

概率论 与数理统计

龙永红 主编

高等教育出版社

内容简介

本书是教育部“ 高等教育面向 21 世纪教学内容和课程体系改革计划 ”的研究成果，是面向 21 世纪的高等院校经济学学科门类和管理学学科门类的数学基础课教材之一。

全书以经管类学生易于接受的方式介绍了概率论与数理统计的基本内容，重点介绍了概率论与数理统计的方法及其在经济、管理中的应用并附有 A、B 两级习题及参考答案，还附有常用统计分布表。

本书在内容安排上还考虑了经管类学生将来考研的需要，也适合于考研学生复习备考之用。

图书在版编目（CIP）数据

概率论与数理统计/ 龙永红主编. —北京：高等教育出版社，2000
(经济管理学科数学基础/ 范培华，胡显佑主编)
ISBN 7-04-009442-8

. 概... . 龙... . 概率论-应用-经济管理
数理统计-应用-经济管理 . F224.7

中国版本图书馆 CIP 数据核字（2000）第 86786 号

责任编辑 李 陶 封面设计 杨立新 责任绘图 尹 莉
版式设计 马静茹 责任校对 刘青田 责任印制

概率论与数理统计
龙永红 主编

出版发行 高等教育出版社
社 址 北京市东城区沙滩后街 55 号 邮政编码 100009
电 话 010-64054588 传 真 010-64014048
网 址 <http://www.hep.edu.cn>
<http://www.hep.com.cn>
经 销 新华书店北京发行所
印 刷

开 本	787× 960 1/16	版 次	年 月第 版
印 张	20.25	印 次	年 月第 次印刷
字 数	370 000	定 价	17.30 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

前 言

1996 年原国家教委开始组织实施“高等教育面向 21 世纪教学内容和课程体系改革计划”其中子项目经济学门类数学基础课和管理学门类数学基础课研究分别由中国人民大学和北京大学承担。考虑到这两大学科门类数学基础课程的共同点，教育部又将这两个子项目整合为“经济管理学类专业数学基础课程设置与教学内容改革研究”，集中力量合作研究，并成立了以魏权龄教授和范培华教授为项目主持人的课题组。两年多来，课题组对国内外高等院校同类专业数学基础课程的现状进行了调查研究，编写了教学大纲，组织了多次有关课程体系、课程内容的研讨会。其中，于 1997 年 7 月在长春召开的中国数量经济学会年会上，全国 40 余所院校的教师就经济管理类专业的数学基础课、数量经济分析课程的体系、课程设置、内容等进行了深入的讨论；1998 年 4 月，教育部在京召开了管理类专业面向 21 世纪教学内容和课程体系改革的研讨会上，初步确定了数学基础课应包括微积分、线性代数和概率统计三门课程，共 16 学分。其中，“微积分”8 学分，“线性代数”3 学分，“概率统计”5 学分。

在调查研究和充分讨论的基础上，课题组拟定了《经济管理学学科数学基础教学大纲》(草案)，并邀请北京地区部分高校就该大纲进行了讨论。

受教育部委托，北京大学光华管理学院和中国人民大学信息学院共同承担了编写经济管理学学科数学基础系列教材的任务。整套教材分为《微积分》、《线性代数》和《概率论与数理统计》3 个分册，由魏权龄教授任编写组顾问，范培华教授、胡显佑教授任主编。这套教材的《微积分》分册由朱来义教授主编，参加编写的有朱来义、吴岚、范培华和严守权；《线性代数》分册由卢刚副教授主编，参加编写的有卢刚、胡显佑、崔兆鸣；《概率论与数理统计》分册由龙永红副教授主编，参加编写的有龙永红、张贻兰、成世学、王明进。

根据高等教育面向 21 世纪教学内容和课程体系改革总体目标的要求，我们在编写这套教材时，主要考虑了下述问题：

1. 为适应我国在 21 世纪社会主义建设和经济发展的需要，培养“厚基础、宽口径、高素质”的人才，基础课，特别是数学基础课不应削弱，而应适当加强。

2. 考虑到目前绝大多数综合性大学、工科院校都设立了经济或管理学科的有关专业，但各校、各专业方向对数学基础的要求有一定的差异。这套教材应照顾到多数院校教学的实际情况，便于教师和学生使用。

3. 作为一门数学基础课的教材, 我们首先注意保持数学学科本身的科学性、系统性, 但在引入一些概念时尽可能采用学生易于接受的方式叙述, 对个别冗长, 繁琐的推理则略去, 而更突出有关理论、方法的应用和经济数学模型的介绍.

4. 作为经济管理学科各专业的数学基础教材, 我们注意了专业后继课程的需要, 并考虑学生继续深造的需要, 教材的各章均配备了 A, B 两组习题. 一般, 达到 A 组习题的水平, 就已经符合本课程的基本要求. B 组习题是为数学基础要求较高的专业或学生准备的. 各章中打有 “*” 号 (或小字排版) 的内容是为对数学基础要求较高的院校或专业编写的, 可以作为选学内容或学生自学用.

1999 年 12 月, 由教育部高教司聘请了有关专家对教材的初稿进行了审定. 参加审稿会的有: 北京航空航天大学李心灿教授、清华大学胡金德教授、南开大学周概容教授、(以下以姓氏笔划为序) 湖南财经学院苏醒教授、北方交通大学季文铎教授、中央财政金融大学单立波教授、华侨大学龚德恩教授、中南财金大学彭勇行教授. 他们对教材初稿提出了许多中肯的建议和具体的修改意见, 这对于完善教材是非常有益的, 在此向参加审定会的各位教授表示诚挚的谢意.

在各次研讨会上, 全国各高校的许多同行都对这一项目和教材提出了极有价值的建议. 在此向有关院校的老师表示衷心感谢. 在教材编写过程中, 我们得到了教育部高教司的大力支持, 得到高教出版社有关部门的协助, 在此一并致谢.

范培华 胡显佑

2000 年 3 月

目 录

第一章	随机事件与概率	1
§ 1.1	随机事件	1
§ 1.2	随机事件的概率	8
§ 1.3	古典概型与几何概型	12
§ 1.4	条件概率	17
§ 1.5	事件的独立性	24
习题一	(A)	29
习题一	(B)	32
第二章	随机变量的分布与数字特征	33
§ 2.1	随机变量及其分布	33
§ 2.2	随机变量的数学特征	41
§ 2.3	常用的离散型分布	50
§ 2.4	常用的连续型分布	56
§ 2.5	随机变量函数的分布	62
习题二	(A)	66
习题二	(B)	69
第三章	随机向量	70
§ 3.1	随机向量的分布	70
§ 3.2	条件分布与随机变量的独立性	78
§ 3.3	随机向量的函数的分布与数学期望	86
§ 3.4	随机向量的数字特征	94
§ 3.5	大数定律与中心极限定理	103
习题三	(A)	109
习题三	(B)	113
第四章	数理统计的基础知识	115
§ 4.1	总体与样本	115
§ 4.2	统计量	119
§ 4.3	常用的统计分布	121
§ 4.4	抽样分布	131
习题四	(A)	136
习题四	(B)	138
第五章	参数估计	139

2	目 录	
§ 5.1	点估计概述	139
§ 5.2	极大似然法	143
§ 5.3	矩法	148
§ 5.4	置信区间	150
§ 5.5	正态总体参数的置信区间	156
习题五 (A)	163
习题五 (B)	165
第六章	假设检验	166
§ 6.1	假设检验概述	166
§ 6.2	单正态总体的参数假设检验	171
§ 6.3	双正态总体的参数假设检验	178
§ 6.4	关于一般总体数学期望的假设检验	190
* § 6.5	拟合优度 χ^2 检验法	194
习题六 (A)	205
习题六 (B)	207
第七章	方差分析	210
§ 7.1	问题的提出	210
§ 7.2	单因素方差分析	212
§ 7.3	双因素方差分析	218
习题七 (A)	228
习题七 (B)	229
第八章	回归分析	230
§ 8.1	一元线性回归模型及其参数估计	230
§ 8.2	一元线性回归模型的检验	237
§ 8.3	一元线性回归的预测与控制	245
§ 8.4	一元非线性问题的线性化	249
§ 8.5	多元线性回归分析	254
习题八 (A)	260
习题八 (B)	262
第九章	主成分分析与典型相关分析	263
§ 9.1	主成分分析	263
§ 9.2	典型相关分析	272
习题九	277
习题参考答案	279
常用统计分布表	295
附表 1	泊松分布概率值表	295
附表 2	标准正态分布函数值表	298
附表 3	χ^2 分布上侧分位数表	300

附表 4 F 分布上侧分位数表	302
附表 5 t 分布上侧分位数表	312
主要参考文献	313

第 1 章

随机事件与概率

概率论是研究随机现象的规律性的数学学科. 为了对随机现象的有关问题作出明确的数学阐述, 像其他数学学科一样, 概率论具有自己的严格的概念体系和严密的逻辑结构. 本章重点介绍概率论的两个最基本的概念: 随机事件及其概率, 主要包括: 随机事件和随机事件的概率的定义、古典概型与几何概型、条件概率、乘法公式、全概率公式与贝叶斯公式、以及事件的独立性等. 这些内容是进一步学习概率论的基础.

§ 1.1 随机事件

一、随机现象

在自然界和人类社会生活中普遍存在着两类现象, 一类是在一定条件下必然出现的现象, 称为确定性现象. 例如, 一物体从高度为 h (米) 处垂直下落, 则必然在 $\frac{2h}{g}$ 秒后落到地面, 其中 $g = 9.8$ (米/秒²) 为重力加速度. 我们已学过的数学学科均以确定性现象为研究对象, 并为研究其量的规律性提供工具和方法. 另一类则是我们事先无法准确预知其结果的现象, 称为随机现象. 例如, 投掷一枚硬币, 我们不能事先预知将出现正面还是反面. 在实际中, 我们经常要面对和处理随机现象, 比如, 明天是否会下雨? 某种股票明天价格是多少? 电视机价格是否会在近期下调? 这些问题往往事先均不能得到明确的答案, 然而它们却往往与我们的切身利益密切相关. 概率论将以随机现象为研究对象.

二、随机现象的统计规律性

由于随机现象的结果事先不能预知, 初看起来, 随机现象毫无规律可言. 然而人们发现同一随机现象在大量重复出现时, 其每种可能的结果出现的频率却具有稳定性, 从而表明随机现象也有其固有的量的规律性. 人们把随机现象在大量重复出现时所表现出来的量的规律性称为随机现象的统计规律性.

为了对随机现象的统计规律性进行研究,人们往往要对随机现象进行观察,我们把对随机现象的观察称为随机试验,并简称为试验.例如,某射手对固定目标进行射击;观察某地区夏季暴雨次数;观察某电话交换台每日收到的呼叫次数等均为随机试验.一般地,一个随机试验要求满足下列特点:

- (1) 可重复性: 试验原则上可在相同条件下重复进行;
- (2) 可观察性: 试验结果是可观察的,所有可能的结果是明确的;
- (3) 随机性: 每次试验将要出现的结果是不确定的,事先无法准确预知.

历史上,研究随机现象统计规律性的最著名的试验是投掷硬币的试验.我们知道,投掷一枚均匀硬币时,事先无法准确预知将出现正面还是反面.但是,当人们重复投掷上千次时,却发现出现正面和反面的次数大致相等,即各自占总试验次数的比例(即频率)大致等于 0.5,而且随着试验次数的增加,这一比例会更加稳定地靠近 0.5,表 1.1 列出了历史上一些试验的记录.

表 1.1 历史上投掷硬币试验的记录

试验者	投掷次数 (n)	正面次数 (r _n)	正面频率 $\frac{r_n}{n}$
De Morgan	2 048	1 061	0.5181
Buffon	4 040	2 048	0.5069
Pearson K	12 000	6 019	0.5016
Pearson K	24 000	12 012	0.5005

三、样本空间

正如前面指出的,一个随机试验将要出现的结果是不确定的,但其所有可能结果是明确的.我们把随机试验的每一个可能结果称为一个样本点,因而一个随机试验的所有样本点也是明确的,它们的全体,称为样本空间,通常用 Ω 表示. Ω 中的点,即样本点,用 ω 表示.

例 1.1 在投掷一枚硬币观察其出现正面还是反面的试验中,有两个样本点: 正面、反面. 样本空间为

$\Omega = \{\text{正面}, \text{反面}\}$

记 $\omega_1 = \text{正面}$, $\omega_2 = \text{反面}$, 则样本空间可表示为:

$\Omega = \{\omega_1, \omega_2\}$

例 1.2 在投掷一枚骰子,观察其出现的点数的实验中,有 6 个样本点: 1 点, 2 点, ...6 点. 样本空间为

$\Omega = \{1 \text{ 点}, 2 \text{ 点}, \dots, 6 \text{ 点}\}$

或干脆将样本点分别简记为: 1, 2, ..., 6, 相应地, 样本空间记为

$\Omega = \{1, 2, \dots, 6\}$

例 1.3 观察某电话交换台在一天内收到的呼叫次数, 其样本点有可数无

穷多个: i 次, $i = 0, 1, 2, \dots$, 样本空间为:

$$\Omega = \{0 \text{ 次}, 1 \text{ 次}, 2 \text{ 次}, \dots\}$$

或简记为 $\Omega = \{0, 1, 2, \dots\}$.

例 1.4 观察一个新灯泡的寿命, 其样本点也有无穷多个 (且不可数!): t 小时, $0 \leq t < +\infty$, 样本空间为

$$\Omega = \{t \text{ 小时} \mid 0 \leq t < +\infty\},$$

或简记为

$$\Omega = \{t \mid 0 \leq t < +\infty\} = [0, +\infty).$$

四、随机事件

在随机试验中, 人们除了关心试验的结果本身外, 往往还关心试验的结果是否具备某一指定的可观察的特征, 概率论中将这一可观察的特征称为一个事件. 例如, 投掷一枚骰子, 我们也许会关心出现的点数是否为偶数, “点数为偶数”就是一个事件. 同样, “点数小于 7”也是一个事件. 但这两个事件有着根本的区别, 前者在随机试验中可能发生也可能不发生, 这样的事件称为随机事件. 后者在试验中是必然发生的, 这样的事件称为必然事件. 与必然事件完全对立的是, 在试验中一定不发生的事件, 称为不可能事件. 比如在上述试验中“点数不小于 7”是不可能事件. 虽然必然事件与不可能事件是完全对立的, 但它们有一个共同的特点, 那就是在试验之前我们能够准确预知其是否发生, 因而均不是随机事件, 通常称之为确定性事件. 概率论研究的是随机事件, 但为方便起见常常将必然事件和不可能事件视为随机事件的极端情形, 并将随机事件简称为事件, 通常记作 A, B, \dots 等.

例 1.5 在投掷一枚骰子的试验中, 分别记

“点数为 6”为 A

“点数小于 5”为 B

“点数小于 5 的偶数”为 C

则 A, B, C 均为事件, 但事件 A 的结构最为简单, 它对应于一个惟一的可能结果, 即样本点, 这样的事件称为基本事件. 在本例中, 共有 6 个基本事件 (对应于 6 个样本点): “点数为 1”“点数为 2”..., “点数为 6”. 基本事件的称谓缘于相对其他事件而言, 它们是最基本的, 其他事件均可由它们复合而成, 而它们自身又不能再分解成其他事件. 事件 B 和 C 均不是基本事件, 它们分别可以由一些基本事件复合而成. 比如事件 C 可由“点数为 2”和“点数为 4”两个基本事件复合而成.

五、事件的集合表示

根据定义, 样本空间 Ω 是随机试验的所有可能结果——即样本点 ω 的全体, 因而样本空间实际上是所有样本点构成的集合, 相应的每一样本点是该集合中的元素. 而一个事件是由具有该事件所要求的特征的那些可能结果所构成, 所以一个事件对应于 Ω 中具有相应特征的样本点 (元素) 构成的集合, 它是 Ω 的一个子集, 于是任何一个事件, 我们可以用 Ω 的某一子集来表示, 通常用符号 A, B, \dots 等来记. 某事件发生, 就是属于该集合的某一样本点在试验中出现. 如果记 ω 为试验中出现的样本点, 那么事件 A 当且仅当 $\omega \in A$ 时发生.

例 1.6 在例 1.5 中, 样本空间为 $\Omega = \{1, 2, 3, 4, 5, 6\}$, 事件 B 和 C 则可分别表示为:

$$B = \{1, 2, 3, 4\}$$

$$C = \{2, 4\}$$

由于样本空间 Ω 包含所有可能结果, 试验结果必是其中之一, 所以样本空间作为一个事件是必然发生的, 即为必然事件, 今后用 Ω 表示必然事件. 空集 \emptyset 作为 Ω 的子集不含有任何样本点, 不管试验的结果是什么, \emptyset 作为一个事件总不会发生, 因而是不可能事件. 今后用 \emptyset 来表示不可能事件.

六、事件间的关系与运算

在一个随机试验中, 一般有很多随机事件, 为了通过对简单事件的研究来掌握复杂事件, 我们需要研究事件之间的关系和事件之间的一些运算. 前面引进的事件的集合表示, 为这一任务提供极大的便利.

1. 事件的包含

如果事件 A 发生必然导致 B 发生, 即属于 A 的每一个样本点一定也属于 B , 则称事件 B 包含事件 A , 或称事件 A 包含于事件 B , 或称 A 是 B 的子事件. 记作

$$B \supset A \text{ 或 } A \subset B$$

显然, 事件 $A \subset B$ 的含义与集合论中的含义是一致的. 对任意事件 A , 易知

$$A \subset A$$

2. 事件的相等

如果事件 A 包含事件 B , 事件 B 也包含事件 A , 则称事件 A 与 B 相等 (或等价), 记作 $A = B$. 易见, 相等的两个事件总是同时发生或同时不发生. 更直接的表述是, $A = B$ 是指 A 与 B 所含的样本点完全相同, 这等同于集合的相等.

3. 事件的并 (或和)

“事件 A 与 B 至少有一个发生”这一事件称作事件 A 与 B 的并 (或和), 记

作 $A \cup B$, 或 $A + B$. 显然, 事件 $A \cup B$ 是由 A 和 B 的样本点共同构成的事件, 这与集合的并集的含义是一致的.

例 1.7 在投掷一枚骰子的试验中, 记

$A = \text{“ 点数为奇数 ”}$

$B = \text{“ 点数小于 5 ”}$

则 $A \cap B = \{1, 2, 3, 4, 5\}$.

4. 事件的交 (或积)

“事件 A 和 B 都发生”这一事件称为事件 A 与 B 的交 (或积), 记作 $A \cap B$ (或 AB). 显然, $A \cap B$ 实际上是由 A 和 B 的公共样本点所构成, 这与集合的交的含义一致.

在例 1.7 中, 事件 A 与 B 的交为:

$$A \cap B = \{1, 3\}.$$

5. 事件的差

“事件 A 发生而 B 不发生”这一事件称为事件 A 与 B 的差, 记作 $A - B$.

在例 1.7 中, 事件 A 与 B 的差为:

$$A - B = \{5\}.$$

6. 互不相容事件

若事件 A 与 B 不可能同时发生, 也就是说, AB 是不可能事件, 即 $AB = \emptyset$, 则称事件 A 与 B 是互不相容事件.

比如, 在投掷一枚骰子的试验中, “点数小于 3”和“点数大于 4”这两个事件是互不相容事件.

7. 对立事件

“事件 A 不发生”, 这一事件称为事件 A 的对立事件, 记作 \bar{A} , 易见, $\bar{\bar{A}} = A$, 且 $\bar{A} = A^c$. 根据定义, 在一次试验中, 如果 A 发生, 则 \bar{A} 一定不发生, 如果 A 不发生, 则 \bar{A} 一定发生, 也就是说 A 与 \bar{A} 一定也只能发生其中之一, 因而有

$$A \cap \bar{A} = \emptyset, \quad A \cup \bar{A} = \Omega$$

例如, 在投掷一枚骰子的试验中记 A 为事件“点数为偶数”, 则 \bar{A} 为事件“点数为奇数”.

8. 有限个或可数个事件的并与交

设有 n 个事件 A_1, A_2, \dots, A_n , 则称“ A_1, A_2, \dots, A_n 至少有一个发生”这一事件为事件 A_1, A_2, \dots, A_n 的并, 记作 $A_1 \cup A_2 \cup \dots \cup A_n$, 或 $\bigcup_{i=1}^n A_i$. 称“ A_1, A_2, \dots, A_n 都发生”这一事件为事件 A_1, A_2, \dots, A_n 的交, 记作 $A_1 A_2 \dots A_n$, 或 $\bigcap_{i=1}^n A_i$.

设有可数个事件 $A_1, A_2, \dots, A_n, \dots$, 则称 “ $A_1, A_2, \dots, A_n, \dots$, 至少有一个发生” 这一事件为事件 $A_1, A_2, \dots, A_n, \dots$ 的并, 记作 $\bigcup_{i=1}^{\infty} A_i$, 称 “ $A_1, A_2, \dots, A_n, \dots$ 都发生” 这一事件为事件 $A_1, A_2, \dots, A_n, \dots$ 的交, 记作 $\bigcap_{i=1}^{\infty} A_i$.

9. 完备事件组

设 $A_1, A_2, \dots, A_n, \dots$ 是有限或可数个事件, 如果其满足:

$$(1) A_i A_j = \emptyset, \quad i \neq j, \quad i, j = 1, 2, \dots$$

$$(2) \bigcup_{i=1}^{\infty} A_i = \Omega$$

则称 $A_1, A_2, \dots, A_n, \dots$ 是一个完备事件组. 显然, A 与 \bar{A} 构成一个完备事件组.

10. 事件的关系与运算的文氏图

上述关于事件的各种关系与运算可直观地用图形(文氏图)来表示(见图 1.1)

例 1.8 考察某一位同学在一次数学考试中的成绩, 分别用 A, B, C, D, P, F 表示下列各事件(括号中表示成绩所处的范围):

A —— 优秀 ($[90, 100]$), D —— 及格 ($[60, 70]$),

B —— 良好 ($[80, 90]$), P —— 通过 ($[60, 100]$),

C —— 中等 ($[70, 80]$), F —— 未通过 ($[0, 60]$),

则 A, B, C, D, F 是两两不相容事件; P 与 F 是互为对立的事件, 即有 $P = \bar{F}$; A, B, C, D 均为 P 的子事件, 且有 $P = A \cup B \cup C \cup D$.

例 1.9 甲、乙、丙三人各射一次靶, 记 A —— “甲中靶”, B —— “乙中靶”, C —— “丙中靶”, 则可用上述三个事件的运算来分别表示下列各事件:

(1) “甲未中靶” —— \bar{A} ;

(2) “甲中靶而乙未中靶” —— $A\bar{B}$;

(3) “三人中只有丙未中靶” —— $\bar{A}\bar{B}C$;

(4) “三人中恰好有一人中靶” —— $A\bar{B}\bar{C} + \bar{A}B\bar{C} + \bar{A}\bar{B}C$;

(5) “三人中至少有一人中靶” —— $A + B + C$;

(6) “三人中至少有一人未中靶” —— $\bar{A} + \bar{B} + \bar{C}$;

(7) “三人中恰有两人中靶” —— $AB\bar{C} + A\bar{B}C + \bar{A}BC$;

(8) “三人中至少两人中靶” —— $AB + AC + BC$;

(9) “三人均未中靶” —— $\bar{A}\bar{B}\bar{C}$;

(10) “三人中至多一人中靶” —— $\bar{A}\bar{B}\bar{C} + \bar{A}\bar{B}C + \bar{A}B\bar{C} + A\bar{B}\bar{C}$;

(11) “三人中至多两人中靶” —— $\bar{A}\bar{B}\bar{C} + \bar{A}\bar{B}C + \bar{A}B\bar{C} + A\bar{B}\bar{C}$;

用其他事件的运算来表示一个事件, 方法往往不惟一(比如例 1.9 中的 (6) 和 (11) 实际上是同一事件) 读者应学会用不同方法表达同一事件, 在解

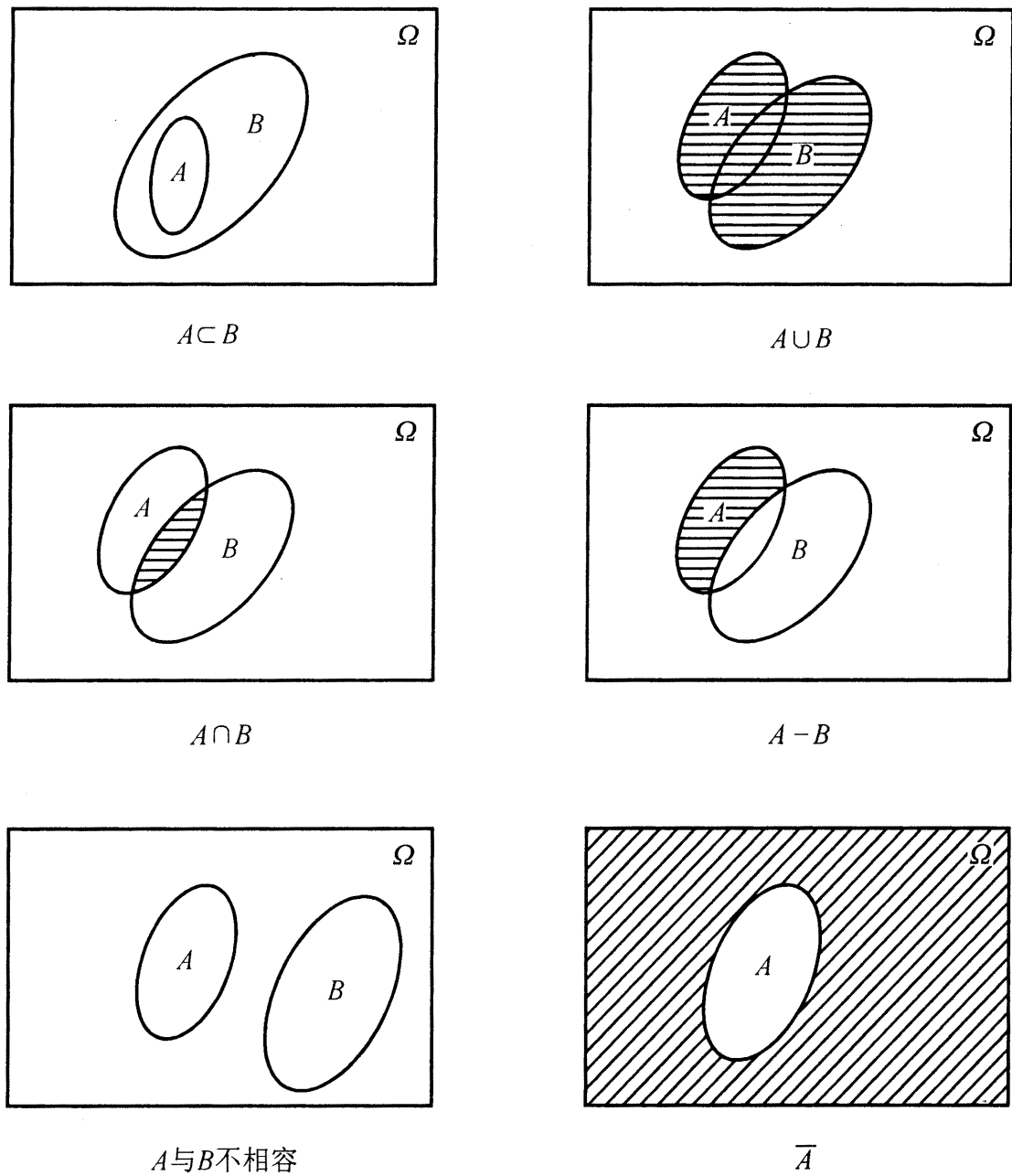


图 1.1 事件的关系与运算的文氏图

决具体问题时，特别是进行概率计算时，往往要根据需要选择其中的一种方法。例 1.9 中许多事件不止一种表达方法，读者可对此进行讨论。

七、随机事件的运算律

我们学过集合的运算律，事件也有相应的运算律，归纳于下：

1. 关于求和运算

- (1) $A \cup B = B \cup A$ (交换律)
- (2) $(A \cup B) \cup C = A \cup (B \cup C)$ $A \cup B \cup C$ (结合律)

2. 关于求交运算

- (1) $A \cap B = B \cap A$ (交换律)
- (2) $(A \cap B) \cap C = A \cap (B \cap C)$ $A \cap B \cap C$ (结合律)

3. 关于求和与求交运算的混合

$$(1) A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \quad (\text{第一分配律})$$

$$(2) A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad (\text{第二分配律})$$

4. 关于求对立事件的运算

$$(\bar{\bar{A}}) = A \quad (\text{自反律})$$

5. 关于和及交事件的对立事件

$$(1) \overline{A \cap B} = \bar{A} \cup \bar{B} \quad (\text{第一对偶律})$$

$$(2) \overline{A \cup B} = \bar{A} \cap \bar{B} \quad (\text{第二对偶律})$$

上述各运算律可以推广到有限个和可数个事件的情形, 读者可通过复习集合论的知识自行给出.

§ 1.2 随机事件的概率

一、概率和频率解释

我们知道, 在一次随机试验中, 一个随机事件是否会发生, 事先不能确定. 但是, 我们可以问, 在一次试验中, 事件 A 发生的可能性有多大? 比如, 投掷一枚均匀硬币, 我们不能肯定是否会出现正面. 但由于硬币是均匀的, 我们有理由认为, 出现正面和出现反面的可能性相同, 均为 $\frac{1}{2}$. 但可能性大小究竟意味着什么呢? 简单地说, 它反映了一次试验中事件 A 出现的机会. 然而这种机会在一次试验的实际结果中无法体现. 可以想像, 如果 A 发生的机会越大, 那么在大量重复试验它将出现得越频繁, 也就是说出现的频率会越大. 在投掷一枚均匀硬币的试验中, 如果果真出现正面和反面的机会均等, 即可能性均为 $\frac{1}{2}$, 那么在大量重复试验中, 出现正面和反面的频率会接近. 事实正是如此, 正如前一节指出的, 大量重复投掷一枚均匀硬币, 出现正面和反面的频率会接近一个稳定值 $\frac{1}{2}$. 可见频率的稳定值与事件发生的可能性大小存在内在必然的联系. 一方面频率的稳定性说明事件发生的可能性大小确实是一种客观存在, 另一方面, 频率的稳定值对事件发生的可能性大小提供了经验解释. 为此, 我们引入下列定义:

定义 1.1 随机事件 A 发生的可能性大小的度量(数值), 称为事件 A 发生的概率, 记作 $P(A)$.

正如前面指出的, 一个事件 A 发生的可能性的的大小——概率, 在经验上表现为大量重复试验中事件 A 发生的频率的稳定值. 因而频率的稳定值为概率的

含义提供了一种经验上的直观. 但频率的稳定值本身并不是概率的本质, 不能作为概率的定义. 一个事件的概率是由事件本身特征所决定的客观存在, 就好比一根木棒有它的长度一样. 频率的稳定值是概率的外在的必然表现, 当进行大量重复试验时, 频率会接近稳定值, 因而, 频率可用来作为概率的估计, 就好比是测定概率的“尺子”, 随着试验次数的增加, 测定的精度会越来越高.

二、从频率的性质看概率的性质

记一个事件 A 在 n 次重复试验中, 发生的次数为 $r_n(A)$, 则其发生的频率 $f_n(A)$ 为

$$f_n(A) = \frac{r_n(A)}{n}$$

设与试验有关的全体事件的集合为 F , 通常称 F 为事件域. 随着 A 取遍 F 中的任意事件, $f_n(A)$ 便成为定义在 F 上关于 A 的函数. 容易证明, 作为一个函数, $f_n(A)$ 满足下列性质:

- (1) $f_n(\Omega) = 1$;
- (2) 对任意事件 A , 有 $f_n(A) \geq 0$;
- (3) 对任意两两不相容的事件 $A_1, A_2, \dots, A_n, \dots$, 有

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

上述各性质的证明是简单的, 我们略去. 但值得指出的是, $f_n(A)$ 还满足许多其他性质, 比如, 比较显然的性质有: $f_n(\emptyset) = 0, f_n(A) \leq 1$. 然而这些性质均可由上述三条性质导出, 所以上述三条性质是反映频率特征的核心性质.

同频率一样, 记事件 A 发生的概率为 $P(A)$, 随着 A 取遍任意事件, $P(A)$ 则可视作定义在全体事件构成的集合, 即事件域 F 上的一个函数. 根据概率的频率解释, 概率可视作频率的稳定值, 从而应具有频率的相应性质, 即

- (1) $P(\Omega) = 1$;
- (2) 对任意事件 A , 有 $P(A) \geq 0$;
- (3) 对任意可数个两两不相容的事件 $A_1, A_2, \dots, A_n, \dots$, 有

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

三、概率的公理化定义

任何一个数学概念都是对现实世界的抽象, 这种抽象使得其具有广泛的适应性, 并成为进一步数学推理的基础. 前面指出, 概率的频率解释为概率提供了经验基础, 但不能作为一个严格的数学定义, 它没能抓住“概率”这一概念

的抽象本质. 如果人们对概率的认识只停留在这一简单的直观上, 那么人们对概率论的研究便只能停留在对一些肤浅的问题的零散研究上, 就会使得概率论的研究和应用带有很大的局限性. 从人们研究概率论有关问题开始算起, 经过近三个世纪的漫长探索历程, 伴随着数学的公理化潮流, 人们才真正完整地解决概率的严格数学定义. 其功劳归于前苏联著名的数学家柯尔莫哥洛夫, 他在 1933 年发表的《概率论的基本概念》一书中系统地表述了现在已被广泛接受的概率的公理化体系, 第一次将概率论建立在严密的逻辑基础上. 概率论从此确立了它作为一门严格的数学分支的地位. 这一突破性的进展, 是一个里程碑, 它将概率论推向一个全新的发展阶段.

我们知道, 一个事件 A 的概率实际上是赋予事件 A 的一个实数值, 记作 $P(A)$, 那么当 A 在事件域 F 中变化时 $P(A)$ 便成为事件域 F 上的一个函数, 记作 $P(\cdot)$. 概率的公理化定义并不考虑每一个事件 A 对应的概率 $P(A)$ 是怎么确定的, 值为多大 (这依赖于每一个具体的实际问题的结构), 而要求作为一个整体, 函数 $P(\cdot)$ 应满足一些必要的条件——公理, 这些公理是从概率的现实直观中抽象出来的.

定义 1.2 设 Ω 是一个样本空间, 定义在 Ω 的事件域 F 上的一个实值函数 $P(\cdot)$ 称为 Ω 上的一个概率测度, 如果它满足下列三条公理:

公理 1 $P(\Omega) = 1$;

公理 2 对任意事件 A , 有 $P(A) \geq 0$;

公理 3 对任意可数个两两不相容的事件 $A_1, A_2, \dots, A_n, \dots$, 有

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

其中, 对任意给定的具体事件 A , 函数值 $P(A)$ 称为事件 A 的概率. 此外, 一个配有概率测度 $P(\cdot)$ 的样本空间 Ω 称为一个概率空间, 记作 (Ω, F, P) .

概率测度还满足其他许多与直观相符的性质, 这些性质均可由上述定义中的三个公理推导出来.

四、概率测度的其他性质

由概率的公理化定义中的三条公理出发, 可以推导出概率测度的许多其他性质, 这些性质有助于我们对概率概念的进一步了解, 同时, 它们也是概率计算的重要基础.

性质 1 $P(\emptyset) = 0$.

严格地讲, 概率空间还应指明相应的事件域 F , 因而概率空间一般表达为 (Ω, F, P) , 在本书中, 由于很少对事件域本身进行研究, 因而省掉事件域 F .

证明 取 $A_i = \{ \omega \mid \omega \text{ 在第 } i \text{ 次试验中失败} \}$, $i = 1, 2, \dots$, 显然这是一列两两不相容的事件, 且 $\bigcup_{i=1}^{\infty} A_i = \Omega$, 由公理 3 知

$$P(\Omega) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{\infty} P(\omega_i)$$

由于 $P(\Omega)$ 为实数, 故必有 $P(\omega_i) = 0$.

性质 2 (有限可加性) 若 A_1, A_2, \dots, A_n 是两两不相容的, 则有

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

上述性质只须取 $A_i = \{\omega_i\}$, $i = n+1, n+2, \dots$, 即可由公理 3 及性质 1 导出. 下面各性质的证明均留作练习.

性质 3 $P(\bar{A}) = 1 - P(A)$

性质 4 $P(A - B) = P(A) - P(AB)$

特别地 若 $A \subset B$, 则

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

(2) $P(A) \leq P(B)$

性质 5 $0 \leq P(A) \leq 1$

性质 6 $P(A \cup B) = P(A) + P(B) - P(AB)$

性质 6 可以推广到任意有限个事件的并的情形, 这里我们给出了三个事件的情形, 更一般的情形读者可以自行讨论.

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

例 1.10 观察某地区未来 5 天的天气情况, 记 A_i 为事件: “有 i 天不下雨”, 已知 $P(A_i) = iP(A_0)$, $i = 1, 2, 3, 4, 5$. 求下列各事件的概率:

(1) 5 天均下雨;

(2) 至少一天不下雨;

(3) 至多三天不下雨.

解 显然 A_0, A_1, \dots, A_5 是两两不相容事件且 $\bigcup_{i=0}^5 A_i = \Omega$, 从而

$$1 = P(\Omega) = P\left(\bigcup_{i=0}^5 A_i\right) = \sum_{i=0}^5 P(A_i) = P(A_0) + \sum_{i=1}^5 iP(A_0) = 16P(A_0),$$

于是可求得

$$P(A_0) = \frac{1}{16}, \quad P(A_i) = \frac{i}{16}, \quad (i = 1, 2, 3, 4, 5)$$

记 (1), (2), (3) 中三个事件分别为 A, B, C , 则

$$(1) P(A) = P(A_0) = \frac{1}{16},$$

$$(2) P(B) = P\left(\bigcup_{i=1}^5 A_i\right) = 1 - P(A_0) = \frac{15}{16},$$

$$(3) P(C) = P\left(\bigcup_{i=0}^3 A_i\right) = P(A_0) + P(A_1) + P(A_2) + P(A_3) = \frac{7}{16}.$$

例 1.11 已知 $P(A) = 0.5, P(AB) = 0.2, P(B) = 0.4$,

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

解 (1) 因为 $AB + A\overline{B} = B$, 且 AB 与 $A\overline{B}$ 是不相容的, 故有

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

于是

$$P(A\overline{B}) = P(B) - P(AB) = 0.4 - 0.2 = 0.2$$

$$(2) P(A) = 1 - P(\overline{A}) = 1 - 0.5 = 0.5,$$

$$P(A - B) = P(A) - P(AB) = 0.5 - 0.2 = 0.3,$$

$$(3) P(A \cup B) = P(A) + P(B) - P(AB) = 0.5 + 0.4 - 0.2 = 0.7,$$

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

§ 1.3 古典概型与几何概型

一、古典概型

概率的公理化定义给出了概率的严格的数学定义, 人们在解决具体问题时, 希望明确相应的概率测度是什么, 一般而言, 这是很困难的. 这一节里, 我们介绍两类特殊情形. 首先介绍古典概型, 它是一类最简单的概率模型. 因为其简单, 曾经是概率论发展早期的主要研究对象, 也正因为此, 而今被冠以“古典概型”.

古典概型是指满足下面两个假设条件的概率模型:

- (1) 随机试验只有有限个可能结果;
- (2) 每一个可能结果发生的可能性相同.

这两个条件在数学上可表述为:

- (1) 样本空间有限, 记 $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$;
- (2) 每一个基本事件的概率相同, 即 $P\{\omega_1\} = P\{\omega_2\} = \dots = P\{\omega_n\}$.

根据概率的公理化定义知

$$1 = P(\Omega) = P\left(\bigcup_{i=1}^n \{\omega_i\}\right) = \sum_{i=1}^n P\{\omega_i\},$$

于是由上述假设 (2), 知

$$P\{\omega_i\} = \frac{1}{n}, (i = 1, 2, \dots, n)$$

在古典概型的假设下,我们可以确定其相应的概率测度,即对任意事件 A , 确定其概率 $P(A)$. 设 A 为任意一个事件, 并令 A 中含有 m 个样本点, 分别为 i_1, i_2, \dots, i_m , 即

$$A = \{i_1, i_2, \dots, i_m\} = \bigcup_{j=1}^m \{i_j\}.$$

由公理化定义中的公理 3 知

$$P(A) = \sum_{j=1}^m P\{i_j\} = \frac{m}{n}.$$

因而, 古典概型的概率测度可表述为:

$$P(A) = \frac{A \text{ 中元素个数}}{\text{中元素个数}} = \frac{\text{使 } A \text{ 发生的基本事件数}}{\text{基本事件总数}},$$

其中 A 为 Ω 中的任意事件.

例 1.12 一个袋子中装有 10 个大小相同的球, 其中 3 个黑球, 7 个白球, 求

(1) 从袋子中任取一球, 这个球是黑球的概率;

(2) 从袋子中任取两球, 刚好一个白球一个黑球的概率以及两个球全是黑球的概率.

解 (1) 10 个球中任取一个, 共有 $C_{10}^1 = 10$ 种取法, 10 个球中有 3 个黑球, 取到黑球的取法有 $C_3^1 = 3$ 种, 从而根据古典概率计算, 事件 A : “取到的球为黑球” 的概率为

$$P(A) = \frac{C_3^1}{C_{10}^1} = \frac{3}{10}.$$

(2) 10 个球中任取两球的取法有 C_{10}^2 种, 其中刚好一个白球, 一个黑球的取法有 $C_3^1 \cdot C_7^1$ 种取法, 两个球均是黑球的取法有 C_3^2 种, 记 B 为事件 “刚好取到一个白球一个黑球”, C 为事件 “两个球均为黑球” 则

$$P(B) = \frac{C_3^1 C_7^1}{C_{10}^2} = \frac{21}{45} = \frac{7}{15},$$

$$P(C) = \frac{C_3^2}{C_{10}^2} = \frac{3}{45} = \frac{1}{15}.$$

例 1.13 将标号为 1, 2, 3, 4 的四个球随意地排成一行, 求下列各事件的概率:

(1) 各球自左至右或自右至左恰好排成 1, 2, 3, 4 的顺序;

(2) 第 1 号球排在最右边或最左边;

(3) 第 1 号球与第 2 号球相邻;

(4) 第 1 号球排在第 2 号球的右边 (不一定相邻).

解 将 4 个球随意地排成一行有 $4! = 24$ 种排法, 即基本事件总数为 24. 记

(1), (2), (3), (4) 的事件分别为 A、B、C、D.

(1) A 中有两种排法, 故有

$$P(A) = \frac{2}{24} = \frac{1}{12}.$$

(2) B 中有 $2 \times (3!) = 12$ 种排法, 故有

$$P(B) = \frac{12}{24} = \frac{1}{2}.$$

(3) 先将第 1, 2 号球排在任意相邻两个位置, 共有 2×3 种排法, 其余两个球可在其余两个位置任意排放, 共有 $2!$ 种排法, 因而 C 有 $2 \times 3 \times 2 = 12$ 种排法, 故

$$P(C) = \frac{12}{24} = \frac{1}{2}.$$

(4) 第 1 号球排在第 2 号球的右边的每一种排法, 交换第 1 号球和第 2 号球的位置便对应于第 1 号球排在第 2 号球的左边的一种排法, 反之亦然. 因而第 1 号球排在第 2 号球的右边与第 1 号球排在第 2 号球的左边的排法种数相同, 各占总排法数的 $\frac{1}{2}$, 故有 $P(D) = \frac{1}{2}$.

例 1.14 将 n 个球随意地放入 N 个箱子中 ($N \geq n$), 其中每个球都等可能地放入任意一个箱子, 求下列各事件的概率:

- (1) 指定的 n 个箱子各放一球;
- (2) 每个箱子最多放入一球;
- (3) 某指定的箱子不空;
- (4) 某指定的箱子恰好放入 k ($k \leq n$) 个球.

解 将 n 个球随意地放入 N 个箱子, 共有 N^n 种放法, 记 (1), (2), (3), (4) 的事件分别为 A, B, C, D.

(1) 将 n 个球放进指定的 n 个箱子, 每个箱子一个球, 其放法有 $n!$ 种, 故有

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

(2) 每个箱子最多放入一球等价于将 n 个球放进任意的 n 个箱子中, 每箱一个球, 其放法有 $C_N^n \cdot (n!)$ (或记作 P_N^n) 种, 于是

$$P(B) = \frac{P_N^n}{N^n}.$$

(3) 由于 C 的对立事件 \bar{C} 表示“指定的箱子是空的”, 它等价于将 n 个球全部放到其余 $N-1$ 个箱子中, 共有 $(N-1)^n$ 种放法, 从而

$$P(\bar{C}) = \frac{(N-1)^n}{N^n}, \quad P(C) = 1 - P(\bar{C}) = \frac{N^n - (N-1)^n}{N^n}.$$

(4) 先任取 k 个球 (有 C_n^k 种取法) 放入指定的箱子中, 然后将其余 $n-k$ 个

球随意地放入其余 $N - 1$ 个箱子 (共有 $(N - 1)^{n-k}$ 种放法), 于是某指定的箱子恰好放入 k 个球的放法有 $C_n^k (N - 1)^{n-k}$ 种, 故有

$$P(D) = \frac{C_n^k (N - 1)^{n-k}}{N^n}.$$

例 1.15 一个袋子中装有 $a + b$ 个球, 其中 a 个黑球, b 个白球, 随意地每次从中取出一球 (不放回), 求下列各事件的概率:

- (1) 第 i 次取到的是黑球;
- (2) 第 i 次才取到黑球;
- (3) 前 i 次中能取到黑球.

