立的随机变量, 其概率密度分别为

$$f_X(x) = \begin{cases} 1, & 0 \leq x \leq 1, \\ 0, & \text{其他.} \end{cases} \quad f_Y(y) = \begin{cases} e^{-y}, & y > 0, \\ 0, & \text{其他.} \end{cases}$$

求随机变量 $Z = X + Y$ 的概率密度.

23. 某种商品一周的需求量是一个随机变量, 其概率密度为

$$f(t) = \begin{cases} te^{-t}, & t > 0, \\ 0, & t \leq 0. \end{cases}$$

设各周的需求量是相互独立的. 求 (1) 两周, (2) 三周的需求量的概率密度.

24. 设随机变量 (X, Y) 的概率密度为

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

(1) 问 X 和 Y 是否相互独立?

(2) 求 $Z = X + Y$ 的概率密度.

25. 设随机变量 X, Y 相互独立, 且具有相同的分布, 它们概率密度均为

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

求 $Z = X + Y$ 的概率密度.

26. 设随机变量 X, Y 相互独立, 它们的概率密度均为

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

求 $Z = Y/X$ 的概率密度.

27. 设随机变量 X, Y 相互独立, 它们都在区间 $(0, 1)$ 上服从均匀分布. A 是以 X, Y 为边长的矩形的面积, 求 A 的概率密度.

28. 设 X, Y 是相互独立的随机变量, 它们都服从正态分布 $N(0, \sigma^2)$. 试验证随机变量 $Z = \sqrt{X^2 + Y^2}$ 的概率密度为

$$f_Z(z) = \begin{cases} \frac{z}{\sigma^2} e^{-z^2/(2\sigma^2)}, & z \geq 0, \\ 0, & \text{其他.} \end{cases}$$

我们称 Z 服从参数为 σ ($\sigma > 0$) 的瑞利(Rayleigh)分布.

29. 设随机变量 (X, Y) 的概率密度为

$$f(x, y) = \begin{cases} be^{-(x+y)}, & 0 < x < 1, 0 < y < \infty, \\ 0, & \text{其他.} \end{cases}$$

(1) 试确定常数 b .

(2) 求边缘概率密度 $f_X(x), f_Y(y)$.

(3) 求函数 $U = \max\{X, Y\}$ 的分布函数.

30. 设某种型号的电子元件的寿命(以小时计)近似地服从正态分布 $N(160, 20^2)$, 随机地选取 4 只, 求其中没有一只寿命小于 180 的概率.

31. 对某种电子装置的输出测量了 5 次, 得到结果为 X_1, X_2, X_3, X_4, X_5 . 设它们是相互独立的随机变量且都服从参数 $\sigma = 2$ 的瑞利分布.

(1) 求 $Z = \max\{X_1, X_2, X_3, X_4, X_5\}$ 的分布函数.

(2) 求 $P\{Z > 4\}$.

32. 设随机变量 X, Y 相互独立, 且服从同一分布, 试证明:

$$P\{a < \min\{X, Y\} \leq b\} = [P\{X > a\}]^2 - [P\{X > b\}]^2 \quad (a \leq b).$$

33. 设 X, Y 是相互独立的随机变量, 其分布律分别为

$$P\{X = k\} = p(k), \quad k = 0, 1, 2, \dots,$$

$$P\{Y = r\} = q(r), \quad r = 0, 1, 2, \dots.$$

证明随机变量 $Z = X + Y$ 的分布律为

$$P\{Z = i\} = \sum_{k=0}^i p(k)q(i-k), \quad i = 0, 1, 2, \dots.$$

34. 设 X, Y 是相互独立的随机变量, $X \sim \pi(\lambda_1), Y \sim \pi(\lambda_2)$. 证明 $Z = X + Y \sim \pi(\lambda_1 + \lambda_2)$.

35. 设 X, Y 是相互独立的随机变量, $X \sim b(n_1, p), Y \sim b(n_2, p)$. 证明

$$Z = X + Y \sim b(n_1 + n_2, p).$$

36. 设随机变量 (X, Y) 的分布律为

$\begin{matrix} X \\ Y \end{matrix}$	0	1	2	3	4	5
0	0.00	0.01	0.03	0.05	0.07	0.09
1	0.01	0.02	0.04	0.05	0.06	0.08
2	0.01	0.03	0.05	0.05	0.05	0.06
3	0.01	0.02	0.04	0.06	0.06	0.05

- (1) 求 $P\{X=2|Y=2\}, P\{Y=3|X=0\}$.
- (2) 求 $V=\max\{X,Y\}$ 的分布律.
- (3) 求 $U=\min\{X,Y\}$ 的分布律.
- (4) 求 $W=X+Y$ 的分布律.



第四章 随机变量的数字特征

上一章介绍了随机变量的分布函数、概率密度和分布律,它们都能完整地描述随机变量,但在某些实际或理论问题中,人们感兴趣于某些能描述随机变量某一种特征的常数,例如,一篮球队上场比赛的运动员的身高是一个随机变量,人们常关心上场运动员的平均身高.一个城市一户家庭拥有汽车的辆数是一个随机变量,在考察城市的交通情况时,人们关心户均拥有汽车的辆数.评价棉花的质量时,既需要注意纤维的平均长度,又需要注意纤维长度与平均长度的偏离程度,平均长度较大,偏离程度较小,质量就较好.这种由随机变量的分布所确定的,能刻画随机变量某一方面的特征的常数统称为数字特征,它在理论和实际应用中都很重要.本章将介绍几个重要的数字特征:数学期望、方差、相关系数和矩.

§1 数学期望

先看一个例子.一射手进行打靶练习,规定射入区域 e_2 (图4-1) 得 2 分,射入区域 e_1 得 1 分,脱靶,即射入区域 e_0 , 得 0 分. 射手一次射击得分 X 是一个随机变量. 设 X 的分布律为

$$P\{X=k\}=p_k, k=0,1,2.$$

现在射击 N 次,其中得 0 分的有 a_0 次,得 1 分的有 a_1 次,得 2 分的有 a_2 次, $a_0+a_1+a_2=N$. 他射击 N 次得分的总和为 $a_0 \times 0 + a_1 \times 1 + a_2 \times 2$. 于是平均一次射击的得分数为

$$\frac{a_0 \times 0 + a_1 \times 1 + a_2 \times 2}{N} = \sum_{k=0}^2 k \frac{a_k}{N}.$$

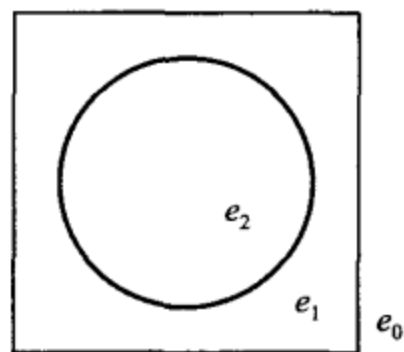


图 4-1

这里, a_k/N 是事件 $\{X=k\}$ 的频率. 在第五章将会讲到, 当 N 很大时, a_k/N 在一定意义下接近于事件 $\{X=k\}$ 的概率 p_k . 就是说, 在试验次数很大时, 随机变量 X 的观察值的算术平均 $\sum_{k=0}^2 ka_k/N$ 在一定意义下接近于 $\sum_{k=0}^2 kp_k$. 我们称 $\sum_{k=0}^2 kp_k$ 为随机变量 X 的数学期望或均值. 一般, 有以下的定义.

定义 设离散型随机变量 X 的分布律为

$$P\{X=x_k\}=p_k, \quad k=1,2,\dots.$$

若级数

$$\sum_{k=1}^{\infty} x_k p_k$$

绝对收敛, 则称级数 $\sum_{k=1}^{\infty} x_k p_k$ 的和为随机变量 X 的数学期望, 记为 $E(X)$. 即

$$E(X) = \sum_{k=1}^{\infty} x_k p_k. \quad (1.1)$$

设连续型随机变量 X 的概率密度为 $f(x)$, 若积分

$$\int_{-\infty}^{\infty} x f(x) dx$$

绝对收敛, 则称积分 $\int_{-\infty}^{\infty} x f(x) dx$ 的值为随机变量 X 的数学期望, 记为 $E(X)$. 即

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx. \quad (1.2)$$

数学期望简称期望, 又称为均值.

数学期望 $E(X)$ 完全由随机变量 X 的概率分布所确定. 若 X 服从某一分布, 也称 $E(X)$ 是这一分布的数学期望.

例 1 某医院当新生儿诞生时, 医生要根据婴儿的皮肤颜色、肌肉弹性、反应的敏感性、心脏的搏动等方面的情况进行评分, 新生儿的得分 X 是一个随机变量. 据以往的资料表明 X 的分布律为

X	0	1	2	3	4	5	6	7	8	9	10
p_k	0.002	0.001	0.002	0.005	0.02	0.04	0.18	0.37	0.25	0.12	0.01

试求 X 的数学期望 $E(X)$.

解 $E(X) = 0 \times 0.002 + 1 \times 0.001 + 2 \times 0.002 + 3 \times 0.005 + 4 \times 0.02$
 $+ 5 \times 0.04 + 6 \times 0.18 + 7 \times 0.37 + 8 \times 0.25 + 9 \times 0.12 + 10 \times 0.01$
 $= 7.15(\text{分})$

这意味着, 若考察医院出生的很多新生儿, 例如 1000 个, 那么一个新生儿的平均得分约 7.15 分, 1000 个新生儿共得分约 7150 分. $\square$

例 2 有两个相互独立工作的电子装置, 它们的寿命(以小时计) X_k ($k=1, 2$) 服从同一指数分布, 其概率密度为

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

若将这两个电子装置串联连接组成整机, 求整机寿命(以小时计) N 的数学期望.

解 $X_k (k=1, 2)$ 的分布函数为

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

由第三章 § 5 的 (5.12) 式 $N = \min\{X_1, X_2\}$ 的分布函数为

$$F_{\min}(x) = 1 - [1 - F(x)]^2 = \begin{cases} 1 - e^{-2x/\theta}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

因而 N 的概率密度为

$$f_{\min}(x) = \begin{cases} \frac{2}{\theta} e^{-2x/\theta}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

于是 N 的数学期望为

$$E(N) = \int_{-\infty}^{\infty} x f_{\min}(x) dx = \int_0^{\infty} \frac{2x}{\theta} e^{-2x/\theta} dx = \frac{\theta}{2}. \quad \square$$

例 3 按规定, 某车站每天 8:00~9:00, 9:00~10:00 都恰有一辆客车到站, 但到站的时刻是随机的, 且两者到站的时间相互独立. 其规律为

到站时刻	8:10	8:30	8:50
	9:10	9:30	9:50
概率	$\frac{1}{6}$	$\frac{3}{6}$	$\frac{2}{6}$

一旅客 8:20 到车站, 求他候车时间的数学期望.

解 设旅客的候车时间为 X (以分计). X 的分布律为

X	10	30	50	70	90
p_k	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{6} \times \frac{1}{6}$	$\frac{1}{6} \times \frac{3}{6}$	$\frac{1}{6} \times \frac{2}{6}$

在上表中, 例如

$$P\{X=70\} = P(AB) = P(A)P(B) = \frac{1}{6} \times \frac{3}{6},$$

其中 A 为事件“第一班车在 8:10 到站”, B 为“第二班车在 9:30 到站”. 候车时间的数学期望为

$$\begin{aligned} E(X) &= 10 \times \frac{3}{6} + 30 \times \frac{2}{6} + 50 \times \frac{1}{36} + 70 \times \frac{3}{36} + 90 \times \frac{2}{36} \\ &= 27.22(\text{分}). \end{aligned} \quad \square$$

例 4 某商店对某种家用电器的销售采用先使用后付款的方式. 记使用寿命为 X (以年计), 规定:

$X \leq 1$, 一台付款 1 500 元;

$1 < X \leq 2$, 一台付款 2 000 元;

$2 < X \leq 3$, 一台付款 2 500 元;

$X > 3$, 一台付款 3 000 元.

设寿命 X 服从指数分布, 概率密度为

$$f(x) = \begin{cases} \frac{1}{10} e^{-x/10}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

试求该商店一台这种家用电器收费 Y 的数学期望.

解 先求出寿命 X 落在各个时间区间的概率. 即有

$$P\{X \leq 1\} = \int_0^1 \frac{1}{10} e^{-x/10} dx = 1 - e^{-0.1} = 0.0952,$$

$$P\{1 < X \leq 2\} = \int_1^2 \frac{1}{10} e^{-x/10} dx = e^{-0.1} - e^{-0.2} = 0.0861,$$

$$P\{2 < X \leq 3\} = \int_2^3 \frac{1}{10} e^{-x/10} dx = e^{-0.2} - e^{-0.3} = 0.0779,$$

$$P\{X > 3\} = \int_3^{\infty} \frac{1}{10} e^{-x/10} dx = e^{-0.3} = 0.7408.$$

一台家用电器收费 Y 的分布律为

Y	1 500	2 000	2 500	3 000
p_k	0.0952	0.0861	0.0779	0.7408

得 $E(Y) = 2\,732.15$, 即平均一台收费 2732.15 元. $\square$

例 5 在一个人数很多的团体中普查某种疾病, 为此要抽验 N 个人的血, 可以用两种方法进行. (i) 将每个人的血分别去验, 这就需验 N 次. (ii) 按 k 个人一组进行分组, 把从 k 个人抽来的血混合在一起进行检验, 如果这混合血液呈阴性反应, 就说明 k 个人的血都呈阴性反应, 这样, 这 k 个人的血就只需验一次. 若呈阳性, 则再对这 k 个人的血液分别进行化验. 这样, k 个人的血总共要化验 $k+1$ 次. 假设每个人化验呈阳性的概率为 p , 且这些人的试验反应是相互独立的. 试说明当 p 较小时, 选取适当的 k , 按第二种方法可以减少化验的次数. 并说明 k 取什么值时最适宜.

解 各人的血呈阴性反应的概率为 $q = 1 - p$. 因而 k 个人的混合血呈阴性反应的概率为 q^k , k 个人的混合血呈阳性反应的概率为 $1 - q^k$.

设以 k 个人为一组时, 组内每人化验的次数为 X , 则 X 是一个随机变量, 其分布律为

X	$\frac{1}{k}$	$\frac{k+1}{k}$
p_k	q^k	$1 - q^k$

X 的数学期望为

$$E(X) = \frac{1}{k}q^k + \left(1 + \frac{1}{k}\right)(1 - q^k) = 1 - q^k + \frac{1}{k}.$$

N 个人平均需化验的次数为

$$N\left(1 - q^k + \frac{1}{k}\right).$$

由此可知, 只要选择 k 使

$$1 - q^k + \frac{1}{k} < 1,$$

则 N 个人平均需化验的次数 $< N$. 当 p 固定时, 我们选取 k 使得

$$L = 1 - q^k + \frac{1}{k}$$

小于 1 且取到最小值, 这时就能得到最好的分组方法.

例如, $p = 0.1$, 则 $q = 0.9$, 当 $k = 4$ 时, $L = 1 - q^k + \frac{1}{k}$ 取到最小值. 此时得到最好的分组方法. 若 $N = 1000$, 此时以 $k = 4$ 分组, 则按第二种方法平均只需化验

$$1000\left(1 - 0.9^4 + \frac{1}{4}\right) = 594(\text{次}).$$

这样平均来说, 可以减少 40% 的工作量. □

例 6 设 $X \sim \pi(\lambda)$, 求 $E(X)$.

解 X 的分布律为

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

X 的数学期望为

$$E(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda,$$

即 $E(X) = \lambda$. □

例 7 设 $X \sim U(a, b)$, 求 $E(X)$.

解 X 的概率密度为

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

X 的数学期望为

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2}.$$

即数学期望位于区间 (a, b) 的中点. □

我们经常需要求随机变量的函数的数学期望, 例如飞机机翼受到压力 $W =$

kV^2 (V 是风速, $k>0$ 是常数) 的作用, 需要求 W 的数学期望, 这里 W 是随机变量 V 的函数. 这时, 可以通过下面的定理来求 W 的数学期望.

定理 设 Y 是随机变量 X 的函数: $Y=g(X)$ (g 是连续函数).

(i) 如果 X 是离散型随机变量, 它的分布律为 $P\{X=x_k\}=p_k, k=1, 2, \dots$,

若 $\sum_{k=1}^{\infty} g(x_k) p_k$ 绝对收敛, 则有

$$E(Y) = E[g(X)] = \sum_{k=1}^{\infty} g(x_k) p_k. \quad (1.3)$$

(ii) 如果 X 是连续型随机变量, 它的概率密度为 $f(x)$, 若 $\int_{-\infty}^{\infty} g(x) f(x) dx$ 绝对收敛, 则有

$$E(Y) = E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx. \quad (1.4)$$

定理的重要意义在于当我们求 $E(Y)$ 时, 不必算出 Y 的分布律或概率密度, 而只需利用 X 的分布律或概率密度就可以了, 定理的证明超出了本书的范围. 我们只对下述特殊情况加以证明.

证 设 X 是连续型随机变量, 且 $y=g(x)$ 满足第二章 §5 中定理的条件.

由第二章 §5 中的 (5.2) 式知道随机变量 $Y=g(X)$ 的概率密度为

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

于是

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{\alpha}^{\beta} y f_X[h(y)] |h'(y)| dy.$$

当 $h'(y)$ 恒 > 0 时

$$E(Y) = \int_{\alpha}^{\beta} y f_X[h(y)] h'(y) dy = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

当 $h'(y)$ 恒 < 0 时

$$\begin{aligned} E(Y) &= - \int_{\alpha}^{\beta} y f_X[h(y)] h'(y) dy \\ &= - \int_{\infty}^{-\infty} g(x) f(x) dx = \int_{-\infty}^{\infty} g(x) f(x) dx. \end{aligned}$$

综合上两式, (1.4) 式得证. □

上述定理还可以推广到两个或两个以上随机变量的函数的情况.

例如, 设 Z 是随机变量 X, Y 的函数 $Z=g(X, Y)$ (g 是连续函数), 那么, Z 是一个一维随机变量. 若二维随机变量 (X, Y) 的概率密度为 $f(x, y)$, 则有

$$E(Z) = E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy, \quad (1.5)$$

这里设上式右边的积分绝对收敛. 又若 (X, Y) 为离散型随机变量, 其分布律为 $P\{X=x_i, Y=y_j\}=p_{ij}, i, j=1, 2, \dots$, 则有

$$E(Z) = E[g(X, Y)] = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} g(x_i, y_j) p_{ij}, \quad (1.6)$$

这里设上式右边的级数绝对收敛.

例 8 设风速 V 在 $(0, a)$ 上服从均匀分布, 即具有概率密度

$$f(v) = \begin{cases} \frac{1}{a}, & 0 < v < a, \\ 0, & \text{其他.} \end{cases}$$

又设飞机机翼受到的正压力 W 是 V 的函数: $W = kV^2$ ($k > 0$, 常数), 求 W 的数学期望.

解 由 (1.4) 式有

$$E(W) = \int_{-\infty}^{\infty} kv^2 f(v) dv = \int_0^a kv^2 \frac{1}{a} dv = \frac{1}{3} ka^2. \quad \square$$

例 9 设随机变量 (X, Y) 的概率密度

$$f(x, y) = \begin{cases} \frac{3}{2x^3 y^2}, & \frac{1}{x} < y < x, x > 1, \\ 0, & \text{其他.} \end{cases}$$

求数学期望 $E(Y), E\left(\frac{1}{XY}\right)$.

解 由 (1.5) 式得

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) dy dx = \int_1^{\infty} \int_{\frac{1}{x}}^x \frac{3}{2x^3 y} dy dx \\ &= \frac{3}{2} \int_1^{\infty} \frac{1}{x^3} \left[\ln y \right]_{\frac{1}{x}}^x dx = 3 \int_1^{\infty} \frac{\ln x}{x^3} dx \\ &= \left[-\frac{3}{2} \frac{\ln x}{x^2} \right]_1^{\infty} + \frac{3}{2} \int_1^{\infty} \frac{1}{x^3} dx = \frac{3}{4}. \end{aligned}$$

$$E\left(\frac{1}{XY}\right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{xy} f(x, y) dy dx = \int_1^{\infty} dx \int_{\frac{1}{x}}^x \frac{3}{2x^4 y^3} dy = \frac{3}{5}. \quad \square$$

例 10 某公司计划开发一种新产品市场, 并试图确定该产品的产量. 他们估计出售一件产品可获利 m 元, 而积压一件产品导致 n 元的损失. 再者, 他们预测销售量 Y (件) 服从指数分布, 其概率密度为

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

问若要获得利润的数学期望最大, 应生产多少件产品 (m, n, θ 均为已知)?

解 设生产 x 件, 则获利 Q 是 x 的函数

$$Q=Q(x)=\begin{cases} mY-n(x-Y), & Y<x, \\ mx, & Y\geq x. \end{cases}$$

Q 是随机变量,它是 Y 的函数,其数学期望为

$$\begin{aligned} E(Q) &= \int_0^{\infty} Qf_Y(y)dy \\ &= \int_0^x [my - n(x-y)] \frac{1}{\theta} e^{-y/\theta} dy \\ &\quad + \int_x^{\infty} mx \frac{1}{\theta} e^{-y/\theta} dy \\ &= (m+n)\theta - (m+n)\theta e^{-x/\theta} - nx. \end{aligned}$$

$$\text{令} \quad \frac{d}{dx}E(Q) = (m+n)e^{-x/\theta} - n = 0,$$

$$\text{得} \quad x = -\theta \ln \frac{n}{m+n}.$$

$$\text{而} \quad \frac{d^2}{dx^2}E(Q) = \frac{-(m+n)}{\theta} e^{-x/\theta} < 0,$$

故知当 $x = -\theta \ln \frac{n}{m+n}$ 时 $E(Q)$ 取极大值,且可知这也是最大值.

$$\text{例如,若 } f_Y(y) = \begin{cases} \frac{1}{10\,000} e^{-\frac{y}{10\,000}}, & y > 0, \\ 0, & y \leq 0, \end{cases}$$

且有 $m=500$ 元, $n=2000$ 元,则

$$x = -10\,000 \ln \frac{2\,000}{500+2\,000} = 2\,231.4.$$

取 $x=2231$ 件. □

例 11 某甲与其他三人参与一个项目的竞拍,价格以千美元计,价格高者获胜.若甲中标,他就将此项目以 10 千美元转让给他人.可认为其他三人的竞拍价是相互独立的,且都在 7~11 千美元之间均匀分布.问甲应如何报价才能使获益的数学期望为最大(若甲中标必须将此项目以他自己的报价买下).

解 设 X_1, X_2, X_3 是其他三人的报价,按题意 X_1, X_2, X_3 相互独立,且在区间(7,11)上服从均匀分布.其分布函数为

$$F(u) = \begin{cases} 0, & u < 7, \\ \frac{u-7}{4}, & 7 \leq u < 11, \\ 1, & u \geq 11. \end{cases}$$

以 Y 记三人最大出价,即 $Y = \max\{X_1, X_2, X_3\}$. Y 的分布函数为

$$F_Y(u) = \begin{cases} 0, & u < 7, \\ \left(\frac{u-7}{4}\right)^3, & 7 \leq u < 11, \\ 1, & u \geq 11. \end{cases}$$

若甲的报价为 x , 按题意 $7 \leq x \leq 10$, 知甲能赢得这一项目的概率为

$$p = P\{Y \leq x\} = F_Y(x) = \left(\frac{x-7}{4}\right)^3 \quad (7 \leq x \leq 10).$$

以 $G(x)$ 记甲的赚钱数, $G(x)$ 是一个随机变量, 它的分布律为

$G(x)$	$10-x$	0
概率	$\left(\frac{x-7}{4}\right)^3$	$1 - \left(\frac{x-7}{4}\right)^3$

于是甲赚钱数的数学期望为

$$E[G(x)] = \left(\frac{x-7}{4}\right)^3 (10-x).$$

$$\text{令 } \frac{d}{dx} E[G(x)] = \frac{1}{4^3} [(x-7)^2 (37-4x)] = 0,$$

$$\text{得 } x = 37/4, x = 7 \text{ (舍去).}$$

$$\text{又知 } \left. \frac{d^2}{dx^2} E[G(x)] \right|_{x=37/4} < 0.$$

故知当甲的报价为 $x = 37/4$ 千美元时, 他赚钱数的数学期望达到极大值, 还可知这也是最大值. $\square$

现在来证明数学期望的几个重要性质^① (以下设所遇到的随机变量的数学期望存在).

1° 设 C 是常数, 则有 $E(C) = C$.

2° 设 X 是一个随机变量, C 是常数, 则有

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

3° 设 X, Y 是两个随机变量, 则有

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

这一性质可以推广到任意有限个随机变量之和的情况.

4° 设 X, Y 是相互独立的随机变量, 则有

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

这一性质可以推广到任意有限个相互独立的随机变量之积的情况.

证 1°、2° 由读者自己证明. 我们来证 3° 和 4°.

^① 这里我们只对连续型随机变量的情况加以证明, 读者只要将证明中的“积分”用“和式”代替, 就能得到离散型随机变量情况的证明.

设二维随机变量 (X, Y) 的概率密度为 $f(x, y)$. 其边缘概率密度为 $f_X(x)$, $f_Y(y)$. 由(1.5)式

$$\begin{aligned} E(X+Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y) f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) dx dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) dx dy \\ &= E(X) + E(Y). \end{aligned}$$

3°得证.

又若 X 和 Y 相互独立,

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_X(x) f_Y(y) dx dy \\ &= \left[\int_{-\infty}^{\infty} x f_X(x) dx \right] \left[\int_{-\infty}^{\infty} y f_Y(y) dy \right] = E(X)E(Y). \end{aligned}$$

4°得证. □

例 12 一民航送客车载有 20 位旅客自机场开出, 旅客有 10 个车站可以下车. 如到达一个车站没有旅客下车就不停车. 以 X 表示停车的次数, 求 $E(X)$ (设每位旅客在各个车站下车是等可能的, 并设各位旅客是否下车相互独立).

解 引入随机变量

$$X_i = \begin{cases} 0, & \text{在第 } i \text{ 站没有人下车,} \\ 1, & \text{在第 } i \text{ 站有人下车,} \end{cases} \quad i=1, 2, \dots, 10.$$

易知

$$X = X_1 + X_2 + \dots + X_{10}.$$

现在来求 $E(X)$.

按题意, 任一旅客在第 i 站不下车的概率为 $\frac{9}{10}$, 因此 20 位旅客都不在第 i 站下车的概率为 $\left(\frac{9}{10}\right)^{20}$, 在第 i 站有人下车的概率为 $1 - \left(\frac{9}{10}\right)^{20}$, 也就是

$$P\{X_i=0\} = \left(\frac{9}{10}\right)^{20}, P\{X_i=1\} = 1 - \left(\frac{9}{10}\right)^{20}, i=1, 2, \dots, 10.$$

由此

$$E(X_i) = 1 - \left(\frac{9}{10}\right)^{20}, i=1, 2, \dots, 10.$$

进而

$$\begin{aligned} E(X) &= E(X_1 + X_2 + \dots + X_{10}) \\ &= E(X_1) + E(X_2) + \dots + E(X_{10}) \\ &= 10 \left[1 - \left(\frac{9}{10}\right)^{20} \right] = 8.784 (\text{次}). \end{aligned}$$

本题是将 X 分解成数个随机变量之和, 然后利用随机变量和的数学期望等

于随机变量数学期望之和来求数学期望的,这种处理方法具有一定的普遍意义. $\square$

例 13 设一电路中电流 I (A) 与电阻 R (Ω) 是两个相互独立的随机变量, 其概率密度分别为

$$g(i) = \begin{cases} 2i, & 0 \leq i \leq 1, \\ 0, & \text{其他}, \end{cases} \quad h(r) = \begin{cases} \frac{r^2}{9}, & 0 \leq r \leq 3, \\ 0, & \text{其他}. \end{cases}$$

试求电压 $V = IR$ 的均值.

解

$$E(V) = E(IR) = E(I)E(R)$$

$$\begin{aligned} &= \left[\int_{-\infty}^{\infty} ig(i) di \right] \left[\int_{-\infty}^{\infty} rh(r) dr \right] \\ &= \left(\int_0^1 2i^2 di \right) \left(\int_0^3 \frac{r^3}{9} dr \right) = \frac{3}{2} (\text{V}). \end{aligned}$$

$\square$

§ 2 方 差

先从例子说起. 例如, 有一批灯泡, 知其平均寿命是 $E(X) = 1\,000$ (小时). 仅由这一指标我们还不能判定这批灯泡的质量好坏. 事实上, 有可能其中绝大部分灯泡的寿命都在 $950 \sim 1\,050$ 小时; 也有可能其中约有一半是高质量的, 它们的寿命大约有 $1\,300$ 小时, 另一半却是质量很差的, 其寿命大约只有 700 小时. 为要评定这批灯泡质量的好坏, 还需进一步考察灯泡寿命 X 与其均值 $E(X) = 1\,000$ 的偏离程度. 若偏离程度较小, 表示质量比较稳定. 从这个意义上来说, 我们认为质量较好. 前面也曾提到在检验棉花的质量时, 既要注意纤维的平均长度, 还要注意纤维长度与平均长度的偏离程度. 由此可见, 研究随机变量与其均值的偏离程度是十分必要的. 那么, 用怎样的量去度量这个偏离程度呢? 容易看到

$$E\{|X - E(X)|\}$$

能度量随机变量与其均值 $E(X)$ 的偏离程度. 但由于上式带有绝对值, 运算不方便, 为运算方便起见, 通常用量

$$E\{[X - E(X)]^2\}$$

来度量随机变量 X 与其均值 $E(X)$ 的偏离程度.

定义 设 X 是一个随机变量, 若 $E\{[X - E(X)]^2\}$ 存在, 则称 $E\{[X - E(X)]^2\}$ 为 X 的方差, 记为 $D(X)$ 或 $\text{Var}(X)$, 即

$$D(X) = \text{Var}(X) = E\{[X - E(X)]^2\}. \quad (2.1)$$

在应用上还引入量 $\sqrt{D(X)}$, 记为 $\sigma(X)$, 称为标准差或均方差.

按定义, 随机变量 X 的方差表达了 X 的取值与其数学期望的偏离程度. 若 $D(X)$ 较小意味着 X 的取值比较集中在 $E(X)$ 的附近, 反之, 若 $D(X)$ 较大则表示 X 的取值较分散. 因此, $D(X)$ 是刻画 X 取值分散程度的一个量, 它是衡量 X 取值分散程度的一个尺度.

由定义知, 方差实际上就是随机变量 X 的函数 $g(X) = (X - E(X))^2$ 的数学期望. 于是对于离散型随机变量, 按(1.3)式有

$$D(X) = \sum_{k=1}^{\infty} [x_k - E(X)]^2 p_k, \quad (2.2)$$

其中 $P\{X=x_k\}=p_k, k=1, 2, \dots$ 是 X 的分布律.

对于连续型随机变量, 按(1.4)式有

$$D(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x) dx, \quad (2.3)$$

其中 $f(x)$ 是 X 的概率密度.

随机变量 X 的方差可按下列公式计算.

$$D(X) = E(X^2) - [E(X)]^2. \quad (2.4)$$

证 由数学期望的性质 $1^\circ, 2^\circ, 3^\circ$ 得

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

□

例 1 设随机变量 X 具有数学期望 $E(X) = \mu$, 方差 $D(X) = \sigma^2 \neq 0$. 记

$$X^* = \frac{X - \mu}{\sigma},$$

则
$$E(X^*) = \frac{1}{\sigma} E(X - \mu) = \frac{1}{\sigma} [E(X) - \mu] = 0;$$

$$\begin{aligned} D(X^*) &= E(X^{*2}) - [E(X^*)]^2 = E\left[\left(\frac{X - \mu}{\sigma}\right)^2\right] \\ &= \frac{1}{\sigma^2} E[(X - \mu)^2] = \frac{\sigma^2}{\sigma^2} = 1. \end{aligned}$$

即 $X^* = \frac{X - \mu}{\sigma}$ 的数学期望为 0, 方差为 1. X^* 称为 X 的标准化变量. □

例 2 设随机变量 X 具有(0-1)分布, 其分布律为

$$P\{X=0\}=1-p, \quad P\{X=1\}=p.$$

求 $D(X)$.

解

$$E(X) = 0 \cdot (1-p) + 1 \cdot p = p,$$

$$E(X^2) = 0^2 \cdot (1-p) + 1^2 \cdot p = p.$$

由(2.4)式

$$D(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1-p). \quad \square$$

例 3 设随机变量 $X \sim \pi(\lambda)$, 求 $D(X)$.

解 随机变量 X 的分布律为

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

上节例 6 已算得 $E(X) = \lambda$, 而

$$\begin{aligned} E(X^2) &= E[X(X-1) + X] = E[X(X-1)] + E(X) \\ &= \sum_{k=0}^{\infty} k(k-1) \frac{\lambda^k e^{-\lambda}}{k!} + \lambda = \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda \\ &= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda = \lambda^2 + \lambda, \end{aligned}$$

所以方差

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

由此可知, 泊松分布的数学期望与方差相等, 都等于参数 λ . 因为泊松分布只含一个参数 λ , 只要知道它的数学期望或方差就能完全确定它的分布了. $\square$

例 4 设随机变量 $X \sim U(a, b)$, 求 $D(X)$.

解 X 的概率密度为

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

上节例 7 已算得 $E(X) = \frac{a+b}{2}$. 方差为

$$\begin{aligned} D(X) &= E(X^2) - [E(X)]^2 \\ &= \int_a^b x^2 \frac{1}{b-a} dx - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}. \end{aligned} \quad \square$$

例 5 设随机变量 X 服从指数分布, 其概率密度为

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

其中 $\theta > 0$, 求 $E(X)$, $D(X)$.

解

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) dx = \int_0^{\infty} x \frac{1}{\theta} e^{-x/\theta} dx \\ &= -x e^{-x/\theta} \Big|_0^{\infty} + \int_0^{\infty} e^{-x/\theta} dx = \theta, \\ E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^{\infty} x^2 \frac{1}{\theta} e^{-x/\theta} dx \\ &= -x^2 e^{-x/\theta} \Big|_0^{\infty} + \int_0^{\infty} 2x e^{-x/\theta} dx = 2\theta^2, \end{aligned}$$

于是

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

即有

$$E(X)=\theta, \quad D(X)=\theta^2.$$

□

现在来证明方差的几个重要性质(以下设所遇到的随机变量其方差存在).

1° 设 C 是常数, 则 $D(C)=0$.

2° 设 X 是随机变量, C 是常数, 则有

$$D(CX)=C^2 D(X), \quad D(X+C)=D(X).$$

3° 设 X, Y 是两个随机变量, 则有

$$D(X+Y)=D(X)+D(Y)+2E\{(X-E(X))(Y-E(Y))\}. \quad (2.5)$$

特别, 若 X, Y 相互独立, 则有

$$D(X+Y)=D(X)+D(Y). \quad (2.6)$$

这一性质可以推广到任意有限多个相互独立的随机变量之和的情况.

4° $D(X)=0$ 的充要条件是 X 以概率 1 取常数 $E(X)$, 即

$$P\{X=E(X)\}=1.$$

证 1° $D(C)=E\{[C-E(C)]^2\}=0$.

2° $D(CX)=E\{[CX-E(CX)]^2\}=C^2 E\{[X-E(X)]^2\}=C^2 D(X)$.

$D(X+C)=E\{[X+C-E(X+C)]^2\}=E\{[X-E(X)]^2\}=D(X)$.

3° $D(X+Y)=E\{[(X+Y)-E(X+Y)]^2\}$
 $=E\{[(X-E(X))+(Y-E(Y))]^2\}$
 $=E\{(X-E(X))^2\}+E\{(Y-E(Y))^2\}$
 $+2E\{[X-E(X)][Y-E(Y)]\}$
 $=D(X)+D(Y)+2E\{[X-E(X)][Y-E(Y)]\}.$

上式右端第三项:

$$\begin{aligned} & 2E\{[X-E(X)][Y-E(Y)]\} \\ &= 2E\{XY-XE(Y)-YE(X)+E(X)E(Y)\} \\ &= 2\{E(XY)-E(X)E(Y)-E(Y)E(X)+E(X)E(Y)\} \\ &= 2\{E(XY)-E(X)E(Y)\}. \end{aligned}$$

若 X, Y 相互独立, 由数学期望的性质 4° 知道上式右端为 0, 于是

$$D(X+Y)=D(X)+D(Y).$$

4° 充分性. 设 $P\{X=E(X)\}=1$, 则有 $P\{X^2=[E(X)]^2\}=1$, 于是

$$D(X)=E(X^2)-[E(X)]^2=0.$$

必要性的证明写在切比雪夫不等式证明的后面. □

例 6 设随机变量 $X \sim b(n, p)$, 求 $E(X), D(X)$.

解 由二项分布的定义知, 随机变量 X 是 n 重伯努利试验中事件 A 发生的次数, 且在每次试验中 A 发生的概率为 p . 引入随机变量

$$X_k = \begin{cases} 1, & A \text{ 在第 } k \text{ 次试验发生,} \\ 0, & A \text{ 在第 } k \text{ 次试验不发生,} \end{cases} \quad k=1, 2, \dots, n.$$

易知 $X = X_1 + X_2 + \cdots + X_n$, (2.7)

由于 X_k 只依赖于第 k 次试验, 而各次试验相互独立, 于是 $X_1, X_2, \cdots, X_n$ 相互独立, 又知 $X_k, k=1, 2, \cdots, n$ 服从同一 $(0-1)$ 分布

X_k	0	1
p_k	$1-p$	p

(2.7) 式表明以 n, p 为参数的二项分布变量, 可分解成为 n 个相互独立且都服从以 p 为参数的 $(0-1)$ 分布的随机变量之和.

由例 2 知 $E(X_k) = p, D(X_k) = p(1-p), k=1, 2, \cdots, n$. 故知

$$E(X) = E\left(\sum_{k=1}^n X_k\right) = \sum_{k=1}^n E(X_k) = np.$$

又由于 $X_1, X_2, \cdots, X_n$ 相互独立, 得

$$D(X) = D\left(\sum_{k=1}^n X_k\right) = \sum_{k=1}^n D(X_k) = np(1-p).$$

即 $E(X) = np, D(X) = np(1-p)$. □

例 7 设随机变量 $X \sim N(\mu, \sigma^2)$, 求 $E(X), D(X)$.

解 先求标准正态变量

$$Z = \frac{X - \mu}{\sigma}$$

的数学期望和方差. Z 的概率密度为

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

$$\text{于是 } E(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-t^2/2} dt = \frac{-1}{\sqrt{2\pi}} e^{-t^2/2} \Big|_{-\infty}^{\infty} = 0,$$

$$\begin{aligned} D(Z) &= E(Z^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-t^2/2} dt \\ &= \frac{-1}{\sqrt{2\pi}} t e^{-t^2/2} \Big|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} dt = 1. \end{aligned}$$

因 $X = \mu + \sigma Z$, 即得

$$E(X) = E(\mu + \sigma Z) = \mu,$$

$$D(X) = D(\mu + \sigma Z) = D(\sigma Z) = \sigma^2 D(Z) = \sigma^2.$$

这就是说, 正态分布的概率密度中的两个参数 μ 和 σ 分别就是该分布的数学期望和均方差, 因而正态分布完全可由它的数学期望和方差所确定.

再者, 由上一章 §5 中例 1 知道, 若 $X_i \sim N(\mu_i, \sigma_i^2), i=1, 2, \cdots, n$, 且它们相互独立, 则它们的线性组合: $C_1 X_1 + C_2 X_2 + \cdots + C_n X_n$ ($C_1, C_2, \cdots, C_n$ 是不全为 0 的常数) 仍然服从正态分布, 于是由数学期望和方差的性质知道

$$C_1 X_1 + C_2 X_2 + \cdots + C_n X_n \sim N\left(\sum_{i=1}^n C_i \mu_i, \sum_{i=1}^n C_i^2 \sigma_i^2\right) \quad (2.8)$$

这一重要结果.

例如, 若 $X \sim N(1, 3)$, $Y \sim N(2, 4)$ 且 X, Y 相互独立, 则 $Z = 2X - 3Y$ 也服从正态分布, 而 $E(Z) = 2 \times 1 - 3 \times 2 = -4$, $D(Z) = D(2X - 3Y) = 4D(X) + 9D(Y) = 48$. 故有 $Z \sim N(-4, 48)$. $\square$

例 8 设活塞的直径 (以 cm 计) $X \sim N(22.40, 0.03^2)$, 气缸的直径 $Y \sim N(22.50, 0.04^2)$, X, Y 相互独立. 任取一只活塞, 任取一只气缸, 求活塞能装入气缸的概率.

解 按题意需求 $P\{X < Y\} = P\{X - Y < 0\}$. 由于

$$X - Y \sim N(-0.10, 0.0025),$$

故有 $P\{X < Y\} = P\{X - Y < 0\}$

$$\begin{aligned} &= P\left\{\frac{(X - Y) - (-0.10)}{\sqrt{0.0025}} < \frac{0 - (-0.10)}{\sqrt{0.0025}}\right\} \\ &= \Phi\left(\frac{0.10}{0.05}\right) = \Phi(2) = 0.9772. \end{aligned} \quad \square$$

下面介绍一个重要的不等式.

定理 设随机变量 X 具有数学期望 $E(X) = \mu$, 方差 $D(X) = \sigma^2$, 则对于任意正数 ϵ , 不等式

$$P\{|X - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{\epsilon^2} \quad (2.9)$$

成立.

这一不等式称为切比雪夫 (Chebyshev) 不等式.

证 我们只就连续型随机变量的情况来证明. 设 X 的概率密度为 $f(x)$, 则有 (如图 4-2)

$$\begin{aligned} P\{|X - \mu| \geq \epsilon\} &= \int_{|x - \mu| \geq \epsilon} f(x) dx \leq \int_{|x - \mu| \geq \epsilon} \frac{|x - \mu|^2}{\epsilon^2} f(x) dx \\ &\leq \frac{1}{\epsilon^2} \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \frac{\sigma^2}{\epsilon^2}. \end{aligned} \quad \square$$

切比雪夫不等式也可以写成如下的形式:

$$P\{|X - \mu| < \epsilon\} \geq 1 - \frac{\sigma^2}{\epsilon^2}. \quad (2.10)$$

切比雪夫不等式给出了在随机变量的分布未知, 而只知道 $E(X)$ 和 $D(X)$ 的情况下估计概率 $P\{|X - E(X)| < \epsilon\}$ 的界限. 例

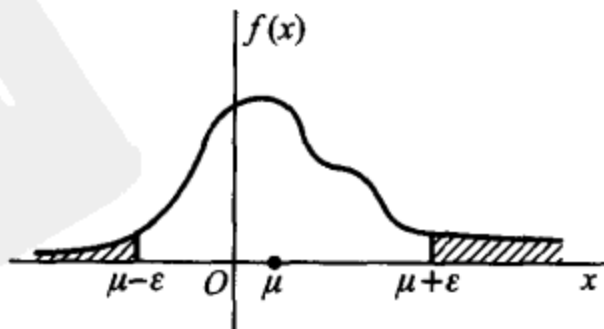


图 4-2

如在(2.9)式中分别取 $\epsilon = 3\sqrt{D(X)}, 4\sqrt{D(X)}$ 得到

$$P\{|X-E(X)| < 3\sqrt{D(X)}\} \geq 0.8889,$$

$$P\{|X-E(X)| < 4\sqrt{D(X)}\} \geq 0.9375.$$

这个估计是比较粗糙的^①, 如果已经知道随机变量的分布时, 那么所需求的概率可以确切地计算出来, 也就没有必要利用这一不等式来作估计了.

方差性质 4° 必要性的证明:

设 $D(X)=0$, 要证 $P\{X=E(X)\}=1$.

证 用反证法 假设 $P\{X=E(X)\} < 1$, 则对于某一个数 $\epsilon > 0$, 有 $P\{|X-E(X)| \geq \epsilon\} > 0$, 但由切比雪夫不等式, 对于任意 $\epsilon > 0$, 由(2.9)式因 $\sigma^2 = 0$, 有

$$P\{|X-E(X)| \geq \epsilon\} = 0,$$

矛盾, 于是 $P\{X=E(X)\}=1$. □

在书末附表 1 中列出了多种常用的随机变量的数学期望和方差, 供读者查用.

§ 3 协方差及相关系数

对于二维随机变量 (X, Y) , 我们除了讨论 X 与 Y 的数学期望和方差以外, 还需讨论描述 X 与 Y 之间相互关系的数字特征. 本节讨论有关这方面的数字特征.

在本章 § 2 方差性质 3° 的证明中, 我们已经看到, 如果两个随机变量 X 和 Y 是相互独立的, 则

$$E\{[X-E(X)][Y-E(Y)]\} = 0.$$

这意味着当 $E\{[X-E(X)][Y-E(Y)]\} \neq 0$ 时, X 与 Y 不相互独立, 而是存在着一定的关系的.

定义 量 $E\{[X-E(X)][Y-E(Y)]\}$ 称为随机变量 X 与 Y 的协方差. 记为 $\text{Cov}(X, Y)$, 即

$$\text{Cov}(X, Y) = E\{[X-E(X)][Y-E(Y)]\}.$$

而

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

称为随机变量 X 与 Y 的相关系数.

由定义, 即知

^① 例如若 $X \sim U(0, 8)$, 则有 $E(X)=4, D(X)=16/3$, 由切比雪夫不等式(2.10), $P\{|X-4| < 4\} \geq 1 - 1/3 = 2/3$, 但准确的结果是 $P\{|X-4| < 4\} = 1$.

$$\text{Cov}(X, Y) = \text{Cov}(Y, X), \quad \text{Cov}(X, X) = D(X).$$

由上述定义及(2.5)式知道, 对于任意两个随机变量 X 和 Y , 下列等式成立:

$$D(X+Y) = D(X) + D(Y) + 2\text{Cov}(X, Y). \quad (3.1)$$

将 $\text{Cov}(X, Y)$ 的定义式展开, 易得

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y). \quad (3.2)$$

我们常常利用这一式子计算协方差.

协方差具有下述性质:

1° $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$, a, b 是常数.

2° $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$.

(证明由读者自己来完成.)

下面我们来推导 ρ_{XY} 的两条重要性质, 并说明 ρ_{XY} 的含义.

考虑以 X 的线性函数 $a + bX$ 来近似表示 Y . 我们以均方误差

$$\begin{aligned} e &= E[(Y - (a + bX))^2] \\ &= E(Y^2) + b^2 E(X^2) + a^2 - 2bE(XY) + 2abE(X) - 2aE(Y) \end{aligned} \quad (3.3)$$

来衡量以 $a + bX$ 近似表达 Y 的好坏程度. e 的值越小表示 $a + bX$ 与 Y 的近似程度越好. 这样, 我们就取 a, b 使 e 取到最小. 下面就来求最佳近似式 $a + bX$ 中的 a, b . 为此, 将 e 分别关于 a, b 求偏导数, 并令它们等于零, 得

$$\begin{cases} \frac{\partial e}{\partial a} = 2a + 2bE(X) - 2E(Y) = 0, \\ \frac{\partial e}{\partial b} = 2bE(X^2) - 2E(XY) + 2aE(X) = 0. \end{cases}$$

解得

$$b_0 = \frac{\text{Cov}(X, Y)}{D(X)},$$

$$a_0 = E(Y) - b_0 E(X) = E(Y) - E(X) \frac{\text{Cov}(X, Y)}{D(X)}.$$

将 a_0, b_0 代入(3.3)式得

$$\begin{aligned} \min_{a, b} E\{[Y - (a + bX)]^2\} &= E\{[Y - (a_0 + b_0 X)]^2\} \\ &= (1 - \rho_{XY}^2) D(Y) \textcircled{1}. \end{aligned} \quad (3.4)$$

① $E[(Y - (a_0 + b_0 X))^2] = D[Y - a_0 - b_0 X] + [E(Y - a_0 - b_0 X)]^2$

$$= D(Y - b_0 X) - \left[\frac{1}{2} \frac{\partial e}{\partial a} \bigg|_{\substack{a=a_0 \\ b=b_0}} \right]^2 = D(Y - b_0 X) + 0$$

$$= D(Y) + b_0^2 D(X) - 2b_0 \text{Cov}(X, Y) = D(Y) + \frac{\text{Cov}^2(X, Y)}{D(X)} - 2 \frac{\text{Cov}^2(X, Y)}{D(X)}$$

$$= D(Y) \left[1 - \frac{\text{Cov}^2(X, Y)}{D(X)D(Y)} \right] = (1 - \rho_{XY}^2) D(Y).$$

由(3.4)式容易得到下述定理:

定理 1° $|\rho_{XY}| \leq 1$.

2° $|\rho_{XY}| = 1$ 的充要条件是, 存在常数 a, b 使

$$P\{Y = a + bX\} = 1.$$

证 1° 由(3.4)式与 $E\{[Y - (a_0 + b_0 X)]^2\}$ 及 $D(Y)$ 的非负性, 得知 $1 - \rho_{XY}^2 \geq 0$, 亦即 $|\rho_{XY}| \leq 1$.

2° 若 $|\rho_{XY}| = 1$ 由(3.4)式得

$$E\{[Y - (a_0 + b_0 X)]^2\} = 0.$$

从而 $0 = E\{[Y - (a_0 + b_0 X)]^2\} = D[Y - (a_0 + b_0 X)] + [E(Y - (a_0 + b_0 X))]^2$,
故有

$$D[Y - (a_0 + b_0 X)] = 0,$$

$$E[Y - (a_0 + b_0 X)] = 0.$$

又由方差的性质 4° 知

$$P\{Y - (a_0 + b_0 X) = 0\} = 1, \text{ 即 } P\{Y = a_0 + b_0 X\} = 1.$$

反之, 若存在常数 a^*, b^* 使

$$P\{Y = a^* + b^* X\} = 1, \text{ 即 } P\{Y - (a^* + b^* X) = 0\} = 1,$$

于是

$$P\{[Y - (a^* + b^* X)]^2 = 0\} = 1.$$

即得

$$E\{[Y - (a^* + b^* X)]^2\} = 0.$$

故有 $0 = E\{[Y - (a^* + b^* X)]^2\} \geq \min_{a, b} E\{[Y - (a + bX)]^2\}$
 $= E\{[Y - (a_0 + b_0 X)]^2\} = (1 - \rho_{XY}^2) D(Y).$

即得

$$|\rho_{XY}| = 1. \quad \square$$

由(3.4)知, 均方误差 e 是 $|\rho_{XY}|$ 的严格单调减少函数, 这样 ρ_{XY} 的含义就很明显了. 当 $|\rho_{XY}|$ 较大时 e 较小, 表明 X, Y (就线性关系来说) 联系较紧密. 特别当 $|\rho_{XY}| = 1$ 时, 由定理中的 2°, X, Y 之间以概率 1 存在着线性关系. 于是 ρ_{XY} 是一个可以用来表征 X, Y 之间线性关系紧密程度的量. 当 $|\rho_{XY}|$ 较大时, 我们通常说 X, Y 线性相关的程度较好; 当 $|\rho_{XY}|$ 较小时, 我们说, X, Y 线性相关的程度较差.

当 $\rho_{XY} = 0$ 时, 称 X 和 Y 不相关.

假设随机变量 X, Y 的相关系数 ρ_{XY} 存在. 当 X 和 Y 相互独立时, 由数学期望的性质 4° 及(3.2)式知 $\text{Cov}(X, Y) = 0$, 从而 $\rho_{XY} = 0$, 即 X, Y 不相关. 反之, 若 X, Y 不相关, X 和 Y 却不一定相互独立 (见例 1). 上述情况, 从“不相关”和“相互独立”的含义来看是明显的. 这是因为不相关只是就线性关系来说的, 而相互独立是就一般关系而言的.

不过, 从例 2 可以看到, 当 (X, Y) 服从二维正态分布时, X 和 Y 不相关与 X 和 Y 相互独立是等价的.

例 1 设 (X, Y) 的分布律为

$X \backslash Y$	-2	-1	1	2	$P\{Y=i\}$
1	0	1/4	1/4	0	1/2
4	1/4	0	0	1/4	1/2
$P\{X=i\}$	1/4	1/4	1/4	1/4	1

易知 $E(X)=0, E(Y)=5/2, E(XY)=0$, 于是 $\rho_{XY}=0$, X, Y 不相关. 这表示 X, Y 不存在线性关系. 但, $P\{X=-2, Y=1\}=0 \neq P\{X=-2\}P\{Y=1\}$, 知 X, Y 不是相互独立的. 事实上, X 和 Y 具有关系: $Y=X^2$, Y 的值完全可由 X 的值所确定. $\square$

例 2 设 (X, Y) 服从二维正态分布, 它的概率密度为

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

我们来求 X 和 Y 的相关系数.

在第三章 § 2 例 3 中已经知道 (X, Y) 的边缘概率密度为

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

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

故知 $E(X)=\mu_1, E(Y)=\mu_2, D(X)=\sigma_1^2, D(Y)=\sigma_2^2$. 而

$$\begin{aligned} \text{Cov}(X, Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_1)(y-\mu_2) f(x, y) dx dy \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_1)(y-\mu_2) \\ &\quad \times \exp \left[\frac{-1}{2(1-\rho^2)} \left(\frac{y-\mu_2}{\sigma_2} - \rho \frac{x-\mu_1}{\sigma_1} \right)^2 - \frac{(x-\mu_1)^2}{2\sigma_1^2} \right] dy dx. \end{aligned}$$

令 $t = \frac{1}{\sqrt{1-\rho^2}} \left(\frac{y-\mu_2}{\sigma_2} - \rho \frac{x-\mu_1}{\sigma_1} \right), u = \frac{x-\mu_1}{\sigma_1}$, 则有

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\sigma_1\sigma_2\sqrt{1-\rho^2}tu + \rho\sigma_1\sigma_2u^2) e^{-(u^2+t^2)/2} dt du \\ &= \frac{\rho\sigma_1\sigma_2}{2\pi} \left(\int_{-\infty}^{\infty} u^2 e^{-\frac{u^2}{2}} du \right) \left(\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt \right) \\ &\quad + \frac{\sigma_1\sigma_2\sqrt{1-\rho^2}}{2\pi} \left(\int_{-\infty}^{\infty} ue^{-\frac{u^2}{2}} du \right) \left(\int_{-\infty}^{\infty} te^{-\frac{t^2}{2}} dt \right) \end{aligned}$$

$$= \frac{\rho\sigma_1\sigma_2}{2\pi} \sqrt{2\pi} \cdot \sqrt{2\pi},$$

即有

$$\text{Cov}(X, Y) = \rho\sigma_1\sigma_2.$$

于是

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \rho.$$

□

这就是说, 二维正态随机变量 (X, Y) 的概率密度中的参数 ρ 就是 X 和 Y 的相关系数, 因而二维正态随机变量的分布完全可由 X, Y 各自的数学期望、方差以及它们的相关系数所确定.

在第三章 §4 中已经讲过, 若 (X, Y) 服从二维正态分布, 那么 X 和 Y 相互独立的充要条件为 $\rho = 0$. 现在知道 $\rho = \rho_{XY}$, 故知对于二维正态随机变量 (X, Y) 来说, X 和 Y 不相关与 X 和 Y 相互独立是等价的.

§4 矩、协方差矩阵

本节先介绍随机变量的另外几个数字特征. 设 (X, Y) 是二维随机变量.

定义 设 X 和 Y 是随机变量, 若

$$E(X^k), \quad k=1, 2, \dots$$

存在, 称它为 X 的 k 阶原点矩, 简称 k 阶矩.

若

$$E\{[X - E(X)]^k\}, \quad k=2, 3, \dots$$

存在, 称它为 X 的 k 阶中心矩.

若

$$E(X^k Y^l), \quad k, l=1, 2, \dots$$

存在, 称它为 X 和 Y 的 $k+l$ 阶混合矩.

若

$$E\{[X - E(X)]^k [Y - E(Y)]^l\}, \quad k, l=1, 2, \dots$$

存在, 称它为 X 和 Y 的 $k+l$ 阶混合中心矩.

显然, X 的数学期望 $E(X)$ 是 X 的一阶原点矩, 方差 $D(X)$ 是 X 的二阶中心矩, 协方差 $\text{Cov}(X, Y)$ 是 X 和 Y 的二阶混合中心矩.

下面介绍 n 维随机变量的协方差矩阵. 先从二维随机变量讲起.

二维随机变量 (X_1, X_2) 有四个二阶中心矩 (设它们都存在), 分别记为

$$c_{11} = E\{[X_1 - E(X_1)]^2\},$$

$$c_{12} = E\{[X_1 - E(X_1)][X_2 - E(X_2)]\},$$

$$c_{21} = E\{[X_2 - E(X_2)][X_1 - E(X_1)]\},$$

$$c_{22} = E\{[X_2 - E(X_2)]^2\}.$$

将它们排成矩阵的形式

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}.$$

这个矩阵称为随机变量 (X_1, X_2) 的协方差矩阵.

设 n 维随机变量 $(X_1, X_2, \dots, X_n)$ 的二阶混合中心矩

$$c_{ij} = \text{Cov}(X_i, X_j) = E\{[X_i - E(X_i)][X_j - E(X_j)]\}, i, j = 1, 2, \dots, n$$

都存在, 则称矩阵

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}$$

为 n 维随机变量 $(X_1, X_2, \dots, X_n)$ 的协方差矩阵. 由于 $c_{ij} = c_{ji}$ ($i \neq j; i, j = 1, 2, \dots, n$), 因而上述矩阵是一个对称矩阵.

一般, n 维随机变量的分布是不知道的, 或者是太复杂, 以致在数学上不易处理, 因此在实际应用中协方差矩阵就显得重要了.

本节的最后, 介绍 n 维正态随机变量的概率密度. 我们先将二维正态随机变量的概率密度改写成另一种形式, 以便将它推广到 n 维随机变量的场合中去. 二维正态随机变量 (X_1, X_2) 的概率密度为

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ \frac{-1}{2(1-\rho^2)} \left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right] \right\}.$$

现在将上式中花括号内的式子写成矩阵形式, 为此引入下面的列矩阵

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}.$$

(X_1, X_2) 的协方差矩阵为

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

它的行列式 $\det \mathbf{C} = \sigma_1^2\sigma_2^2(1-\rho^2)$, $\mathbf{C}$ 的逆矩阵为

$$\mathbf{C}^{-1} = \frac{1}{\det \mathbf{C}} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}.$$

经过计算可知 (这里矩阵 $(\mathbf{X}-\boldsymbol{\mu})^T$ 是 $(\mathbf{X}-\boldsymbol{\mu})$ 的转置矩阵)

$$\begin{aligned} & (\mathbf{X}-\boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{X}-\boldsymbol{\mu}) \\ &= \frac{1}{\det \mathbf{C}} (x_1-\mu_1 \quad x_2-\mu_2) \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \\ &= \frac{1}{1-\rho^2} \left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right]. \end{aligned}$$

于是 (X_1, X_2) 的概率密度可写成

$$f(x_1, x_2) = \frac{1}{(2\pi)^{2/2} (\det \mathbf{C})^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right\}.$$

上式容易推广到 n 维正态随机变量 $(X_1, X_2, \dots, X_n)$ 的情况.

引入列矩阵

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \text{ 和 } \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}.$$

n 维正态随机变量 $(X_1, X_2, \dots, X_n)$ 的概率密度定义为

$$\begin{aligned} f(x_1, x_2, \dots, x_n) \\ = \frac{1}{(2\pi)^{n/2} (\det \mathbf{C})^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right\}, \end{aligned}$$

其中 $\mathbf{C}$ 是 $(X_1, X_2, \dots, X_n)$ 的协方差矩阵.

n 维正态随机变量具有以下四条重要性质(证略):

1° n 维正态随机变量 $(X_1, X_2, \dots, X_n)$ 的每一个分量 $X_i, i=1, 2, \dots, n$ 都是正态随机变量;反之,若 $X_1, X_2, \dots, X_n$ 都是正态随机变量,且相互独立,则 $(X_1, X_2, \dots, X_n)$ 是 n 维正态随机变量.

2° n 维随机变量 $(X_1, X_2, \dots, X_n)$ 服从 n 维正态分布的充要条件是 $X_1, X_2, \dots, X_n$ 的任意的线性组合

$$l_1 X_1 + l_2 X_2 + \dots + l_n X_n$$

服从一维正态分布(其中 $l_1, l_2, \dots, l_n$ 不全为零).

3° 若 $(X_1, X_2, \dots, X_n)$ 服从 n 维正态分布,设 $Y_1, Y_2, \dots, Y_k$ 是 X_j ($j=1, 2, \dots, n$) 的线性函数,则 $(Y_1, Y_2, \dots, Y_k)$ 也服从多维正态分布.

这一性质称为正态变量的线性变换不变性.

4° 设 $(X_1, X_2, \dots, X_n)$ 服从 n 维正态分布,则“ $X_1, X_2, \dots, X_n$ 相互独立”与“ $X_1, X_2, \dots, X_n$ 两两不相关”是等价的.

n 维正态分布在随机过程和数理统计中常会遇到.

小结

随机变量的数字特征是由随机变量的分布确定的,能描述随机变量某一个方面的特征的常数.最重要的数字特征是数学期望和方差.数学期望 $E(X)$ 描述随机变量 X 取值的平均大小,方差 $D(X) = E\{[X - E(X)]^2\}$ 描述随机变量 X 与它自己的数学期望 $E(X)$ 的偏离程度.数学期望和方差在应用和理论上都非常重要.

要掌握随机变量的函数 $Y = g(X)$ 的数学期望 $E(Y) = E[g(X)]$ 的计算公式(1.3)和

(1.4). 这两个公式的意义在于当我们求 $E(Y)$ 时, 不必先求出 $Y=g(X)$ 的分布律或概率密度, 而只需利用 X 的分布律或概率密度就可以了, 这样做的好处是明显的.

要掌握数学期望和方差的性质. 提请读者注意的是:

(1) 当 X_1, X_2 独立或 X_1, X_2 不相关时, 才有 $E(X_1 X_2) = E(X_1)E(X_2)$;

(2) 设 C 为常数, 则有 $D(CX) = C^2 D(X)$, 右边的系数是 C^2 , 不是 C ;

(3) $D(X_1 + X_2) = D(X_1) + D(X_2) + 2\text{Cov}(X_1, X_2)$, 当 X_1, X_2 独立或 X_1, X_2 不相关时才有

$$D(X_1 + X_2) = D(X_1) + D(X_2).$$

例如, 若 X_1, X_2 独立, 则有 $D(2X_1 - 3X_2) = 4D(X_1) + 9D(X_2)$.

相关系数 ρ_{XY} 有时也称为线性相关系数, 它是一个可以用来描述随机变量 (X, Y) 的两个分量 X, Y 之间的线性关系紧密程度的数字特征. 当 $|\rho_{XY}|$ 较小时 X, Y 的线性相关的程度较差; 当 $\rho_{XY} = 0$ 时称 X, Y 不相关. 不相关是指 X, Y 之间不存在线性关系, X, Y 不相关, 它们还可能存在除线性关系之外的关系 (参见 §3 例 1). 又由于 X, Y 相互独立是指 X, Y 的一般关系而言的, 因此有以下的结论: X, Y 相互独立则 X, Y 一定不相关; 反之, 若 X, Y 不相关则 X, Y 不一定相互独立.

特别, 对于二维正态随机变量 (X, Y) , X 和 Y 不相关与 X 和 Y 相互独立是等价的. 而二元正态随机变量的相关系数 ρ_{XY} 就是参数 ρ . 于是, 用 “ $\rho = 0$ ” 是否成立来检验 X, Y 是否相互独立是很方便的.

切比雪夫不等式给出了在随机变量 X 的分布未知, 只知道 $E(X)$ 和 $D(X)$ 的情况下, 对事件 $\{|X - E(X)| \leq \epsilon\}$ 概率的下限的估计.

重要术语及主题

数学期望 随机变量函数的数学期望 数学期望的性质

方差 标准差 方差的性质 标准化的随机变量 协方差 相关系数 相关系数的性质

X, Y 不相关 切比雪夫不等式 几种重要分布的数学期望和方差 矩 协方差矩阵

习题

1. (1) 在下列句子中随机地取一个单词, 以 X 表示取到的单词所包含的字母个数, 写出 X 的分布律并求 $E(X)$.

“THE GIRL PUT ON HER BEAUTIFUL RED HAT”.

(2) 在上述句子的 30 个字母中随机地取一个字母, 以 Y 表示取到的字母所在单词所包含的字母数, 写出 Y 的分布律并求 $E(Y)$.

(3) 一人掷骰子, 如得 6 点则掷第 2 次, 此时得分为 6 + 第二次得到的点数; 否则得分为他第一次掷得的点数, 且不能再掷, 求得分 X 的分布律及 $E(X)$.

2. 某产品的次品率为 0.1, 检验员每天检验 4 次. 每次随机地取 10 件产品进行检验, 如发现其中的次品数多于 1, 就去调整设备. 以 X 表示一天中调整设备的次数, 试求 $E(X)$. (设诸产品是否为次品是相互独立的.)

3. 有 3 只球, 4 个盒子, 盒子的编号为 1, 2, 3, 4. 将球逐个独立地, 随机地放入 4 个盒子中去. 以 X 表示其中至少有一只球的盒子的最小号码 (例如 $X=3$ 表示第 1 号, 第 2 号盒子是

空的,第3个盒子至少有一只球),试求 $E(X)$.

4. (1) 设随机变量 X 的分布律为 $P\left\{X=(-1)^{j+1}\frac{3^j}{j}\right\}=\frac{2}{3^j}, j=1,2,\dots$, 说明 X 的数学期望不存在.

(2) 一盒中装有一只黑球,一只白球,作摸球游戏,规则如下:一次从盒中随机摸一只球,若摸到白球,则游戏结束,摸到黑球放回再放入一只黑球,然后再从盒中随机地摸一只球.试说明要游戏结束的摸球次数 X 的数学期望不存在.

5. 设在某一规定的时间间隔里,某电气设备用于最大负荷的时间 X (以 min 计)是一个随机变量,其概率密度为

$$f(x) = \begin{cases} \frac{1}{1500^2}x, & 0 \leq x \leq 1500, \\ \frac{-1}{1500^2}(x-3000), & 1500 < x \leq 3000, \\ 0, & \text{其他.} \end{cases}$$

求 $E(X)$.

6. (1) 设随机变量 X 的分布律为

X	-2	0	2
p_k	0.4	0.3	0.3

求 $E(X), E(X^2), E(3X^2+5)$.

(2) 设 $X \sim \pi(\lambda)$, 求 $E[1/(X+1)]$.

7. (1) 设随机变量 X 的概率密度为

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

求 (i) $Y=2X$; (ii) $Y=e^{-2X}$ 的数学期望.

(2) 设随机变量 $X_1, X_2, \dots, X_n$ 相互独立,且都服从 $(0,1)$ 上的均匀分布 (i) 求 $U=\max\{X_1, X_2, \dots, X_n\}$ 的数学期望, (ii) 求 $V=\min\{X_1, X_2, \dots, X_n\}$ 的数学期望.

8. 设随机变量 (X, Y) 的分布律为

$Y \backslash X$	X	1	2	3
	Y			
-1		0.2	0.1	0.0
0		0.1	0.0	0.3
1		0.1	0.1	0.1

(1) 求 $E(X), E(Y)$.

(2) 设 $Z=Y/X$, 求 $E(Z)$.

(3) 设 $Z=(X-Y)^2$, 求 $E(Z)$.

9. (1) 设随机变量 (X, Y) 的概率密度为

$$f(x, y) = \begin{cases} 12y^2, & 0 \leq y \leq x \leq 1, \\ 0, & \text{其他.} \end{cases}$$

求 $E(X), E(Y), E(XY), E(X^2 + Y^2)$.

(2) 设随机变量 X, Y 的联合密度为

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

求 $E(X), E(Y), E(XY)$.

10. (1) 设随机变量 $X \sim N(0, 1), Y \sim N(0, 1)$ 且 X, Y 相互独立. 求 $E[X^2 / (X^2 + Y^2)]$.

(2) 一飞机进行空投物资作业, 设目标点为原点 $O(0, 0)$, 物资着陆点为 (X, Y) , X, Y 相互独立, 且设 $X \sim N(0, \sigma^2), Y \sim N(0, \sigma^2)$, 求原点到点 (X, Y) 间距离的数学期望.

11. 一工厂生产的某种设备的寿命 X (以年计) 服从指数分布, 概率密度为

$$f(x) = \begin{cases} \frac{1}{4} e^{-x/4}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

工厂规定, 出售的设备若在售出一年之内损坏可予以调换. 若工厂售出一台设备赢利 100 元, 调换一台设备厂方需花费 300 元. 试求厂方出售一台设备净赢利的数学期望.

12. 某车间生产的圆盘直径在区间 (a, b) 服从均匀分布, 试求圆盘面积的数学期望.

13. 设电压 (以 V 计) $X \sim N(0, 9)$. 将电压施加于一检波器, 其输出电压为 $Y = 5X^2$, 求输出电压 Y 的均值.

14. 设随机变量 X_1, X_2 的概率密度分别为

$$f_1(x) = \begin{cases} 2e^{-2x}, & x > 0, \\ 0, & x \leq 0, \end{cases} \quad f_2(x) = \begin{cases} 4e^{-4x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

(1) 求 $E(X_1 + X_2), E(2X_1 - 3X_2^2)$.

(2) 又设 X_1, X_2 相互独立, 求 $E(X_1 X_2)$.

15. 将 n 只球 (1~ n 号) 随机地放进 n 个盒子 (1~ n 号) 中去, 一个盒子装一只球. 若一只球装入与球同号的盒子中, 称为一个配对. 记 X 为总的配对数, 求 $E(X)$.

16. 若有 n 把看上去样子相同的钥匙, 其中只有一把能打开门上的锁, 用它们去试开门上的锁. 设取到每只钥匙是等可能的. 若每把钥匙试开一次后除去, 试用下面两种方法求试开次数 X 的数学期望.

(1) 写出 X 的分布律.

(2) 不写出 X 的分布律.

17. 设 X 为随机变量, C 是常数, 证明 $D(X) < E[(X-C)^2]$, 对于 $C \neq E(X)$. (由于 $D(X) = E\{[X-E(X)]^2\}$, 上式表明 $E[(X-C)^2]$ 当 $C=E(X)$ 时取到最小值.)

18. 设随机变量 X 服从瑞利分布, 其概率密度为

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

其中 $\sigma > 0$ 是常数. 求 $E(X), D(X)$.

19. 设随机变量 X 服从 Γ 分布, 其概率密度为

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

其中 $\alpha > 0, \beta > 0$ 是常数. 求 $E(X), D(X)$.

20. 设随机变量 X 服从几何分布, 其分布律为

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

其中 $0 < p < 1$ 是常数. 求 $E(X), D(X)$.

21. 设长方形的高(以 m 计) $X \sim U(0, 2)$, 已知长方形的周长(以 m 计)为 20. 求长方形面积 A 的数学期望和方差.

22. (1) 设随机变量 X_1, X_2, X_3, X_4 相互独立, 且有 $E(X_i) = i, D(X_i) = 5 - i, i = 1, 2, 3,$

4. 设 $Y = 2X_1 - X_2 + 3X_3 - \frac{1}{2}X_4$. 求 $E(Y), D(Y)$.

(2) 设随机变量 X, Y 相互独立, 且 $X \sim N(720, 30^2), Y \sim N(640, 25^2)$, 求 $Z_1 = 2X + Y, Z_2 = X - Y$ 的分布, 并求概率 $P\{X > Y\}, P\{X + Y > 1400\}$.

23. 五家商店联营, 它们每两周售出的某种农产品的数量(以 kg 计)分别为 X_1, X_2, X_3, X_4, X_5 . 已知 $X_1 \sim N(200, 225), X_2 \sim N(240, 240), X_3 \sim N(180, 225), X_4 \sim N(260, 265), X_5 \sim N(320, 270), X_1, X_2, X_3, X_4, X_5$ 相互独立.

(1) 求五家商店两周的总销售量的均值和方差.

(2) 商店每隔两周进货一次, 为了使新的供货到达前商店不会脱销的概率大于 0.99, 问商店的仓库应至少储存多少千克该产品?

24. 卡车装运水泥, 设每袋水泥重量 X (以 kg 计) 服从 $N(50, 2.5^2)$, 问最多装多少袋水泥使总重量超过 2000 的概率不大于 0.05.

25. 设随机变量 X, Y 相互独立, 且都服从 $(0, 1)$ 上的均匀分布.

(1) 求 $E(XY), E(X/Y), E[\ln(XY)], E[|Y - X|]$.

(2) 以 X, Y 为边长作一长方形, 以 A, C 分别表示长方形的面积和周长, 求 A 和 C 的相关系数.

26. (1) 设随机变量 X_1, X_2, X_3 相互独立, 且有 $X_1 \sim b(4, 1/2), X_2 \sim b(6, 1/3), X_3 \sim b(6, 1/3)$, 求 $P\{X_1 = 2, X_2 = 2, X_3 = 5\}, E(X_1 X_2 X_3), E(X_1 - X_2), E(X_1 - 2X_2)$.

(2) 设 X, Y 是随机变量, 且有 $E(X) = 3, E(Y) = 1, D(X) = 4, D(Y) = 9$, 令 $Z = 5X - Y + 15$, 分别在下列 3 种情况下求 $E(Z)$ 和 $D(Z)$.

(i) X, Y 相互独立, (ii) X, Y 不相关, (iii) X 与 Y 的相关系数为 0.25.

27. 下列各对随机变量 X 和 Y , 问哪几对是相互独立的? 哪几对是不相关的.

(1) $X \sim U(0, 1), Y = X^2$.

(2) $X \sim U(-1, 1), Y = X^2$.

(3) $X = \cos V, Y = \sin V, V \sim U(0, 2\pi)$.

若 (X, Y) 的概率密度为 $f(x, y)$,

(4) $f(x, y) = \begin{cases} x + y, & 0 < x < 1, 0 < y < 1, \\ 0, & \text{其他.} \end{cases}$

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

28. 设二维随机变量 (X, Y) 的概率密度为

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

试验证 X 和 Y 是不相关的, 但 X 和 Y 不是相互独立的.

29. 设随机变量 (X, Y) 的分布律为

X Y	-1	0	1
-1	1/8	1/8	1/8
0	1/8	0	1/8
1	1/8	1/8	1/8

验证 X 和 Y 是不相关的, 但 X 和 Y 不是相互独立的.

30. 设 A 和 B 是试验 E 的两个事件, 且 $P(A) > 0, P(B) > 0$, 并定义随机变量 X, Y 如下:

$$X = \begin{cases} 1, & \text{若 } A \text{ 发生,} \\ 0, & \text{若 } A \text{ 不发生,} \end{cases} \quad Y = \begin{cases} 1, & \text{若 } B \text{ 发生,} \\ 0, & \text{若 } B \text{ 不发生.} \end{cases}$$

证明若 $\rho_{XY} = 0$, 则 X 和 Y 必定相互独立.

31. 设随机变量 (X, Y) 具有概率密度

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

求 $E(X), E(Y), \text{Cov}(X, Y)$.

32. 设随机变量 (X, Y) 具有概率密度

$$f(x, y) = \begin{cases} \frac{1}{8}(x+y), & 0 \leq x \leq 2, 0 \leq y \leq 2, \\ 0, & \text{其他.} \end{cases}$$

求 $E(X), E(Y), \text{Cov}(X, Y), \rho_{XY}, D(X+Y)$.

33. 设随机变量 $X \sim N(\mu, \sigma^2), Y \sim N(\mu, \sigma^2)$, 且设 X, Y 相互独立, 试求 $Z_1 = \alpha X + \beta Y$ 和 $Z_2 = \alpha X - \beta Y$ 的相关系数 (其中 α, β 是不为零的常数).

34. (1) 设随机变量 $W = (aX + 3Y)^2, E(X) = E(Y) = 0, D(X) = 4, D(Y) = 16, \rho_{XY} = -0.5$. 求常数 a 使 $E(W)$ 为最小, 并求 $E(W)$ 的最小值.

(2) 设随机变量 (X, Y) 服从二维正态分布, 且有 $D(X) = \sigma_X^2, D(Y) = \sigma_Y^2$. 证明当 $a^2 = \sigma_X^2 / \sigma_Y^2$ 时, 随机变量 $W = X - aY$ 与 $V = X + aY$ 相互独立.

35. 设随机变量 (X, Y) 服从二维正态分布, 且 $X \sim N(0, 3), Y \sim N(0, 4)$, 相关系数 $\rho_{XY} = -1/4$, 试写出 X 和 Y 的联合概率密度.

36. 已知正常男性成人血液中, 每一毫升白细胞数平均是 7 300, 均方差是 700. 利用切比雪夫不等式估计每毫升含白细胞数在 5 200 ~ 9 400 之间的概率 p .

37. 对于两个随机变量 V, W , 若 $E(V^2), E(W^2)$ 存在, 证明

$$[E(VW)]^2 \leq E(V^2)E(W^2). \quad (A)$$

这一不等式称为柯西-施瓦茨(Cauchy-Schwarz)不等式.

提示:考虑实变量 t 的函数

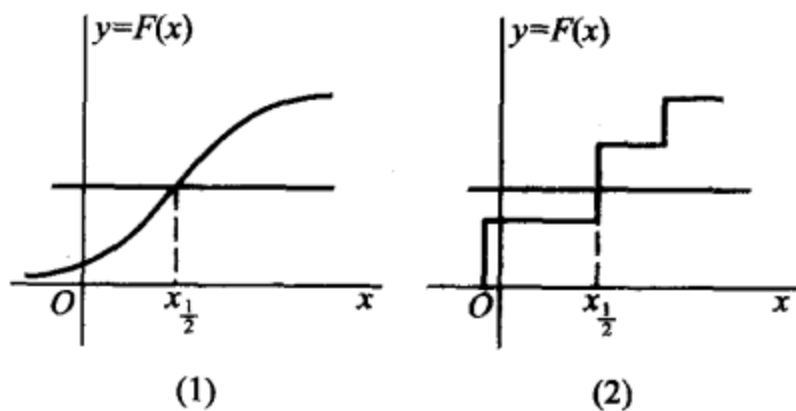
$$q(t) = E[(V+tW)^2] = E(V^2) + 2tE(VW) + t^2E(W^2).$$

38. 中位数.

对于任意随机变量 X , 满足以下两式

$$P\{X \leq x\} \geq \frac{1}{2}, \quad P\{X \geq x\} \leq \frac{1}{2}$$

的 x 称为 X 的中位数, 记为 $x_{\frac{1}{2}}$ 或 M . 它是反映集中位置的一个数字特征, 中位数总是存在的. 画出 X 的分布函数 $F(x)$ 的图. 如果 $F(x)$ 连续, 那么 $x_{\frac{1}{2}}$ 是方程 $F(x) = \frac{1}{2}$ 的解 (如题 38 图(1)), 如果 $F(x)$ 有跳跃点 (见题 38 图(2)), 用垂直于横轴的线段联结后, 得一连续曲线, 它与直线 $y=1/2$ 的交点的横坐标即为 $x_{\frac{1}{2}}$.



题 38 图

(1) 设 X 的概率密度为

$$f(x) = \begin{cases} 2e^{-2x}, & x \geq 0, \\ 0, & \text{其他.} \end{cases}$$

试求 X 的中位数 M .

(2) 设 X 服从柯西分布, 其概率密度为

$$f(x) = \frac{b}{\pi[(x-a)^2 + b^2]}, \quad b > 0.$$

试求 X 的中位数 M .

第五章 大数定律及中心极限定理

极限定理是概率论的基本理论,在理论研究和应用中起着重要的作用,其中最重要的是称为“大数定律”与“中心极限定理”的一些定理.大数定律是叙述随机变量序列的前一些项的算术平均值在某种条件下收敛到这些项的均值的算术平均值;中心极限定理则是确定在什么条件下,大量随机变量之和的分布逼近于正态分布.本章介绍几个大数定理和中心极限定理.

§1 大数定律

第一章曾讲过,大量试验证实,随机事件 A 的频率 $f_n(A)$ 当重复试验的次数 n 增大时总呈现出稳定性,稳定在某一个常数的附近.频率的稳定性是概率定义的客观基础.本节我们将对频率的稳定性作出理论的说明.

弱大数定理(辛钦大数定理) 设 $X_1, X_2, \dots$ 是相互独立^①,服从同一分布的随机变量序列,且具有数学期望 $E(X_k) = \mu$ ($k=1, 2, \dots$). 作前 n 个变量的算术平均 $\frac{1}{n} \sum_{k=1}^n X_k$, 则对于任意 $\epsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_{k=1}^n X_k - \mu \right| < \epsilon \right\} = 1. \quad (1.1)$$

证 我们只在随机变量的方差 $D(X_k) = \sigma^2$ ($k=1, 2, \dots$) 存在这一条件下证明上述结果. 因为

$$E\left(\frac{1}{n} \sum_{k=1}^n X_k\right) = \frac{1}{n} \sum_{k=1}^n E(X_k) = \frac{1}{n} (n\mu) = \mu,$$

又由独立性得

$$D\left(\frac{1}{n} \sum_{k=1}^n X_k\right) = \frac{1}{n^2} \sum_{k=1}^n D(X_k) = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n},$$

由切比雪夫不等式(见第四章(2.9)式)得

$$1 \geq P \left\{ \left| \frac{1}{n} \sum_{k=1}^n X_k - \mu \right| < \epsilon \right\} \geq 1 - \frac{\sigma^2/n}{\epsilon^2}.$$

在上式中令 $n \rightarrow \infty$, 即得

^① 是指对于任意 $n > 1$, $X_1, X_2, \dots, X_n$ 是相互独立的.

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_{k=1}^n X_k - \mu \right| < \varepsilon \right\} = 1. \quad \square$$

$\left\{ \left| \frac{1}{n} \sum_{k=1}^n X_k - \mu \right| < \varepsilon \right\}$ 是一个随机事件. 等式(1.1)式表明, 当 $n \rightarrow \infty$ 时这个事件的概率趋于 1. 即对于任意正数 ε , 当 n 充分大时, 不等式 $\left| \frac{1}{n} \sum_{k=1}^n X_k - \mu \right| < \varepsilon$ 成立的概率很大. 通俗地说, 辛钦大数定理是说, 对于独立同分布且具有均值 μ 的随机变量 $X_1, \dots, X_n$, 当 n 很大时它们的算术平均 $\frac{1}{n} \sum_{k=1}^n X_k$ 很可能接近于 μ .

设 $Y_1, Y_2, \dots, Y_n, \dots$ 是一个随机变量序列, a 是一个常数. 若对于任意正数 ε , 有

$$\lim_{n \rightarrow \infty} P \{ |Y_n - a| < \varepsilon \} = 1,$$

则称序列 $Y_1, Y_2, \dots, Y_n, \dots$ 依概率收敛于 a , 记为

$$Y_n \xrightarrow{P} a.$$

依概率收敛的序列有以下性质.

设 $X_n \xrightarrow{P} a, Y_n \xrightarrow{P} b$, 又设函数 $g(x, y)$ 在点 (a, b) 连续, 则

$$g(X_n, Y_n) \xrightarrow{P} g(a, b). \quad (\text{证略})$$

这样, 上述定理又可叙述为:

弱大数定理(辛钦大数定理) 设随机变量 $X_1, X_2, \dots, X_n, \dots$ 相互独立, 服从同一分布且具有数学期望 $E(X_k) = \mu$ ($k=1, 2, \dots$), 则序列 $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$ 依概率收敛于 μ , 即 $\bar{X} \xrightarrow{P} \mu$.

下面介绍辛钦大数定理的一个重要推论.

伯努利大数定理 设 f_A 是 n 次独立重复试验中事件 A 发生的次数, p 是事件 A 在每次试验中发生的概率, 则对于任意正数 $\varepsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{f_A}{n} - p \right| < \varepsilon \right\} = 1 \quad (1.2)$$

或

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{f_A}{n} - p \right| \geq \varepsilon \right\} = 0. \quad (1.2)'$$

证 因为 $f_A \sim b(n, p)$, 由第四章 §2 例 6, 有

$$f_A = X_1 + X_2 + \dots + X_n,$$

其中, $X_1, X_2, \dots, X_n$ 相互独立, 且都服从以 p 为参数的 $(0-1)$ 分布, 因而 $E(X_k) = p$ ($k=1, 2, \dots, n$), 由(1.1)式即得

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_{k=1}^n X_k - p \right| < \epsilon \right\} = 1,$$

即

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{f_A}{n} - p \right| \geq \epsilon \right\} = 0. \quad \square$$

伯努利大数定理的结果表明,对于任意 $\epsilon > 0$,只要重复独立试验的次数 n 充分大,事件 $\left\{ \left| \frac{f_A}{n} - p \right| \geq \epsilon \right\}$ 是一个小概率事件,由实际推断原理知(见第一章 § 4),这一事件实际上几乎是不发生的,即在 n 充分大时事件 $\left\{ \left| \frac{f_A}{n} - p \right| < \epsilon \right\}$ 实际上几乎是必定要发生的,亦即对于给定的任意小的正数 ϵ ,在 n 充分大时,事件“频率 $\frac{f_A}{n}$ 与概率 p 的偏差小于 ϵ ”实际上几乎是必定要发生的.这就是我们所说的频率稳定性的真正含义.由实际推断原理,在实际应用中,当试验次数很大时,便可以用事件的频率来代替事件的概率.

§ 2 中心极限定理

在客观实际中有许多随机变量,它们是由大量的相互独立的随机因素的综合影响所形成的.而其中每一个别因素在总的影响中所起的作用都是微小的.这种随机变量往往近似地服从正态分布.这种现象就是中心极限定理的客观背景.本节只介绍三个常用的中心极限定理.

定理一(独立同分布的中心极限定理) 设随机变量 $X_1, X_2, \dots, X_n, \dots$ 相互独立,服从同一分布,且具有数学期望和方差: $E(X_k) = \mu, D(X_k) = \sigma^2 > 0$ ($k = 1, 2, \dots$),则随机变量之和 $\sum_{k=1}^n X_k$ 的标准化变量

$$Y_n = \frac{\sum_{k=1}^n X_k - E\left(\sum_{k=1}^n X_k\right)}{\sqrt{D\left(\sum_{k=1}^n X_k\right)}} = \frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\sigma}$$

的分布函数 $F_n(x)$ 对于任意 x 满足

$$\begin{aligned} \lim_{n \rightarrow \infty} F_n(x) &= \lim_{n \rightarrow \infty} P \left\{ \frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\sigma} \leq x \right\} \\ &= \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \Phi(x). \end{aligned} \quad (2.1)$$

证明略.