解 因为所考虑的事件涉及到取球的次序, 所以基本事件也应考虑顺序, $(a + b)$ 次取球的总取法为 $(a + b)!$, 记 (1), (2), (3) 中的事件分别为 A, B, C .

(1) 第 i 次取到的黑球可以是 a 个黑球中的任意一个, 选定其中一个以后, 其他各次取球必在 $a + b - 1$ 个球中任意选取, 共有 $(a + b - 1)!$ 种取法, 从而 A 中包含的取法有 $a \cdot [(a + b - 1)!]$ 种, 故

$$P(A) = \frac{a[(a + b - 1)!]}{(a + b)!} = \frac{a}{a + b}.$$

(2) 第 i 次才取到的黑球可以是 a 个黑球中的任意一个, 第 1 到第 $i - 1$ 次是在 b 个白球中任选 $i - 1$ 个 (共有 P_b^{i-1} 种取法) 其他各次在剩下的 $a + b - i$ 个球中任意选取 (共有 $(a + b - i)!$), 于是 B 所含的总取法为 $a \cdot P_b^{i-1} \cdot [(a + b - i)!]$, 故

$$P(B) = \frac{a \cdot P_b^{i-1} \cdot [(a + b - i)!]}{(a + b)!} = \frac{a \cdot P_b^{i-1}}{P_{a+b}^i}.$$

(3) 直接考虑事件 C 比较复杂, 先考虑其对立事件 \bar{C} : “前 i 次未取到黑球”, 显然 \bar{C} 包含的取法有 $P_b^i [(a + b - i)!]$ 于是

$$P(\bar{C}) = \frac{P_b^i [(a + b - i)!]}{(a + b)!} = \frac{P_b^i}{P_{a+b}^i} = \frac{C_b^i}{C_{a+b}^i},$$

故

$$P(C) = 1 - P(\bar{C}) = 1 - \frac{C_b^i}{C_{a+b}^i}.$$

值得指出的是例 1.15 的问题也可以理解为将 $(a + b)$ 个球随意地排序或理解为将 $(a + b)$ 个球随意地放入 $(a + b)$ 个箱子中, 每个箱子放入一个球. 此外, 从例 1.15 的结果, 我们还看出以下几点.

(1) 第 i 次取到的是黑球的概率与 i 无关, 均为 $\frac{a}{a + b}$, 自然, 这也等于第 1 次取到的是黑球的概率, 而后者的概率为 $\frac{a}{a + b}$ 则是显然的. 这一结果说明了实际生活中抽签的公平性: 一组签中有若干好签, 若干坏签, 不管你先抽还是后

抽，抽到好签的概率总是相同的。

(2) 从例中的结果来看，由于我们考虑的事件只涉及到前 i 次，所以我们可以只考虑取 i 次球，计算的结果与前面的结果一致。

(3) 对于例中的 (3)，我们按只考虑取 i 次球来解答时，问题中的事件并不涉及这 i 次中的次序问题，因而也可以不考虑顺序，例中的答案正说明了这一点。

最后指出，上面所举各例均是古典概率计算中的典型问题，每个例子均是一类问题的一个代表。我们有意识地编写成“取球”或“放球”问题，是为了使其更有代表性，这里的“球”可以理解为其他任何事物，只要所考虑的问题的结构与相应例子相同，我们便可套用例子的作法，读者可以通过本章提供的练习来领会这一点。

二、几何概型

古典概型是关于有限等可能结果的随机试验的概率模型。现在考虑样本空间为一线段、平面区域或空间立体等的等可能随机试验的概率模型。

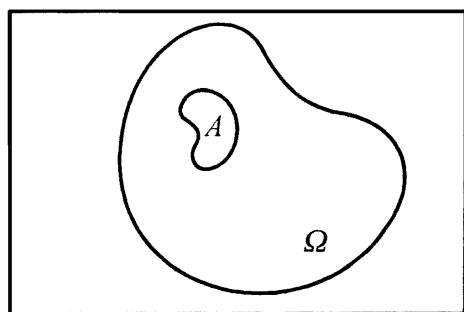


图 1.2 几何概型图示

如果我们在一个面积为 $S(\Omega)$ 的区域 Ω 中等可能地任意投点(如图 1.2). 这里“等可能”的确切含义是: 点落入 Ω 中任意区域 A 的可能性大小与区域 A 的面积 $S(A)$ 成正比, 而与其位置和形状无关. 将“点落入区域 A ”这一事件仍记为 A , 则有

$$P(A) = tS(A)$$

其中 t 为常数. 于是由

$$P(\Omega) = tS(\Omega) = 1$$

知 $t = \frac{1}{S(\Omega)}$. 进而得

$$P(A) = \frac{S(A)}{S(\Omega)} \quad (1.1)$$

由 (1.1) 定义的概率通常称为几何概率. 容易证明几何概率满足概率的公理化定义. 值得指出的是, 如果在一个线段上投点, 则几何概率 (1.1) 中的面积应改为长度; 如果在一个空间立体上投点, 则面积应改为体积.

例 1.16 某人午觉醒来, 发觉表停了, 他打开收音机, 想听电台报时, 设电台每正点时报时一次, 求他(她)等待时间短于 10 分钟的概率.

解 以分钟为单位, 记上一次报时时刻为 0, 则下一次报时时刻为 60, 于是这个人打开收音机的时间必在 $(0, 60)$, 记“等待时间短于 10 分钟”为事件 A , 则有 $\Omega = (0, 60)$, $A = (50, 60)$, 于是

$$P(A)=\frac{10}{60}=\frac{1}{6}.$$

例 1.17 (会面问题) 甲、乙两人相约在 7 点到 8 点之间在某地会面, 先到者等候另一人 20 分钟, 过时就离开. 如果每个人可在指定的一小时内任意时刻到达, 试计算二人能够会面的概率.

解 记 7 点为计算时刻的 0 时, 以分钟为单位, x, y 分别记甲、乙达到指定地点的时刻, 则样本空间为

$$=\{(x, y) \in [0, 60] \times [0, 60]\}.$$

以 A 表示事件“两人能会面”, 则显然有

$$A=\{(x, y) \in [0, 60] \times [0, 60] \mid |x-y| \leq 20\}$$

(如图 1.3)

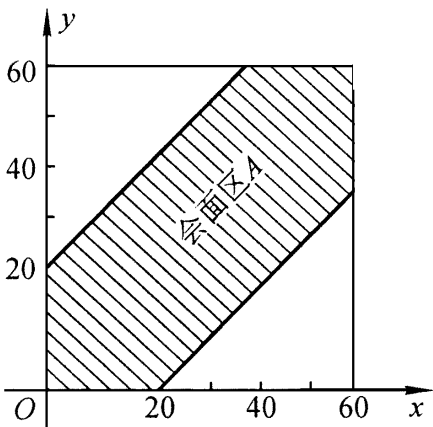


图 1.3 会面问题图示

根据题意, 这是一个几何概型问题, 于是

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

§ 1.4 条件概率

一、引出条件概率的例子

当我们拥有关于试验结果的额外信息, 比如, 已知一个事件 A 已经发生时, 我们可能需要对另一事件 B 发生的可能性大小重新作出度量, 先看一个例子.

例 1.18 一批同型号产品由甲、乙两厂生产, 产品结构如下表:

数 量 等 级		厂 别		合 计
		甲 厂	乙 厂	
合格品		475	644	1 119
次 品		25	56	81
合 计		500	700	1 200

从这批产品中随意地取一件, 则这件产品为次品的概率为

$$\frac{81}{1200}=6.83\%$$

现在假设被告知取出的产品是甲厂生产的, 那么这件产品为次品的概率是多大呢? 回答这一问题并不困难. 当我们被告知取出的产品是甲厂生产的, 我们不能肯定的是该件产品是甲厂生产的 500 件中的哪一件, 由于 500 件中有 25 件次

品, 自然我们可得出, 在已知取出的产品是甲厂生产的条件下, 它是次品的概率为

$$\frac{25}{500} = 5\%$$

记“取出的产品是甲厂生产的”这一事件为 A , “取出的产品为次品”这一事件为 B . 我们在前面实际上已计算了 $P(B)$, 同时我们也算出了在“已知 A 发生”的条件下, B 发生的概率, 这个概率称为在 A 发生的条件下, B 发生的条件概率, 记作 $P(B|A)$. 在本例中, 我们注意到:

$$P(B|A) = \frac{25}{500} = \frac{25/1200}{500/1200} = \frac{P(AB)}{P(A)}$$

事实上, 我们容易验证, 对一般的古典概型, 只要 $P(A) > 0$, 总有

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

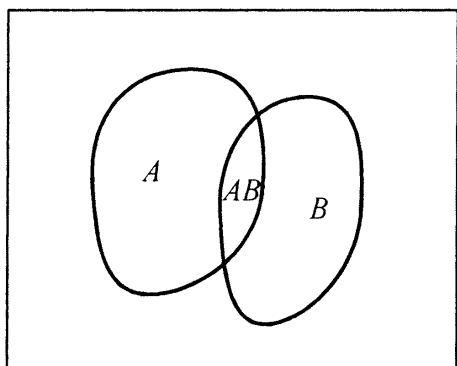


图 1.4 几何概型

在几何概型中 (以平面区域情形为例), 在平面上的有界区域、内等可能地投点 (如图 1.4), 若已知 A 发生, 则 B 发生的概率为

$$P(B|A) = \frac{S(AB)}{S(A)} = \frac{S(AB)/S(\cdot)}{S(A)/S(\cdot)} = \frac{P(AB)}{P(A)}.$$

可见, 在古典概型和几何概型这两类“等可能”概率模型中总有

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

事实上, 还可以验证这一关系对频率也成立. 由这些共性中得到启发, 我们在一般的概率空间中引入条件概率的数学定义.

二、条件概率的数学定义

定义 1.3 给定概率空间 (Ω, \mathcal{F}, P) , A, B 是其上的两个事件, 且 $P(A) > 0$, 则称

$$P(B|A) = \frac{P(AB)}{P(A)} \quad (1.2)$$

为已知事件 A 发生的条件下, 事件 B 发生的条件概率. 随着 B 在事件域 \mathcal{F} 中变化, $P(B|A)$ 便成为 \mathcal{F} 上的函数, 我们称之为在已知 A 发生的条件下的条件概率测度.

不难验证, 对给定的事件 A , $P(A) > 0$, 条件概率测度 $P(\cdot|A)$ 满足概率的三条公理:

$$(1) P(\Omega|A) = 1$$

(2) 对任意事件 B , 有 $P(B|A) \geq 0$

(3) 对任意可数个两两不相容的事件 $A_1, A_2, \dots, A_n, \dots$, 有

$$P\left(\bigcup_{i=1}^{\infty} A_i | A\right) = \sum_{i=1}^{\infty} P(A_i | A)$$

由此可见, 对给定的概率空间 (Ω, \mathcal{F}, P) 及事件 A , 如果 $P(A) > 0$, 则条件概率 $P(\cdot | A)$ 也是一个概率测度. 特别地, 当 $A = \Omega$ 时, $P(\cdot | \Omega)$ 就是原来的概率测度 $P(\cdot)$. 此外, 由于 $P(\cdot | A)$ 也是概率测度, 它也满足概率测度的其他性质.

在计算条件概率时, 有时从试验的结构可直接得到条件概率. 有时则需要利用 (1.2) 式来计算条件概率的值. 下列例子中的两个问题分别对应于这两种情况.

例 1.19 一袋中装有 10 个球, 其中 3 个黑球, 7 个白球, 先后两次从袋中各取一球 (不放回)

(1) 已知第一次取出的是黑球, 求第二次取出的仍是黑球的概率;

(2) 已知第二次取出的是黑球, 求第一次取出的也是黑球的概率.

解 记 A_i 为事件 “第 i 次取到的是黑球” ($i = 1, 2$)

(1) 在已知 A_1 发生, 即第一次取到的是黑球的条件下, 第二次取球就在剩下的 2 个黑球、7 个白球共 9 个球中任取一个, 根据古典概率计算, 取到黑球的概率为 $\frac{2}{9}$, 即有

$$P(A_2 | A_1) = \frac{2}{9}.$$

(2) 在已知 A_2 发生, 即第二次取到的是黑球的条件下, 第一次取球发生在第二次取球之前, 问题的结构不像 (1) 那么直观. 采用 (1.2) 式计算 $P(A_1 | A_2)$ 更方便一些. 因为

$$P(A_1 A_2) = \frac{P_3^2}{P_{10}^2} = \frac{1}{15}, \quad P(A_2) = \frac{3}{10}$$

由 (1.2) 可得

$$P(A_1 | A_2) = \frac{P(A_1 A_2)}{P(A_2)} = \frac{2}{9}.$$

例 1.19 中两个问题的结果表现出一种有趣的现象: 已知第二次取到的是黑球的条件下, 第一次取到的是黑球的概率与已知第一次取到的是黑球的条件下, 第二次取到的是黑球的概率相同, 这一结论具有一般性, 作为练习, 读者可以考虑 a 个黑球、 b 个白球的情形. 如何理解这一现象呢? 我们可以对例 1.19 的结果作出如下解释:

尽管第一次取球时, 可能取到的是 10 个球中的一个, 但当我们得知第二次取到的是黑球之后, 我们反过来知道第一次取球必是 2 个黑球、7 个白球, 共 9

个球中的一个, 从而结果跟 (1) 相同.

三、乘法公式

由条件概率定义中的 (1.2) 式立即可导出下列公式:

$$P(A \cap B) = P(A)P(B|A) \quad (P(A) > 0) \quad (1.3)$$

对称地, 如果 $P(B) > 0$, 由

$$P(A \cap B) = \frac{P(A \cap B)}{P(B)} P(B) \quad (1.4)$$

可得

$$P(A \cap B) = P(B)P(A|B) \quad (P(B) > 0) \quad (1.5)$$

式(1.3)和(1.5)通常称为两个事件交的概率的乘法公式. 在有些问题中, 条件概率 $P(B|A)$ 或 $P(A|B)$ 容易得到. 于是可用乘法公式计算交的概率 $P(A \cap B)$.

例 1.20 某批产品中, 甲厂生产的产品占 60%, 已知甲厂的产品的次品率为 10%, 从这批产品中随意地抽取一件, 求该件产品是甲厂生产的次品的概率.

解 该 A 表示事件“产品是甲厂生产的”, B 表示事件“产品是次品”, 由题设知

$$P(A) = 60\% \quad P(B|A) = 10\%$$

根据乘法公式, 有

$$P(A \cap B) = P(A)P(B|A) = 60\% \times 10\% = 6\%$$

例 1.21 一袋中装 10 个球, 其中 3 个黑球、7 个白球, 先后两次从中随意各取一球 (不放回), 求两次取到的均为黑球的概率.

这一概率, 我们曾用古典概型方法计算过, 这里我们使用乘法公式来计算. 在本例中, 问题本身提供了分两步完成一个试验的结构, 这恰恰与乘法公式的形式相对应, 合理地利用问题本身的结构来使用乘法公式往往是使问题得到简化的关键.

解 设 A_i 表示事件“第 i 次取到的是黑球” ($i = 1, 2$), 则 $A_1 \cap A_2$ 表示事件“两次取到的均为黑球”. 由题设知:

$$P(A_1) = \frac{3}{10} \quad P(A_2|A_1) = \frac{2}{9}$$

于是根据乘法公式, 有

$$P(A_1 \cap A_2) = P(A_1)P(A_2|A_1) = \frac{3}{10} \times \frac{2}{9} = \frac{1}{15}.$$

最后, 我们指出, 乘法公式 (1.3) 和 (1.5) 可以推广到有限个事件的交的概率的乘法公式:

$$P(A_1 A_2 \dots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 A_2) \dots P(A_n|A_1 A_2 \dots A_{n-1}) \quad (1.6)$$

四、全概率公式

人们在计算某一较复杂的事件的概率时，有时根据事件在不同情况或不同原因或不同途径下发生而将它分解成两个或若干个互不相容的部分的并，分别计算每一部分的概率，然后求和，我们先看两个例子。

例 1.22 一袋中有 10 个球，其中 3 个黑球，7 个白球，从中先后随意各取一球（不放回），求第二次取到的是黑球的概率。

这一概率，我们前面在古典概型中也计算过，这里我们用一种新的方法来计算。将事件“第二次取到的是黑球”根据第一次取球的情况分解成两个互不相容的部分，分别计算其概率再求和。

解 记 A_i 为事件“第 i 次取到的是黑球” ($i=1,2$)，则有

$$\begin{aligned} P(A_2) &= P(A_1A_2) + P(\bar{A}_1A_2) \\ &= P(A_1)P(A_2|A_1) + P(\bar{A}_1)P(A_2|\bar{A}_1) \end{aligned}$$

由题设易知

$$\begin{aligned} P(A_1) &= \frac{3}{10}, & P(\bar{A}_1) &= \frac{7}{10}, \\ P(A_2|A_1) &= \frac{2}{9}, & P(A_2|\bar{A}_1) &= \frac{3}{9}, \end{aligned}$$

于是有

$$P(A_2) = \frac{3}{10} \times \frac{2}{9} + \frac{7}{10} \times \frac{3}{9} = \frac{3}{10}.$$

例 1.23 人们为了解一支股票未来一定时期内价格的变化，往往会去分析影响股票价格的基本因素，比如利率的变化。现在假设人们经分析估计利率下调的概率为 60%，利率不变的概率为 40%。根据经验，人们估计，在利率下调的情况下，该支股票价格上涨的概率为 80%，而在利率不变的情况下，其价格上涨的概率为 40%，求该支股票将上涨的概率。

解 记 A 为事件“利率下调”，那么 \bar{A} 即为“利率不变”，记 B 为事件“股票价格上涨”。据题设知

$$\begin{aligned} P(A) &= 60\%, & P(\bar{A}) &= 40\%, \\ P(B|A) &= 80\%, & P(B|\bar{A}) &= 40\%, \end{aligned}$$

于是

$$\begin{aligned} P(B) &= P(AB) + P(\bar{A}B) \\ &= P(A)P(B|A) + P(\bar{A})P(B|\bar{A}) \\ &= 60\% \times 80\% + 40\% \times 40\% \\ &= 64\% \end{aligned}$$

上面两个不同的问题所采取的思路是完全一样的，要求一个事件 B 的概

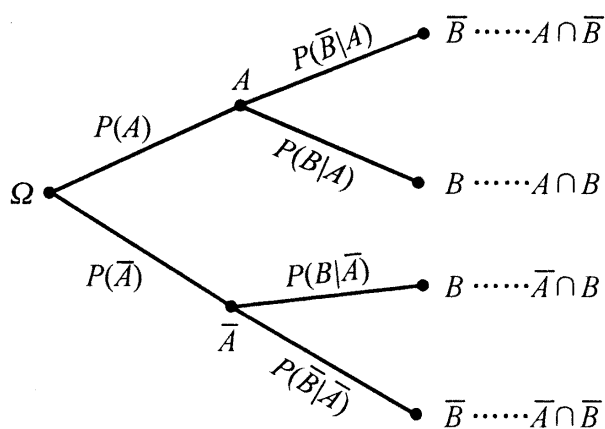


图 1.5 概率树

率，首先根据另一事件 A 发生与否两种情况（即 A 和 \bar{A} ）将事件 B 分解成两个不相容的部分： AB 和 $A\bar{B}$ ，分别计算这两部分的概率。而且人们事先能够得到每一种情况（ A 和 \bar{A} ）发生的概率以及在每一种情况下事件 B 发生的概率，即条件概率 $P(B|A)$ 和 $P(B|\bar{A})$ 。可由乘法公式分别计算两部分的概率。这一结构可以通过图 1.5 所示的概率树来描述。

将上述方法一般化，便得到下面的定理。

定理 1.5（全概率公式） 设 $\{A_i\}$ 是一列有限或可数个两两不相容的非零概率事件，且 $\bigcup_i A_i = \Omega$ ，则对任意事件 B ，有

$$P(B) = \sum_i P(A_i)P(B|A_i) \quad (1.7)$$

$$\begin{aligned} \text{证明} \quad P(B) &= P(B \cap \Omega) = P[B \cap (\bigcup_i A_i)] \\ &= P[\bigcup_i (B \cap A_i)] = \sum_i P(B \cap A_i) \\ &= \sum_i P(A_i)P(B|A_i). \end{aligned}$$

例 1.22 和例 1.23 实际上是定理 1.1 的特殊情形。

五、贝叶斯公式

利用全概率公式，人们可以通过综合分析一个事件发生的不同原因、情况或途径及其可能性来求得该事件发生的概率。现在我们来考虑与之完全相反的问题。观察到一个事件已经发生，我们要考察所观察到的事件发生的各种原因，情况或途径的可能性，这里先通过一个简单的例子加以说明。

例 1.24（续例 1.22） 例 1.22 中，如果我们观察到第二次取到的球是黑球，求第一次取到的是黑球的概率。

这一问题在例 1.19 (2) 中曾作出过解答，现在，我们换一个思路来考虑这一问题。正如例 1.22 那样，我们可以将“第二次取到的球为黑球”这一事件分解为两种情况下发生，那里利用全概率公式算得“第二次取到的球为黑球”的概率。现在的问题是，假设我们已经观察到“第二次取到的球为黑球”，但我们不知道是在第一次取到的球为黑球的情况下第二次取到的是黑球，还是在第一次取到的球为白球的情况下第二次取到的是黑球，要求“第一次取到的是黑球”这种“情况”发生的概率。

解 设“第一次取到的是黑球”这一事件为 A ，“第二次取到的是黑球”这一事件为 B ，则问题归结为求条件概率 $P(A|B)$ 。

根据条件概率的定义，有

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

由乘法公式知

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

由全概率公式知

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

于是得

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\bar{A})P(B|\bar{A})} \quad (1.8)$$

据题设易知

$$P(A) = \frac{3}{10}, \quad P(B|A) = \frac{2}{9}$$

$$P(\bar{A}) = \frac{7}{10}, \quad P(B|\bar{A}) = \frac{3}{9}$$

从而得

$$P(A|B) = \frac{\frac{3}{10} \times \frac{2}{9}}{\frac{3}{10} \times \frac{2}{9} + \frac{7}{10} \times \frac{3}{9}} = \frac{2}{9}.$$

上述例子中的式 (1.8) 可以推广，我们将其表述为一个一般的公式，称为贝叶斯公式，它是以英国数学家贝叶斯 (Bayes Thomas) 命名的。

定理 1.2 设 $\{A_i\}$ 是有限或可数个两两不相容的非零概率事件，且 $\sum_{i=1}^n A_i = B$ ，则对任意事件 B ， $P(B) > 0$ ，有

$$P(A_i|B) = \frac{P(A_i B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^n P(A_j)P(B|A_j)} \quad (1.9)$$

由全概公式和乘法公式，该定理的证明是容易的，略去。

例 1.25 设某批产品中，甲、乙、丙三厂生产的产品分别占 45%，35%，20%，各厂的产品的次品率分别为 4%，2%，5%，现从中任取一件，

(1) 求取到的是次品的概率；

(2) 经检验发现取到的产品为次品，求该产品是甲厂生产的概率。

解 记“该产品为甲厂生产的”这一事件为 A_1 ；“该产品为乙厂生产的”这一事件为 A_2 ；“该产品为丙厂生产的”这一事件为 A_3 ；“该产品是次品”这一事件为 B 。由题设知：

$$P(A_1) = 45\%, \quad P(A_2) = 35\%, \quad P(A_3) = 20\%,$$

$$P(B|A_1) = 4\%, \quad P(B|A_2) = 2\%, \quad P(B|A_3) = 5\%.$$

(1) 由全概公式得

$$\begin{aligned} P(B) &= \sum_{i=1}^3 P(A_i)P(B|A_i) \\ &= 45\% \times 4\% + 35\% \times 2\% + 20\% \times 5\% \\ &= 3.5\% \end{aligned}$$

(2) 由贝叶斯公式(或条件概率定义)得:

$$P(A_1|B) = \frac{P(A_1B)}{P(B)} = \frac{P(A_1)P(B|A_1)}{P(B)} = \frac{45\% \times 4\%}{3.5\%} = 51.4\%.$$

§ 1.5 事件的独立性

一、两个事件的独立性

前一节曾指出,考察同一试验的两个事件,有时一个事件的发生与否会影响另一事件发生的概率.但有时,一个事件的发生与否并不影响另一事件发生的概率.比如,投掷一枚硬币和投掷一枚骰子组成一个试验中,硬币是否出现正面,不会影响骰子出现点数为 5 的概率.在数学上,一个事件 B 发生与否对另一事件 A 发生的概率没有任何影响,可表述为:

$$P(A|B) = P(A) \quad (1.10)$$

其中 $P(B) > 0$, 并称 A 独立于 B. 同样, 如果

$$P(B|A) = P(B) \quad (1.11)$$

其中 $P(A) > 0$, 则称 B 独立于 A.

由于 $P(A) > 0, P(B) > 0$ 时, (1.10) 和 (1.11) 均等价于

$$P(AB) = P(A)P(B) \quad (1.12)$$

因而, 此时, A 独立于 B 等价于 B 独立于 A, 故通常称 A 与 B 相互独立. 注意到 $P(A) = 0$, 或 $P(B) = 0$ 时 (1.12) 恒成立, 为了使独立性概念包括零概率事件的情形, 我们以后采用下列定义:

定义 1.4 设 (Ω, \mathcal{F}, P) 是一个概率空间, A, B 是其上的两个事件, 如果

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

则称 A 与 B 相互独立, 简称 A 与 B 独立.

例 1.26 投掷一枚均匀的骰子

(1) 设 A 表示事件“点数小于 5”, B 表示事件“点数为奇数”, 则有

$$P(A) = \frac{4}{6}, \quad P(B) = \frac{3}{6}, \quad P(AB) = \frac{2}{6} = \frac{1}{3}$$

由于

$$P(AB) = P(A)P(B) = \frac{1}{3}$$

故 A 与 B 独立.

(2) 设 A 表示事件“点数小于 4”, B 同(1), 则有 $P(A) = \frac{3}{6}$, $P(B) = \frac{3}{6}$, $P(AB) = \frac{2}{6} = \frac{1}{3}$

由于

$$P(AB) \neq P(A)P(B)$$

故 A 与 B 不独立.

二、有限个事件的独立性

定义 1.5 如果 n 个事件 ($n \geq 2$): A_1, A_2, \dots, A_n 中任意两个事件均相互独立, 即对任意 $1 \leq i < j \leq n$, 均有

$$P(A_i A_j) = P(A_i)P(A_j) \quad (1.13)$$

则称 n 个事件 A_1, A_2, \dots, A_n 两两独立.

两两独立的概念只涉及到 n 个事件中每对事件之间的相互关系, 这种关系不涉及到第三者. 但我们通常还会碰到需要同时考虑多个事件之间的关系, 比如, 如果三个事件 A_1, A_2, A_3 两两独立, 那么粗略地讲, 其中任意一个事件发生与否对另一事件发生的概率没有影响, 即 $P(A_i | A_j) = P(A_i)$, $i \neq j$, $i, j = 1, 2, 3$. 但是, 有例子可说明其中两个事件的发生可能会影响另一事件的概率, 比如, 可能会出现:

$$P(A_1 | A_2 A_3) \neq P(A_1)$$

为此, 我们引入一个比“两两独立”更强的独立性关系—— n 个事件相互独立.

定义 1.6 设 A_1, A_2, \dots, A_n 为 n 个事件 ($n \geq 2$) 如果其中任何 k 个事件 ($2 \leq k \leq n$): $A_{i_1}, A_{i_2}, \dots, A_{i_k}$ ($1 \leq i_1 < i_2 < \dots < i_k \leq n$) 均有

$$P(A_{i_1} A_{i_2} \dots A_{i_k}) = P(A_{i_1})P(A_{i_2}) \dots P(A_{i_k}) \quad (1.14)$$

则称 A_1, A_2, \dots, A_n 相互独立.

显然, 当 $n=2$ 时, 两两独立与相互独立是同一概念. 但一般地, 当 $n>2$ 时, 相互独立性是比两两独立性更强的性质. 直观地讲, n 个事件相互独立, 那么其中任何一个或多个事件的发生不会对其他事件发生的概率产生影响.

此外, 从定义 1.5 容易看出, 如果 n 个事件相互独立, 则其中任意 k 个事件 ($2 \leq k \leq n$) 也相互独立.

三、相互独立性的性质

事件之间具有独立性的情况下，很多概率计算将变得简单，其中经常用到下面两个性质.

性质 1 如果 n 个事件 A_1, A_2, \dots, A_n 相互独立，则将其中任何 m ($1 \leq m \leq n$) 个事件改为相应的对立事件，形成的新的 n 个事件仍然相互独立.

性质 2 如果 n 个事件 A_1, A_2, \dots, A_n 相互独立，则有

$$P\left(\bigcap_{i=1}^n A_i\right) = 1 - \prod_{i=1}^n P(A_i) = 1 - \prod_{i=1}^n (1 - P(A_i)) \quad (1.15)$$

性质 1 的证明只需对 $m=1$ 证明，然后使用数学归纳法. 性质 2 是性质 1 的直接推论. 详细证明过程，留作练习.

例 1.27 甲、乙、丙三人各射一次靶，他们各自中靶与否相互独立，且已知他们各自中靶的概率分别为 0.5, 0.6, 0.8，求下列事件的概率.

(1) 恰有一人中靶

(2) 至少有一个中靶

解 设 A_i ($i=1, 2, 3$) 分别表示甲、乙、丙中靶三个事件，则“恰有 1 人中靶”这一事件可表示为： $A_1A_2A_3 + A_1A_2A_3 + A_1A_2A_3$ ，“至少有 1 人中靶”这一事件可表示为： $A_1 + A_2 + A_3$.

$$\begin{aligned} (1) & P(A_1A_2A_3 + A_1A_2A_3 + A_1A_2A_3) \\ &= P(A_1A_2A_3) + P(A_1A_2A_3) + P(A_1A_2A_3) \\ &= P(A_1)P(A_2)P(A_3) + P(A_1)P(A_2)P(A_3) + P(A_1)P(A_2)P(A_3) \\ &= 0.5 \times (1 - 0.6) \times (1 - 0.8) + (1 - 0.5) \times 0.6 \times (1 - 0.8) + \\ & \quad (1 - 0.5)(1 - 0.6) \times 0.8 \\ &= 0.26 \end{aligned}$$

$$\begin{aligned} (2) P(A_1 + A_2 + A_3) &= 1 - P(A_1)P(A_2)P(A_3) \\ &= 1 - 0.5 \times 0.4 \times 0.2 \\ &= 0.96 \end{aligned}$$

四、伯努利 (Bernoulli) 概型

有时，我们需要研究的问题涉及到多个甚至无穷多个试验，这多个或无穷多个试验通常称为一个试验序列.

定义 1.7 一个试验序列称为一个独立试验序列，如果它的各试验的结果之间是相互独立.

我们在实际中经常碰到一类特殊的试验，它只有两个可能结果，这样的试验称为伯努利试验. 比如，投掷一枚硬币时观察其出现正面还是反面，抽取一

件产品考察其是正品还是次品等.

有些试验的结果虽然不只两个,但我们可能对试验感兴趣的是某事件 A 是否发生,那么我们可以把 A 作为一个结果, \bar{A} 作为另一结果,从而也可以将试验归结为伯努利试验. 比如,一个灯泡的寿命. 它可以取不小于 0 的任何数值,但有时根据需要,我们将寿命大于 1000 小时的灯泡当作合格品,而把寿命不大于 1000 小时的灯泡当作次品,我们感兴趣的可能是灯泡是合格品还是次品.

定义 1.8 一个试验序列称为伯努利试验序列, 如果它是由一个伯努利试验独立重复进行形成的试验序列. 特别地, 由一个伯努利试验独立重复 n 次形成的试验序列称为 n 重伯努利试验.

设在一次伯努利试验中, $P(A) = p, P(\bar{A}) = q$, 其中 $p > 0, q > 0$, 且 $p + q = 1$, 那么相应的 n 重伯努利试验, 根据定义, 其特点为: 事件 A 在每次试验中发生的概率均为 p , 且不受其他各次试验中 A 是否发生的影响.

定理 1.3 (伯努利定理) 在一次试验中, 事件 A 发生的概率为 p ($0 < p < 1$), 则在 n 重伯努利试验中, 事件 A 恰好发生 k 次的概率 (记作 $b(k; n, p)$) 为:

$$b(k; n, p) = C_n^k p^k q^{n-k}$$

其中 $q = 1 - p$.

证明 记“第 i 次试验中事件 A 发生”这一事件为 $A_i, i = 1, 2, \dots, n$, 则“事件 A 恰好发生 k 次” (记作 B_k) 是下列 C_n^k 个两两不相容事件的并:

$$A_{i_1} A_{i_2} \dots A_{i_k} \bar{A}_{j_1} \bar{A}_{j_2} \dots \bar{A}_{j_{n-k}}$$

其中 i_1, i_2, \dots, i_k 是取遍 $1, 2, \dots, n$ 中的任意 k 个数 (共有 C_n^k 种取法), j_1, j_2, \dots, j_{n-k} 是取走 i_1, i_2, \dots, i_k 后剩下的 $n - k$ 个数.

而对任意取出的 i_1, i_2, \dots, i_k , 根据独立性及 $P(A_i) = p$, 有

$$\begin{aligned} & P(A_{i_1} A_{i_2} \dots A_{i_k} \bar{A}_{j_1} \bar{A}_{j_2} \dots \bar{A}_{j_{n-k}}) \\ &= P(A_{i_1}) P(A_{i_2}) \dots P(A_{i_k}) P(\bar{A}_{j_1}) \dots P(\bar{A}_{j_{n-k}}) \\ &= p^k q^{n-k} \end{aligned}$$

故有

$$b(k; n, p) = P(B_k) = C_n^k p^k q^{n-k}$$

定理 1.4 在伯努利试验序列中, 设每次试验中事件 A 发生的概率为 p , “事件 A 在第 k 次试验中才首次发生” ($k \geq 1$) 这一事件的概率为

$$g(k, p) = q^{k-1} p$$

证明 “事件 A 在第 k 次试验中才首次发生” 等价于在前 k 次试验组成的 k 重伯努利试验中“事件 A 在前 $k-1$ 次试验中均不发生而第 k 次试验中事件 A 发生”, 于是根据定理 1.3 的证明可知定理 1.4 成立.

例 1.28 一袋中装有 10 个球, 其中 3 个黑球 7 个白球, 每次从中随意取出一球, 取后放回.

(1) 如果共取 10 次, 求 10 次中能取到黑球的概率及 10 次中恰好取到 3 次黑球的概率.

(2) 如果未取到黑球就一直取下去, 直到取到黑球为止, 求恰好要取 3 次的概率及至少要取 3 次的概率.

解 记 A_i 为事件“第 i 次取到的是黑球”则 $P(A_i) = \frac{3}{10}$, $i = 1, 2, \dots$,

(1) 记 B 为事件“10 次中能取到黑球”, B_k 为事件“10 次中恰好取到 k 次黑球” $k = 0, 1, \dots, 10$. 于是有

$$P(B) = 1 - P(\bar{B}) = 1 - P(B_0) = 1 - \left(\frac{7}{10}\right)^{10},$$

$$P(B_3) = C_{10}^3 \left(\frac{3}{10}\right)^3 \left(\frac{7}{10}\right)^7.$$

(2) 记 C 为“恰好要取 3 次”, D 为“至少要取 3 次”则由定理 1.4 知

$$P(C) = \left(\frac{7}{10}\right)^2 \cdot \frac{3}{10},$$

$$P(D) = P(A_1 A_2) = P(A_1)P(A_2) = \left(\frac{3}{10}\right)^2.$$

例 1.29 一辆飞机场的交通车载有 25 名乘客途经 9 个站, 每位乘客都等可能在这 9 站中任意一站下车 (且不受其他乘客下车与否的影响), 交通车只有在有乘客下车时才停车, 求交通车在第 i 站停车的概率以及在第 i 站不停车的条件下第 j 站停车的概率, 并判断“第 i 站停车”与“第 j 站停车”两个事件是否独立.

解 记 A_k 为“第 k 位乘客在第 i 站下车” $k = 1, 2, \dots, 25$. 考察每一位乘客在第 i 站是否下车, 可视为一个 25 重的伯努利试验, 记 B 为“第 i 站停车”, C 为“第 j 站停车”, 则 B 、 C 分别等价于“第 i 站有人下车”和“第 j 站有人下车”于是有

$$P(B) = 1 - \left(\frac{8}{9}\right)^{25}, \quad P(C) = 1 - \left(\frac{8}{9}\right)^{25}.$$

在 B 不发生 (即 \bar{B} 发生) 的条件下, 每位乘客均等可能地在第 i 站以外的 8 个站中任意一站下车, 于是每位乘客在第 j 站下车的概率为 $\frac{1}{8}$, 故有

$$P(C|\bar{B}) = 1 - \left(\frac{7}{8}\right)^{25}.$$

由于 $P(C|\bar{B}) \neq P(C)$, 故 B 与 C 不独立, 从而 B 与 C 不独立.

习 题 一

(A)

1. 写出下列各试验的样本空间:

(1) 掷两个骰子, 分别观察其出现的点数;

(2) 观察一支股票某日的价格 (收盘价);

(3) 一人射靶三次, 观察其中靶次数;

(4) 一袋中装有 10 个同型号的零件, 其中 3 个合格 7 个不合格, 每次从中随意取出一个, 不合格便放回去, 直到取到合格的零件为止, 观察所抽取的次数.

2. 试问第 1 题中你所给出的样本空间中, 哪些遵循古典概型?

3. 掷一个骰子, 观察其出现的点数, A 表示“出现奇数点”, B 表示“出现的点数小于 5”, C 表示“出现的点数是小于 5 的偶数”, 用集合列举法表示下列事件: \bar{A} , A , B , C , $A + B$, $A - B$, $B - A$, AB , AC , $A + B$.

4. 写出例 1.9 中 (5), (8), (10), (11) 等事件的其他表示方法.

5. 互不相容事件与对立事件的区别何在? 说出下列各对事件之间的关系.

(1) $x = a$ 与 $x \neq a$;(2) $x > 20$ 与 $x \leq 20$;(3) $x > 20$ 与 $x < 18$;(4) $x > 20$ 与 $x \leq 22$;

(5) “20 件产品全是合格品”与“20 件产品中恰有一件是废品”;

(6) “20 件产品全是合格品”与“20 件产品中至少有一件是废品”;

(7) “20 件产品全是合格品”与“20 件产品中至多有一件是废品”.

6. 某人用步枪射击目标 5 次, A_i 表示“第 i 次射击击中目标” ($i = 1, 2, 3, 4, 5$). B_i 表示“5 次射击中击中目标 i 次” ($i = 0, 1, 2, 3, 4, 5$), 用文字叙述下列各事件, 并指出各对事件之间的关系.(1) $\bigcap_{i=1}^5 A_i$ 与 $\bigcap_{i=1}^5 B_i$;(2) $\bigcap_{i=2}^5 A_i$ 与 $\bigcap_{i=2}^5 B_i$;(3) $\bigcap_{i=1}^2 A_i$ 与 $\bigcap_{i=3}^5 A_i$;(4) $\bigcap_{i=1}^2 B_i$ 与 $\bigcap_{i=3}^5 B_i$;(5) $\bigcap_{i=0}^2 B_i$ 与 $\bigcap_{i=3}^5 B_i$;(6) $A_1 A_2 A_3 A_4 A_5$ 与 B_3 ;(7) A_1 与 B_5 ;(8) $\bigcap_{i=1}^5 A_i$ 与 B_5 .

7. 证明下列关系式:

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

8. 由概率的公理化定义推导概率测度的其他性质: 性质 3 —— 性质 6 以及性质 6 的推论.

9. 已知 $P(A) = 0.4$, $P(B) = 0.25$, $P(A - B) = 0.25$, 求 $P(AB)$, $P(A \cup B)$, $P(B - A)$, $P(\overline{AB})$.
10. 已知 $P(A) = 0.4$, $P(BA) = 0.2$, $P(CA \cup B) = 0.1$, 求 $P(A \cup B \cup C)$
11. N 件产品中有 N_1 件次品, 从中任取 n 件 (不放回), 其中 $1 \leq n \leq N$. (1) 求其中恰有 k 件 ($k \leq n$ 且 $k \leq N_1$) 次品的概率; (2) 求其中有次品的概率; (3) 如果 $N_1 = 2$, $n = 2$, 求其中至少有两件次品的概率.
12. 一个班共有 30 名同学, 其中有 6 名女生, 假设他们到校先后次序的所有模式都有同样的可能性. (1) 求男生均比女生先到校的概率; (2) 求班上李明和王菲两位同学中, 李明比王菲先到校的概率.
13. 某班级有 n 个同学 ($n \leq 365$), 求至少有两位同学的生日在同一天概率 (设一年按 365 天计).
14. 一辆飞机场的交通车载有 25 名乘客, 途经 9 个站, 每位乘客都等可能在 9 个站中任意一站下车, 交通车只在有乘客下车时才停车, 求下列各事件的概率.
- (1) 交通车在第 i 站停车;
 - (2) 交通车在第 i 站和第 j 站至少有一站停车;
 - (3) 交通车在第 i 站和第 j 站均停车;
 - (4) 在第 i 站有 3 人下车;
15. 两封信随机地投入 4 个邮筒, 求前两个邮筒没有信的概率及第一个邮筒恰有一封信的概率.
16. 求例 1.15 中, 前 i 次中恰好取到 k 个黑球的概率.
17. 一串钥匙, 共有 10 把, 其中有 4 把能打开门, 因开门者忘记哪些能打开门, 便逐把试开, 求下列事件的概率.
- (1) 第 3 把钥匙能打开门;
 - (2) 第 3 把钥匙才打开门;
 - (3) 最多试 3 把钥匙就能打开门.
18. 连续投掷一枚均匀硬币 10 次, 求其中有 3 次是正面的概率.
19. 一个袋子中装有 5 个红球, 3 个白球, 2 个黑球, 从中任取 3 个球, 求其中恰有一个红球, 一个白球和一个黑球的概率.
20. 将 13 个分别写有 A、A、A、C、E、H、I、I、M、M、N、T、T 的卡片随意地排成一行, 求恰好排单词 "MATHEMATICIAN" 的概率.
21. 某公共汽车站每隔 10 分钟有一辆汽车到达, 一位乘客到达汽车站的时间是任意的, 求他等候时间不超过 3 分钟的概率.
22. 两艘轮船都要停靠同一泊位, 它们可能在一昼夜的任意时间到达, 设两船停靠泊位的时间分别需要 1 小时与两小时, 求一艘轮船停靠泊位时, 需要等待空出码头的概率.
23. 一袋中装有 a 个黑球, b 个白球先后两次从袋中各取一球 (不放回).
- (1) 已知第一次取出的是黑球, 求第二次取出的仍是黑球的概率;
 - (2) 已知第二次取出的是黑球, 求第一次取出的也是黑球的概率;
 - (3) 已知取出的两个球中有一个是黑球, 求另一个也是黑球的概率.

24. 已知 A 是概率空间 (Ω, \mathcal{F}, P) 上的事件, $P(A) > 0$, 证明:

(1) 如果 $B \subset A$, 则 $P(B|A) = 1$;

(2) 如果 B_1, B_2 是 (Ω, \mathcal{F}, P) 上两个事件, 且 $B_1 B_2 = \emptyset$, 则 $P(B_1 \cup B_2|A) = P(B_1|A) + P(B_2|A)$.

25. 一个家庭中有两个小孩, (1) 已知其中有一个是女孩, 求另一个也是女孩的概率;

(2) 已知第一胎是女孩, 求第二胎也是女孩的概率.

26. 掷两个均匀的骰子, 它们的点数各不相同, 求其中有一个点数为 4 的概率.

27. 10 个考题签中有 4 题难答. 3 人参加抽签, 甲先抽, 乙次之, 丙最后. 求下列事件的概率:

(1) 甲抽到难答签;

(2) 甲未抽到难答签而乙抽到难答签;

(3) 甲、乙、丙均抽到难答签.

28. 有外形相同的球分装 3 个盒子, 每盒 10 个球, 其中第一个盒子中有 7 个球标有字母 A, 3 个球标有字母 B; 第二个盒子中有红球和白球各 5 个; 第三个盒子则有红球 8 个, 白球 2 个. 试验按如下规则进行: 先在第一个盒子中任取一球, 若取得标有字母 A 的球, 则在第二个盒子中任取一球; 若第一次取到标有字母 B 的球, 则在第三个盒子中任取一球. 如果第三次取到的是红球, 则称试验成功, 求试验成功的概率.

29. 某商店收进甲厂生产的产品 30 箱, 乙厂生产的同种产品 20 箱, 甲厂产品每箱装 100 个, 废品率为 0.06, 乙厂产品每箱 120 个废品率为 0.05

(1) 任取一箱, 从中任取一个产品, 求其为废品的概率;

(2) 若将所有产品开箱混装, 任取一个其为废品的概率.

30. 12 个乒乓球中有 9 个新球, 3 个旧球第一次比赛, 取出 3 个球, 用完以后放回去, 第二次比赛又从中取出 3 个球. (1) 求第二次取出的三个球中有 2 个新球的概率; (2) 若第二次取出的 3 个球中有 2 个新球, 求第一次取到的 3 个球中恰有一个新球的概率.

31. 血液试验 ELISA (Enzyme Linked Immunosorbent Assay 酶连接免疫吸附测定) 是现今检验艾滋病病毒的一种流行方法. 假定 ELISA 试验能正确测出确实带有病毒的人中的 95% 存在艾滋病病毒, 又把不带病毒的人中的 1% 不正确地识别为存在病毒. 又假定在总人口 1 000 人中大约有 1 人确实带有艾滋病病毒, 如果对某人的检验结果呈阳性 (即认为带有病毒), 那么他真的带有艾滋病病毒的概率有多大? 如果被检测者属于“高感染人群”中的一员, 而估计这一高感染人群中大约 100 人中有 1 人带有病毒, 那么检测为阳性的人, 真的带有艾滋病病毒的概率有多大?

32. 一男子到闹市区去, 他遇到背后袭击并被枪劫, 他断言凶手是个白人, 然而当调查这一案件的法院在可比较的光照条件下多次重复展现现场情况时, 受害者正确识别袭击者种族的次数约占 80%, 袭击者确实是白人的概率是 0.8 吗? 试给出说明.

33. 一个均匀的四面体, 其第一面染红色第二面染白色, 第三面染黑色, 而第四面染红、白、黑三种颜色, 以 A、B、C 分别记投掷一次四面体, 底面出现红、白、黑的三个事件, 判断 A、B、C 是否两两独立? 是否相互独立?

34. 设 $P(A) = 0$ 或 1, 证明 A 与其他任何事件 B 相互独立.

35. 设 $P(A) > 0, P(B) > 0$, 且 A, B 互不相容, 那么 A 与 B 相互独立吗? 为什么?
36. A, B, C 相互独立, 证明 A, B, C 亦相互独立.
37. 证明相互独立性的性质 1 和性质 2.
38. 加工一个产品要经过三道工序, 第一、二、三道工序不出废品的概率分别为 0.9, 0.95, 0.8, 若假定各工序是否出废品是独立的, 求经过三道工序生产出的是废品的概率.
39. 一个自动报警器由雷达和计算机两部分组成, 两部分有任何一个失灵, 这个报警器就失灵, 若使用 100 小时后, 雷达失灵的的概率为 0.1, 计算机失灵的的概率为 0.3, 若两部分失灵与否独立, 求这个报警器使用 100 小时而不失灵的的概率.
40. 高射炮向敌机发射三发炮弹 (每弹击中与否相互独立), 以每发炮弹击中敌机的概率均为 0.3, 又知若敌机中一弹, 其坠落的概率为 0.2; 若敌机中两弹, 其坠落的概率为 0.6; 若中三弹则必然坠落. (1) 求敌机被击落的概率; (2) 若敌机被击落, 求它中两弹的概率.

习 题 一

(B)

1. 为了减少比赛场次, 把 20 个球队分成两组, 每组 10 队, 进行比赛, 求最强的两队被分在同一组的概率, 及最强两队分在不同组的概率.
2. 若 n 个人站成一行, 其中有 A, B 两人, 问夹在 A, B 之间恰有 r 个人的概率是多少? 如果 n 个人围成一个圆圈, 求从 A 到 B 的顺时针方向, A, B 之间恰有 r 个人的概率.
3. (巴拿赫问题) 某数学家有两盒火柴, 每一盒装有 N 根. 每次使用时, 他在任一盒中取一根, 问他发现一盒空, 而另一盒还有 k 根火柴的概率是多少?
4. 有 k 个坛子, 每一个装有 n 个球, 分别编号为 1 至 n , 今从每个坛子中任取一球, 问 m 是所取的球中的最大编号的概率.
5. 一根长为 1 的棍子在任意两点折断, 试计算得到的三段能围成三角形的概率.
6. 今有两名射手轮流对同一目标射击, 甲射手命中概率为 p_1 , 乙射手命中概率为 p_2 , 甲先射, 谁先命中该得胜, 求甲、乙得胜的概率各是多少?
7. 设有来自三个地区的 10 名、15 名、25 名考生的报名表, 其中女生的报名表分别为 3 份、7 份、5 份. 随机地取一个地区的报名表, 从中先后抽取两份.
- (1) 求先抽到的一份是女生表的概率 p
- (2) 已知后抽到的一份是男生表, 求先抽到的是女生表的概率 q .
8. 在一通讯渠道中, 能可能传送字符 AAAA, BBBB, CCCC 三者之一, 由于通讯噪声干扰, 正确接收到被传送字母的概率为 0.6, 而接收到其他两个字母的概率均为 0.2, 若前后字母是否被歪曲互不影响.
- (1) 求收到字符 ABCA 的概率;
- (2) 若收到字符为 ABCA, 问被传送字符为 AAAA 的概率是多大.

第 2 章

随机变量的分布与数字特征

对于一个随机试验,人们除了对某些特定的事件发生的概率感兴趣以外,往往还会关心某个与随机试验的结果相联系的变量.由于这一变量的取值依赖于试验结果,而试验结果是不确定的,所以这一变量的取值也是不确定的,这种变量因而被称为随机变量.对于随机变量,人们无法准确预知其确切取值,但人们可以研究其取值的统计规律性.对一个随机变量的统计规律性的完整描述被称为随机变量的分布,本章将介绍的两类随机变量——离散型和连续型随机变量及其分布.但在多数实际应用中,人们很难知道一个随机变量的真实分布,这时人们希望用一些综合指标来反映随机变量的统计规律中的某些重要特征,这些指标被称为随机变量的数字特征.最常用的数字特征有“数学期望”和“方差”.随机变量的数字特征不仅容易估计,而且为随机变量的取值提供了综合评价,这恰恰是解决一些概率应用问题的必要前提.

§ 2.1 随机变量及其分布

一、随机变量的概念

在一些随机试验中,试验的结果本身就是由数量来表示的.比如,投掷一个骰子,观察其出现的点数,可能的结果可分别由 1, 2, 3, 4, 5, 6, 来表示;再比如,观察一个灯泡的使用寿命,实际使用寿命可能是 $[0, +\infty)$ 中的任何一个实数.在另一些随机试验中,我们可能根据问题的需要对每一个可能结果指定一个数量.比如,投掷一枚硬币进行打赌时,如果规定投掷者在硬币出现正面时赢 1 元钱,出现反面时输 1 元钱,则可对“出现正面”指定一个数 1,对“出现反面”指定一个数 -1. 无论哪种情形,其共同点是:对每一个可能结果,有惟一一个实数与之对应.这种对应关系实际定义了样本空间 Ω 上的函数,通常记作 $X = X(\omega)$, $\omega \in \Omega$. 这与微积分中定义的函数概念本质上并无区别,只不过在微积分中的函数 $y = f(x)$, 其自变量 x 通常是实数,而且只关注 y 对 x 的依赖关系,并不关心 x 的取值是否是确定的.在这里函数 $X = X(\omega)$, $\omega \in \Omega$, 其

“自变量”在样本空间中“取值”(即定义域为 Ω),我们主要关心的不确定性及其统计规律所导致的因变量 X 的不确定性及其统计规律,这种取值依赖于一个试验的结果而具有不确定性的变量称为随机变量,其正式数学定义如下:

定义 2.1 定义在概率空间 (Ω, \mathcal{F}, P) 上,取值为实数的函数: $X = X(\omega)$,
称为 (Ω, \mathcal{F}, P) 上的一个随机变量.

随机变量 X 的取值由样本点 ω 决定.反过来, X 取某一特定值 a 的那些样本点的全体构成 Ω 的一个子集,即 $\{\omega \in \Omega | X(\omega) = a\}$.同样,设 I 为实数集 R 的一个子区间,使得 X 的值落在 I 上的那些样本点全体也是 Ω 的一个子集.为了研究随机变量 X 的统计规律,我们均假设这些子集是随机事件,也假设这些事件的可数并、交,及补都是事件,并称这些事件为随机变量 X 生成的事件.为表达简便,今后在事件的表示中省掉 Ω ,比如 $\{X = a\}$, $\{a < X \leq b\}$ 等.

在掷骰子的试验中,其出现的点数记为随机变量 X ,作为样本空间 $\Omega = \{1, 2, 3, 4, 5, 6\}$ 上的函数 X 定义为:

$$X(\omega) = \begin{cases} 1, & \omega = 1, \\ 2, & \omega = 2, \\ \dots \end{cases}$$

在投掷一枚硬币进行打赌时,出现正面时投掷者赢 1 元钱,出现反面时输 1 元钱,记赢钱数为随机变量 X ,则 X 作为样本空间 $\Omega = \{\text{正面}, \text{反面}\}$ 上的函数定义为

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

二、离散型随机变量的概率分布

在所有的随机变量中,有一类随机变量最简单,它只有有限个或可数个可能取值.

定义 2.2 设 X 是定义在概率空间 (Ω, \mathcal{F}, P) 上的一个随机变量,如果 X 的全部可能取值只有有限或可数个,则称 X 是一个离散型随机变量.

设 x_1, x_2, \dots 是 X 的所有可能取值,我们知道对每个 x_i , $\{X = x_i\}$ 是 Ω 上的一个随机事件,人们往往关心这些事件发生的可能性,也即 X 取每一个可能值 x_i 的概率.

定义 2.3 设 X 是离散型随机变量,其全部可能取值为 $\{x_i, i = 1, 2, \dots\}$,记

$$p(x_i) = P\{X = x_i\} \quad i = 1, 2, \dots \quad (2.1)$$

则称 $\{p(x_i), i = 1, 2, \dots\}$ 为 X 的概率分布.有时也将 $p(x_i)$ 记为 p_i ,用下列表格形式来表示并称之为 X 的概率分布表.

X	x_1	x_2	...	x_i	...
P	p_1	p_2	...	p_i	...

容易看出，任何一个离散型随机变量的概率分布 $\{p(x_i)\}$ 必然满足下列性质：

(1) $p(x_i) \geq 0, \quad i=1, 2, \dots,$ (2.2)

(2) $\sum_i p(x_i) = 1$ (2.3)

例 2.1 投掷一枚均匀硬币，设 X 为一次投掷中出现正面的次数，即

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

则有

$$P\{X=1\} = P\{\text{出现正面}\} = \frac{1}{2}$$

$$P\{X=0\} = P\{\text{出现反面}\} = \frac{1}{2}$$

于是 X 的概率分布为

X	1	0
P_i	$\frac{1}{2}$	$\frac{1}{2}$

或表示为 $X = \begin{cases} 1, & \frac{1}{2}; \\ 0, & \frac{1}{2}. \end{cases}$

例 2.2 设离散型随机变量 X 的概率分布为

(1) $P\{X=i\} = a \cdot \left(\frac{2}{3}\right)^i, \quad i=1, 2, 3;$

(2) $P\{X=i\} = a \cdot \left(\frac{2}{3}\right)^i, \quad i=1, 2, \dots.$

分别求上述各式中的常数 a .

解(1) 由于

$$1 = \sum_{i=1}^3 P\{X=i\} = \sum_{i=1}^3 a \cdot \left(\frac{2}{3}\right)^i = a \cdot \frac{2}{3} \cdot \frac{38}{27},$$

故有, $a = \frac{27}{38}$.

(2) 由于

$$1 = \sum_{i=1}^{\infty} P\{X=i\} = \sum_{i=1}^{\infty} a \cdot \left(\frac{2}{3}\right)^i = a \cdot \frac{\frac{2}{3}}{1 - \frac{2}{3}} = 2a,$$

故有, $a = \frac{1}{2}$.

一旦知道一个离散型随机变量 X 的概率分布 $\{p(x_i)\}$ ，我们便可求得 X 所

生成的任何事件的概率，特别地

$$\begin{aligned} P\{a \leq X \leq b\} &= P\left(\bigcup_{a \leq x_i \leq b} \{X = x_i\}\right) = \sum_{a \leq x_i \leq b} P\{X = x_i\} \\ &= \sum_{a \leq x_i \leq b} p(x_i) \quad (a \leq b). \end{aligned} \quad (2.4)$$

一般地， I 是一个区间，则

$$P\{X \in I\} = \sum_{x_i \in I} p(x_i) \quad (2.5)$$

例 2.3 设 X 的概率分布由例 2.2 (1) 给出，则有：

$$P\{X < 1\} = 0,$$

$$P\{X \leq 1\} = P\{X = 1\} = \frac{27}{38} \times \frac{2}{3} = \frac{9}{19},$$

$$P\{X < 2\} = P\{X = 1\} = \frac{9}{19},$$

$$P\{X < 2.5\} = P\{X = 1\} + P\{X = 2\} = \frac{9}{19} + \frac{27}{38} \times \frac{4}{9} = \frac{15}{19},$$

$$P\{X \leq 3\} = P\{X = 1\} + P\{X = 2\} + P\{X = 3\} = 1,$$

$$P\{X \leq 4\} = P\{X = 1\} + P\{X = 2\} + P\{X = 3\} = 1.$$

三、分布函数

离散型随机变量的概率分布为离散型随机变量的统计规律提供了一目了然的描述。然而对那些取值非可数的随机变量，比如，测量的误差，灯泡的寿命等，我们如果同离散型随机变量一样，通过罗列取每一个值及其相应的概率来描述一个随机变量会遇到不可克服的困难。其一，这类随机变量的非可数个取值无法一一列举出来；其二，从下面的例子可看到取连续值的随机变量，它取某个特定值的概率往往是 0。不过，对取连续值的随机变量，我们往往关心的是它的取值落在一定范围（即区间或区间的并）的概率，而不关心它取某个特定值的概率。因此，对这类随机变量，我们希望能够对其取值落于任何一个区间上的概率给出描述。

例 2.4 等可能地在数轴上的有界区间 $[a, b]$ 上投点，记 X 为落点的位置（数轴上的坐标）则 X 是样本空间 $\Omega = [a, b]$ 上的函数：

$$X(\omega) = \omega, \quad \omega \in [a, b], \quad (2.6)$$

根据几何概型，我们知道，对任意 $c \in [a, b]$ ，有

$$P\{X = c\} = P\{\omega = c\} = 0.$$

而对任意 $B = (c, d] \subset [a, b]$ ，有

$$P\{c < X \leq d\} = P\{\text{落点在 } B \text{ 中}\} = \frac{d - c}{b - a}. \quad (2.7)$$

另一方面, 由于

$$P\{c < X \leq d\} = P\{X \leq d\} - P\{X \leq c\}, \quad (2.8)$$

为给出 X 取值于任意区间上的概率, 我们实际上只要给出所有 X 取值于形如 $(-\infty, x]$ 的区间上的概率 $P\{X \leq x\}$ 即可.

记

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

当 x 取遍 $(-\infty, +\infty)$ 上的一切实数时, $F(x)$ 便成为定义在 $(-\infty, +\infty)$ 上的函数, 一旦知道了这个函数 $F(x)$, 我们便可得到相应的随机变量取值于任何区间的概率.

上述例子的讨论中 (2.8) 式适合于任意随机变量, 为此, 我们引入下列定义:

定义 2.4 设 X 是一随机变量, 则称函数

$$F(x) = P\{X \leq x\} \quad x \in (-\infty, +\infty) \quad (2.9)$$

为随机变量 X 的分布函数, 记作 $X \sim F(x)$.

例 2.5 求例 2.4 中的随机变量 X 的分布函数.

解: 当 $x < a$ 时, $\{X \leq x\}$ 是不可能事件, 于是,

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

当 $a \leq x < b$ 时, 由于 $\{X \leq x\} = \{a \leq X \leq x\}$, 且 $[a, x] \subset [a, b]$, 于是,

$$F(x) = P\{X \leq x\} = P\{a \leq X \leq x\} = \frac{x-a}{b-a}.$$

当 $x \geq b$ 时, 由于 $\{X \leq x\} = \{a \leq X \leq b\}$, 于是

$$F(x) = P\{X \leq x\} = P\{a \leq X \leq b\} = \frac{b-a}{b-a} = 1.$$

综上, 可得 X 的分布函数为:

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

根据定义 2.4, 可以导出一个随机变量的分布函数必然满足下列性质:

(1) 单调性. 若 $x_1 < x_2$, 则 $F(x_1) \leq F(x_2)$;

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

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

(3) 右连续性. $F(x+) = F(x)$

上述性质中, 单调性容易证明, 留作练习. 其他性质的证明超出了本书的要求, 略去.

另一方面, 如果一个函数 $F(x)$ 满足上述三条性质, 则可以证明, 它一定是某一随机变量 X 的分布函数, 因此, 通常将满足上述三条性质的函数都称为分布函数.

四、离散型随机变量的分布函数

一个离散型随机变量的分布也可由分布函数来描述, 事实上我们将看到其概率分布与分布函数能够相互确定.

例 2.6 设 X 由例 2.1 给出, 求其分布函数.

解: X 只有两个可能取值, 其概率分布为:

$$P\{X=0\}=P\{X=1\}=\frac{1}{2}.$$

于是, 当 $x < 0$ 时,

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

当 $0 \leq x < 1$ 时,

$$F(x)=P\{X \leq x\}=P\{X=0\}=\frac{1}{2}.$$

当 $x \geq 1$ 时,

$$F(x)=P\{X \leq x\}=P\{X=0\}+P\{X=1\}=\frac{1}{2}+\frac{1}{2}=1.$$

综上, X 的分布函数为:

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

如图 2.1, 观察上例中的 X 的分布函数, 我们发现 $F(x)$ 是一个阶梯形的函数,

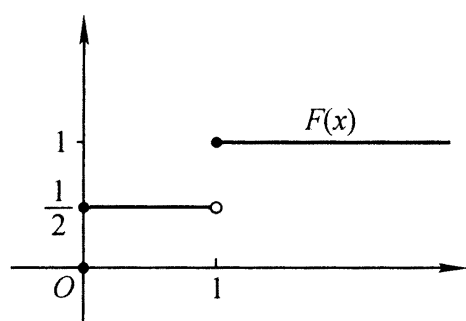


图 2.1 例 2.6 的分布函数

它在 X 的可能取值点 $0, 1$ 处发生跳跃, 跳跃高度等于相应点处的概率, 而在两个相邻跳跃点之间分布函数值保持不变. 这一特征实际上是所有离散型随机变量的共同特征, 而且反过来, 如果一个随机变量 X 的分布函数 $F(x)$ 是阶梯型函数, 则 X 一定是一个离散型随机变量, 其概率分布可由分布函数 $F(x)$ 惟一确定: $F(x)$ 的跳跃点全体构成 X 的所有可能取值, 每一跳跃点处的跳跃高度则是 X 在相应点处的概率.

例 2.7 设随机变量 X 的分布函数为

$$F(x) = \begin{cases} 0, & x < 1, \\ \frac{9}{19}, & 1 \leq x < 2, \\ \frac{15}{19}, & 2 \leq x < 3, \\ 1, & x \geq 3. \end{cases}$$

由于 $F(x)$ 是一个阶梯型函数, 故知 X 是一个离散型随机变量, $F(x)$ 的跳跃点分别为 1, 2, 3, 对应的跳跃高度分别为 $\frac{9}{19}$, $\frac{6}{19}$, $\frac{4}{19}$, 如图 2.2.

故 X 的概率分布为

$$P\{X=1\} = \frac{9}{19}, P\{X=2\} = \frac{6}{19}, P\{X=3\} = \frac{4}{19}.$$

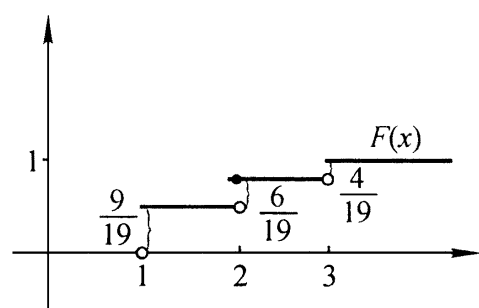


图 2.2 例 2.7 的分布函数

五、连续型随机变量及其概率密度

在例 2.4 中, 我们得到了在 $[a, b]$ 上等可能投落点的位置 X 的分布函数, 现在, 我们换一角度来考虑. 由于 X 在 $[a, b]$ 上等可能取值, 而 X 在 $[a, b]$ 上取值的概率为 1, 我们可将这一概率 1 视为均匀分布在 $[a, b]$ 的每一点上, 为此可求得概率 1 在区间 $[a, b]$ 上的平均值: $\frac{1}{b-a}$, 这个平均值称为 X 在 $[a, b]$ 上的概率密度. 这里, 由于 X 在 $[a, b]$ 上等可能取值, 因而 X 在 $[a, b]$ 上的每一点有相同的密度 $\frac{1}{b-a}$. 而且易见, 对任意 $[c, d] \subset [a, b]$, 有

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

由于 $[a, b]$ 以外的点是 X 不可能取值点, 我们记其密度为 0, 令

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{其他,} \end{cases} \quad (2.10)$$

称 $f(x)$ 为 X 的概率密度函数. 容易验证, X 的分布函数可表示为:

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

对一般情形, 我们引入下列定义

定义 2.5 一个随机变量 X 称为连续型随机变量, 如果存在一个非负可积函数 $f(x)$, 使得:

$$F(x) = P\{X \leq x\} = \int_{-\infty}^x f(t) dt. \quad (2.11)$$

并称 $f(x)$ 为 X 的概率密度函数, 或简称为密度函数.

根据定义 2.5 及分布函数的性质, 易知密度函数具有下列性质:

$$(1) f(x) \geq 0 \quad x \in (-\infty, +\infty);$$

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

反过来, 可以证明, 一个函数满足上述两个性质, 一定可以作为某一连续型随机变量的密度函数.

对于一个给定的连续型随机变量 X , 如果已知其密度函数, 根据定义 2.5, 自然可以求得其分布函数, 同时, 可以通过密度函数的积分来求 X 的取值落于任意区间上的概率:

$$P\{a < X \leq b\} = F(b) - F(a) = \int_a^b f(x) dx. \quad (2.12)$$

在几何上, $F(x)$ 和 $P\{a < X \leq b\}$ 可以用相应的曲边梯形的面积来表示 (如图 2.3)

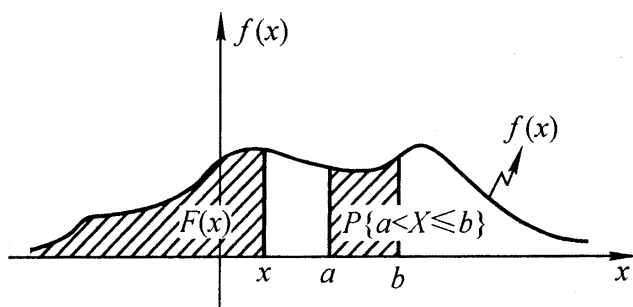


图 2.3 密度函数

由 (2.12) 式还可得知, 对任意实数 x , 有

$$P\{X = x\} = 0. \quad (2.13)$$

此外, 由 (2.11) 式知, 在 $f(x)$ 的连续点处, 有

$$F'(x) = f(x). \quad (2.14)$$

它为我们从分布函数出发确定密度函数提供了途径.

例 2.8 设 X 是在 $[a, b]$ 上等可能投点的位置, 在例 2.5 中, 我们已求得 X 的分布函数, 试由分布函数求其密度函数.

解 在例 2.5 中, 我们已求出其分布函数为

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

于是其密度函数为

$$f(x) = F'(x) = \begin{cases} 0, & x < a, \\ \frac{1}{b-a}, & a < x < b, \\ 0, & x > b. \end{cases} \quad (2.15)$$

在 $x = a$ 和 $x = b$ 处, $F(x)$ 的导数不存在, 可补充定义这两点的密度为:

$$f(x) = \frac{1}{b-a}, \quad x = a \text{ 或 } x = b. \quad (2.16)$$

(原则上, 补充定义 $f(a)$ 和 $f(b)$ 为其他非负值也是可以的, 因为改变密度函数个别点处的值不影响其在区间上的积分值.)

上面求得的结果, 与 (2.10) 一致: 随机变量 X 在 $[a, b]$ 上的密度为一个常数 $\frac{1}{b-a}$, 而在 $[a, b]$ 以外的密度为 0, 我们将把这样的随机变量称为在 $[a, b]$ 上服从均匀分布, 即指 X 取 $[a, b]$ 上的每一点是“等可能的”, 这正好是对我们在例 2.4 中定义的 X 的“等可能性”的一个更为明确的表述.

§ 2.2 随机变量的数字特征

一、离散型随机变量的数学期望

我们知道, 离散型随机变量的统计规律可以由其概率分布完全描述, 但在许多实际问题中, 这种完全的描述并不使人感到方便, 为了某些目的, 我们希望用一些数字指标——数字特征来反映随机变量的统计规律的某些层面, 数学期望就是其中之一, 它主要反映随机变量取值的平均水平.

例如, 观察一名射手 20 次射击的成绩如下:

中靶环数(x_i)	0	1	2	3	4	5	6	7	8	9	10
频数(n_i)	1	2	1	2	3	3	2	1	2	2	1
频率(f_i)	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{3}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{1}{20}$

人们常常使用“平均中靶环数”来对射手的射击水平作出综合评价, 记平均中靶环数为 \bar{x} , 则有:

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=0}^{10} x_i n_i}{n} = \sum_{i=0}^{10} x_i f_i \\ &= 0 \times \frac{1}{20} + 1 \times \frac{2}{20} + 2 \times \frac{1}{20} + \dots + 9 \times \frac{2}{20} + 10 \times \frac{1}{20} \\ &= 5.\end{aligned}$$

我们知道, 当试验次数增大时, 频率的稳定值就是概率, 那么完整描述该射手真实水平的是其射中各环数的概率分布, 相应地, 观察到的平均中靶环数 \bar{x} , 随试验次数增大也将趋于一个稳定值. 设中靶环数 X (观察之前为随机变量) 的概率分布为:

$$P\{X=i\}=p_i, \quad i=0, 1, 2, \dots, 10,$$

则 \bar{x} 的稳定值为

$$\sum_{i=0}^{10} x_i p_i.$$

它是对该射手的真实水平的综合评价.

定义 2.6 若离散型随机变量 X 的可能值为 $x_i (i=1, 2, \dots)$, 其概率分布为

$$P\{X=x_i\}=p_i, \quad i=1, 2, \dots,$$

则当

$$\sum_{i=1}^{\infty} x_i p_i < \infty \quad (2.17)$$

时, 称 X 的数学期望存在, 并且其数学期望记作 EX , 定义为

$$EX = \sum_{i=1}^{\infty} x_i p_i. \quad (2.18)$$

在上述定义中, 既然数学期望定义为 $EX = \sum_{i=1}^{\infty} x_i p_i$, 那么只要 $\sum_{i=1}^{\infty} x_i p_i$ 收敛就可以了, 为什么还要有条件

$$\sum_{i=1}^{\infty} x_i p_i < \infty,$$

是不是有点多余呢? 我们知道, 离散型随机变量可依某种次序一一列举的, 对同一随机变量, 它的取值的列举次序可以不同, 但当改变次序时, 其数学期望不应该改变, 这意味着改变 $\sum_{i=1}^{\infty} x_i p_i$ 的求和次序, 其收敛性及和不应改变, 为此必须要

求 $\sum_{i=1}^{\infty} x_i p_i$ 绝对收敛.

例 2.9 设盒中有 5 个球, 其中 2 个白球, 3 个黑球, 从中随意抽取 3 个球. 记 X 为抽取到的白球数, 求 EX .

解 X 只可能取 0, 1, 2, 这三个实数值, 而且根据古典概型计算, 有

$$P\{X=0\} = \frac{C_3^3}{C_5^3} = \frac{1}{10},$$

$$P\{X=1\} = \frac{C_3^2 C_2^1}{C_5^3} = \frac{6}{10},$$

$$P\{X=2\} = \frac{C_3^1 C_2^2}{C_5^3} = \frac{3}{10}.$$

于是

$$EX = 0 \times \frac{1}{10} + 1 \times \frac{6}{10} + 2 \times \frac{3}{10} = 1.2.$$

二、连续型随机变量的数学期望

在微积分中, 我们知道对离散取值 (可数或有限个) 的函数——数列, 我

们可以考虑其和，而对连续取值（即值域为一个区间）的函数，我们不能简单地考虑函数值的和，这时函数的和的概念被扩充为定积分．回顾函数在 $[a, b]$ 上的定积分的定义，我们知道定积分实际上是和式的极限，受此启发，我们来考虑连续型随机变量的数学期望．

设 X 是连续型随机变量，密度函数为 $f(x)$ ，为简单起见，设 $f(x)$ 只在有限区间 $[a, b]$ 上取不为零的值，即对一切 $x \notin [a, b]$ ，有 $f(x) = 0$ ．取分点：

$$a = x_0 < x_1 < \dots < x_{n+1} = b,$$

则 X 落在区间 $x_i = (x_i, x_{i+1})$ 中的概率为

$$P\{X \in x_i\} = \int_{x_i}^{x_{i+1}} f(x) dx. \tag{2.19}$$

当 Δx 很小时，

$$P\{X \in x_i\} \approx f(x_i) \Delta x. \tag{2.20}$$

这时，概率分布：

x_i	x_0	x_1	\dots	x_n
p_i	$f(x_0) \Delta x_0$	$f(x_1) \Delta x_1$	\dots	$f(x_n) \Delta x_n$

可视为 X 的离散近似，服从上述分布的离散型随机变量的数学期望为

$$\sum_{i=0}^n x_i f(x_i) \Delta x_i \tag{2.21}$$

它近似地表达了连续型随机变量 X 的平均值——数学期望，当分点越来越密时，近似会越来越好．根据定积分的定义，上述和式（2.21）以定积分：

$$\int_a^b x f(x) dx = \lim_{\Delta x \rightarrow 0} \sum_{i=0}^n x_i f(x_i) \Delta x_i$$

为极限，于是这一定积分的值便是 X 的精确的数学期望值．如果 X 在无穷区间上取值，上述定义还应扩充到广义积分，这里就不再讨论了．一般地，我们给出下列定义：

定义 2.7 若 X 为连续型随机变量， $f(x)$ 为其密度函数，如果

$$\int_{-\infty}^{+\infty} |x| f(x) dx < +\infty,$$

则称 X 的数学期望存在，并且其数学期望记作 EX ，定义为

$$EX = \int_{-\infty}^{+\infty} x f(x) dx. \tag{2.22}$$

例 2.10 设随机变量 X 的密度函数为

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

求 EX .

解 因为 $f(x)$ 只在有限区间 $[-\frac{\pi}{2}, \frac{\pi}{2}]$ 上不为 0, 且在该区间上为连续函数, 所以 EX 存在, 且

$$EX = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} xf(x) dx = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} x \cos^2 x dx,$$

根据奇函数的性质知, $EX = 0$,

例 2.11 设随机变量 X 的密度函数为

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

求 EX .

解 显然 EX 存在, 且

$$\begin{aligned} EX &= \int_{-\infty}^{+\infty} xf(x) dx = \int_0^1 x^2 dx + \int_1^2 x(2-x) dx \\ &= \left. \frac{x^3}{3} \right|_0^1 + \left. \left(2x - \frac{1}{2}x^2 \right) \right|_1^2 = 1. \end{aligned}$$

三、随机变量函数的数学期望

在许多实际问题中, 我们除了对某一随机变量进行研究以外, 往往还要对该随机变量的函数进行研究. 设 X 是一个随机变量, $g(x)$ 是任意实函数, 则 X 与 $g(x)$ 复合成

$$Y = g(X).$$

显然 Y 也是一个随机变量, 但值得指出的是 $g(x)$ 本身是一个确定性的函数, 即 X 与 Y 的关系是确定性的, 这意味着当 X 取定某值时, Y 的值将由函数关系 $g(x)$ 惟一确定. 正因为此, Y 的随机性完全由 X 的随机性所决定, 进而, Y 的分布原则上由 X 的分布所确定. 然而, 从 X 的分布出发求 Y 的分布并非易事. 关于随机变量函数的分布, 我们将在本章 § 2.4 中讨论, 这里, 我们先讨论随机变量函数的数学期望. $Y = g(X)$ 作为一个随机变量, 其数学期望按定义应根据其分布来计算, 然而下面的结论表明, 我们不必求 Y 的分布, 而根据 X 的分布即可求其函数 $Y = g(X)$ 的数学期望, 这显然为计算随机变量函数的数学期望提供了极大的方便.

定理 2.1 设 X 是一个随机变量, $g(x)$ 是一个实函数.

(1) 若 X 为离散型随机变量, 概率分布为

$$P\{X = x_i\} = p_i \quad i = 1, 2, \dots$$

且 $\sum_{i=1}^{\infty} g(x_i) p_i < \infty$, 则 $Eg(X)$ 存在, 且

$$Eg(X) = \sum_{i=1}^{\infty} g(x_i) p_i \quad (2.23)$$

(2) 若 X 为连续型随机变量, $f(x)$ 是其密度函数, 且 $\int_{-\infty}^{+\infty} g(x) f(x) dx < \infty$, 则 $Eg(X)$ 存在, 且

$$Eg(X) = \int_{-\infty}^{+\infty} g(x) f(x) dx \quad (2.24)$$

证明(这里仅就离散型情形给以证明, 连续型的证明已超过本书要求, 读者可与离散型对比来理解相应结果)

令 $Y = g(X)$, 则 Y 仍然是一个离散型随机变量, 设其可能取值为 $y_j, j = 1, 2, \dots$. 于是

$$\begin{aligned} P\{Y = y_j\} &= P\left(\bigcup_{i: g(x_i)=y_j} \{X = x_i\}\right) \\ &= \sum_{i: g(x_i)=y_j} P\{X = x_i\} \end{aligned}$$

由数学期望定义有:

$$\begin{aligned} Eg(X) &= \sum_{j=1}^{\infty} y_j P\{Y = y_j\} \\ &= \sum_{j=1}^{\infty} y_j \sum_{i: g(x_i)=y_j} P\{X = x_i\} \\ &= \sum_{j=1}^{\infty} \sum_{i: g(x_i)=y_j} g(x_i) P\{X = x_i\} \\ &= \sum_{i=1}^{\infty} g(x_i) P\{X = x_i\} \end{aligned}$$

例 2.12 设 X 如例 2.9, 即 X 的概率分布如右表, 求 $E(X - EX)^2$.

解 由例 2.9 已解得 $EX = 1.2$, 于是根据定理 2.1(1)有:

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

$$\begin{aligned} E(X - EX)^2 &= (0 - 1.2)^2 \times \frac{1}{10} + (1 - 1.2)^2 \times \frac{6}{10} + (2 - 1.2)^2 \times \frac{3}{10} \\ &= 1.44 \times \frac{1}{10} + 0.04 \times \frac{6}{10} + 0.64 \times \frac{3}{10} \\ &= 0.36 \end{aligned}$$

例 2.13 设 X 的密度函数如例 2.11, 即

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

求 $E(X)$ 及 $E(X^2)$

解 由例 2.11 已算得 $E(X) = 1$, 于是由定理 2.1 (2) 有:

$$\begin{aligned} E(X^2) - E(X)^2 &= E(X^2 - 1) = \int_{-\infty}^{+\infty} (x^2 - 1)f(x) dx \\ &= \int_0^1 (x^2 - 1)x dx + \int_1^2 (x^2 - 1)(2-x) dx \\ &= \int_0^1 (1-x)x dx + \int_1^2 (x-1)(2-x) dx \\ &= \int_0^1 (1-x)x dx + 2 \int_1^2 (x-1) dx \\ &= \frac{1}{3} \end{aligned}$$

四、数学期望的性质

将数学期望的一些基本性质概括出来可能会为许多问题的考虑提供方便, 下面列举的性质, 对离散型和连续型随机变量而言, 是定理 2.1 的直接推论, 但事实上, 它们对一般的随机变量也都成立.

(1) 对任意常数 a , 有 $Ea = a$

(2) 设 c_1, c_2 为任意实数, $g_1(x), g_2(x)$ 为任意实函数, 如果 $Eg_1(X), Eg_2(X)$ 均存在, 则

$$E[c_1g_1(X) + c_2g_2(X)] = c_1Eg_1(X) + c_2Eg_2(X) \quad (2.25)$$

(3) 如果 EX 存在, 则对任意实数 a , 有

$$E(X + a) = EX + a \quad (2.26)$$

证明(这里仅对离散型给出证明, 连续型的证明与之类似, 读者可以自行证明)

(1) 取 $g(x) = a$, 由定理 2.1 即得.

(2) 首先, 由

$$c_1g_1(x) + c_2g_2(x) = c_1g_1(x) + c_2g_2(x)$$

可知

$$c_1g_1(x_i) + c_2g_2(x_i) = c_1g_1(x_i) + c_2g_2(x_i)$$

因为 $Eg_1(X), Eg_2(X)$ 存在, 故由上式得知 $E[c_1g_1(X) + c_2g_2(X)]$ 存在. 令 $g(x) = c_1g_1(x) + c_2g_2(x)$, 代入(2.23)即可证得(2.25)成立.

(3) 在(2)中: 取 $g_1(x) = x, g_2(x) = a$, 即得(3)

例 2.14 设 EX, EX^2 均存在, 证明

$$E(X - EX)^2 = EX^2 - (EX)^2 \quad (2.27)$$

证明 因为 $(X - EX)^2 = X^2 - 2X \cdot EX + (EX)^2$, 于是由(2.25)得:

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

五、随机变量的方差

随机变量的数学期望是对随机变量取值水平的综合评价, 在许多问题中, 我们还需要了解随机变量的其他特征. 比如, 在投资决策中, 我们选择投资某一项目或购买某种资产(如股票、债券等), 我们不仅关心其未来的收益水平, 还关心其未来收益的不确定性程度, 前者通常用数学期望来度量, 后者通常称为风险程度, 有许多种衡量方法, 最简单、直观的方法就是用方差来度量. 一个随机变量的方差, 粗略地讲, 反映随机变量偏离其中心—数学期望的平均偏离程度.

定义 2.8 设 X 为一个随机变量, 其数学期望 EX 存在, 则称 $X - EX$ 为 X 的离差, 进一步, 如果 $E(X - EX)^2$ 也存在, 则称 $E(X - EX)^2$ 为随机变量 X 的方差, 记作 DX 或 $\text{Var}X$, 并称 \sqrt{DX} 为 X 的标准差.

由于 X 是一个随机变量, 其离差 $X - EX$ 因而也是一个随机变量, 平均来讲, X 的正、负离差相互抵消, 因而 $E(X - EX) = 0$, 为考虑 X 对 EX 的偏离程度, 我们必须消除符号的影响, 为此用 $(X - EX)^2$ 来衡量 X 对 EX 的偏离, 从而方差 $DX = E(X - EX)^2$ 即为 X 对 EX 的平均偏离. 当然为消除离差中的符号, 我们也可考虑使用绝对离差 $|X - EX|$, 但由于 $E|X - EX|$ 中绝对值不便于处理, 所以人们习惯于使用方差来作为随机变量偏离其期望的偏离程度的度量.

根据随机变量函数的期望的计算方法(定理 2.1), 在(2.23)和(2.24)中, 取 $g(x) = (x - EX)^2$, 即可计算方差:

若 X 为离散型随机变量, 其概率分布为: $P\{X = x_i\} = p_i, i = 1, 2, \dots$, 则

$$DX = E(X - EX)^2 = \sum_i (x_i - EX)^2 p_i \quad (2.28)$$

若 X 为连续型随机变量, $f(x)$ 为其密度函数, 则

$$DX = E(X - EX)^2 = \int_{-\infty}^{+\infty} (x - EX)^2 f(x) dx \quad (2.29)$$

此外, 由例 2.14, 也可通过(2.27)来计算方差, 即

$$DX = EX^2 - (EX)^2 \quad (2.30)$$

由数学期望的性质, 容易导出方差的一些基本性质:

设 X 的方差 DX 存在, a 为任意常数, 则

$$(1) D a = 0; \quad (2.31)$$

$$(2) D(X + a) = DX; \quad (2.32)$$

$$(3) D(aX) = a^2 DX. \quad (2.33)$$

例 2.15 设 X 的分布如例 2.12, 例 2.12 实际上已算得 X 的方差 $DX = E(X - EX)^2 = 0.36$, 其中 $EX = 1.2$ 在例 2.9 中算得. 现在我们改用 (2.30) 式来计算方差:

$$EX^2 = 0^2 \times \frac{1}{10} + 1^2 \times \frac{6}{10} + 2^2 \times \frac{3}{10} = 1.8,$$

$$DX = EX^2 - (EX)^2 = 1.8 - (1.2)^2 = 0.36.$$

例 2.16 设 X 的密度函数如例 2.11, 即

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

求 DX .

解 在例 2.11 中, 我们已算得 $EX = 1$, 又

$$\begin{aligned} EX^2 &= \int_{-\infty}^{+\infty} x^2 f(x) dx = \int_0^1 x^3 dx + \int_1^2 x^2 (2-x) dx \\ &= \left. \frac{x^4}{4} \right|_0^1 + \left. \left(\frac{2}{3} x^3 - \frac{x^4}{4} \right) \right|_1^2 = \frac{7}{6}, \end{aligned}$$

从而

$$DX = EX^2 - (EX)^2 = \frac{7}{6} - 1 = \frac{1}{6}.$$

例 2.17 X 为一随机变量, 方差存在, 令

$$l(C) = E(X - C)^2 \quad (2.34)$$

证明: 当且仅当 $C = EX$ 时, $l(C)$ 达到最小值, 此时最小值为 DX .

$$\begin{aligned} \text{证明} \quad l(C) &= E(X - C)^2 = E[(X - EX) + (EX - C)]^2 \\ &= E[(X - EX)^2 + 2(X - EX)(EX - C) + (EX - C)^2] \\ &= E(X - EX)^2 + (EX - C)^2 \\ &= DX + (EX - C)^2 \end{aligned}$$

显然, 当且仅当 $C = EX$ 时, 最后一个不等式的等号成立, 故有 $l(C)$ 在 $C = EX$ 时达到最小值, 且最小值为 DX .

例 2.17 实际上给出了数学期望的预测含义: 随机变量 X 的取值是不确定的, 如果我们企图用一个常数 C 来对 X 作点值预测, 未来 X 的实际取值与该预测值 C 存在偏差 $X - C$, 平均意义下的偏差程度用 $E(X - C)^2$ (通常称之为均方误差) 来衡量, 人们当然希望预测的偏差程度越小越好, 于是最好的预测 C 应该使得 $E(X - C)^2$ 达到最小, 于是例 2.16 实际上指出 $C = EX$ 是均方误差最小意义下 X 的最好的点值预测, 且最小的均方误差为 DX .

六、随机变量的矩与切比雪夫不等式

数学期望和方差可以纳入到一个更一般的概念范畴之中, 那就是随机变量的矩.

定义 2.9 X 为一随机变量, $k > 0$ 如果 EX^k 存在 (即 $E|X|^k < \infty$), 则称 EX^k 为 X 的 k 阶原点矩, 称 $E|X|^k$ 为 X 的 k 阶绝对矩.

定理 2.2 随机变量 X 的 t 阶矩存在, 则其 s 阶矩 ($0 < s \leq t$) 也存在.

证明 设 X 为连续型 (离散型类似), 其密度函数为 $f(x)$, 则

$$\begin{aligned} E|X|^k &= \int_{-\infty}^{-1} |x|^k f(x) dx + \int_{-1}^{\infty} |x|^k f(x) dx \\ &= \int_{-\infty}^{-1} f(x) dx + \int_{-1}^{\infty} |x|^k f(x) dx \\ &= P\{|X| \geq 1\} + E|X|^k I_{\{|X| < 1\}} \end{aligned}$$

推论 设 k 为正整数, 如果 EX^k 存在, 则 $E(X+C)^k$ 存在, 特别地, $E(X-EX)^k$ 存在.

定义 2.10 X 为一随机变量, k 为正整数, 如果 EX^k 存在, 则称 $E(X-EX)^k$ 为 X 的 k 阶中心矩, 称 $E|X-EX|^k$ 为 X 的 k 阶绝对中心矩.

显然数学期望 EX 是 X 的一阶原点矩, 方差 $DX = E(X-EX)^2$ 是 X 的二阶中心矩. 而且, 根据定理 2.2 及其推论知: 如果 $EX^2 < \infty$, 则 X 的数学期望和方差均存在.

接下来, 我们介绍一类矩的不等式

定理 2.3 设 $h(x)$ 是 x 的一个非负函数, X 是一个随机变量, 且 $Eh(X)$ 存在, 则对任意 $\varepsilon > 0$, 有

$$P\{h(X) \geq \varepsilon\} \leq \frac{Eh(X)}{\varepsilon}. \quad (2.35)$$

证明 这里仅证明连续型, 离散型类似可证.