这就是说,均值为 μ , 方差为 $\sigma^2 > 0$ 的独立同分布的随机变量 $X_1, X_2, \dots, X_n$ 之和 $\sum_{k=1}^n X_k$ 的标准化变量, 当 n 充分大时, 有

$$\frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\sigma} \stackrel{\text{近似地}}{\sim} N(0, 1). \quad (2.2)$$

在一般情况下, 很难求出 n 个随机变量之和 $\sum_{k=1}^n X_k$ 的分布函数, (2.2) 式表明, 当 n 充分大时, 可以通过 $\Phi(x)$ 给出其近似的分布. 这样, 就可以利用正态分布对 $\sum_{k=1}^n X_k$ 作理论分析或作实际计算, 其好处是明显的.

将(2.2)式左端改写成 $\frac{\frac{1}{n} \sum_{k=1}^n X_k - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$, 这样, 上述结果可写成: 当 n 充分大时,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \stackrel{\text{近似地}}{\sim} N(0, 1) \quad \text{或} \quad \bar{X} \stackrel{\text{近似地}}{\sim} N(\mu, \sigma^2/n). \quad (2.3)$$

这是独立同分布中心极限定理结果的另一个形式. 这就是说, 均值为 μ , 方差为 $\sigma^2 > 0$ 的独立同分布的随机变量 $X_1, X_2, \dots, X_n$ 的算术平均 $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$, 当 n 充分大时近似地服从均值为 μ , 方差为 σ^2/n 的正态分布. 这一结果是数理统计中大样本统计推断的基础.

定理二(李雅普诺夫(Lyapunov)定理) 设随机变量 $X_1, X_2, \dots, X_n, \dots$ 相互独立, 它们具有数学期望和方差

$$E(X_k) = \mu_k, \quad D(X_k) = \sigma_k^2 > 0, \quad k=1, 2, \dots,$$

记

$$B_n^2 = \sum_{k=1}^n \sigma_k^2.$$

若存在正数 δ , 使得当 $n \rightarrow \infty$ 时,

$$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^n E\{|X_k - \mu_k|^{2+\delta}\} \rightarrow 0,$$

则随机变量之和 $\sum_{k=1}^n X_k$ 的标准化变量

$$Z_n = \frac{\sum_{k=1}^n X_k - E\left(\sum_{k=1}^n X_k\right)}{\sqrt{D\left(\sum_{k=1}^n X_k\right)}} = \frac{\sum_{k=1}^n X_k - \sum_{k=1}^n \mu_k}{B_n}$$

的分布函数 $F_n(x)$ 对于任意 x , 满足

$$\begin{aligned}\lim_{n \rightarrow \infty} F_n(x) &= \lim_{n \rightarrow \infty} P \left\{ \frac{\sum_{k=1}^n X_k - \sum_{k=1}^n \mu_k}{B_n} \leq x \right\} \\ &= \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \Phi(x).\end{aligned}\quad (2.4)$$

证明略.

定理二表明, 在定理的条件下, 随机变量

$$Z_n = \frac{\sum_{k=1}^n X_k - \sum_{k=1}^n \mu_k}{B_n}$$

当 n 很大时, 近似地服从正态分布 $N(0, 1)$. 由此, 当 n 很大时,

$\sum_{k=1}^n X_k = B_n Z_n + \sum_{k=1}^n \mu_k$ 近似地服从正态分布 $N(\sum_{k=1}^n \mu_k, B_n^2)$. 这就是说, 无论各

个随机变量 X_k ($k=1, 2, \dots$) 服从什么分布, 只要满足定理的条件, 那么它们的

和 $\sum_{k=1}^n X_k$ 当 n 很大时, 就近似地服从正态分布. 这就是为什么正态随机变量在概

率论中占有重要地位的一个基本原因. 在很多问题中, 所考虑的随机变量可以表示成很多个独立的随机变量之和, 例如, 在任一指定时刻, 一个城市的耗电量是大量用户耗电量的总和; 一个物理实验的测量误差是由许多观察不到的、可加的微小误差所合成的, 它们往往近似地服从正态分布.

下面介绍另一个中心极限定理, 它是定理一的特殊情况.

定理三(棣莫弗—拉普拉斯(De Moivre-Laplace)定理) 设随机变量 η_n ($n=1, 2, \dots$) 服从参数为 n, p ($0 < p < 1$) 的二项分布, 则对于任意 x , 有

$$\lim_{n \rightarrow \infty} P \left\{ \frac{\eta_n - np}{\sqrt{np(1-p)}} \leq x \right\} = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \Phi(x). \quad (2.5)$$

证 由第四章 § 2 例 6 知可以将 η_n 分解成为 n 个相互独立、服从同一(0-1)分布的诸随机变量 $X_1, X_2, \dots, X_n$ 之和, 即有

$$\eta_n = \sum_{k=1}^n X_k,$$

其中 X_k ($k=1, 2, \dots, n$) 的分布律为

$$P\{X_k = i\} = p^i (1-p)^{1-i}, \quad i=0, 1.$$

由于 $E(X_k) = p, D(X_k) = p(1-p)$ ($k=1, 2, \dots, n$), 由定理一得

$$\lim_{n \rightarrow \infty} P \left\{ \frac{\eta_n - np}{\sqrt{np(1-p)}} \leq x \right\} = \lim_{n \rightarrow \infty} P \left\{ \frac{\sum_{k=1}^n X_k - np}{\sqrt{np(1-p)}} \leq x \right\}$$

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

这个定理表明,正态分布是二项分布的极限分布.当 n 充分大时,我们可以利用 (2.5) 式来计算二项分布的概率.下面举几个关于中心极限定理应用的例子.

例 1 一加法器同时收到 20 个噪声电压 V_k ($k=1,2,\dots,20$), 设它们是相互独立的随机变量,且都在区间 $(0,10)$ 上服从均匀分布.记 $V = \sum_{k=1}^{20} V_k$, 求 $P\{V > 105\}$ 的近似值.

解 易知 $E(V_k) = 5, D(V_k) = 100/12$ ($k=1,2,\dots,20$). 由定理一, 随机变量

$$Z = \frac{\sum_{k=1}^{20} V_k - 20 \times 5}{\sqrt{100/12} \sqrt{20}} = \frac{V - 20 \times 5}{\sqrt{100/12} \sqrt{20}}$$

近似服从正态分布 $N(0,1)$, 于是

$$\begin{aligned} P\{V > 105\} &= P\left\{\frac{V - 20 \times 5}{(10/\sqrt{12}) \sqrt{20}} > \frac{105 - 20 \times 5}{(10/\sqrt{12}) \sqrt{20}}\right\} \\ &= P\left\{\frac{V - 100}{(10/\sqrt{12}) \sqrt{20}} > 0.387\right\} \\ &= 1 - P\left\{\frac{V - 100}{(10/\sqrt{12}) \sqrt{20}} \leq 0.387\right\} \\ &\approx 1 - \int_{-\infty}^{0.387} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = 1 - \Phi(0.387) = 0.348. \end{aligned}$$

即有

$$P\{V > 105\} \approx 0.348. \quad \square$$

例 2 一船舶在某海区航行, 已知每遭受一次波浪的冲击, 纵摇角大于 3° 的概率为 $p=1/3$, 若船舶遭受了 90 000 次波浪冲击, 问其中有 29 500~30 500 次纵摇角度大于 3° 的概率是多少?

解 我们将船舶每遭受一次波浪冲击看作是一次试验, 并假定各次试验是独立的. 在 90 000 次波浪冲击中纵摇角度大于 3° 的次数记为 X , 则 X 是一个随机变量, 且有 $X \sim b(90\,000, 1/3)$. 其分布律为

$$P\{X=k\} = \binom{90\,000}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{90\,000-k}, \quad k=0,1,\dots,90\,000.$$

所求的概率为

$$P\{29\,500 \leq X \leq 30\,500\} = \sum_{k=29\,500}^{30\,500} \binom{90\,000}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{90\,000-k}.$$

要直接计算是麻烦的,我们利用棣莫弗—拉普拉斯定理来求它的近似值. 即有

$$\begin{aligned} P\{29\,500 \leq X \leq 30\,500\} \\ &= P\left\{ \frac{29\,500 - np}{\sqrt{np(1-p)}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{30\,500 - np}{\sqrt{np(1-p)}} \right\} \\ &\approx \int_{\frac{29\,500 - np}{\sqrt{np(1-p)}}}^{\frac{30\,500 - np}{\sqrt{np(1-p)}}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \Phi\left(\frac{30\,500 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{29\,500 - np}{\sqrt{np(1-p)}}\right), \end{aligned}$$

其中 $n=90\,000$, $p=1/3$. 即有

$$P\{29\,500 \leq X \leq 30\,500\} \approx \Phi\left(\frac{5\sqrt{2}}{2}\right) - \Phi\left(-\frac{5\sqrt{2}}{2}\right) = 0.999\,5. \quad \square$$

例 3 对于一个学生而言,来参加家长会的家长人数是一个随机变量,设一个学生无家长、1 名家长、2 名家长来参加会议的概率分别为 0.05、0.8、0.15. 若学校共有 400 名学生,设各学生参加会议的家长人数相互独立,且服从同一分布.

(1) 求参加会议的家长人数 X 超过 450 的概率;

(2) 求有 1 名家长来参加会议的学生人数不多于 340 的概率.

解 (1) 以 X_k ($k=1,2,\dots,400$) 记第 k 个学生来参加会议的家长人数,则 X_k 的分布律为

X_k	0	1	2
p_k	0.05	0.8	0.15

易知 $E(X_k) = 1.1$, $D(X_k) = 0.19$, $k=1,2,\dots,400$. 而 $X = \sum_{k=1}^{400} X_k$. 由定理一, 随机变量

$$\frac{\sum_{k=1}^{400} X_k - 400 \times 1.1}{\sqrt{400} \sqrt{0.19}} = \frac{X - 400 \times 1.1}{\sqrt{400} \sqrt{0.19}}$$

近似服从正态分布 $N(0,1)$, 于是

$$\begin{aligned} P\{X > 450\} &= P\left\{ \frac{X - 400 \times 1.1}{\sqrt{400} \sqrt{0.19}} > \frac{450 - 400 \times 1.1}{\sqrt{400} \sqrt{0.19}} \right\} \\ &= 1 - P\left\{ \frac{X - 400 \times 1.1}{\sqrt{400} \sqrt{0.19}} \leq 1.147 \right\} \\ &\approx 1 - \Phi(1.147) = 0.125\,1. \end{aligned}$$

(2) 以 Y 记有一名家长参加会议的学生人数, 则 $Y \sim b(400, 0.8)$, 由定理三

$$\begin{aligned} P\{Y \leq 340\} \\ &= P\left\{ \frac{Y - 400 \times 0.8}{\sqrt{400 \times 0.8 \times 0.2}} \leq \frac{340 - 400 \times 0.8}{\sqrt{400 \times 0.8 \times 0.2}} \right\} \end{aligned}$$

$$=P\left\{\frac{Y-400\times 0.8}{\sqrt{400\times 0.8\times 0.2}}\leq 2.5\right\}\approx\Phi(2.5)=0.9938.$$

□

小结

人们在长期实践中认识到频率具有稳定性,即当试验次数不断增大时,频率稳定在一个数的附近.这一事实显示了可以用一个数来表征事件发生的可能性的.这使人们认识到概率是客观存在的,进而由频率的性质的启发和抽象给出了概率的定义,因而频率的稳定性是概率定义的客观基础.伯努利大数定理则以严密的数学形式论证了频率的稳定性.

中心极限定理表明,在相当一般的条件下,当独立随机变量的个数不断增加时,其和的分布趋于正态分布.这一事实阐明了正态分布的重要性.也揭示了为什么在实际应用中会经常遇到正态分布,也就是揭示了产生正态分布变量的源泉.另一方面,它提供了独立同分布随机变量之和 $\sum_{k=1}^n X_k$ (其中 X_k 的方差存在) 的近似分布,只要和式中加项的个数充分大,就可以不必考虑和式中的随机变量服从什么分布,都可以用正态分布来近似,这在应用上是有效和重要的.

中心极限定理的内容包含极限,因而称它为极限定理是很自然的.又由于它在统计中的重要性,称它为中心极限定理,这是波利亚(Polya)在1920年取的名字.

重要术语及主题

依概率收敛 伯努利大数定理 辛钦大数定理 独立同分布的中心极限定理 李雅普诺夫中心极限定理 棣莫弗-拉普拉斯中心极限定理

习题

1. 据以往经验,某种电器元件的寿命服从均值为100 h的指数分布,现随机地取16只,设它们的寿命是相互独立的.求这16只元件的寿命的总和大于1920 h的概率.

2. (1) 一保险公司有10 000个汽车投保人,每个投保人索赔金额的数学期望为280美元,标准差为800美元,求索赔总金额超过2 700 000美元的概率.

(2) 一公司有50张签约保险单,各张保险单的索赔金额为 $X_i, i=1,2,\dots,50$ (以千美元计)服从韦布尔(Weibull)分布,均值 $E(X_i)=5$, 方差 $D(X_i)=6$, 求50张保险单索赔的合计金额大于300的概率(设各保险单索赔金额是相互独立的).

3. 计算器在进行加法时,将每个加数舍入最靠近它的整数,设所有舍入误差相互独立且在 $(-0.5, 0.5)$ 上服从均匀分布.

(1) 将1 500个数相加,问误差总和的绝对值超过15的概率是多少?

(2) 最多可有几个数相加使得误差总和的绝对值小于10的概率不小于0.90?

4. 设各零件的重量都是随机变量,它们相互独立,且服从相同的分布,其数学期望为0.5 kg,均方差为0.1 kg,问5 000只零件的总重量超过2 510 kg的概率是多少?

5. 有一批建筑房屋用的木柱,其中80%的长度不小于3 m,现从这批木柱中随机地取100根,求其中至少有30根短于3 m的概率.

6. 一工人修理一台机器需两个阶段,第一阶段所需时间(小时)服从均值为0.2的指数

分布,第二阶段服从均值为 0.3 的指数分布,且与第一阶段独立.现有 20 台机器需要修理,求他在 8 h 内完成的概率.

7. 一食品店有三种蛋糕出售,由于售出哪一种蛋糕是随机的,因而售出一只蛋糕的价格是一个随机变量,它取 1 元、1.2 元、1.5 元各个值的概率分别为 0.3、0.2、0.5.若售出 300 只蛋糕.

(1) 求收入至少 400 元的概率.

(2) 求售出价格为 1.2 元的蛋糕多于 60 只的概率.

8. 一复杂的系统由 100 个相互独立起作用的部件所组成,在整个运行期间每个部件损坏的概率为 0.10.为了使整个系统起作用,至少必须有 85 个部件正常工作,求整个系统起作用的概率.

9. 已知在某十字路口,一周事故发生数的数学期望为 2.2,标准差为 1.4.

(1) 以 $\bar{X}$ 表示一年(以 52 周计)此十字路口事故发生数的算术平均,试用中心极限定理求 $\bar{X}$ 的近似分布,并求 $P\{\bar{X} < 2\}$.

(2) 求一年事故发生数小于 100 的概率.

10. 某种小汽车氧化氮的排放量的数学期望为 0.9 g/km,标准差为 1.9 g/km,某汽车公司有这种小汽车 100 辆,以 $\bar{X}$ 表示这些车辆氧化氮排放量的算术平均,问当 L 为何值时 $\bar{X} > L$ 的概率不超过 0.01.

11. 随机地选取两组学生,每组 80 人,分别在两个实验室里测量某种化合物的 pH. 各人测量的结果是随机变量,它们相互独立,服从同一分布,数学期望为 5,方差为 0.3,以 $\bar{X}, \bar{Y}$ 分别表示第一组和第二组所得结果的算术平均.

(1) 求 $P\{4.9 < \bar{X} < 5.1\}$.

(2) 求 $P\{-0.1 < \bar{X} - \bar{Y} < 0.1\}$.

12. 一公寓有 200 户住户,一户住户拥有汽车辆数 X 的分布律为

X	0	1	2
p_k	0.1	0.6	0.3

问需要多少车位,才能使每辆汽车都具有一个车位的概率至少为 0.95.

13. 某种电子器件的寿命(小时)具有数学期望 μ (未知),方差 $\sigma^2 = 400$.为了估计 μ ,随机地取 n 只这种器件,在时刻 $t=0$ 投入测试(测试是相互独立的)直到失效,测得其寿命为 $X_1, X_2, \dots, X_n$,以 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 作为 μ 的估计,为使 $P\{|\bar{X} - \mu| < 1\} \geq 0.95$,问 n 至少为多少?

14. 某药厂断言,该厂生产的某种药品对于医治一种疑难血液病的治愈率为 0.8,医院任意抽查 100 个服用此药品的病人,若其中多于 75 人治愈,就接受此断言,否则就拒绝此断言.

(1) 若实际上此药品对这种疾病的治愈率是 0.8.问接受这一断言的概率是多少?

(2) 若实际上此药品对这种疾病的治愈率为 0.7,问接受这一断言的概率是多少?

第六章 样本及抽样分布

前面五章我们讲述了概率论的基本内容,随后的四章将讲述数理统计.数理统计是具有广泛应用的一个数学分支,它以概率论为理论基础,根据试验或观察得到的数据,来研究随机现象,对研究对象的客观规律性作出种种合理的估计和判断.

数理统计的内容包括:如何收集、整理数据资料;如何对所得的数据资料进行分析、研究,从而对所研究的对象的性质、特点作出推断.后者就是我们所说的统计推断问题.本书只讲述统计推断的基本内容.

在概率论中,我们所研究的随机变量,它的分布都是假设已知的,在这一前提下去研究它的性质、特点和规律性,例如求出它的数字特征,讨论随机变量函数的分布,介绍常用的各种分布等.在数理统计中,我们研究的随机变量,它的分布是未知的,或者是不完全知道的,人们是通过对所研究的随机变量进行重复独立的观察,得到许多观察值,对这些数据进行分析,从而对所研究的随机变量的分布作出种种推断的.

本章我们介绍总体、随机样本及统计量等基本概念,并着重介绍几个常用统计量及抽样分布.

§ 1 随 机 样 本

我们知道,随机试验的结果很多是可以用数来表示的,另有一些试验的结果虽是定性的,但总可以将它数量化.例如,检验某个学校学生的血型这一试验,其可能结果有O型、A型、B型、AB型4种,是定性的.如果分别以1,2,3,4依次记这4种血型,那么试验的结果就能用数来表示了.

在数理统计中,我们往往研究有关对象的某一项数量指标(例如研究某种型号灯泡的寿命这一数量指标).为此,考虑与这一数量指标相联系的随机试验,对这一数量指标进行试验或观察.我们将试验的全部可能的观察值称为**总体**,这些值不一定都不相同,数目上也不一定是有限的,每一个可能观察值称为**个体**.总体中所包含的个体的个数称为**总体的容量**.容量为有限的称为**有限总体**,容量为无限的称为**无限总体**.

例如在考察某大学一年级男生的身高这一试验中,若一年级男生共2 000人,每个男生的身高是一个可能观察值,所形成的总体中共含2 000个可

能观察值,是一个有限总体.又如考察某一湖泊中某种鱼的含汞量,所得总体也是有限总体.观察并记录某一地点每天(包括以往、现在和将来)的最高气温,或者测量一湖泊任一地点的深度,所得总体是无限总体.有些有限总体,它的容量很大,我们可以认为它是一个无限总体.例如,考察全国正在使用的某种型号灯泡的寿命所形成的总体,由于可能观察值的个数很多,就可以认为是无限总体.

总体中的每一个个体是随机试验的一个观察值,因此它是某一随机变量 X 的值,这样,一个总体对应于一个随机变量 X . 我们对总体的研究就是对一个随机变量 X 的研究, X 的分布函数和数字特征就称为总体的分布函数和数字特征.今后将不区分总体与相应的随机变量,笼统称为总体 X .

例如,我们检验自生产线出来的零件是次品还是正品,以 0 表示产品为正品,以 1 表示产品为次品.设出现次品的概率为 p (常数),那么总体是由一些“1”和一些“0”所组成,这一总体对应于一个具有参数为 p 的(0-1)分布:

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

的随机变量.我们就将它说成是(0-1)分布总体.意指总体中的观察值是(0-1)分布随机变量的值.又如上述灯泡寿命这一总体是指数分布总体,意指总体中的观察值是指数分布随机变量的值.

在实际中,总体的分布一般是未知的,或只知道它具有某种形式而其中包含着未知参数.在数理统计中,人们都是通过从总体中抽取一部分个体,根据获得的数据来对总体分布作出推断的.被抽出的部分个体叫做总体的一个样本.

所谓从总体抽取一个个体,就是对总体 X 进行一次观察并记录其结果.我们在相同的条件下对总体 X 进行 n 次重复的、独立的观察.将 n 次观察结果按试验的次序记为 $X_1, X_2, \dots, X_n$. 由于 $X_1, X_2, \dots, X_n$ 是对随机变量 X 观察的结果,且各次观察是在相同的条件下独立进行的,所以有理由认为 $X_1, X_2, \dots, X_n$ 是相互独立的,且都是与 X 具有相同分布的随机变量.这样得到的 $X_1, X_2, \dots, X_n$ 称为来自总体 X 的一个简单随机样本, n 称为这个样本的容量.以后如无特别说明,所提到的样本都是指简单随机样本.

当 n 次观察一经完成,我们就得到一组实数 $x_1, x_2, \dots, x_n$,它们依次是随机变量 $X_1, X_2, \dots, X_n$ 的观察值,称为样本值.

对于有限总体,采用放回抽样就能得到简单随机样本,但放回抽样使用起来不方便,当个体的总数 N 比要得到的样本的容量 n 大得多时,在实际中可将不放回抽样近似地当作放回抽样来处理.

至于无限总体,因抽取一个个体不影响它的分布,所以总是用不放回抽样.例如,在生产过程中,每隔一定时间抽取一个个体,抽取 n 个就得到一个简单随机样本,实验室中的记录,水文、气象等观察资料都是样本.试制新产品得到的样品的质量指标,也常被认为是样本.

综合上述,我们给出以下的定义.

定义 设 X 是具有分布函数 F 的随机变量,若 $X_1, X_2, \dots, X_n$ 是具有同一分布函数 F 的、相互独立的随机变量,则称 $X_1, X_2, \dots, X_n$ 为从分布函数 F (或总体 F 、或总体 X) 得到的容量为 n 的简单随机样本,简称样本,它们的观察值 $x_1, x_2, \dots, x_n$ 称为样本值,又称为 X 的 n 个独立的观察值.

也可以将样本看成是一个随机向量,写成 $(X_1, X_2, \dots, X_n)$, 此时样本值相应地写成 $(x_1, x_2, \dots, x_n)$. 若 $(x_1, x_2, \dots, x_n)$ 与 $(y_1, y_2, \dots, y_n)$ 都是相应于样本 $(X_1, X_2, \dots, X_n)$ 的样本值,一般来说它们是不相同的.

由定义得:若 $X_1, X_2, \dots, X_n$ 为 F 的一个样本,则 $X_1, X_2, \dots, X_n$ 相互独立,且它们的分布函数都是 F , 所以 $(X_1, X_2, \dots, X_n)$ 的分布函数为

$$F^*(x_1, x_2, \dots, x_n) = \prod_{i=1}^n F(x_i).$$

又若 X 具有概率密度 f , 则 $(X_1, X_2, \dots, X_n)$ 的概率密度为

$$f^*(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i).$$

§ 2 直方图和箱线图

为了研究总体分布的性质,人们通过试验得到许多观察值,一般来说这些数据是杂乱无章的. 为了利用它们进行统计分析,将这些数据加以整理,还常借助于表格或图形对它们加以描述. 本节将通过例子对连续型随机变量 X 引入“频率直方图.”接着介绍数据的“箱线图”. 它们使人们对总体 X 的分布有一个粗略的了解.

(一) 直方图

例 1 下面列出了 84 个伊特拉斯坎(Etruscan)人男子的头颅的最大宽度(mm), 现在来画这些数据的“频率直方图”.

141	148	132	138	154	142	150	146	155	158
150	140	147	148	144	150	149	145	149	158
143	141	144	144	126	140	144	142	141	140
145	135	147	146	141	136	140	146	142	137
148	154	137	139	143	140	131	143	141	149
148	135	148	152	143	144	141	143	147	146
150	132	142	142	143	153	149	146	149	138
142	149	142	137	134	144	146	147	140	142
140	137	152	145						

解 这些数据杂乱无章,先要将它们进行整理. 这些数据的最小值、最大值分别为 126、158,即所有数据落在区间 $[126, 158]$ 上,现取区间 $[124.5, 159.5]$,它能覆盖区间 $[126, 158]$. 将区间 $[124.5, 159.5]$ 等分为 7 个小区间^①,小区间的长度记为 Δ , $\Delta = (159.5 - 124.5)/7 = 5$. Δ 称为组距. 小区间的端点称为组限. 数出落在每个小区间内的数据的频数 f_i , 算出频率 f_i/n ($n=84, i=1, 2, \dots, 7$) 如下表:

组 限	频 数 f_i	频率 f_i/n	累积频率
124.5~129.5	1	0.011 9	0.011 9
129.5~134.5	4	0.047 6	0.059 5
134.5~139.5	10	0.119 1	0.178 6
139.5~144.5	33	0.392 9	0.571 5
144.5~149.5	24	0.285 7	0.857 2
149.5~154.5	9	0.107 1	0.952 4
154.5~159.5	3	0.035 7	1

现在自左至右依次在各个小区间上作以 $\frac{f_i}{n}/\Delta$ 为高的小矩形. 如图 6-1 所示这样的图形叫频率直方图. 显然这种小矩形的面积就等于数据落在该小区间的频率 f_i/n . 由于当 n 很大时, 频率接近于概率, 因而一般来说, 每个小区间上的小矩形面积接近于概率密度曲线之下该小区间之上的曲边梯形的面积. 于是, 一般来说, 直方图的外廓曲线接近于总体 X 的概率密度曲线. 从本例的直方图看(图 6-1), 它有一个峰, 中间高, 两头低, 比较对称. 看起来样本很像来自某一正态总体 X (在第八章中将进一步讨论). 从直方图上还可以估计 X 落在某一区间的概率, 例如从图上看到有 51.2% 的人最大头颅宽度落在区间 $(134.5, 144.5)$ 之内, 最大头颅宽度小于 129.5 的仅占 1.1% 等等.

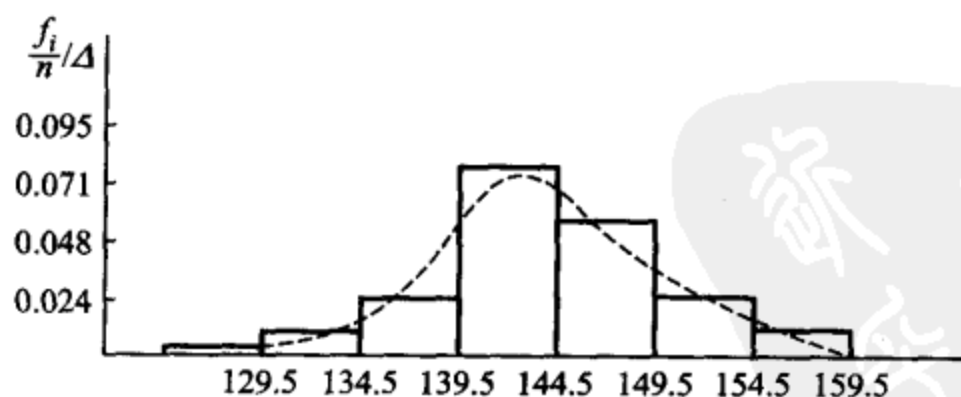


图 6-1

① 作直方图时,先取一个区间,其下限比最小的数据稍小,其上限比最大的数据稍大,然后将这一区间分为 k 个小区间,通常当 n 较大时 k 取 $10 \sim 20$, 当 $n < 50$ 时则 k 取 $5 \sim 6$. 若 k 取得过大,则会出现某些小区间内频数为零的情况(一般应设法避免). 分点通常取比数据精度高一位,以免数据落在分点上.

(二) 箱线图

先介绍样本分位数.

定义 设有容量为 n 的样本观察值 $x_1, x_2, \dots, x_n$, 样本 p 分位数 ($0 < p < 1$) 记为 x_p , 它具有以下的性质: (1) 至少有 np 个观察值小于或等于 x_p ; (2) 至少有 $n(1-p)$ 个观察值大于或等于 x_p .

样本 p 分位数可按以下法则求得. 将 $x_1, x_2, \dots, x_n$ 按自小到大的次序排列成 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$.

1° 若 np 不是整数, 则只有一个数据满足定义中的两点要求, 这一数据位于大于 np 的最小整数处, 即为位于 $[np] + 1$ 处的数. 例如, $n=12, p=0.9, np=10.8, n(1-p)=1.2$, 则 x_p 的位置应满足至少有 10.8 个数据 $\leq x_p$ (x_p 应位于第 11 或大于第 11 处); 且至少有 1.2 个数据 $\geq x_p$ (x_p 应位于第 11 或小于第 11 处), 故 x_p 应位于第 11 处.

2° 若 np 是整数. 例如在 $n=20, p=0.95$ 时, x_p 的位置应满足至少有 19 个数据 $\leq x_p$ (x_p 应位于第 19 或大于第 19 处) 且至少有 1 个数据 $\geq x_p$ (x_p 应位于第 20 或小于第 20 处), 故第 19 或第 20 的数据均符合要求, 就取这两个数的平均值作为 x_p . 综上,

$$x_p = \begin{cases} x_{([np]+1)}, & \text{当 } np \text{ 不是整数,} \\ \frac{1}{2}[x_{(np)} + x_{(np+1)}], & \text{当 } np \text{ 是整数.} \end{cases}$$

特别, 当 $p=0.5$ 时, 0.5 分位数 $x_{0.5}$ 也记为 Q_2 或 M , 称为样本中位数, 即有

$$x_{0.5} = \begin{cases} x_{([\frac{n}{2}]+1)}, & \text{当 } n \text{ 是奇数,} \\ \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}], & \text{当 } n \text{ 是偶数.} \end{cases}$$

易知, 当 n 是奇数时中位数 $x_{0.5}$ 就是 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 这一数组最中间的一个数; 而当 n 是偶数时中位数 $x_{0.5}$ 就是 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 这一数组中最中间两个数的平均值.

0.25 分位数 $x_{0.25}$ 称为第一四分位数, 又记为 Q_1 ; 0.75 分位数 $x_{0.75}$ 称为第三四分位数, 又记为 Q_3 . $x_{0.25}, x_{0.5}, x_{0.75}$ 在统计中是很有用的.

例 2 设有一组容量为 18 的样本值如下 (已经过排序)

122 126 133 140 145 145 149 150 157
162 166 175 177 177 183 188 199 212

求样本分位数: $x_{0.2}, x_{0.25}, x_{0.5}$.

解 (1) 因为 $np = 18 \times 0.2 = 3.6$, $x_{0.2}$ 位于第 $[3.6] + 1 = 4$ 处, 即有 $x_{0.2} = x_{(4)} = 140$.

(2) 因为 $np = 18 \times 0.25 = 4.5$, $x_{0.25}$ 位于第 $[4.5] + 1 = 5$ 处, 即有 $x_{0.25} = 145$.

(3) 因为 $np = 18 \times 0.5 = 9$, $x_{0.5}$ 是这组数中间两个数的平均值, 即有

$$x_{0.5} = \frac{1}{2}(157 + 162) = 159.5.$$

下面介绍箱线图.

数据集的箱线图是由箱子和直线组成的图形, 它是基于以下 5 个数的图形概括: 最小值 Min, 第一四分位数 Q_1 , 中位数 M , 第三四分位数 Q_3 和最大值 Max. 它的作法如下:

(1) 画一水平数轴, 在轴上标上 Min, Q_1 , M , Q_3 , Max. 在数轴上方画一个上、下侧平行于数轴的矩形箱子, 箱子的左右两侧分别位于 Q_1 , Q_3 的上方. 在 M 点的上方画一条垂直线段. 线段位于箱子内部.

(2) 自箱子左侧引一条水平线直至最小值 Min; 在同一水平高度自箱子右侧引一条水平线直至最大值. 这样就将箱线图作好了, 如图 6-2 所示. 箱线图也可以沿垂直数轴来作. 自箱线图可以形象地看出数据集的以下重要性质.

① 中心位置: 中位数所在的位置就是数据集的中心.

② 散布程度: 全部数据都落在 $[\text{Min}, \text{Max}]$ 之内, 在区间 $[\text{Min}, Q_1]$, $[Q_1, M]$, $[M, Q_3]$, $[Q_3, \text{Max}]$ 的数据个数各占 $1/4$. 区间较短时, 表示落在该区间的点较集中, 反之较为分散.

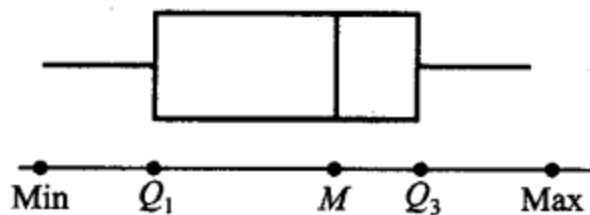


图 6-2

(3) 关于对称性: 若中位数位于箱子的中间位置, 则数据分布较为对称. 又若 Min 离 M 的距离较 Max 离 M 的距离大, 则表示数据分布向左倾斜, 反之表示数据向右倾斜, 且能看出分布尾部的长短.

例 3 以下是 8 个病人的血压(收缩压, mmHg)数据(已经过排序), 试作出箱线图.

102 110 117 118 122 123 132 150

解 因 $np = 8 \times 0.25 = 2$, 故 $Q_1 = \frac{1}{2}(110 + 117) = 113.5$.

因 $np = 8 \times 0.5 = 4$, 故 $x_{0.5} = Q_2 = \frac{1}{2}(118 + 122) = 120$.

因 $np = 8 \times 0.75 = 6$, 故 $x_{0.75} = Q_3 = \frac{1}{2}(123 + 132) = 127.5$.

Min = 102, Max = 150, 作出箱线图如图 6-3 所示.

例 4 下面分别给出了 25 个男子和 25 个女子的肺活量(以升计, 数据已经

过排序)

女子组	2.7	2.8	2.9	3.1	3.1	3.1	3.2	3.4	3.4
	3.4	3.4	3.4	3.5	3.5	3.5	3.6	3.7	3.7
	3.7	3.8	3.8	4.0	4.1	4.2	4.2		
男子组	4.1	4.1	4.3	4.3	4.5	4.6	4.7	4.8	4.8
	5.1	5.3	5.3	5.3	5.4	5.4	5.5	5.6	5.7
	5.8	5.8	6.0	6.1	6.3	6.7	6.7		

试分别画出这两组数据的箱线图.

解 女子组 $\text{Min}=2.7, \text{Max}=4.2, M=3.5$,

因 $np=25 \times 0.25=6.25, Q_1=3.2$.

因 $np=25 \times 0.75=18.75, Q_3=3.7$.

男子组 $\text{Min}=4.1, \text{Max}=6.7, M=5.3$,

因 $np=25 \times 0.25=6.25, Q_1=4.7$.

因 $np=25 \times 0.75=18.75, Q_3=5.8$.

作出箱线图如图 6-4 所示.

□

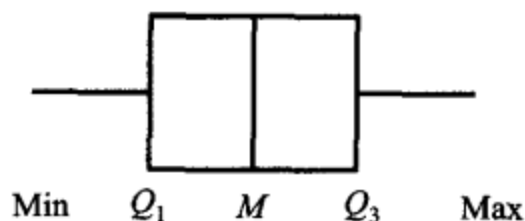


图 6-3

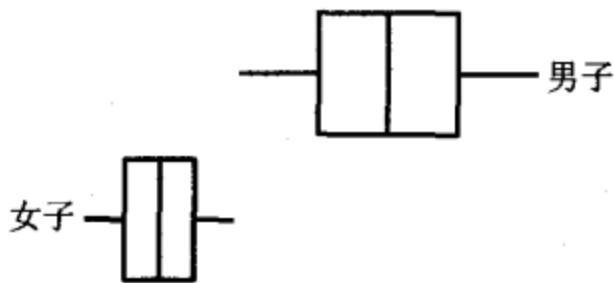


图 6-4

箱线图特别适用于比较两个或两个以上数据集的性质,为此,我们将几个数据集的箱线图画在同一个数轴上.例如在例 3 中可以明显地看到男子的肺活量要比女子大,男子的肺活量较女子的肺活量为分散.

在数据集中某一个观察值不寻常地大于或小于该数集中的其他数据,称为疑似异常值.疑似异常值的存在,会对随后的计算结果产生不适当的影响.检查疑似异常值并加以适当的处理是十分重要的.箱线图只要稍加修改,就能用来检测数据集是否存在疑似异常值.

第一四分位数 Q_1 与第三四分位数 Q_3 之间的距离: $Q_3 - Q_1$ 记为 IQR ,称为四分位数间距.若数据小于 $Q_1 - 1.5IQR$ 或大于 $Q_3 + 1.5IQR$,就认为它是疑似异常值.我们将上述箱线图的作法(1)、(2)、(3)作如下的改变:

(1')同(1)

(2')计算 $IQR = Q_3 - Q_1$,若一个数据小于 $Q_1 - 1.5IQR$ 或大于 $Q_3 +$

1. $5IQR$, 则认为它是一个疑似异常值. 画出疑似异常值, 并以 * 表示.

(3') 自箱子左侧引一水平线段直至数据集中除去疑似异常值后的最小值, 又自箱子右侧引一水平线直至数据集中除去疑似异常值后的最大值. 按(1')、(2')、(3')作出的图形称为修正箱线图.

例5 下面给出了某医院 21 个病人的住院时间(以天计), 试画出修正箱线图(数据已经过排序).

1 2 3 3 4 4 5 6 6 7 7 9 9
10 12 12 13 15 18 23 55

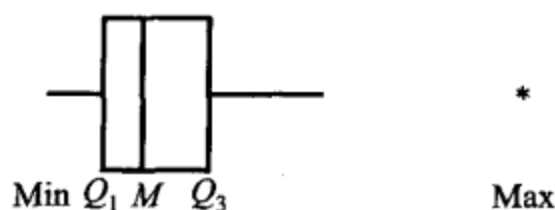


图 6-5

解 $\text{Min}=1, \text{Max}=55, M=7,$

因 $21 \times 0.25 = 5.25$, 得 $Q_1 = 4,$

又 $21 \times 0.75 = 15.75$, 得 $Q_3 = 12,$

故 $IQR = Q_3 - Q_1 = 8,$

$Q_3 + 1.5IQR = 12 + 1.5 \times 8 = 24, Q_1 - 1.5IQR = 4 - 12 = -8.$

观察值 $55 > 24$, 故 55 是疑似异常值, 且仅此一个疑似异常值. 作出修正箱线图如图 6-5 所示. 可见数据分布不对称, 而向右倾斜, 在中位数的右边较为分散. $\square$

数据集中, 疑似异常值的产生源于(1)数据的测量、记录或输入计算机时的错误;(2)数据来自不同的总体;(3)数据是正确的, 但它只体现小概率事件. 当检测出疑似异常值时, 人们需对疑似异常值出现的原因加以分析. 如果是由于测量或记录的错误, 或某些其他明显的原因造成的, 将这些疑似异常值从数据集中丢弃就可以了. 然而当出现的原因无法解释时要作出丢弃或保留这些值的决策无疑是困难的, 此时我们在对数据集作分析时尽量选用稳健的方法, 使得疑似异常值对我们的结论的影响较小. 例如我们采用中位数来描述数据集的中心趋势, 而不使用数据集的平均值, 因为后者受疑似异常值的影响较大.

§3 抽样分布

样本是进行统计推断的依据. 在应用时, 往往不是直接使用样本本身, 而是针对不同的问题构造样本的适当函数, 利用这些样本的函数进行统计推断.

定义 设 $X_1, X_2, \dots, X_n$ 是来自总体 X 的一个样本, $g(X_1, X_2, \dots, X_n)$ 是 $X_1, X_2, \dots, X_n$ 的函数, 若 g 中不含未知参数, 则称 $g(X_1, X_2, \dots, X_n)$ 是一统计量.

因为 $X_1, X_2, \dots, X_n$ 都是随机变量, 而统计量 $g(X_1, X_2, \dots, X_n)$ 是随机变量的函数, 因此统计量是一个随机变量. 设 $x_1, x_2, \dots, x_n$ 是相应于样本 $X_1, X_2, \dots, X_n$ 的样本值, 则称 $g(x_1, x_2, \dots, x_n)$ 是 $g(X_1, X_2, \dots, X_n)$ 的观察值.

下面列出几个常用的统计量. 设 $X_1, X_2, \dots, X_n$ 是来自总体 X 的一个样本, $x_1, x_2, \dots, x_n$ 是这一样本的观察值. 定义

样本平均值

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i;$$

样本方差

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right);$$

样本标准差

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2};$$

样本 k 阶(原点)矩

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k, \quad k = 1, 2, \dots;$$

样本 k 阶中心矩

$$B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k, \quad k = 2, 3, \dots.$$

它们的观察值分别为

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i;$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right);$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2};$$

$$\alpha_k = \frac{1}{n} \sum_{i=1}^n x_i^k, \quad k = 1, 2, \dots;$$

$$b_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k, \quad k = 2, 3, \dots.$$

这些观察值仍分别称为样本均值、样本方差、样本标准差、样本 k 阶(原点)矩以及样本 k 阶中心矩.

我们指出, 若总体 X 的 k 阶矩 $E(X^k) \stackrel{\text{记成}}{=} \mu_k$ 存在, 则当 $n \rightarrow \infty$ 时, $A_k \xrightarrow{P}$

$\mu_k, k=1, 2, \dots$. 这是因为 $X_1, X_2, \dots, X_n$ 独立且与 X 同分布, 所以 $X_1^k, X_2^k, \dots, X_n^k$ 独立且与 X^k 同分布. 故有

$$E(X_1^k) = E(X_2^k) = \dots = E(X_n^k) = \mu_k.$$

从而由第五章的辛钦大数定理知

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{P} \mu_k, \quad k=1, 2, \dots.$$

进而由第五章中关于依概率收敛的序列的性质知道

$$g(A_1, A_2, \dots, A_k) \xrightarrow{P} g(\mu_1, \mu_2, \dots, \mu_k),$$

其中 g 为连续函数. 这就是下一章所要介绍的矩估计法的理论根据.