设 X 的密度函数为 $f(x)$, 则

$$\begin{aligned} Eh(X) &= \int_{-\infty}^{+\infty} h(x) f(x) dx \\ &= \int_{h(x) \geq \varepsilon} h(x) f(x) dx + \int_{h(x) < \varepsilon} h(x) f(x) dx \\ &\geq \int_{h(x) \geq \varepsilon} \varepsilon f(x) dx \\ &= \varepsilon \int_{h(x) \geq \varepsilon} f(x) dx \\ &= \varepsilon P\{h(X) \geq \varepsilon\} \end{aligned}$$

推论 1 (马尔可夫不等式) 设 X 的 k 阶矩存在, 即 $E|X|^k < \infty$, 则对任

意 $\epsilon > 0$ 有

$$P\{|X - EX| \geq \epsilon\} \leq \frac{E(X - EX)^2}{\epsilon^2} \quad (2.36)$$

推论 2 (切比雪夫不等式) 设 X 的方差存在, 则对任意 $\epsilon > 0$ 有

$$P\{|X - EX| \geq \epsilon\} \leq \frac{DX}{\epsilon^2} \quad (2.37)$$

推论 3 随机变量 X 的方差为 0 当且仅当存在一个常数 a , 使得 $P\{X = a\} = 1$.

证明 充分性显然, 下证必要性.

首先注意到

$$\{X - EX \geq 0\} = \bigcap_{n=1}^{\infty} \{X - EX \geq \frac{1}{n}\}$$

从而有:

$$P\{X - EX \geq 0\} = P\bigcap_{n=1}^{\infty} \{X - EX \geq \frac{1}{n}\} \\ = \lim_{n \rightarrow \infty} P\{X - EX \geq \frac{1}{n}\}$$

由于 $DX = 0$, 对一切 $n \geq 1$, 据切比雪夫不等式, 有

$$P\{X - EX \geq \frac{1}{n}\} \leq \frac{DX}{1/n^2} = 0$$

从而得:

$$P\{X - EX \geq 0\} = 0.$$

所以

$$P\{X - EX \leq 0\} = 1,$$

即推论 3 得证, 且其中常数 a 即为 EX .

§ 2.3 常用的离散型分布

一、退化分布

在所有分布中, 最简单的分布是退化分布. 一个随机变量 X 以概率 1 取某一常数, 即

$$P\{X = a\} = 1$$

则称 X 服从 a 处的退化分布. 由定理 2.3 的推论 3 知, X 服从退化分布的充要条件是 $DX = 0$. 且若 X 服从 a 处的退化分布, 则 $EX = a$.

退化分布之所以称为退化分布是因为其取值几乎是确定的, 即这样的随机

变量退化成了一个确定的常数.

二、两点分布

另一个简单分布是两点分布. 一个随机变量只有两个可能取值, 设其分布为:

$$P\{X = x_1\} = p, P\{X = x_2\} = 1 - p \quad (0 < p < 1) \quad (2.38)$$

则称 X 服从 x_1, x_2 处参数为 p 的两点分布. 容易求得

$$EX = px_1 + (1 - p)x_2 \quad (2.39)$$

$$DX = p(1 - p)(x_1 - x_2)^2 \quad (2.40)$$

特别地, 如果 X 服从 $x_1 = 1, x_2 = 0$ 处的参数为 p 两点分布, 即

$$P\{X = 1\} = p, P\{X = 0\} = 1 - p \quad (0 < p < 1) \quad (2.41)$$

通常简称为 X 服从参数为 p 的两点分布或称 X 服从参数 p 的 0—1 分布, 也称 X 是参数为 p 的伯努利随机变量. 此时

$$EX = p, DX = p(1 - p) \quad (2.42)$$

在实际中, 一个两点分布的随机变量通常根据某试验中一特定事件 A 的发生与否构造出来. 例如, 设 $P(A) = p, P(\bar{A}) = 1 - p$, 则随机变量

$$X(\omega) = \begin{cases} x_1, & A \text{ (即 } A \text{ 发生)} \\ x_2, & \bar{A} \text{ (即 } A \text{ 不发生)} \end{cases} \quad (2.43)$$

便服从 x_1, x_2 处参数为 p 的两点分布. 特别地在一次试验中, 观察 A 是否发生, 记 A 发生的次数为 X , 则 X 要么取值为 1, 要么取值为 0, 于是 X 服从参数为 p 的 0—1 分布.

三、 n 个点上的均匀分布

有一类特殊的随机变量, 它共有 n 个不同的可能取值, 且取每一个值的可能性相同, 即有

$$P\{X = x_i\} = \frac{1}{n}, \quad i = 1, 2, \dots, n, \quad (2.44)$$

则称 X 服从 n 个点 $\{x_1, x_2, \dots, x_n\}$ 上的均匀分布.

容易算得:

$$EX = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \quad (2.45)$$

$$DX = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2.46)$$

我们可以将古典概型与均匀分布联系起来. 古典概型中, 试验共有 n 个不同的可能结果, 且每个结果出现的可能性相同, 设 $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, 则 $P\{\omega_i\} =$

$\frac{1}{n}$, $i = 1, 2, \dots, n$. 如果随机变量 X 是 上的一一对应的函数, 那么 X 便服从均匀分布. 一个简单的例子是: 设 X 表示投掷一枚均匀的骰子, 出现的点数, 此时 $\Omega = \{1, 2, \dots, 6\}$, 令

$$X(\omega) = \omega,$$

则 X 服从 $\{1, 2, \dots, 6\}$ 上的均匀分布.

四、二项分布

在 n 重伯努利试验中, 每次试验中事件 A 发生的概率为 p ($0 < p < 1$), 记 X 为 n 次试验中事件 A 发生的次数, 则 X 的可能取值为 $0, 1, 2, \dots, n$. 且对每一 k , $0 \leq k \leq n$, 事件 $\{X = k\}$ 即为“在 n 次试验中事件 A 恰好发生 k 次”, 根据伯努利概型, 有

$$P\{X = k\} = C_n^k p^k (1-p)^{n-k} \quad k = 0, 1, 2, \dots, n. \quad (2.47)$$

一般地, 如果一个随机变量 X 的概率分布由 (2.47) 给出, 则称 X 服从参数为 n, p 的二项分布, 并记作 $X \sim b(n, p)$, 且记 $b(k; n, p) = C_n^k p^k (1-p)^{n-k}$. 事实上 $b(k; n, p)$, $k = 0, 1, 2, \dots, n$ 是二项式 $(p + q)^n$ 的展开式中各项 (其中 $q = 1 - p$), 这也正是上述分布称为“二项分布”的原因.

显然, 当 $n = 1$ 时, 二项分布 $b(1, p)$ 实际上就是参数为 p 的 0—1 分布.

例 2.18 一个袋子中装有 N 个球, 其中 N_1 个白球, N_2 个黑球 ($N_1 + N_2 = N$), 每次从中任取一球, 查看完其颜色后再放回去, 一共取 n 次, 求取到的白球数 X 的分布.

解 每次取球视为一次试验, n 次取球视为 n 重贝努利试验, 每次取球, 取到白球的概率为 $p = \frac{N_1}{N}$, 故 $X \sim b(n, \frac{N_1}{N})$, 其分布为:

$$P\{X = k\} = C_n^k \left(\frac{N_1}{N}\right)^k \left(\frac{N_2}{N}\right)^{n-k} \quad 0 \leq k \leq n \quad (2.48)$$

接下来, 我们来求二项分布的数学期望和方差

$$\begin{aligned} EX &= \sum_{k=0}^n kb(k; n, p) = \sum_{k=1}^n k p C_n^k p^{k-1} q^{n-k} \\ &= np \sum_{k=1}^n C_{n-1}^{k-1} p^{k-1} q^{(n-1)-(k-1)} \\ &= np \sum_{k=0}^{n-1} C_{n-1}^k p^k q^{(n-1)-k} \\ &= np \sum_{k=0}^{n-1} b(k; n-1, p) \\ &= np \end{aligned} \quad (2.49)$$

其中最后一个等式, 是因为 $b(k; n-1, p)$, $(0 \leq k \leq n-1)$ 是以 $n-1$ 和 p 为参数的二项分布. 类似的计算可得:

$$\begin{aligned}
 EX^2 &= \sum_{k=0}^n k^2 b(k; n, p) \\
 &= \sum_{k=0}^n [k(k-1) + k] b(k; n, p) \\
 &= \sum_{k=2}^n k(k-1) b(k; n, p) + \sum_{k=0}^n k b(k; n, p) \\
 &= n(n-1)p^2 \sum_{k=2}^n b(k-2; n-2, p) + np \\
 &= n(n-1)p^2 \sum_{k=0}^{n-2} b(k; n-2, p) + np \\
 &= n(n-1)p^2 + np = n^2 p^2 + npq \quad (2.50)
 \end{aligned}$$

于是

$$DX = EX^2 - (EX)^2 = n^2 p^2 + npq - n^2 p^2 = npq \quad (2.51)$$

五、几何分布

在独立重复试验中, 事件 A 发生的概率为 p , 设 X 为直到 A 发生为止所进行的试验的次数, 显然 X 的可能取值是全体正整数, 且由定理 1.4 知其分布为

$$P\{X=k\} = q^{k-1}p = g(k, p) \quad k=1 \quad (2.52)$$

由于 $g(k, p) = q^{k-1}p$ 是一个几何数列 (或称等比数列), 因而将以 (2.52) 为概率分布的随机变量称为服从参数为 p 的几何分布.

几何分布的数学期望和方差的计算可利用无穷级数求和的常规方法——转化为求幂级数的和函数, 下面给出计算结果, 计算过程留作练习.

$$EX = \sum_{n=1}^{\infty} npq^{n-1} = \frac{1}{p} \quad (2.53)$$

$$EX^2 = \sum_{n=1}^{\infty} n^2 p q^{n-1} = \frac{2q}{p^2} + \frac{1}{p} \quad (2.54)$$

$$DX = EX^2 - (EX)^2 = \frac{q}{p^2} \quad (2.55)$$

例 2.19 设 X 服从几何分布, 则对任何两个正整数 m, n , 有

$$P\{X > m+n | X > m\} = P\{X > n\} \quad (2.56)$$

证明 由 $P\{X > m+n | X > m\} = \frac{P\{X > m+n\}}{P\{X > m\}}$, 据 (2.52) 知

$$P\{X > m\} = \sum_{k=m+1}^{\infty} q^{k-1}p = q^m \sum_{j=1}^{\infty} q^{j-1}p = q^m$$

同理, 有

$$P\{X > m+n\} = q^{m+n}, P\{X > n\} = q^n$$

于是得:

$$P\{X > m+n | X > m\} = \frac{q^{m+n}}{q^m} = q^n = P\{X > n\}.$$

式 (2.56) 通常称为几何分布的无记忆性, 意指几何分布对过去的 m 次失败的信息在后面的计算中被遗忘了. 事实上还可以证明: 一个取正整数值的随机变量, 如果具有无记忆性, 即满足 (2.56), 则一定服从几何分布. 可见, 无记忆性实际上是几何分布的一个特征性质.

六、超几何分布

一个袋子中装有 N 个球, 其中 N_1 个白球, N_2 个黑球 ($N = N_1 + N_2$), 从中不放回地抽取 n 个球, X 表示取到白球的数目, 那么 X 的分布容易根据古典概型计算 (参考例 1.12) 得到:

$$P\{X = k\} = \frac{C_{N_1}^k C_{N_2}^{n-k}}{C_N^n} \quad 0 \leq k \leq n \quad (2.57)$$

这里约定: 当 $a < b$ 时, $C_a^b = 0$. 以 (2.57) 为概率分布的随机变量通常称为服从超几何分布.

比较例 2.18 与上述取球问题, 惟一的区别是: 前者是放回的, 后者是不放回的. 在实际中, 当 N 很大时, 且 N_1 和 N_2 均较大, 而 n 相对很小时, 通常将不放回近似地当作放回来处理, 从而用二项分布 (见例 2.18) 来近似超几何分布, 即

$$\frac{C_{N_1}^k C_{N_2}^{n-k}}{C_N^n} \approx C_n^k \left(\frac{N_1}{N}\right)^k \left(\frac{N_2}{N}\right)^{n-k} \quad (2.58)$$

这一近似关系的严格数学表述是: 当 $N \rightarrow \infty$ 时, $N_1 \rightarrow pN$, $N_2 \rightarrow (1-p)N$, 且 $\frac{N_1}{N} \rightarrow p$, $\frac{N_2}{N} \rightarrow 1-p$, 则对任意给定的 n 和 k , 有

$$\lim_{N \rightarrow \infty} \frac{C_{N_1}^k C_{N_2}^{n-k}}{C_N^n} = C_n^k p^k (1-p)^{n-k} \quad (2.59)$$

最后, 我们给出超几何分布的数学期望和方差, 略去其计算过程:

$$EX = n \cdot \frac{N_1}{N}, \quad (2.60)$$

$$D(X) = n \cdot \frac{N_1}{N} \cdot \frac{N_2}{N} \cdot \frac{N-n}{N-1}. \quad (2.61)$$

七、泊松 (Poisson) 分布

如果一个随机变量 X 的概率分布为:

$$P\{X=k\} = \frac{\lambda^k}{k!} e^{-\lambda} \quad (k=0, 1, 2, \dots) \quad (2.62)$$

其中 $\lambda > 0$ 为参数, 则称 X 服从参数为 λ 的泊松分布, 记作 $X \sim P(\lambda)$.

首先我们注意到, 概率分布 (2.62) 中 $\frac{\lambda^k}{k!}$ 实际上是 e^λ (关于 λ 的函数) 的幂级数展开式的一般项, 因而有

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

这说明 (2.62) 确实是一个概率分布. 此外容易算得:

$$EX = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = \lambda. \quad (2.63)$$

同样方法可得:

$$EX^2 = \lambda^2 + \lambda.$$

从而有

$$DX = EX^2 - (EX)^2 = \lambda. \quad (2.64)$$

泊松分布是实际中经常遇到的一类分布, 比如, 电话交换台在一给定时间内收到用户的呼叫次数, 售票口到达的顾客人数, 保险公司在一定时期内被索赔的次数等等, 均可近似地用泊松分布来描述.

例 2.20 某商店根据过去的销售记录知道某种商品每月的销售量可以用参数为 $\lambda = 10$ 的泊松分布来描述, 为了以 95% 以上的概率保证不脱销, 问商店在月底应存多少件该种商品 (设只在月底进货)?

解 设该商店每月销售该商品的件数为 X , 月底存货为 a 件, 则当 $X \leq a$ 时就不会脱销, 据题意, 要求 a 使得

$$P\{X \leq a\} \geq 0.95$$

由于已知 X 服从参数为 $\lambda = 10$ 的泊松分布, 上式于是即为:

$$\sum_{k=0}^a \frac{10^k}{k!} e^{-10} \geq 0.95$$

由附录的泊松分布表知

$$\begin{aligned} \sum_{k=0}^{14} \frac{10^k}{k!} e^{-10} &= 0.9166 < 0.95 \\ \sum_{k=0}^{15} \frac{10^k}{k!} e^{-10} &= 0.9513 > 0.95 \end{aligned}$$

于是, 这家商店只要在月底保证存货不低于 15 件就能以 95% 以上的概率保证下个月该种商品不会脱销.

下面的定理给出了二项分布与泊松分布间的近似关系.

定理 2.4 (泊松定理) 在 n 重伯努利试验中, 事件 A 在每次试验中发生的概率为 p_n (注意这与试验的次数 n 有关), 如果 $n \rightarrow \infty$ 时, $np_n \rightarrow \lambda$ ($\lambda > 0$ 为常

数), 则对任意给定的 k , 有:

$$\lim_{n \rightarrow \infty} b(k, n, p_n) = \frac{e^{-\lambda} \lambda^k}{k!} \quad (2.65)$$

该定理的证明略去. 由该定理, 我们可以将二项分布用泊松分布来近似: 当二项分布 $b(n, p)$ 的参数 n 很大, 而 p 很小时, 可以将它用参数为 $\lambda = np$ 的泊松分布来近似, 即有

$$b(k; n, p) \approx \frac{(np)^k}{k!} e^{-np} \quad (2.66)$$

例 2.21 纺织厂女工照顾 800 个纺锭, 每一纺锭在某一短时间内发生断头的概率为 0.005 (设短时间内最多只发生一次断头), 求在这段时间内总共发生的断头次数超过 2 的概率.

解 设 X 为 800 个纺锭在该段时间内发生的断头次数, 则 $X \sim b(800, 0.005)$, 它可近似于参数为 $\lambda = 800 \times 0.005 = 4$ 的泊松分布, 从而有

$$P\{0 \leq X \leq 2\} = \sum_{k=0}^2 P\{X = k\} = \sum_{k=0}^2 b(k; 800, 0.005)$$

$$= \sum_{k=0}^2 \frac{4^k}{k!} e^{-4} = e^{-4} \left(1 + 4 + \frac{16}{2!} \right) \approx 0.2381$$

从而

$$P\{X > 2\} = 1 - P\{0 \leq X \leq 2\} = 1 - 0.2381 = 0.7619.$$

§ 2.4 常用的连续型分布

一、均匀分布

均匀分布是连续型分布中最简单的一种分布, 它用来描述一个随机变量在一个区间上取每一个值的可能性均等的分布规律, 它是离散型情形 n 个点上的均匀分布在连续型情形的推广. 在离散型情形, n 个点的均匀分布描述为随机变量在 n 个点中取值, 取每一点的概率相同. 在连续型情形, $[a, b]$ 上的均匀分布, 则描述为该区间上每一点的概率密度相同, 或等价地描述为取值落在该区间中每一个子区间上的概率与子区间的长度成正比, 在例 2.5 及例 2.8 中, 我们实际上已给出了详细的讨论. 在第一章中讨论的几何概型, 如果用随机变量 X 来描述落点的位置, 则 X 服从均匀分布. 下面明确给出均匀分布的定义.

一个随机变量 X , 如果其密度函数为:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{其他} \end{cases} \quad (2.67)$$

则称 X 服从 $[a, b]$ 上的均匀分布, 记作 $X \sim U[a, b]$.

$[a, b]$ 上的均匀分布的惟一特征是密度函数在 $[a, b]$ 以外为 0, 而在 $[a, b]$ 上为常数. 根据密度函数的性质:

$$\int_{-\infty}^{+\infty} f(x) dx = \int_a^b f(x) dx = 1$$

这个常数必然是 $\frac{1}{b-a}$.

由 (2.67) 容易导出均匀分布的分布函数为:

$$F(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \leq x \leq b, \\ 1, & x > b, \end{cases} \quad (2.68)$$

且可算得其数学期望和方差分别为

$$EX = \frac{a+b}{2}, \quad (2.69)$$

$$DX = \frac{(b-a)^2}{12}. \quad (2.70)$$

二、指数分布

指数分布通常用来描述对某一事件发生的等待时间, 比如, 乘客在公共汽车站等车的时间, 灯泡的使用寿命 (等待用坏的时间), 电话交换台收到两次呼叫的时间间隔. 在离散型分布中, 我们知道, 几何分布用来描述伯努利试验中, 直到某事件 A 发生为止共进行的试验次数, 如果将每次试验视为经历一个单位时间, 那么直到事件 A 发生为止进行的试验的次数可视为直到 A 发生为止的等待时间 (离散时间). 在这个意义上, 指数分布可视为离散型情形的几何分布在连续型情形的推广. 下面给出指数分布的定义.

一个随机变量 X , 如果其密度函数为

$$f(x) = \begin{cases} e^{-\lambda x} & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (2.71)$$

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

由 (2.71) 容易导出 X 的分布函数为

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0, & x < 0. \end{cases} \quad (2.72)$$

且可算得其数学期望和方差分别为:

$$EX = \frac{1}{\lambda}, \quad (2.73)$$

$$DX = \frac{1}{2}. \quad (2.74)$$

与几何分布类似,无记忆性是连续型随机变量变量的指数分布的特征性质,即有下列结果:

定理 2.5 非负连续型随机变量 X 服从参数为 λ 的指数分布的充要条件是:对任意正实数 r 和 s , 有

$$P\{X > r + s | X > s\} = P\{X > r\} \quad (2.75)$$

该定理的必要性的证明留作练习,充分性的证明超出了本书的要求,略去.

例 2.22 某元件的寿命 X 服从指数分布,已知其平均寿命为 1 000 小时,求 3 个这样的元件使用 1 000 小时,至少已有一个损坏的概率.

解 由题设知, $EX = 1000$ 小时,于是该指数分布的参数为

$$\lambda = \frac{1}{EX} = \frac{1}{1000}.$$

从而 X 的分布函数为

$$F(x) = \begin{cases} 1 - e^{-\frac{x}{1000}}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

由此得知:

$$P\{X > 1000\} = 1 - P\{X \leq 1000\} = 1 - F(1000) = e^{-1}$$

各元件的寿命是否超过 1000 小时是独立的,于是 3 个元件使用 1000 小时都未损坏的概率为 e^{-3} ,从而至少有一个已损坏的概率为 $1 - e^{-3}$.

三、正态分布

一个连续型随机变量 X , 如果其密度函数为

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x < +\infty) \quad (2.76)$$

其中 μ 为常数,且 $\sigma > 0$. 则称 X 服从参数为 μ 和 σ^2 的正态分布,记作 $X \sim N(\mu, \sigma^2)$.

首先,根据泊松积分:

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi} \quad (2.77)$$

不难验证 $\int_{-\infty}^{+\infty} f(x) dx = 1$,并可计算 $N(\mu, \sigma^2)$ 的数学期望和方差(留作练习)

$$EX = \mu \quad DX = \sigma^2 \quad (2.78)$$

可见,正态分布的两个参数实际上分别为其数学期望和方差.

其次,让我们考察一下正态分布的密度函数 $f(x)$ 的特征.如图 2.4 所示,

$\varphi(x)$ 具有钟型的图象, 且以 x 轴为渐近线; 关于 $x = \mu$ 对称, 在 $x = \mu$ 处达到函数最大值 $\frac{1}{\sqrt{2\pi}\sigma}$.

当 $\mu = 0, \sigma^2 = 1$ 时, 即 $X \sim N(0, 1)$, 称 X 服从标准正态分布, 其密度函数记作 $\phi(x)$, 即

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (2.79)$$

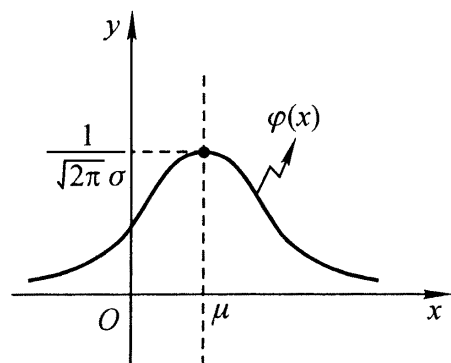


图 2.4 正态分布的密度函数

正态分布的“钟型”特征与实际中很多随机变量的“中间大, 两头小”的分布规律相吻合, 比如考察一群人的身高, 其高度作为一个随机变量, 分布的特点是, 在平均身高附近的人较多, 特别高和特别矮的人较少. 一个班的一次考试成绩, 测量误差等均有类似的特征. 正态分布是概率论中最重要的分布, 高斯在研究误差理论时曾用它来刻画误差, 所以很多著作中亦称之为高斯分布. 进一步的理论研究表明, 一个变量如果受到大量的独立因素的影响 (无主导因素), 则它一般服从正态分布, 这一问题将在下一章进行讨论.

1. 正态分布的分布函数.

对许多随机变量, 利用密度函数可求出分布函数的表达式. 然而, 由于正态分布的密度函数 $\phi(x)$ 的原函数没有初等表达式, 因而其分布函数 (记作 $\Phi(x)$)

$$\Phi(x) = \int_{-\infty}^x \phi(t) dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \quad (2.80)$$

不能表示为初等函数. 对于任意给定 x , 我们需要利用数值计算方法来求上述广义积分的近似值, 然而对每一个具体的正态分布及每个给定 x 都进行近似计算显然极为不便, 解决这一问题的方法是, 将标准正态分布的分布函数 (记作 $\Phi(x)$) 在各点的值计算出来制成表, 在实际计算时可直接查表, 而一般正态分布, 则可经过适当变换转化为标准正态分布来计算.

2. 标准正态分布表.

在附录中列出了标准正态分布的密度函数值表和分布函数值表, 但表中只列出 $x \geq 0$ 时 $\phi(x)$ 和 $\Phi(x)$ 的值, 这是因为由正态分布的对称性 (严格地讲, 密度函数 $\phi(x)$ 关于 $x = 0$ 对称!) 可以导出 $\phi(x)$ 和 $\Phi(x)$ 在 $x < 0$ 时的值.

对于 $\phi(x)$ 而言, 直接由其对称性有

$$\phi(-x) = \phi(x)$$

因而, 当 $x < 0$ 时, $\phi(x) = \phi(-x)$, 在表中查 $\phi(-x)$ 即得 $\phi(x)$.

对于 $\Phi(x)$, 由于 $\phi(x)$ 关于 $x = 0$ 对称, 因而有:

$$\Phi(-x) = \int_{-\infty}^{-x} \phi(t) dt = \int_{+\infty}^{-x} \phi(t) dt = 1 - \int_{-\infty}^x \phi(t) dt = 1 - \Phi(x) \quad (2.81)$$

或写作

$$\Phi(x) + \Phi(-x) = 1 \quad (2.82)$$

特别地, 有 $\Phi(x) = 0.5$, 当 $x < 0$ 时, $\Phi(x) = 1 - \Phi(-x)$, 查表得 $\Phi(-x)$, 即可得 $\Phi(x)$.

例 2.23 设 $X \sim N(0, 1)$, (1) 求 $P\{X \leq 1.96\}$, $P\{X \leq -1.96\}$, $P\{-1.96 \leq X \leq 1.96\}$, $P\{-1 < X \leq 2\}$. (2) 已知 $P\{X \leq a\} = 0.7019$, $P\{-a \leq X \leq b\} = 0.9242$, $P\{X \leq c\} = 0.2981$, 求 a, b, c .

解 (1) 直接查表可得

$$P\{X \leq 1.96\} = \Phi(1.96) = 0.975,$$

根据 $\Phi(x)$ 的对称性, 有

$$P\{X \leq -1.96\} = \Phi(-1.96) = 1 - \Phi(1.96) = 1 - 0.975 = 0.025,$$

$$\begin{aligned} P\{-1.96 \leq X \leq 1.96\} &= P\{-1.96 \leq X \leq 1.96\} = \Phi(1.96) - \Phi(-1.96) \\ &= 2\Phi(1.96) - 1 = 2 \times 0.975 - 1 = 0.95, \end{aligned}$$

$$\begin{aligned} P\{-1 < X \leq 2\} &= \Phi(2) - \Phi(-1) = \Phi(2) - [1 - \Phi(1)] = \Phi(2) + \Phi(1) - 1 \\ &= 0.97725 + 0.8413 - 1 = 0.81855, \end{aligned}$$

(2) 直接查表可得 $a = 0.53$, 由

$$P\{-a \leq X \leq b\} = 2\Phi(b) - 1 = 0.9242,$$

得

$$\Phi(b) = \frac{1}{2}(1 + 0.9242) = 0.9621$$

查表即得 $b = 1.78$.

由于 $P\{X \leq c\} = 0.2981 < 0.5$, 所以 $c < 0$, 根据对称性, 有

$$\Phi(-c) = 1 - \Phi(c) = 0.7019,$$

查表得 $-c = 0.53$, $c = -0.53$.

3. 一般正态分布与标准正态分布的关系.

我们先看一个有关正态分布的一般性结论.

定理 2.6 设 $X \sim N(\mu, \sigma^2)$, $Y = aX + b$, a, b 为常数, 且 $a \neq 0$, 则 $Y \sim N(a\mu + b, a^2\sigma^2)$

证明 记 Y 的分布函数为 $F_Y(x)$, 密度函数为 $f_Y(x)$, X 的分布函数为 $F_X(x)$, 密度函数为 $f_X(x)$, 则有:

$$F_Y(x) = P\{Y \leq x\} = P\{aX + b \leq x\} \quad (2.83)$$

当 $a > 0$ 时,

$$F_Y(x) = P\left\{X \leq \frac{x-b}{a}\right\} = F_X\left(\frac{x-b}{a}\right) \quad (2.84)$$

$$f_Y(x) = F'_Y(x) = \frac{1}{a} f_X\left(\frac{x-b}{a}\right) = \frac{1}{a} f_X\left(\frac{x-b}{a}\right) \quad (2.85)$$

当 $a < 0$ 时,

$$F_Y(x) = P\left\{X \leq \frac{x-b}{a}\right\} = 1 - \frac{x-b}{a} \quad (2.86)$$

$$f_Y(x) = F_Y'(x) = -\frac{1}{a} \frac{x-b}{a} = -\frac{1}{a} \frac{x-b}{a} \quad (2.87)$$

综上, 有

$$\begin{aligned} f_Y(x) &= \frac{1}{|a|} \frac{x-b}{a} = \frac{1}{2|a|} e^{-\frac{\left(\frac{x-b}{a} - \mu\right)^2}{2}} \\ &= \frac{1}{2|a|} e^{-\frac{[x - (a\mu + b)]^2}{2a^2}} \end{aligned} \quad (2.88)$$

因而 $Y \sim N(a\mu + b, a^2)$.

推论 1 如果 $X \sim N(\mu, \sigma^2)$, 则 $\frac{X-\mu}{\sigma} \sim N(0, 1)$.

通常称为 X 的标准化.

推论 2 $X \sim N(\mu, \sigma^2)$ 的充要条件是存在一个随机变量 $Z \sim N(0, 1)$, 使得 $X = Z\sigma + \mu$

推论 3 设 $X \sim N(\mu, \sigma^2)$, $F(x)$, $f(x)$ 分别为其分布函数与密度函数, $\Phi(x)$, $\phi(x)$ 是标准正态分布的分布函数和密度函数, 则有

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \quad (2.89)$$

$$f(x) = \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right) \quad (2.90)$$

推论 1 和推论 2 留给读者作为练习. 推论 3 的证明, 可利用推论 2 及定理 2.6 的证明中的 (2.84) 和 (2.85).

4. 一般正态分布的概率计算.

前面讨论了一般正态分布与标准正态分布的关系, 特别是定理 2.6 的推论 1, 为一般正态分布的概率计算提供了有效的途径. 一般地, 对于一般正态分布有关问题, 尤其是概率计算, 我们都可以转化为标准正态来解决, 下面通过一些例子加以说明.

例 2.24 已知 $X \sim N(8, 0.5^2)$, 求 (1) $\Phi(9)$, $\Phi(7)$; (2) $P\{7.5 \leq X \leq 10\}$; (3) $P\{X \leq 8\}$; (4) $P\{X \leq 9\}$

解 (1) $\Phi(9) = P\{X \leq 9\} = P\left\{\frac{X-8}{0.5} \leq \frac{9-8}{0.5}\right\} = P\left\{\frac{X-8}{0.5} \leq 2\right\}$
 $= \Phi(2) = 0.97725;$

$\Phi(7) = P\{X \leq 7\} = P\left\{\frac{X-8}{0.5} \leq \frac{7-8}{0.5}\right\} = P\left\{\frac{X-8}{0.5} \leq -2\right\}$
 $= \Phi(-2) = 1 - \Phi(2) = 0.02275;$

$$\begin{aligned}
 (2) P\{7.5 \leq X \leq 10\} &= P\left\{\frac{7.5-8}{0.5} \leq \frac{X-8}{0.5} \leq \frac{10-8}{0.5}\right\} = P\{-1 \leq \frac{X-8}{0.5} \leq 4\} \\
 &= \Phi(4) - \Phi(-1) = \Phi(4) + \Phi(1) - 1 \\
 &= 0.946833 + 0.8413 - 1 = 0.8413;
 \end{aligned}$$

$$\begin{aligned}
 (3) P\{8 \leq X \leq 11\} &= P\left\{\frac{8-8}{0.5} \leq \frac{X-8}{0.5} \leq \frac{11-8}{0.5}\right\} = P\{0 \leq \frac{X-8}{0.5} \leq 6\} \\
 &= \Phi(6) - \Phi(0) = 1 - 0.5 = 0.5;
 \end{aligned}$$

$$\begin{aligned}
 (4) P\{9 \leq X \leq 9.5\} &= P\{8.5 \leq X \leq 9.5\} = P\left\{-1 \leq \frac{X-8.5}{0.5} \leq 1\right\} \\
 &= \Phi(1) - \Phi(-1) = 0.8413 - 0.1587 = 0.6826.
 \end{aligned}$$

例 2.25 某种型号电池的寿命 X 近似服从正态分布 $N(\mu, \sigma^2)$, 已知其寿命在 250 小时以上的概率和寿命不超过 350 小时的概率均为 92.36%, 为使其寿命在 $\mu - x$ 和 $\mu + x$ 之间的概率不小于 0.9, x 至少为多大?

解 由 $P\{X > 250\} = P\{X < 350\}$, 根据密度函数关于 $x = \mu$ 对称, 有 $\mu = \frac{250+350}{2} = 300$, 又由

$$P\{X < 350\} = P\left\{\frac{X-300}{\sigma} < \frac{350-300}{\sigma}\right\} = \Phi\left(\frac{50}{\sigma}\right) = 0.9236$$

查表得 $\frac{50}{\sigma} = 1.43$, 于是 $\sigma = 35$. 故 $X \sim N(300, 35^2)$, 又

$$P\{\mu - x \leq X \leq \mu + x\} = P\left\{-\frac{x}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{x}{\sigma}\right\} = 2\Phi\left(\frac{x}{\sigma}\right) - 1 \geq 0.9,$$

即 $\Phi\left(\frac{x}{\sigma}\right) \geq 0.95$, 查表得 $\frac{x}{\sigma} = 1.645$, 于是 $x = 1.645 \times 35 = 57.58$.

§ 2.5 随机变量函数的分布

一、随机变量的函数

在上一节中, 我们在讨论正态分布时, 对服从 $N(\mu, \sigma^2)$ 的随机变量 X , 引进了线性变换 $Y = aX + b$, $a \neq 0$. 事实上, Y 是随机变量 X 的函数, 对 X 的每一个取值, Y 有一个唯一取值与之对应. 由于 X 是随机变量, 其取值是不确定的, 因而 Y 的取值也随之不确定, 即 Y 也是一个随机变量.

一般地, 如果存在一个函数 $g(x)$, 使得随机变量 X, Y 满足:

$$Y = g(X) \quad (2.91)$$

则称随机变量 Y 是随机变量 X 的函数. 在微积分中, 也讨论变量间的函数关系, 但在那里, 我们主要研究函数关系中的确定性特征, 比如因变量随自变量变化而变化的变化率——导数. 在概率论中, 我们主要研究的是随机变量函数的随机性特征, 即由自变量的统计规律出发研究因变量的统计规律性. 如何从自变

量 X 的统计规律导出其函数 $Y = g(X)$ 的统计规律呢? 一般地, 对任意区间 (或区间的并) B , 令 $C = \{x \in \mathbb{R} : g(x) \in B\}$, 则

$$\{Y \in B\} = \{g(X) \in B\} = \{X \in C\} \quad (2.92)$$

从而

$$P\{Y \in B\} = P\{g(X) \in B\} = P\{X \in C\} \quad (2.93)$$

(2.93) 说明, X 的统计规律确实决定了 Y 的统计规律.

例 2.26 设 X 是一随机变量, 且 $Y = X^2$, 则对任意 $x \geq 0$, 有

$$P\{Y \leq x\} = P\{X^2 \leq x\} = P\{-\sqrt{x} \leq X \leq \sqrt{x}\} \quad (2.94)$$

$$P\{Y = x\} = P\{X^2 = x\} = P(\{X = \sqrt{x}\} \cup \{X = -\sqrt{x}\}) \quad (2.95)$$

为对随机变量 X 的函数 Y 的统计规律进行完整描述, 我们希望能从 X 的分布出发导出 Y 的分布, (2.93) 式实际上为此提供了可行的途径. 本节仍然只对离散型和连续型这两种情形进行讨论.

二、离散型随机变量函数的分布

离散型随机变量 X 的函数 $Y = g(X)$ 显然还是离散型随机变量, 因此, 我们希望由 X 的概率分布出发导出 Y 的概率分布. 先看一个简单的例子.

例 2.27 设随机变量 X 的分布为

$$P\{X = -1\} = \frac{1}{4}, P\{X = 0\} = \frac{1}{2}, P\{X = 1\} = \frac{1}{4}.$$

求 $Y = X^2$ 的分布.

解 我们注意到 Y 的可能取值为 0, 1, 根据 (2.92), 有

$$P\{Y = 0\} = P\{X^2 = 0\} = P\{X = 0\} = \frac{1}{2}$$

$$\begin{aligned} P\{Y = 1\} &= P\{X^2 = 1\} = P(\{X = 1\} \cup \{X = -1\}) = P\{X = 1\} + P\{X = -1\} \\ &= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}. \end{aligned}$$

上面的例子虽然简单, 却反映了求离散型随机变量函数的概率分布的一般方法: 先根据自变量 X 的可能取值确定因变量 Y 的所有可能取值, 然后对 Y 的每一个可能取值 y_i , $i = 1, 2, \dots$, 确定相应的 $C_i = \{x_j \in \mathbb{R} : g(x_j) = y_i\}$, 于是正如 (2.92) 和 (2.93) 一样, 有:

$$\{Y = y_i\} = \{g(X) = y_i\} = \{X \in C_i\} \quad (2.96)$$

$$P\{Y = y_i\} = P\{X \in C_i\} = \sum_{x_j \in C_i} P\{X = x_j\} \quad (2.97)$$

从而求得 Y 的概率分布. 此外, 上述过程还说明 Y 的概率分布完全由 X 的概率分布所确定.

三、连续型随机变量函数的分布

一般说来, 连续型随机变量的函数不一定是连续型随机变量, 这时讨论随机变量函数的分布的目标是导出其分布函数, 这里我们主要讨论连续型随机变量的函数还是连续型的情形, 这时我们希望不仅求出随机变量函数的分布函数, 而且还希望求出其密度函数. 不过只要求出分布函数, 密度函数也就不成问题了. 我们将要说明导出随机变量函数的分布函数的一般方法, 并介绍一些典型的随机变量函数的例子.

在上一节中讨论正态分布时, 我们引入了服从正态分布 $N(\mu, \sigma^2)$ 的随机变量 X 的线性变换: $Y = aX + b$, 定理 2.6 导出了 Y 的分布仍是正态分布, 参数分别为 $a\mu + b$ 和 $a^2\sigma^2$, 实际上其推导过程体现了求连续型随机变量函数分布的一般方法.

一般地, 已知 X 的分布函数 $F_X(x)$ 或密度函数 $f_X(x)$, 为求 $Y = g(X)$ 的分布函数, 正如 (2.93) 所示,

$$F_Y(x) = P\{Y \leq x\} = P\{g(X) \leq x\} = P\{X \in C_x\} \quad (2.98)$$

其中 $C_x = \{t \in \mathbb{R} : g(t) \leq x\}$.

而 $P\{X \in C_x\}$ 往往可由 X 的分布函数 $F_X(x)$ 来表达或用其密度函数 $f_X(x)$ 的积分来表达:

$$P\{X \in C_x\} = \int_{C_x} f_X(t) dt \quad (2.99)$$

进而, Y 的密度函数, 可直接从 $F_Y(x)$ 导出.

例 2.28 设 X 是一个连续型随机变量, 其分布函数 $F(x)$ 是严格单调递增的, 则 $F(X)$ 服从 $[0, 1]$ 上的均匀分布.

证明 由于 $F(x)$ 是严格单调递增函数, 因而其反函数存在, 记作 $F^{-1}(x)$. 记 $Y = F(X)$ 的分布函数为 $F_Y(x)$, 则

$$F_Y(x) = P\{Y \leq x\} = P\{F(X) \leq x\}$$

由于 $F(x)$ 的值域为 $[0, 1]$, 故当 $x < 0$ 时, $F_Y(x) = P\{F(X) \leq x\} = 0$, 当 $x > 1$ 时, $F_Y(x) = P\{F(X) \leq x\} = 1$, 而当 $0 \leq x \leq 1$ 时, 由于 $F(x)$ 严格单调递增, 故有

$$F_Y(x) = P\{F(X) \leq x\} = P\{X \leq F^{-1}(x)\} = F(F^{-1}(x)) = x \quad \text{综上,}$$

$$F_Y(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

即 $Y = F(X)$ 服从 $[0, 1]$ 上的均匀分布.

例 2.29 中的结论对 $F(x)$ 非严格单调递增的情形同样成立, 证明略有不同, 这里就不讨论了.

例 2.29 ($\chi^2(1)$ 分布) 设 $X \sim N(0, 1)$, 求 $Y = X^2$ 的密度函数.

解 记 Y 的分布函数为 $F_Y(x)$, 则 $F_Y(x) = P\{Y \leq x\} = P\{X^2 \leq x\}$

显然, 当 $x < 0$ 时,

$$F_Y(x) = P\{X^2 \leq x\} = 0$$

当 $x \geq 0$ 时,

$$F_Y(x) = P\{X^2 \leq x\} = P\{-\sqrt{x} \leq X \leq \sqrt{x}\} = 2\Phi(\sqrt{x}) - 1$$

从而 $Y = X^2$ 的分布函数为

$$F_Y(x) = \begin{cases} 2\Phi(\sqrt{x}) - 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

于是其密度函数为

$$\begin{aligned} f_Y(x) = F'_Y(x) &= \begin{cases} \frac{1}{\sqrt{x}} \phi(\sqrt{x}) & x \geq 0 \\ 0 & x < 0 \end{cases} \\ &= \begin{cases} \frac{1}{2\sqrt{x}} e^{-\frac{x}{2}} & x \geq 0 \\ 0 & x < 0 \end{cases} \end{aligned} \quad (2.100)$$

以 (2.100) 为密度函数的随机变量称为服从 $\chi^2(1)$ 分布, 它是一类更广泛的分布——自由度为 n 的 χ^2 分布—— $\chi^2(n)$ 在 $n=1$ 时的特例. 关于 $\chi^2(n)$ 分布的细节将在第四章中给出.

例 2.30 (对数正态分布) 随机变量 X 称为服从参数为 μ, σ^2 的对数正态分布, 如果 $Y = \ln X$ 服从正态分布 $N(\mu, \sigma^2)$. 试求对数正态分布的密度函数.

解 由于 $Y = \ln X \sim N(\mu, \sigma^2)$, 等价地有 $X = e^Y$, $Y \sim N(\mu, \sigma^2)$, 于是, 当 $x > 0$ 时,

$$F_X(x) = P\{X \leq x\} = P\{e^Y \leq x\} = P\{Y \leq \ln x\} = \Phi\left(\frac{\ln x - \mu}{\sigma}\right)$$

当 $x \leq 0$ 时, 显然 $F_X(x) = 0$. 继而可得 X 的密度函数为

$$\begin{aligned} f_X(x) = F'_X(x) &= \begin{cases} \frac{1}{x} \phi\left(\frac{\ln x - \mu}{\sigma}\right) & x > 0 \\ 0 & x \leq 0 \end{cases} \\ &= \begin{cases} \frac{1}{2\sigma x} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} & x > 0 \\ 0 & x \leq 0 \end{cases} \end{aligned}$$

在实际中, 通常用对数正态分布来描述价格的分布, 特别是在金融市场的理论研究中, 如著名的期权定价公式——Black-Scholes 公式, 以及许多实证研究都用对数正态分布来描述金融资产的价格. 设某种资产当前价格为 P_0 , 考虑单期投资问题, 到期时该资产的价格为一个随机变量, 记作 P_1 , 设投资于该资产的连续复合收益率为 r , 则有

$$P_1 = P_0 e^r$$

从而

$$r = \ln \frac{P_1}{P_0} = \ln P_1 - \ln P_0$$

注意到 P_0 为当前价格, 是已知常数, 因而假设价格 P_1 服从对数正态分布实际上等价于假设连续复合收益率 r 服从正态分布.

例 2.30 (对数正态分布的矩) 设 X 服从参数为 μ 和 σ^2 的对数正态分布, 即 $Y = \ln X \sim N(\mu, \sigma^2)$, 则由 $X = e^Y$, 有

$$\begin{aligned} EX^k &= Ee^{kY} = \int_{-\infty}^{+\infty} e^{ky} \cdot \frac{1}{\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sigma} \exp \left[-\frac{(y-\mu)^2}{2\sigma^2} - \frac{2k\sigma^2(y-\mu) + k^2\sigma^4}{2\sigma^2} \right] + \\ &\quad k\mu + \frac{k^2\sigma^2}{2} dy = e^{k\mu + \frac{k^2\sigma^2}{2}} \end{aligned}$$

特别地, $k=1$ 时,

$$EX = e^{\mu + \frac{\sigma^2}{2}}$$

进而有,

$$DX = EX^2 - (EX)^2 = e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2} = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1).$$

习 题 二

(A)

1. 甲、乙两人分别拥有赌本 30 元和 20 元, 他们利用投掷一枚均匀硬币进行赌博, 约定如果出现正面, 甲赢 10 元、乙输 10 元. 如果出现反面, 则甲输 10 元、乙赢 10 元, 分别用随机变量表示投掷一次后甲、乙两人的赌本, 并求其概率分布和分布函数, 画出分布函数的图形.

2. 离散型随机变量 X 的概率分布为:

$$(1) P\{X=i\} = a2^i, \quad i=1, 2, \dots, 100;$$

$$(2) P\{X=i\} = 2a^i, \quad i=1, 2, \dots,$$

分别求 (1) (2) 中的 a 的值.

3. 设随机变量 X 的分布函数为

$$F(x) = \begin{cases} 0, & x < -5 \\ \frac{1}{5}, & -5 \leq x < -2 \\ \frac{3}{10}, & -2 \leq x < 0 \\ \frac{1}{2}, & 0 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$

求 X 的概率分布.