经验分布函数 我们还可以作出与总体分布函数 $F(x)$ 相应的统计量——经验分布函数. 它的作法如下: 设 $X_1, X_2, \dots, X_n$ 是总体 F 的一个样本, 用 $S(x)$, $-\infty < x < \infty$ 表示 $X_1, X_2, \dots, X_n$ 中不大于 x 的随机变量的个数. 定义经验分布函数 $F_n(x)$ 为

$$F_n(x) = \frac{1}{n} S(x), \quad -\infty < x < \infty.$$

对于一个样本值, 那么经验分布函数 $F_n(x)$ 的观察值是很容易得到的 ($F_n(x)$ 的观察值仍以 $F_n(x)$ 表示). 例如

(1) 设总体 F 具有一个样本值 1, 2, 3, 则经验分布函数 $F_3(x)$ 的观察值为

$$F_3(x) = \begin{cases} 0, & \text{若 } x < 1, \\ \frac{1}{3}, & \text{若 } 1 \leq x < 2, \\ \frac{2}{3}, & \text{若 } 2 \leq x < 3, \\ 1, & \text{若 } x \geq 3. \end{cases}$$

(2) 设总体 F 具有一个样本值 1, 1, 2, 则经验分布函数 $F_3(x)$ 的观察值为

$$F_3(x) = \begin{cases} 0, & \text{若 } x < 1, \\ \frac{2}{3}, & \text{若 } 1 \leq x < 2, \\ 1, & \text{若 } x \geq 2. \end{cases}$$

一般, 设 $x_1, x_2, \dots, x_n$ 是总体 F 的一个容量为 n 的样本值. 先将 $x_1, x_2, \dots, x_n$ 按自小到大的次序排列, 并重新编号. 设为

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

则经验分布函数 $F_n(x)$ 的观察值为

$$F_n(x) = \begin{cases} 0, & \text{若 } x < x_{(1)}, \\ \frac{k}{n}, & \text{若 } x_{(k)} \leq x < x_{(k+1)}, \\ 1, & \text{若 } x \geq x_{(n)}. \end{cases}$$

对于经验分布函数 $F_n(x)$, 格里汶科 (Glivenko) 在 1933 年证明了以下的结果: 对于任一实数 x , 当 $n \rightarrow \infty$ 时 $F_n(x)$ 以概率 1 一致收敛于分布函数 $F(x)$, 即

$$P\{\lim_{n \rightarrow \infty} \sup_{-\infty < x < \infty} |F_n(x) - F(x)| = 0\} = 1.$$

因此, 对于任一实数 x 当 n 充分大时, 经验分布函数的任一个观察值 $F_n(x)$ 与总体分布函数 $F(x)$ 只有微小的差别, 从而在实际上可当作 $F(x)$ 来使用^①.

统计量的分布称为抽样分布. 在使用统计量进行统计推断时常需知道它的分布. 当总体的分布函数已知时, 抽样分布是确定的, 然而要求出统计量的精确分布, 一般来说是困难的. 本节介绍来自正态总体的几个常用统计量的分布.

(一) χ^2 分布

设 $X_1, X_2, \dots, X_n$ 是来自总体 $N(0, 1)$ 的样本, 则称统计量

$$\chi^2 = X_1^2 + X_2^2 + \dots + X_n^2 \quad (3.1)$$

服从自由度为 n 的 χ^2 分布, 记为 $\chi^2 \sim \chi^2(n)$.

此处, 自由度是指 (3.1) 式右端包含的独立变量的个数.

$\chi^2(n)$ 分布的概率密度为

$$f(y) = \begin{cases} \frac{1}{2^{n/2} \Gamma(n/2)} y^{n/2-1} e^{-y/2}, & y > 0, \\ 0, & \text{其他.} \end{cases} \quad (3.2)$$

$f(y)$ 的图形如图 6-6 所示.

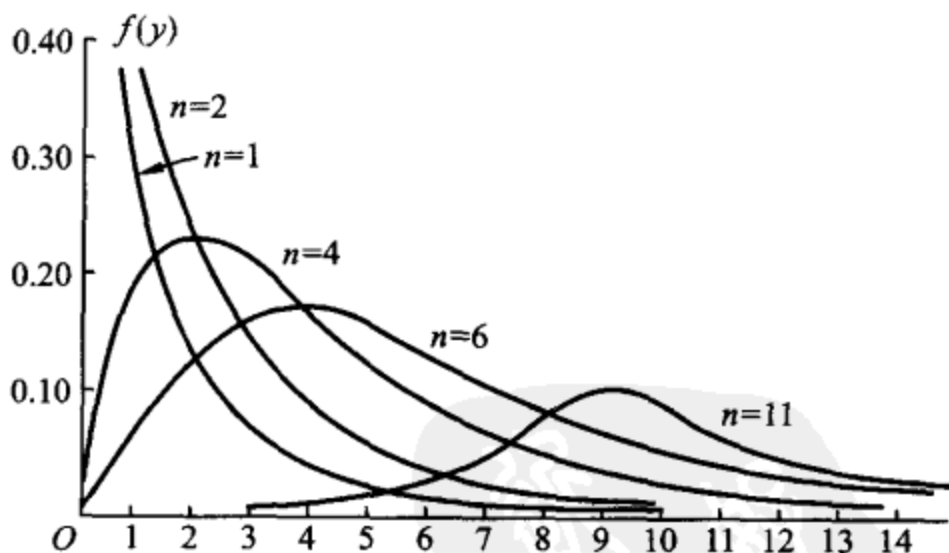


图 6-6

现在来推求 (3.2) 式.

^① 对于任意固定的 x , $-\infty < x < \infty$, $S(x) \sim b(n, F(x))$, 从而可知对于固定的 x , $E[F_n(x)] = E\left[\frac{S(x)}{n}\right] = \frac{1}{n}E[S(x)] = \frac{1}{n}[nF(x)] = F(x)$.

首先由第二章 §5 例 3 及第三章 §5 例 3 知 $\chi^2(1)$ 分布即为 $\Gamma\left(\frac{1}{2}, 2\right)$ 分布.

现 $X_i \sim N(0, 1)$, 由定义 $X_i^2 \sim \chi^2(1)$, 即 $X_i^2 \sim \Gamma\left(\frac{1}{2}, 2\right), i=1, 2, \dots, n$. 再由 $X_1, X_2, \dots, X_n$ 的独立性知 $X_1^2, X_2^2, \dots, X_n^2$ 相互独立, 从而由 Γ 分布的可加性(见第三章 §5 例 3)知

$$\chi^2 = \sum_{i=1}^n X_i^2 \sim \Gamma\left(\frac{n}{2}, 2\right), \quad (3.3)$$

即得 χ^2 的概率密度如(3.2)式所示. □

根据 Γ 分布的可加性易得 χ^2 分布的可加性如下:

χ^2 分布的可加性 设 $\chi_1^2 \sim \chi^2(n_1), \chi_2^2 \sim \chi^2(n_2)$, 并且 χ_1^2, χ_2^2 相互独立, 则有

$$\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2). \quad (3.4)$$

χ^2 分布的数学期望和方差 若 $\chi^2 \sim \chi^2(n)$, 则有

$$E(\chi^2) = n, \quad D(\chi^2) = 2n. \quad (3.5)$$

事实上, 因 $X_i \sim N(0, 1)$, 故

$$E(X_i^2) = D(X_i) = 1,$$

$$D(X_i^2) = E(X_i^4) - [E(X_i^2)]^2 = 3 - 1 = 2, i=1, 2, \dots, n.$$

于是

$$E(\chi^2) = E\left(\sum_{i=1}^n X_i^2\right) = \sum_{i=1}^n E(X_i^2) = n,$$

$$D(\chi^2) = D\left(\sum_{i=1}^n X_i^2\right) = \sum_{i=1}^n D(X_i^2) = 2n.$$

χ^2 分布的分位点 对于给定的正数 $\alpha, 0 < \alpha < 1$, 称满足条件

$$P\{\chi^2 > \chi_\alpha^2(n)\} = \int_{\chi_\alpha^2(n)}^{\infty} f(y) dy = \alpha \quad (3.6)$$

的点 $\chi_\alpha^2(n)$ 为 $\chi^2(n)$ 分布的上 α 分位点, 如图 6-7 所示. 对于不同的 α, n , 上 α 分位点的值已制成表格, 可以查用(参见附表 5). 例如对于 $\alpha=0.1, n=25$, 查得 $\chi_{0.1}^2(25) = 34.382$. 但该表只详列到 $n=40$ 为止, 费希尔(R. A. Fisher)曾证明, 当 n 充分大时, 近似地有

$$\chi_\alpha^2(n) \approx \frac{1}{2}(z_\alpha + \sqrt{2n-1})^2, \quad (3.7)$$

其中 z_α 是标准正态分布的上 α 分位点. 利用(3.7)式可以求得当 $n > 40$ 时 $\chi^2(n)$ 分布的上 α 分位点的近似值.

例如, 由(3.7)式可得 $\chi_{0.05}^2(50) \approx \frac{1}{2}(1.645 + \sqrt{99})^2 = 67.221$ (由更详细的表得 $\chi_{0.05}^2(50) = 67.505$).

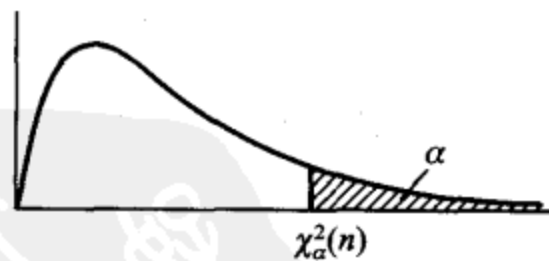


图 6-7

(二) t 分布

设 $X \sim N(0, 1)$, $Y \sim \chi^2(n)$, 且 X, Y 相互独立, 则称随机变量

$$t = \frac{X}{\sqrt{Y/n}} \quad (3.8)$$

服从自由度为 n 的 t 分布. 记为 $t \sim t(n)$.

t 分布又称学生氏 (Student) 分布. $t(n)$ 分布的概率密度函数为

$$h(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{\pi n} \Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}, \quad -\infty < t < \infty \quad (3.9)$$

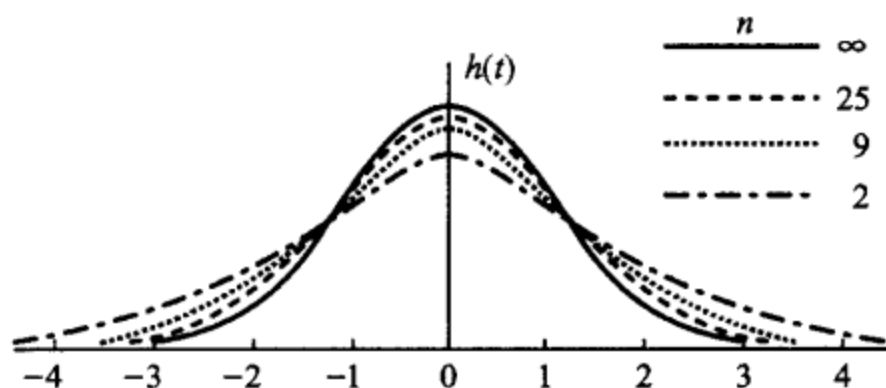


图 6-8

(证略). 图 6-8 中画出了 $h(t)$ 的图形. $h(t)$ 的图形关于 $t=0$ 对称, 当 n 充分大时其图形类似于标准正态变量概率密度的图形. 事实上, 利用 Γ 函数的性质可得

$$\lim_{n \rightarrow \infty} h(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \quad (3.10)$$

故当 n 足够大时 t 分布近似于 $N(0, 1)$ 分布. 但对于较小的 n , t 分布与 $N(0, 1)$ 分布相差较大 (见附表 2 与附表 4).

t 分布的分位点 对于给定的 $\alpha, 0 < \alpha < 1$, 称满足条件

$$P\{t > t_\alpha(n)\} = \int_{t_\alpha(n)}^{\infty} h(t) dt = \alpha \quad (3.11)$$

的点 $t_\alpha(n)$ 为 $t(n)$ 分布的上 α 分位点 (如图 6-9).

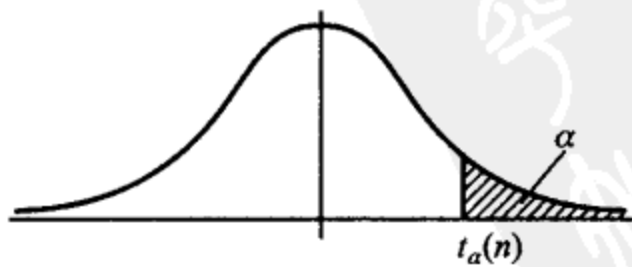


图 6-9

由 t 分布上 α 分位点的定义及 $h(t)$ 图形的对称性知

$$t_{1-\alpha}(n) = -t_{\alpha}(n). \quad (3.12)$$

t 分布的上 α 分位点可自附表 4 查得. 在 $n > 45$ 时, 对于常用的 α 的值, 就用正态近似

$$t_{\alpha}(n) \approx z_{\alpha}. \quad (3.13)$$

(三) F 分布

设 $U \sim \chi^2(n_1)$, $V \sim \chi^2(n_2)$, 且 U, V 相互独立, 则称随机变量

$$F = \frac{U/n_1}{V/n_2} \quad (3.14)$$

服从自由度为 (n_1, n_2) 的 F 分布, 记为 $F \sim F(n_1, n_2)$.

$F(n_1, n_2)$ 分布的概率密度为

$$\psi(y) = \begin{cases} \frac{\Gamma[(n_1+n_2)/2](n_1/n_2)^{n_1/2} y^{(n_1/2)-1}}{\Gamma(n_1/2)\Gamma(n_2/2)[1+(n_1 y/n_2)]^{(n_1+n_2)/2}}, & y > 0, \\ 0, & \text{其他.} \end{cases} \quad (3.15)$$

(证略). 图 6-10 中画出了 $\psi(y)$ 的图形.

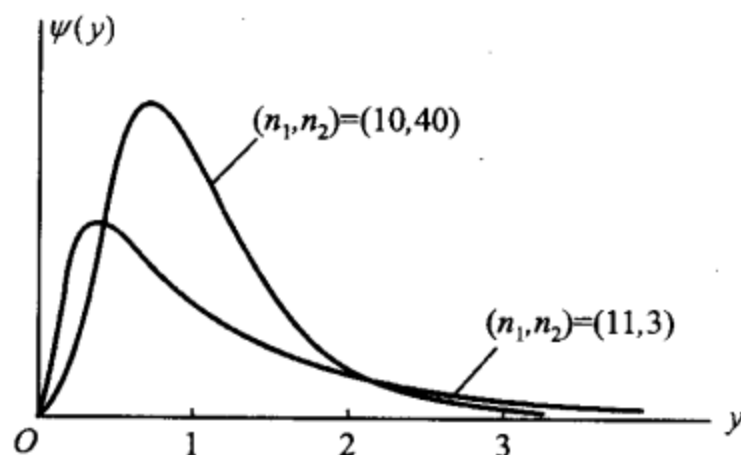


图 6-10

由定义可知, 若 $F \sim F(n_1, n_2)$, 则

$$\frac{1}{F} \sim F(n_2, n_1). \quad (3.16)$$

F 分布的分位点 对于给定的 α , $0 < \alpha < 1$, 称满足条件

$$P\{F > F_{\alpha}(n_1, n_2)\} = \int_{F_{\alpha}(n_1, n_2)}^{\infty} \psi(y) dy = \alpha \quad (3.17)$$

的点 $F_{\alpha}(n_1, n_2)$ 为 $F(n_1, n_2)$ 分布的上 α 分位点(图 6-11). F 分布的上 α 分位点有表格可查(见附表 6).

F 分布的上 α 分位点有如下的性质①:

$$F_{1-\alpha}(n_1, n_2) = \frac{1}{F_{\alpha}(n_2, n_1)}. \quad (3.18)$$

(3.18) 式常用来求 F 分布表中未列出的常用的上 α 分位点. 例如,

$$F_{0.95}(12, 9) = \frac{1}{F_{0.05}(9, 12)} = \frac{1}{2.80} = 0.357.$$

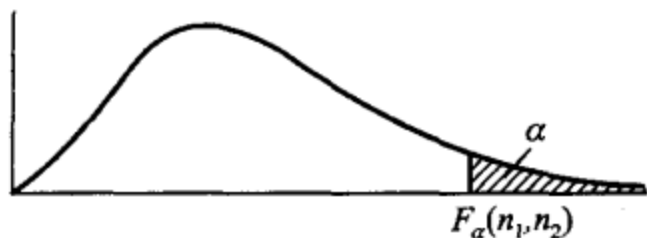


图 6-11

(四) 正态总体的样本均值与样本方差的分布

设总体 X (不管服从什么分布, 只要均值和方差存在) 的均值为 μ , 方差为 σ^2 , $X_1, X_2, \dots, X_n$ 是来自 X 的一个样本, $\bar{X}, S^2$ 分别是样本均值和样本方差, 则有

$$E(\bar{X}) = \mu, \quad D(\bar{X}) = \sigma^2/n. \quad (3.19)$$

$$\begin{aligned} \text{而} \quad E(S^2) &= E\left[\frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right)\right] = \frac{1}{n-1}\left[\sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2)\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^n (\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2)\right] = \sigma^2, \end{aligned}$$

$$\text{即} \quad E(S^2) = \sigma^2. \quad (3.20)$$

进而, 设 $X \sim N(\mu, \sigma^2)$, 由第四章 §2 的 (2.8) 式知 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 也服从正态分布, 于是得到以下的定理:

定理一 设 $X_1, X_2, \dots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本, $\bar{X}$ 是样本均值, 则有

$$\bar{X} \sim N(\mu, \sigma^2/n).$$

对于正态总体 $N(\mu, \sigma^2)$ 的样本均值 $\bar{X}$ 和样本方差 S^2 , 有以下两个重要定

① (3.18) 式的证明如下: 若 $F \sim F(n_1, n_2)$, 按定义

$$\begin{aligned} 1-\alpha &= P\{F > F_{1-\alpha}(n_1, n_2)\} = P\left\{\frac{1}{F} < \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\} \\ &= 1 - P\left\{\frac{1}{F} \geq \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\} = 1 - P\left\{\frac{1}{F} > \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\}, \end{aligned}$$

于是

$$P\left\{\frac{1}{F} > \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\} = \alpha. \quad (1)$$

再由 $\frac{1}{F} \sim F(n_2, n_1)$ 知

$$P\left\{\frac{1}{F} > F_{\alpha}(n_2, n_1)\right\} = \alpha. \quad (2)$$

比较 (1), (2) 两式得

$$\frac{1}{F_{1-\alpha}(n_1, n_2)} = F_{\alpha}(n_2, n_1), \text{ 即 } F_{1-\alpha}(n_1, n_2) = \frac{1}{F_{\alpha}(n_2, n_1)}.$$

理.

定理二 设 $X_1, X_2, \dots, X_n$ 是来自总体 $N(\mu, \sigma^2)$ 的样本, $\bar{X}, S^2$ 分别是样本均值和样本方差, 则有

$$1^\circ \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1); \quad (3.21)$$

2° $\bar{X}$ 与 S^2 相互独立.

定理的证明见本章末附录.

定理三 设 $X_1, X_2, \dots, X_n$ 是来自总体 $N(\mu, \sigma^2)$ 的样本, $\bar{X}, S^2$ 分别是样本均值和样本方差, 则有

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1). \quad (3.22)$$

证 由定理一、定理二

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

且两者独立. 由 t 分布的定义知

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \bigg/ \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} \sim t(n-1).$$

化简上式左边, 即得(3.22)式. □

对于两个正态总体的样本均值和样本方差有以下的定理.

定理四 设 $X_1, X_2, \dots, X_{n_1}$ 与 $Y_1, Y_2, \dots, Y_{n_2}$ 分别是来自正态总体 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 的样本, 且这两个样本相互独立^①. 设 $\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i$, $\bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i$ 分别是这两个样本的样本均值; $S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$, $S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$ 分别是这两个样本的样本方差, 则有

$$1^\circ \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1);$$

2° 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时,

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

其中

$$S_w^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \quad S_w = \sqrt{S_w^2}.$$

^① 是指随机向量 $(X_1, X_2, \dots, X_{n_1})$ 与 $(Y_1, Y_2, \dots, Y_{n_2})$ 相互独立.

证 1° 由定理二

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1), \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1).$$

由假设 S_1^2, S_2^2 相互独立, 则由 F 分布的定义知

$$\frac{(n_1-1)S_1^2 / (n_1-1)\sigma_1^2}{(n_2-1)S_2^2 / (n_2-1)\sigma_2^2} \sim F(n_1-1, n_2-1),$$

即

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1).$$

2° 易知 $\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$, 即有

$$U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1).$$

又由给定条件知

$$\frac{(n_1-1)S_1^2}{\sigma^2} \sim \chi^2(n_1-1), \quad \frac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_2-1),$$

且它们相互独立, 故由 χ^2 分布的可加性知

$$V = \frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_1+n_2-2).$$

由本章附录 2° 知 U 与 V 相互独立. 从而按 t 分布的定义知

$$\frac{U}{\sqrt{V/(n_1+n_2-2)}} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1+n_2-2). \quad \square$$

本节所介绍的几个分布以及四个定理, 在下面各章中都起着重要的作用. 应注意, 它们都是在总体为正态这一基本假定下得到的.

小结

在数理统计中往往研究有关对象的某一项数量指标. 对这一数量指标进行试验或观察, 将试验的全部可能的观察值称为总体, 每个观察值称为个体. 总体中的每一个个体是某一随机变量 X 的值, 因此一个总体对应一个随机变量 X . 我们将不区分总体与相应的随机变量 X , 笼统称为总体 X . 随机变量 X 服从什么分布, 就称总体服从什么分布. 在实际中遇到的总体往往是有限总体, 它对应一个离散型随机变量. 当总体中包含的个体的个数很大时, 在理论上可以认为它是一个无限总体. 我们说某种型号的灯泡寿命总体服从指数分布, 是指无限总体而言的. 又如我们说某一年龄段的男性儿童的身高服从正态分布, 也是指无限总体而言的. 无限总体是人们对具体事物的抽象. 无限总体的分布的形式较为简明, 便于在数学上进行处理, 使用方便.

在相同的条件下, 对总体 X 进行 n 次重复的、独立的观察, 得到 n 个结果 $X_1, X_2, \dots, X_n$,

称随机变量 $X_1, X_2, \dots, X_n$ 为来自总体 X 的简单随机样本,它具有两条性质:

- 1° $X_1, X_2, \dots, X_n$ 都与总体具有相同的分布;
- 2° $X_1, X_2, \dots, X_n$ 相互独立.

我们就是利用来自样本的信息推断总体,得到有关总体分布的种种结论的.

样本 $X_1, X_2, \dots, X_n$ 的函数 $g(X_1, X_2, \dots, X_n)$, 若不包含未知参数, 则称为统计量. 统计量是一个随机变量, 它是完全由样本所确定的. 统计量是进行统计推断的工具. 样本均值

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$$

和样本方差

$$S^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2$$

是两个最重要的统计量. 统计量的分布称为抽样分布. 下面是三个来自正态分布的抽样分布:

χ^2 分布, t 分布, F 分布.

这三个分布称为统计学的三大分布, 它们在数理统计中有着广泛的应用. 对于这三个分布, 要求读者掌握它们的定义和密度函数图形的轮廓, 还会使用分位点表写出分位点.

关于样本均值 $\bar{X}$ 、样本方差 S^2 , 有以下的结果.

1. 设 $X_1, X_2, \dots, X_n$ 是来自总体 X (不管服从什么分布, 只要它的均值和方差存在) 的样本, 且有 $E(X) = \mu, D(X) = \sigma^2$, 则有

$$E(\bar{X}) = \mu, \quad D(\bar{X}) = \sigma^2/n.$$

2. 设总体 $X \sim N(\mu, \sigma^2)$, $X_1, X_2, \dots, X_n$ 是来自 X 的样本, 则有

$$1^\circ \bar{X} \sim N(\mu, \sigma^2/n);$$

$$2^\circ \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1);$$

$$3^\circ \bar{X} \text{ 与 } S^2 \text{ 相互独立};$$

$$4^\circ \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

3. 对于两个正态总体 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$, 有定理四的重要结果.

■ 重要术语及主题

总体 简单随机样本 统计量

χ^2 分布、 t 分布、 F 分布的定义及它们的密度函数图形轮廓

上 α 分位点 $F_{1-\alpha}(n_1, n_2) = \frac{1}{F_\alpha(n_2, n_1)}$

小结中关于样本均值、样本方差的重要结果

附录

1° 定理二的证明

令 $Z_i = \frac{X_i - \mu}{\sigma}, i = 1, 2, \dots, n$, 则由定理二的假设知, $Z_1, Z_2, \dots, Z_n$ 相互独立, 且都服从

$N(0, 1)$ 分布, 而

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i = \frac{\bar{X} - \mu}{\sigma};$$

$$\begin{aligned} \frac{(n-1)S^2}{\sigma^2} &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^n \left[\frac{(X_i - \mu) - (\bar{X} - \mu)}{\sigma} \right]^2 \\ &= \sum_{i=1}^n (Z_i - \bar{Z})^2 = \sum_{i=1}^n Z_i^2 - n\bar{Z}^2. \end{aligned}$$

取一 n 阶正交矩阵 $A = (a_{ij})$, 其中第一行的元素均为 $1/\sqrt{n}$. 作正交变换

$$Y = AZ,$$

其中

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad Z = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}.$$

由于 $Y_i = \sum_{j=1}^n a_{ij} Z_j, i = 1, 2, \dots, n$. 故 $Y_1, Y_2, \dots, Y_n$ 仍为正态变量. 由 $Z_i \sim N(0, 1), i = 1, 2, \dots, n$ 知

$$E(Y_i) = E\left(\sum_{j=1}^n a_{ij} Z_j\right) = \sum_{j=1}^n a_{ij} E(Z_j) = 0.$$

又由 $\text{Cov}(Z_i, Z_j) = \delta_{ij}$ ($\delta_{ij} = 0$, 当 $i \neq j$; $\delta_{ij} = 1$, 当 $i = j$), $i, j = 1, 2, \dots, n$, 知

$$\begin{aligned} \text{Cov}(Y_i, Y_k) &= \text{Cov}\left(\sum_{j=1}^n a_{ij} Z_j, \sum_{l=1}^n a_{kl} Z_l\right) \\ &= \sum_{j=1}^n \sum_{l=1}^n a_{ij} a_{kl} \text{Cov}(Z_j, Z_l) = \sum_{j=1}^n a_{ij} a_{kj} = \delta_{ik} \end{aligned}$$

(由正交矩阵的性质), 故 $Y_1, Y_2, \dots, Y_n$ 两两不相关. 又由于 n 维随机变量 $(Y_1, Y_2, \dots, Y_n)$ 是由 n 维正态随机变量 $(X_1, X_2, \dots, X_n)$ 经由线性变换而得到的, 因此, $(Y_1, Y_2, \dots, Y_n)$ 也是 n 维正态随机变量 (参见第 4 章 § 4). 于是由 $Y_1, Y_2, \dots, Y_n$ 两两不相关可推得 $Y_1, Y_2, \dots, Y_n$ 相互独立 (参见第 4 章 § 4), 且有 $Y_i \sim N(0, 1), i = 1, 2, \dots, n$. 而

$$Y_1 = \sum_{j=1}^n a_{1j} Z_j = \sum_{j=1}^n \frac{1}{\sqrt{n}} Z_j = \sqrt{n} \bar{Z};$$

$$\sum_{i=1}^n Y_i^2 = Y^T Y = (AZ)^T (AZ) = Z^T (A^T A) Z = Z^T I Z = Z^T Z = \sum_{i=1}^n Z_i^2,$$

于是
$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n Z_i^2 - n\bar{Z}^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=2}^n Y_i^2.$$

由于 $Y_2, \dots, Y_n$ 相互独立, 且 $Y_i \sim N(0, 1), i = 2, 3, \dots, n$, 知 $\sum_{i=2}^n Y_i^2 \sim \chi^2(n-1)$. 从而证得

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

再者, $\bar{X} = \sigma \bar{Z} + \mu = \frac{\sigma Y_1}{\sqrt{n}} + \mu$ 仅依赖于 Y_1 , 而 $S^2 = \frac{\sigma^2}{n-1} \sum_{i=2}^n Y_i^2$ 仅依赖于 $Y_2, Y_3, \dots, Y_n$. 再

由 $Y_1, Y_2, \dots, Y_n$ 的独立性, 推知 $\bar{X}$ 与 S^2 相互独立.

2° 定理二的推广

定理二中 $\bar{X}$ 与 S^2 相互独立这一结论, 还能推广到多个同方差正态总体的情形. 例如, 对于两个同方差正态总体的情形. 设 $\bar{X}, \bar{Y}, S_1^2, S_2^2$ 是定理四 2° 中所说的正态总体 $N(\mu_1, \sigma^2), N(\mu_2, \sigma^2)$ 的样本均值和样本方差. 只要引入正交矩阵

$$T = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix},$$

其中 A_i 为 n_i 阶正交矩阵, 其第一行元素都是 $1/\sqrt{n_i}$ ($i=1, 2$), 与上面同样的做法, 考察向量

$$Z = TV$$

各分量的独立性, 其中

$$V^T = (V_1, V_2, \dots, V_n),$$

$$V_i = (X_i - \mu_1)/\sigma, i=1, 2, \dots, n_1,$$

$$V_{n_1+j} = (Y_j - \mu_2)/\sigma, j=1, 2, \dots, n_2, n_1 + n_2 = n.$$

就可证得 $\bar{X}, \bar{Y}, S_1^2, S_2^2$ 相互独立.

对于 m ($m \geq 2$) 个同方差的正态总体的情形, 设 $\bar{X}_i, S_i^2$ 分别是总体 $N(\mu_i, \sigma^2), i=1, 2, \dots, m$ 的样本均值和样本方差, 且设各样本相互独立, 则 $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_m, S_1^2, S_2^2, \dots, S_m^2$ 相互独立.

习题

1. 在总体 $N(52, 6.3^2)$ 中随机抽取一容量为 36 的样本, 求样本均值 $\bar{X}$ 落在 50.8 到 53.8 之间的概率.

2. 在总体 $N(12, 4)$ 中随机抽一容量为 5 的样本 X_1, X_2, X_3, X_4, X_5 .

(1) 求样本均值与总体均值之差的绝对值大于 1 的概率.

(2) 求概率 $P\{\max\{X_1, X_2, X_3, X_4, X_5\} > 15\}; P\{\min\{X_1, X_2, X_3, X_4, X_5\} < 10\}$.

3. 求总体 $N(20, 3)$ 的容量分别为 10, 15 的两独立样本均值差的绝对值大于 0.3 的概率.

4. (1) 设样本 $X_1, X_2, \dots, X_6$ 来自总体 $N(0, 1)$, $Y = (X_1 + X_2 + X_3)^2 + (X_4 + X_5 + X_6)^2$, 试确定常数 C 使 CY 服从 χ^2 分布.

(2) 设样本 $X_1, X_2, \dots, X_5$ 来自总体 $N(0, 1)$, $Y = \frac{C(X_1 + X_2)}{(X_3^2 + X_4^2 + X_5^2)^{1/2}}$, 试确定常数 C 使 Y 服从 t 分布.

(3) 已知 $X \sim t(n)$, 求证 $X^2 \sim F(1, n)$.

5. (1) 已知某种能力测试的得分服从正态分布 $N(\mu, \sigma^2)$, 随机取 10 个人参与这一测试. 求他们得分的联合概率密度, 并求这 10 个人得分的平均值小于 μ 的概率.

(2) 在(1)中设 $\mu = 62, \sigma^2 = 25$, 若得分超过 70 就能得奖, 求至少有一人得奖的概率.

6. 设总体 $X \sim b(1, p)$, $X_1, X_2, \dots, X_n$ 是来自 X 的样本.

(1) 求 $(X_1, X_2, \dots, X_n)$ 的分布律.

(2) 求 $\sum_{i=1}^n X_i$ 的分布律.

(3) 求 $E(\bar{X}), D(\bar{X}), E(S^2)$.

7. 设总体 $X \sim \chi^2(n)$, $X_1, X_2, \dots, X_{10}$ 是来自 X 的样本, 求 $E(\bar{X}), D(\bar{X}), E(S^2)$.

8. 设总体 $X \sim N(\mu, \sigma^2)$, $X_1, X_2, \dots, X_{10}$ 是来自 X 的样本.

(1) 写出 $X_1, X_2, \dots, X_{10}$ 的联合概率密度.

(2) 写出 $\bar{X}$ 的概率密度.

9. 设在总体 $N(\mu, \sigma^2)$ 中抽得一容量为 16 的样本, 这里 μ, σ^2 均未知.

(1) 求 $P\{S^2/\sigma^2 \leq 2.041\}$, 其中 S^2 为样本方差.

(2) 求 $D(S^2)$.

10. 下面列出了 30 个美国 NBA 球员的体重(以磅计, 1 磅 = 0.454kg)数据. 这些数据是从美国 NBA 球队 1990—1991 赛季的花名册中抽样得到的.

225	232	232	245	235	245	270	225	240	240
217	195	225	185	200	220	200	210	271	240
220	230	215	252	225	220	206	185	227	236

(1) 画出这些数据的频率直方图(提示: 最大和最小观察值分别为 271 和 185, 区间 $[184.5, 271.5]$ 包含所有数据, 将整个区间分为 5 等份, 为计算方便, 将区间调整为 $(179.5, 279.5)$).

(2) 作出这些数据的箱线图.

11. 截尾均值 设数据集包含 n 个数据, 将这些数据自小到大排序为

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)},$$

删去 $100\alpha\%$ 个数值小的数, 同时删去 $100\alpha\%$ 个数值大的数, 将留下的数据取算术平均, 记为 $\bar{x}_\alpha$, 即

$$\bar{x}_\alpha = \frac{x_{([n\alpha]+1)} + \dots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

其中 $[n\alpha]$ 是小于或等于 $n\alpha$ 的最大整数(一般取 α 为 $0.1 \sim 0.2$). $\bar{x}_\alpha$ 称为 $100\alpha\%$ 截尾均值. 例如对于第 10 题中的数据, 取 $\alpha = 0.1$, 则有 $[n\alpha] = [30 \times 0.1] = 3$, 得 $100 \times 0.1\%$ 截尾均值

$$\bar{x}_\alpha = \frac{200 + 200 + \dots + 245 + 245}{30 - 6} = 225.4167.$$

若数据来自某一总体的样本, 则 $\bar{x}_\alpha$ 是一个统计量. $\bar{x}_\alpha$ 不受样本的极端值的影响. 截尾均值在实际应用问题中是常会用到的.

试求第 10 题的 30 个数据的 $\alpha = 0.2$ 的截尾均值.



第七章 参数估计

统计推断的基本问题可以分为两大类,一类是估计问题,另一类是假设检验问题.本章讨论总体参数的点估计和区间估计.

§1 点估计

设总体 X 的分布函数的形式已知,但它的一个或多个参数未知,借助于总体 X 的一个样本来估计总体未知参数的值的问题称为参数的点估计问题.

例 1 在某炸药制造厂,一天中发生着火现象的次数 X 是一个随机变量,假设它服从以 $\lambda > 0$ 为参数的泊松分布,参数 λ 为未知.现有以下的样本值,试估计参数 λ .

着火次数 k	0	1	2	3	4	5	6	
发生 k 次着火的天数 n_k	75	90	54	22	6	2	1	$\Sigma=250$

解 由于 $X \sim \pi(\lambda)$, 故有 $\lambda = E(X)$. 我们自然想到用样本均值来估计总体的均值 $E(X)$. 现由已知数据计算得到

$$\begin{aligned}\bar{x} &= \frac{\sum_{k=0}^6 kn_k}{\sum_{k=0}^6 n_k} \\ &= \frac{1}{250} [0 \times 75 + 1 \times 90 + 2 \times 54 + 3 \times 22 + 4 \times 6 + 5 \times 2 + 6 \times 1] = 1.22,\end{aligned}$$

即 $E(X) = \lambda$ 的估计为 1.22. □

点估计问题的一般提法如下: 设总体 X 的分布函数 $F(x; \theta)$ ①的形式为已知, θ 是待估参数. $X_1, X_2, \dots, X_n$ 是 X 的一个样本, $x_1, x_2, \dots, x_n$ 是相应的一个样本值. 点估计问题就是要构造一个适当的统计量 $\hat{\theta}(X_1, X_2, \dots, X_n)$, 用它的观察值 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 作为未知参数 θ 的近似值. 我们称 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 为 θ 的估计量, 称 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 为 θ 的估计值. 在不致混淆的情况下统称估计量和估

① 多于一个未知参数时,可同样讨论.

计值为估计,并都简记为 $\hat{\theta}$.由于估计量是样本的函数.因此对于不同的样本值, θ 的估计值一般是不相同的.

例如在例1中,我们用样本均值来估计总体均值.即有估计量

$$\hat{\lambda} = E(\hat{X}) = \frac{1}{n} \sum_{k=1}^n X_k, \quad n = 250.$$

估计值

$$\hat{\lambda} = E(\hat{X}) = \frac{1}{n} \sum_{k=1}^n x_k = 1.22.$$

下面介绍两种常用的构造估计量的方法:矩估计法和最大似然估计法.

(一) 矩估计法

设 X 为连续型随机变量,其概率密度为 $f(x; \theta_1, \theta_2, \dots, \theta_k)$,或 X 为离散型随机变量,其分布律为 $P\{X=x\} = p(x; \theta_1, \theta_2, \dots, \theta_k)$,其中 $\theta_1, \theta_2, \dots, \theta_k$ 为待估参数, $X_1, X_2, \dots, X_n$ 是来自 X 的样本.假设总体 X 的前 k 阶矩

$$\mu_l = E(X^l) = \int_{-\infty}^{\infty} x^l f(x; \theta_1, \theta_2, \dots, \theta_k) dx \quad (X \text{ 连续型})$$

或

$$\mu_l = E(X^l) = \sum_{x \in R_X} x^l p(x; \theta_1, \theta_2, \dots, \theta_k) \quad (X \text{ 离散型})$$

$$l=1, 2, \dots, k$$

(其中 R_X 是 X 可能取值的范围)存在.一般来说,它们是 $\theta_1, \theta_2, \dots, \theta_k$ 的函数.基于样本矩

$$A_l = \frac{1}{n} \sum_{i=1}^n X_i^l$$

依概率收敛于相应的总体矩 $\mu_l (l=1, 2, \dots, k)$,样本矩的连续函数依概率收敛于相应的总体矩的连续函数(见第六章§3),我们就用样本矩作为相应的总体矩的估计量,而以样本矩的连续函数作为相应的总体矩的连续函数的估计量.这种估计方法称为矩估计法.矩估计法的具体做法如下:设

$$\begin{cases} \mu_1 = \mu_1(\theta_1, \theta_2, \dots, \theta_k), \\ \mu_2 = \mu_2(\theta_1, \theta_2, \dots, \theta_k), \\ \vdots \\ \mu_k = \mu_k(\theta_1, \theta_2, \dots, \theta_k). \end{cases}$$

这是一个包含 k 个未知参数 $\theta_1, \theta_2, \dots, \theta_k$ 的联立方程组.一般来说,可以从中解出 $\theta_1, \theta_2, \dots, \theta_k$,得到

$$\begin{cases} \theta_1 = \theta_1(\mu_1, \mu_2, \dots, \mu_k), \\ \theta_2 = \theta_2(\mu_1, \mu_2, \dots, \mu_k), \\ \vdots \\ \theta_k = \theta_k(\mu_1, \mu_2, \dots, \mu_k). \end{cases}$$

以 A_i 分别代替上式中的 $\mu_i, i=1, 2, \dots, k$, 就以

$$\hat{\theta}_i = \theta_i(A_1, A_2, \dots, A_k), i = 1, 2, \dots, k$$

分别作为 $\theta_i, i=1, 2, \dots, k$ 的估计量, 这种估计量称为矩估计量. 矩估计量的观察值称为矩估计值.

例 2 设总体 X 在 $[a, b]$ 上服从均匀分布, a, b 未知. $X_1, X_2, \dots, X_n$ 是来自 X 的样本, 试求 a, b 的矩估计量.

解

$$\begin{aligned}\mu_1 &= E(X) = (a+b)/2, \\ \mu_2 &= E(X^2) = D(X) + [E(X)]^2 \\ &= (b-a)^2/12 + (a+b)^2/4.\end{aligned}$$

即

$$\begin{cases} a+b = 2\mu_1, \\ b-a = \sqrt{12(\mu_2 - \mu_1^2)}. \end{cases}$$

解这一方程组得

$$a = \mu_1 - \sqrt{3(\mu_2 - \mu_1^2)}, \quad b = \mu_1 + \sqrt{3(\mu_2 - \mu_1^2)}.$$

分别以 A_1, A_2 代替 μ_1, μ_2 , 得到 a, b 的矩估计量分别为 (注意到 $\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$):

$$\hat{a} = A_1 - \sqrt{3(A_2 - A_1^2)} = \bar{X} - \sqrt{\frac{3}{n} \sum_{i=1}^n (X_i - \bar{X})^2},$$

$$\hat{b} = A_1 + \sqrt{3(A_2 - A_1^2)} = \bar{X} + \sqrt{\frac{3}{n} \sum_{i=1}^n (X_i - \bar{X})^2}. \quad \square$$

例 3 设总体 X 的均值 μ 及方差 σ^2 都存在, 且有 $\sigma^2 > 0$. 但 μ, σ^2 均为未知. 又设 $X_1, X_2, \dots, X_n$ 是来自 X 的样本. 试求 μ, σ^2 的矩估计量.

解
$$\begin{cases} \mu_1 = E(X) = \mu, \\ \mu_2 = E(X^2) = D(X) + [E(X)]^2 = \sigma^2 + \mu^2. \end{cases}$$

解得

$$\begin{cases} \mu = \mu_1, \\ \sigma^2 = \mu_2 - \mu_1^2. \end{cases}$$

分别以 A_1, A_2 代替 μ_1, μ_2 , 得 μ 和 σ^2 的矩估计量分别为

$$\hat{\mu} = A_1 = \bar{X},$$

$$\hat{\sigma}^2 = A_2 - A_1^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

所得结果表明, 总体均值与方差的矩估计量的表达式不因不同的总体分布而异.

例如, $X \sim N(\mu, \sigma^2)$, μ, σ^2 未知, 即得 μ, σ^2 的矩估计量为

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2. \quad \square$$

(二) 最大似然估计法

若总体 X 属离散型, 其分布律 $P\{X=x\}=p(x;\theta), \theta \in \Theta$ 的形式为已知, θ 为待估参数, Θ 是 θ 可能取值的范围. 设 $X_1, X_2, \dots, X_n$ 是来自 X 的样本, 则 $X_1, X_2, \dots, X_n$ 的联合分布律为

$$\prod_{i=1}^n p(x_i; \theta).$$

又设 $x_1, x_2, \dots, x_n$ 是相应于样本 $X_1, X_2, \dots, X_n$ 的一个样本值. 易知样本 $X_1, X_2, \dots, X_n$ 取到观察值 $x_1, x_2, \dots, x_n$ 的概率, 亦即事件 $\{X_1=x_1, X_2=x_2, \dots, X_n=x_n\}$ 发生的概率为

$$L(\theta) = L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n p(x_i; \theta), \theta \in \Theta. \quad (1.1)$$

这一概率随 θ 的取值而变化, 它是 θ 的函数, $L(\theta)$ 称为样本的似然函数 (注意, 这里 $x_1, x_2, \dots, x_n$ 是已知的样本值, 它们都是常数).

关于最大似然估计法, 我们有以下的直观想法: 现在已经取到样本值 $x_1, x_2, \dots, x_n$ 了, 这表明取到这一样本值的概率 $L(\theta)$ 比较大. 我们当然不会考虑那些不能使样本 $x_1, x_2, \dots, x_n$ 出现的 $\theta \in \Theta$ 作为 θ 的估计, 再者, 如果已知当 $\theta = \theta_0 \in \Theta$ 时使 $L(\theta)$ 取很大值, 而 Θ 中的其他 θ 的值使 $L(\theta)$ 取很小值, 我们自然认为取 θ_0 作为未知参数 θ 的估计值, 较为合理. 由费希尔 (R. A. Fisher) 引进的最大似然估计法, 就是固定样本观察值 $x_1, x_2, \dots, x_n$, 在 θ 取值的可能范围 Θ 内挑选使似然函数 $L(x_1, x_2, \dots, x_n; \theta)$ 达到最大的参数值 $\hat{\theta}$, 作为参数 θ 的估计值. 即取 $\hat{\theta}$ 使

$$L(x_1, x_2, \dots, x_n; \hat{\theta}) = \max_{\theta \in \Theta} L(x_1, x_2, \dots, x_n; \theta). \quad (1.2)$$

这样得到的 $\hat{\theta}$ 与样本值 $x_1, x_2, \dots, x_n$ 有关, 常记为 $\hat{\theta}(x_1, x_2, \dots, x_n)$, 称为参数 θ 的最大似然估计值, 而相应的统计量 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 称为参数 θ 的最大似然估计量.

若总体 X 属连续型, 其概率密度 $f(x; \theta), \theta \in \Theta$ 的形式已知, θ 为待估参数, Θ 是 θ 可能取值的范围. 设 $X_1, X_2, \dots, X_n$ 是来自 X 的样本, 则 $X_1, X_2, \dots, X_n$ 的联合密度为

$$\prod_{i=1}^n f(x_i, \theta).$$

设 $x_1, x_2, \dots, x_n$ 是相应于样本 $X_1, X_2, \dots, X_n$ 的一个样本值, 则随机点 $(X_1, X_2, \dots, X_n)$ 落在点 $(x_1, x_2, \dots, x_n)$ 的邻域 (边长分别为 $dx_1, dx_2, \dots, dx_n$ 的 n 维立方体) 内的概率近似地为

$$\prod_{i=1}^n f(x_i; \theta) dx_i. \quad (1.3)$$

其值随 θ 的取值而变化. 与离散型的情况一样, 我们取 θ 的估计值 $\hat{\theta}$ 使概率

(1.3) 取到最大值, 但因子 $\prod_{i=1}^n dx_i$ 不随 θ 而变, 故只需考虑函数

$$L(\theta) = L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta) \quad (1.4)$$

的最大值. 这里 $L(\theta)$ 称为样本的似然函数. 若

$$L(x_1, x_2, \dots, x_n; \hat{\theta}) = \max_{\theta \in \Theta} L(x_1, x_2, \dots, x_n; \theta),$$

则称 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 为 θ 的最大似然估计值, 称 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 为 θ 的最大似然估计量.

这样, 确定最大似然估计量的问题就归结为微分学中的求最大值的问题了.

在很多情形下, $p(x; \theta)$ 和 $f(x; \theta)$ 关于 θ 可微, 这时 $\hat{\theta}$ 常可从方程

$$\frac{d}{d\theta} L(\theta) = 0 \quad (1.5)$$

解得^①. 又因 $L(\theta)$ 与 $\ln L(\theta)$ 在同一 θ 处取到极值, 因此, θ 的最大似然估计 $\hat{\theta}$ 也可以从方程

$$\frac{d}{d\theta} \ln L(\theta) = 0 \quad (1.6)$$

求得, 而从后一方程求解往往比较方便. (1.6) 称为对数似然方程.

例 4 设 $X \sim b(1, p)$. $X_1, X_2, \dots, X_n$ 是来自 X 的一个样本, 试求参数 p 的最大似然估计量.

解 设 $x_1, x_2, \dots, x_n$ 是相应于样本 $X_1, X_2, \dots, X_n$ 的一个样本值. X 的分布律为

$$P\{X=x\} = p^x (1-p)^{1-x}, \quad x=0, 1.$$

故似然函数为

$$L(p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i},$$

而

$$\ln L(p) = \left(\sum_{i=1}^n x_i \right) \ln p + \left(n - \sum_{i=1}^n x_i \right) \ln(1-p),$$

令

$$\frac{d}{dp} \ln L(p) = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} = 0,$$

^① 这里没有提到 $L(\theta)$ 取最大值的充分条件, 但对于具体的函数是容易讨论的.

解得 p 的最大似然估计值

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}.$$

p 的最大似然估计量为

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$

我们看到这一估计量与矩估计量是相同的. □

最大似然估计法也适用于分布中含多个未知参数 $\theta_1, \theta_2, \dots, \theta_k$ 的情况. 这时, 似然函数 L 是这些未知参数的函数. 分别令

$$\frac{\partial}{\partial \theta_i} L = 0, \quad i = 1, 2, \dots, k$$

$$\text{或令} \quad \frac{\partial}{\partial \theta_i} \ln L = 0, \quad i = 1, 2, \dots, k. \quad (1.7)$$

解上述由 k 个方程组成的方程组, 即可得到各未知参数 $\theta_i (i=1, 2, \dots, k)$ 的最大似然估计值 $\hat{\theta}_i$. (1.7) 称为对数似然方程组.

例 5 设 $X \sim N(\mu, \sigma^2)$, μ, σ^2 为未知参数, $x_1, x_2, \dots, x_n$ 是来自 X 的一个样本值. 求 μ, σ^2 的最大似然估计量.

解 X 的概率密度为

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right],$$

似然函数为

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right] \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right]. \end{aligned}$$

$$\text{而} \quad \ln L = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

$$\text{令} \quad \begin{cases} \frac{\partial}{\partial \mu} \ln L = \frac{1}{\sigma^2} \left(\sum_{i=1}^n x_i - n\mu \right) = 0, \\ \frac{\partial}{\partial \sigma^2} \ln L = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (x_i - \mu)^2 = 0. \end{cases}$$

由前一式解得 $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$, 代入后一式得 $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$. 因此得 μ 和 σ^2 的最大似然估计量分别为

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

它们与相应的矩估计量相同. □

例 6 设总体 X 在 $[a, b]$ 上服从均匀分布, a, b 未知, $x_1, x_2, \dots, x_n$ 是一个样本值. 试求 a, b 的最大似然估计量.

解 记 $x_{(1)} = \min\{x_1, x_2, \dots, x_n\}$, $x_{(n)} = \max\{x_1, x_2, \dots, x_n\}$. X 的概率密度是

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

似然函数为

$$L(a, b) = \begin{cases} \frac{1}{(b-a)^n}, & a \leq x_1, x_2, \dots, x_n \leq b, \\ 0, & \text{其他.} \end{cases}$$

由于 $a \leq x_1, x_2, \dots, x_n \leq b$, 等价于 $a \leq x_{(1)}, x_{(n)} \leq b$. 似然函数可写成

$$L(a, b) = \begin{cases} \frac{1}{(b-a)^n}, & a \leq x_{(1)}, \quad b \geq x_{(n)}, \\ 0, & \text{其他.} \end{cases}$$

于是对于满足条件 $a \leq x_{(1)}, b \geq x_{(n)}$ 的任意 a, b 有

$$L(a, b) = \frac{1}{(b-a)^n} \leq \frac{1}{(x_{(n)} - x_{(1)})^n}.$$

即 $L(a, b)$ 在 $a = x_{(1)}, b = x_{(n)}$ 时取到最大值 $(x_{(n)} - x_{(1)})^{-n}$. 故 a, b 的最大似然估计值为

$$\hat{a} = x_{(1)} = \min_{1 \leq i \leq n} x_i, \quad \hat{b} = x_{(n)} = \max_{1 \leq i \leq n} x_i.$$

a, b 的最大似然估计量为

$$\hat{a} = \min_{1 \leq i \leq n} X_i, \quad \hat{b} = \max_{1 \leq i \leq n} X_i. \quad \square$$

此外, 最大似然估计具有下述性质: 设 θ 的函数 $u = u(\theta)$, $\theta \in \Theta$ 具有单值反函数 $\theta = \theta(u)$, $u \in \mathcal{U}$. 又假设 $\hat{\theta}$ 是 X 的概率分布中参数 θ 的最大似然估计, 则 $\hat{u} = u(\hat{\theta})$ 是 $u(\theta)$ 的最大似然估计. 这一性质称为最大似然估计的不变性.

事实上, 因为 $\hat{\theta}$ 是 θ 的最大似然估计, 于是有

$$L(x_1, x_2, \dots, x_n; \hat{\theta}) = \max_{\theta \in \Theta} L(x_1, x_2, \dots, x_n, \theta),$$

其中 $x_1, x_2, \dots, x_n$ 是 X 的一个样本值, 考虑到 $\hat{u} = u(\hat{\theta})$, 且有 $\hat{\theta} = \theta(\hat{u})$, 上式可写成

$$L(x_1, x_2, \dots, x_n; \theta(\hat{u})) = \max_{u \in \mathcal{U}} L(x_1, x_2, \dots, x_n; \theta(u)).$$

这就证明了 $\hat{u} = u(\hat{\theta})$ 是 $u(\theta)$ 的最大似然估计.

当总体分布中含有多个未知参数时,也具有上述性质.例如,在例 5 中已得到 σ^2 的最大似然估计为

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

函数 $u = u(\sigma^2) = \sqrt{\sigma^2}$ 有单值反函数 $\sigma^2 = u^2 (u \geq 0)$, 根据上述性质, 得到标准差 σ 的最大似然估计为

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

我们还要提到的是, 对数似然方程 (1.6) 或对数似然方程组 (1.7) 除了一些简单的情况外, 往往没有有限函数形式的解, 这就需要用数值方法求近似解. 常用的算法是牛顿-拉弗森 (Newton-Raphson) 算法, 对于 (1.7) 有时也用拟牛顿算法, 它们都是迭代算法, 读者可参考有关的参考书.

§ 2 基于截尾样本的最大似然估计

在研究产品的可靠性时, 需要研究产品寿命 T 的各种特征. 产品寿命 T 是一个随机变量, 它的分布称为寿命分布. 为了对寿命分布进行统计推断, 就需要通过产品的寿命试验, 以取得寿命数据.

一种典型的寿命试验是, 将随机抽取的 n 个产品在时间 $t=0$ 时, 同时投入试验, 直到每个产品都失效. 记录每一个产品的失效时间, 这样得到的样本 (即由所有产品的失效时间 $0 \leq t_1 \leq t_2 \leq \cdots \leq t_n$ 所组成的样本) 叫完全样本. 然而产品的寿命往往较长, 由于时间和财力的限制, 我们不可能得到完全样本, 于是就考虑截尾寿命试验. 截尾寿命试验常用的有两种: 一种是定时截尾寿命试验. 假设将随机抽取的 n 个产品在时间 $t=0$ 时同时投入试验, 试验进行到事先规定的截尾时间 t_0 停止. 如试验截止时共有 m 个产品失效, 它们的失效时间分别为

$$0 \leq t_1 \leq t_2 \leq \cdots \leq t_m \leq t_0,$$

此时 m 是一个随机变量, 所得的样本 $t_1, t_2, \cdots, t_m$ 称为定时截尾样本. 另一种是定数截尾寿命试验. 假设将随机抽取的 n 个产品在时间 $t=0$ 时同时投入试验, 试验进行到有 m 个 (m 是事先规定的, $m < n$) 产品失效时停止. m 个失效产品的失效时间分别为

$$0 \leq t_1 \leq t_2 \leq \cdots \leq t_m,$$

这里 t_m 是第 m 个产品的失效时间, t_m 是随机变量. 所得的样本 $t_1, t_2, \cdots, t_m$ 称为定数截尾样本. 用截尾样本来进行统计推断是可靠性研究中常见的问题.

设产品的寿命分布是指数分布, 其概率密度为

$$f(t) = \begin{cases} \frac{1}{\theta} e^{-t/\theta}, & t > 0, \\ 0, & t \leq 0, \end{cases}$$

$\theta > 0$ 未知. 设有 n 个产品投入定数截尾试验, 截尾数为 m , 得定数截尾样本 $0 \leq t_1 \leq t_2 \leq \cdots \leq t_m$, 现在要利用这一样本来估计未知参数 θ (即产品的平均寿命). 在时间区间 $[0, t_m]$ 有 m 个产品失效, 而有 $n-m$ 个产品在 t_m 时尚未失效, 即有 $n-m$ 个产品的寿命超过 t_m . 我们用最大似然估计法来估计 θ , 为了确定似然函数, 需要知道上述观察结果出现的概率. 我们知道一个产品在 $(t_i, t_i + dt_i]$ 失效的概率近似地为 $f(t_i) dt_i = \frac{1}{\theta} e^{-t_i/\theta} dt_i, i=1, 2, \dots, m$, 其余 $n-m$ 个产品寿命超过 t_m 的概率为 $\left(\int_{t_m}^{\infty} \frac{1}{\theta} e^{-t/\theta} dt \right)^{n-m} = (e^{-t_m/\theta})^{n-m}$, 故上述观察结果出现的概率近似地为

$$\begin{aligned} & \binom{n}{m} \left(\frac{1}{\theta} e^{-t_1/\theta} dt_1 \right) \left(\frac{1}{\theta} e^{-t_2/\theta} dt_2 \right) \cdots \left(\frac{1}{\theta} e^{-t_m/\theta} dt_m \right) (e^{-t_m/\theta})^{n-m} \\ &= \binom{n}{m} \frac{1}{\theta^m} e^{-\frac{1}{\theta}[t_1+t_2+\cdots+t_m+(n-m)t_m]} dt_1 dt_2 \cdots dt_m, \end{aligned}$$

其中 $dt_1, \dots, dt_m$ 为常数. 因忽略一个常数因子不影响 θ 的最大似然估计, 故可取似然函数为

$$L(\theta) = \frac{1}{\theta^m} e^{-\frac{1}{\theta}[t_1+t_2+\cdots+t_m+(n-m)t_m]}.$$

对数似然函数为

$$\ln L(\theta) = -m \ln \theta - \frac{1}{\theta} [t_1 + t_2 + \cdots + t_m + (n-m)t_m].$$

$$\text{令 } \frac{d}{d\theta} \ln L(\theta) = -\frac{m}{\theta} + \frac{1}{\theta^2} [t_1 + t_2 + \cdots + t_m + (n-m)t_m] = 0.$$

于是得到 θ 的最大似然估计为

$$\hat{\theta} = \frac{s(t_m)}{m}.$$

其中 $s(t_m) = t_1 + t_2 + \cdots + t_m + (n-m)t_m$ 称为总试验时间, 它表示直至时刻 t_m 为止 n 个产品的试验时间的总和.

对于定时截尾样本

$$0 \leq t_1 \leq t_2 \leq \cdots \leq t_m \leq t_0$$

(其中 t_0 是截尾时间), 与上面的讨论类似, 可得似然函数为

$$L(\theta) = \frac{1}{\theta^m} e^{-\frac{1}{\theta}[t_1+t_2+\cdots+t_m+(n-m)t_0]},$$

θ 的最大似然估计为

$$\hat{\theta} = \frac{s(t_0)}{m},$$

其中 $s(t_0) = t_1 + t_2 + \cdots + t_m + (n-m)t_0$ 称为总试验时间, 它表示直至时刻 t_0 为止 n 个产品的试验时间的总和.

例 设电池的寿命服从指数分布, 其概率密度为

$$f(t) = \begin{cases} \frac{1}{\theta} e^{-t/\theta}, & t > 0, \\ 0, & t \leq 0, \end{cases}$$

$\theta > 0$ 未知. 随机地取 50 只电池投入寿命试验, 规定试验进行到其中有 15 只失效时结束试验, 测得失效时间(小时)为

115 119 131 138 142 147 148 155
158 159 163 166 167 170 172

试求电池的平均寿命 θ 的最大似然估计.

解 $n=50, m=15, s(t_{15}) = 115 + 119 + \cdots + 170 + 172 + (50-15) \times 172 = 8270$, 得 θ 的最大似然估计为

$$\hat{\theta} = \frac{8270}{15} = 551.33(\text{小时}).$$

□

§ 3 估计量的评选标准

自前一节可以看到, 对于同一参数, 用不同的估计方法求出的估计量可能不相同, 如第一节的例 2 和例 6. 而且, 很明显, 原则上任何统计量都可以作为未知参数的估计量. 我们自然会问, 采用哪一个估计量为好呢? 这就涉及用什么样的标准来评价估计量的问题. 下面介绍几个常用的标准.

(一) 无偏性

设 $X_1, X_2, \dots, X_n$ 是总体 X 的一个样本, $\theta \in \Theta$ 是包含在总体 X 的分布中的待估参数, 这里 Θ 是 θ 的取值范围.

无偏性 若估计量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 的数学期望 $E(\hat{\theta})$ 存在, 且对于任意 $\theta \in \Theta$ 有

$$E(\hat{\theta}) = \theta, \quad (3.1)$$

则称 $\hat{\theta}$ 是 θ 的无偏估计量.

估计量的无偏性是说对于某些样本值, 由这一估计量得到的估计值相对于真值来说偏大, 有些则偏小. 反复将这一估计量使用多次, 就“平均”来说其偏差为零. 在科学技术中 $E(\hat{\theta}) - \theta$ 称为以 $\hat{\theta}$ 作为 θ 的估计的系统误差. 无偏估计的实际意义就是无系统误差.

例如, 设总体 X 的均值为 μ , 方差 $\sigma^2 > 0$ 均未知, 由第六章(3.19)、(3.20)知

$$E(\bar{X}) = \mu, \quad E(S^2) = \sigma^2.$$

这就是说不论总体服从什么分布, 样本均值 $\bar{X}$ 是总体均值 μ 的无偏估计; 样本

方差 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 是总体方差的无偏估计. 而估计量 $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 却不是 σ^2 的无偏估计, 因此我们一般取 S^2 作为 σ^2 的估计量.

例 1 设总体 X 的 k 阶矩 $\mu_k = E(X^k)$ ($k \geq 1$) 存在, 又设 $X_1, X_2, \dots, X_n$ 是 X 的一个样本. 试证明不论总体服从什么分布, k 阶样本矩 $A_k = \frac{1}{n} \sum_{i=1}^n X_i^k$ 是 k 阶总体矩 μ_k 的无偏估计量.

证 $X_1, X_2, \dots, X_n$ 与 X 同分布, 故有

$$E(X_i^k) = E(X^k) = \mu_k, \quad i = 1, 2, \dots, n.$$

即有

$$E(A_k) = \frac{1}{n} \sum_{i=1}^n E(X_i^k) = \mu_k. \quad (3.2) \square$$

例 2 设总体 X 服从指数分布, 其概率密度为

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} e^{-x/\theta}, & x > 0, \\ 0, & \text{其他}, \end{cases}$$

其中参数 $\theta > 0$ 为未知, 又设 $X_1, X_2, \dots, X_n$ 是来自 X 的样本, 试证 $\bar{X}$ 和 $nZ = n(\min\{X_1, X_2, \dots, X_n\})$ 都是 θ 的无偏估计量.

证 因为 $E(\bar{X}) = E(X) = \theta$, 所以 $\bar{X}$ 是 θ 的无偏估计量. 而 $Z = \min\{X_1, X_2, \dots, X_n\}$ 具有概率密度

$$f_{\min}(x; \theta) = \begin{cases} \frac{n}{\theta} e^{-nx/\theta}, & x > 0, \\ 0, & \text{其他}. \end{cases}$$

故知

$$E(Z) = \frac{\theta}{n},$$

$$E(nZ) = \theta.$$

即 nZ 也是参数 θ 的无偏估计量. □

由此可见一个未知参数可以有不同的无偏估计量. 事实上, 在本例中 $X_1, X_2, \dots, X_n$ 中的每一个都可以作为 θ 的无偏估计量.

(二) 有效性

现在来比较参数 θ 的两个无偏估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$, 如果在样本容量 n 相同的情况下, $\hat{\theta}_1$ 的观察值较 $\hat{\theta}_2$ 更密集在真值 θ 的附近, 我们就认为 $\hat{\theta}_1$ 较 $\hat{\theta}_2$ 为理想. 由于

方差是随机变量取值与其数学期望(此时数学期望 $E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta$) 的偏离程度的度量, 所以无偏估计以方差小者为好. 这就引出了估计量的有效性这一概念.

有效性 设 $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \dots, X_n)$ 与 $\hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \dots, X_n)$ 都是 θ 的无偏估计量, 若对于任意 $\theta \in \Theta$, 有

$$D(\hat{\theta}_1) \leq D(\hat{\theta}_2)$$

且至少对于某一个 $\theta \in \Theta$ 上式中的不等号成立, 则称 $\hat{\theta}_1$ 较 $\hat{\theta}_2$ 有效.

例 3(续例 2) 试证当 $n > 1$ 时, θ 的无偏估计量 $\bar{X}$ 较 θ 的无偏估计量 nZ 有效.

证 由于 $D(X) = \theta^2$, 故有 $D(\bar{X}) = \theta^2/n$. 再者, 由于 $D(Z) = \theta^2/n^2$, 故有 $D(nZ) = \theta^2$. 当 $n > 1$ 时 $D(nZ) > D(\bar{X})$, 故 $\bar{X}$ 较 nZ 有效. $\square$

(三) 相合性

前面讲的无偏性与有效性都是在样本容量 n 固定的前提下提出的. 我们自然希望随着样本容量的增大, 一个估计量的值稳定于待估参数的真值. 这样, 对估计量又有下述相合性的要求.

相合性 设 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 为参数 θ 的估计量, 若对于任意 $\theta \in \Theta$, 当 $n \rightarrow \infty$ 时 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 依概率收敛于 θ , 则称 $\hat{\theta}$ 为 θ 的相合估计量.

即, 若对于任意 $\theta \in \Theta$ 都满足: 对于任意 $\epsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P\{|\hat{\theta} - \theta| < \epsilon\} = 1,$$

则称 $\hat{\theta}$ 是 θ 的相合估计量.

例如由第六章 § 3 知, 样本 $k(k \geq 1)$ 阶矩是总体 X 的 k 阶矩 $\mu_k = E(X^k)$ 的相合估计量, 进而若待估参数 $\theta = g(\mu_1, \mu_2, \dots, \mu_k)$, 其中 g 为连续函数, 则 θ 的矩估计量 $\hat{\theta} = g(\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_k) = g(A_1, A_2, \dots, A_k)$ 是 θ 的相合估计量.

由最大似然估计法得到的估计量, 在一定条件下也具有相合性. 其详细讨论已超出本书范围, 从略.

相合性是对一个估计量的基本要求, 若估计量不具有相合性, 那么不论将样本容量 n 取得多么大, 都不能将 θ 估计得足够准确, 这样的估计量是不可取的.

上述无偏性、有效性、相合性是评价估计量的一些基本标准, 其他的标准这里就不讲了.

§4 区间估计

对于一个未知量,人们在测量或计算时,常不以得到近似值为满足,还需估计误差,即要求知道近似值的精确程度(亦即所求真值所在的范围).类似地,对于未知参数 θ ,除了求出它的点估计 $\hat{\theta}$ 外,我们还希望估计出一个范围,并希望知道这个范围包含参数 θ 真值的可信程度.这样的范围通常以区间的形式给出,同时还给出此区间包含参数 θ 真值的可信程度.这种形式的估计称为区间估计,这样的区间即所谓置信区间.现在我们引入置信区间的定义.

置信区间 设总体 X 的分布函数 $F(x;\theta)$ 含有一个未知参数 $\theta, \theta \in \Theta$ (Θ 是 θ 可能取值的范围),对于给定值 α ($0 < \alpha < 1$),若由来自 X 的样本 $X_1, X_2, \dots, X_n$ 确定的两个统计量 $\underline{\theta} = \underline{\theta}(X_1, X_2, \dots, X_n)$ 和 $\bar{\theta} = \bar{\theta}(X_1, X_2, \dots, X_n)$ ($\underline{\theta} < \bar{\theta}$),对于任意 $\theta \in \Theta$ 满足

$$P\{\underline{\theta}(X_1, X_2, \dots, X_n) < \theta < \bar{\theta}(X_1, X_2, \dots, X_n)\} \geq 1 - \alpha, \quad (4.1)$$

则称随机区间 $(\underline{\theta}, \bar{\theta})$ 是 θ 的置信水平为 $1 - \alpha$ 的置信区间, $\underline{\theta}$ 和 $\bar{\theta}$ 分别称为置信水平为 $1 - \alpha$ 的双侧置信区间的置信下限和置信上限, $1 - \alpha$ 称为置信水平.

当 X 是连续型随机变量时,对于给定的 α ,我们总是按要求 $P\{\underline{\theta} < \theta < \bar{\theta}\} = 1 - \alpha$ 求出置信区间.而当 X 是离散型随机变量时,对于给定的 α ,常常找不到区间 $(\underline{\theta}, \bar{\theta})$ 使得 $P\{\underline{\theta} < \theta < \bar{\theta}\}$ 恰为 $1 - \alpha$.此时我们去找区间 $(\underline{\theta}, \bar{\theta})$ 使得 $P\{\underline{\theta} < \theta < \bar{\theta}\}$ 至少为 $1 - \alpha$,且尽可能地接近 $1 - \alpha$.

(4.1)式的含义如下:若反复抽样多次(各次得到的样本的容量相等,都是 n).每个样本值确定一个区间 $(\underline{\theta}, \bar{\theta})$,每个这样的区间要么包含 θ 的真值,要么不包含 θ 的真值(参见图7-1).按伯努利大数定理,在这么多的区间中,包含 θ 真值的约占 $100(1 - \alpha)\%$,不包含 θ 真值的约仅占 $100\alpha\%$.例如,若 $\alpha = 0.01$,反复抽样1000次,则得到的1000个区间中不包含 θ 真值的约仅为10个.

例1 设总体 $X \sim N(\mu, \sigma^2)$, σ^2 为已知, μ 为未知,设 $X_1, X_2, \dots, X_n$ 是来自 X 的样本,求 μ 的置信水平为 $1 - \alpha$ 的置信区间.

解 我们知道 $\bar{X}$ 是 μ 的无偏估计.且有

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1). \quad (4.2)$$

$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ 所服从的分布 $N(0, 1)$ 不依赖于任何未知参数.按标准正态分布的上 α 分位点的定义,有(参见图7-2)

$$P\left\{\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}\right\} = 1 - \alpha, \quad (4.3)$$

$$\text{即} \quad P\left\{\bar{X} - \frac{\sigma}{\sqrt{n}}z_{\alpha/2} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right\} = 1 - \alpha. \quad (4.4)$$

这样,我们就得到了 μ 的一个置信水平为 $1-\alpha$ 的置信区间

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}}z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right). \quad (4.5)$$

这样的置信区间常写成

$$\left(\bar{X} \pm \frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right). \quad (4.6)$$

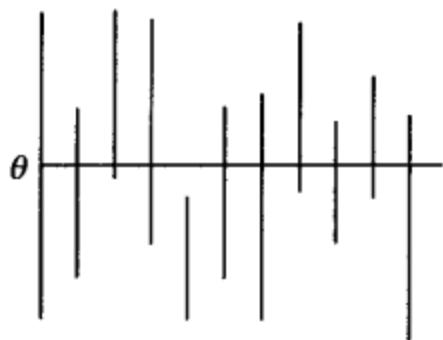


图 7-1

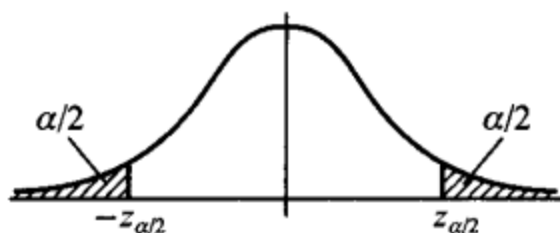


图 7-2

如果取 $1-\alpha=0.95$, 即 $\alpha=0.05$, 又若 $\sigma=1, n=16$, 查表得 $z_{\alpha/2}=z_{0.025}=1.96$. 于是我们得到一个置信水平为 0.95 的置信区间

$$\left(\bar{X} \pm \frac{1}{\sqrt{16}} \times 1.96\right), \text{ 即 } (\bar{X} \pm 0.49). \quad (4.7)$$

再者,若由一个样本值算得样本均值的观察值 $\bar{x}=5.20$, 则得到一个区间

$$(5.20 \pm 0.49), \text{ 即 } (4.71, 5.69).$$

注意,这已经不是随机区间了. 但我们仍称它为置信水平为 0.95 的置信区间. 其含义是: 若反复抽样多次, 每个样本值 ($n=16$) 按 (4.7) 式确定一个区间, 按上面的解释, 在这么多的区间中, 包含 μ 的约占 95%, 不包含 μ 的约占 5%. 现在抽样得到区间 (4.71, 5.69), 则该区间属于那些包含 μ 的区间的可信程度为 95%, 或“该区间包含 μ ”这一陈述的可信程度为 95%. $\square$

置信水平为 $1-\alpha$ 的置信区间并不是唯一的. 以例 1 来说, 若给定 $\alpha=0.05$, 则又有

$$P\left\{-z_{0.04} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{0.01}\right\} = 0.95,$$

$$\text{即} \quad P\left\{\bar{X} - \frac{\sigma}{\sqrt{n}}z_{0.01} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}z_{0.04}\right\} = 0.95.$$

$$\text{故} \quad \left(\bar{X} - \frac{\sigma}{\sqrt{n}}z_{0.01}, \bar{X} + \frac{\sigma}{\sqrt{n}}z_{0.04}\right) \quad (4.8)$$

也是 μ 的置信水平为 0.95 的置信区间. 我们将它与 (4.5) 中令 $\alpha=0.05$ 所得的置信水平为 0.95 的置信区间相比较, 可知由 (4.5) 所确定的区间的长度为 $2 \times \frac{\sigma}{\sqrt{n}} z_{0.025} = 3.92 \times \frac{\sigma}{\sqrt{n}}$, 这一长度要比区间 (4.8) 的长度 $\frac{\sigma}{\sqrt{n}} (z_{0.04} + z_{0.01}) = 4.08 \times$

$\frac{\sigma}{\sqrt{n}}$ 为短. 置信区间短表示估计的精度高. 故由 (4.5) 给出的区间较 (4.8) 为优. 易知, 像 $N(0,1)$ 分布那样其概率密度的图形是单峰且对称的情况, 当 n 固定时, 以形如 (4.5) 那样的区间其长度为最短, 我们自然选用它.

参考例 1 可得寻求未知参数 θ 的置信区间的具体做法如下.

1° 寻求一个样本 $X_1, X_2, \dots, X_n$ 和 θ 的函数 $W=W(X_1, X_2, \dots, X_n; \theta)$, 使得 W 的分布不依赖于 θ 以及其他未知参数, 称具有这种性质的函数 W 为枢轴量.

2° 对于给定的置信水平 $1-\alpha$, 定出两个常数 a, b 使得

$$P\{a < W(X_1, X_2, \dots, X_n; \theta) < b\} = 1 - \alpha.$$

若能从 $a < W(X_1, X_2, \dots, X_n; \theta) < b$ 得到与之等价的 θ 的不等式 $\underline{\theta} < \theta < \bar{\theta}$, 其中 $\underline{\theta} = \underline{\theta}(X_1, X_2, \dots, X_n)$, $\bar{\theta} = \bar{\theta}(X_1, X_2, \dots, X_n)$ 都是统计量. 那么 $(\underline{\theta}, \bar{\theta})$ 就是 θ 的一个置信水平为 $1-\alpha$ 的置信区间.

函数 $W(X_1, X_2, \dots, X_n; \theta)$ 的构造, 通常可以从 θ 的点估计着手考虑. 常用的正态总体的参数的置信区间可以用上述步骤推得.

§ 5 正态总体均值与方差的区间估计

(一) 单个总体 $N(\mu, \sigma^2)$ 的情况

设已给定置信水平为 $1-\alpha$, 并设 $X_1, X_2, \dots, X_n$ 为总体 $N(\mu, \sigma^2)$ 的样本. $\bar{X}$, S^2 分别是样本均值和样本方差.

1. 均值 μ 的置信区间

(1) σ^2 为已知, 此时由 § 4 例 1 采用 (4.2) 中的枢轴量 $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$, 已得到 μ 的一个置信水平为 $1-\alpha$ 的置信区间为

$$\left(\bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right). \quad (5.1)$$

(2) σ^2 为未知, 此时不能使用 (5.1) 给出的区间, 因其中含未知参数 σ . 考虑到 S^2 是 σ^2 的无偏估计, 将 (4.2) 中的 σ 换成 $S = \sqrt{S^2}$, 由第六章 § 3 定理三, 知

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1), \quad (5.2)$$

并且右边的分布 $t(n-1)$ 不依赖于任何未知参数. 使用 $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ 作为枢轴量可得 (参见

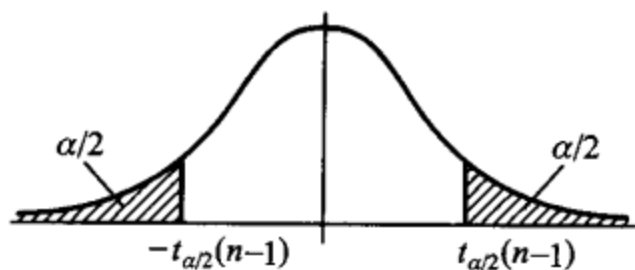


图7-3)

图 7-3

$$P\left\{-t_{\alpha/2}(n-1) < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}(n-1)\right\} = 1 - \alpha, \quad (5.3)$$

即

$$P\left\{\bar{X} - \frac{S}{\sqrt{n}}t_{\alpha/2}(n-1) < \mu < \bar{X} + \frac{S}{\sqrt{n}}t_{\alpha/2}(n-1)\right\} = 1 - \alpha.$$

于是得 μ 的一个置信水平为 $1-\alpha$ 的置信区间

$$\left(\bar{X} \pm \frac{S}{\sqrt{n}}t_{\alpha/2}(n-1)\right). \quad (5.4)$$

例 1 有一大批糖果. 现从中随机地取 16 袋, 称得重量 (以 g 计) 如下:

506 508 499 503 504 510 497 512
514 505 493 496 506 502 509 496

设袋装糖果的重量近似地服从正态分布, 试求总体均值 μ 的置信水平为 0.95 的置信区间.

解 这里 $1-\alpha=0.95$, $\alpha/2=0.025$, $n-1=15$, $t_{0.025}(15)=2.1315$, 由给出的数据算得 $\bar{x}=503.75$, $s=6.2022$. 由 (5.4) 式得均值 μ 的一个置信水平为 0.95 的置信区间为

$$\left(503.75 \pm \frac{6.2022}{\sqrt{16}} \times 2.1315\right),$$

即

$$(500.4, 507.1).$$

这就是说估计袋装糖果重量的均值在 500.4g 与 507.1g 之间, 这个估计的可信程度为 95%. 若以此区间内任一值作为 μ 的近似值, 其误差不大于 $\frac{6.2022}{\sqrt{16}} \times 2.1315 \times 2 = 6.61(\text{g})$, 这个误差估计的可信程度为 95%. $\square$

在实际问题中, 总体方差 σ^2 未知的情况居多, 故区间 (5.4) 较区间 (5.1) 有更大的实用价值.

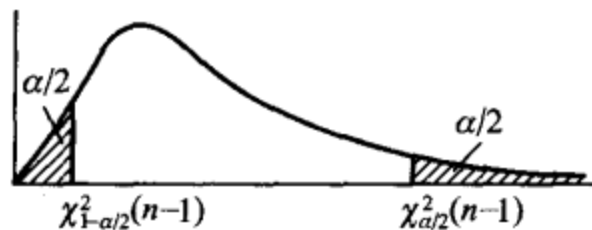
2. 方差 σ^2 的置信区间

此处, 根据实际问题的需要, 只介绍 μ 未知的情况.

σ^2 的无偏估计为 S^2 , 由第六章 §3 定理二知

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1), \quad (5.5)$$

并且上式右端的分布不依赖于任何未知参数, 取 $\frac{(n-1)S^2}{\sigma^2}$ 作为枢轴量, 即得 (参见图



7-4)

图 7-4

$$P\left\{\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1)\right\} = 1-\alpha, \quad (5.6)$$

即
$$P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right\} = 1-\alpha. \quad (5.6)'$$

这就得到方差 σ^2 的一个置信水平为 $1-\alpha$ 的置信区间

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right). \quad (5.7)$$

由 (5.6)' 式, 还可得到标准差 σ 的一个置信水平为 $1-\alpha$ 的置信区间

$$\left[\frac{\sqrt{n-1}S}{\sqrt{\chi^2_{\alpha/2}(n-1)}}, \frac{\sqrt{n-1}S}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}}\right]. \quad (5.8)$$

注意, 在密度函数不对称时, 如 χ^2 分布和 F 分布, 习惯上仍是取对称的分位点 (如图 7-4 中的上分位点 $\chi^2_{1-\alpha/2}(n-1)$ 与 $\chi^2_{\alpha/2}(n-1)$) 来确定置信区间的。

例 2 求例 1 中总体标准差 σ 的置信水平为 0.95 的置信区间。

解 现在 $\alpha/2 = 0.025$, $1-\alpha/2 = 0.975$, $n-1 = 15$, 查表得 $\chi^2_{0.025}(15) = 27.488$, $\chi^2_{0.975}(15) = 6.262$, 又 $s = 6.2022$, 由 (5.8) 式得所求的标准差 σ 的一个置信水平为 0.95 的置信区间为

$$(4.58, 9.60).$$

□

(二) 两个总体 $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ 的情况

在实际中常遇到下面的问题: 已知产品的某一质量指标服从正态分布, 但由于原料、设备条件、操作人员不同, 或工艺过程的改变等因素, 引起总体均值、总体方差有所改变. 我们需要知道这些变化有多大, 这就需要考虑两个正态总体均值差或方差比的估计问题。

设已给定置信水平为 $1-\alpha$, 并设 $X_1, X_2, \dots, X_{n_1}$ 是来自第一个总体的样本; $Y_1, Y_2, \dots, Y_{n_2}$ 是来自第二个总体的样本, 这两个样本相互独立. 且设 $\bar{X}, \bar{Y}$ 分别为第一、第二个总体的样本均值, S_1^2, S_2^2 分别是第一、第二个总体的样本方差。

1. 两个总体均值差 $\mu_1 - \mu_2$ 的置信区间

(1) σ_1^2, σ_2^2 均为已知. 因 $\bar{X}, \bar{Y}$ 分别为 μ_1, μ_2 的无偏估计, 故 $\bar{X} - \bar{Y}$ 是 $\mu_1 - \mu_2$ 的无偏估计. 由 $\bar{X}, \bar{Y}$ 的独立性以及 $\bar{X} \sim N(\mu_1, \sigma_1^2/n_1), \bar{Y} \sim N(\mu_2, \sigma_2^2/n_2)$ 得

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

或

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1), \quad (5.9)$$

取(5.9)左边的函数为枢轴量, 即得 $\mu_1 - \mu_2$ 的一个置信水平为 $1 - \alpha$ 的置信区间

$$\left(\bar{X} - \bar{Y} \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right). \quad (5.10)$$

(2) $\sigma_1^2 = \sigma_2^2 = \sigma^2$, 但 σ^2 为未知. 此时, 由第六章 §3 定理四

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2). \quad (5.11)$$

取(5.11)左边的函数为枢轴量, 可得 $\mu_1 - \mu_2$ 的一个置信水平为 $1 - \alpha$ 的置信区间为

$$\left(\bar{X} - \bar{Y} \pm t_{\alpha/2}(n_1 + n_2 - 2) S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right). \quad (5.12)$$

此处
$$S_w^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, S_w = \sqrt{S_w^2}. \quad (5.13)$$

例 3 为比较 I, II 两种型号步枪子弹的枪口速度, 随机地取 I 型子弹 10 发, 得到枪口速度的平均值为 $\bar{x}_1 = 500$ m/s, 标准差 $s_1 = 1.10$ m/s, 随机地取 II 型子弹 20 发, 得到枪口速度的平均值为 $\bar{x}_2 = 496$ m/s, 标准差 $s_2 = 1.20$ m/s. 假设两总体都可认为近似地服从正态分布. 且由生产过程可认为方差相等. 求两总体均值差 $\mu_1 - \mu_2$ 的一个置信水平为 0.95 的置信区间.

解 按实际情况, 可认为分别来自两个总体的样本是相互独立的. 又因由假设两总体的方差相等, 但数值未知, 故可用(5.12)式求均值差的置信区间. 由于 $1 - \alpha = 0.95$, $\alpha/2 = 0.025$, $n_1 = 10$, $n_2 = 20$, $n_1 + n_2 - 2 = 28$, $t_{0.025}(28) = 2.0484$. $s_w^2 = (9 \times 1.10^2 + 19 \times 1.20^2)/28$, $s_w = \sqrt{s_w^2} = 1.1688$, 故所求的两总体均值差 $\mu_1 - \mu_2$ 的一个置信水平为 0.95 的置信区间是

$$\left(\bar{x}_1 - \bar{x}_2 \pm s_w \times t_{0.025}(28) \sqrt{\frac{1}{10} + \frac{1}{20}}\right) = (4 \pm 0.93),$$

即

$$(3.07, 4.93).$$

本题中得到的置信区间的下限大于零, 在实际中我们就认为 μ_1 比 μ_2 大. $\square$

例 4 为提高某一化学生产过程的得率, 试图采用一种新的催化剂. 为慎重起见, 在实验工厂先进行试验. 设采用原来的催化剂进行了 $n_1 = 8$ 次试验, 得到得率的平均值 $\bar{x}_1 = 91.73$, 样本方差 $s_1^2 = 3.89$; 又采用新的催化剂进行了 $n_2 = 8$

次试验, 得到得率的平均值 $\bar{x}_2 = 93.75$, 样本方差 $s_2^2 = 4.02$. 假设两总体都可认为服从正态分布, 且方差相等, 两样本独立. 试求两总体均值差 $\mu_1 - \mu_2$ 的置信水平为 0.95 的置信区间.

解 现在

$$s_w^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 3.96, s_w = \sqrt{3.96}.$$

由(5.12)式得所求的置信区间为

$$\left(\bar{x}_1 - \bar{x}_2 \pm t_{0.025}(14) s_w \sqrt{\frac{1}{8} + \frac{1}{8}} \right) = (-2.02 \pm 2.13),$$

即 $(-4.15, 0.11)$.

由于所得置信区间包含零, 在实际中我们就认为采用这两种催化剂所得的得率的均值没有显著差别. $\square$

2. 两个总体方差比 σ_1^2/σ_2^2 的置信区间

我们仅讨论总体均值 μ_1, μ_2 均为未知的情况, 由第六章 §3 定理四

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1), \quad (5.14)$$

并且分布 $F(n_1 - 1, n_2 - 1)$ 不依赖任何未知参数. 取 $\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}$ 为枢轴量得

$$P \left\{ F_{1-\alpha/2}(n_1 - 1, n_2 - 1) < \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} < F_{\alpha/2}(n_1 - 1, n_2 - 1) \right\} = 1 - \alpha, \quad (5.15)$$

即

$$P \left\{ \frac{S_1^2}{S_2^2} \frac{1}{F_{\alpha/2}(n_1 - 1, n_2 - 1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} \frac{1}{F_{1-\alpha/2}(n_1 - 1, n_2 - 1)} \right\} = 1 - \alpha. \quad (5.15)'$$

于是得 σ_1^2/σ_2^2 的一个置信水平为 $1 - \alpha$ 的置信区间为

$$\left(\frac{S_1^2}{S_2^2} \frac{1}{F_{\alpha/2}(n_1 - 1, n_2 - 1)}, \frac{S_1^2}{S_2^2} \frac{1}{F_{1-\alpha/2}(n_1 - 1, n_2 - 1)} \right). \quad (5.16)$$

例 5 研究由机器 A 和机器 B 生产的钢管的内径(单位: mm), 随机抽取机器 A 生产的管子 18 只, 测得样本方差 $s_1^2 = 0.34$; 抽取机器 B 生产的管子 13 只, 测得样本方差 $s_2^2 = 0.29$. 设两样本相互独立, 且设由机器 A, 机器 B 生产的管子的内径分别服从正态分布 $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$, 这里 $\mu_i, \sigma_i^2 (i=1, 2)$ 均未知. 试求方差比 σ_1^2/σ_2^2 的置信水平为 0.90 的置信区间.