4. 一批产品共 10 件, 其中 7 件正品, 3 件次品, 每次从中任取一件, 求下面两种情形下, 直到取到正品为止所需抽取次数号的概率分布: (1) 每次取出后再放回去; (2) 每次取出后不放回.

5. 设 X 的分布函数如第 3 题所示, 求下列概率: $P\{X > -3\}$, $P\{-3 \leq X \leq 3\}$, $P\{X + 1 \leq 2\}$.

6. 设 X 的分布函数为

$$F(x) = \begin{cases} 0, & x \leq 0 \\ Ax^2, & 0 < x \leq 1 \\ 1, & x > 1 \end{cases}$$

求 A 及 $P\{0.5 < X \leq 0.8\}$.

7. 函数 $F(x) = \frac{1}{1+x^2}$ 是否可作为某一随机变量的分布函数, 如果

- (1) $-\infty < x < +\infty$; (2) $0 < x < +\infty$, 其他场合适当定义;
(3) $-\infty < x < 0$, 其他场合适当定义.

8. X 的分布函数分别如下, 判断它是否为连续型随机变量, 如果是, 则求其密度函数.

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

9. 已知连续型随机变量 X 的密度函数为

$$(1) f(x) = ae^{-ax}, \quad (2) f(x) = \begin{cases} x, & 0 \leq x < 1 \\ 2-x, & 1 \leq x < a \\ 0, & \text{其他} \end{cases}$$

求 a 及分布函数 $F(x)$, $P\{-1 < X \leq \frac{2}{2}\}$, $P\{\frac{2}{2} < X \leq \frac{2}{2}\}$, $P\{X > 1\}$.

10. 设随机变量 X 的密度函数关于 $x = \mu$ 对称, 证明其分布函数满足以下性质:

$$F(\mu + x) + F(\mu - x) = 1, \quad -\infty < x < +\infty.$$

11. 求第 2 题中 (2) 及第 3 题的随机变量 X 的数学期望 EX .

12. 已知投资某一项目的收益率 R 是一随机变量, 其分布为:

R	1%	2%	3%	4%	5%	6%
p_i	0.1	0.1	0.2	0.3	0.2	0.1

一位投资者在该项目上投资 10 万元, 求他预期获得多少收入? 收入的方差是多大?

13. 一张贴现债券 (期中不付息, 期末还本付息的债券) 承诺到期还本付息共偿还 1 100 元, 根据分析, 市场上同类债券的收益率为一随机变量, 记作 $K\%$, 设 K 的密度函数为:

$$f(x) = \begin{cases} \frac{1}{5}, & 0 \leq x \leq 5 \\ 0, & \text{其他} \end{cases}$$

求这张债券现在平均值多少钱.

14. 在连续型情形证明 (2.25).
15. 利用数学期望的性质, 证明方差的性质 (2.31) — (2.33).
16. 利用定理 2.3 的证明方法, 分别就离散型和连续型情形直接证明 “切比雪夫不等式” 即 (2.37).
17. 假设一厂家生产的每台仪器以 0.7 的概率可直接出厂, 以概率 0.3 需进一步调试, 经调试后以概率 0.8 可出厂, 0.2 的概率不合格而不能出厂, 现该厂生产 n ($n \geq 2$) 台仪器 (设仪器生产过程相互独立) 求 (1) 能出厂的仪器数 X 的分布列; (2) n 台仪器能全部出厂的概率; (3) 至少有两台不能出厂的概率; (4) 不能出厂的仪器数的期望和方差.
18. 自动生产线在调整后出现废品的概率为 P , 当在生产过程中出现废品时, 立即重新进行调整, 求在两次调整之间生产的合格品数 X 的分布列.
19. 设 X 服从泊松分布, 已知 $P\{X=1\}=2P\{X=2\}$, 求 $E X, D X, E X^2, P\{X=3\}$.
20. 在某公共汽车站甲、乙、丙三人分别等 1, 2, 3 路公共汽车, 设每个人等车时间 (单位: 分钟) 均服从 $[0, 5]$ 上的均匀分布, 求 3 人中至少有两人等车时间不超过 2 分钟的概率.
21. 某保险公司设置某一险种, 规定每一保单有效期为一年, 有效理赔一次, 每个保单收取保费 500 元, 理赔额为 20 000 元. 据估计每个保单索赔概率为 0.05, 设公司共卖出这种保单 800 个, 求该公司在该险种上获得的平均利润.
22. 由指数分布的密度函数导出指数分布的分布函数以及数学期望和方差.
23. 给出定理 2.5 中的必要性的证明.
24. 3 个电子元件并联成一个系统, 只有当 3 个元件损坏两个或两个以上时, 系统便报废, 已知电子元件的寿命服从参数为 $\frac{1}{1000}$ 的指数分布, 求系统的寿命超过 1 000 小时的概率.
25. 导出正态分布 $N(\mu, \sigma^2)$ 的数学期望和方差.
26. 设测量的随机误差 $X \sim N(0, 10^2)$, 试求 100 次独立重复测量, 至少有三次测量误差的绝对值大于 19.6 的概率, 并用泊松分布求 X 的近似值.
27. 已知电源电压 X 服从正态分布 $N(220, 25^2)$, 在电源电压处于 $X \leq 200V, 200V < X < 240V, X \geq 240V$ 三种情况下, 某电子元件损坏的概率分别 0.1, 0.01, 0.2, 试求该电子元件损坏的概率 X ; (2) 该电子元件损坏时, 电源电压在 $200V \sim 240V$ 的概率.
28. 某班数学考试成绩呈正态分布 $N(70, 100)$, 老师将最高成绩的 5% 定为优秀, 那么成绩为优秀的最少成绩是多少?
29. 测量一圆的半径 R , 其概率分布为

R	10	11	12	13
p_i	0.1	0.4	0.3	0.2

求圆的面积 S 和周长 L 的分布.

30. 设 X 服从 $[a, b]$ 上的均匀分布, 证明 $X + Y$ ($Y > 0$) 服从 $[a + Y, b + Y]$ 上的均匀分布.
31. 设 X 服从 $[-1, 1]$ 上的均匀分布, 求 X^2 的分布函数和密度函数.

32. X 服从参数为 1 的指数分布, 求 $Y = X + \quad (\quad > 0)$ 的分布函数和密度函数.
33. 对球直径作测量, 设其服从 $[a, b]$ 上的均匀分布, 求球的体积的均值.
34. 设例 2.26 中 \quad 服从 $[500, 1500]$ 上的均匀分布. 求平均损失 EL .
35. X 服从参数为 2 的指数分布, 求 $Y = 1 - e^{-2X}$ 的分布函数和密度函数.
36. 已知某股票的一年以后价格 X 服从对数正态分布, 当前价格为 10 元, 且 $EX = 15$, $DX = 4$. 求其连续复合年收益率的分布.

37. 设 \quad 为非负随机变量, 密度函数为 $f(x)$, 证明 \quad 的密度函数为:

$$f(y) = \begin{cases} 2yf(y^2), & y > 0 \\ 0 & y = 0 \end{cases}$$

38. 设随机变量 X 的密度函数为 $f_X(x)$, 令 $Y = aX + b$, ($a > 0$) 证明: Y 的密度函数为

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

习 题 二

(B)

1. 设 $F(x)$ 是一个连续型随机变量的分布函数, $a > 0$, 证明: $\int_{-\infty}^{+\infty} [F(x+a) - F(x)] dx = a$.
2. 某型号电子管的寿命 \quad 服从指数分布, 如果它的平均寿命为 $E = 1000$ 小时. (1) 一个该型号的旧电子管的寿命记为 \quad , 求 \quad 的密度函数; (2) 一个系统由 n 个该型号的电子管并联组成, 求该系统的寿命 X 的密度函数; (3) 一个系统由 n 个该型号电子管串联而成, 求该系统的寿命 Y 的密度函数.
3. 某商场经统计发现顾客对某商品的日需求量 $X \sim N(\mu, \sigma^2)$, 且日平均需求量 $\mu = 40$ (件), 销售机会在 30(件) ~ 50(件) 之间的概率为 0.5. 若进货不足每件损失利润 70 元. 进货过量每件损失 100 元, 求日最优进货量.
4. 设袋中装有 m 只颜色各不相同的球, 有放回地摸取 n 次, 摸到的球的颜色种数为 \quad , 求证: $E = m \left(1 - \left(1 - \frac{1}{m}\right)^n\right)$.
5. 利用概率论的思想证明: $\int_a^b f(x) dx \leq (b-a) \int_a^b f^2(x) dx$, 其中 $f(x)$ 在 $[a, b]$ 上连续.

第 3 章

随 机 向 量

一个随机变量是定义在概率空间上的一个函数，因而，它实际上是一个与某一特定试验相联系的变量. 这一章, 我们将把讨论扩展到多个随机变量——随机向量，重点讨论二维随机向量. 对二维随机向量的讨论很容易被扩展到二维以上的随机向量. 本章的主要内容包括随机向量的（联合）分布，独立性以及数字特征. 在最后，我们还将专门介绍在概率论的理论和应用中都占有重要地位的极限理论.

§ 3.1 随机向量的分布

一、随机向量及其分布函数

定义 3.1 设 X_1, X_2, \dots, X_n 是定义在概率空间 (Ω, \mathcal{F}, P) 上的 n 个随机变量，则称 (X_1, X_2, \dots, X_n) 是 (Ω, \mathcal{F}, P) 上的一个 n 维随机向量.

对于任何一个随机变量，我们都可以用分布函数来描述其统计特性，同样，我们也引入 n 维随机向量的分布函数.

定义 3.2 设 (X_1, X_2, \dots, X_n) 是 (Ω, \mathcal{F}, P) 上的一个 n 维随机向量，则称 n 元函数：

$$F(x_1, x_2, \dots, x_n) = P\{X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n\} \tag{3.1}$$

为随机向量 (X_1, X_2, \dots, X_n) 的分布函数或 n 个随机变量 X_1, X_2, \dots, X_n 的联合分布函数.

为简单起见，我们主要讨论二维情形，设 (X, Y) 为一个二维随机向量，对给定的实向量 (x, y) , $F(x, y) = P\{X \leq x, Y \leq y\}$ 实际上是 (X, Y) 取值于区域 $\{(t, s) \in \mathcal{R}^2 : t \leq x, s \leq y\}$ (见图 3.1 中的阴影部分) 的概率. 而根据概率的加法法则， (X, Y) 取值于图 3.2 中的区域 $\{x_1 < X \leq x_2, y_1 < Y \leq y_2\}$, 可用分布函数来表示：

$$\begin{aligned} P\{x_1 < X \leq x_2, y_1 < Y \leq y_2\} &= P\{(X, Y) \in (x_1, x_2] \times (y_1, y_2]\} \\ &= P\{(X, Y) \in (x_1, x_2] \times \mathcal{R}\} - P\{(X, Y) \in (x_1, x_2] \times (-\infty, y_1]\} \end{aligned}$$

$$- P\{(X, Y) \in (x_1, x_2) \times (y_1, y_2)\} + P\{(X, Y) \in (x_1, x_2) \times (y_1, +\infty)\} \\ = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1). \quad (3.2)$$

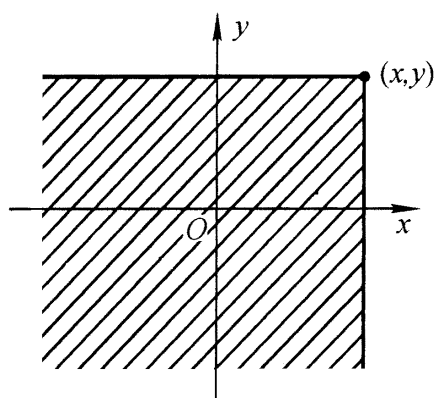


图 3.1

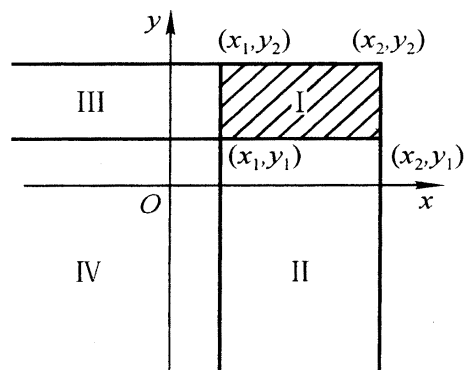


图 3.2

与一维情形类似，由（联合）分布函数的定义可导出其如下性质：

- (1) $0 \leq F(x, y) \leq 1$;
- (2) $F(x, y)$ 关于 x 和 y 均单调非降、右连续;

$$(3) F(-\infty, y) = \lim_{x \rightarrow -\infty} F(x, y) = 0,$$

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

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

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

此外，如果 (X, Y) 的分布函数 $F(x, y)$ 已知，则由 $F(x, y)$ 可导出 X 和 Y 各自的分布函数 $F_X(x)$ 和 $F_Y(y)$ ：

$$F_X(x) = P\{X \leq x\} = P\{X \leq x, Y < +\infty\} = F(x, +\infty) \quad (3.3)$$

同理，

$$F_Y(y) = P\{Y \leq y\} = F(+\infty, y) \quad (3.4)$$

通常称 $F_X(x)$ 和 $F_Y(y)$ 为（联合）分布函数 $F(x, y)$ 的边缘分布函数。

一般地，设 n 维随机向量 (X_1, X_2, \dots, X_n) 的分布函数为 $F(x_1, x_2, \dots, x_n)$ ，则称

$$F_i(x_i) = F(+\infty, \dots, +\infty, x_i, +\infty, \dots, +\infty) \quad (3.5) \\ i = 1, 2, \dots, n$$

为 $F(x_1, x_2, \dots, x_n)$ 的边缘分布函数。

二、离散型随机向量的概率分布

定义 3.3 如果二维随机向量 (X, Y) 只取有限个或可数个值，则称 (X, Y) 为二维离散型随机向量。

显然, (X, Y) 为二维离散型随机向量, 当且仅当 X, Y 均为离散型随机变量. 自然, 对二维离散型随机向量的最直接的描述是给出其取每一可能值的概率, 为此, 我们引入下列定义.

定义 3.4 设随机向量 (X, Y) 的所有可能取值为 (x_i, y_j) , $i, j = 1, 2, \dots$, 如果已知

$$P \{X = x_i, Y = y_j\} = p_{ij}, \quad i, j = 1, 2, \dots, \tag{3.6}$$

则称 (3.6) 为随机向量 (X, Y) 的概率分布, 或 X 和 Y 的联合概率分布.

容易看出 p_{ij} 满足下列性质:

- (1) $p_{ij} \geq 0, \quad i, j = 1, 2, \dots,$
- (2) $\sum_i \sum_j p_{ij} = 1.$

为了直观, 有时也将联合概率分布用表格形式表示 (见表 3.1), 并称之为联合概率分布表.

表 3.1 联合概率分布表

<div><div></div><div>X</div></div>	Y					P {X = x _i }
	y ₁	y ₂	...	y _j	...	
x ₁	p ₁₁	p ₁₂	...	p _{1j}	...	$\sum_j p_{1j}$
x ₂	p ₂₁	p ₂₂	...	p _{2j}	...	$\sum_j p_{2j}$
x _i	p _{i1}	p _{i2}	...	p _{ij}	...	$\sum_j p_{ij}$
P {Y = y _j }	$\sum_i p_{i1}$	$\sum_i p_{i2}$...	$\sum_i p_{ij}$...	

由 X 和 Y 的联合概率的分布, 可以求出 X, Y 各自的概率分布 $p_i^X, i = 1, 2, \dots; p_j^Y, j = 1, 2, \dots,$

$$\begin{aligned} p_i^X &= P \{X = x_i\} = P \{ \sum_j \{X = x_i, Y = y_j\} \} \\ &= \sum_j P \{X = x_i, Y = y_j\} = \sum_j p_{ij}, i = 1, 2, \dots \end{aligned} \tag{3.7}$$

同理,

$$p_j^Y = P \{Y = y_j\} = \sum_i p_{ij} \quad j = 1, 2, \dots \tag{3.8}$$

通常称 (3.7), (3.8) 为联合概率分布 $P \{X = x_i, Y = y_j\} = p_{ij}, i, j = 1, 2, \dots$ 的边缘概率分布. 在联合概率分布表中, 边缘分布分别列在表中的最后一行和最后一列, 它们分别等于联合概率分布的行和或列和, 见表 3.1.

例 3.1 将两封信随意地投入 3 个邮筒, 设 X, Y 分别表示投入第 1, 2 号

邮筒中信的数目，求 X 和 Y 的联合概率分布及边缘概率分布.

解 X, Y 各自的可能取值显然均为 $0, 1, 2$ ，由题设知， (X, Y) 取 $(1, 2), (2, 1), (2, 2)$ 均不可能，因而相应的概率均为 0 ，我们将其标在联合概率分布表中相应位置. (X, Y) 取其他值的概率可由古典概型计算，由于对称性，我们实际上只需计算下列概率：

$$P\{X=0, Y=0\}=\frac{1}{3^2}=\frac{1}{9},$$

$$P\{X=0, Y=1\}=\frac{2}{3^2}=\frac{2}{9},$$

$$P\{X=1, Y=1\}=\frac{2}{3^2}=\frac{2}{9}.$$

边缘概率分布可直接在联合概率分布表中计算，其中 X 的概率分布由行和产生， Y 的概率分布由列和产生.（见下表）

$X \backslash Y$	0	1	2	p_i^X
0	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{4}{9}$
1	$\frac{2}{9}$	$\frac{2}{9}$	0	$\frac{4}{9}$
2	$\frac{1}{9}$	0	0	$\frac{1}{9}$
p_j^Y	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{1}{9}$	

对离散型随机向量而言，联合概率分布不仅比联合分布函数更加直观，而且它能够更加方便地确定 (X, Y) 取值于任何区域 D 上的概率，事实上，有

$$P\{(X, Y) \in D\}=\sum_{(x_i, y_j) \in D} p_{ij} . \tag{3.9}$$

特别地，由联合概率分布可以确定联合分布函数：

$$F\{(x, y)\}=P\{X \leq x, Y \leq y\}=\sum_{x_i \leq x, y_j \leq y} p_{ij} . \tag{3.10}$$

例 3.2 设 (X, Y) 的概率分布由右表给出，求 $P\{X \leq 0, Y=0\}$ ， $P\{X=0, Y=0\}$ ， $P\{XY=0\}$ ， $P\{X=Y\}$ ， $P\{X \neq Y\}$ ， $P\{X \leq 0, Y \leq 0\}$.

$X \backslash Y$	-1	0	2
0	0.1	0.2	0
1	0.3	0.05	0.1
2	0.15	0	0.1

$$\begin{aligned} \text{解 } P\{X \leq 0, Y=0\} &= P\{X=1, Y=0\}+P\{X=2, Y=0\} \\ &= 0.05+0=0.05, \end{aligned}$$

$$\begin{aligned} P\{X=0, Y=0\} &= P\{X=0, Y=-1\} + P\{X=0, Y=0\} \\ &= 0.1 + 0.2 = 0.3, \end{aligned}$$

$$\begin{aligned} P\{XY=0\} &= P(\{X=0\} \cup \{Y=0\}) \\ &= P\{X=0\} + P\{X \neq 0, Y=0\} \\ &= 0.1 + 0.2 + 0 + 0.05 = 0.35 \end{aligned}$$

$$\begin{aligned} P\{X=Y\} &= P\{X=0, Y=0\} + P\{X=2, Y=2\} \\ &= 0.2 + 0.1 = 0.3, \end{aligned}$$

$$\begin{aligned} P\{X \neq Y\} &= P\{X=0, Y=0\} + P\{X=1, Y=-1\} + P\{X=2, \\ &\quad Y=2\} \\ &= 0.2 + 0.3 + 0.1 = 0.6. \end{aligned}$$

三、连续型随机向量的概率密度函数

定义 3.5 设 (X, Y) 为二维随机向量, 分布函数为 $F(X, Y)$, 如果存在一个非负可积的二元函数 $f(x, y)$, 使得对任意实向量 (x, y) , 有

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(s, t) ds dt, \quad (3.11)$$

则称 (X, Y) 为二维连续型随机向量, 并称 $f(x, y)$ 为 (X, Y) 的概率密度函数 (简称密度函数), 或 X 与 Y 的联合密度函数.

从定义 3.5, 易知联合密度函数 $f(x, y)$, 满意下列性质:

- (1) $f(x, y) \geq 0$,
- (2) $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1$.

反过来, 满足上述两条性质的函数 $f(x, y)$, 我们都称为联合密度函数, 因为可以证明, 它一定是某两个随机变量的联合密度函数. 此外, 通过密度函数, 可以表达随机向量取值于任何平面区域的概率:

- (3) 若 D 是平面上的一个区域, 则

$$P\{(X, Y) \in D\} = \int_D f(x, y) dx dy \quad (3.12)$$

特别地, 由 (3.12), 边缘分布函数 $F_X(x)$ 可表示为:

$$\begin{aligned} F_X(x) &= P\{X \leq x\} = P\{X \leq x, Y < +\infty\} \\ &= \int_{-\infty}^x \int_{-\infty}^{+\infty} f(s, t) ds dt \\ &= \int_{-\infty}^x \left(\int_{-\infty}^{+\infty} f(s, t) dt \right) ds \end{aligned} \quad (3.13)$$

由 (3.13) 知, X 是连续型随机变量, 且其密度函数为:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy. \quad (3.14)$$

同理, Y 是连续型随机变量, 其密度函数为:

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx. \quad (3.15)$$

通常称 (3.14), (3.15) 中的 $f_X(x)$ 和 $f_Y(y)$ 为 (X, Y) 或联合密度函数 $f(x, y)$ 的边缘密度函数.

例 3.3 (均匀分布) 设 G 是平面上的一个有界区域, 其面积记作 $S(G)$, 二维随机向量 (X, Y) 只在 G 中取值, 且取 G 中的每一点是“等可能的”, 即在 G 中每一点的概率密度相同. 于是, (X, Y) 的密度函数为:

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

由密度函数的性质:

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

可得: $C = \frac{1}{S(G)}$, 即有

$$f(x, y) = \begin{cases} \frac{1}{S(G)}, & (x, y) \in G, \\ 0, & \text{其他.} \end{cases} \quad (3.16)$$

通常, 如果一个二维随机向量 (X, Y) 以 (3.16) 为密度函数, 则称 (X, Y) 服从区域 G 上的均匀分布.

由 (3.12), 若 (X, Y) 服从 G 上的均匀分布, 则对任何平面区域 D , 有:

$$\begin{aligned} P\{(X, Y) \in D\} &= \int_D f(x, y) dx dy = \int_{D \cap G} \frac{1}{S(G)} dx dy \\ &= \frac{S(D \cap G)}{S(G)} \end{aligned} \quad (3.17)$$

其中 $S(D \cap G)$ 是区域 $D \cap G$ 的面积.

例 3.4 设随机向量 (X_1, Y_1) 的密度函数 $f(x, y)$, (X_2, Y_2) 的密度函数 $g(x, y)$ 分别为

$$\begin{aligned} f(x, y) &= \begin{cases} k_1 xy, & 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0, & \text{其他;} \end{cases} \\ g(x, y) &= \begin{cases} k_2 xy, & 0 \leq x \leq y \leq 1, \\ 0, & \text{其他.} \end{cases} \end{aligned}$$

求 (1) 参数 k_1 的值及 (X_1, Y_1) 的边缘密度;

(2) 参数 k_2 的值及 (X_2, Y_2) 的边缘密度.

解 (1) 由密度函数的性质, 有

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = \int_0^1 \int_0^1 k_1 xy dx dy = 1,$$

由此易得 $k_1 = 4$, 从而

$$f(x, y) = \begin{cases} 4xy, & 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0, & \text{其他.} \end{cases}$$

记 (X_1, Y_1) 的边缘密度函数分别为 $f_{x_1}(x), f_{y_1}(y)$, 则由

$$f_{x_1}(x) = \int_{-\infty}^{+\infty} f(x, y) dy,$$

得: 当 $0 \leq x \leq 1$ 时,

$$f_{x_1}(x) = \int_0^1 4xy dy = 2x.$$

当 $x < 0$ 或 $x > 1$ 时,

$$f_{x_1}(x) = 0.$$

即

$$f_{x_1}(x) = \begin{cases} 2x, & 0 \leq x \leq 1, \\ 0, & \text{其他.} \end{cases}$$

根据对称性, 有

$$f_{y_1}(y) = \begin{cases} 2y, & 0 \leq y \leq 1, \\ 0, & \text{其他.} \end{cases}$$

(2) 由密度函数的性质, 有

$$\begin{aligned} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) dx dy &= \int_0^1 \int_0^1 k_2 xy dx dy = \int_0^1 dx \int_x^1 k_2 xy dy \\ &= \frac{1}{8} k_2 = 1 \end{aligned}$$

从而, $k_2 = 8$, 故知

$$g(x, y) = \begin{cases} 8xy, & 0 \leq x \leq y \leq 1, \\ 0, & \text{其他.} \end{cases}$$

记 (X_2, Y_2) 的边缘密度函数为 $g_{x_2}(x), g_{y_2}(y)$, 则由

$$g_{x_2}(x) = \int_{-\infty}^{+\infty} g(x, y) dy$$

得: 当 $0 \leq x \leq 1$ 时,

$$g_{x_2}(x) = \int_x^1 8xy dy = 4x(1 - x^2),$$

当 $x < 0$ 或 $x > 1$ 时,

$$g_{x_2}(x) = 0,$$

即

$$g_{x_2}(x) = \begin{cases} 4x(1 - x^2), & 0 \leq x \leq 1, \\ 0, & \text{其他.} \end{cases}$$

同理, 可得

$$g_{Y_2}(y) = \begin{cases} 4y^3, & 0 \leq y \leq 1, \\ 0, & \text{其他.} \end{cases}$$

四、二元正态分布

在讨论一元随机变量时，我们曾指出，一元正态分布是实际应用中最常见的分布之一。对二维随机向量，在实际中会常用到二元正态分布：设随机向量 (X, Y) 的密度函数为：

$$\varphi(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2\rho} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]} \quad (3.18)$$

其中 $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$ 均为参数，且 $\sigma_1^2 > 0, \sigma_2^2 > 0, |\rho| < 1$ ，则称 (X, Y) 服从参数为 $(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2; \rho)$ 的二元正态分布，记作 $(X, Y) \sim N(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2; \rho)$

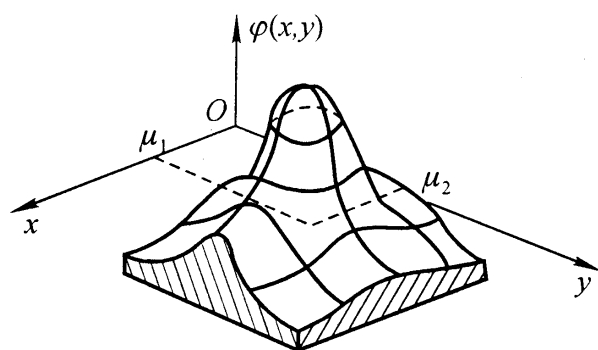


图 3.3 二元正态分布的密度函数

如图 3.3，二元正态分布以 (μ_1, μ_2) 为中心，在中心附近具有较高的密度，离中心越远，密度越小，这与实际中很多现象相吻合。

若 $(X, Y) \sim N(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2; \rho)$ ，记其边缘密度函数分别为 $\varphi_X(x)$ 和 $\varphi_Y(y)$ ，令 $u = \frac{x-\mu_1}{\sigma_1}, t = \frac{y-\mu_2}{\sigma_2}$ ，则

$$\begin{aligned} \varphi_X(x) &= \int_{-\infty}^{+\infty} \varphi(x, y) dy \\ &= \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}[u^2 - 2\rho ut + t^2]} dt \\ &= \frac{1}{2\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \cdot \frac{1}{2\sigma_2\sqrt{1-\rho^2}} \int_{-\infty}^{+\infty} e^{-\frac{(t-\rho u)^2}{2(1-\rho^2)}} dt \\ &= \frac{1}{2\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \end{aligned}$$

可见 $X \sim N(\mu_1, \sigma_1^2)$ 。对称地，可知

$$\varphi_Y(y) = \frac{1}{2\sigma_2} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}},$$

即 $Y \sim N(\mu_2, \sigma_2^2)$ 。

比较联合密度函数 $\varphi(x, y)$ 和边缘密度函数 $\varphi_X(x), \varphi_Y(y)$ ，我们注意到，当且仅当 $\rho = 0$ 时，对一切 (x, y) ，有

$$(x, y) = \begin{pmatrix} x(x) & y(y) \end{pmatrix}. \quad (3.19)$$

上述讨论实际上说明:

(1) 二元正态分布的边缘分布是一元正态分布, 且对应于二元正态分布的前 4 个参数:

(2) 不同的二元正态分布, 比如不同的 ρ , 可以有相同的边缘分布, 因而由边缘分布不能惟一确定联合分布. 为了确定一个二元正态分布的密度函数, 除了知道边缘分布以外, 还须知道参数 ρ 的值. 特别地, 如果 $\rho = 0$, 则联合密度函数可由 (3.19) 确定. 除此之外, 值得进一步指出的是:

(3) 两个边缘分布为正态分布的二维随机向量未必服从二元正态分布 (见习题三, 第 15 题和第 16 题).

§ 3.2 条件分布与随机变量的独立性

一、条件分布与独立性的一般概念

设 X 是一个随机变量, 我们知道, 其统计规律可由其分布函数 $F_X(x) = P\{X \leq x\}$, $-\infty < x < +\infty$ 来完整描述. 现在假设我们观察到一个额外的信息: 事件 A 已经发生, 那么 A 的发生可能会对事件 $\{X \leq x\}$ 发生的概率产生影响. 对每个给定的实数 x , 我们记条件概率 $P\{X \leq x | A\}$ 为 $F(x|A)$, 并称 $F(x|A)$, $-\infty < x < +\infty$ 为在 A 发生的条件下, X 的条件分布函数.

例 3.5 设 X 服从 $[0, 1]$ 上的均匀分布, 求在已知 $X > \frac{1}{2}$ 的条件下, X 条件分布函数.

解 由条件分布函数的定义, 有

$$F(x|X > \frac{1}{2}) = \frac{P\{X \leq x, X > \frac{1}{2}\}}{P\{X > \frac{1}{2}\}}.$$

由于 X 服从 $[0, 1]$ 上的均匀分布, 故 $P\{X > \frac{1}{2}\} = \frac{1}{2}$. 而当 $x \leq \frac{1}{2}$ 时,

$$P\{X \leq x, X > \frac{1}{2}\} = 0$$

当 $x > \frac{1}{2}$ 时,

$$P\{X \leq x, X > \frac{1}{2}\} = F(x) - F(\frac{1}{2}) = F(x) - \frac{1}{2},$$

其中 $F(x)$ 为 X 的分布函数, 我们知道,

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

于是, 当 $x > \frac{1}{2}$ 时

$$P\{X \leq x, X > \frac{1}{2}\} = \begin{cases} x - \frac{1}{2}, & \frac{1}{2} < x \leq 1, \\ \frac{1}{2}, & x > 1. \end{cases}$$

从而可得:

$$F(x|X > \frac{1}{2}) = \begin{cases} 0, & x \leq \frac{1}{2} \\ 2x - 1, & \frac{1}{2} < x \leq 1 \\ 1, & x > 1 \end{cases}$$

设 A 是另一随机变量 Y 所生成的事件: $A = \{Y \leq y\}$, 且 $P\{Y \leq y\} > 0$, 则有

$$F(x|Y \leq y) = \frac{P\{X \leq x, Y \leq y\}}{P\{Y \leq y\}} = \frac{F(x, y)}{F_Y(y)} \quad (3.20)$$

一般, 两个随机变量 X 和 Y 之间存在着相互联系, 因而一个随机变量的取值可能会影响另一随机变量取值的统计规律性. (3.20) 表明联合分布函数包含了 X 与 Y 相互联系的内容.

对给定的 x 和 y , 如果事件 $\{X \leq x\}$ 与事件 $\{Y \leq y\}$ 独立, 则有

$$P\{X \leq x, Y \leq y\} = P\{X \leq x\}P\{Y \leq y\}$$

亦即

$$F(x, y) = F_X(x)F_Y(y) \quad (3.21)$$

此时,

$$F(x|Y \leq y) = F_X(x)$$

如果对任意 x, y , $\{X \leq x\}$ 与 $\{Y \leq y\}$ 均独立, 则称随机变量 X 与 Y 相互独立, 由 (3.21) 我们给出下列定义:

定义 3.6 设随机变量 X, Y 的联合分布函数为 $F(x, y)$, 边缘分布函数分别为 $F_X(x), F_Y(y)$, 如果对任意实数 x 和 y , 恒有

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

则称随机变量 X 和 Y 相互独立.

表面上, 上述定义不太自然, 随机变量 X 与 Y 独立似乎应该要求 X 所生成的任何事件与 Y 所生成的任何事件都相互独立, 其实不然, 我们有下列定理.

定理 3.1 随机变量 X 与 Y 相互独立的充要条件是 X 所生成的任何事件

与 Y 生成的任何事件独立, 即, 对任意实数集 A 和 B , 有

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\} \quad (3.22)$$

定理 3.1 的充分性显然, 必要性的完整证明超出了本书的要求, 作为练习, 读者可以证明 $A = (x_1, x_2], B = (y_1, y_2]$ 这种简单的情形。

定理 3.2 如果随机变量 X 和 Y 相互独立, 则对任意函数 $g_1(x), g_2(y)$, 均有 $g_1(X)$ 与 $g_2(Y)$ 相互独立。

证明: 令 $U = g_1(X), V = g_2(Y)$, 对任意 x, y , 记 $D_x^1 = \{t \in \mathbb{R} : g_1(t) = x\}, D_y^2 = \{t \in \mathbb{R} : g_2(t) = y\}$, 则由定理 3.1, 有

$$\begin{aligned} P\{U = x, V = y\} &= P\{g_1(X) = x, g_2(Y) = y\} \\ &= P\{X \in D_x^1, Y \in D_y^2\} \\ &= P\{X \in D_x^1\}P\{Y \in D_y^2\} \\ &= P\{U = x\}P\{V = y\} \end{aligned}$$

从而由定义 3.6 知 U 与 V 独立。

关于两个随机变量的独立性的概念和讨论可以推广到 n 个随机变量的情形。

定义 3.7 设 X_1, X_2, \dots, X_n 是 n 个随机变量, 其联合分布函数为 $F(x_1, x_2, \dots, x_n)$, 边缘分布函数为 $F_i(x_i), i = 1, 2, \dots, n$, 如果对任意实数 x_1, x_2, \dots, x_n 恒有:

$$F(x_1, x_2, \dots, x_n) = F_1(x_1)F_2(x_2) \dots F_n(x_n)$$

则称 X_1, X_2, \dots, X_n 相互独立

条件分布函数为条件分布和随机变量的独立性提供了一般性描述和讨论, 但应用起来十分不便, 对于本书重点讨论的离散型和连续型的随机变量 (向量), 我们希望有更加直观而方便的描述。

二、离散型随机变量的条件概率分布与独立性

设 (X, Y) 是二维离散型随机向量, 其概率分布为

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

则由条件概率公式, 当 $P\{Y = y_j\} > 0$ 时, 有

$$P\{X = x_i | Y = y_j\} = \frac{P\{X = x_i, Y = y_j\}}{P\{Y = y_j\}} = \frac{p_{ij}}{p_{j\cdot}} \quad (3.23)$$

其中 $P\{X = x_i | Y = y_j\}$ 是在事件 “ $Y = y_j$ ” 发生的条件下, 事件 “ $X = x_i$ ” 发生的条件概率, 通常记作 p_{ij} 。

对固定的 j , $P\{X = x_i | Y = y_j\} = p_{ij}, i = 1, 2, \dots$, 完整地给出了在 $Y = y_j$ 的

条件下, X 取每一个可能值 x_i ($i = 1, 2, \dots$) 的概率, 且不难验证, 数列 $p_{i|j}$, $i = 1, 2, \dots$ 满足概率分布所要求的性质:

$$(1) \quad p_{i|j} \geq 0$$

$$(2) \quad \sum_i p_{i|j} = 1$$

因而, 我们称

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

为已知 $Y = y_j$ 的条件下, X 的条件概率分布.

对称地, 如果 $P\{X = x_i\} > 0$, 那么在 $X = x_i$ 的条件下, Y 的条件概率分布为:

$$P\{Y = y_j | X = x_i\} = p_{j|i} = \frac{p_{ij}}{p_i^X}, \quad j = 1, 2, \dots \quad (3.24)$$

此外, 根据乘法公式或 (3.23), (3.24), 对任意 $i, j = 1, 2, \dots$,

$$p_{ij} = p_i^X \cdot p_{j|i} = p_j^Y \cdot p_{i|j} \quad (3.25)$$

例 3.6 设 X 与 Y 的联合概率分布由例 3.2 给出, 求 $Y = 0$ 时, X 的条件概率分布以及 $X = 0$ 时, Y 的条件概率分布.

解 $P\{Y = 0\} = 0.2 + 0.05 + 0 = 0.25$, 在 $Y = 0$ 时, X 的条件概率分布为:

$$P\{X = 0 | Y = 0\} = \frac{P\{X = 0, Y = 0\}}{P\{Y = 0\}} = \frac{0.2}{0.25} = 0.8$$

$$P\{X = 1 | Y = 0\} = \frac{P\{X = 1, Y = 0\}}{P\{Y = 0\}} = \frac{0.05}{0.25} = 0.2$$

$$P\{X = 2 | Y = 0\} = \frac{P\{X = 2, Y = 0\}}{P\{Y = 0\}} = \frac{0}{0.25} = 0$$

又 $P\{X = 0\} = 0.1 + 0.2 + 0 = 0.3$, 故在 $X = 0$ 时, Y 的条件概率分布为

$$P\{Y = -1 | X = 0\} = \frac{P\{Y = -1, X = 0\}}{P\{X = 0\}} = \frac{0.1}{0.3} = \frac{1}{3}$$

$$P\{Y = 0 | X = 0\} = \frac{P\{Y = 0, X = 0\}}{P\{X = 0\}} = \frac{0.2}{0.3} = \frac{2}{3}$$

$$P\{Y = 2 | X = 0\} = \frac{P\{Y = 2, X = 0\}}{P\{X = 0\}} = \frac{0}{0.3} = 0$$

接下来考虑离散型随机变量的独立性. 由定理 3.1, 如果 X 与 Y 相互独立, 那么对任意 i, j , 必有

$$P\{X = x_i, Y = y_j\} = P\{X = x_i\}P\{Y = y_j\} \quad (3.26)$$

另一方面, 如果对任意 i, j , (3.26) 成立, 则 X 与 Y 相互独立, 这是因为: 由 (3.26), 对任意 x 和 y ,

$$P\{X = x, Y = y\} = P\left(\bigcup_{x_i=x} \{X = x_i, Y = y\}\right)$$

$$\begin{aligned}
 &= \sum_{x_i \in x} P\{X = x_i, Y = y_j\} \\
 &= \sum_{x_i \in x} P\{X = x_i\}P\{Y = y_j\} \\
 &= P\left(\sum_{x_i \in x} \{X = x_i\}\right)P\{Y = y_j\} \\
 &= P\{X \in x\}P\{Y = y_j\},
 \end{aligned}$$

进而, 对任意 x, y , 有

$$\begin{aligned}
 P\{X \in x, Y \in y\} &= P\left(\sum_{y_j \in y} \{X \in x, Y = y_j\}\right) \\
 &= \sum_{y_j \in y} P\{X \in x, Y = y_j\} \\
 &= \sum_{y_j \in y} P\{X \in x\}P\{Y = y_j\} \\
 &= P\{X \in x\}P\left(\sum_{y_j \in y} \{Y = y_j\}\right) \\
 &= P\{X \in x\}P\{Y \in y\}.
 \end{aligned}$$

于是, 我们有下列定理.

定理 3.3 X, Y 是离散型随机变量, 其联合概率分布为 $P\{X = x_i, Y = y_j\} = p_{ij}$, $i, j = 1, 2, \dots$, 边缘概率分布分别为 p_i^X 和 p_j^Y , $i, j = 1, 2, \dots$, 则 X 与 Y 相互独立的充要条件是

$$p_{ij} = p_i^X p_j^Y, \quad i, j = 1, 2, \dots \quad (3.27)$$

例 3.7 设 X 与 Y 的联合概率分布由例 3.2 给出, 判断 X 与 Y 是否相互独立?

解 因为

$$P\{X = 0\} = 0.1 + 0.2 = 0.3,$$

$$P\{Y = -1\} = 0.1 + 0.3 + 0.15 = 0.55$$

而 $P\{X = 0, Y = -1\} = 0.1$, 可见, $P\{X = 0, Y = -1\} \neq P\{X = 0\}P\{Y = -1\}$

所以, X 与 Y 不独立.

在前一节讨论中, 我们得知, 由联合概率分布可以确定边缘概率分布. 但是由边缘概率分布一般不能确定联合概率分布, 比较表 3.2 中的两个不同联合概率分布, 我们注意到它们具有相同的边缘概率分布.

如果已知随机变量 X 与 Y 相互独立, 那么由边缘概率分布可以确定联合概率分布, 因为此时

$$p_{ij} = p_i^X p_j^Y$$

比如, 设 X 的概率分布为

$$P\{X=0\}=P\{X=1\}=P\{X=2\}=\frac{1}{3}$$

Y 的概率分布为

$$P\{Y=0\}=P\{Y=1\}=P\{Y=2\}=\frac{1}{3}$$

且已知 X 与 Y 相互独立, 那么 X 与 Y 的联合概率分布可由 (3.27) 求得, 它实际上就是表 3.2 中右表所列的联合概率分布.

表 3.2 具有相同边缘概率分布的两个不同的联合概率分布

<div><div>Y</div><div>X</div></div>	0	1	2	p_i^X
0	$\frac{2}{9}$	$\frac{1}{9}$	0	$\frac{1}{3}$
1	0	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{1}{3}$
2	$\frac{1}{9}$	0	$\frac{2}{9}$	$\frac{1}{3}$
p_j^Y	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	

<div><div>Y</div><div>X</div></div>	0	1	2	p_i^X
0	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$
1	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$
2	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$
p_j^Y	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	

三、连续型随机变量的条件密度函数与独立性

设 (X, Y) 是连续型随机向量, 分布函数和密度函数分别为 F(x, y) 和 f(x, y). 我们希望考虑在 Y= y 的条件下, X 的条件分布. 但由于 {Y= y} 是一个零概率事件,

$$P\{X \leq x | Y = y\} = \frac{P\{X \leq x, Y = y\}}{P\{Y = y\}} \tag{3.28}$$

的分子、分母均为 0, 因而直接根据条件概率定义来考虑 X 条件分布行不通. 为此, 我们将 (3.28) 视为 $\frac{0}{0}$ 型的未定式, 即有:

$$\begin{aligned} P\{X \leq x | Y = y\} &= \lim_{y \rightarrow 0} P\{X \leq x | y - \epsilon < Y < y + \epsilon\} \\ &= \lim_{y \rightarrow 0} \frac{P\{X \leq x, y - \epsilon < Y < y + \epsilon\}}{P\{y - \epsilon < Y < y + \epsilon\}} \\ &= \lim_{y \rightarrow 0} \frac{\int_{y-\epsilon}^{y+\epsilon} \int_{-\infty}^x f(u, t) du dt}{\int_{y-\epsilon}^{y+\epsilon} f_Y(t) dt} \\ &= \frac{\int_{-\infty}^x f(u, y) du}{f_Y(y)} \end{aligned} \tag{3.29}$$

并且, 对给定的 y, 如果 f_Y(y) > 0, 则称 P{X ≤ x | Y = y} 为 Y = y 的条件下 X 的分布函数, 记作 F_{X|Y}(x|y). 由 (3.29) 知

$$F_{X|Y}(x|y) = \int_{-\infty}^x \frac{f(u, y)}{f_Y(y)} du \quad (3.30)$$

因而, 称 $\frac{f(x, y)}{f_Y(y)}$ 为 $Y=y$ 的条件下, X 的条件密度函数, 记作 $f_{X|Y}(x|y)$.

同样, 对给定的 x , 如果 $f_X(x) > 0$, 那么在 $X=x$ 的条件下, Y 的条件分布函数为

$$F_{Y|X}(y|x) = \int_{-\infty}^y \frac{f(x, t)}{f_X(x)} dt \quad (3.31)$$

条件密度函数为

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} \quad (3.32)$$

由条件密度函数的定义, 我们容易知道, 密度函数有下列乘法公式:

$$f(x, y) = f_X(x)f_{Y|X}(y|x) = f_Y(y)f_{X|Y}(x|y). \quad (3.33)$$

例 3.8 设 (X, Y) 是在 $D = \{(x, y) | x^2 + y^2 \leq 1\}$ 上服从均匀分布的随机向量, 求 $f_{Y|X}(y|x)$ 和 $f_{X|Y}(x|y)$.