解 现在 $n_1 = 18, s_1^2 = 0.34, n_2 = 13, s_2^2 = 0.29, \alpha = 0.10, F_{\alpha/2}(n_1 - 1, n_2 - 1) = F_{0.05}(17, 12) = 2.59, F_{1-\alpha/2}(17, 12) = F_{0.95}(17, 12) = \frac{1}{F_{0.05}(12, 17)} = \frac{1}{2.38}$, 于

是由(5.16)式得 σ_1^2/σ_2^2 的一个置信水平为 0.90 的置信区间为

$$\left(\frac{0.34}{0.29} \times \frac{1}{2.59}, \frac{0.34}{0.29} \times 2.38 \right),$$

即

$$(0.45, 2.79).$$

由于 σ_1^2/σ_2^2 的置信区间包含 1, 在实际中我们就认为 σ_1^2, σ_2^2 两者没有显著差别. $\square$

§ 6 (0-1)分布参数的区间估计

设有一容量 $n > 50$ 的大样本, 它来自(0-1)分布的总体 X , X 的分布律为

$$f(x; p) = p^x (1-p)^{1-x}, x=0, 1, \quad (6.1)$$

其中 p 为未知参数. 现在来求 p 的置信水平为 $1-\alpha$ 的置信区间.

已知(0-1)分布的均值和方差分别为

$$\mu = p, \quad \sigma^2 = p(1-p). \quad (6.2)$$

设 $X_1, X_2, \dots, X_n$ 是一个样本. 因样本容量 n 较大, 由中心极限定理, 知

$$\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} = \frac{n\bar{X} - np}{\sqrt{np(1-p)}} \quad (6.3)$$

近似地服从 $N(0, 1)$ 分布, 于是有

$$P\left\{-z_{\alpha/2} < \frac{n\bar{X} - np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right\} \approx 1 - \alpha. \quad (6.4)$$

而不等式

$$-z_{\alpha/2} < \frac{n\bar{X} - np}{\sqrt{np(1-p)}} < z_{\alpha/2} \quad (6.5)$$

等价于

$$(n + z_{\alpha/2}^2)p^2 - (2n\bar{X} + z_{\alpha/2}^2)p + n\bar{X}^2 < 0. \quad (6.6)$$

记

$$p_1 = \frac{1}{2a}(-b - \sqrt{b^2 - 4ac}), \quad (6.7)$$

$$p_2 = \frac{1}{2a}(-b + \sqrt{b^2 - 4ac}), \quad (6.8)$$

此处 $a = n + z_{\alpha/2}^2$, $b = -(2n\bar{X} + z_{\alpha/2}^2)$, $c = n\bar{X}^2$. 于是由(6.5)式得 p 的一个近似的置信水平为 $1-\alpha$ 的置信区间为

$$(p_1, p_2).$$

例 设自一大批产品的 100 个样品中, 得一级品 60 个, 求这批产品的一级品率 p 的置信水平为 0.95 的置信区间.

解 一级品率 p 是(0-1)分布的参数, 此处 $n=100$, $\bar{x}=60/100=0.6$, $1-\alpha=0.95$, $\alpha/2=0.025$, $z_{\alpha/2}=1.96$, 按(6.7), (6.8)式来求 p 的置信区间, 其中

$$a = n + z_{\alpha/2}^2 = 103.84, b = -(2n\bar{x} + z_{\alpha/2}^2) = -123.84, c = n\bar{x}^2 = 36.$$

于是

$$p_1 = 0.50, \quad p_2 = 0.69.$$

故得 p 的一个置信水平为 0.95 的近似置信区间为

$$(0.50, 0.69).$$

□

§ 7 单侧置信区间

在上述讨论中,对于未知参数 θ ,我们给出两个统计量 $\underline{\theta}, \bar{\theta}$,得到 θ 的双侧置信区间 $(\underline{\theta}, \bar{\theta})$.但在某些实际问题中,例如,对于设备、元件的寿命来说,平均寿命长是我们所希望的,我们关心的是平均寿命 θ 的“下限”;与之相反,在考虑化学药品中杂质含量的均值 μ 时,我们常关心参数 μ 的“上限”.这就引出了单侧置信区间的概念.

对于给定值 $\alpha (0 < \alpha < 1)$,若由样本 $X_1, X_2, \dots, X_n$ 确定的统计量 $\underline{\theta} = \underline{\theta}(X_1, X_2, \dots, X_n)$,对于任意 $\theta \in \Theta$ 满足

$$P\{\theta > \underline{\theta}\} \geq 1 - \alpha, \quad (7.1)$$

称随机区间 $(\underline{\theta}, \infty)$ 是 θ 的置信水平为 $1 - \alpha$ 的单侧置信区间, $\underline{\theta}$ 称为 θ 的置信水平为 $1 - \alpha$ 的单侧置信下限.

又若统计量 $\bar{\theta} = \bar{\theta}(X_1, X_2, \dots, X_n)$,对于任意 $\theta \in \Theta$ 满足

$$P\{\theta < \bar{\theta}\} \geq 1 - \alpha, \quad (7.2)$$

称随机区间 $(-\infty, \bar{\theta})$ 是 θ 的置信水平为 $1 - \alpha$ 的单侧置信区间, $\bar{\theta}$ 称为 θ 的置信水平为 $1 - \alpha$ 的单侧置信上限.

例如对于正态总体 X ,若均值 μ ,方差 σ^2 均为未知,设 $X_1, X_2, \dots, X_n$ 是一个样本,由

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

有(见图 7-5)

$$P\left\{\frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha}(n-1)\right\} = 1 - \alpha,$$

即

$$P\left\{\mu > \bar{X} - \frac{S}{\sqrt{n}} t_{\alpha}(n-1)\right\} = 1 - \alpha.$$

于是得到 μ 的一个置信水平为 $1 - \alpha$ 的单侧置信区间

$$\left(\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha}(n-1), \infty\right). \quad (7.3)$$

μ 的置信水平为 $1 - \alpha$ 的单侧置信下限为

$$\underline{\mu} = \bar{X} - \frac{S}{\sqrt{n}} t_{\alpha}(n-1). \quad (7.4)$$

又由

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

有(见图 7-6)

$$P\left\{\frac{(n-1)S^2}{\sigma^2} > \chi_{1-\alpha}^2(n-1)\right\} = 1-\alpha,$$

即

$$P\left\{\sigma^2 < \frac{(n-1)S^2}{\chi_{1-\alpha}^2(n-1)}\right\} = 1-\alpha.$$

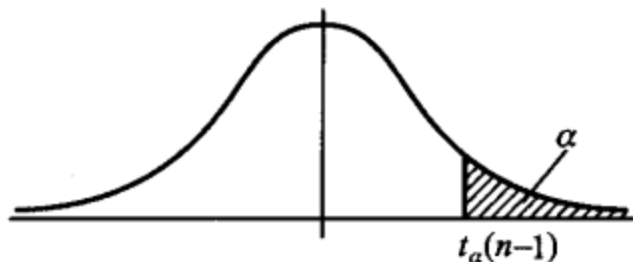


图 7-5

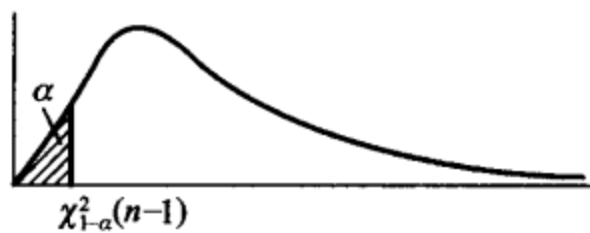


图 7-6

于是得 σ^2 的置信水平为 $1-\alpha$ 的单侧置信区间

$$\left(0, \frac{(n-1)S^2}{\chi_{1-\alpha}^2(n-1)}\right). \quad (7.5)$$

σ^2 的置信水平为 $1-\alpha$ 的单侧置信上限为

$$\bar{\sigma}^2 = \frac{(n-1)S^2}{\chi_{1-\alpha}^2(n-1)}. \quad (7.6)$$

例 从一批灯泡中随机地取 5 只作寿命试验,测得寿命(以 h 计)为

1 050 1 100 1 120 1 250 1 280

设灯泡寿命服从正态分布. 求灯泡寿命平均值的置信水平为 0.95 的单侧置信下限.

解 $1-\alpha=0.95, n=5, t_\alpha(n-1)=t_{0.05}(4)=2.1318, \bar{x}=1\ 160, s^2=9\ 950$.
由(7.4)式得所求单侧置信下限为

$$\underline{\mu} = \bar{x} - \frac{s}{\sqrt{n}} t_\alpha(n-1) = 1\ 065. \quad \square$$

小结

参数估计问题分为点估计和区间估计. 点估计是适当地选择一个统计量作为未知参数的估计(称为估计量),若已取得一样本,将样本值代入估计量,得到估计量的值,以估计量的值作为未知参数的近似值(称为估计值).

本章介绍了两种求点估计的方法:矩估计法和最大似然估计法.

矩估计法的做法是,以样本矩作为总体矩的估计量,而以样本矩的连续函数作为相应的总体矩的连续函数的估计量,从而得到总体未知参数的估计.

最大似然估计法的基本想法是,若已观察到样本 $(X_1, X_2, \dots, X_n)$ 的样本值 $(x_1, x_2, \dots, x_n)$,而取到这一样本值的概率为 p (在离散型的情况),或 $(X_1, X_2, \dots, X_n)$ 落在这一样本值 $(x_1, x_2, \dots, x_n)$ 的邻域内的概率为 p (在连续型的情况),而 p 与未知参数有关,我们就取 θ 的估计值使概率 p 取到最大.

对于一个未知参数可以提出不同的估计量,因此自然提出比较估计量的好坏的问题,这就需要给出评定估计量好坏的标准.估计量是一个随机变量,对于不同的样本值,一般给出参数的不同估计值.因而在考虑估计量的好坏时,应从某种整体性能去衡量,而不能看它在个别样本之下表现如何.本章介绍了三个标准:无偏性、有效性和相合性.相合性是对估计量的一个基本要求,不具备相合性的估计量,我们是不考虑的.

点估计不能反映估计的精度,我们引入了区间估计.置信区间是一个随机区间 $(\underline{\theta}, \bar{\theta})$,它覆盖未知参数具有预先给定的高概率(置信水平),即对于任意 $\theta \in \Theta$,有

$$P\{\underline{\theta} < \theta < \bar{\theta}\} \geq 1 - \alpha.$$

例如,对于正态分布 $N(\mu, \sigma^2)$, σ^2 未知,可得 μ 的一个置信水平为 $1 - \alpha$ 的置信区间为

$$\left(\bar{X} - t_{\alpha/2}(n-1) \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1) \frac{S}{\sqrt{n}}\right), \quad (5.4)$$

就是说这一随机区间覆盖 μ 的概率 $\geq 1 - \alpha$.一旦有了一个样本值 $x_1, x_2, \dots, x_n$,将它代入(5.4),得到一个数字区间

$$\left(\bar{x} - t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}}\right) \text{记成 } (-c, c),$$

$(-c, c)$ 也称为 μ 的置信水平为 $1 - \alpha$ 的置信区间,意指“ $(-c, c)$ 包含 μ ”这一陈述的可信程度为 $1 - \alpha$.如果将这事实写成 $P\{-c < \mu < c\} = 1 - \alpha$ 是错误的,因为 $(-c, c)$ 是一个数字区间,要么有 $\mu \in (-c, c)$,此时 $P\{-c < \mu < c\} = 1$;要么有 $\mu \notin (-c, c)$,此时 $P\{-c < \mu < c\} = 0$.

本章还介绍了单侧置信区间,例如,对于正态分布 $N(\mu, \sigma^2)$, σ^2 未知,可得 μ 的置信水平为 $1 - \alpha$ 的单侧置信区间为

$$(i) \left(-\infty, \bar{X} + t_{\alpha}(n-1) \frac{S}{\sqrt{n}}\right), \text{ 单侧置信上限为 } \bar{\mu} = \bar{X} + t_{\alpha}(n-1) \frac{S}{\sqrt{n}}.$$

$$(ii) \left(\bar{X} - t_{\alpha}(n-1) \frac{S}{\sqrt{n}}, \infty\right), \text{ 单侧置信下限为 } \underline{\mu} = \bar{X} - t_{\alpha}(n-1) \frac{S}{\sqrt{n}}.$$

在形式上,只需将置信区间(5.4)的上下限中的“ $\alpha/2$ ”改成“ α ”,就得到相应的单侧置信上下限了.

■ 重要术语及主题

矩估计量 最大似然估计量

估计量的评选标准:无偏性、有效性、相合性

参数 θ 的置信水平为 $1 - \alpha$ 的置信区间

参数 θ 的单侧置信上限和单侧置信下限

单个正态总体均值、方差的置信区间、单侧置信上限与单侧置信下限

两个正态总体均值差、方差比的置信区间、单侧置信上限与单侧置信下限

表 7-1 正态总体均值、方差的置信区间与单侧置信限(置信水平为 $1-\alpha$)

待估参数	其他参数	枢轴量 W 的分布	置信区间	单侧置信限
一个正态总体	σ^2 已知	$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\left(\bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$	$\bar{\mu} = \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha} \quad \underline{\mu} = \bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha}$
	σ^2 未知	$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$	$\left(\bar{X} \pm \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1) \right)$	$\bar{\mu} = \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha}(n-1) \quad \underline{\mu} = \bar{X} - \frac{S}{\sqrt{n}} t_{\alpha}(n-1)$
	μ 未知	$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left(\frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)} \right)$	$\bar{\sigma}^2 = \frac{(n-1)S^2}{\chi_{1-\alpha}^2(n-1)} \quad \underline{\sigma}^2 = \frac{(n-1)S^2}{\chi_{\alpha}^2(n-1)}$
两个正态总体	σ_1^2, σ_2^2 已知	$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$	$\left(\bar{X} - \bar{Y} \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$	$\overline{\mu_1 - \mu_2} = \bar{X} - \bar{Y} + z_{\alpha} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ $\underline{\mu_1 - \mu_2} = \bar{X} - \bar{Y} - z_{\alpha} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
	$\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知	$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$ $S_w^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$	$\left(\bar{X} - \bar{Y} \pm t_{\alpha/2}(n_1 + n_2 - 2) S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$	$\overline{\mu_1 - \mu_2} = \bar{X} - \bar{Y} + t_{\alpha}(n_1 + n_2 - 2) S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $\underline{\mu_1 - \mu_2} = \bar{X} - \bar{Y} - t_{\alpha}(n_1 + n_2 - 2) S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
	μ_1, μ_2 未知	$F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$	$\left(\frac{S_1^2}{S_2^2} \frac{1}{F_{\alpha/2}(n_1 - 1, n_2 - 1)}, \frac{S_1^2}{S_2^2} \frac{1}{F_{1-\alpha/2}(n_1 - 1, n_2 - 1)} \right)$	$\overline{\frac{\sigma_1^2}{\sigma_2^2}} = \frac{S_1^2}{S_2^2} \frac{1}{F_{1-\alpha}(n_1 - 1, n_2 - 1)}$ $\underline{\frac{\sigma_1^2}{\sigma_2^2}} = \frac{S_1^2}{S_2^2} \frac{1}{F_{\alpha}(n_1 - 1, n_2 - 1)}$

习题

1. 随机地取 8 只活塞环,测得它们的直径为(以 mm 计)

74.001 74.005 74.003 74.001

74.000 73.998 74.006 74.002

试求总体均值 μ 及方差 σ^2 的矩估计值,并求样本方差 s^2 .

2. 设 $X_1, X_2, \dots, X_n$ 为总体的一个样本, $x_1, x_2, \dots, x_n$ 为一相应的样本值. 求下列各总体的概率密度或分布律中的未知参数的矩估计量和矩估计值.

$$(1) f(x) = \begin{cases} \theta c^\theta x^{-(\theta+1)}, & x > c, \\ 0, & \text{其他,} \end{cases} \text{ 其中 } c > 0 \text{ 为已知, } \theta > 1, \theta \text{ 为未知参数.}$$

$$(2) f(x) = \begin{cases} \sqrt{\theta} x^{\sqrt{\theta}-1}, & 0 \leq x \leq 1, \\ 0, & \text{其他,} \end{cases} \text{ 其中 } \theta > 0, \theta \text{ 为未知参数.}$$

$$(3) P\{X=x\} = \binom{m}{x} p^x (1-p)^{m-x}, x=0,1,2,\dots,m, \text{ 其中 } 0 < p < 1, p \text{ 为未知参数.}$$

3. 求上题中各未知参数的最大似然估计值和估计量.

4. (1) 设总体 X 具有分布律

X	1	2	3
p_k	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$

其中 θ ($0 < \theta < 1$) 为未知参数. 已知取得了样本值 $x_1=1, x_2=2, x_3=1$. 试求 θ 的矩估计值和最大似然估计值.

- (2) 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 的最大似然估计量及矩估计量.

- (3) 设随机变量 X 服从以 r, p 为参数的负二项分布, 其分布律为

$$P\{X=x_k\} = \binom{x_k-1}{r-1} p^r (1-p)^{x_k-r}, \quad x_k=r, r+1, \dots,$$

其中 r 已知, p 未知. 设有样本值 $x_1, x_2, \dots, x_n$, 试求 p 的最大似然估计值.

5. 设某种电子器件的寿命(以 h 计) T 服从双参数的指数分布, 其概率密度为

$$f(t) = \begin{cases} \frac{1}{\theta} e^{-(t-c)/\theta}, & t \geq c, \\ 0, & \text{其他,} \end{cases}$$

其中 c, θ ($c, \theta > 0$) 为未知参数. 自一批这种器件中随机地取 n 件进行寿命试验. 设它们的失效时间依次为 $x_1 \leq x_2 \leq \dots \leq x_n$.

- (1) 求 θ 与 c 的最大似然估计值.

- (2) 求 θ 与 c 的矩估计量.

6. 一地质学家为研究密歇根湖湖滩地区的岩石成分, 随机地自该地区取 100 个样品, 每个样品有 10 块石子, 记录了每个样品中属石灰石的石子数. 假设这 100 次观察相互独立, 并

且由过去经验知,它们都服从参数为 $m=10, p$ 的二项分布, p 是这地区一块石子是石灰石的概率. 求 p 的最大似然估计值. 该地质学家所得的数据如下:

样品中属石灰石的石子数 i	0	1	2	3	4	5	6	7	8	9	10
观察到 i 块石灰石的样品个数	0	1	6	7	23	26	21	12	3	1	0

7. (1) 设 $X_1, X_2, \dots, X_n$ 是来自总体 X 的一个样本, 且 $X \sim \pi(\lambda)$, 求 $P\{X=0\}$ 的最大似然估计值.

(2) 某铁路局证实一个扳道员在五年内所引起的严重事故的次数服从泊松分布. 求一个扳道员在五年内未引起严重事故的的概率 p 的最大似然估计. 使用下面 122 个观察值. 下表中, r 表示一扳道员五年中引起严重事故的次数, s 表示观察到的扳道员人数.

r	0	1	2	3	4	5
s	44	42	21	9	4	2

8. (1) 设 $X_1, X_2, \dots, X_n$ 是来自概率密度为

$$f(x; \theta) = \begin{cases} \theta x^{\theta-1}, & 0 < x < 1, \\ 0, & \text{其他} \end{cases}$$

的总体的样本, θ 未知, 求 $U = e^{-1/\theta}$ 的最大似然估计值.

(2) 设 $X_1, X_2, \dots, X_n$ 是来自正态总体 $N(\mu, 1)$ 的样本. μ 未知, 求 $\theta = P\{X > 2\}$ 的最大似然估计值.

(3) 设 $x_1, x_2, \dots, x_n$ 是来自总体 $b(m, \theta)$ 的样本值, 又 $\theta = \frac{1}{3}(1 + \beta)$, 求 β 的最大似然估计值.

9. (1) 验证教材第六章 §3 定理四中的统计量

$$S_w^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

是两总体公共方差 σ^2 的无偏估计量 (S_w^2 称为 σ^2 的合并估计).

(2) 设总体 X 的数学期望为 μ , $X_1, X_2, \dots, X_n$ 是来自 X 的样本, $a_1, a_2, \dots, a_n$ 是任意常数, 验证 $(\sum_{i=1}^n a_i X_i) / \sum_{i=1}^n a_i$ (其中 $\sum_{i=1}^n a_i \neq 0$) 是 μ 的无偏估计量.

10. 设 $X_1, X_2, \dots, X_n$ 是来自总体 X 的一个样本, 设 $E(X) = \mu, D(X) = \sigma^2$.

(1) 确定常数 c , 使 $c \sum_{i=1}^{n-1} (X_{i+1} - X_i)^2$ 为 σ^2 的无偏估计.

(2) 确定常数 c 使 $(\bar{X})^2 - cS^2$ 是 μ^2 的无偏估计 ($\bar{X}, S^2$ 是样本均值和样本方差).

11. 设总体 X 的概率密度为

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} x^{(1-\theta)/\theta}, & 0 < x < 1, \\ 0, & \text{其他.} \end{cases} \quad 0 < \theta < +\infty,$$

$X_1, X_2, \dots, X_n$ 是来自总体 X 的样本.

(1) 验证 θ 的最大似然估计量是 $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n \ln X_i$.

(2) 证明 $\hat{\theta}$ 是 θ 的无偏估计量.

12. 设 X_1, X_2, X_3, X_4 是来自均值为 θ 的指数分布总体的样本, 其中 θ 未知. 设有估计量

$$T_1 = \frac{1}{6}(X_1 + X_2) + \frac{1}{3}(X_3 + X_4),$$

$$T_2 = (X_1 + 2X_2 + 3X_3 + 4X_4)/5,$$

$$T_3 = (X_1 + X_2 + X_3 + X_4)/4.$$

(1) 指出 T_1, T_2, T_3 中哪几个是 θ 的无偏估计量.

(2) 在上述 θ 的无偏估计中指出哪一个较为有效.

13. (1) 设 $\hat{\theta}$ 是参数 θ 的无偏估计, 且有 $D(\hat{\theta}) > 0$, 试证

$$\hat{\theta}^2 = (\hat{\theta})^2 \text{ 不是 } \theta^2 \text{ 的无偏估计.}$$

(2) 试证明均匀分布

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 < x \leq \theta, \\ 0, & \text{其他} \end{cases}$$

中未知参数 θ 的最大似然估计量不是无偏的.

14. 设从均值为 μ , 方差为 $\sigma^2 > 0$ 的总体中分别抽取容量为 n_1, n_2 的两独立样本. $\bar{X}_1$ 和 $\bar{X}_2$ 分别是两样本的均值. 试证: 对于任意常数 a, b ($a+b=1$), $Y=a\bar{X}_1+b\bar{X}_2$ 都是 μ 的无偏估计, 并确定常数 a, b 使 $D(Y)$ 达到最小.

15. 设有 k 台仪器, 已知用第 i 台仪器测量时, 测定值总体的标准差为 σ_i ($i=1, 2, \dots, k$). 用这些仪器独立地对某一物理量 θ 各观察一次, 分别得到 $X_1, X_2, \dots, X_k$. 设仪器都没有系统误差, 即 $E(X_i) = \theta$ ($i=1, 2, \dots, k$). 问 $a_1, a_2, \dots, a_k$ 取何值, 方能使使用 $\hat{\theta} = \sum_{i=1}^k a_i X_i$ 估计 θ

时, $\hat{\theta}$ 是无偏的, 并且 $D(\hat{\theta})$ 最小?

16. 设某种清漆的 9 个样品, 其干燥时间(以 h 计)分别为

$$6.0 \quad 5.7 \quad 5.8 \quad 6.5 \quad 7.0 \quad 6.3 \quad 5.6 \quad 6.1 \quad 5.0$$

设干燥时间总体服从正态分布 $N(\mu, \sigma^2)$. 求 μ 的置信水平为 0.95 的置信区间, (1) 若由以往经验知 $\sigma=0.6$ (h), (2) 若 σ 为未知.

17. 分别使用金球和铂球测定引力常数(单位: $10^{-11} \text{ m}^3 \cdot \text{kg}^{-1} \cdot \text{s}^{-2}$).

(1) 用金球测定观察值为

$$6.683 \quad 6.681 \quad 6.676 \quad 6.678 \quad 6.679 \quad 6.672$$

(2) 用铂球测定观察值为

$$6.661 \quad 6.661 \quad 6.667 \quad 6.667 \quad 6.664$$

设测定值总体为 $N(\mu, \sigma^2)$, μ, σ^2 均为未知. 试就(1), (2)两种情况分别求 μ 的置信水平为 0.9 的置信区间, 并求 σ^2 的置信水平为 0.9 的置信区间.

18. 随机地取某种炮弹 9 发做试验, 得炮口速度的样本标准差 $s=11$ m/s. 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 0.95 的置信区间.

19. 设 $X_1, X_2, \dots, X_n$ 是来自分布 $N(\mu, \sigma^2)$ 的样本, μ 已知, σ 未知.

(1) 验证 $\sum_{i=1}^n (X_i - \mu)^2 / \sigma^2 \sim \chi^2(n)$. 利用这一结果构造 σ^2 的置信水平为 $1-\alpha$ 的置信区间.

(2) 设 $\mu=6.5$, 且有样本值 7.5, 2.0, 12.1, 8.8, 9.4, 7.3, 1.9, 2.8, 7.0, 7.3. 试求 σ 的置信水平为 0.95 的置信区间.

20. 在第 17 题中, 设用金球和用铂球测定时测定值总体的方差相等. 求两个测定值总体均值差的置信水平为 0.90 的置信区间.

21. 随机地从 A 批导线中抽 4 根, 又从 B 批导线中抽 5 根, 测得电阻(Ω)为

A 批导线: 0.143 0.142 0.143 0.137

B 批导线: 0.140 0.142 0.136 0.138 0.140

设测定数据分别来自分布 $N(\mu_1, \sigma^2), N(\mu_2, \sigma^2)$, 且两样本相互独立. 又 μ_1, μ_2, σ^2 均为未知. 试求 $\mu_1 - \mu_2$ 的置信水平为 0.95 的置信区间.

22. 研究两种固体燃料火箭推进器的燃烧率. 设两者都服从正态分布, 并且已知燃烧率的标准差均近似地为 0.05 cm/s, 取样本容量为 $n_1 = n_2 = 20$. 得燃烧率的样本均值分别为 $\bar{x}_1 = 18$ cm/s, $\bar{x}_2 = 24$ cm/s, 设两样本独立. 求两燃烧率总体均值差 $\mu_1 - \mu_2$ 的置信水平为 0.99 的置信区间.

23. 设两位化验员 A, B 独立地对某种聚合物含氯量用相同的方法各做 10 次测定, 其测定值的样本方差依次为 $s_A^2 = 0.5419, s_B^2 = 0.6065$. 设 σ_A^2, σ_B^2 分别为 A, B 所测定的测定值总体的方差. 设总体均为正态的, 且两样本独立. 求方差比 σ_A^2 / σ_B^2 的置信水平为 0.95 的置信区间.

24. 在一批货物的容量为 100 的样本中, 经检验发现有 16 只次品, 试求这批货物次品率的置信水平为 0.95 的置信区间.

25. (1) 求第 16 题中 μ 的置信水平为 0.95 的单侧置信上限.

(2) 求第 21 题中 $\mu_1 - \mu_2$ 的置信水平为 0.95 的单侧置信下限.

(3) 求第 23 题中方差比 σ_A^2 / σ_B^2 的置信水平为 0.95 的单侧置信上限.

26. 为研究某种汽车轮胎的磨损特性, 随机地选择 16 只轮胎, 每只轮胎行驶到磨坏为止, 记录所行驶的路程(以 km 计)如下:

41 250 40 187 43 175 41 010 39 265 41 872 42 654 41 287

38 970 40 200 42 550 41 095 40 680 43 500 39 775 40 400

假设这些数据来自正态总体 $N(\mu, \sigma^2)$, 其中 μ, σ^2 未知, 试求 μ 的置信水平为 0.95 的单侧置信下限.

27. 科学上的重大发现往往是由年轻人做出的. 下面列出了自 16 世纪中叶至 20 世纪早期的十二项重大发现的发现者和他们发现时的年龄:

发现内容	发现者	发现时间	年龄
1. 地球绕太阳运转	哥白尼(Copernicus)	1543	40
2. 望远镜、天文学的基本定律	伽利略(Galileo)	1600	36
3. 运动原理、重力、微积分	牛顿(Newton)	1665	23
4. 电的本质	富兰克林(Franklin)	1746	40

5. 燃烧是与氧气联系着的	拉瓦锡 (Lavoisier)	1774	31
6. 地球是渐进过程演化成的	莱尔 (Lyell)	1830	33
7. 自然选择控制演化的证据	达尔文 (Darwin)	1858	49
8. 光的场方程	麦克斯韦 (Maxwell)	1864	33
9. 放射性	居里 (Curie)	1896	34
10. 量子论	普朗克 (Planck)	1901	43
11. 狭义相对论, $E=mc^2$	爱因斯坦 (Einstein)	1905	26
12. 量子论的数学基础	薛定谔 (Schrödinger)	1926	39

设样本来自正态总体, 试求发现者的平均年龄 μ 的置信水平为 0.95 的单侧置信上限.



第八章 假设检验

§ 1 假设检验

统计推断的另一类重要问题是假设检验问题. 在总体的分布函数完全未知或只知其形式、但不知其参数的情况, 为了推断总体的某些未知特性, 提出某些关于总体的假设. 例如, 提出总体服从泊松分布的假设, 又如, 对于正态总体提出数学期望等于 μ_0 的假设等. 我们要根据样本对所提出的假设作出是接受, 还是拒绝的决策. 假设检验是作出这一决策的过程. 这里, 先结合例子来说明假设检验的基本思想和做法.

例 1 某车间用一台包装机包装葡萄糖. 袋装糖的净重是一个随机变量, 它服从正态分布. 当机器正常时, 其均值为 0.5 kg, 标准差为 0.015 kg. 某日开工后为检验包装机是否正常, 随机地抽取它所包装的糖 9 袋, 称得净重为(kg)

0.497 0.506 0.518 0.524 0.498 0.511 0.520 0.515 0.512

问机器是否正常?

以 μ, σ 分别表示这一天袋装糖的净重总体 X 的均值和标准差. 由于长期实践表明标准差比较稳定, 我们就设 $\sigma = 0.015$. 于是 $X \sim N(\mu, 0.015^2)$, 这里 μ 未知. 问题是根据样本值来判断 $\mu = 0.5$ 还是 $\mu \neq 0.5$. 为此, 我们提出两个相互对立的假设

$$H_0: \mu = \mu_0 = 0.5$$

和

$$H_1: \mu \neq \mu_0.$$

然后, 我们给出一个合理的法则, 根据这一法则, 利用已知样本作出决策是接受假设 H_0 (即拒绝假设 H_1), 还是拒绝假设 H_0 (即接受假设 H_1). 如果作出的决策是接受 H_0 , 则认为 $\mu = \mu_0$, 即认为机器工作是正常的, 否则, 则认为是不正常的.

由于要检验的假设涉及总体均值 μ , 故首先想到是否可借助样本均值 $\bar{X}$ 这一统计量来进行判断. 我们知道, $\bar{X}$ 是 μ 的无偏估计, $\bar{X}$ 的观察值 $\bar{x}$ 的大小在一定程度上反映 μ 的大小. 因此, 如果假设 H_0 为真, 则观察值 $\bar{x}$ 与 μ_0 的偏差 $|\bar{x} - \mu_0|$ 一般不应太大. 若 $|\bar{x} - \mu_0|$ 过分大, 我们就怀疑假设 H_0 的正确性而拒绝 H_0 ,

并考虑到当 H_0 为真时 $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$. 而衡量 $|\bar{x}-\mu_0|$ 的大小可归结为衡量

$\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}}$ 的大小. 基于上面的想法, 我们可适当选定一正数 k , 使当观察值 $\bar{x}$ 满足

$\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} \geq k$ 时就拒绝假设 H_0 , 反之, 若 $\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} < k$, 就接受假设 H_0 .

然而, 由于作出决策的依据是一个样本, 当实际上 H_0 为真时仍可能作出拒绝 H_0 的决策 (这种可能性是无法消除的), 这是一种错误, 犯这种错误的概率记为

$P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\}$ 或 $P_{\mu_0}\{\text{拒绝 } H_0\}$ 或 $P_{\mu \in H_0}\{\text{拒绝 } H_0\}$.

记号 $P_{\mu_0}\{\cdot\}$ 表示参数 μ 取 μ_0 时事件 $\{\cdot\}$ 的概率, $P_{\mu \in H_0}\{\cdot\}$ 表示 μ 取 H_0 规定的值时事件 $\{\cdot\}$ 的概率. 我们无法排除犯这类错误的可能性, 因此自然希望将犯这类错误的概率控制在一定限度之内, 即给出一个较小的数 α ($0 < \alpha < 1$), 使犯这类错误的概率不超过 α , 即使得

$$P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} \leq \alpha. \quad (1.1)$$

为了确定常数 k , 我们考虑统计量 $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$. 由于只允许犯这类错误的概率最大为 α , 令 (1.1) 式右端取等号, 即令

$$P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} = P_{\mu_0}\left\{\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| \geq k\right\} = \alpha,$$

由于当 H_0 为真时, $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$, 由标准正态分布分位点的定义得 (如图 8-1).

$$k = z_{\alpha/2}.$$

因而, 若 Z 的观察值满足

$$|z| = \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| \geq k = z_{\alpha/2},$$

则拒绝 H_0 , 而若

$$|z| = \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| < k = z_{\alpha/2},$$

则接受 H_0 .

例如, 在本例中取 $\alpha = 0.05$, 则有 $k = z_{0.05/2} = z_{0.025} = 1.96$, 又已知 $n = 9$, $\sigma = 0.015$, 再由样本算得 $\bar{x} = 0.511$, 即有

$$\left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| = 2.2 > 1.96,$$



图 8-1

于是拒绝 H_0 , 认为这天包装机工作不正常. $\square$

上例中所采用的检验法则是符合实际推断原理的. 因通常 α 总是取得较小, 一般取 $\alpha=0.01, 0.05$. 因而若 H_0 为真, 即当 $\mu=\mu_0$ 时, $\left\{ \left| \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \right| \geq z_{\alpha/2} \right\}$ 是一个小概率事件, 根据实际推断原理, 就可以认为, 如果 H_0 为真, 则由一次试验得到的观察值 $\bar{x}$, 满足不等式 $\left| \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \right| \geq z_{\alpha/2}$ 几乎是不会发生的. 现在在一次观察中竟然出现了满足 $\left| \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \right| \geq z_{\alpha/2}$ 的 $\bar{x}$, 则我们有理由怀疑原来的假设 H_0 的正确性, 因而拒绝 H_0 . 若出现的观察值 $\bar{x}$ 满足 $\left| \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \right| < z_{\alpha/2}$, 此时没有理由拒绝假设 H_0 , 因此只能接受假设 H_0 .

在上例的做法中, 我们看到当样本容量固定时, 选定 α 后, 数 k 就可以确定, 然后按照统计量 $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ 的观察值的绝对值 $|z|$ 大于等于 k 还是小于 k 来作出决策. 数 k 是检验上述假设的一个门槛值. 如果 $|z| = \left| \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \right| \geq k$, 则称 $\bar{x}$ 与 μ_0 的差异是显著的, 这时拒绝 H_0 ; 反之, 如果 $|z| = \left| \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \right| < k$, 则称 $\bar{x}$ 与 μ_0 的差异是不显著的, 这时接受 H_0 . 数 α 称为显著性水平, 上面关于 $\bar{x}$ 与 μ_0 有无显著差异的判断是在显著性水平 α 之下作出的.

统计量 $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ 称为检验统计量.

前面的检验问题通常叙述成: 在显著性水平 α 下, 检验假设

$$H_0: \mu = \mu_0, \quad H_1: \mu \neq \mu_0. \quad (1.2)$$

也常说成“在显著性水平 α 下, 针对 H_1 检验 H_0 ”. H_0 称为原假设或零假设, H_1 称为备择假设 (意指在原假设被拒绝后可供选择的假设). 我们要进行的工作是, 根据样本, 按上述检验方法作出决策在 H_0 与 H_1 两者之间接受其一.

当检验统计量取某个区域 C 中的值时, 我们拒绝原假设 H_0 , 则称区域 C 为拒绝域, 拒绝域的边界点称为临界点. 如在上例中拒绝域为 $|z| \geq z_{\alpha/2}$, 而 $z = -z_{\alpha/2}, z = z_{\alpha/2}$ 为临界点.

由于检验法则是根据样本作出的, 总有可能作出错误的决策. 如上面所说的那样, 在假设 H_0 实际上为真时, 我们可能犯拒绝 H_0 的错误, 称这类“弃真”的错误为第 I 类错误. 又当 H_0 实际上不真时, 我们也有可能接受 H_0 . 称这类“取伪”的错误为第 II 类错误. 犯第 II 类错误的概率记为

$$P\{\text{当 } H_0 \text{ 不真接受 } H_0\} \text{ 或 } P_{\mu \in H_1}\{\text{接受 } H_0\}.$$

为此,在确定检验法则时,我们应尽可能使犯两类错误的概率都较小.但是,进一步讨论可知,一般来说,当样本容量固定时,若减少犯一类错误的概率,则犯另一类错误的概率往往增大.若要使犯两类错误的概率都减小,除非增加样本容量.在给定样本容量的情况下,一般来说,我们总是控制犯第 I 类错误的概率,使它不大于 α . α 的大小视具体情况而定,通常 α 取 0.1, 0.05, 0.01, 0.005 等值.这种只对犯第 I 类错误的概率加以控制,而不考虑犯第 II 类错误的概率的检验,称为显著性检验.

形如(1.2)式中的备择假设 H_1 , 表示 μ 可能大于 μ_0 , 也可能小于 μ_0 , 称为双边备择假设, 而称形如(1.2)的假设检验为双边假设检验.

有时,我们只关心总体均值是否增大,例如,试验新工艺以提高材料的强度.这时,所考虑的总体的均值应该越大越好.如果我们能判断在新工艺下总体均值较以往正常生产的大,则可考虑采用新工艺.此时,我们需要检验假设

$$H_0: \mu \leq \mu_0, \quad H_1: \mu > \mu_0. \quad (1.3)$$

形如(1.3)的假设检验,称为右边检验.类似地,有时我们需要检验假设

$$H_0: \mu \geq \mu_0, \quad H_1: \mu < \mu_0. \quad (1.4)$$

形如(1.4)的假设检验,称为左边检验.右边检验和左边检验统称为单边检验.

下面来讨论单边检验的拒绝域.

设总体 $X \sim N(\mu, \sigma^2)$, μ 未知、 σ 为已知, $X_1, X_2, \dots, X_n$ 是来自 X 的样本. 给定显著性水平 α . 我们来求检验问题(1.3)

$$H_0: \mu \leq \mu_0, \quad H_1: \mu > \mu_0$$

的拒绝域.

因 H_0 中的全部 μ 都比 H_1 中的 μ 要小, 当 H_1 为真时, 观察值 $\bar{x}$ 往往偏大, 因此, 拒绝域的形式为

$$\bar{x} \geq k \quad (k \text{ 是某一正常数}).$$

下面来确定常数 k , 其做法与例 1 中的做法类似.

$$P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} = P_{\mu \in H_0}\{\bar{X} \geq k\}$$

$$= P_{\mu \leq \mu_0} \left\{ \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{k - \mu_0}{\sigma/\sqrt{n}} \right\}$$

$$\leq P_{\mu \leq \mu_0} \left\{ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{k - \mu_0}{\sigma/\sqrt{n}} \right\}$$

(上式不等号成立是由于 $\mu \leq \mu_0$, $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$, 事件 $\left\{ \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{k - \mu_0}{\sigma/\sqrt{n}} \right\} \subset \left\{ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{k - \mu_0}{\sigma/\sqrt{n}} \right\}$). 要控制 $P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} \leq \alpha$, 只需令

$$P_{\mu \leq \mu_0} \left\{ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{k - \mu_0}{\sigma/\sqrt{n}} \right\} = \alpha. \quad (1.5)$$

由于 $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, 由 (1.5) 得到 $\frac{k - \mu_0}{\sigma/\sqrt{n}} =$

z_α (如图 8-2), $k = \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha$, 即得检验问题

(1.3) 的拒绝域为

$$\bar{x} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha,$$

即

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha. \quad (1.6)$$

类似地, 可得左边检验问题 (1.4)

$$H_0: \mu \geq \mu_0, \quad H_1: \mu < \mu_0$$

的拒绝域为

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha. \quad (1.7)$$

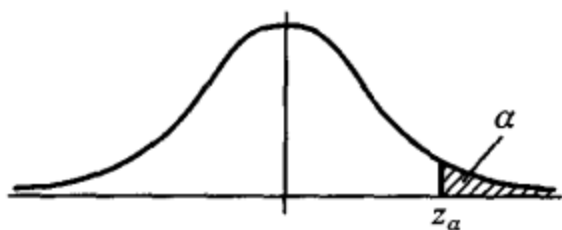


图 8-2

例 2 公司从生产商购买牛奶。公司怀疑生产商在牛奶中掺水以谋利。通过测定牛奶的冰点, 可以检验出牛奶是否掺水。天然牛奶的冰点温度近似服从正态分布, 均值 $\mu_0 = -0.545^\circ\text{C}$, 标准差 $\sigma = 0.008^\circ\text{C}$ 。牛奶掺水可使冰点温度升高而接近于水的冰点温度 (0°C)。测得生产商提交的 5 批牛奶的冰点温度, 其均值为 $\bar{x} = -0.535^\circ\text{C}$, 问是否可以认为生产商在牛奶中掺了水? 取 $\alpha = 0.05$ 。

解 按题意需检验假设

$$H_0: \mu \leq \mu_0 = -0.545 \text{ (即设牛奶未掺水),}$$

$$H_1: \mu > \mu_0 \text{ (即设牛奶已掺水)}$$

这是右边检验问题, 其拒绝域如 (1.6) 式所示, 即为

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{0.05} = 1.645.$$

现在 $z = \frac{-0.535 - (-0.545)}{0.008/\sqrt{5}} = 2.7951 > 1.645$, z 的值落在拒绝域中, 所以我

们在显著性水平 $\alpha = 0.05$ 下拒绝 H_0 , 即认为牛奶商在牛奶中掺了水。□

综上所述, 可得处理参数的假设检验问题的步骤如下:

- 1° 根据实际问题的要求, 提出原假设 H_0 及备择假设 H_1 ;
- 2° 给定显著性水平 α 以及样本容量 n ;
- 3° 确定检验统计量以及拒绝域的形式;
- 4° 按 $P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} \leq \alpha$ 求出拒绝域;

5° 取样, 根据样本观察值作出决策, 是接受 H_0 还是拒绝 H_0 .

下面我们只讨论正态总体参数的假设检验问题.

§ 2 正态总体均值的假设检验

(一) 单个总体 $N(\mu, \sigma^2)$ 均值 μ 的检验

1. σ^2 已知, 关于 μ 的检验 (Z 检验)

在 § 1 中已讨论过正态总体 $N(\mu, \sigma^2)$ 当 σ^2 已知时关于 μ 的检验问题 (1. 2), (1. 3), (1. 4). 在这些检验问题中, 我们都是利用统计量 $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ 来确定拒绝域的. 这种检验法常称为 Z 检验法.