解 由于 (X, Y) 的密度函数为

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

于是其边缘密度函数 $f_X(x)$ 为:

$$f_X(x) = \begin{cases} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} f(x, y) dy = \frac{2\sqrt{1-x^2}}{\pi}, & |x| \leq 1, \\ 0, & \text{其他} \end{cases}$$

对称地, $f_Y(y)$ 为:

$$f_Y(y) = \begin{cases} \frac{2\sqrt{1-y^2}}{\pi}, & |y| \leq 1, \\ 0, & \text{其他} \end{cases}$$

于是, 对一切 $x: |x| \leq 1$, 有

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = \begin{cases} \frac{1}{2\sqrt{1-x^2}}, & |y| \leq \sqrt{1-x^2} \\ 0, & \text{其他} \end{cases}$$

同样, 对一切 $y: |y| \leq 1$, 有

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

例 3.9 设 $(X, Y) \sim N(\mu, \mu; \frac{\sigma^2}{2}, \frac{\sigma^2}{2}; \rho)$, 求 $f_{X|Y}(x|y)$ 和 $f_{Y|X}(y|x)$.

解 由 § 3.1 知, $X \sim N(\mu, \frac{\sigma^2}{2}), Y \sim N(\mu, \frac{\sigma^2}{2})$ 于是:

$$\begin{aligned}
 f_{X|Y}(x|y) &= \frac{f(x, y)}{f_Y(y)} \\
 &= \frac{\frac{1}{2\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2\rho} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right]}}{\frac{1}{2\sigma_2\sqrt{1-\rho^2}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2(1-\rho^2)}}} \\
 &= \frac{1}{2\sigma_1\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left[\frac{x-\mu_1}{\sigma_1} - \frac{y-\mu_2}{\sigma_2\rho} \right]^2} \\
 &= \frac{1}{2\sigma_1\sqrt{1-\rho^2}} e^{-\frac{1}{2\sigma_1^2(1-\rho^2)} \left[x-\mu_1 - \frac{\rho}{\sigma_2}(y-\mu_2) \right]^2}
 \end{aligned}$$

故, 在 $Y=y$ 的条件下, X 服从正态分布 $N\left(\mu_1 + \frac{\rho}{\sigma_2}(y-\mu_2), \sigma_1^2(1-\rho^2)\right)$.

对称地, 在 $X=x$ 的条件下, Y 服从正态分布 $N\left(\mu_2 + \frac{\rho}{\sigma_1}(x-\mu_1), \sigma_2^2(1-\rho^2)\right)$.

接下来考虑连续型随机变量的独立性. 我们希望用密度函数来刻画独立性. 为此, 有下列定理:

定理 3.4 设连续型随机向量 (X, Y) 的密度函数为 $f(x, y)$, 边缘密度函数分别为 $f_X(x)$ 和 $f_Y(y)$, 则 X 与 Y 相互独立的充要条件是

$$f(x, y) = f_X(x)f_Y(y) \quad (3.34)$$

证明 必要性 如果 X 与 Y 相互独立则对任意 x, y , 有

$$\begin{aligned}
 F(x, y) &= F_X(x)F_Y(y) = \int_{-\infty}^x f_X(u) du \int_{-\infty}^y f_Y(t) dt \\
 &= \int_{-\infty}^x \int_{-\infty}^y f_X(u)f_Y(t) dudt,
 \end{aligned}$$

于是 $f_X(x)f_Y(y)$ 是 (X, Y) 的密度函数, 即

$$f(x, y) = f_X(x)f_Y(y).$$

充分性 若 $f(x, y) = f_X(x)f_Y(y)$, 则

$$\begin{aligned}
 F(x, y) &= \int_{-\infty}^x \int_{-\infty}^y f_X(u)f_Y(t) dudt \\
 &= \int_{-\infty}^x f_X(u) du \int_{-\infty}^y f_Y(t) dt \\
 &= F_X(x)F_Y(y)
 \end{aligned}$$

从而 X 与 Y 相互独立.

例 3.10 判断例 3.4 中 X_1 与 Y_1 , X_2 与 Y_2 是否相互独立.

解 由例 3.4 知, X_1 与 Y_1 的联合密度函数及边缘密度函数分别为:

$$f(x, y) = \begin{cases} 4xy, & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{其他} \end{cases}$$

$$f_{x_1}(x) = \begin{cases} 2x, & 0 \leq x \leq 1 \\ 0, & \text{其他} \end{cases}$$

$$f_{y_1}(y) = \begin{cases} 2y, & 0 \leq y \leq 1 \\ 0, & \text{其他} \end{cases}$$

显然有

$$f(x, y) = f_{x_1}(x)f_{y_1}(y)$$

故 X_1 与 Y_1 相互独立.

由例 3.4 知, X_2 与 Y_2 的联合密度函数及边缘密度函数分别为:

$$g(x, y) = \begin{cases} 8xy, & 0 \leq x \leq y \leq 1 \\ 0, & \text{其他} \end{cases}$$

$$g_{x_2}(x) = \begin{cases} 4x(1-x^2), & 0 \leq x \leq 1 \\ 0, & \text{其他} \end{cases}$$

$$g_{y_2}(y) = \begin{cases} 4y^3, & 0 \leq y \leq 1 \\ 0, & \text{其他} \end{cases}$$

对任意满足 $1 > x_0 > y_0 > 0$ 的点 (x_0, y_0) , 有 $g(x_0, y_0) = 0$, 但 $g_{x_2}(x_0)g_{y_2}(y_0) = 16x_0(1-x_0^2)y_0^3 \neq 0$, 即

$$g(x_0, y_0) \neq g_{x_2}(x_0)g_{y_2}(y_0)$$

因而 X_2 与 Y_2 不相互独立.

例 3.11 若 $(X, Y) \sim N(\mu, \mu; \sigma_1^2, \sigma_2^2; \rho)$, 证明 X 与 Y 相互独立的充要条件是 $\rho = 0$.

解 由 § 3.1 知 $X \sim N(\mu, \sigma_1^2)$, $Y \sim N(\mu, \sigma_2^2)$, 于是比较 $N(\mu, \mu; \sigma_1^2, \sigma_2^2; \rho)$ 与 $N(\mu, \sigma_1^2), N_2(\mu, \sigma_2^2)$ 的密度函数: $(x, y), x(x), y(y)$ 易知, 当且仅当 $\rho = 0$ 时,

$$(x, y) = x(x) y(y).$$

§ 3.3 随机向量的函数的分布与数学期望

一、离散型随机向量的函数的分布

设 (X, Y) 是二维离散型随机向量, $g(x, y)$ 是一个二元函数, 则 $g(X, Y)$ 作为 (X, Y) 的函数是一个随机变量. 如果 (X, Y) 的概率分布为

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

记 $z_k, k = 1, 2, \dots$, 为 $Z = g(X, Y)$ 的所有可能取值, 则由 (3.9), Z 的概率分布为:

$$P\{Z = z_k\} = P\{g(X, Y) = z_k\} = \sum_{g(x_i, y_j) = z_k} P\{X = x_i, Y = y_j\} \quad (3.35)$$

$k = 1, 2, \dots$

例 3.12 设 (X, Y) 的概率分布由例 3.2 给出, 求 $U = X + Y$ 及 $V = XY$ 的概率分布.

解 显然 $U = X + Y$ 的可能取值有: $-1, 0, 1, 2, 3, 4$, $V = XY$ 的可能取值有: $-2, -1, 0, 4$. 由 (3.35), U 的概率分布为:

$$P\{U = -1\} = P\{X + Y = -1\} = P\{X = 0, Y = -1\} = 0.1,$$
$$P\{U = 0\} = P\{X + Y = 0\} = P\{X = 0, Y = 0\} + P\{X = 1, Y = -1\}$$
$$= 0.2 + 0.3 = 0.5,$$
$$P\{U = 1\} = P\{X + Y = 1\} = P\{X = 1, Y = 0\} + P\{X = 2, Y = -1\}$$
$$= 0.05 + 0.15 = 0.2,$$
$$P\{U = 2\} = P\{X + Y = 2\} = P\{X = 0, Y = 2\} + P\{X = 2, Y = 0\}$$
$$= 0 + 0 = 0,$$
$$P\{U = 3\} = P\{X + Y = 3\} = P\{X = 1, Y = 2\} = 0.1,$$
$$P\{U = 4\} = P\{X + Y = 4\} = P\{X = 2, Y = 2\} = 0.1.$$

V 的概率分布为

$$P\{V = -2\} = P\{X = 2, Y = -1\} = 0.15,$$
$$P\{V = -1\} = P\{X = 1, Y = -1\} = 0.3,$$
$$P\{V = 0\} = P\{X = 0, Y = -1\} + P\{X = 0, Y = 0\} + P\{X = 0, Y = 2\}$$
$$+ P\{X = 1, Y = 0\} + P\{X = 2, Y = 0\},$$
$$= 0.1 + 0.2 + 0 + 0.05 + 0 = 0.35,$$
$$P\{V = 2\} = P\{X = 1, Y = 2\} = 0.1,$$
$$P\{V = 4\} = P\{X = 2, Y = 2\} = 0.1.$$

我们也可用一种更直接的方式——表上作业法来求上述概率分布. 这种表上作业法一般分为三步:

第一步: 将 (X, Y) 的概率分布表达为下列形式.

$X \backslash Y$	-1	0	2
0	0.1	0.2	0
1	0.3	0.05	0.1
2	0.15	0	0.1

第二步，将每一对 (x_i, y_j) 对应的函数值 $g(x_i, y_j)$ 算出标在对应小方框的右下角($X+Y$ 见下列左表, XY 见下列右表)

$\begin{smallmatrix} Y \\ \backslash X \end{smallmatrix}$	- 1	0	2
0	$\begin{smallmatrix} 0.1 \\ - 1 \end{smallmatrix}$	$\begin{smallmatrix} 0.2 \\ 0 \end{smallmatrix}$	$\begin{smallmatrix} 0 \\ 2 \end{smallmatrix}$
1	$\begin{smallmatrix} 0.3 \\ 0 \end{smallmatrix}$	$\begin{smallmatrix} 0.05 \\ 1 \end{smallmatrix}$	$\begin{smallmatrix} 0.1 \\ 3 \end{smallmatrix}$
2	$\begin{smallmatrix} 0.15 \\ 1 \end{smallmatrix}$	$\begin{smallmatrix} 0 \\ 2 \end{smallmatrix}$	$\begin{smallmatrix} 0.1 \\ 4 \end{smallmatrix}$

$\begin{smallmatrix} Y \\ \backslash X \end{smallmatrix}$	- 1	0	2
0	$\begin{smallmatrix} 0.1 \\ 0 \end{smallmatrix}$	$\begin{smallmatrix} 0.2 \\ 0 \end{smallmatrix}$	$\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}$
1	$\begin{smallmatrix} 0.3 \\ - 1 \end{smallmatrix}$	$\begin{smallmatrix} 0.05 \\ 0 \end{smallmatrix}$	$\begin{smallmatrix} 0.1 \\ 2 \end{smallmatrix}$
2	$\begin{smallmatrix} 0.15 \\ - 2 \end{smallmatrix}$	$\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}$	$\begin{smallmatrix} 0.1 \\ 4 \end{smallmatrix}$

第三步：将第二步表格中 $g(x_i, y_i)$ 的值相同的项合并，即得 $g(X, Y)$ 的概率分布，对本例，有

$P\{Z = -1\} = 0.1, P\{Z = 0\} = 0.5, P\{Z = 1\} = 0.15 + 0.05 = 0.2,$
 $P\{Z = 2\} = 0, P\{Z = 3\} = 0.1, P\{Z = 4\} = 0.1,$
 $P\{Z = -2\} = 0.15, P\{Z = -1\} = 0.3, P\{Z = 0\} = 0.1 + 0.2 + 0.05 = 0.35,$
 $P\{Z = 2\} = 0.1, P\{Z = 4\} = 0.1.$

例 3.13 设 X, Y 是两个相互独立的随机变量, 分别服从参数为 λ_1 和 λ_2 的泊松分布, 求 $Z = X + Y$ 的分布.

解 $P\{Z = k\} = P\{X + Y = k\}$

$$\begin{aligned} &= \sum_{i=0}^k P\{X = i, Y = k - i\} \\ &= \sum_{i=0}^k P\{X = i\}P\{Y = k - i\} \\ &= \sum_{i=0}^k \frac{\lambda_1^i}{i!} e^{-\lambda_1} \frac{\lambda_2^{k-i}}{(k-i)!} e^{-\lambda_2} \\ &= e^{-(\lambda_1 + \lambda_2)} \sum_{i=0}^k \frac{\lambda_1^i \lambda_2^{k-i}}{i! (k-i)!} \\ &= \frac{(\lambda_1 + \lambda_2)^k}{k!} e^{-(\lambda_1 + \lambda_2)}, \\ &k = 0, 1, 2, \dots \end{aligned}$$

可见 $Z = X + Y$ 服从参数为 $\lambda_1 + \lambda_2$ 的泊松分布.

二、连续型随机向量的函数的分布

设 (X, Y) 是二维连续型随机向量, 其概率密度函数为 $f(x, y)$, 令 $g(x, y)$ 为一个二元函数, 则 $g(X, Y)$ 是 (X, Y) 的函数. 我们可以类似于求一元随机

变量函数分布的方法来求 $Z = g(X, Y)$ 的分布.

先求分布函数 $F_Z(z)$

$$\begin{aligned} F_Z(z) &= P\{Z \leq z\} = P\{g(X, Y) \leq z\} = P\{(X, Y) \in D_z\} \\ &= \int_{D_z} f(x, y) dx dy \end{aligned} \quad (3.36)$$

其中 $D_z = \{(x, y) \in \mathbb{R}^2 : g(x, y) \leq z\}$.

继而, 其密度函数 $f_Z(z)$, 对几乎所有的 z , 有

$$f_Z(z) = F'_Z(z) \quad (3.37)$$

例 3.14 (随机变量的和) 设 (X, Y) 的联合密度为 $f(x, y)$, 求 $X + Y$ 的密度函数

解 对任意 z , 令 $D_z = \{(x, y) \in \mathbb{R}^2 : x + y \leq z\}$, 如图 3.4, 则

$$\begin{aligned} F_Z(z) &= P\{Z \leq z\} = P\{X + Y \leq z\} \\ &= \int_{D_z} f(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} dy \int_{-\infty}^{z-y} f(x, y) dx \\ &= \int_{-\infty}^{+\infty} dy \int_{-\infty}^z f(u - y, y) du \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^z f(u - y, y) dy du \quad (3.38) \end{aligned}$$

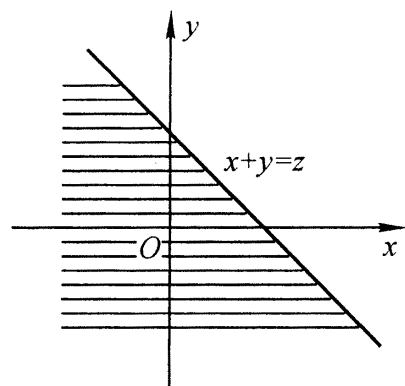


图 3.4 $D_z = \{(x, y) \in \mathbb{R}^2 : x + y \leq z\}$

于是, 有

$$f_Z(z) = \int_{-\infty}^{+\infty} f(z - y, y) dy \quad (3.39)$$

易见, 交换积分次序, 我们亦可得到:

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, z - x) dx \quad (3.40)$$

特别地, 如果 X 与 Y 是相互独立的随机变量, 则由 (3.39) 和 (3.40) 有

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{+\infty} f_X(x) f_Y(z - x) dx \\ &= \int_{-\infty}^{+\infty} f_X(z - y) f_Y(y) dy \end{aligned} \quad (3.41)$$

(3.41) 给出的积分运算通常称为函数 $f_X(x)$ 与 $f_Y(y)$ 的卷积, 记作 $f_X * f_Y(z)$, 由 (3.41) 知:

$$f_Z(z) = f_X * f_Y(z) = f_Y * f_X(z) \quad (3.42)$$

我们把 (3.41) 或 (3.42) 称为卷积公式.

作为卷积公式的一个应用, 下面给出一个重要的例子.

例 3.15 (独立正态随机变量之和) 设随机变量 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, 且 X 与 Y 独立, 证明 $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

$$\begin{aligned}
 \text{证明 } f_{X+Y}(z) &= \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x) dx \\
 &= \int_{-\infty}^{+\infty} \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \cdot e^{-\frac{(z-x-\mu_2)^2}{2\sigma_2^2}} dx \\
 &= \int_{-\infty}^{+\infty} \frac{1}{\sigma_1 \sigma_2 \sqrt{2\pi}} e^{-\frac{1}{2} \left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(z-x-\mu_2)^2}{\sigma_2^2} \right]} dx \\
 &= \int_{-\infty}^{+\infty} \frac{1}{\sigma_1 \sigma_2 \sqrt{2\pi}} e^{-\frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1^2 \sigma_2^2} \left[x - \mu_1 - \frac{\sigma_2^2(z-\mu_1-\mu_2)}{\sigma_1^2 + \sigma_2^2} \right]^2 - \frac{(z-\mu_1-\mu_2)^2}{2(\frac{\sigma_1^2}{\sigma_1^2} + \frac{\sigma_2^2}{\sigma_2^2})}} dx \\
 &= \frac{1}{\sigma_1 \sigma_2 \sqrt{2\pi}} e^{-\frac{(z-\mu_1-\mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2} \left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(z-x-\mu_2)^2}{\sigma_2^2} \right]} dx
 \end{aligned} \tag{3.43}$$

于是证得 $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

例 3.15 实际上得到了正态分布的一个重要性质: 正态分布关于独立和运算具有封闭性, 同时例 3.15 的结论可以写成更一般的形式: X, Y 相互独立且分别服从正态分布 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$, 则其任意非零线性组合仍服从正态分布, 且

$$aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2) \tag{3.44}$$

其中 a, b 不全为 0, 这一结论可以推广到 n 个随机变量的情形, 请读者自己完成.

例 3.16 (随机变量的商) 设二维随机向量 (X, Y) 的密度函数为 $f(x, y)$,

求 $Z = \frac{X}{Y}$ 的密度函数.

解 对任意 z , 令 $D_z = \{(x, y) \mid \frac{x}{y} \leq z\}$ (如图 3.5) 则有

$$\begin{aligned}
 F_Z(z) &= P\left(\frac{X}{Y} \leq z\right) = \iint_{D_z} f(x, y) dx dy \\
 &= \int_0^{+\infty} \int_{-\infty}^{zy} f(x, y) dx dy + \int_{-\infty}^0 \int_{zy}^{-\infty} f(x, y) dx dy
 \end{aligned}$$

于是 Z 的密度函数为

$$\begin{aligned}
 f_Z(z) &= F'_Z(z) \\
 &= \int_0^{+\infty} y f(zy, y) dy - \int_{-\infty}^0 y f(zy, y) dy
 \end{aligned}$$

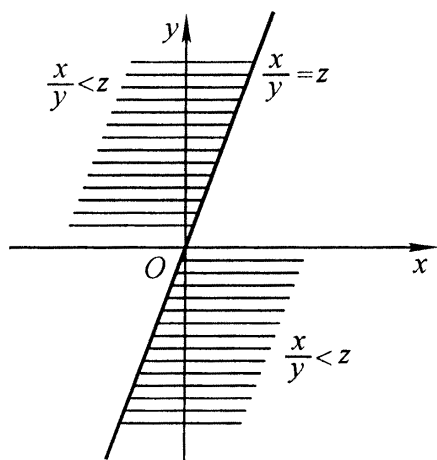


图 3.5 $D_z = \{(x, y) \mid \frac{x}{y} \leq z\}$

$$= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathbb{I}_{\{zy \leq z\}} f(y) dy \quad (3.45)$$

例 3.17 (最大值与最小值) 设 X, Y 分布函数分别为 $F(x), G(x)$, 密度函数分别为 $f(x), g(x)$, 且 X 与 Y 相互独立, 令 $M = \max\{X, Y\}, N = \min\{X, Y\}$ 求 M 和 N 的分布函数与密度函数.

$$\begin{aligned} \text{解 } F_M(z) &= P\{M \leq z\} = P\{X \leq z, Y \leq z\} = P\{X \leq z\}P\{Y \leq z\} \\ &= F(z)G(z) \end{aligned} \quad (3.46)$$

于是 M 的密度函数为:

$$f_M(z) = F'_M(z) = F'(z)G(z) + F(z)G'(z) = f(z)G(z) + F(z)g(z) \quad (3.47)$$

而 N 的分布函数为:

$$\begin{aligned} F_N(z) &= P\{N \leq z\} = P(\{X \leq z\} \cup \{Y \leq z\}) = 1 - P\{X > z, Y > z\} \\ &= 1 - P\{X > z\}P\{Y > z\} = 1 - [1 - F(z)][1 - G(z)] \end{aligned} \quad (3.48)$$

于是 N 的密度函数为:

$$f_N(z) = F'_N(z) = f(z)[1 - G(z)] + g(z)[1 - F(z)] \quad (3.49)$$

例 3.18 设二维随机向量 (X, Y) 在矩形 $G = \{(x, y) \mid 0 \leq x \leq 2, 0 \leq y \leq 1\}$ 上服从均匀分布, 试求边长为 X 和 Y 的矩形面积 S 的密度函数 $f(s)$.

解 二维随机向量 (X, Y) 的密度函数为

$$f(x, y) = \begin{cases} \frac{1}{2}, & (x, y) \in G, \\ 0, & (x, y) \notin G. \end{cases}$$

令 $F(s)$ 为 S 的分布函数, 则

$$F(s) = P\{S \leq s\} = \int_{xy \leq s} f(x, y) dx dy,$$

显然 $s \leq 0$ 时, $F(s) = 0$, $s \geq 2$ 时 $F(s) = 1$, 而当 $0 < s < 2$ 时, 如图 3.6, 有

$$\begin{aligned} \int_{xy \leq s} f(x, y) dx dy &= 1 - \int_s^2 \int_{\frac{s}{x}}^1 \frac{1}{2} dy \\ &= \frac{s}{2} (1 + \ln 2 - \ln s) \end{aligned}$$

于是

$$F(s) = \begin{cases} 0, & s \leq 0, \\ \frac{s}{2} (1 + \ln 2 - \ln s), & 0 < s < 2 \\ 1, & s \geq 2. \end{cases}$$

从而

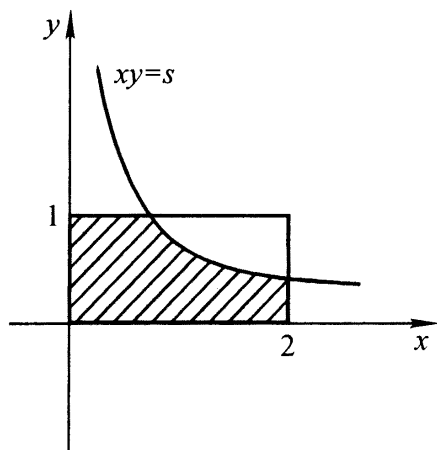


图 3.6

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

三、随机向量函数的数学期望

设随机向量 (X, Y) 的函数 $Z = g(X, Y)$ 的数学期望存在, 为求 EZ , 与一元随机变量函数的数学期望一样, 并不必先求函数 Z 的分布(事实上, 求 Z 的分布往往十分困难). 类似于一元情形, 我们按如下方式来计算 $Z = g(X, Y)$ 的数学期望:

如果 (X, Y) 是二维离散型随机向量, 令其概率分布为 $P\{X = x_i, Y = y_j\} = p_{ij}$, $i, j = 1, 2, \dots$, 则

$$EZ = Eg(X, Y) = \sum_{i,j} g(x_i, y_j) p_{ij} \quad (3.50)$$

如果 (X, Y) 是二维连续型随机向量, 令其密度函数为 $f(x, y)$, 则

$$EZ = Eg(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y) dx dy \quad (3.51)$$

例如, 设 $g(X, Y) = XY$, 则有

$$EXY = \begin{cases} \sum_{i,j} x_i y_j p_{ij}, & \text{若}(X, Y) \text{ 为离散型,} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy f(x, y) dx dy, & \text{若}(X, Y) \text{ 为连续型.} \end{cases} \quad (3.52)$$

例 3.19 设 (X, Y) 的密度函数如例 3.4 的 $g(x, y)$, 即

$$g(x, y) = \begin{cases} 8xy, & 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0, & \text{其他.} \end{cases}$$

求 EXY .

$$\begin{aligned} \text{解 } EXY &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy g(x, y) dx dy \\ &= \int_0^1 dx \int_0^1 xy \cdot 8xy dy \\ &= \frac{4}{9}. \end{aligned}$$

例 3.20 二维离散型随机向量 (X, Y) 的概率分布由例 3.2 给出, 求 EXY .

$$\begin{aligned} \text{解 } EXY &= 0 \times (-1) \times 0.1 + 0 \times 0 \times 0.2 + 0 \times 2 \times 0 \\ &\quad + 1 \times (-1) \times 0.3 + 1 \times 0 \times 0.5 + 1 \times 2 \times 0.1 \\ &\quad + 2 \times (-1) \times 0.15 + 2 \times 0 \times 0 + 2 \times 2 \times 0.1 \\ &= 0. \end{aligned}$$

例 3.21 一商店经销某种商品, 每周进货量 X 与顾客对该商品的需求量 Y 是相互独立的随机变量, 且都服从区间 $[10, 20]$ 上的均匀分布, 商店每售出一单位商品可得利润 1 000 元, 若需求量超过进货量, 商店可从其他商店调剂供

应, 这时每单位商品获利润为 500 元, 试计算此商品经销商经销该种商品每周所获平均利润.

解 设 Z 表示商店每周所获利润, 由题设有:

$$Z = g(X, Y) = \begin{cases} 1000Y, & Y \leq X, \\ 1000X + 500(Y - X) = 500(X + Y), & Y > X. \end{cases}$$

由于 (X, Y) 的密度函数为

$$f(x, y) = \begin{cases} \frac{1}{100}, & 10 \leq x \leq 20, 10 \leq y \leq 20, \\ 0, & \text{其他.} \end{cases}$$

所以有:

$$\begin{aligned} EZ &= \int_{10}^{20} \int_{10}^{20} g(x, y) f(x, y) dx dy \\ &= \int_{10}^{20} dy \int_y^{20} 1000y \cdot \frac{1}{100} dx + \int_{10}^{20} dy \int_{10}^y 500(x + y) \cdot \frac{1}{100} dx \\ &= \int_{10}^{20} y(20 - y) dy + 5 \int_{10}^{20} \left(\frac{3}{2}y^2 - 10y - 50 \right) dy \\ &= \frac{20000}{3} + 5 \times 1500 = 14166.67 (\text{元}). \end{aligned}$$

四、数学期望的进一步性质

在一元随机变量的讨论中, 我们已经给出了数学期望的若干性质, 在那里只涉及到一个随机变量, 下面我们给出的性质则涉及两个或多个随机变量.

(1) 对任意两个随机变量 X, Y , 如果其数学期望均存在, 则 $E(X + Y)$ 存在, 且

$$E(X + Y) = EX + EY; \quad (3.53)$$

(2) 设 X, Y 为任意两个相互独立的随机变量, 数学期望均存在, 则 EXY 存在, 且

$$EXY = EX \cdot EY. \quad (3.54)$$

这两个性质是针对最一般的随机变量给出的, 在本书要求范围内, 我们只就离散型和连续型进行讨论, 这里给出连续型情形的证明, 离散型情形留作练习.

证明 设 (X, Y) 的密度函数为 $f(x, y)$, 边缘密度函数为 $F_X(x)$ 和 $F_Y(y)$, 据 (3.51) 有:

关于存在性的证明, 这里略去, 事实上, 跟后面的证明过程一样, 我们可得 $E(X \cdot Y) = E(X) \cdot E(Y)$. 再由 $E(X + Y) = E(X) + E(Y)$, 即得 $E(X + Y)$ 的存在性. 由 $E(XY) = E(X)E(Y) = E(X)E(Y)$ 可得 EXY 的存在性.

$$\begin{aligned}
 (1) \quad E(X+Y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x+y)f(x,y)dx dy \\
 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xf(x,y)dx dy + \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} yf(x,y)dx dy \\
 &= \int_{-\infty}^{+\infty} x \int_{-\infty}^{+\infty} f(x,y)dy dx + \int_{-\infty}^{+\infty} y \int_{-\infty}^{+\infty} f(x,y)dx dy \\
 &= \int_{-\infty}^{+\infty} xf_x(x)dx + \int_{-\infty}^{+\infty} yf_y(y)dy \\
 &= EX + EY.
 \end{aligned}$$

(2) 由 X, Y 相互独立, 知 $f(x,y) = f_x(x)f_y(y)$, 从而

$$\begin{aligned}
 EXY &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xyf_x(x)f_y(y)dx dy \\
 &= \int_{-\infty}^{+\infty} xf_x(x)dx \int_{-\infty}^{+\infty} yf_y(y)dy \\
 &= EX \cdot EY.
 \end{aligned}$$

上述两个性质可以推广到 n 个变量情形.

(1) X_1, X_2, \dots, X_n 是任意 n 个随机变量; 数学期望均存在, 则 $E(X_1 + X_2 + \dots + X_n)$ 存在, 且

$$E(X_1 + X_2 + \dots + X_n) = EX_1 + EX_2 + \dots + EX_n \quad (3.55)$$

(2) 设 X_1, X_2, \dots, X_n 是 n 个相互独立的随机变量, 且数学期望均存在, 则 $E(X_1 X_2 \dots X_n)$ 存在, 且

$$E(X_1 X_2 \dots X_n) = EX_1 EX_2 \dots EX_n. \quad (3.56)$$

§ 3.4 随机向量的数字特征

一、协方差

同随机变量一样, 我们希望用一些数字指标来从不同角度综合反映随机向量的分布中的某些重要统计特征, 这些数字指标统称为数字特征. 由于随机向量的分布中不仅包含单个分量自身的统计规律性, 还包含了分量之间相互联系的统计规律性. 因而除了各单个分量的数字特征仍是我们需要关注的以外, 我们还需要反映分量之间相互联系的一些数字特征. “协方差”就是这样的数字特征之一.

定义 3.7 设 (X, Y) 为二维随机向量, EX, EY 均存在, 如果 $E[(X - EX)(Y - EY)]$ 存在, 则称其为随机变量 X 与 Y 的协方差, 记作 $\text{cov}(X, Y)$. 即

$$\text{cov}(X, Y) = E[(X - EX)(Y - EY)] \quad (3.57)$$

由 (3.50) 和 (3.51) 知, 如果 (X, Y) 为离散型随机向量, 概率分布为:

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

则 X 与 Y 的协方差可由下式计算:

$$\text{cov}(X, Y) = \sum_{i,j} (x_i - EX)(y_j - EY)p_{ij} \quad (3.58)$$

如果 (X, Y) 是连续型随机向量, 密度函数为 $f(x, y)$, 则 X 与 Y 的协方差可按下列下式计算:

$$\text{cov}(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - EX)(y - EY)f(x, y)dx dy \quad (3.59)$$

此外, 我们常常通过将 (3.57) 化简为

$$\text{cov}(X, Y) = EXY - EXEY \quad (3.60)$$

来计算协方差.

下一定理总结了协方差的一些基本性质, 其证明是容易的, 我们将它留给读者完成.

- 定理 3.5
- (1) $\text{cov}(X, X) = DX$;
 - (2) $\text{cov}(X, Y) = \text{cov}(Y, X)$;
 - (3) $\text{cov}(aX, bY) = ab\text{cov}(X, Y)$, a, b 为任意常数;
 - (4) $\text{cov}(C, X) = 0$, C 为任意常数;
 - (5) $\text{cov}(X_1 + X_2, Y) = \text{cov}(X_1, Y) + \text{cov}(X_2, Y)$
 - (6) 如果 X 与 Y 相互独立, 则 $\text{cov}(X, Y) = 0$

推论 设 X, Y 为任意两个随机变量, 如果其方差均存在, 则 $X + Y$ 的方差也存在, 且

$$D(X + Y) = DX + DY + 2\text{cov}(X, Y) \quad (3.61)$$

特别地, 如果 X 与 Y 相互独立, 则

$$D(X + Y) = DX + DY \quad (3.62)$$

证明 可以证明 (这里略去), 如果 X, Y 的方差存在, 则协方差 $\text{cov}(X, Y)$ 一定存在且满足下列不等式:

$$|\text{cov}(X, Y)| \leq \sqrt{DX} \sqrt{DY} \quad (3.63)$$

由协方差的性质(1)和(5)有:

$$\begin{aligned} D(X + Y) &= \text{cov}(X + Y, X + Y) \\ &= \text{cov}(X, X) + 2\text{cov}(X, Y) + \text{cov}(Y, Y) \\ &= DX + DY + 2\text{cov}(X, Y) \end{aligned}$$

当 X 与 Y 相互独立时, 由协方差的性质(6)知 $\text{cov}(X, Y) = 0$, 从而

$$D(X + Y) = DX + DY.$$

例 3.22 已知离散型随机向量 (X, Y) 的概率分布由例 3.2 给出, 求

$\text{cov}(X, Y)$

解 容易求得 X 的概率分布为

$$P\{X=0\}=0.3, P\{X=1\}=0.45, P\{X=2\}=0.25;$$

Y 的概率分布为

$$P\{Y=-1\}=0.55, P\{Y=0\}=0.25, P\{Y=2\}=0.2,$$

于是有

$$EX=0 \times 0.3+1 \times 0.45+2 \times 0.25=0.95,$$

$$EY=(-1) \times 0.55+0 \times 0.25+2 \times 0.2=-0.15.$$

在例 3.20 中, 我们已算得

$$EXY=0,$$

于是

$$\text{cov}(X, Y)=EXY-EXEY=0.95 \times (-0.15)=-0.1425.$$

例 3.23 设连续型随机向量 (X, Y) 的密度函数由例 3.4 中 $g(x, y)$ 给出, 求 $\text{cov}(X, Y)$ 和 $D(X+Y)$

解 由于 (X, Y) 的密度函数为:

$$g(x, y)=\begin{cases} 8xy, & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{其他} \end{cases}$$

在例 3.4 中, 我们已求得其边缘密度函数分别为:

$$g_X(x)=\begin{cases} 4x(1-x^2), & 0 \leq x \leq 1 \\ 0, & \text{其他} \end{cases}$$

$$g_Y(y)=\begin{cases} 4y^3, & 0 \leq y \leq 1 \\ 0, & \text{其他} \end{cases}$$

于是,

$$EX=\int_0^1 x g_X(x) dx = \int_0^1 x \cdot 4x(1-x^2) dx = \frac{8}{15}$$

$$EY=\int_0^1 y g_Y(y) dy = \int_0^1 y \cdot 4y^3 dy = \frac{4}{5}$$

在例 3.19 中我们已算得

$$EXY=\frac{4}{9}$$

从而得:

$$\text{cov}(X, Y)=EXY-EXEY=\frac{4}{9}-\frac{4}{5} \times \frac{8}{15}=-\frac{4}{225}$$

又,

$$EX^2=\int_0^1 x^2 g_X(x) dx = \int_0^1 x^2 \cdot 4x(1-x^2) dx = \frac{1}{3}$$

$$EY^2 = \int_{-\infty}^{+\infty} y^2 g_Y(y) dy = \int_0^1 y^2 \cdot 4y^3 dy = \frac{2}{3}$$

从而得:

$$DX = EX^2 - (EX)^2 = \frac{1}{3} - \left(\frac{8}{15}\right)^2 = \frac{11}{225}$$

$$DY = EY^2 - (EY)^2 = \frac{2}{3} - \left(\frac{4}{5}\right)^2 = \frac{2}{75}$$

故,

$$\begin{aligned} D(X + Y) &= DX + DY + 2\text{cov}(X, Y) \\ &= \frac{11}{225} + \frac{2}{75} + \frac{8}{225} \\ &= \frac{1}{9} \end{aligned}$$

下面的定理将 (3.61) 和 (3.62) 推广到 n 维随机向量的情形.

定理 3.6 设 (X_1, X_2, \dots, X_n) 是 n 维随机向量, 如果 $X_i, i=1, 2, \dots, n$ 的方差均存在, 则对任意实向量 $(\alpha_1, \alpha_2, \dots, \alpha_n)$, $\sum_{i=1}^n \alpha_i X_i$ 的方差必存在, 且

$$D \sum_{i=1}^n \alpha_i X_i = \sum_{i=1}^n \alpha_i^2 DX_i + 2 \sum_{1 \leq i < j \leq n} \alpha_i \alpha_j \text{cov}(X_i, X_j) \quad (3.64)$$

特别地, 如果 X_1, X_2, \dots, X_n 两两独立, 则

$$D \sum_{i=1}^n \alpha_i X_i = \sum_{i=1}^n \alpha_i^2 DX_i \quad (3.65)$$

该定理的证明与 (3.61) 和 (3.62) 的证明完全类似, 留作练习.

二、协方差矩阵

定义 3.8 设 (X_1, X_2, \dots, X_n) 为一个 n 维随机向量, X_i 的方差 $DX_i, i=1, 2, \dots, n$ 均存在, 则以 $\alpha_{ij} = \text{cov}(X_i, X_j)$ 为第 (i, j) 元素 $i, j=1, 2, \dots, n$ 的矩阵 $(\alpha_{ij})_{n \times n}$ 称为随机向量 (X_1, X_2, \dots, X_n) 的协方差矩阵, 简称协差阵.

如果记 $X = (X_1, X_2, \dots, X_n)$, 其协差阵通常记作 DX . 容易验证 (3.64) 可表为矩阵形式: 对任意实向量 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$, 有:

$$D(\alpha X) = \alpha DX \quad (3.66)$$

由 (3.66) 知, 对任意实向量 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$,

$$\alpha DX = D(\alpha X) \geq 0$$

从而知协差阵 DX 一定是非负定阵.

三、相关系数

从协方差的定义可以看出, 协方差是对两个随机变量的协同变化的度量, 因

而反映了随机变量间相互联系的内容, 然而协方差的值还受各随机变量自身取值水平的影响, 比如 X 和 Y 同时增大到 K 倍, 即 $X_1 = KX$, $Y_1 = KY$, 这时 X_1 与 Y_1 之间的统计关系和 X 与 Y 之间的统计关系应该是一样的, 然而协方差却增大到 K^2 倍, 即

$$\text{cov}(X_1, Y_1) = K^2 \text{cov}(X, Y)$$

为了避免随机变量自身取值水平不影响相互关系的度量, 我们先将每个随机变量标准化, 即取

$$X^* = \frac{X - EX}{\sqrt{DX}}, Y^* = \frac{Y - EY}{\sqrt{DY}}$$

然后计算 $\text{cov}(X^*, Y^*)$, 将 $\text{cov}(X^*, Y^*)$ 作为 X 与 Y 之间的相互关系的一种度量, 易知,

$$\text{cov}(X^*, Y^*) = \frac{\text{cov}(X, Y)}{\sqrt{DX} \sqrt{DY}} \quad (3.67)$$

定义 3.9 设 (X, Y) 是一个二维随机向量, X 和 Y 的方差均存在, 且均为正, 则称

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sqrt{DX} \sqrt{DY}} \quad (3.68)$$

为 X 与 Y 之间的相关系数.

由(3.63)知, 随机变量间的相关系数恒满足:

$$|\rho_{X,Y}| \leq 1$$

例 3.24 设 X, Y 是两个随机变量, 且 $Y = aX + b$, ($a \neq 0$, 及 b 均为常数), DX 存在且不为零, 求 $\rho_{X,Y}$.

解 $\text{cov}(X, Y) = \text{cov}(X, aX + b) = a \text{cov}(X, X) = aDX$,

$$DY = D(aX + b) = a^2 DX,$$

于是

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sqrt{DX} \sqrt{DY}} = \frac{aDX}{\sqrt{DX} \sqrt{a^2 DX}} = \frac{a}{|a|} \quad (3.69)$$

故, 当 $a > 0$ 时, $\rho_{X,Y} = 1$; 当 $a < 0$ 时, $\rho_{X,Y} = -1$.

例 3.24 表明, 当 X 与 Y 之间具有线性函数关系时, 相关系数的绝对值 $|\rho_{X,Y}|$ 达到最大值 1. 事实上例 3.24 中的条件可以略为放宽, 只要

$$P\{Y = aX + b\} = 1 \quad (3.70)$$

即除一个零概率事件以外, X 与 Y 之间具有线性函数关系, 则必有 $|\rho_{X,Y}| = 1$. 通常将零概率事件忽略不计, 如果 (3.70) 成立, 则称 X 与 Y 之间存在线性函数关系, 或简称为 X 与 Y 具有线性关系. 那么 $|\rho_{X,Y}| = 1$ 是否是对 X 与 Y 具有线性关系的完整刻画呢? 下述定理对此作出了肯定的回答.

定理 3.7 设 (X, Y) 是一个二维随机向量, DX, DY 均存在且为正, 则 $\rho_{X,Y} = 1$ 的充要条件是 X 与 Y 具有线性关系, 即, 存在常数 $a \neq 0$ 及常数 b , 使得

$$P\{Y = aX + b\} = 1$$

而且, 当 $a > 0$ 时, $\rho_{X,Y} = 1$; 当 $a < 0$ 时 $\rho_{X,Y} = -1$.

该定理的证明略去, 有兴趣的读者可以参考有关文献.

与 $\rho_{X,Y} = 1$ 完全相反的情形是 $\rho_{X,Y} = 0$. 比如, 当随机变量 X 与 Y 独立时, 设 DX, DY 均存在且为正, 则由定理 3.5(6) 知 $\text{cov}(X, Y) = 0$, 从而 $\rho_{X,Y} = 0$. 但当 $\rho_{X,Y} = 0$ 时, X 与 Y 不一定相互独立 (下面的例 3.25 将对此说明). 事实上, 两个随机变量相互独立表明随机变量取值之间不存在任何联系, 而 $\rho_{X,Y} = 0$ 上表明 X 与 Y 之间不存在线性联系, 此时, 称 X 与 Y 不相关.

易见, 如果 DX, DY 均存在且为正, 那么 X 与 Y 不相关等价下列任何一个条件:

- (1) $\text{cov}(X, Y) = 0$
- (2) $EXY = EXEY$
- (3) $D(X + Y) = DX + DY$.

例 3.25 设 θ 服从 $[-\frac{\pi}{2}, \frac{\pi}{2}]$ 上的均匀分布,

$$X = \sin \theta, \quad Y = \cos \theta$$

判断 X 与 Y 是否不相关, 是否独立.

解 由于

$$\begin{aligned} EX &= \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \sin \theta d\theta = 0, EY = \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \cos \theta d\theta = 0, \\ DX &= \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \sin^2 \theta d\theta = \frac{1}{2}, DY = EY^2 = \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \cos^2 \theta d\theta = \frac{1}{2} \\ EXY &= \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \sin \theta \cos \theta d\theta = 0 \end{aligned}$$

因此, $EXY = EXEY$, 从而 X 与 Y 不相关. 但由于 X 与 Y 满足关系:

$$X^2 + Y^2 = 1$$

所以 X 与 Y 不独立.

介于 $\rho_{X,Y} = 1$ 和 $\rho_{X,Y} = 0$ 这两个极端情形之间的一般情形是 $0 < \rho_{X,Y} < 1$, 此时, X 与 Y 之间的关系既不是准确的线性函数关系, 但也不是没有任何线性联系, 比如, 设 X 与 Y 满足下列关系:

$$Y = aX + b + Z \quad (3.71)$$

其中 $a \neq 0$, 及 b 均为常数, Z 是与 X 独立的零均值的随机变量, 显然 Y 与 X 之间不是线性函数关系, 因为 Y 的取值不能由 X 的取值所决定, 但 Y 与 X 之间也存在一定的线性联系, 当给定 X 的取值时, Y 的取值部分地由 X 的线性函数 aX

+ b 所确定, 但 Y 的取值同时还受与 X 独立的因素(变量) 所影响. 作为练习, 读者可以验证, 当 X 与 Y 满足 (3.71) 时, 其相关系数 $\rho_{X,Y}$ 满足: $0 < \rho_{X,Y} < 1$.

综上, 相关系数在某种意义上度量了两个随机变量之间的线性联系的程度, 随着 $\rho_{X,Y}$ 从 0 增加到 1, 这种线性联系的程度越来越高.

四、条件数学期望

由于随机变量之间存在相互联系, 一个随机变量的取值可能会对另一随机变量的分布产生影响, 这种影响会在数字特征上得到反映. 下面我们将要讨论的就是, 当已知一个随机变量的取值时, 与之相联系的另一随机变量的平均水平(数学期望).

定义 3.10 (1) 设 (X, Y) 是离散型随机向量, 在 $Y = y_j$ 的条件下, X 的条件概率分布为:

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

如果

$$\sum_i x_i p_{ij} < +\infty$$

则称 X 在 $Y = y_j$ 的条件下的条件数学期望存在, 并称

$$E[X | Y = y_j] = \sum_i x_i p_{ij} \quad (3.72)$$

为 X 在 $Y = y_j$ 的条件下的条件数学期望.

(2) 设 (X, Y) 为连续型随机向量, 在 $Y = y$ 的条件下, X 的条件密度函数为 $f_{X|Y}(x|y)$, 如果

$$\int_{-\infty}^{+\infty} x f_{X|Y}(x|y) dx < +\infty$$

则称 X 在 $Y = y$ 的条件下的条件数学期望存在, 并称

$$E[X | Y = y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y) dx \quad (3.73)$$

为 X 在 $Y = y$ 的条件下的条件数学期望.

例 3.26 设 (X, Y) 的概率分布由例 3.2 给出, 求出 $Y = 0$ 的条件下, X 的条件数学期望.

解 在例 3.6 中, 我们已求得 $Y = 0$ 的条件下, X 的条件概率分布为:

$$P\{X = 0 | Y = 0\} = 0.8 \quad P\{X = 1 | Y = 0\} = 0.2 \quad P\{X = 2 | Y = 0\} = 0$$

于是

$$E[X | Y = 0] = 0 \times 0.8 + 1 \times 0.2 + 2 \times 0 = 0.2$$

例 3.27 记 (X, Y) 的密度函数如例 3.8 所示, 求 $E[X | Y = y] (0 \leq y < 1)$

解 由例 3.8 知, 当 $|y| < 1$ 时

$$f_{X|Y}(x|y) = \begin{cases} \frac{1}{2} \frac{1}{1-y^2}, & |x| \leq 1-y^2; \\ 0, & \text{其他.} \end{cases}$$

从而, 当 $|y| < 1$ 时,

$$E[X|Y=y] = \int_{-1-y^2}^{1-y^2} x f_{X|Y}(x|y) dx = \int_{-1-y^2}^{1-y^2} \frac{1}{2} \frac{x}{1-y^2} dx = 0$$

容易验证, 条件数学期望具有数学期望所满足的所有性质:

$$(1) C \text{ 为常数, 则 } E[C|Y=y] = C \quad (3.74)$$

(2) k_1, k_2 为常数, 且 $E[X_i|Y=y], i=1, 2$, 均存在, 则

$$E[k_1 X_1 + k_2 X_2 | Y=y] = k_1 E[X_1 | Y=y] + k_2 E[X_2 | Y=y] \quad (3.75)$$

$$(3) \text{ 如果 } X \text{ 与 } Y \text{ 独立, 则 } E[X|Y=y] = EX \quad (3.76)$$

$E[X|Y=y]$ 还可以看成 Y 的函数, 当 Y 取 y 时, 其函数值为 $E[X|Y=y]$, 我们将这一函数记作 $E[X|Y]$, 由于它是随机变量 Y 的函数, 因而它也是一个随机变量, 根据前面列举的 $E[X|Y=y]$ 的性质, 可得 $E[X|Y]$ 的下列性质:

$$(1) C \text{ 为常数, 则 } E[C|Y] = C \quad (3.77)$$

$$(2) k_1, k_2 \text{ 为常数, 则 } E[k_1 X_1 + k_2 X_2 | Y] = k_1 E[X_1 | Y] + k_2 E[X_2 | Y], \quad (3.78)$$

$$(3) X \text{ 与 } Y \text{ 独立, 则 } E[X|Y] = EX. \quad (3.79)$$

除此之外, 我们还可得到下列性质:

$$(4) g(x) \text{ 是一个任意函数, 则 } E[g(Y)X|Y] = g(Y)E[X|Y], \text{ 特别地, 有 } E[g(Y)|Y] = g(Y). \quad (3.80)$$

$$(5) \text{ 全期望公式. } E(E[X|Y]) = EX. \quad (3.81)$$

证明: (4) 由于当 $Y=y$ 时, $g(Y) = g(y)$ (为常数), 从而有

$$E[g(Y)X|Y=y] = E[g(y)X|Y=y] = g(y)E[X|Y=y]$$

于是得

$$E[g(Y)X|Y] = g(Y)E[X|Y]$$

(5) 我们只证连续型情形, 离散型留作练习.

记 $g(y) = E[X|Y=y]$, 从而 $g(Y) = E[X|Y]$,

$$\begin{aligned} E(E[X|Y]) &= E(g(Y)) = \int_{-\infty}^{+\infty} g(y) f_Y(y) dy \\ &= \int_{-\infty}^{+\infty} E[X|Y=y] f_Y(y) dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f_{X|Y}(x|y) dx f_Y(y) dy \end{aligned}$$

$$\begin{aligned}
 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [x f_{X|Y}(x|y) f_Y(y)] dx dy \\
 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [x f(x, y)] dx dy \\
 &= \int_{-\infty}^{+\infty} x \int_{-\infty}^{+\infty} f(x, y) dy dx \\
 &= \int_{-\infty}^{+\infty} x f_X(x) dx \\
 &= EX
 \end{aligned}$$

例 3.29 设 $(X, Y) \sim N(\mu, \mu, \sigma_1^2, \sigma_2^2, \rho)$ 求 $E[X|Y]$

解 由于在 $Y=y$ 的条件下, X 的条件分布为

$$N\left(\mu + \frac{1}{\sigma_2^2} (y - \mu), \sigma_1^2 (1 - \rho^2)\right)$$

$$\text{从而, } E[X|Y=y] = \mu + \frac{1}{\sigma_2^2} (y - \mu) \quad (3.82)$$

于是得:

$$E[X|Y] = \mu + \frac{1}{\sigma_2^2} (Y - \mu) \quad (3.83)$$

例 3.30 设 X, Y 满足下列关系:

$$Y = aX + b + \epsilon \quad (3.84)$$

其中 ϵ 是一个与 X 独立的, 期望为 0 的随机变量, 求 $E[Y|X]$.

解: $E[Y|X] = E(aX + b + \epsilon|X)$

$$= E[aX|X] + E[b|X] + E[\epsilon|X]$$

$$= aX + b + E[\epsilon]$$

$$= aX + b \quad (3.85)$$

四、条件期望的预测含义

在数学期望的讨论中, 我们曾指出, 在均方误差最小意义下, 数学期望 EX 实际上是对 X 的最优点值预测, 即

$$E(X - EX)^2 = \min_C E(X - C)^2 \quad (3.86)$$

在许多实际问题中, 为对某随机变量 X 进行预测时, 我们能对一个与 X 有关的随机变量 Y 进行观察, 通常称 Y 为一个信息变量. 一旦得到 Y 的观察值, 由于 X 与 Y 存在一定的统计联系, 我们便能利用这种联系来对 X 进行预测, 其预测值依 Y 的取值而定, 也就是说 X 的预测值是 Y 的函数 $g(Y)$ (通常称为预测函数). 为寻求一个合理的预测, 实际上就是寻求一个适当的函数 g_0 , 使得在一定准则下 $g_0(Y)$ 是 X 的最优预测. 比如在均方误差最小意义下, $g_0(Y)$ 作为 X 的最优预测应满足:

$$E[X - g_0(Y)]^2 = \min_g E[X - g(Y)]^2 \quad (3.87)$$

另一方面, 当我们得到 Y 的观察值 y 时, 利用 X 与 Y 的统计联系, 我们重新将 X 的分布修正为, 在 $Y=y$ 的条件下 X 的条件分布. 在这一新的分布下, 直观上容易想象, 在均方误差最小意义下 X 的新的合理预测应该是新的分布下的数学期望, 即 $E[X|Y=y]$, 可见 $E[X|Y]$ 应该是一个合理的预测函数, 严格地讲, 由 (3.87), 为说明其合理性, 我们需要证明:

$$E(X - E(X|Y))^2 = \min_g E(X - g(Y))^2$$

其证明过程大致与(3.86)的证明相似:

$$\begin{aligned} E(X - g(Y))^2 &= E((X - E[X|Y]) + (E[X|Y] - g(Y)))^2 \\ &= E(X - E[X|Y])^2 + 2E(X - E[X|Y])(E[X|Y] - g(Y)) \\ &\quad + E(E[X|Y] - g(Y))^2 \end{aligned}$$

又由 (3.81) 和 (3.80) 可知

$$\begin{aligned} &E(X - E[X|Y])(E[X|Y] - g(Y)) \\ &= E(E[(X - E[X|Y])(E[X|Y] - g(Y))|Y]) \\ &= E((E[X|Y] - g(Y))E[(X - E[X|Y])|Y]) \\ &= E((E[X|Y] - g(Y))(E[X|Y] - E[X|Y])) \\ &= 0 \end{aligned}$$

从而有:

$$\begin{aligned} E(X - g(Y))^2 &= E(X - E[X|Y])^2 + E(E[X|Y] - g(Y))^2 \\ &\quad + E(X - E[X|Y])^2 \end{aligned}$$

§ 3.5 大数定律与中心极限定理

一、依概率收敛

在微积分中, 收敛性及其极限是一个基本而重要的概念. 一个数列 a_n 收敛到 a , 记作

$$\lim_n a_n = a$$

是指对任意 $\epsilon > 0$, 总存在正整数 N , 当 $n > N$ 时恒有

$$|a_n - a| < \epsilon$$

在概率论中, 我们研究的对象是随机变量, 因而, 我们需要考虑随机变量序列的收敛性. 设 $X_1, X_2, \dots, X_n, \dots$ 是一列随机变量, 如果我们以数列的极限完全相同的方式来定义随机变量序列的收敛性, 那么 $X_1, X_2, \dots, X_n, \dots$ 收敛到一个随机变量 X 是指, 对任意 $\epsilon > 0$, 存在正整数 N , 当 $n > N$ 时, 恒有

$$|X_n - X| < \epsilon$$

由于 X_n, X 均为随机变量, 于是 $|X_n - X|$ 也是随机变量, 要求一个随机变量取值小于给定的足够小的 ϵ 未免太苛刻了, 这种苛刻的收敛性关系, 在概率论的理论和应用问题只能是一些罕见的特例, 因而对概率论中进一步问题的研究意义并不大. 为此, 我们需要对上述定义进行修正, 以适合随机变量本身的特征. 一种比较普遍的修正方式是: 我们并不要求 $n > N$ 时, $|X_n - X| < \epsilon$ 恒成立, 而只要求 n 足够大时, 出现

$$|X_n - X| < \epsilon$$

的可能性可以任意小, 于是有下列定义.

定义 3.13 设 $X, X_1, X_2, \dots, X_n, \dots$ 是一列随机变量, 如果对任意 $\epsilon > 0$, 恒有

$$\lim_n P\{|X_n - X| < \epsilon\} = 1 \quad (3.88)$$

则称 $\{X_n\}$ 依概率收敛到 X , 记作 $X_n \xrightarrow{P} X$ 或 $P - \lim_n X_n = X$

二、大数定律

在第一章中, 我们曾指出, 如果一个事件 A 的概率为 p , 那么大量重复试验中事件 A 发生的频率将逐渐稳定 (靠近) 到概率 p , 这只是一种直观的说法. 我们将对这一直观给出严格的数学表述和论证. 记 n 次试验中, 事件 A 发生的次数为 μ_n , 那么事件 A 在 n 次试验中发生的频率为 $\frac{\mu_n}{n}$, μ_n 和 $\frac{\mu_n}{n}$ 均为随机变量, 我们将 $\frac{\mu_n}{n}$ 逐渐稳定到常数 p (A 的概率) 表述为: $\frac{\mu_n}{n} \xrightarrow{P} p$, 下面的定理说明了这一表述的正确性, 从而为 “频率稳定到概率” 这一经验事实提供了理论依据.

定理 3.7 (伯努利大数定律) 设 μ_n 是 n 重伯努利试验中事件 A 发生的次数, 已知在每次试验中 A 发生的概率为 p ($0 < p < 1$), 则对任意 $\epsilon > 0$ 有

$$\lim_n P\left\{\left|\frac{\mu_n}{n} - p\right| < \epsilon\right\} = 1 \quad (3.89)$$

即 $\frac{\mu_n}{n} \xrightarrow{P} p$ 或 $P - \lim_n \frac{\mu_n}{n} = p$

证明: 显然, $E \frac{\mu_n}{n} = p$, $D \frac{\mu_n}{n} = \frac{1}{n^2} D(\mu_n) = \frac{pq}{n}$, ($q = 1 - p$) 由切比雪夫不等式, 对任意给定 $\epsilon > 0$,

$$P\left|\frac{\mu_n}{n} - p\right| > \frac{pq}{n^2}$$

从而有:

$$0 \leq \lim_{n \rightarrow \infty} P \left| \frac{\mu_n}{n} - p \right| > \lim_{n \rightarrow \infty} \frac{pq}{n^2} = 0$$

故有

$$\lim_{n \rightarrow \infty} P \left| \frac{\mu_n}{n} - p \right| > \epsilon = 0$$

如果记

$$\xi_i = \begin{cases} 1 & \text{第 } i \text{ 次试验中 } A \text{ 发生} \\ 0 & \text{第 } i \text{ 次试验中 } A \text{ 不发生} \end{cases} \quad i = 1, 2, \dots, n$$

则 $\mu_n = \sum_{i=1}^n \xi_i$, $E \xi_i = p$, 从而 (3.105) 可改写为:

$$\lim_{n \rightarrow \infty} P \left| \frac{1}{n} \sum_{i=1}^n \xi_i - \frac{1}{n} E \sum_{i=1}^n \xi_i \right| < \epsilon = 1.$$

其中 ξ_i ($i = 1, 2, \dots$) 是相互独立服从同一个 0—1 分布的随机变量序列. 我们不难将其推广到更一般的情形, 得到下列定理.

定理 3.8 (切比雪夫大数定律) 设 $\xi_1, \xi_2, \dots, \xi_n, \dots$ 是一列两两不相关的随机变量, 它们的数学期望 $E \xi_i$ 和方差 $D \xi_i$ 均存在, 且方差有界, 即存在常数 C , 使得 $D \xi_i \leq C$ ($i = 1, 2, \dots$), 则对任意 $\epsilon > 0$ 有

$$\lim_{n \rightarrow \infty} P \left| \frac{1}{n} \sum_{i=1}^n \xi_i - \frac{1}{n} \sum_{i=1}^n E \xi_i \right| < \epsilon = 1 \quad (3.90)$$

证明 由切比雪夫不等式, 有

$$\begin{aligned} P \left| \frac{1}{n} \sum_{i=1}^n \xi_i - \frac{1}{n} \sum_{i=1}^n E \xi_i \right| \geq \epsilon & \leq \frac{D \left(\frac{1}{n} \sum_{i=1}^n \xi_i \right)}{\epsilon^2} \\ & = \frac{D \sum_{i=1}^n \xi_i}{n^2 \epsilon^2} = \frac{1}{n^2 \epsilon^2} \sum_{i=1}^n D(\xi_i) \leq \frac{n \cdot C}{n^2 \epsilon^2} = \frac{C}{n \epsilon^2} \rightarrow 0 \end{aligned}$$

从而

$$\lim_{n \rightarrow \infty} P \left| \frac{1}{n} \sum_{i=1}^n \xi_i - \frac{1}{n} \sum_{i=1}^n E \xi_i \right| < \epsilon = 1.$$

推论 1 设 $\xi_1, \xi_2, \dots, \xi_n, \dots$ 是一列独立同分布的随机变量, 其数学期望和方差均存在, 记 $E \xi_i = \mu$ 则对任意 $\epsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P \left| \frac{1}{n} \sum_{i=1}^n \xi_i - \mu \right| < \epsilon = 1 \quad (3.91)$$

即

$$\frac{1}{n} \sum_{i=1}^n \xi_i \xrightarrow{P} \mu$$

推论 1 中要求方差存在, 实际上这一条件可以去掉, 相应的结论仍然成立.

定理 3.9 (辛钦大数定律) 设 $\xi_1, \xi_2, \dots, \xi_n, \dots$ 是一列相互独立同分布

的随机变量, 且数学期望存在, 记 $E\, \xi = \mu$ 则有

$$\lim_{n \rightarrow \infty} P \left| \frac{1}{n} \sum_{i=1}^n \xi_i - \mu \right| < \varepsilon = 1 \quad (3.92)$$

伯努利大数定律表明 当 n 很大时, 事件发生的频率会“靠近”其概率. 而辛钦大数定律则表明, n 次观察的算术平均值 $\frac{1}{n} \sum_{i=1}^n \xi_i$ 会“靠近”它的期望值 $\mu = E\, \xi$, 这为估计期望值提供了一条实际可行的途径.

三、中心极限定理

我们知道正态分布在概率论中占有极其重要的地位, 实际中许多随机现象均可用正态分布来描述, 为什么会这样呢? 其原因与高斯对误差的研究中使用正态分布来描述误差的分布是相似的, 我们不妨通过“误差”的考察来加以说明. 以一门炮的射程为例, 一门炮生产出来以后其制造技术和工艺等内在因素决定了其射程的基准值, 但在每次射击中, 由于震动会造成误差 ξ_1 , 每发炮弹外形上的细小差别会造成误差 ξ_2 , 每发炮弹内炸药数量或质量上的微小差异会造成误差 ξ_3 , 炮弹在前进中遇到空气流速的微小扰动而造成误差 ξ_4 等等许多原因, 每种原因引起微小误差, 有的为正, 有的为负, 都是随机的, 而炮弹射程的总误差 ξ 是许多这种随机小误差的总和, 即

$$\xi = \sum_{i=1}^n \xi_i$$

而且这许多小误差 ξ_i 可以看成是相互独立的, 因此要讨论独立随机变量和的分布问题. 中心极限定理要研究的正是大量的独立随机变量和的近似分布问题, 其结论将告诉我们, 实际上近似服从正态分布. 继而射程等于基准值加上总误差 ξ , 也近似服从正态分布. 实际中, 许多随机现象与上面的例子类似, 我们考虑某个量受许多随机因素 (主导因素除外) 的共同影响而随机取值, 那么它的分布便会近似服从正态分布, 为了使问题简单并便于掌握, 我们在这里讨论的是中心极限定理中比较特殊的情形——独立同分布的随机变量和.

定理 3.10 (林德伯格-勒维) 设 $\xi_1, \xi_2, \dots, \xi_n, \dots$ 是一列独立同分布的随机变量, 且 $E\, \xi_i = \mu, D(\xi_i) = \sigma^2 > 0, i = 1, 2, \dots$, 则有

$$\lim_{n \rightarrow \infty} P \frac{\sum_{i=1}^n \xi_i - n\mu}{\sigma \sqrt{n}} \leq x = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt \quad (3.93)$$

该定理的证明需要进一步的知识, 超出了本书要求, 在此不作介绍.

定理 3.10 实际上说明独立同分布随机变量和 $\sum_{i=1}^n \xi_i$ 的标准化在 $n \rightarrow \infty$ 时渐近服从标准正态 $N(0, 1)$, 通常记作

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}} \stackrel{a}{\sim} N(0, 1) \quad (3.94)$$

由此得知, 当 n 充分大时, $\sum_{i=1}^n X_i \stackrel{a}{\sim} N(n\mu, n^2)$, $\frac{1}{n} \sum_{i=1}^n X_i \stackrel{a}{\sim} N\left(\mu, \frac{1}{n}\right)$

我们来比较一下大数定律与中心极限定理: 大数定律实际上告诉我们, 当 n 趋向于无穷大时, 独立同分布随机变量的算术平均值 $\frac{1}{n} \sum_{i=1}^n X_i$ 依概率收敛到其

期望值 μ . 即对任意指定的 $\varepsilon > 0$, 有 $P\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| > \varepsilon \rightarrow 0$, 那么, 对固定的

$\varepsilon > 0$, n 充分大时, $\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| > \varepsilon$ 的概率究竟有多大呢? 大数定律没有告诉我们任何内容, 但中心极限定理表明:

$$\frac{1}{n} \sum_{i=1}^n X_i - \mu \stackrel{a}{\sim} N\left(0, \frac{1}{n}\right)$$

因而

$$\begin{aligned} P\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| > \varepsilon &= 1 - P\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| \leq \varepsilon \\ &= 1 - P\left|\frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\sqrt{\frac{1}{n}}}\right| \leq \frac{\varepsilon}{\sqrt{\frac{1}{n}}} \\ &= 1 - 2\Phi\left(\frac{\varepsilon}{\sqrt{\frac{1}{n}}}\right) = 2\left[1 - \Phi\left(\frac{\varepsilon}{\sqrt{\frac{1}{n}}}\right)\right] \end{aligned} \quad (3.95)$$

其中 $\Phi(x)$ 是 $N(0, 1)$ 的分布函数, 且 $\lim_{x \rightarrow \infty} \Phi(x) = 1$. 因而(3.95) 不仅导出

$P\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| > \varepsilon \rightarrow 0$, 而且得到了对较大的 n , $\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| > \varepsilon$ 的概率近似值可见中心极限定理比大数定律的结论更加“精细”.

例 3.31 一盒同型号螺丝钉共有 100 个, 已知该型号的螺丝钉的重量是一个随机变量, 期望值是 100g 标准差是 10g, 求一盒螺丝钉的重量超过 10.2kg 的概率.

解 设 X_i 为第 i 个螺丝钉的重量, $i = 1, 2, \dots, 100$ 且它们之间独立同分布,

于是一盒螺丝钉的重量为 $\sum_{i=1}^{100} X_i$, 且由 $E X_i = 100$, $D(X_i) = 10$, 知 $E \sum_{i=1}^{100} X_i = 100 \times E X_i = 10000$, $D \sum_{i=1}^{100} X_i = 100$, 由中心极限定理有

$$\begin{aligned}
 P\{ > 10200\} &= P \frac{-\frac{10000}{100}}{100} > \frac{10200 - \frac{10000}{100}}{100} \\
 &= P \frac{-\frac{10000}{100}}{100} > 2 \\
 &= 1 - P \frac{-\frac{10000}{100}}{100} \leq 2 \\
 &= 1 - \Phi(2) = 1 - 0.97725 = 0.02275
 \end{aligned}$$

作为林德伯格-勒维中心极限定理的推论，我们给出历史上著名的德莫佛-拉普拉斯中心极限定理.

定理 3.11 设 $X_n \sim B(n, p)$, $0 < p < 1$, 则

$$\lim_{n \rightarrow \infty} P \left\{ \frac{X_n - np}{\sqrt{npq}} \leq x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt \quad (3.96)$$

定理 3.11 表明，当 n 充分大时，二项分布可用正态分布来近似. 比如，设 $X_n \sim B(n, p)$ ，要计算 $P\{a \leq X_n \leq b\} = \sum_{a \leq k \leq b} C_n^k p^k q^{n-k}$ ，当 n 很大时，这个计算量是相当大的，现在，根据定理 3.11，由于 $\frac{X_n - np}{\sqrt{npq}} \xrightarrow{d} N(0, 1)$ 或等价地 $X_n \xrightarrow{d} N(np, npq)$ ，于是可以近似地用正态分布来计算上述概率. 即：

$$\begin{aligned}
 P\{a \leq X_n \leq b\} &= P \left\{ \frac{a - np}{\sqrt{npq}} \leq \frac{X_n - np}{\sqrt{npq}} \leq \frac{b - np}{\sqrt{npq}} \right\} \\
 &\approx \Phi \left(\frac{b - np}{\sqrt{npq}} \right) - \Phi \left(\frac{a - np}{\sqrt{npq}} \right),
 \end{aligned}$$

只要查一查标准正态分布函数表就很容易得到 $P\{a \leq X_n \leq b\}$ 的相当精确的值.

例 3.32 设某电站供电网有 10 000 盏电灯，夜晚每盏灯开灯的概率为 0.7，而假定开关时间彼此独立，估计夜晚同时开着的灯的盏数在 6 800 与 7 200 之间的概率.

解 表示在夜晚同时开着的灯的盏数，则 $X \sim B(10000, 0.7)$ ，于是 $E = 10000 \times 0.7 = 7000$, $D = 10000 \times 0.7 \times 0.3 = 2100$. 由中心极限定理，有

$$\begin{aligned}
 P\{6800 < X < 7200\} &= P \left\{ \frac{-200}{\sqrt{2100}} < \frac{X - 7000}{\sqrt{2100}} < \frac{200}{\sqrt{2100}} \right\} \\
 &= P \left\{ \left| \frac{X - 7000}{\sqrt{2100}} \right| < 4.36 \right\} \\
 &= 2\Phi(4.36) - 1 = 0.99999.
 \end{aligned}$$

例 3.33 某仪器由 n 个电子元件组成，每个电子元件的寿命服从 $[0, 1000]$ 上的均匀分布（单位：小时）当有 20% 的元件烧坏时，仪器便报废，求为使该仪器的寿命超过 100 小时的概率不低于 0.95， n 至少为多大？

解 设 X 表示仪器的寿命， X_i 表示第 i 个电子元件的寿命，记 $A_i = \{X_i$

100}, 表示 n 个事件 $A_i \quad i=1, 2, \dots, n$ 中发生的个数, 由于 $P\{A_i\}=0.9$, 故 $\sim B(n, 0.9)$, 显然 $\{X \geq 100\} = \frac{1}{n} \geq 0.8$, 于是有:

$$\begin{aligned} P\{X \geq 100\} &= P\left\{\frac{X}{n} \geq 0.8\right\} = P\left\{\frac{X - 0.9n}{\sqrt{n \cdot 0.9 \cdot 0.1}} \geq \frac{0.8n - 0.9n}{\sqrt{n \cdot 0.9 \cdot 0.1}}\right\} \\ &= 1 - \Phi\left(\frac{0.1n}{\sqrt{n \cdot 0.9 \cdot 0.1}}\right) = 1 - \Phi\left(\frac{\sqrt{n}}{3}\right) \geq 0.95 \end{aligned}$$

从而 $\frac{\sqrt{n}}{3} \geq 1.64$, $n \geq 25$.

习 题 三

(A)

1. 设 $F(x, y)$ 是一个二维随机向量 (X, Y) 的分布函数, $x_1 < x_2, y_1 < y_2$, 证明:

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

2. 一个袋内装有 5 个白球, 3 个红球. 第一次从袋内任意取一个球, 不放回, 第二次又从袋内任意取两个球, X_i 表示第 i 次取到的白球数, $i=1, 2$. 求 (1) (X_1, X_2) 的分布及边缘分布; (2) $P\{X_1=0, X_2=0\}, P\{X_1=X_2\}, P\{XY=0\}$.

3. 设二维随机向量 (X_1, Y_1) 及 (X_2, Y_2) 的密度函数 $f(x, y)$ 及 $g(x, y)$ 分别为:

$$\begin{aligned} f(x, y) &= \begin{cases} k_1 e^{-3x-4y}, & x > 0, y > 0 \\ 0, & \text{其他} \end{cases} \\ g(x, y) &= \begin{cases} k_2 e^{-3x-4y}, & x > y > 0 \\ 0, & \text{其他} \end{cases} \end{aligned}$$

求 (1) 常数 k_1, k_2 , (2) 边缘密度函数.

4. 设 (X, Y) 服从 $G = \{(x, y) \mid x \geq 2, 0 \leq y \leq 1\}$ 上的均匀分布, 求 (1) (X, Y) 的密度函数及分布函数; (2) X 和 Y 的边缘密度函数和边缘分布函数; (3) $P\{Y < X^2\}$

5. 设 (X, Y) 服从 $G = \{(x, y) \mid y > x > 0\}$ 上的均匀分布, 求 (1) (X, Y) 的密度函数, (2) X 和 Y 的边缘密度函数.

6. 已知 X 和 Y 的分布函数 $F_X(x), F_Y(y)$ 分别为:

$$\begin{aligned} F_X(x) &= \begin{cases} 0, & x < 0 \\ \frac{x}{2}, & 0 \leq x \leq 2 \\ 1, & x > 2 \end{cases}, & F_Y(y) &= \begin{cases} 0, & y < 1 \\ y-1, & 1 \leq y \leq 2 \\ 1, & y > 2 \end{cases} \end{aligned}$$

且 X 与 Y 相互独立, (1) 求 (X, Y) 的分布函数; (2) 令 $U = X^2, V = Y^2$, 求 (U, V) 的分布函数 $G(u, v)$; (3) 求 $P\{X < 1, Y > \frac{3}{2}\}$

7. 设 X 与 Y 独立, 证明: 对任意实数 $x_1, x_2, y_1, y_2 (x_1 < x_2; y_1 < y_2)$, 事件 $\{x_1 < X < x_2\}$ 与

事件 $\{y_1 < Y \leq y_2\}$ 独立.

8. 求第 2 题中 $X_2 = 1$ 的条件下 X_1 的条件分布.

9. 已知 (X, Y) 为离散型分布, X 的分布为: $P\{X = x_i\} = p_i^X, i = 1, 2, \dots$, 对每个 $i, i = 1, 2, \dots$, 在 $X = x_i$ 的条件下, Y 的条件分布为 $P\{Y = y_j | X = x_i\} = p_{ji|i}, j = 1, 2, \dots$. (1) 求 (X, Y) 的分布 $P\{X = x_i, Y = y_j\} = p_{ij}, i, j = 1, 2, \dots$; (2) 求 Y 的分布 $P\{Y = y_j\} = p_j^Y$; (3) 求 $Y = y_j$ 的条件下 X 的条件分布 $P\{X = x_i | Y = y_j\} = p_{i|j}, i = 1, 2, \dots$.

10. 已知 (X, Y) 的分布及边缘分布如下表:

$\begin{matrix} Y \\ X \end{matrix}$	- 1	0	1	p_i^X
0	p_{11}	p_{12}	p_{13}	$\frac{1}{2}$
1	0	p_{22}	0	$\frac{1}{2}$
p_j^Y	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	

(1) 求 (X, Y) 的联合分布表中 $p_{11}, p_{12}, p_{13}, p_{22}$ 的值;

(2) 判断 X 与 Y 是否独立.

11. 设 X 与 Y 独立, 下表列出了二维随机向量 (X, Y) 的分布、边缘分布中的部分概率值, 试将其余概率值填入表中空白处.

$\begin{matrix} Y \\ X \end{matrix}$	y_1	y_2	y_3	p_i^X
x_1		$\frac{1}{8}$		
x_2	$\frac{1}{8}$			
p_j^Y	$\frac{1}{6}$			

12. 设 (X, Y) 是二维离散型随机向量, 其分布为 $P\{X = x_i, Y = y_j\} = p_{ij}, i = 1, 2, \dots, m, j = 1, 2, \dots, n$, 称 $(p_{ij})_{m \times n}$ 为联合概率矩阵. 证明: X 与 Y 相互独立的充要条件是 $(p_{ij})_{m \times n}$ 的秩为 1.

13. 独立投掷一枚均匀骰子两次, 记 B, C 为两次中各出现的点数, 求一元二次方程 $x^2 + Bx + C = 0$ 有实根的概率 p 和有重根的概率 q .

14. 求第 3 题中的条件密度函数: $f_{X_1|X_1}(x|y), f_{Y_1|X_1}(y|x), g_{X_2|X_2}(x|y), g_{Y_2|X_2}(y|x)$.

15. 设区域 D 是由直线 $y = x - 1, y = x + 1, x = 2$ 及生标轴围成的区域. (X, Y) 服从区域 D 上的均匀分布. 求条件密度函数 $f_{Y|X}(y|x)$ 和 $f_{X|Y}(x|y)$.

16. 设 $D = \{(x, y) | a \leq x \leq b, (x) < y < (x)\}$, 其中 $(x), (x)$ 是两个任意函数, 且 $(x) < (x)$. 在 $[a, b]$ 上恒成立, (X, Y) 服从 D 上的均匀分布. 证明条件分布为均匀分布.

17. 设 (X, Y) 为连续型随机向量, 已知 X 的密度函数 $f_X(x)$ 及对一切 x , 在 $X = x$ 的条件

下 Y 的条件密度 $f_{Y|X}(y|x)$, (1) 求密度函数 $f(x, y)$; (2) 求 Y 的密度函数 $f_Y(y)$; (3) 求条件密度函数 $f_{X|Y}(x|y)$.

- 18. 判断第 3 题中 X_1 与 Y_1 , X_2 与 Y_2 分别是否独立.
- 19. 判断第 5 题中 X 与 Y 是否独立.
- 20. 已知 X 服从 $[0, 1]$ 上的均匀分布, $Y \sim N(0, 1)$, 且 X 与 Y 相互独立, 求 (X, Y) 的密度函数.
- 21. 举例说明边缘密度不能确定联合密度 (以二元正态为例).
- 22. 如果 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ 那么 (X, Y) 一定服从二元正态分布吗? 分析下面的例子:

$$(X, Y) \sim f(x, y) = \frac{1}{2} e^{-\frac{x^2+y^2}{2}} (1 + \sin x \sin y)$$

- 23. (X, Y) 如例 3.1 所示. (1) 求 $U = X + Y$, $V = X - Y$ 各自的分布, (2) 求 (U, V) 的分布.
- 24. (B, C) 如第 13 题所示, 求 $W = B + C$ 的分布.
- 25. 证明: 如果 $X_i \sim P(\lambda_i)$ (参数为 λ_i 的泊松分布), $i = 1, 2, \dots, n$, 且相互独立, 则 $\sum_{i=1}^n X_i \sim P(\sum_{i=1}^n \lambda_i)$.
- 26. 设 X_1, X_2 均服从参数为 λ 的指数分布, 且相互独立, 求 $X_1 + X_2$ 的密度函数.
- 27. 已知 (X, Y) 为一个二维随机向量, $X_1 = X + Y$, $X_2 = X - Y$, (X_1, X_2) 的密度函数为:

$$(x_1, x_2) = \frac{1}{2\sqrt{3}} e^{-\frac{1}{2} \left(\frac{(x_1-4)^2}{3} + (x_2-2)^2 \right)}$$

分别求 X 和 Y 的密度函数.

- 28. 将 (3.44) 推广到 n 个相互独立的随机变量的情形.
- 29. 已知 (X, Y) 服从 $G = \{(x, y) | 0 < x < 2, 0 < y < 1\}$ 上的均匀分布, 求 $U = \frac{X}{Y}$ 的分布函数和密度函数.
- 30. 设 (X, Y) 的分布为:

<div><div>Y</div><div>X</div></div>			
	- 1	0	1
0	0.3	0	0.3
1	0.1	0.2	0.1

求 $E(XY), E(X + Y)$

- 31. 在一次拍卖中, 两人竞买一幅名画, 拍卖以暗标形式进行, 并以最高价成交. 设两人的出价相互独立且均服从 $[1, 2]$ 上的均匀分布, 求这幅画的期望成交价.
- 32. 对离散型情形证明 (3.53) 及 (3.54).
- 33. 一袋中装有 60 个黑球, 40 个红球, 从中任取 20 个, 求取到的红球数的期望值.
- 34. 求例 1.29 中, 交通车的停车次数的数学期望.
- 35. 证明定理 3.3.

36. 求第 2 题中 X_1 与 X_2 的协方差 $\text{cov}(X_1, X_2)$
37. 求第 3 题中 (X_1, Y_1) 和 (X_2, Y_2) 的协方差 $\text{cov}(X_1, Y_1), \text{cov}(X_2, Y_2)$.
38. 求第 27 题中的 $\text{cov}(X, Y), \text{cov}(X_1, X_2)$
39. 设 (X, Y) 服从 $G = \{(x, y) | x^2 + y^2 = 1\}$ 上的均匀分布, 讨论 X 与 Y 的独立性与相关性.
40. 求例 3.23 中 X 与 Y 的相关系数 $\rho_{X,Y}$.
41. 两支股票 A 和 B, 在一个给定时期内的收益率 r_A, r_B 均为随机变量, 已知 r_A 和 r_B 的协差阵为 V :

$$V = \begin{pmatrix} 16 & 6 \\ 6 & 9 \end{pmatrix}$$

现将一笔资金按比例 $x, 1-x$ 分别投资到股票 A 和 B 上形成一个投资组合 P, 记其收益率为 r_P .

- (1) 求 r_A 和 r_B 的相关系数;
- (2) 求 D_{r_P} ;
- (3) 在不允许卖空的情况下(即 $0 \leq x \leq 1$), x 为何值时 D_{r_P} 最小, 何时 $D_{r_P} = \min(D_{r_A}, D_{r_B})$.

42. 两种证券 A, B 的收益率为 r_A 和 r_B , 人们常用收益率的方差来衡量证券的风险, 收益率的方差为正的证券称为风险证券. 如果 A, B 均为风险证券, 且 $\rho_{AB} \leq 1$, 证明 A 与 B 的任意投资组合 P (允许卖空) 必然也是风险证券. 若 $\rho_{AB} > 1$, 何时能得到无风险组合? 当 ρ_{AB} 满足什么条件时, 我们能在不允许卖空的情况下, 得到比 A, B 的风险都小的投资组合?

43. 已知某支股票价格变化率 r 与银行利率 r_f 存在一定的联系, 设 r 和 r_f 的联合分布如下:

$\begin{matrix} r \\ \backslash \\ r_f \end{matrix}$	- 3%	1%	2%	3%	4%	5%	6%	7%
1%	0.015	0.015	0.045	0.09	0.03	0.06	0.03	0.015
1.5%	0.025	0.05	0.1	0.15	0.075	0.05	0.025	0.025
2%	0.06	0.04	0.03	0.02	0.02	0.02	0.01	0

- (1) 求该股票价格的平均变化率;
- (2) 如果已知利率 $r_f = 1.5\%$, 求股票价格的平均变化率.
44. 求第 40 题中 $E[Y|X=x]$ 和 $E[Y|X]$.
45. 设 (X, Y) 服从 $D = \{(x, y) | x^2 + y^2 = 2x\}$ 上的均匀分布, 求 $E[X|Y]$ 和 $E[Y|X]$
46. 计算机有 120 个终端, 每个终端在一小时内平均有 3 分钟使用打印机, 假定各终端使用打印机与否相互独立, 求至少有 10 个终端同时使用打印机的概率.
47. 某车间有同型号机床 200 部, 每部开动的概率为 0.7, 假定各机床开关是相互独立的, 开动时每部要消耗电能 15 个单位, 问电厂最少要供应该车间多少单位电能, 才能以 95% 的概率保证不致因供电不足而影响生产?
48. 计算机在进行加法时, 每个加数取整数 (按四舍五入取最为接近的整数), 设所有加数的取整误差是相互独立的, 且它们服从 $[-0.5, 0.5]$ 上的均匀分布.

- (1) 若将 300 个数相加, 求误差总和绝对值超过 15 的概率;
 (2) 至多 n 个数加在一起, 其误差总和的绝对值小于 10 的概率为 0.9.

习 题 三

(B)

1. 一个盒子内放有 12 个大小相同的球, 其中有 5 个红球, 4 个白球, 3 个黑球. 第一次随机地摸出 2 个球, 观察后不放回, 第二次随机地摸出 3 个球, 记 X_i 表示第 i 次摸到的红球的数目, $i=1, 2$; Y_j 表示第 j 次摸到的白球数, 求

- (1) (X_1, X_2) 的分布;
 (2) X_2 的分布;
 (3) 在分别已知 $X_2=j$ ($j=0, 1, 2, 3$) 时, X_1 的条件分布;
 (4) 两次摸到的红球总数 Y 的分布;
 (5) (X_1, Y_1) 及 (X_1, Y_2) 的分布.

2. 一辆机场交通车载有 25 名乘客途经 9 个站, 每位乘客都等可能在这 9 个站中任意一站下车 (且不受其他乘客下车与否的影响), 交通车只在有乘客下车时才停车, 令随机变量 Y_i 表示在第 i 站下车的乘客数, $i=1, 2, \dots, 9$, X_i 在有乘客下车时取值为 1, 否则取值为 0. 求

- (1) (Y_i, Y_j) ($i \neq j$) 的分布及边缘分布, 并判断 Y_i 与 Y_j 是否独立;
 (2) (X_i, X_j) ($i \neq j$) 的分布及边缘分布, 并判断 X_i 与 X_j 是否独立;
 (3) $\text{cov}(X_i, X_j)$, $\text{cov}(X_i, Y_j)$;
 (4) 交通车停车次数 X 的方差.

3. 设 $X \sim b(25, p_1)$, $Y \sim b(25-X, p_2)$, 求

- (1) 已知 $X=k$ ($k=0, 1, 2, \dots, 25$) 时, Y 的条件分布.
 (2) (X, Y) 的分布
 (3) $E[Y], E[Y^2]$, $E[Y^2|X]$;
 (4) EY, DY .

4. 设 (X, Y) 的联合分布为:

$$P\{X=K, Y=S\} = \frac{N!}{K! S! (N-K-S)!} p_1^K p_2^S (1-p_1-p_2)^{N-K-S} \\ (K=0, S=0, \text{且 } K+S \leq N)$$

其中 N 是正整数, p_1, p_2 是 $(0, 1)$ 上的实数, 且 $p_1 + p_2 < 1$.

(1) 证明: $X \sim b(N, p_1)$, $Y \sim b(N, p_2)$;

(2) 证明: 在 $Y=S$ 的条件下, X 的条件分布为 $b(N-S, \frac{p_1}{1-p_2})$, 在 $X=K$ 的条件下

Y 的条件分布为 $b(N-K, \frac{p_2}{1-p_1})$;

(3) 求第三题中 Y 的边缘分布及在 $Y=S$ 的条件下 X 的条件分布.

5. Y 服从参数 X 的指数分布, 而 X 是服从 $[1, 2]$ 上的均匀分布的随机变量.

- (1) 求 (X, Y) 的密度函数;
- (2) 求 Y 的边缘密度函数;
- (3) 求已知 $Y=y$ 时 X 的条件密度函数;
- (4) $Y=1$ 时 X 的条件期望;
- (5) 求 $P\{Y \leq X\}$.
6. 设 (X, Y) 服从 $D = \{(x, y) | 0 \leq y \leq 1, y \leq x \leq 3-y\}$ 上的均匀分布.

- (1) 求 X, Y 的边缘密度函数, 并判断 X, Y 是否独立;
- (2) 求 (X, Y) 的协方差阵, 判断 X 与 Y 是否相关.
- (3) 求密度函数 $f_{Y|X}(y|x)$ 和 $f_{X|Y}(x|y)$;
- (4) 求 $E[Y|X]$ 和 $E[X|Y]$.

7. 设 $X_1 \sim p(\lambda_1), X_2 \sim p(\lambda_2)$ 证明在 $X_1 + X_2 = n$ ($n = 1, 2, \dots$) 的条件下, X_j 的条件分布为 $b(n, \frac{\lambda_j}{\lambda_1 + \lambda_2})$.

8. 假设随机变量 Y 服从参数为 $\lambda = 1$ 的指数分布, 随机变量

$$X_k = \begin{cases} 0, & \text{若 } Y \leq k \\ 1, & \text{若 } Y > k \end{cases} \quad (k = 1, 2).$$

求 (1) (X_1, X_2) 的联合分布;

- (2) $\text{cov}(X_1, X_2), \rho_{X_1, X_2}$.

9. 假设二维随机变量 (X, Y) 在矩形 $G = \{(x, y) | 0 \leq x \leq 2, 0 \leq y \leq 1\}$ 上服从均匀分布, 记

$$U = \begin{cases} 0 & \text{若 } X \leq Y \\ 1 & \text{若 } X > Y \end{cases} \quad V = \begin{cases} 0 & \text{若 } X \leq 2Y \\ 1 & \text{若 } X > 2Y \end{cases}$$

- (1) 求 (U, V) 的分布;
- (2) 求 (U, V) 的相关系数.

10. 设 $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, 利用条件期望 $E[X|Y] = \mu_1 + \frac{\rho\sigma_1}{\sigma_2}(Y - \mu_2)$, 证明 $\rho_{XY} = \rho$.

11. 设 $D[Y|X] = E[(Y - E[Y|X])^2|X]$, 通常称之为 Y 关于 X 的条件方差. 证明:

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

12. 设 X, Y 的方差存在, 相关系数为 $\rho_{X,Y}$, 现用 X 对 Y 作线性预测. 证明在均方误差最小意义下的最优线性预测为:

$$L(X) = EY + \rho_{X,Y} \cdot \frac{\sigma_Y}{\sigma_X}(X - EX)$$

13. 设二维随机向量 (X, Y) 服从 $D = \{(x, y) | 0 \leq x \leq 1, 0 \leq y \leq 2\}$ 上的均匀分布, 求 (1) $P\{3X \leq Y\}$; (2) $Z = \min\{X, Y\}$ 的密度函数.

第 4 章

数理统计的基础知识

从本章开始，我们将讨论另一主题：数理统计。数理统计是研究统计工作的一般原理和方法的科学，它主要阐述搜集、整理、分析统计数据，并据以对研究对象进行统计推断的理论和方法，是统计学的核心和基础。本章将介绍数理统计的基本概念：总体、样本、统计量与抽样分布。抽样分布泛指统计量服从的分布。为研究抽样分布，必须首先引入一些在前三章中未提及的概率分布，从而在§ 4.3 节较为详尽地介绍了三种常用的统计分布： χ^2 分布、F 分布与 t 分布。鉴于统计应用中最常见的总体服从正态分布，因此本章将重点介绍正态总体的抽样分布。数理统计与前三章阐述的概率论的基本概念与方法有密切的联系，概率论是研究数理统计的基本工具，因此熟悉与掌握前三章的内容是学习以下各章的必要前提。

§ 4.1 总体与样本

一、总体与总体分布

总体是具有一定的共同属性的研究对象全体。总体的大小与范围由具体研究与考察的目的确定。譬如，为了了解某校一年级学生“高等数学”的学习情况，该校学习“高等数学”的全体一年级学生便构成了待研究的总体。又如要了解北京市男大学生的身高和体重的分布情况，北京市的全体男大学生便组成了总体。一旦总体确定了，便称总体的每一个别成员为个体。个体与总体的关系，即集合论中元素与集合之间的关系。统计学中关心的不是每个个体的所有具体特性，而仅仅是它的某一项或某几项数量指标。在上述第一个例子中，数量指标可取为该校一年级学生“高等数学”的期末考试成绩。在第二个例子中，北京市男大学生的身高与体重是数量指标。再如，为了掌握某工厂某日生产的产品质量情况，该日生产出来的全部产品便构成总体。如产品是灯泡，可选取灯泡的使用寿命为数量指标，如产品是钢筋，则可选取钢筋的强度为数量指标。对于选定的数量指标 x （可以是向量）而言，每个个体所取的值是不同的。在

试验中,抽取了若干个个体就观察到了 X 的这样或那样的数值,因而这一数量指标 X 是一个随机变量(或向量),而 X 的分布就完全描述了总体中我们所关心的这一数量指标的分布情况. 由于我们关心的正是此数量指标,因此我们以后就把总体与数量指标 X 可能取值的全体所组成的集合等同起来,并把数量指标 X 的分布称为总体的分布,由此导出下述定义:

定义 4.1 统计学中称随机变量(或向量) X 为总体,并把随机变量(或向量) X 的分布称为总体分布.