2. σ^2 未知, 关于 μ 的检验 (t 检验)

设总体 $X \sim N(\mu, \sigma^2)$, 其中 μ, σ^2 未知, 我们来求检验问题

$$H_0: \mu = \mu_0, \quad H_1: \mu \neq \mu_0$$

的拒绝域 (显著性水平为 α).

设 $X_1, X_2, \dots, X_n$ 是来自总体 X 的样本. 由于 σ^2 未知, 现在不能利用 $\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ 来确定拒绝域了. 注意到 S^2 是 σ^2 的无偏估计, 我们用 S 来代替 σ , 采用

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

作为检验统计量. 当观察值 $|t| = \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right|$ 过分大时就拒绝 H_0 , 拒绝域的形式为

$$|t| = \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \geq k.$$

由第六章 § 3 定理三知, 当 H_0 为真时, $\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$, 故由

$$P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} = P_{\mu_0} \left\{ \left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| \geq k \right\} = \alpha$$

得 $k = t_{\alpha/2}(n-1)$, 即得拒绝域为

$$|t| = \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \geq t_{\alpha/2}(n-1). \quad (2.1)$$

对于正态总体 $N(\mu, \sigma^2)$, 当 σ^2 未知时, 关于 μ 的单边检验的拒绝域在表 8.1 中给出.

上述利用 t 统计量得出的检验法称为 t 检验法.

在实际中,正态总体的方差常为未知,所以我们常用 t 检验法来检验关于正态总体均值的检验问题.

例 1 某种元件的寿命 X (以 h 计)服从正态分布 $N(\mu, \sigma^2)$, μ, σ^2 均未知. 现测得 16 只元件的寿命如下:

159 280 101 212 224 379 179 264
222 362 168 250 149 260 485 170

问是否有理由认为元件的平均寿命大于 225 h?

解 按题意需检验

$$H_0: \mu \leq \mu_0 = 225, \quad H_1: \mu > 225.$$

取 $\alpha = 0.05$. 由表 8.1 知此检验问题的拒绝域为

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \geq t_{\alpha}(n-1).$$

现在 $n = 16$, $t_{0.05}(15) = 1.753$. 又算得 $\bar{x} = 241.5$, $s = 98.7259$, 即有

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = 0.6685 < 1.753.$$

t 没有落在拒绝域中, 故接受 H_0 , 即认为元件的平均寿命不大于 225 h. $\square$

(二) 两个正态总体均值差的检验(t 检验)

我们还可以用 t 检验法检验具有相同方差的两正态总体均值差的假设. 设 $X_1, X_2, \dots, X_{n_1}$ 是来自正态总体 $N(\mu_1, \sigma^2)$ 的样本, $Y_1, Y_2, \dots, Y_{n_2}$ 是来自正态总体 $N(\mu_2, \sigma^2)$ 的样本, 且设两样本独立. 又分别记它们的样本均值为 $\bar{X}, \bar{Y}$, 记样本方差为 S_1^2, S_2^2 . 设 μ_1, μ_2, σ^2 均为未知, 要特别引起注意的是, 在这里假设两总体的方差是相等的. 现在来求检验问题:

$$H_0: \mu_1 - \mu_2 = \delta, \quad H_1: \mu_1 - \mu_2 \neq \delta$$

(δ 为已知常数) 的拒绝域. 取显著性水平为 α .

引用下述 t 统计量作为检验统计量:

$$t = \frac{(\bar{X} - \bar{Y}) - \delta}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

其中
$$S_w^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \quad S_w = \sqrt{S_w^2}.$$

当 H_0 为真时, 由第六章 §3 定理四知 $t \sim t(n_1 + n_2 - 2)$. 与单个总体的 t 检验法相仿, 其拒绝域的形式为

$$\left| \frac{(\bar{x} - \bar{y}) - \delta}{s_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \geq k.$$

由

$$P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} = P_{\mu_1 - \mu_2 = \delta} \left\{ \left| \frac{(\bar{X} - \bar{Y}) - \delta}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \geq k \right\} = \alpha$$

可得 $k = t_{\alpha/2}(n_1 + n_2 - 2)$. 于是得拒绝域为

$$|t| = \frac{|(\bar{x} - \bar{y}) - \delta|}{s_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq t_{\alpha/2}(n_1 + n_2 - 2). \quad (2.2)$$

关于均值差的两个单边检验问题的拒绝域在表 8.1 中给出. 常用的是 $\delta = 0$ 的情况.

当两个正态总体的方差均为已知(不一定相等)时, 我们可用 Z 检验法来检验两正态总体均值差的假设问题, 见表 8.1.

例 2 用两种方法(A 和 B)测定冰自 -0.72°C 转变为 0°C 的水的融化热(以 cal/g 计). 测得以下的数据:

方法 A: 79.98 80.04 80.02 80.04 80.03 80.03

80.04 79.97 80.05 80.03 80.02 80.00 80.02

方法 B: 80.02 79.94 79.98 79.97 79.97 80.03 79.95 78.97

设这两个样本相互独立, 且分别来自正态总体 $N(\mu_1, \sigma^2)$ 和 $N(\mu_2, \sigma^2)$, μ_1, μ_2, σ^2 均未知. 试检验假设(取显著性水平 $\alpha = 0.05$)

$$H_0: \mu_1 - \mu_2 \leq 0, H_1: \mu_1 - \mu_2 > 0.$$

解 分别画出对应于方法 A 和方法 B 的数据的箱线图, 如图 8-3. 这两种方法所得的结果是有明显差异的, 现在来检验上述假设.

$$n_1 = 13, \bar{x}_A = 80.02, s_A^2 = 0.024^2$$

$$n_2 = 8, \bar{x}_B = 79.98, s_B^2 = 0.03^2$$

$$s_w^2 = \frac{12 \times s_A^2 + 7 \times s_B^2}{19} = 0.0007178.$$

$$t = \frac{\bar{x}_A - \bar{x}_B}{s_w \sqrt{1/13 + 1/8}} = 3.33 > t_{0.05}(13 + 8 - 2) = 1.7291.$$

故拒绝 H_0 , 认为方法 A 比方法 B 测得的融化热要大. $\square$

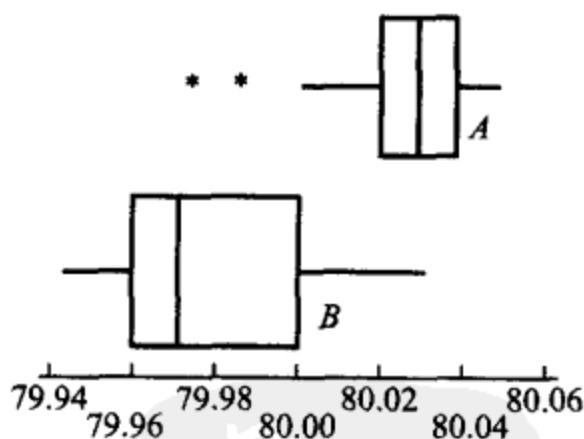


图 8-3

(三) 基于成对数据的检验(t 检验)

有时为了比较两种产品、两种仪器、两种方法等的差异, 我们常在相同的条件下做对比试验, 得到一批成对的观察值. 然后分析观察数据作出推断. 这种方

法常称为逐对比较法.

例 3 有两台光谱仪 I_x, I_y , 用来测量材料中某种金属的含量, 为鉴定它们的测量结果有无显著的差异, 制备了 9 件试块(它们的成分、金属含量、均匀性等均各不相同), 现在分别用这两台仪器对每一试块测量一次, 得到 9 对观察值如下.

$x(\%)$	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
$y(\%)$	0.10	0.21	0.52	0.32	0.78	0.59	0.68	0.77	0.89
$d=x-y(\%)$	0.10	0.09	-0.12	0.18	-0.18	0.11	0.12	0.13	0.11

问能否认为这两台仪器的测量结果有显著的差异(取 $\alpha=0.01$)?

解 本题中的数据是成对的, 即对同一试块测出一对数据. 我们看到一对与另一对之间的差异是由各种因素, 如材料成分、金属含量、均匀性等因素引起的. 由于各试块的特性有广泛的差别, 就不能将仪器 I_x 对 9 个试块的测量结果(即表中第一行)看成是同分布随机变量的观察值. 因而表中第一行不能看成是一个样本的样本值. 同样, 表中第二行也不能看成是一个样本的样本值. 再者, 对于每一对数据而言, 它们是同一试块用不同仪器 I_x, I_y 测得的结果, 因此, 它们不是两个独立的随机变量的观察值. 综上所述, 我们不能用表 8.1 中第 4 栏的检验法来作检验. 而同一对中两个数据的差异则可看成是仅由这两台仪器性能的差异所引起的, 这样, 局限于各对中两个数据来比较就能排除种种其他因素, 而只考虑单独由仪器的性能所产生的影响. 从而能比较这两台仪器的测量结果是否有显著的差异.

一般, 设有 n 对相互独立的观察结果: $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, 令 $D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, \dots, D_n = X_n - Y_n$, 则 $D_1, D_2, \dots, D_n$ 相互独立. 又由于 $D_1, D_2, \dots, D_n$ 是由同一因素所引起的, 可认为它们服从同一分布. 今假设 $D_i \sim N(\mu_D, \sigma_D^2), i=1, 2, \dots, n$. 这就是说 $D_1, D_2, \dots, D_n$ 构成正态总体 $N(\mu_D, \sigma_D^2)$ 的一个样本, 其中 μ_D, σ_D^2 未知. 我们需要基于这一样本检验假设:

- (1) $H_0: \mu_D = 0, H_1: \mu_D \neq 0$;
- (2) $H_0: \mu_D \leq 0, H_1: \mu_D > 0$;
- (3) $H_0: \mu_D \geq 0, H_1: \mu_D < 0$.

分别记 $D_1, D_2, \dots, D_n$ 的样本均值和样本方差的观察值为 $\bar{d}, s_D^2$, 按表 8.1 第 2 栏中关于单个正态总体均值的 t 检验. 知检验问题(1), (2), (3)的拒绝域分别为(显著性水平为 α):

$$|t| = \left| \frac{\bar{d}}{s_D / \sqrt{n}} \right| \geq t_{\alpha/2}(n-1),$$

$$t = \frac{\bar{d}}{s_D/\sqrt{n}} \geq t_\alpha(n-1),$$

$$t = \frac{\bar{d}}{s_D/\sqrt{n}} \leq -t_\alpha(n-1).$$

现在反过来讨论本例的检验问题. 先作出同一试块分别由仪器 I_x, I_y 测得的结果之差, 列于上表的第三行. 按题意需检验假设

$$H_0: \mu_D = 0, H_1: \mu_D \neq 0.$$

现在 $n=9, t_{\alpha/2}(8) = t_{0.005}(8) = 3.3554$ 即知拒绝域为

$$|t| = \left| \frac{\bar{d}}{s_D/\sqrt{n}} \right| \geq 3.3554.$$

由观察值得 $\bar{d}=0.06, s_D=0.1227, |t| = \frac{0.06}{0.1227/\sqrt{9}} = 1.467 < 3.3554$. 现 $|t|$ 的

值不落在拒绝域内, 故接受 H_0 , 认为两台仪器的测量结果并无显著差异. $\square$

例 4 做以下的实验以比较人对红光或绿光的反应时间(以 s 计). 实验在点亮红光或绿光的同时, 启动计时器, 要求受试者见到红光或绿光点亮时, 就按下按钮, 切断计时器, 这就能测得反应时间. 测量的结果如下表:

红光(x)	0.30	0.23	0.41	0.53	0.24	0.36	0.38	0.51
绿光(y)	0.43	0.32	0.58	0.46	0.27	0.41	0.38	0.61
$d=x-y$	-0.13	-0.09	-0.17	0.07	-0.03	-0.05	0.00	-0.10

设 $D_i = X_i - Y_i$ ($i=1, 2, \dots, 8$) 是来自正态总体 $N(\mu_D, \sigma_D^2)$ 的样本, μ_D, σ_D^2 均未知. 试检验假设(取显著性水平 $\alpha=0.05$)

$$H_0: \mu_D \geq 0, H_1: \mu_D < 0.$$

解 现在 $n=8, \bar{x}_d = -0.0625, s_d = 0.0765$, 而

$$\frac{\bar{x}_d}{s_d/\sqrt{8}} = -2.311 < -t_{0.05}(7) = -1.8946,$$

故拒绝 H_0 , 认为 $\mu_D < 0$, 即认为人对红光的反应时间小于对绿光的反应时间, 也就是人对红光的反应要比绿光快. $\square$

§ 3 正态总体方差的假设检验

现在来讨论有关正态总体方差的假设检验问题. 以下分单个总体和两个总体的情况来讨论.

(一) 单个总体的情况

设总体 $X \sim N(\mu, \sigma^2)$, μ, σ^2 均未知, $X_1, X_2, \dots, X_n$ 是来自 X 的样本. 要求检

验假设(显著性水平为 α)

$$H_0: \sigma^2 = \sigma_0^2, \quad H_1: \sigma^2 \neq \sigma_0^2,$$

σ_0^2 为已知常数.

由于 S^2 是 σ^2 的无偏估计, 当 H_0 为真时, 观察值 s^2 与 σ_0^2 的比值 $\frac{s^2}{\sigma_0^2}$ 一般来说应在 1 附近摆动, 而不应过分大于 1 或过分小于 1. 由第六章 §3 定理二知当 H_0 为真时

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1),$$

我们取

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

作为检验统计量, 如上所说知道上述检验问题的拒绝域具有以下的形式:

$$\frac{(n-1)s^2}{\sigma_0^2} \leq k_1 \quad \text{或} \quad \frac{(n-1)s^2}{\sigma_0^2} \geq k_2,$$

此处 k_1, k_2 的值由下式确定:

$$\begin{aligned} & P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} \\ &= P_{\sigma_0^2} \left\{ \left(\frac{(n-1)S^2}{\sigma_0^2} \leq k_1 \right) \cup \left(\frac{(n-1)S^2}{\sigma_0^2} \geq k_2 \right) \right\} = \alpha. \end{aligned}$$

为计算方便起见, 习惯上取

$$P_{\sigma_0^2} \left\{ \frac{(n-1)S^2}{\sigma_0^2} \leq k_1 \right\} = \frac{\alpha}{2}, \quad P_{\sigma_0^2} \left\{ \frac{(n-1)S^2}{\sigma_0^2} \geq k_2 \right\} = \frac{\alpha}{2},$$

故得 $k_1 = \chi_{1-\alpha/2}^2(n-1)$, $k_2 = \chi_{\alpha/2}^2(n-1)$. 于是得拒绝域为

$$\frac{(n-1)s^2}{\sigma_0^2} \leq \chi_{1-\alpha/2}^2(n-1) \quad \text{或} \quad \frac{(n-1)s^2}{\sigma_0^2} \geq \chi_{\alpha/2}^2(n-1) \text{ ①.} \quad (3.1)$$

下面来求单边检验问题(显著性水平为 α)

$$H_0: \sigma^2 \leq \sigma_0^2, \quad H_1: \sigma^2 > \sigma_0^2 \quad (3.2)$$

的拒绝域. 因 H_0 中的全部 σ^2 都比 H_1 中的 σ^2 要小, 当 H_1 为真时, S^2 的观察值 s^2 往往偏大, 因此拒绝域的形式为

$$s^2 \geq k.$$

下面来确定常数 k .

$$\begin{aligned} P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} &= P_{\sigma^2 \leq \sigma_0^2} \{S^2 \geq k\} \\ &= P_{\sigma^2 \leq \sigma_0^2} \left\{ \frac{(n-1)S^2}{\sigma_0^2} \geq \frac{(n-1)k}{\sigma_0^2} \right\} \end{aligned}$$

① 这里指的是 $\frac{(n-1)s^2}{\sigma_0^2} \leq \chi_{1-\alpha/2}^2(n-1)$ 与 $\frac{(n-1)s^2}{\sigma_0^2} \geq \chi_{\alpha/2}^2(n-1)$ 的并集.

$$\leq P_{\sigma^2 \leq \sigma_0^2} \left\{ \frac{(n-1)S^2}{\sigma^2} \geq \frac{(n-1)k}{\sigma_0^2} \right\} \quad (\text{因为 } \sigma^2 \leq \sigma_0^2).$$

要控制 $P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} \leq \alpha$, 只需令

$$P_{\sigma^2 \leq \sigma_0^2} \left\{ \frac{(n-1)S^2}{\sigma^2} \geq \frac{(n-1)k}{\sigma_0^2} \right\} = \alpha. \quad (3.3)$$

因 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, 由 (3.3) 得 $\frac{(n-1)k}{\sigma_0^2}$

$= \chi_{\alpha}^2(n-1)$ (见图 8-4). 于是 $k = \frac{\sigma_0^2}{n-1} \chi_{\alpha}^2$

$(n-1)$, 得检验问题 (3.2) 的拒绝域为 $s^2 \geq$

$\frac{\sigma_0^2}{n-1} \chi_{\alpha}^2(n-1)$, 即

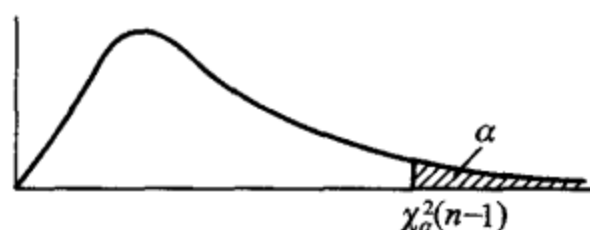


图 8-4

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \geq \chi_{\alpha}^2(n-1). \quad (3.4)$$

类似地, 可得左边检验问题

$$H_0: \sigma^2 \geq \sigma_0^2, \quad H_1: \sigma^2 < \sigma_0^2$$

的拒绝域为

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \leq \chi_{1-\alpha}^2(n-1). \quad (3.5)$$

以上检验法称为 χ^2 检验法.

表 8-1 正态总体均值、方差的检验法 (显著性水平为 α)

	原假设 H_0	检验统计量	备择假设 H_1	拒绝域
1	$\mu \leq \mu_0$ $\mu \geq \mu_0$ $\mu = \mu_0$ $(\sigma^2 \text{ 已知})$	$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$\mu > \mu_0$ $\mu < \mu_0$ $\mu \neq \mu_0$	$z \geq z_{\alpha}$ $z \leq -z_{\alpha}$ $ z \geq z_{\alpha/2}$
2	$\mu \leq \mu_0$ $\mu \geq \mu_0$ $\mu = \mu_0$ $(\sigma^2 \text{ 未知})$	$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$	$\mu > \mu_0$ $\mu < \mu_0$ $\mu \neq \mu_0$	$t \geq t_{\alpha}(n-1)$ $t \leq -t_{\alpha}(n-1)$ $ t \geq t_{\alpha/2}(n-1)$
3	$\mu_1 - \mu_2 \leq \delta$ $\mu_1 - \mu_2 \geq \delta$ $\mu_1 - \mu_2 = \delta$ $(\sigma_1^2, \sigma_2^2 \text{ 已知})$	$Z = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$\mu_1 - \mu_2 > \delta$ $\mu_1 - \mu_2 < \delta$ $\mu_1 - \mu_2 \neq \delta$	$z \geq z_{\alpha}$ $z \leq -z_{\alpha}$ $ z \geq z_{\alpha/2}$

续表

	原假设 H_0	检验统计量	备择假设 H_1	拒绝域
4	$\mu_1 - \mu_2 \leq \delta$ $\mu_1 - \mu_2 \geq \delta$ $\mu_1 - \mu_2 = \delta$ $(\sigma_1^2 = \sigma_2^2 = \sigma^2 \text{ 未知})$	$t = \frac{\bar{X} - \bar{Y} - \delta}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $S_w^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$	$\mu_1 - \mu_2 > \delta$ $\mu_1 - \mu_2 < \delta$ $\mu_1 - \mu_2 \neq \delta$	$t \geq t_\alpha(n_1 + n_2 - 2)$ $t \leq -t_\alpha(n_1 + n_2 - 2)$ $ t \geq t_{\alpha/2}(n_1 + n_2 - 2)$
5	$\sigma^2 \leq \sigma_0^2$ $\sigma^2 \geq \sigma_0^2$ $\sigma^2 = \sigma_0^2$ $(\mu \text{ 未知})$	$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$	$\sigma^2 > \sigma_0^2$ $\sigma^2 < \sigma_0^2$ $\sigma^2 \neq \sigma_0^2$	$\chi^2 \geq \chi_\alpha^2(n-1)$ $\chi^2 \leq \chi_{1-\alpha}^2(n-1)$ $\chi^2 \geq \chi_{\alpha/2}^2(n-1)$ 或 $\chi^2 \leq \chi_{1-\alpha/2}^2(n-1)$
6	$\sigma_1^2 \leq \sigma_2^2$ $\sigma_1^2 \geq \sigma_2^2$ $\sigma_1^2 = \sigma_2^2$ $(\mu_1, \mu_2 \text{ 未知})$	$F = \frac{S_1^2}{S_2^2}$	$\sigma_1^2 > \sigma_2^2$ $\sigma_1^2 < \sigma_2^2$ $\sigma_1^2 \neq \sigma_2^2$	$F \geq F_\alpha(n_1 - 1, n_2 - 1)$ $F \leq F_{1-\alpha}(n_1 - 1, n_2 - 1)$ $F \geq F_{\alpha/2}(n_1 - 1, n_2 - 1)$ 或 $F \leq F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$
7	$\mu_D \leq 0$ $\mu_D \geq 0$ $\mu_D = 0$ (成对数据)	$t = \frac{\bar{D} - 0}{S_D / \sqrt{n}}$	$\mu_D > 0$ $\mu_D < 0$ $\mu_D \neq 0$	$t \geq t_\alpha(n-1)$ $t \leq -t_\alpha(n-1)$ $ t \geq t_{\alpha/2}(n-1)$

例 1 某厂生产的某种型号的电池,其寿命(以 h 计)长期以来服从方差 $\sigma^2 = 5\,000$ 的正态分布,现有一批这种电池,从它的生产情况来看,寿命的波动性有所改变.现随机取 26 只电池,测出其寿命的样本方差 $s^2 = 9\,200$.问根据这一数据能否推断这批电池的寿命的波动性较以往的有显著的变化(取 $\alpha = 0.02$)?

解 本题要求在水平 $\alpha = 0.02$ 下检验假设

$$H_0: \sigma^2 = 5\,000, \quad H_1: \sigma^2 \neq 5\,000.$$

现在 $n = 26$, $\chi_{\alpha/2}^2(n-1) = \chi_{0.01}^2(25) = 44.314$, $\chi_{1-\alpha/2}^2(25) = \chi_{0.99}^2(25) = 11.524$. $\sigma_0^2 = 5\,000$, 由(3.1)式拒绝域为

$$\frac{(n-1)s^2}{\sigma_0^2} \geq 44.314 \quad \text{或} \quad \frac{(n-1)s^2}{\sigma_0^2} \leq 11.524.$$

由观察值 $s^2 = 9\,200$ 得 $\frac{(n-1)s^2}{\sigma_0^2} = 46 > 44.314$, 所以拒绝 H_0 , 认为这批电池寿命的波动性较以往的有显著的变化. □

(二) 两个总体的情况

设 $X_1, X_2, \dots, X_{n_1}$ 是来自总体 $N(\mu_1, \sigma_1^2)$ 的样本, $Y_1, Y_2, \dots, Y_{n_2}$ 是来自总体 $N(\mu_2, \sigma_2^2)$ 的样本, 且两样本独立. 其样本方差分别为 S_1^2, S_2^2 . 且设 $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ 均为未知. 现在需要检验假设 (显著性水平为 α)

$$H_0: \sigma_1^2 \leq \sigma_2^2, \quad H_1: \sigma_1^2 > \sigma_2^2. \quad (3.6)$$

当 H_0 为真时, $E(S_1^2) = \sigma_1^2 \leq \sigma_2^2 = E(S_2^2)$, 当 H_1 为真时 $E(S_1^2) = \sigma_1^2 > \sigma_2^2 = E(S_2^2)$.

当 H_1 为真时, 观察值 $\frac{S_1^2}{S_2^2}$ 有偏大的趋势, 故拒绝域具有形式

$$\frac{s_1^2}{s_2^2} \geq k,$$

常数 k 确定如下:

$$\begin{aligned} P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} &= P_{\sigma_1^2 \leq \sigma_2^2} \left\{ \frac{S_1^2}{S_2^2} \geq k \right\} \\ &\leq P_{\sigma_1^2 \leq \sigma_2^2} \left\{ \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \geq k \right\} \quad (\text{因为 } \sigma_1^2/\sigma_2^2 \leq 1). \end{aligned}$$

要控制 $P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} \leq \alpha$, 只需令

$$P_{\sigma_1^2 \leq \sigma_2^2} \left\{ \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \geq k \right\} = \alpha. \quad (3.7)$$

由第六章 § 3 定理四知 $\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$, 由 (3.7) 式得 $k = F_\alpha(n_1 - 1, n_2 - 1)$. 即得检验问题 (3.6) 的拒绝域为

$$F = \frac{s_1^2}{s_2^2} \geq F_\alpha(n_1 - 1, n_2 - 1). \quad (3.8)$$

上述检验法称为 F 检验法. 关于 σ_1^2, σ_2^2 的另外两个检验问题的拒绝域在表 8.1 中给出.

例 2 设 § 2 例 2 中的两个样本分别来自总体 $N(\mu_A, \sigma_A^2), N(\mu_B, \sigma_B^2)$, 且两样本独立. 试检验 $H_0: \sigma_A^2 = \sigma_B^2, H_1: \sigma_A^2 \neq \sigma_B^2$, 以说明我们假设 $\sigma_A^2 = \sigma_B^2$ 是合理的. (取显著性水平 $\alpha = 0.01$)

解 此处 $n_1 = 13, n_2 = 8, \alpha = 0.01$, 拒绝域为

$$\frac{s_A^2}{s_B^2} \geq F_{0.005}(12, 7) = 8.18,$$

$$\text{或} \quad \frac{s_A^2}{s_B^2} \leq F_{0.005}(12, 7) = \frac{1}{F_{0.005}(7, 12)} = \frac{1}{5.52} = 0.18.$$

现在 $s_A^2 = 0.024^2, s_B^2 = 0.03^2, s_A^2/s_B^2 = 0.64$,

$$0.18 < 0.64 < 8.18,$$

故接受 H_0 , 认为两总体方差相等. 两总体方差相等也称两总体具有方差齐性, 这也表明 § 2 例 2 中假设两总体方差相等是合理的. $\square$

§ 4 置信区间与假设检验之间的关系

置信区间与假设检验之间有明显的联系, 先考察置信区间与双边检验之间的对应关系. 设 $X_1, \dots, X_n$ 是一个来自总体的样本, $x_1, \dots, x_n$ 是相应的样本值, Θ 是参数 θ 的可能取值范围.

设 $(\underline{\theta}(X_1, \dots, X_n), \bar{\theta}(X_1, \dots, X_n))$ 是参数 θ 的一个置信水平为 $1-\alpha$ 的置信区间, 则对于任意 $\theta \in \Theta$, 有

$$P_{\theta} \{ \underline{\theta}(X_1, \dots, X_n) < \theta < \bar{\theta}(X_1, \dots, X_n) \} \geq 1-\alpha, \quad (4.1)$$

考虑显著性水平为 α 的双边检验

$$H_0: \theta = \theta_0, \quad H_1: \theta \neq \theta_0. \quad (4.2)$$

由 (4.1) 式

$$P_{\theta_0} \{ \underline{\theta}(X_1, \dots, X_n) < \theta_0 < \bar{\theta}(X_1, \dots, X_n) \} \geq 1-\alpha,$$

即有

$$P_{\theta_0} \{ (\theta_0 \leq \underline{\theta}(X_1, \dots, X_n)) \cup (\theta_0 \geq \bar{\theta}(X_1, \dots, X_n)) \} \leq \alpha.$$

按显著性水平为 α 的假设检验的拒绝域的定义, 检验 (4.2) 的拒绝域为

$$\theta_0 \leq \underline{\theta}(x_1, \dots, x_n) \quad \text{或} \quad \theta_0 \geq \bar{\theta}(x_1, \dots, x_n);$$

接受域为

$$\underline{\theta}(x_1, \dots, x_n) < \theta_0 < \bar{\theta}(x_1, \dots, x_n).$$

这就是说, 当我们要检验假设 (4.2) 时, 先求出 θ 的置信水平为 $1-\alpha$ 的置信区间 $(\underline{\theta}, \bar{\theta})$, 然后考察区间 $(\underline{\theta}, \bar{\theta})$ 是否包含 θ_0 , 若 $\theta_0 \in (\underline{\theta}, \bar{\theta})$ 则接受 H_0 , 若 $\theta_0 \notin (\underline{\theta}, \bar{\theta})$, 则拒绝 H_0 .

反之, 对于任意 $\theta_0 \in \Theta$, 考虑显著性水平为 α 的假设检验问题

$$H_0: \theta = \theta_0, \quad H_1: \theta \neq \theta_0,$$

假设它的接受域为

$$\underline{\theta}(x_1, \dots, x_n) < \theta_0 < \bar{\theta}(x_1, \dots, x_n),$$

即有

$$P_{\theta_0} \{ \underline{\theta}(X_1, \dots, X_n) < \theta_0 < \bar{\theta}(X_1, \dots, X_n) \} \geq 1-\alpha.$$

由 θ_0 的任意性, 由上式知对于任意 $\theta \in \Theta$, 有

$$P_{\theta} \{ \underline{\theta}(X_1, \dots, X_n) < \theta < \bar{\theta}(X_1, \dots, X_n) \} \geq 1-\alpha.$$

因此 $(\underline{\theta}(X_1, \dots, X_n), \bar{\theta}(X_1, \dots, X_n))$ 是参数 θ 的一个置信水平为 $1-\alpha$ 的置信区间.

这就是说,为要求出参数 θ 的置信水平为 $1-\alpha$ 的置信区间,我们先求出显著性水平为 α 的假设检验问题: $H_0: \theta = \theta_0, H_1: \theta \neq \theta_0$ 的接受域 $\underline{\theta}(x_1, \dots, x_n) < \theta_0 < \bar{\theta}(x_1, \dots, x_n)$, 那么, $(\underline{\theta}(X_1, \dots, X_n), \bar{\theta}(X_1, \dots, X_n))$ 就是 θ 的置信水平为 $1-\alpha$ 的置信区间.

还可验证,置信水平为 $1-\alpha$ 的单侧置信区间 $(-\infty, \bar{\theta}(X_1, \dots, X_n))$ 与显著性水平为 α 的左边检验问题 $H_0: \theta \geq \theta_0, H_1: \theta < \theta_0$ 有类似的对应关系. 即若已求得单侧置信区间 $(-\infty, \bar{\theta}(X_1, \dots, X_n))$, 则当 $\theta_0 \in (-\infty, \bar{\theta}(x_1, \dots, x_n))$ 时接受 H_0 , 当 $\theta_0 \notin (-\infty, \bar{\theta}(x_1, \dots, x_n))$ 时拒绝 H_0 . 反之, 若已求得检验问题 $H_0: \theta \geq \theta_0, H_1: \theta < \theta_0$ 的接受域为 $-\infty < \theta_0 \leq \bar{\theta}(x_1, \dots, x_n)$, 则可得 θ 的一个单侧置信区间 $(-\infty, \bar{\theta}(X_1, \dots, X_n))$.

置信水平为 $1-\alpha$ 的单侧置信区间 $(\underline{\theta}(X_1, \dots, X_n), \infty)$ 与显著性水平为 α 的右边检验问题 $H_0: \theta \leq \theta_0, H_1: \theta > \theta_0$ 也有类似的对应关系. 即若已求得单侧置信区间 $(\underline{\theta}(X_1, \dots, X_n), \infty)$, 则当 $\theta_0 \in (\underline{\theta}(x_1, \dots, x_n), \infty)$ 时接受 H_0 , 当 $\theta_0 \notin (\underline{\theta}(x_1, \dots, x_n), \infty)$ 时拒绝 H_0 . 反之, 若已求得检验问题 $H_0: \theta \leq \theta_0, H_1: \theta > \theta_0$ 的接受域为 $\underline{\theta}(x_1, \dots, x_n) \leq \theta_0 < \infty$, 则可得 θ 的一个单侧置信区间 $(\underline{\theta}(X_1, \dots, X_n), \infty)$.

例 1 设 $X \sim N(\mu, 1)$, μ 未知, $\alpha = 0.05, n = 16$, 且由一样本算得 $\bar{x} = 5.20$, 于是得到参数 μ 的一个置信水平为 0.95 的置信区间

$$\begin{aligned} (\bar{x} - \frac{1}{\sqrt{16}} z_{0.025}, \bar{x} + \frac{1}{\sqrt{16}} z_{0.025}) &= (5.20 - 0.49, 5.20 + 0.49) \\ &= (4.71, 5.69). \end{aligned}$$

现在考虑检验问题 $H_0: \mu = 5.5, H_1: \mu \neq 5.5$. 由于 $5.5 \in (4.71, 5.69)$, 故接受 H_0 . □

例 2 数据如上例. 试求右边检验问题 $H_0: \mu \leq \mu_0, H_1: \mu > \mu_0$ 的接受域, 并求 μ 的单侧置信下限 ($\alpha = 0.05$).

解 检验问题的拒绝域为 $z = \frac{\bar{x} - \mu_0}{1/\sqrt{16}} \geq z_{0.05}$, 或即 $\mu_0 \leq 4.79$. 于是检验问题的接受域为 $\mu_0 > 4.79$. 这样就得到 μ 的单侧置信区间 $(4.79, \infty)$, 单侧置信下限 $\mu = 4.79$. □

§ 5 样本容量的选取

以上我们在进行假设检验时,总是根据问题的要求,预先给出显著性水平以控制犯第 I 类错误的概率,而犯第 II 类错误的概率则依赖于样本容量的选择. 在

一些实际问题中,我们除了希望控制犯第 I 类错误的概率外,往往还希望控制犯第 II 类错误的概率.在这一节,我们将阐明如何选取样本的容量使得犯第 II 类错误的概率控制在预先给定的限度之内.为此,我们引入施行特征函数.

定义 若 C 是参数 θ 的某检验问题的一个检验法,

$$\beta(\theta) = P_{\theta}(\text{接受 } H_0) \quad (5.1)$$

称为检验法 C 的施行特征函数或 OC 函数,其图形称为 OC 曲线.

由定义知,若此检验法的显著性水平为 α ,那么当真值 $\theta \in H_0$ 时, $\beta(\theta)$ 就是作出正确判断(即 H_0 为真时接受 H_0)的概率,故此时 $\beta(\theta) \geq 1 - \alpha$;而当 $\theta \in H_1$ 时,则 $\beta(\theta)$ 就是犯第 II 类错误的概率,而 $1 - \beta(\theta)$ 是作出正确判断(即 H_0 为不真时拒绝 H_0)的概率.函数 $1 - \beta(\theta)$ 称为检验法 C 的功效函数.当 $\theta^* \in H_1$ 时,值 $1 - \beta(\theta^*)$ 称为检验法 C 在点 θ^* 的功效.它表示当参数 θ 的真值为 θ^* 时,检验法 C 作出正确判断的概率.

本书只介绍正态总体均值的检验法的 OC 函数及其图形.

1. Z 检验法的 OC 函数

右边检验问题, $H_0: \mu \leq \mu_0$, $H_1: \mu > \mu_0$ 的 OC 函数是

$$\begin{aligned} \beta(\mu) &= P_{\mu}(\text{接受 } H_0) = P_{\mu}\left\{\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha}\right\} \\ &= P_{\mu}\left\{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right\} = \Phi(z_{\alpha} - \lambda), \end{aligned} \quad (5.2)$$

其中 $\lambda = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$. 其图形如图 8-5 所示. 此

OC 函数 $\beta(\mu)$ 有如下性质:

(1) 它是 $\lambda = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$ 的单调递减连续函数;

(2) $\lim_{\mu \rightarrow \mu_0^+} \beta(\mu) = 1 - \alpha$, $\lim_{\mu \rightarrow \infty} \beta(\mu) = 0$.

由 $\beta(\mu)$ 的连续性可知,当参数的真值 $\mu(\mu > \mu_0)$ 在 μ_0 附近时,检验法的功效很低,即 $\beta(\mu)$ 的值很大,亦即犯第 II 类错误的概率很大.因为 α 通常取得比较小,而不管 σ 多么小, n 多么大,只要 n 给定,总存在 μ_0 附近的点 $\mu(\mu > \mu_0)$ 使 $\beta(\mu)$ 几乎等于 $1 - \alpha$.

这表明,无论样本容量 n 多么大,要想对所有 $\mu \in H_1$,即真值为 H_1 所规定的任一点时,控制犯第 II 类错误的概率都很小是不可能的.但是我们可以使用 OC 函数 $\beta(\mu)$ 以确定样本容量 n ,使当真值 $\mu \geq \mu_0 + \delta$ ($\delta > 0$ 为取定的值)时,犯第 II 类错误的概率不超过给定的 β .这是由于 $\beta(\mu)$ 是 μ 的递减函数,故当 $\mu \geq \mu_0 + \delta$

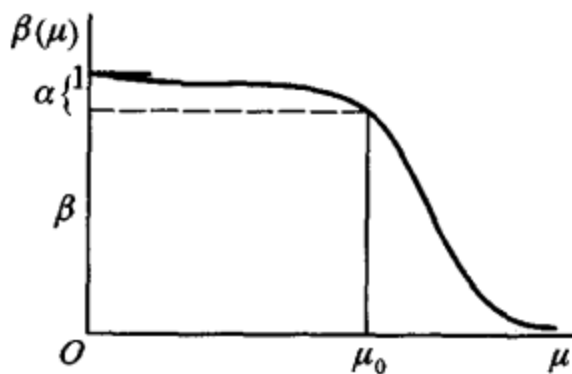


图 8-5

时有

$$\beta(\mu_0 + \delta) \geq \beta(\mu).$$

于是只要 $\beta(\mu_0 + \delta) = \Phi(z_\alpha - \sqrt{n}\delta/\sigma) \leq \beta$, 亦即只要 n 满足

$$z_\alpha - \sqrt{n}\delta/\sigma \leq -z_\beta$$

即可. 这就是说, 只要

$$\sqrt{n} \geq \frac{(z_\alpha + z_\beta)\sigma}{\delta}, \quad (5.3)$$

就能使当 $\mu \in H_1$ 且 $\mu \geq \mu_0 + \delta$ 时 (即真值 $\mu \geq \mu_0 + \delta$ 时) 犯第 II 类错误的概率不超过 β .

类似地, 可得左边检验问题 $H_0: \mu \geq \mu_0, H_1: \mu < \mu_0$ 的 OC 函数为

$$\beta(\mu) = \Phi(z_\alpha + \lambda), \quad \lambda = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}. \quad (5.4)$$

当真值 $\mu \geq \mu_0$ 时 $\beta(\mu)$ 为作出正确判断的概率; 当真值 $\mu < \mu_0$ 时, $\beta(\mu)$ 给出犯第 II 类错误的概率. 只要样本容量 n 满足

$$\sqrt{n} \geq \frac{(z_\alpha + z_\beta)\sigma}{\delta}, \quad (5.5)$$

就能使当 $\mu \in H_1$ 且 $\mu \leq \mu_0 - \delta$ ($\delta > 0$, 为取定的值) 时, 犯第 II 类错误的概率不超过给定的值 β .

双边检验问题 $H_0: \mu = \mu_0, H_1: \mu \neq \mu_0$ 的 OC 函数是

$$\begin{aligned} \beta(\mu) &= P_\mu(\text{接受 } H_0) = P_\mu \left\{ -z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2} \right\} \\ &= P_\mu \left\{ -\lambda - z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < -\lambda + z_{\alpha/2} \right\} = \Phi(z_{\alpha/2} - \lambda) - \Phi(-z_{\alpha/2} - \lambda) \\ &= \Phi(z_{\alpha/2} - \lambda) + \Phi(z_{\alpha/2} + \lambda) - 1, \lambda = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}. \end{aligned} \quad (5.6)$$

OC 曲线如图 8-6 所示. 注意 $\beta(\mu)$ 是 $|\lambda|$ 的严格单调下降函数.

在双边检验问题中, 若要求对 H_1 中满足 $|\mu - \mu_0| \geq \delta > 0$ 的 μ 处的函数值 $\beta(\mu) \leq \beta$, 则需解超越方程

$$\begin{aligned} \beta &= \Phi(z_{\alpha/2} - \sqrt{n}\delta/\sigma) + \\ &\quad \Phi(z_{\alpha/2} + \sqrt{n}\delta/\sigma) - 1 \end{aligned}$$

才能确定 n . 通常因 n 较大, 故总可以认为

$z_{\alpha/2} + \sqrt{n}\delta/\sigma \geq 4$, 于是 $\Phi(z_{\alpha/2} + \sqrt{n}\delta/\sigma) \approx 1$, 故近似地有

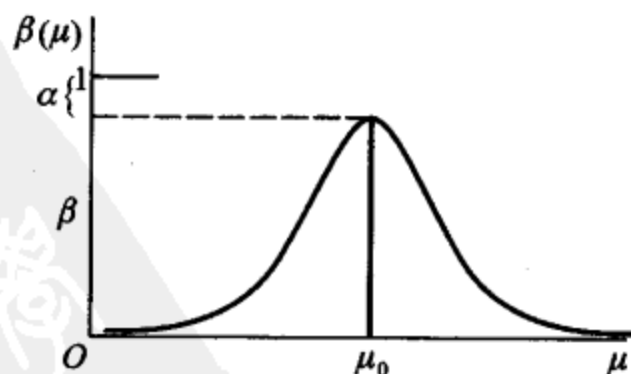


图 8-6

$$\beta \approx \Phi(z_{\alpha/2} - \sqrt{n}\delta/\sigma).$$

由此知只要样本容量 n 满足

$$z_{\alpha/2} - \sqrt{n}\delta/\sigma \leq -z_{\beta},$$

即只要 n 满足

$$\sqrt{n} \geq (z_{\alpha/2} + z_{\beta}) \frac{\sigma}{\delta}, \quad (5.7)$$

就能使当 $\mu \in H_1$ 且 $|\mu - \mu_0| \geq \delta$ ($\delta > 0$, 为取定的值) 时, 犯第 II 类错误的概率不超过给定的值 β .

例 1 (工业产品质量抽验方案) 设有一大批产品, 产品质量指标 $X \sim N(\mu, \sigma^2)$. 以 μ 小者为佳, 厂方要求所确定的验收方案对高质量的产品 ($\mu \leq \mu_0$) 能以高概率 $1 - \alpha$ 为买方所接受. 买方则要求低质产品 ($\mu \geq \mu_0 + \delta, \delta > 0$) 能以高概率 $1 - \beta$ 被拒绝. α, β 由厂方与买方协商给出. 并采取一次抽样以确定该批产品是否为买方所接受. 问应怎样安排抽样方案. 已知 $\mu_0 = 120, \delta = 20$, 且由工厂长期经验知 $\sigma^2 = 900$. 又经商定 α, β 均取为 0.05.