对于上述定义,作三点说明. 首先,表示总体的 X 既可以是随机变量,也可以是随机向量. 如果当事者关心的不是个体的一项数量指标,而是两项或两项以上的数量指标时, X 便是随机向量. 但为简化讨论,本书只限于考察一项数量指标的情形. 这样,今后凡总体指的皆是随机变量.

其次,有时个体的特性很难用数量指标来描述. 例如,服装厂生产的各式时装,玩具厂生产的儿童玩具等. 这时产品的检验部门通常将产品分成若干等级. 譬如说,划分成一、二、三等品和等外品四个等级. 当出现这样的情形时,我们仍可用一个随机变量 X 来表示产品的质量. 它可取 1、2、3 和 4 四个值,分别视产品为一、二、三等和等外品而定. 一般说来,个体的定性指标皆可类似地通过上述方式转化成一个数量指标,从而也就可设定一个随机变量来表示所研究的总体.

第三,总体分布就是设定的表示总体的随机变量 X 的分布. 总体的分布,一般说来是未知的. 有时虽已知总体分布的类型(如正态分布、伯努利分布等),但不知这些分布中所含的参数(如 μ , σ^2 ; p 等),有时甚至连分布所属的类型也不能肯定. 统计学的主要任务是对总体的未知分布进行推断.

二、样本与样本分布

前面指出,作为统计研究对象的总体的分布一般说来是未知的. 为了获取对于总体的分布的知识,一般的方法是对总体进行抽样观察. 通过观察可得到总体 X 的一组数值 (x_1, x_2, \dots, x_n) , 其中每一 x_i 是从总体中抽取的某一个体的数量指标 X 的观察值. 统计学的任务就是提供科学的方法,借助这组观察值对未知的总体分布进行合理的推断. 有时,我们还需进行多次抽样观察. 显然,再作一次抽样观察所得的一组值 (x_1, x_2, \dots, x_n) 往往和前一次得到的观察值有别. 这样,考虑问题时,就不能把每次抽样观察得到的值看成是一组确定的数值. 一种合理的解释是把它看成随机向量 (X_1, X_2, \dots, X_n) 的一次实现值. 既然抽样观察的目的是为了对总体的分布进行各种分析推断,因而要求抽取的样本能很好地反映总体的特性. 由此导出对样本的下述定义.

定义 4.2 称 (X_1, X_2, \dots, X_n) 为总体 X 的简单随机样本,若 $X_1, X_2,$

..., X_n 是独立同分布的随机变量, 且与总体 X 同分布. 样本中所含分量的个数 n 称为是该样本的容量.

要求样本中的每一分量 X_i 与总体 X 同分布, 表明抽样观察时, 每一个体都是从同一总体中抽取的. 要求样本中诸分量是独立的, 则表明每一观察结果既不影响其它观察结果, 也不受其它观察结果的影响. 获得上述简单随机样本的方法称为简单随机抽样.

显然, 简单随机样本是一种非常理想化的样本. 在实际中要获得真正意义下的简单随机样本并非易事. 严格地说, 现实的总体总是有限的. 在有限的总体中, 若是有放回地抽样, 且能保证抽样的方式是充分随机的, 那么可得到一个简单随机样本. 但实际中通常是不放回的抽样, 样本中诸分量既不是同分布的, 也不是相互独立的. 不过, 如果所考察的总体规模很大, 无放回抽样与有放回抽样的区别就很小, 此时可近似地把所得的样本看成是一简单随机样本. 本书不准备对抽样方法作详细介绍, 故以下恒假定所考虑的样本均为简单随机样本, 并简称为样本.

对于样本, 我们应持有双重的理解. 在未观察具体的抽样结果时, 应把样本 (X_1, X_2, \dots, X_n) 视为一随机向量. 在理论上探讨统计方法时, 通常都是如此. 但在观察具体的抽样结果后, 样本便理解为所得的一组具体的观察值 (x_1, x_2, \dots, x_n) . 在实际统计应用中, 通常使用样本观察值. 今后约定, 以大写的英文字母 X_i 表示随机变量, 而以相应的小写英文字母 x_i 表示它的观察值, 并称样本 (X_1, X_2, \dots, X_n) 的一组具体的观察值 (x_1, x_2, \dots, x_n) 为样本值, 全体样本值组成的集合称为样本空间. 显然, 容量为 n 的样本的样本空间是 n 维实空间 R^n 中的一个子集.

设总体 X 的分布函数为 $F(x)$, 则由定义 4.2 知, 样本 (X_1, X_2, \dots, X_n) 的分布函数为

$$F(x_1, x_2, \dots, x_n) = \prod_{i=1}^n F(x_i) \quad (4.1)$$

并称之为样本分布.

特别地, 若总体 X 为连续型随机变量, 其密度函数为 $f(x)$, 则样本的密度函数为

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i). \quad (4.2)$$

并分别称 $f(x)$ 与 $f(x_1, x_2, \dots, x_n)$ 为总体密度与样本密度.

如总体 X 为离散型随机变量, 概率分布为 $p(x) = P\{X = x_i\}$, x 取遍 X 所有可能取值, 则样本的概率分布为

$$p(x_1, x_2, \dots, x_n) = P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\} = \prod_{i=1}^n p(x_i),$$

并分别称 $p(x)$ 与 $p(x_1, x_2, \dots, x_n)$ 为离散总体密度与离散样本密度.

例 4.1 称总体 X 为正态总体, 如它服从正态分布. 正态总体是统计应用中最常见的总体. 现设总体 X 服从正态分布 $N(\mu^2)$, 则其样本密度由下式给出:

$$\begin{aligned} f(x_1, x_2, \dots, x_n) &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{1}{2\sigma^2} (x_i - \mu)^2 \\ &= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2. \end{aligned} \quad (4.3)$$

例 4.2 称总体 X 为伯努利总体, 如果它服从以 $p(0 < p < 1)$ 为参数的伯努利分布, 即

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

不难算出其样本 (X_1, X_2, \dots, X_n) 的概率分布为

$$P\{X_1=i_1, X_2=i_2, \dots, X_n=i_n\}=p^{s_n}(1-p)^{n-s_n} \quad (4.4)$$

其中 $i_k(1 \leq k \leq n)$ 取 1 或 0, 而 $s_n=i_1+i_2+\dots+i_n$, 它恰等于样本中取值为 1 的分量之总数.

服从伯努利分布的总体也具有较广泛的应用背景. 概率 p 通常可视为某一实际总体 (如工厂的某一批产品) 中具有某一特征 (如废品) 的个体所占的比例, 亦称为比率. 从总体中随机抽取一个个体, 可视为一个随机试验, 试验结果可用一随机变量 X 来刻画: 若恰好抽到具有该特征的个体, 记 $X=1$; 否则, 记 $X=0$. 这样, X 便服从以 p 为参数的伯努利分布. 通常参数 p 是未知的, 故需通过抽样对其作统计推断.

例 4.3 设总体 X 服从参数为 λ 的泊松分布, (X_1, X_2, \dots, X_n) 为其样本, 则样本的概率分布为

$$\begin{aligned} P\{X_1=i_1, X_2=i_2, \dots, X_n=i_n\} &= \prod_{k=1}^n P\{X=i_k\} \\ &= \prod_{k=1}^n \frac{\lambda^{i_k}}{i_k!} e^{-\lambda} = \frac{\lambda^{s_n}}{i_1! i_2! \dots i_n!} e^{-n\lambda}, \end{aligned} \quad (4.5)$$

其中 $i_k(1 \leq k \leq n)$ 取非负整数, 而 $s_n=i_1+i_2+\dots+i_n$.

实际应用中, 对总体的分布通常不是完全无知的. 在很多情形, 根据知识与经验已能肯定分布所属的类型. 例如, 通常认为钢筋强度服从正态分布, 描述产品是否合格的随机变量自然是服从伯努利分布的; 再如记录电话呼唤次数的随机变量通常认为是服从泊松分布的. 这样, 对于总体的分布, 仅有参数是未知的. 一旦能确定总体分布的参数, 总体分布就完全已知了. 于是, 我们也可从不同的视角来看待由 (4.3)、(4.4) 与 (4.5) 式刻画的样本密度或离散样本密度, 即把这些式子中的 (x_1, x_2, \dots, x_n) 或 (i_1, i_2, \dots, i_n) 视为是一组样本值. 样本值是观察后得到的具体数值, 是已知的, 但诸参数 $(\mu^2; p$ 或

) 则是未知的. 反过来, 要由这些观察到的样本值来推断这些未知的参数值. 从这一视角出发, 统计学中又常把诸如由 (4.3) — (4.5) 确定的函数称为是未知参数的似然函数. 关于似然函数的概念将在第五章的 § 5.2 节中作详细介绍.

三、统计推断问题简述

前面的阐述已使我们对统计学要解决的问题有了一个概要的认识, 即借助总体 X 的一个样本 (X_1, X_2, \dots, X_n) , 对总体 X 的未知分布进行推断. 我们把这类问题统称为统计推断问题. 不过, 由于总体分布是未知的, 从而样本分布也不能完全确定. 这样, 为利用样本对未知的总体分布进行推断, 还需借助样本构造一些合适的统计量, 即样本的函数, 再利用所构造的统计量的“良好”性质, 对总体分布所属的类型, 或总体分布中所含的未知参数进行统计推断. 为此, 我们将在下一节对统计量展开深入的讨论.

§ 4.2 统计量

一、统计量的定义

如前所述, 当得到总体的一个样本 (X_1, X_2, \dots, X_n) 后, 为了由样本来推断总体, 往往先利用样本构造出一些具有“良好”性质的统计量. 再由这些统计量来推断未知总体. 显然, 广义地讲, 统计量可以是样本的任一函数. 不过, 鉴于构造统计量的目的是用来推断未知的总体分布, 因此在构造统计量的时候, 自然就不应包含总体分布中的未知参数. 由此便引出关于统计量的如下定义.

定义 4.3 设 (X_1, X_2, \dots, X_n) 为总体 X 的一个样本, 称此样本的任一不含总体分布未知参数的函数为该样本的统计量.

例 4.4 设总体 X 服从正态分布, $EX = 5$, $DX = \sigma^2$, σ^2 未知. (X_1, X_2, \dots, X_n) 为总体 X 的一个样本, 令

$$S_n = X_1 + X_2 + \dots + X_n, \quad \bar{X} = \frac{S_n}{n},$$

则 S_n 与 \bar{X} 均为样本 (X_1, X_2, \dots, X_n) 的统计量. 但若令

$$U = \frac{n(\bar{X} - 5)}{\sigma},$$

则 U 不是该样本的统计量, 因 U 的表示式中含有总体分布中的未知参数 σ .

对于一个给定的样本, 根据统计量的定义, 尽管可以构造出很多统计量来, 但常用的统计量并不多. 以下将介绍统计学中常用的一些统计量.

二、常用的统计量

样本均值与样本方差是最常用的统计量. 它们又分别是样本原点矩与样本中心矩的特例, 故可统称为样本的矩统计量. 顺序统计量与矩统计量有别, 它不能表为样本的显式函数. 以下恒设 (X_1, X_2, \dots, X_n) 为总体 X 的一个样本.

1. 样本均值

称样本的算术平均值为样本均值, 记为 \bar{X} , 即

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n).$$

2. 样本方差

样本方差是用来描述样本中诸分量与样本均值的均方差异的. 它有两种定义方式. 较直观的定义是

$$S_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2,$$

并称 S_0^2 为样本的未修正样本方差.

统计学中更常用的是另一种定义, 即

$$S^2 = \frac{n}{n-1} S_0^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$$

并称 S^2 为样本的修正样本方差.

初看起来, S^2 没有 S_0^2 那么自然, 但用于统计推断之目的, S^2 比 S_0^2 具有更好的统计性质, 比如, 当用 S^2 或 S_0^2 来作为总体 X 的方差 σ^2 的估计时, S^2 满足 $ES^2 = \sigma^2$, 而 $ES_0^2 < \sigma^2$, 因而用 S^2 估计 σ^2 不会产生系统的偏差, 第五章将对此作出讨论. 由于今后使用的主要是修正样本方差, 故以下简称修正样本方差为样本方差.

3. 样本标准差

正如总体的方差与标准差的关系一样, 样本标准差 S 定义为样本方差的平方根, 即

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

4. 样本原点矩

记

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k, k = 1, 2, \dots,$$

并称 A_k 为样本的 k 阶原点矩. 显然, 一阶原点矩即为样本均值. 因此可把样本原点矩视为样本均值概念的推广.

5. 样本中心矩

记

$$B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k, k = 1,$$

并称 B_k 为样本的 k 阶中心矩. 显然, 二阶中心矩即为未修正样本方差. 因此可把样本中心矩视为未修正样本方差概念的推广.

上述五种统计量可统称为样本的矩统计量, 或简称为样本矩. 它们皆可表为样本的显式函数. 若注意到函数关系强调的仅是变量之间的依赖关系, 故除样本矩外, 尚可定义如下一些统计量, 它们不能表为样本的显式函数.

6. 顺序统计量

设 (X_1, X_2, \dots, X_n) 为总体 X 的一个样本, 将样本中的诸分量按由小到大的次序排列成

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)},$$

则称 $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ 为样本的一组顺序统计量, 称 $X_{(i)}$ 为样本的第 i 个顺序统计量. 特别地, 称 $X_{(1)}$ 与 $X_{(n)}$ 分别为样本极小值与样本极大值, 并称 $X_{(n)} - X_{(1)}$ 为样本的极差.

三、枢轴量

前面已提及, 样本的统计量中是不应包含总体分布的未知参数的. 不过, 有时在统计推断问题中, 会遇到下述情形. 设 (X_1, X_2, \dots, X_n) 为总体 X 的样本, 现需对总体分布中某一未知参数 θ 进行推断 (总体分布也许还含其他的未知参数). 这时常需构造样本的一个仅含未知参数 θ , 但不再含其他未知参数的函数 $U(X_1, X_2, \dots, X_n; \theta)$. 而如此构造的样本函数, 尽管含未知参数 θ , 却服从一个已知的分布. 利用这一已知分布的知识, 可对未知参数 θ 进行统计推断. 我们将这种含有未知参数, 但其分布却已知的样本函数称为枢轴量.

例 4.5 设总体 $X \sim N(\mu, \sigma_0^2)$, 其中 σ_0^2 已知, μ 未知, (X_1, X_2, \dots, X_n) 为总体 X 的一个样本, 令

$$U = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0}.$$

样本的上述函数 U 中尽管含未知参数 μ 但不论如何, 总有 $U \sim N(0, 1)$. 故 U 是一枢轴量, 可用来对未知参数 μ 作统计推断.

§ 4.3 常用的统计分布

前面已指出, 当取得总体 X 的样本 (X_1, X_2, \dots, X_n) 后, 通常是借助样本的统计量 (或枢轴量) 对未知的总体分布进行推断的. 为了实现推断的目的

必须进一步确定相应的统计量（或枢轴量）所服从的分布. 这样就有必要补充一些在本书概率论部分中未曾提及，但在统计学中却经常用到的分布. 本节将依次介绍 χ^2 分布、F 分布与 t 分布. 鉴于这些分布在统计学中的重要性，通常统称其为常用的统计分布.

一、分位数

在统计推断中, 经常用到统计分布的一类数字特征——分位数. 在即将讨论一些常用的统计分布之前, 我们首先给出分位数的一般概念.

定义 4.4 设随机变量 X 的分布函数为 $F(x)$, 对给定的实数 $(0 < \alpha < 1)$, 如果实数 F_α 满足:

$$P\{X > F_\alpha\} = \alpha$$
 (4.6)

即 $1 - F(F_\alpha) = \alpha$ 或 $F(F_\alpha) = 1 - \alpha$ (4.7)

则称 F_α 为随机变量 X 的分布的水平 α 的上侧分位数. 或直接称为分布 (函数) $F(x)$ 的水平 α 的上侧分位数.

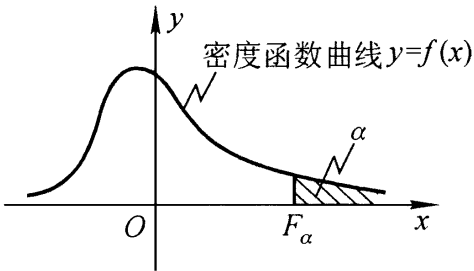


图 4-1 上侧分位数

显然, 如果 $F(x)$ 是严格单调增的, 那么其水平 α 的上侧分位数 F_α 为:

$$F_\alpha = F^{-1}(1 - \alpha)$$
 (4.8)

当 X 是连续型随机变量时, 设其密度函数为 $f(x)$, 则其水平 α 的上侧分位数 F_α 满足:

$$\int_{F_\alpha}^{+\infty} f(x)dx = \alpha$$
 (4.9)

在图形上 (如图 4-1), 介于密度函数曲线下方, x 轴上方与垂直直线 $x = F_\alpha$ 右方之间的阴影区域的面积恰等于 α .

例如, 标准正态分布 $N(0, 1)$ 的水平 α 的上侧分位数通常记作 u_α , 则 u_α 满足:

$$1 - \Phi(u_\alpha) = \alpha$$
 (4.10)

即 $\Phi(u_\alpha) = 1 - \alpha$ (4.11)

图 4-2 给出了标准正态分布的水平 α 的上侧分位数的图示.

一般讲, 直接求解分位数是很困难的, 对常见的统计分布, 在本书附录中给出了分布函数值表或分位数表, 通过查表, 可以很方便地得到分位数的值. 比如, 对给定的 α , 由 (4.11) 查标准正态分布的分布函数值表, 可得到 u_α 的值. 另外根据第二章对正态分布的讨论, 一般正态分布可借助标准化化为标准正态分布来考虑.

对于像标准正态分布那样的对称分布 (密度函数为偶函数, 关于 y 轴对

称!), 统计学中还用到另一种分位数——双侧分位数.

定义 4.5 设 X 是对称分布的随机变量, 其分布函数为 $F(x)$, 对给定的实数 α , 如果实数 T 满足:

$$P\{X \geq T\} = \alpha \quad (4.12)$$

即

$$F(T) = 1 - \alpha \quad (4.13)$$

则称 T 为随机变量 X 的分布的水平 α 的双侧分位数, 也简称为分位数, 或直接称为分布 (函数) $F(x)$ 的水平 α 的分位数.

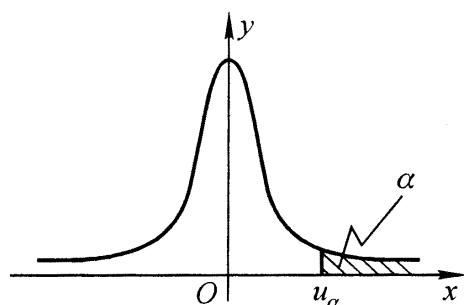


图 4-2 标准正态分布的上侧分位数

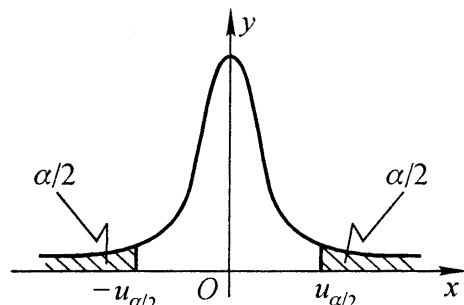


图 4-3 标准正态分布的水平 α 的双侧分位数

由于对称性, (4.13) 可改写为:

$$F(T) = 1 - \frac{\alpha}{2} \quad (4.14)$$

或

$$1 - F(T) = \frac{\alpha}{2} \quad (4.15)$$

可见, 水平 α 的分位数实际等于水平 $\frac{\alpha}{2}$ 的上侧分位数. 即有:

$$T = F_{1-\alpha/2} \quad (4.16)$$

图 4-3 以标准正态分布为例给出了双侧分位数的图示.

例 4.6 设 $\alpha = 0.05$, 求标准正态分布的水平 0.05 的上侧分位数和双侧分位数.

解 由于

$$F(u_{0.05}) = 1 - 0.05 = 0.95,$$

查标准正态分布函数值表可得

$$u_{0.05} = 1.645$$

而水平 0.05 的双侧分位数为 $u_{0.025}$, 它满足:

$$F(u_{0.025}) = 1 - 0.025 = 0.975$$

查表得:

$$u_{0.025} = 1.96.$$

二、 χ^2 分布

在第二章中, 我们曾证明: 若 $X \sim N(0, 1)$, 则 X^2 的密度函数为:

$$f(x) = \frac{1}{2} x^{-\frac{1}{2}} e^{-\frac{1}{2}x}, x > 0. \quad (4.17)$$

这一结果实际上是下述命题的特例.

命题 4.1 设 X_1, X_2, \dots, X_n 是 n 个相互独立的随机变量, 且 $X_i \sim N(0, 1)$, $i = 1, 2, \dots, n$, 则

$$X = X_1^2 + X_2^2 + \dots + X_n^2$$

的密度函数为:

$$f(x; n) = \frac{1}{2^{\frac{n}{2}} \frac{\Gamma(\frac{n}{2})}{\sqrt{\pi}}} x^{\frac{n}{2}-1} e^{-\frac{1}{2}x}, x > 0. \quad (4.18)$$

其中 $\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x} dx (a > 0)$ 是 (伽马) 函数.

* 证明 由 (4.17) 知, 当 $n = 1$ 时, (4.18) 成立. 使用数学归纳法. 设 $n = k$ 时, (4.18) 成立, 令

$$X = X_1^2 + X_2^2 + \dots + X_k^2, \quad X_{k+1} = X_1^2 + \dots + X_k^2 + X_{k+1}^2 = X + Y$$

由归纳假设及 (4.17) 知, X 的密度函数分别为

$$f(x) = \frac{1}{2^{\frac{k}{2}} \frac{\Gamma(\frac{k}{2})}{\sqrt{\pi}}} x^{\frac{k}{2}-1} e^{-\frac{1}{2}x}, x > 0;$$

$$f(y) = \frac{1}{2^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2})}{\sqrt{\pi}}} y^{-\frac{1}{2}} e^{-\frac{1}{2}y}, y > 0.$$

由于 X 与 Y 皆为非负的随机变量, 且相互独立, 由第三章的卷积公式知, 当 $z > 0$ 时, $X + Y$ 的密度函数可按下式计算:

$$\begin{aligned} f(z) &= \int_0^z f(z-y)f(y)dy \\ &= \frac{1}{2^{\frac{k+1}{2}} \frac{\Gamma(\frac{k}{2})}{\sqrt{\pi}} \frac{\Gamma(\frac{1}{2})}{\sqrt{\pi}}} \int_0^z (z-y)^{\frac{k}{2}-1} e^{-\frac{1}{2}(z-y)} y^{-\frac{1}{2}} e^{-\frac{1}{2}y} dy \\ &= \frac{e^{-\frac{1}{2}z}}{2^{\frac{k+1}{2}} \frac{\Gamma(\frac{k}{2})}{\sqrt{\pi}} \frac{\Gamma(\frac{1}{2})}{\sqrt{\pi}}} \int_0^z (z-y)^{\frac{k}{2}-1} y^{-\frac{1}{2}} dy \\ &= \frac{e^{-\frac{1}{2}z} z^{\frac{k+1}{2}-1}}{2^{\frac{k+1}{2}} \frac{\Gamma(\frac{k}{2})}{\sqrt{\pi}} \frac{\Gamma(\frac{1}{2})}{\sqrt{\pi}}} \int_0^1 (1-\frac{y}{z})^{\frac{k}{2}-1} \frac{y}{z}^{-\frac{1}{2}} d\frac{y}{z} \end{aligned}$$

$$\begin{aligned}
 &= \frac{e^{-\frac{1}{2}z} z^{\frac{k+1}{2}-1}}{2^{\frac{k+1}{2}} \frac{k}{2} \frac{1}{2}} \int_0^1 (1-t)^{\frac{k}{2}-1} t^{\frac{1}{2}-1} dt \quad \text{令 } t = \frac{y}{z} \\
 &= \frac{\frac{k}{2} \frac{1}{2}}{2^{\frac{k+1}{2}} \frac{k}{2} \frac{1}{2}} z^{\frac{k+1}{2}-1} e^{-\frac{1}{2}z} \\
 &= \frac{1}{2^{\frac{k+1}{2}} \frac{k+1}{2}} z^{\frac{k+1}{2}-1} e^{-\frac{1}{2}z}
 \end{aligned}$$

其中倒数第二个等式中使用了贝塔函数的定义:

$$(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx,$$

在最后一个等式中, 使用了贝塔函数与伽马函数间的关系:

$$(a, b) = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)}.$$

这表明, 当 $n = k+1$ 时, (4.18) 式也成立. 从而命题证毕.

定义 4.6 一个随机变量称为服从以 n 为自由度的 χ^2 分布, 如果其密度函数由 (4.18) 给出, 记作 $X \sim \chi^2(n)$

当自由度 $n \geq 3$ 时, χ^2 分布密度函数的曲线皆为单峰曲线. 曲线从原点开始递增, 在 $x = n-2$ 处达最大值, 然后递减, 渐近于 x 轴. 显然, 函数图形关于垂直线 $x = n-2$ 不对称. 随着自由度 n 的增大, 曲线的峰值向右移动, 图形变得比较平缓并趋于对称, 因此可用正态分布来近似. 当自由度 $n=2$ 时, 曲线在 $x=0$ 处达最大值, 然后递减; 当自由度 $n=1$ 时, 曲线在 $x=0$ 处取无穷大值, 这时 y 轴为其垂直渐近线. 在后两种情形, x 轴也同为这两种密度曲线的水平渐近线. 图 4-4 给出了 $n=1, 4, 10, 20$ 时, χ^2 分布的密度函数曲线.

由定义 4.6 与正态分布的性质, 不难推出下述命题.

命题 4.2

(1) 若 $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, 且 X 与 Y 相互独立, 则 $(X+Y) \sim \chi^2(m+n)$.

(2) 若 $X \sim \chi^2(n)$, 则 $E[X] = n$, $D[X] = 2n$.

证明 设 X_1, X_2, \dots, X_{m+n} 独立同分布, 且均服从标准正态分布.

(1) 由于 $X \sim \chi^2(m)$, 据定义 4.6 与命题 4.1, X 与 $X_1^2 + X_2^2 + \dots + X_m^2$ 同分布, Y 与 $X_{m+1}^2 + X_{m+2}^2 + \dots + X_{m+n}^2$ 同分布, 再由 X 与 Y 独立

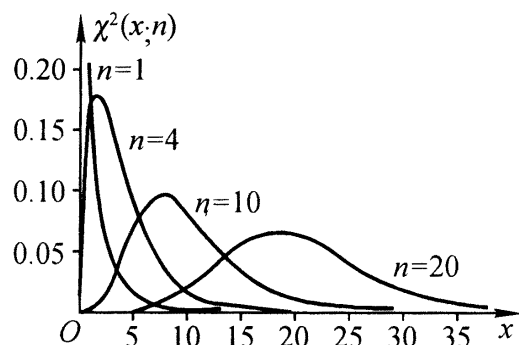


图 4-4 χ^2 分布的密度函数曲线

知, $X + Y$ 与 $X_1^2 + X_2^2 + \dots + X_{m+n}^2$ 同分布, 从而得 $X + Y \sim \chi^2(m+n)$.

(2) 设 X_1, X_2, \dots, X_n 相互独立且均服从标准正态分布, 则由 $X \sim \chi^2(n)$ 知 X 与 $X_1^2 + X_2^2 + \dots + X_n^2$ 同分布, 于是

$$EX = E \sum_{i=1}^n X_i^2 = \sum_{i=1}^n EX_i^2 = \sum_{i=1}^n DX_i = n.$$

此外, 由于 $E[X_i^4] = 3$ (见习题四 (A) 的第 9 题), 便知

$$D[X_i^2] = E[X_i^4] - (E[X_i^2])^2 = 3 - 1 = 2$$

再因 X_1, X_2, \dots, X_n 相互独立, 即得

$$D[X] = D \sum_{i=1}^n X_i^2 = \sum_{i=1}^n D[X_i^2] = 2n.$$

上述命题中第一个结论实际上说明 χ^2 分布同正态分布一样具有可加性.

由于 χ^2 分布是常用的统计分布, 但又难于利用其密度函数 $f(x; n)$ 进行直接计算, 通常也为其制定了统计用表. 附表 3 中对自由度 $n = 45$ (或 $n = 50$) 的 χ^2 分布给出了水平 α 的上侧分位数 $\chi^2_\alpha(n)$ 之值, 由前述关于一般分布的上侧分位数的定义, 当 $X \sim \chi^2(n)$ 时,

$$P\{X > \chi^2_\alpha(n)\} = P\{X < \chi^2_{1-\alpha}(n)\} = \alpha.$$

因 $f(x; n)$ 不是对称函数, 故对 χ^2 分布而言, 不存在双侧分位数, 但在以后统计推断中, 将用到等式:

$$P(\{X < \chi^2_{1-\frac{\alpha}{2}}(n)\} \cap \{X > \chi^2_{\frac{\alpha}{2}}(n)\}) = 0.$$

或

$$P\{\chi^2_{1-\frac{\alpha}{2}}(n) < X < \chi^2_{\frac{\alpha}{2}}(n)\} = 1 - \alpha.$$

例如, 设 $X \sim \chi^2(10)$, 取水平 $\alpha = 0.05$, 查表可知

$$P\{X < 3.940\} = P\{X > 18.307\} = 0.05,$$

$$P\{3.247 \leq X \leq 20.483\} = 0.95.$$

当自由度 n 充分大 (如 $n > 45$ 或 $n > 50$) 时, χ^2 分布可近似地看作正态分布, 于是由正态分布的分位数可近似地求得 χ^2 分布的分位数.

三、F 分布

接下来, 我们考虑另一个常用的统计分布.

设 $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, 且 X 与 Y 相互独立, 记

$$Z = \frac{(X/m)}{(Y/n)} = \frac{n}{m} \frac{X}{Y}, \quad (4.19)$$

显然 Z 是 X 和 Y 的函数, 进而可视为相互独立标准正态分布随机变量 X_1, X_2, \dots, X_{m+n} 的函数.

下述命题给出了 F 分布的密度函数的表示式.

命题 4.3 设 Z 由 (4.19) 所定义, 则 Z 的密度函数为:

$$f(x; m, n) = \frac{1}{\frac{m}{2}, \frac{n}{2}} \frac{m}{n} \frac{m}{n} x^{\frac{m}{2}-1} \left(1 + \frac{m}{n} x\right)^{-\frac{1}{2}(m+n)}, x > 0. \quad (4.20)$$

* 证明 首先, 因 $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, 由定义 4.6 知, X 与 Y 的密度函数分别为

$$f_X(x) = \frac{1}{2^{\frac{m}{2}} \frac{m}{2}} x^{\frac{m}{2}-1} e^{-\frac{1}{2}x}, x > 0;$$

$$f_Y(y) = \frac{1}{2^{\frac{n}{2}} \frac{n}{2}} y^{\frac{n}{2}-1} e^{-\frac{1}{2}y}, y > 0.$$

设 $Z_0 = \frac{X}{Y}$, 从而 $Z = \frac{n}{m} Z_0$.

由于 X 与 Y 皆为非负的随机变量, 且相互独立, 由第三章的例 3.12 知, 当 $z > 0$ 时, 随机变量 Z_0 的密度函数可按下式计算:

$$\begin{aligned} f_{Z_0}(z) &= \int_0^\infty f_X(yz) f_Y(y) y dy \\ &= \frac{1}{2^{\frac{m+n}{2}} \frac{m}{2} \frac{n}{2}} \int_0^\infty (yz)^{\frac{m}{2}-1} e^{-\frac{1}{2}yz} y^{\frac{n}{2}-1} e^{-\frac{1}{2}y} y dy \\ &= \frac{z^{\frac{m}{2}-1}}{2^{\frac{m+n}{2}} \frac{m}{2} \frac{n}{2}} \int_0^\infty y^{\frac{m+n}{2}-1} e^{-\frac{y}{2}(z+1)} dy \\ &= \frac{z^{\frac{m}{2}-1} (1+z)^{-\frac{m+n}{2}}}{\frac{m}{2} \frac{n}{2}} \int_0^\infty t^{\frac{m+n}{2}-1} e^{-t} dt \quad \text{令 } t = \frac{y(z+1)}{2} \\ &= \frac{\frac{m+n}{2}}{\frac{m}{2} \frac{n}{2}} z^{\frac{m}{2}-1} (1+z)^{-\frac{m+n}{2}} \\ &= \frac{1}{\frac{m}{2}, \frac{n}{2}} z^{\frac{m}{2}-1} (1+z)^{-\frac{m+n}{2}} \end{aligned}$$

再由于 $Z = \frac{n}{m} Z_0$, 当 $z > 0$ 时, 即知随机变量 Z 的密度函数可表为

$$f_Z(z) = \frac{m}{n} f_{Z_0} \left(\frac{m}{n} z \right) = \frac{1}{\frac{m}{2}, \frac{n}{2}} \frac{m}{n} \frac{m}{n} z^{\frac{m}{2}-1} \left(1 + \frac{m}{n} z\right)^{-\frac{m+n}{2}}.$$

定义 4.7 一个随机变量 X 称为服从第一自由度为 m , 第二自由度为 n 的

F 分布, 记作 $X \sim F(m, n)$, 如果其密度函数由 (4.20) 给出. 此外, 由命题 4.3 不难推知, 若 $X \sim F(m, n)$, 则 $X^{-1} \sim F(n, m)$.

F 分布的密度函数曲线也为单峰曲线, 当第一自由度 $m \geq 3$ 时, 曲线在

$$x^* = \frac{m-2}{m} \cdot \frac{n}{n+2}$$

处达最大值. 显见 $x^* < 1$, 换言之, 图形的峰值恒在小于 1 处取到. 此外, 不难看出, 当两个自由度 m 与 n 都变得越来越大时, x^* 接近 1, 从而函数图形就在非常接近 1 的地方达到最高点. 图 4-5 给出了若干 F 分布的密度函数曲线.

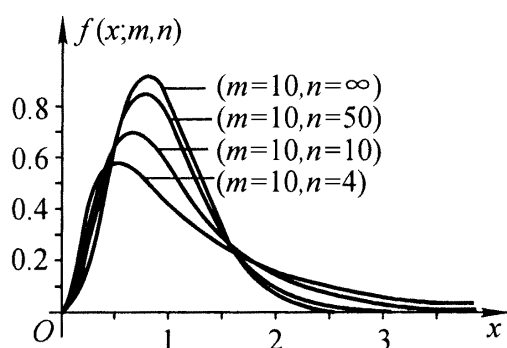


图 4-5 F 分布的密度函数曲线

由于 F 分布是常用的统计分布, 且也难于利用其密度函数进行直接计算, 因此对它也制出了统计用表. 附表 4 中对充分小的 α 值列出了 F 分布的水平 α 的上侧分位数 $F_{1-\alpha}(m, n)$ 之

值, 根据上侧分位数的定义, 当 $X \sim F(m, n)$ 时,

$$P\{X > F_{1-\alpha}(m, n)\} = P\{X < F_{1-\alpha}(m, n)\} = 1 - \alpha.$$

由于 X 为非负随机变量, 其密度函数 $f(x; m, n)$ 自然不是对称函数. 这样, 对 F 分布而言, 也不存在双侧分位数. 但在统计推断中将使用关系式:

$$P(\{X < F_{1-\frac{\alpha}{2}}(m, n)\} \cap \{X > F_{\frac{\alpha}{2}}(m, n)\}) = \alpha.$$

或

$$P\{F_{1-\frac{\alpha}{2}}(m, n) < X < F_{\frac{\alpha}{2}}(m, n)\} = 1 - \alpha.$$

作为例子, 设 $X \sim F(5, 10)$, 查表 4 知

$$P\{X > 3.33\} = 0.05, P\{X > 4.24\} = 0.025.$$

又设 $Y \sim F(10, 5)$, 查表可得

$$P\{Y > 4.74\} = 0.05, P\{Y > 6.62\} = 0.025.$$

由于附表 4 仅对充分小的 α 值给出了 F 分布的上侧分位数, 当 α 接近于 1 时, 可利用下式求出所需的上侧分位数:

$$F_{\alpha}(m, n) = \frac{1}{F_{1-\alpha}(n, m)}, \quad (4.21)$$

这是因为由 $X \sim F(m, n)$, 可推知 $X^{-1} \sim F(n, m)$.

例如, 由 (4.21) 式知,

$$F_{0.95}(m, n) = \frac{1}{F_{0.05}(n, m)}, F_{0.975}(m, n) = \frac{1}{F_{0.025}(n, m)}$$

这样, 当 $X \sim F(5, 10)$ 时, 查表可知

$$P\{X < \frac{1}{4.74}\} = 0.05,$$

$$P \frac{1}{6.62} X \leq 4.24 = 0.95.$$

四、t 分布

前面介绍的 χ^2 分布和 F 分布均可视为正态分布随机变量函数的分布, 现在将介绍的分布——t 分布, 与 F 分布有紧密联系, 因此追根溯源, 也可把 t 分布视为正态随机变量函数的分布.

设 $X \sim N(0, 1)$, $Y \sim \chi^2(n)$, 且 X 与 Y 相互独立. 记

$$T = \frac{X}{Y/n}, \quad (4.22)$$

由于 $Y \sim \chi^2(n)$, 仅从 (4.22) 式便可断言, 随机变量 T 可视为正态随机变量的函数. 事实上进一步由 (4.22) 可知

$$T^2 = \frac{X^2}{Y/n} \sim F(1, n).$$

这表明 t 分布与 F 分布有紧密联系. 我们将利用这一关系简便地导出 (4.22) 定义的随机变量 T 的密度函数的表示式.

命题 4.4 由 (4.22) 所定义的随机变量 T 的密度函数为:

$$f_T(x; n) = \frac{1}{\frac{1}{2} \sqrt{\frac{n}{2}}} \frac{1}{n} \left(1 + \frac{x^2}{n} \right)^{-\frac{n+1}{2}}, \quad -\infty < x < \infty. \quad (4.23)$$

* 证明 首先注意到, T 的表示式 (4.22) 中的分母为非负连续型随机变量, 且与 X 独立; 而 X 服从标准正态分布, 从而具有对称的密度函数. 因此, 由两独立随机变量商的密度函数公式可推出随机变量 T 的密度函数也是对称函数.

其次, 以 $f_T(t)$ 与 $f_{|T|}(t)$ 分别表示 T 与 $|T|$ 的密度函数, 由于 T 具有对称的密度函数, 不难证明, 当 $t > 0$ 时, $f_T(t) = \frac{1}{2} f_{|T|}(t)$ (见习题四 (B) 的第 7 题).

现设 $F = \frac{X^2}{Y/n}$, 则 $F \sim F(1, n)$, 且由命题 4.3 知, 随机变量 F 的密度函数为

$$f_F(x) = \frac{1}{\frac{1}{2} \sqrt{\frac{n}{2}}} \frac{1}{n} \left(\frac{x}{n} \right)^{-\frac{1}{2}} \left(1 + \frac{x}{n} \right)^{-\frac{n+1}{2}}, \quad x > 0.$$

再注意到

$$|T| = \sqrt{\frac{X^2}{Y/n}} = \sqrt{F}$$

由第二章习题二 (A) 的第 37 题便知, 当 $t > 0$ 时, 应有:

$$f_{\text{OT}}(t) = 2tf_F(t^2) = \frac{2t}{\frac{1}{2}, \frac{n}{2}} \frac{1}{n} \frac{t^2}{n}^{-\frac{1}{2}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

$$= \frac{2}{\frac{1}{2}, \frac{n}{2}} \frac{1}{n} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

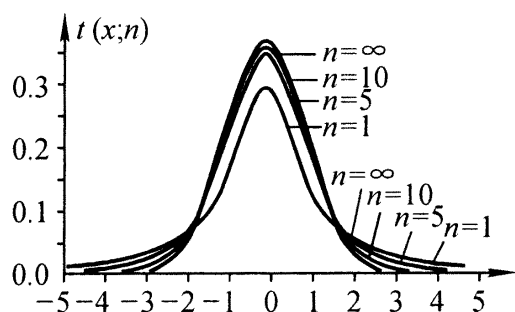
于是, 当 $t > 0$ 时,

$$f_T(t) = \frac{1}{2} f_{\text{OT}}(t) = \frac{1}{\frac{1}{2}, \frac{n}{2}} \frac{1}{n} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

这表明, 当 $x > 0$ 时, (4.23) 式是成立的. 再由于 $f_T(t)$ 是对称函数, 即知当 $x < 0$ 时, (4.23) 式也成立.

定义 4.8. 一个随机变量 X 称为服从自由度为 n 的 t 分布, 记作 $X \sim t(n)$, 如果它的密度函数由 (4.23) 给出.

t 分布的密度曲线也为单峰曲线, 但关于 y 轴对称, 且在 $x = 0$ 处取到最大



值. x 轴为其水平渐近线. 图 4-6 给出了自由度 $n = 1, 5, 10, \infty$ 时, t 分布的密度函数曲线.

当自由度 n 很大时, t 分布接近于标准正态分布, 这是因为

$$\lim_{n \rightarrow \infty} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} = e^{-\frac{1}{2}x^2}.$$

图 4-6 t 分布的密度函数曲线

图 4-6 中, $n = \infty$ 时的 t 分布的密度函数曲线, 即为标准正态分布的密度函数曲线. 从图

4-6 中看出, 与标准正态密度曲线相比, t 分布的曲线以较慢速率趋于 x 轴. 换言之, t 分布的尾部比标准正态分布的尾部有更大的概率.

附表 5 对于充分小的 α 值给出了 t 分布的水平 α 的上侧分位数 $t(n)$ 之值. 由于 t 分布具有对称的密度函数, 当 α 接近 1 时, 可按下式求出相应的上侧分位数:

$$t(n) = -t_{1-\alpha}(n), \quad (4.24)$$

根据上侧分位数的定义, 如 $X \sim t(n)$, 不难推出:

$$P\{X > t(n)\} = P\{X < -t(n)\} = \alpha.$$

再由于 t 分布具有对称的密度函数, 从而具有双侧分位数, 即有

$$P\{|X| > t_{\alpha/2}(n)\} = \alpha.$$

其中 $t_{\alpha/2}$ 即为水平 α 的双侧分位数.

例如, 设 $X \sim t(8)$, 取水平 $\alpha = 0.05$, 查表可知 $t(8) = 1.860$, $t_{\alpha/2}(8) = 2.306$, 故有

$$P\{X > 1.860\} = P\{X < -1.860\} = P\{X > 2.306\} = 0.05.$$

此外, 由于自由度 n 充分大时, t 分布近似于标准正态分布, 故有 $t(n) \rightarrow u$, 其中 u 为标准正态分布的上侧分位数.

§ 4.4 抽样分布

总体的分布往往是未知的, 或是部分地未知的. 根据实际问题的需要, 有时需对总体未知的重要数字特征 (如总体数学期望、总体方差) 或总体分布中所含的未知参数进行统计推断. 这类问题称为参数统计推断. 在参数统计推断问题中, 经常需要利用总体的样本构造出合适的统计量 (或枢轴量), 并使其服从或渐近地服从已知的确定分布. 统计学中泛称统计量 (或枢轴量) 的分布为抽样分布. 讨论抽样分布的途径有两个. 一是精确地求出抽样分布, 并称相应的统计推断为小样本统计推断; 另一种方式是让样本容量趋于无穷, 并求出抽样分布的极限分布. 然后, 在样本容量充分大时, 再利用该极限分布作为抽样分析的近似分布, 既而对未知参数进行统计推断, 因此称相应的统计推断为大样本统计推断, 本节重点讨论正态总体的抽样分布, 属小样本统计范畴; 此外, 也简要地讨论了一般总体的某些抽样分布的极限分布, 属大样本统计范畴.

一、正态总体的抽样分布

前一节中已反复强调指出, 三种常用的统计分布 (χ^2 分布、 F 分布与 t 分布) 均可视为正态随机变量函数的分布. 这些分布的讨论为正态总体的抽样分布作了必要的准备. 不过, 在一般地讨论正态分布的抽样分布以前, 我们还需首先介绍一个涉及正态总体样本均值与样本方差的抽样分布的定理. 它是讨论正态总体抽样分布的一个基础性定理.

定理 4.1 设总体 $X \sim N(\mu, \sigma^2)$, (X_1, X_2, \dots, X_n) 是其容量为 n 的一个样本, \bar{X} 与 S^2 分别为此样本的样本均值与样本方差, 则有

$$(1) \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right);$$

$$(2) \frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1);$$

$$(3) \bar{X} \text{ 与 } S^2 \text{ 相互独立.}$$

上述定理中, 结论 (1) 的证明是容易的, 请读者自己完成.

结论 (2) 与 (3) 的严格证明需要用到关于多重积分的变量替换公式, 此外还要利用正交矩阵的一些性质, 数学推导的技巧性很强, 从而在正文中略去其证明. 有兴趣的读者可试着根据习题四 (B) 中 4-6 题