解 检验问题可表达为

$$H_0: \mu \leq \mu_0, \quad H_1: \mu > \mu_0, \quad (5.8)$$

且要求当 $\mu \geq \mu_0 + \delta$ 时能以 $1 - \beta = 0.95$ 的概率拒绝 H_0 . 由 Z 检验, 拒绝域为

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{\alpha},$$

故 OC 函数为

$$\begin{aligned} \beta(\mu) &= P_{\mu} \left\{ \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha} \right\} = P_{\mu} \left\{ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \right\} \\ &= \Phi \left(z_{\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \right). \end{aligned} \quad (5.9)$$

现要求当 $\mu \geq \mu_0 + \delta$ 时 $\beta(\mu) \leq \beta$. 因 $\beta(\mu)$ 是 μ 的递减函数, 故只需 $\beta(\mu_0 + \delta) = \beta$ 即可. 此时, 由 (5.9) 式可得

$$\sqrt{n} \geq \frac{(z_{\alpha} + z_{\beta})\sigma}{\delta}.$$

按给定的数据算得 $n \geq 24.35$, 故取 $n = 25$. 且当 $\bar{x}$ 满足 $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{\alpha} = z_{0.05} = 1.645$

时, 即当 $\bar{x} \geq 129.87$ 时, 买方就拒绝这批产品, 而当 $\bar{x} < 129.87$ 时, 买方接受这批产品. $\square$

2. t 检验法的 OC 函数

右边检验问题 $H_0: \mu \leq \mu_0, H_1: \mu > \mu_0$ 的 t 检验法的 OC 函数是

$$\beta(\mu) = P_{\mu}(\text{接受 } H_0) = P_{\mu} \left\{ \frac{\bar{X} - \mu_0}{S/\sqrt{n}} < t_{\alpha}(n-1) \right\}, \quad (5.10)$$

其中变量

$$\frac{\bar{X}-\mu_0}{S/\sqrt{n}} = \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} + \lambda \right) / \left(\frac{S}{\sigma} \right), \quad \lambda = \frac{\mu-\mu_0}{\sigma/\sqrt{n}}. \quad (5.11)$$

我们称变量 $\frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ 服从非中心参数为 λ 、自由度为 $n-1$ 的非中心 t 分布. 在 $\lambda=0$ 时, 它是通常的 $t(n-1)$ 变量.

若给定 α, β 以及 $\delta > 0$, 则可从书末附表 7 查得所需容量 n , 使得当 $\mu \in H_1$ 且 $\frac{\mu-\mu_0}{\sigma} \geq \delta$ 时犯第 II 类错误的概率不超过 β .

若给定 α, β 以及 $\delta > 0$, 对于左边检验问题 $H_0: \mu \geq \mu_0, H_1: \mu < \mu_0$ 的 t 检验法, 也可从附表 7 查得所需容量 n , 使得当 $\mu \in H_1$ 且 $\frac{\mu-\mu_0}{\sigma} \leq -\delta$ 时犯第 II 类错误的概率不超过 β . 对于双边检验问题 $H_0: \mu = \mu_0, H_1: \mu \neq \mu_0$ 的 t 检验法也可从附表 7 查得所需容量 n , 使得当 $\mu \in H_1$, 且 $\frac{|\mu-\mu_0|}{\sigma} \geq \delta$ 时犯第 II 类错误的概率不超过 β .

例 2 考虑在显著性水平 $\alpha=0.05$ 下进行 t 检验

$$H_0: \mu \leq 68, \quad H_1: \mu > 68.$$

(1) 要求在 H_1 中 $\mu \geq \mu_1 = 68 + \sigma$ 时犯第 II 类错误的概率不超过 $\beta=0.05$. 求所需的样本容量.

(2) 若样本容量为 $n=30$, 问在 H_1 中 $\mu = \mu_1 = 68 + 0.75\sigma$ 时犯第 II 类错误的概率是多少?

解 (1) 此处 $\alpha=\beta=0.05, \mu_0=68, \delta = \frac{\mu_1-\mu_0}{\sigma} = \frac{(68+\sigma)-68}{\sigma} = 1$, 查附表 7 得 $n=13$.

(2) 现在 $\alpha=0.05, n=30, \delta = \frac{\mu_1-\mu_0}{\sigma} = \frac{(68+0.75\sigma)-68}{\sigma} = 0.75$, 查附表 7, 得 $\beta=0.01$. □

例 3 考虑在显著性水平 $\alpha=0.05$ 下进行 t 检验

$$H_0: \mu = 14, \quad H_1: \mu \neq 14.$$

要求在 H_1 中 $\frac{|\mu-14|}{\sigma} \geq 0.4$ 时犯第 II 类错误的概率不超过 $\beta=0.1$, 求所需样本容量.

解 此处 $\alpha=0.05, \beta=0.1, \delta=0.4$, 查附表 7 得 $n=68$. □

在实际问题中, 有时只给出 α, β 及 $|\mu_1-\mu_0|$ 的值, 而需要确定所需的样本容量 n . 这时由于 σ 未知, 不能确定 $\delta = |\mu_1-\mu_0|/\sigma$ 的值, 因而不能直接查表以确定

样本容量. 此时可采用下述近似方法. 先适当取一值 n_1 , 抽取容量为 n_1 的样本, 根据这一样本计算 s^2 的值, 以 s^2 作为 σ^2 的估计, 算出 δ 的近似值. 由 α, β, δ 的值查附表 7 定出样本的容量, 记为 n_2 . 若 $n_1 \geq n_2$, 则取 n_1 作为所求的容量, 即取 $n = n_1$. 否则, 再抽 $n_2 - n_1$ 个独立观察值与原来抽得的观察值合并, 重新计算 δ 的近似值. 然后用 δ 的新近似值和 α, β 查附表 7, 再次定出样本容量. 记为 n_3 . 若 $n_2 \geq n_3$, 则取 $n = n_2$, 否则再按上法重复进行. 一般, 只需试少数几次就可得到所求的样本容量 n .

下面考虑两个正态总体均值差的 t 检验.

若两个正态总体 $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ 中 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 而 σ^2 未知. 在均值差 $\mu_1 - \mu_2$ 的检验问题 $H_0: \mu_1 - \mu_2 = 0, H_1: \mu_1 - \mu_2 \neq 0$ (或 $H_0: \mu_1 - \mu_2 \leq 0, H_1: \mu_1 - \mu_2 > 0$ 或 $H_0: \mu_1 - \mu_2 \geq 0, H_1: \mu_1 - \mu_2 < 0$) 的 t 检验法中, 当分别自两个总体取得的相互独立的样本其容量 $n_1 = n_2 = n$ 时, 给定 α, β 以及 $\delta = |\mu_1 - \mu_2|/\sigma$ 的值后可以查附表 8 得到所需样本容量, 使当 $|\mu_1 - \mu_2|/\sigma \geq \delta$ 时犯第 II 类错误的概率小于或等于 β . 当仅给出 α, β 以及 $|\mu_1 - \mu_2|$ 的值时, 可按类似于上面所说的方法处理.

例 4 需要比较两种汽车用的燃料的辛烷值, 得数据:

燃料 A	81	84	79	76	82	83	84	80	79	82	81	79
燃料 B	76	74	78	79	80	79	82	76	81	79	82	78

燃料的辛烷值越高, 燃料质量越好. 因燃料 B 较燃料 A 价格便宜, 因此, 若两者辛烷值相同时, 则使用燃料 B; 但若含量的均值差 $\mu_A - \mu_B \geq 5$, 则使用燃料 A. 设两总体的分布均可认为是正态的, 而两个样本相互独立. 问应采用哪种燃料 (取 $\alpha = 0.01, \beta = 0.01$)?

解 按题意需要在显著性水平 $\alpha = 0.01$ 下检验假设

$$H_0: \mu_A - \mu_B \leq 0, \quad H_1: \mu_A - \mu_B > 0,$$

并要求在 $\mu_A - \mu_B \geq 5$ 时, 犯第 II 类错误的概率不超过 $\beta = 0.01$.

所取的样本容量为 $n_A = n_B = 12$, 且有 $\bar{x}_A = 80.83, \bar{x}_B = 78.67, s_A^2 = 5.61, s_B^2 = 6.06$. 经显著性水平为 0.1 的 F 检验知, 可认为两总体的方差相等, 即有 $\sigma_A^2 = \sigma_B^2$, 记为 σ^2 . 因 $n_1 = n_2$, 取 $\hat{\sigma}^2 = (s_A^2 + s_B^2)/2 = 5.835$ 作为 σ^2 的点估计, 取 $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$, 于是 $\delta = 5/\hat{\sigma} = 2.07$, 查表, 当 $\alpha = 0.01, \beta = 0.01, \delta = 2.07$ 时 $n \geq 8$. 现 $n = 12$, 故已近似地满足要求. 而右边检验的拒绝域为

$$t = \frac{\bar{x}_A - \bar{x}_B}{s_w \sqrt{1/n_A + 1/n_B}} \geq t_{0.01}(n_1 + n_2 - 2) = 2.5083.$$

由样本观察值算得 $t = 2.19 < 2.5083$, 故接受 H_0 , 即采用燃料 B. □

§ 6 分布拟合检验

上面介绍的各种检验法都是在总体分布形式为已知的前提下进行讨论的.

但在实际问题中,有时不能知道总体服从什么类型的分布,这时就需要根据样本来检验关于分布的假设.本节介绍 χ^2 拟合检验法.它可以用来检验总体是否具有某一个指定的分布或属于某一个分布族,还介绍专用于检验分布是否为正态的“偏度、峰度检验法”.

(一)单个分布的 χ^2 拟合检验法

设总体 X 的分布未知, $x_1, x_2, \dots, x_n$ 是来自 X 的样本值.我们来检验假设

$$H_0: \text{总体 } X \text{ 的分布函数为 } F(x). \quad (6.1)$$

$$H_1: \text{总体 } X \text{ 的分布函数不是 } F(x) \text{ ①}$$

其中设 $F(x)$ 不含未知参数.(也常以分布律或概率密度代替 $F(x)$).

下面来定义检验统计量.将在 H_0 下 X 可能取值的全体 Ω 分成互不相交的子集 $A_1, A_2, \dots, A_k$, 以 f_i ($i=1, 2, \dots, k$) 记样本观察值 $x_1, x_2, \dots, x_n$ 中落在 A_i 的个数,这表示事件 $A_i = \{X \text{ 的值落在子集 } A_i \text{ 内}\}$ 在 n 次独立试验中发生 f_i 次,于是在这 n 次试验中事件 A_i 发生的频率为 f_i/n . 另一方面,当 H_0 为真时,我们可以根据 H_0 中所假设的 X 的分布函数来计算事件 A_i 的概率,得到 $p_i = P(A_i)$, $i=1, 2, \dots, k$. 频率 f_i/n 与概率 p_i 会有差异,但一般来说,当 H_0 为真,且试验的次数又甚多时,这种差异不应太大,因此 $\left(\frac{f_i}{n} - p_i\right)^2$ 不应太大.我们采用形如

$$\sum_{i=1}^k C_i \left(\frac{f_i}{n} - p_i\right)^2 \quad (6.2)$$

的统计量来度量样本与 H_0 中所假设的分布的吻合程度,其中 C_i ($i=1, 2, \dots, k$) ②为给定的常数.皮尔逊证明,如果选取 $C_i = n/p_i$ ($i=1, 2, \dots, k$),则由(6.2)定义的统计量具有下述定理中所述的简单性质.于是我们就采用

$$\chi^2 = \sum_{i=1}^k \frac{n}{p_i} \left(\frac{f_i}{n} - p_i\right)^2 = \sum_{i=1}^k \frac{f_i^2}{np_i} - n \quad (6.3)$$

作为检验统计量.

定理 若 n 充分大($n \geq 50$),则当 H_0 为真时统计量(6.3)近似服从 $\chi^2(k-1)$ 分布.(证略)

据以上的讨论,当 H_0 为真时,(6.3)式中的 χ^2 不应太大,如 χ^2 过分大就拒绝 H_0 ,拒绝域的形式为

$$\chi^2 \geq G \quad (G \text{ 为正常数}).$$

对于给定的显著性水平 α ,确定 G 使

① 在这里备择假设 H_1 可以不必写出.

② 在每一项前乘以 C_i ,是为了能够适当选择 C_i ,使得统计量(6.2)有一个理想的极限分布.

$$P\{\text{当 } H_0 \text{ 为真时拒绝 } H_0\} = P_{H_0}\{\chi^2 \geq G\} = \alpha.$$

由上述定理得 $G = \chi_{\alpha}^2(k-1)$. 即当样本观察值使(6.3)式中的 χ^2 的值有

$$\chi^2 \geq \chi_{\alpha}^2(k-1),$$

则在显著性水平 α 下拒绝 H_0 ; 否则就接受 H_0 . 这就是单个分布的 χ^2 拟合检验法.

χ^2 拟合检验法是基于上述定理得到的, 所以使用时必须注意 n 不能小于 50. 另外 np_i 不能太小, 应有 $np_i \geq 5$, 否则应适当合并 A_i , 以满足这个要求(见下例).

例 1 下表列出了某一地区在夏季的一个月中由 100 个气象站报告的雷暴雨的次数.

i	0	1	2	3	4	5	≥ 6
f_i	22	37	20	13	6	2	0
A_i	A_0	A_1	A_2	A_3	A_4	A_5	A_6

其中 f_i 是报告雷暴雨次数为 i 的气象站数. 试用 χ^2 拟合检验法检验雷暴雨的次数 X 是否服从均值 $\lambda=1$ 的泊松分布(取显著性水平 $\alpha=0.05$).

解 按题意需检验假设

$$H_0: P\{X=i\} = \frac{\lambda^i e^{-\lambda}}{i!} = \frac{e^{-1}}{i!}, \quad i=0, 1, \dots.$$

在 H_0 下 X 所有可能取的值为 $\Omega = \{0, 1, 2, \dots\}$, 将 Ω 分成如表所示的两两不相交的子集 $A_0, A_1, \dots, A_6$, 则有 $P\{X=i\}$ 为

$$p_i = P\{X=i\} = \frac{e^{-1}}{i!}, \quad i=0, 1, \dots, 5.$$

例如

$$p_0 = P\{X=0\} = e^{-1} = 0.36788,$$

$$p_3 = P\{X=3\} = \frac{e^{-1}}{3!} = 0.06131,$$

$$p_6 = P\{X \geq 6\} = 1 - \sum_{i=0}^5 p_i = 0.059.$$

$$n=100.$$

表 8-2 例 1 的 χ^2 拟合检验计算表

A_i	f_i	p_i	np_i	$f_i^2/(np_i)$
$A_0: \{X=0\}$	22	e^{-1}	36.788	13.16
$A_1: \{X=1\}$	37	e^{-1}	36.788	37.21
$A_2: \{X=2\}$	20	$e^{-1}/2$	18.394	21.75
$A_3: \{X=3\}$	13	$e^{-1}/6$	6.131	54.92
$A_4: \{X=4\}$	6	$e^{-1}/24$	1.533	
$A_5: \{X=5\}$	2	$e^{-1}/120$	0.307	
$A_6: \{X \geq 6\}$	0	$1 - \sum_{i=0}^5 p_i$	0.059	

$$\Sigma = 127.04$$

计算结果如表 8-2 所示,其中有些 $np_i < 5$ 的组予以适当合并,使得每组均有 $np_i \geq 5$,如表中第 4 列花括号所示.并组后 $k=4$, χ^2 的自由度为 $k-1=4-1=3$. $\chi_{0.05}^2(k-1) = \chi_{0.05}^2(3) = 7.815$. 现在 $\chi^2 = 127.04 - 100 = 27.04 > 7.815$,故在显著性水平 0.05 下拒绝 H_0 ,认为样本不是来自均值 $\lambda=1$ 的泊松分布. $\square$

例 2 在研究牛的毛色与牛角的有无,这样两对性状分离现象时,用黑色无角牛与红色有角牛杂交,子二代出现黑色无角牛 192 头,黑色有角牛 78 头,红色无角牛 72 头,红色有角牛 18 头,共 360 头,问这两对性状是否符合孟德尔遗传规律中 9:3:3:1 的遗传比例?

解 现将题中的数据列表如下

序号	1	2	3	4
种类	黑色无角	黑色有角	红色无角	红色有角
数量	192	78	72	18
A_i	A_1	A_2	A_3	A_4

以 X 记各种牛的序号,按题意需检验各类牛的头数符合比例 9:3:3:1,即 $(9/16):(3/16):(3/16):(1/16)$. 需检验假设: $H_0: X$ 的分布律为

X	1	2	3	4
p_k	9/16	3/16	3/16	1/16

取显著性水平为 0.1. 所需计算列在表 8-3 中($n=360$).

表 8-3 例 2 的 χ^2 检验计算表

A_i	f_i	p_i	np_i	$f_i^2/(np_i)$
A_1	192	9/16	$360 \times 9/16 = 202.5$	$192^2/202.5 = 182.04$
A_2	78	3/16	$360 \times 3/16 = 67.5$	$78^2/67.5 = 90.13$
A_3	72	3/16	$360 \times 3/16 = 67.5$	$72^2/67.5 = 76.8$
A_4	18	1/16	$360 \times 1/16 = 22.5$	$18^2/22.5 = 14.4$

$$\Sigma = 363.37$$

现在 $\chi^2 = 363.37 - 360 = 3.37$, $k=4$, $\chi_{0.1}^2(4-1) = 6.251 > 3.37$,故接受 H_0 ,认为两性状符合孟德尔遗传规律中 9:3:3:1 的遗传比例. $\square$

(二) 分布族的 χ^2 拟合检验

在(一)中要检验的原假设是 H_0 : 总体 X 的分布函数是 $F(x)$, 其中 $F(x)$ 是已知的, 这种情况是不多的. 我们经常遇到的所需检验的原假设是

$$H_0: \text{总体 } X \text{ 的分布函数是 } F(x; \theta_1, \dots, \theta_r), \quad (6.4)$$

其中 F 的形式已知, 而 $\theta = (\theta_1, \theta_2, \dots, \theta_r)$ 是未知参数, 它们在某一个范围取值. 在 $F(x; \theta_1, \theta_2, \dots, \theta_r)$ 中当参数 $\theta_1, \theta_2, \dots, \theta_r$ 取不同的值时, 就得到不同的分布, 因而 $F(x; \theta_1, \theta_2, \dots, \theta_r)$ 代表一族分布. (6.4) 中的 H_0 表示总体 X 的分布属于分

布族 $F(x; \theta_1, \theta_2, \dots, \theta_r)$. 采用类似(一)中的方法来定义检验统计量, 将在 H_0 下 X 可能取值的全体 Ω 分成 k ($k > r + 1$) 个互不相交的子集 $A_1, A_2, \dots, A_k$, 以 f_i ($i = 1, 2, \dots, k$) 记样本观察值 $x_1, x_2, \dots, x_n$ 落在 A_i 的个数, 则事件 $A_i = \{X \text{ 的值落在 } A_i \text{ 内}\}$ 的频率为 f_i/n . 另一方面, 当 H_0 为真时, 由 H_0 所假设的分布函数来计算 $P(A_i)$, 得到 $P(A_i) = p_i(\theta_1, \theta_2, \dots, \theta_r) = p_i(\theta) = p_i$. 此时, 需先利用样本求出未知参数的最大似然估计(在 H_0 下), 以估计值作为参数值, 求出 p_i 的估计值 $\hat{p}_i = \hat{P}(A_i)$, 在(6.3)式中以 $\hat{p}_i$ 代替 p_i , 取

$$\chi^2 = \sum_{i=1}^k \frac{f_i^2}{n\hat{p}_i} - n \quad (6.5)$$

作为检验假设 H_0 的统计量. 可以证明, 在某些条件下, 在 H_0 为真时近似地有

$$\chi^2 = \sum_{i=1}^k \frac{f_i^2}{n\hat{p}_i} - n \sim \chi^2(k - r - 1)$$

与在(一)中一样可得假设检验问题(6.4)的拒绝域为

$$\chi^2 \geq \chi_{\alpha}^2(k - r - 1), \quad (6.6)$$

α 为显著性水平. 以上就是用来检验分布族的 χ^2 拟合检验法.

例 3 在一实验中, 每隔一定时间观察一次由某种铀所放射的到达计数器上的 α 粒子数 X , 共观察了 100 次, 得结果如下表:

表 8-4 铀放射的到达计数器上的 α 粒子数的实验记录

i	0	1	2	3	4	5	6	7	8	9	10	11	≥ 12
f_i	1	5	16	17	26	11	9	9	2	1	2	1	0
A_i	A_0	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}

其中 f_i 是观察到有 i 个 α 粒子的次数. 从理论上考虑知 X 应服从泊松分布

$$P\{X = i\} = \frac{\lambda^i e^{-\lambda}}{i!}, \quad i = 0, 1, 2, \dots \quad (6.7)$$

问(6.7)式是否符合实际(取 $\alpha = 0.05$)? 即在显著性水平 0.05 下检验假设

H_0 : 总体 X 服从泊松分布

$$P\{X = i\} = \frac{\lambda^i e^{-\lambda}}{i!}, \quad i = 0, 1, 2, \dots$$

解 因在 H_0 中参数 λ 未具体给出, 所以先估计 λ . 由最大似然估计法得 $\hat{\lambda} = \bar{x} = 4.2$. 在 H_0 假设下, 即在 X 服从泊松分布的假设下, X 所有可能取的值为 $\Omega = \{0, 1, 2, \dots\}$, 将 Ω 分成如表 8-4 所示的两两不相交的子集 $A_0, A_1, \dots, A_{12}$. 则 $P\{X = i\}$ 有估计

$$\hat{p}_i = \hat{P}\{X = i\} = \frac{4.2^i e^{-4.2}}{i!}, \quad i = 0, 1, \dots$$

例如

$$\hat{p}_0 = \hat{P}\{X = 0\} = e^{-4.2} = 0.015,$$

$$\hat{p}_3 = \hat{P}\{X=3\} = \frac{4 \cdot 2^3 e^{-4.2}}{3!} = 0.185,$$

$$\hat{p}_{12} = \hat{P}\{X \geq 12\} = 1 - \sum_{i=0}^{11} \hat{p}_i = 0.002.$$

表 8-5 例 3 的 χ^2 拟合检验计算表

A_i	f_i	$\hat{p}_i$	$n\hat{p}_i$	$f_i^2/n\hat{p}_i$
A_0	1	0.015	1.5	4.615
A_1	5	0.063	6.3	
A_2	16	0.132	13.2	19.394
A_3	17	0.185	18.5	15.622
A_4	26	0.194	19.4	34.845
A_5	11	0.163	16.3	7.423
A_6	9	0.114	11.4	7.105
A_7	9	0.069	6.9	11.739
A_8	2	0.036	3.6	5.538
A_9	1	0.017	1.7	
A_{10}	2	0.007	0.7	
A_{11}	1	0.003	0.3	
A_{12}	0	0.002	0.2	

 $\Sigma = 106.281$

计算结果如表 8-5 所示,其中有些 $n\hat{p}_i < 5$ 的组予以适当合并,使得每组均有 $n\hat{p}_i \geq 5$,如表中第四列花括号所示.此处,并组后 $k=8$,但因在计算概率时,估计了一个参数 λ ,故 $r=1$, χ^2 的自由度为 $8-1-1=6$. $\chi_{0.05}^2(k-r-1) = \chi_{0.05}^2(6) = 12.592$,现在 $\chi^2 = 106.281 - 100 = 6.281 < 12.592$,故在显著性水平 0.05 下接受 H_0 .即认为样本来自泊松分布总体.也就是说认为理论上的结论是符合实际的. $\square$

注意 本题答案是“接受 H_0 ,认为总体 X 的分布属于泊松分布族,即认为 $X \sim \pi(\lambda)$ ”,亦即“认为必有某一个参数 λ_0 , $X \sim \pi(\lambda_0)$ ”,而不能将答案误写成“ X 服从以 $\lambda=4.2$ 为参数的泊松分布”.

例 4 自 1965 年 1 月 1 日至 1971 年 2 月 9 日共 2 231 天中,全世界记录到里氏震级 4 级和 4 级以上地震计 162 次,统计如下:

相继两次地震 间隔天数 x	0~4	5~9	10~14	15~19	20~24	25~29	30~34	35~39	≥ 40
出现的频数	50	31	26	17	10	8	6	6	8 ^①

① 这里 8 个数值是 40, 43, 44, 49, 58, 60, 81, 109.

试检验相继两次地震间隔的天数 X 服从指数分布 ($\alpha=0.05$).

解 按题意需检验假设

H_0 : X 的概率密度为

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta}, & x > 0, \\ 0, & x \leq 0. \end{cases} \quad (6.8)$$

在这里, H_0 中的参数 θ 未给出, 先由最大似然估计法求得 θ 的估计为 $\hat{\theta} = \bar{x} = \frac{223.1}{162} = 13.77$. 在 H_0 下, X 可能取值的全体 Ω 为区间 $[0, \infty)$. 将区间 $[0, \infty)$ 分为 $k=9$ 个互不重叠的小区间: $A_1 = [0, 4.5]$, $A_2 = (4.5, 9.5]$, $\dots$, $A_9 = (39.5, \infty)$, 如表 8-6 第二列所示. 若 H_0 为真, X 的分布函数的估计为

$$\hat{F}(x) = \begin{cases} 1 - e^{-x/13.77}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

由上式可得概率 $p_i = P(A_i)$ 的估计:

$$\hat{p}_i = \hat{P}(A_i) = \hat{P}\{a_i < X \leq a_{i+1}\} = \hat{F}(a_{i+1}) - \hat{F}(a_i).$$

例如

$$\hat{p}_2 = \hat{P}(A_2) = \hat{P}\{4.5 < X \leq 9.5\} = \hat{F}(9.5) - \hat{F}(4.5) = 0.2196,$$

而

$$\hat{p}_9 = \hat{P}(A_9) = 1 - \sum_{i=1}^8 \hat{F}(A_i) = 0.0568.$$

将计算结果列表如下.

表 8-6 例 4 的 χ^2 检验计算表

A_i	f_i	$\hat{p}_i$	$n\hat{p}_i$	$f_i^2/(n\hat{p}_i)$
$A_1: 0 \leq x \leq 4.5$	50	0.2788	45.1656	55.3519
$A_2: 4.5 < x \leq 9.5$	31	0.2196	35.5752	27.0132
$A_3: 9.5 < x \leq 14.5$	26	0.1527	24.7374	27.3270
$A_4: 14.5 < x \leq 19.5$	17	0.1062	17.2044	16.7980
$A_5: 19.5 < x \leq 24.5$	10	0.0739	11.9718	8.3530
$A_6: 24.5 < x \leq 29.5$	8	0.0514	8.3268	7.6860
$A_7: 29.5 < x \leq 34.5$	6	0.0358	5.7996	6.2073
$A_8: 34.5 < x \leq 39.5$	6	0.0248	4.0176	14.8269
$A_9: 39.5 < x < \infty$	8	0.0568	9.2016	

$\Sigma = 163.5633$

现在 $\chi^2 = 163.5633 - 162 = 1.5633$, 因为 $\chi_{0.05}^2(k-r-1) = \chi_{0.05}^2(8-1-1) = \chi_{0.05}^2(6) = 12.592 > 1.5633$, 故在显著性水平 0.05 下接受 H_0 , 认为 X 服从指

数分布. □

例 5 对于第六章 § 2 例 1 中的数据, 试检验它们是否来自正态总体 X (取显著性水平 $\alpha=0.1$).

解 需检验假设

$H_0: X$ 的概率密度为

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

因在 H_0 中未给出 μ, σ^2 的数值. 需先估计 μ, σ^2 . 由最大似然估计法得 μ, σ^2 的估计值为 $\hat{\mu}=143.8, \hat{\sigma}^2=(6.0)^2$. 我们将在 H_0 下 X 可能取值的区间 $(-\infty, \infty)$ 分为 7 个小区间, 并取事件 A_i 如表 8-7 中第一列所示. 若 H_0 为真 X 的概率密度的估计为

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi} \times 6} e^{-\frac{(x-143.8)^2}{2 \times 6^2}}, \quad -\infty < x < \infty.$$

按上式并查标准正态分布的分布函数表即可得概率 $P(A_i)$ 的估计. 例如

$$\begin{aligned} \hat{p}_2 &= \hat{P}(A_2) = \hat{P}\{129.5 < X \leq 134.5\} \\ &= \Phi\left(\frac{134.5 - 143.8}{6}\right) - \Phi\left(\frac{129.5 - 143.8}{6}\right) \\ &= \Phi(-1.55) - \Phi(-2.38) = 0.0519. \end{aligned}$$

将计算结果列表如下:

表 8-7 例 5 的 χ^2 检验计算表

A_i	f_i	$\hat{p}_i$	$n\hat{p}_i$	$f_i^2/n\hat{p}_i$
$A_1: x \leq 129.5$	1	0.0087	0.73	4.91
$A_2: 129.5 < x \leq 134.5$	4	0.0519	4.36	
$A_3: 134.5 < x \leq 139.5$	10	0.1752	14.72	6.79
$A_4: 139.5 < x \leq 144.5$	33	0.3120	26.21	41.55
$A_5: 144.5 < x \leq 149.5$	24	0.2811	23.61	24.40
$A_6: 149.5 < x \leq 154.5$	9	0.1336	11.22	10.02
$A_7: 154.5 < x < \infty$	3	0.0375	3.15	

$\Sigma = 87.67$

现在 $\chi^2 = 87.67 - 84 = 3.67$, 因为 $\chi_{0.1}^2(k-r-1) = \chi_{0.1}^2(5-2-1) = \chi_{0.1}^2(2) = 4.605 > 3.67$, 故在水平 0.1 下接受 H_0 , 即认为数据来自正态分布总体. □

(三) 偏度、峰度检验

根据第五章关于中心极限定理的论述知道, 正态分布随机变量是较广泛地

存在的,因此,当研究一连续型总体时,人们往往先考察它是否服从正态分布.上面介绍的 χ^2 拟合检验法虽然是检验总体分布的较一般的方法,但用它来检验总体的正态性时,犯第 II 类错误的概率往往较大.为此,统计学家们对检验正态总体的种种方法进行了比较,根据奥野忠一等人 20 世纪 70 年代进行的大量模拟计算的结果,认为正态性检验方法中,总的来说,以“偏度、峰度检验法”及“夏皮罗-威尔克法”较为有效.在这里我们仅介绍偏度、峰度检验法.

随机变量 X 的偏度和峰度指的是 X 的标准化变量 $[X-E(X)]/\sqrt{D(X)}$ 的三阶矩和四阶矩:

$$\nu_1 = E\left[\left(\frac{X-E(X)}{\sqrt{D(X)}}\right)^3\right] = \frac{E[(X-E(X))^3]}{(D(X))^{3/2}},$$

$$\nu_2 = E\left[\left(\frac{X-E(X)}{\sqrt{D(X)}}\right)^4\right] = \frac{E[(X-E(X))^4]}{(D(X))^2}.$$

当随机变量 X 服从正态分布时, $\nu_1=0$ 且 $\nu_2=3$.

设 $X_1, X_2, \dots, X_n$ 是来自总体 X 的样本,则 ν_1, ν_2 的矩估计量分别是

$$G_1 = B_3/B_2^{3/2}, \quad G_2 = B_4/B_2^2,$$

其中 B_k ($k=2,3,4$) 是样本 k 阶中心矩,并分别称 G_1, G_2 为样本偏度和样本峰度.

若总体 X 为正态变量,则可证当 n 充分大时,近似地有

$$G_1 \sim N\left(0, \frac{6(n-2)}{(n+1)(n+3)}\right), \quad (6.9)$$

$$G_2 \sim N\left(3 - \frac{6}{n+1}, \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}\right). \quad (6.10)$$

设 $X_1, X_2, \dots, X_n$ 是来自总体 X 的样本,现在来检验假设

$$H_0: X \text{ 为正态总体.}$$

记

$$\sigma_1 = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}, \quad \sigma_2 = \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}},$$

$\mu_2 = 3 - \frac{6}{n+1}$, $U_1 = G_1/\sigma_1$, $U_2 = (G_2 - \mu_2)/\sigma_2$. 当 H_0 为真且 n 充分大时,近似地有

$$U_1 \sim N(0, 1), \quad U_2 \sim N(0, 1).$$

由第六章 §3 知样本偏度 G_1 、样本峰度 G_2 分别依概率收敛于总体偏度 ν_1 和总体峰度 ν_2 . 因此当 H_0 为真且 n 充分大时,一般来说, G_1 与 $\nu_1=0$ 的偏离不应太大,而 G_2 与 $\nu_2=3$ 的偏离不应太大. 故从直观来看当 $|U_1|$ 的观察值 $|u_1|$ 或 $|U_2|$ 的观察值 $|u_2|$ 过大时就拒绝 H_0 . 取显著性水平为 α , H_0 的拒绝域为

$$|u_1| \geq k_1 \quad \text{或} \quad |u_2| \geq k_2, \quad (6.11)$$

其中 k_1, k_2 由以下两式确定:

$$P_{H_0}\{|U_1| \geq k_1\} = \frac{\alpha}{2}, \quad P_{H_0}\{|U_2| \geq k_2\} = \frac{\alpha}{2}.$$

这里记号 $P_{H_0}\{\cdot\}$ 表示当 H_0 为真时事件 $\{\cdot\}$ 的概率, 即有 $k_1 = z_{\alpha/4}, k_2 = z_{\alpha/4}$. 于是得拒绝域为

$$|u_1| \geq z_{\alpha/4} \quad \text{或} \quad |u_2| \geq z_{\alpha/4}. \quad (6.12)$$

下面来验证当 n 充分大时上述检验法近似地满足显著性水平为 α 的要求. 事实上当 n 充分大时有

$$\begin{aligned} P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} \\ &= P_{H_0}\{(|U_1| \geq z_{\alpha/4}) \cup (|U_2| \geq z_{\alpha/4})\} \\ &\leq P_{H_0}\{|U_1| \geq z_{\alpha/4}\} + P_{H_0}\{|U_2| \geq z_{\alpha/4}\} = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha. \end{aligned}$$

例 6 试用偏度、峰度检验法检验例 5 中的数据是否来自正态总体 (取 $\alpha = 0.1$).

解 现在来检验假设

H_0 : 数据来自正态总体.

$$\text{这里 } \alpha = 0.1, n = 84, \sigma_1 = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}} = 0.2579,$$

$$\sigma_2 = \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}} = 0.4892,$$

$\mu_2 = 3 - \frac{6}{n+1} = 2.9294$. 下面来计算样本中心矩 B_2, B_3, B_4 , 计算时可利用以下关系式:

$$B_2 = A_2 - A_1^2, \quad B_3 = A_3 - 3A_2A_1 + 2A_1^3,$$

$$B_4 = A_4 - 4A_3A_1 + 6A_2A_1^2 - 3A_1^4,$$

其中 $A_k = \frac{1}{n} \sum_{i=1}^n X_i^k$ ($k=1, 2, 3, 4$) 为 k 阶样本矩. 经计算得 A_k ($k=1, 2, 3, 4$),

B_k ($k=2, 3, 4$) 的观察值分别为

$$A_1 = 143.7738, \quad A_2 = 20706.13, \quad A_3 = 2987099,$$

$$A_4 = 4.316426 \times 10^8, \quad B_2 = 35.2246, \quad B_3 = -28.5, \quad B_4 = 3840.$$

样本偏度和样本峰度的观察值分别为

$$g_1 = -0.1363, \quad g_2 = 3.0948.$$

而 $z_{\alpha/4} = z_{0.025} = 1.96$. 由 (6.11) 式, 拒绝域为

$$|u_1| = |g_1/\sigma_1| \geq 1.96 \quad \text{或} \quad |u_2| = |g_2 - \mu_2|/\sigma_2 \geq 1.96.$$

现算得 $|u_1| = 0.5285 < 1.96, |u_2| = 0.3381 < 1.96$, 故接受 H_0 , 认为数据来自

正态分布的总体. □

上述检验法称为偏度、峰度检验法. 使用这一检验法时样本容量以大于 100 为宜.

§ 7 秩和检验

本节介绍一种有效的、且使用方便的检验方法——秩和检验法.

设有两个连续型总体, 它们的概率密度函数分别为 $f_1(x)$, $f_2(x)$, 均为未知, 但已知

$$f_1(x) = f_2(x-a), \quad a \text{ 为未知常数}, \quad (7.1)$$

即 f_1 与 f_2 至多只差一平移. 我们要检验下述各项假设

$$H_0: a=0, H_1: a<0. \quad (7.2)$$

$$H_0: a=0, H_1: a>0. \quad (7.3)$$

$$H_0: a=0, H_1: a \neq 0. \quad (7.4)$$

特别, 若两总体的均值存在, 分别记作 μ_1, μ_2 , 则由于 f_1, f_2 至多只差一平移, 故有

$$\mu_2 = \mu_1 - a.$$

此时, 上述各项假设分别等价于

$$H_0: \mu_1 = \mu_2, H_1: \mu_1 < \mu_2. \quad (7.2)'$$

$$H_0: \mu_1 = \mu_2, H_1: \mu_1 > \mu_2. \quad (7.3)'$$

$$H_0: \mu_1 = \mu_2, H_1: \mu_1 \neq \mu_2. \quad (7.4)'$$

现在来介绍威尔柯克斯(Frank Wilcoxon)提出的秩和检验法以检验上述假设. 为此, 先引入秩的概念.

秩 设 X 为一总体, 将一容量为 n 的样本观察值按自小到大的次序编号排列成

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)}, \quad (7.5)$$

称 $x_{(i)}$ 的足标 i 为 $x_{(i)}$ 的秩, $i=1, 2, \cdots, n$.

现设自 1, 2 两总体分别抽取容量为 n_1, n_2 的样本, 且设两样本独立. 这里总假定 $n_1 \leq n_2$. 我们将这 $n_1 + n_2$ 个观察值放在一起, 按自小到大的次序排列, 求出每个观察值的秩, 然后将属于第 1 个总体的样本观察值的秩相加, 其和记为 R_1 , 称为第 1 样本的秩和. 其余观察值的秩的总和记作 R_2 , 称为第 2 样本的秩和. 显然 R_1, R_2 是离散型的随机变量, 且有

$$R_1 + R_2 = \frac{1}{2}(n_1 + n_2)(n_1 + n_2 + 1). \quad (7.6)$$

所以, R_1, R_2 中的一个确定后另一个随之而定. 这样, 我们只要考虑统计量 R_1

即可.

现在来解决双边检验问题(7.4). 对此, 先作直观分析. 当 H_0 为真时, 即有 $f_1(x) = f_2(x)$, 这时两个独立样本实际上是来自同一个总体. 因而第 1 个样本中诸元素的秩应该随机地、分散地在自然数 $1 \sim n_1 + n_2$ 中取值, 一般来说不应过分集中取较小的或较大的值. 考虑到

$$\frac{1}{2}n_1(n_1 + 1) \leq R_1 \leq \frac{1}{2}n_1(n_1 + 2n_2 + 1),$$

即知当 H_0 为真时秩和 R_1 一般来说不应取太靠近上述不等式两端的值. 因而, 当 R_1 的观察值 r_1 过分大或过分小时, 我们都拒绝 H_0 .

据以上分析, 对于双边检验(7.4), 在给定显著性水平 α 下, H_0 的拒绝域为

$$r_1 \leq C_U\left(\frac{\alpha}{2}\right) \quad \text{或} \quad r_1 \geq C_L\left(\frac{\alpha}{2}\right),$$

其中临界点 $C_U\left(\frac{\alpha}{2}\right)$ 是满足 $P_{a=0}\left\{R_1 \leq C_U\left(\frac{\alpha}{2}\right)\right\} \leq \frac{\alpha}{2}$ 的最大整数, 而 $C_L\left(\frac{\alpha}{2}\right)$ 是满足 $P_{a=0}\left\{R_1 \geq C_L\left(\frac{\alpha}{2}\right)\right\} \leq \frac{\alpha}{2}$ 的最小整数. 而犯第 I 类错误的概率为

$$P_{a=0}\left\{R_1 \leq C_U\left(\frac{\alpha}{2}\right)\right\} + P_{a=0}\left\{R_1 \geq C_L\left(\frac{\alpha}{2}\right)\right\} \leq \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.$$

如果知道 R_1 的分布, 则临界点 $C_U\left(\frac{\alpha}{2}\right), C_L\left(\frac{\alpha}{2}\right)$ 是不难求得的. 下面以 $n_1 = 3, n_2 = 4$ 的情况为例来说明求临界点的方法.

当 $n_1 = 3, n_2 = 4$ 时, 第 1 个样本中各观察值的秩的不同取法共有 $\binom{3+4}{3} = 35$ 种. 现将这 35 种情况列表如下.

秩	R_1	秩	R_1	秩	R_1	秩	R_1	秩	R_1
123	6	136	10	167	14	247	13	356	14
124	7	137	11	234	9	256	13	357	15
125	8	145	10	235	10	257	14	367	16
126	9	146	11	236	11	267	15	456	15
127	10	147	12	237	12	345	12	457	16
134	8	156	12	245	11	346	13	467	17
135	9	157	13	246	12	347	14	567	18

由于这 35 种情况的出现是等可能的, 由上表容易求得 R_1 的分布律和分布函数如下:

R_1	6	7	8	9	10	11	12
$P\{R_1=r_1\}$	1/35	1/35	2/35	3/35	4/35	4/35	5/35
$P\{R_1 \leq r_1\}$	1/35	2/35	4/35	7/35	11/35	15/35	20/35
R_1	13	14	15	16	17	18	
$P\{R_1=r_1\}$	4/35	4/35	3/35	2/35	1/35	1/35	
$P\{R_1 \leq r_1\}$	24/35	28/35	31/35	33/35	34/35	1	

于是,对于不同的 α 值,容易写出检验问题(7.4)的临界点和拒绝域.例如,给定 $\alpha=0.2$.由上表知

$$P_{a=0}\{R_1 \leq 7\} = 2/35 < 0.1 = \frac{\alpha}{2},$$

$$P_{a=0}\{R_1 \geq 17\} = 2/35 < 0.1 = \frac{\alpha}{2}.$$

即有 $C_U(0.1)=7, C_L(0.1)=17$.故当 $n_1=3, n_2=4$,在显著性水平 0.2 下检验问题(7.4)的拒绝域为

$$r_1 \leq 7 \quad \text{或} \quad r_1 \geq 17.$$

此时,犯第 I 类错误的概率为

$$P_{a=0}\{R_1 \leq 7\} + P_{a=0}\{R_1 \geq 17\} = 2/35 + 2/35 = 0.114.$$

类似地可得左边检验(7.2)的拒绝域为(显著性水平为 α)

$$r_1 \leq C_U(\alpha),$$

此处,临界点 $C_U(\alpha)$ 是满足 $P_{a=0}\{R_1 \leq C_U(\alpha)\} \leq \alpha$ 的最大整数.而右边检验问题的拒绝域为(显著性水平为 α)

$$r_1 \geq C_L(\alpha),$$

此处,临界点 $C_L(\alpha)$ 是满足 $P_{a=0}\{R_1 \geq C_L(\alpha)\} \leq \alpha$ 的最小整数.

例如,若给定 $\alpha=0.1$,抽取的样本容量为 $n_1=3, n_2=4$,则由上表知检验问题(7.3)的拒绝域为

$$r_1 \geq 17.$$

此时犯第 I 类错误的概率为 $2/35 < 0.1$.

书末附表 9 中列出了 n_1 和 n_2 自 2 到 10 为止的 n_1, n_2 的各种组合的临界点,以及相应的犯第 I 类错误的概率.

例 1 为查明某种血清是否会抑制白血病,选取患白血病已到晚期的老鼠 9 只,其中有 5 只接受这种治疗,另 4 只则不做这种治疗.设两样本相互独立.从试验开始时计算,其存活时间(以月计)如下: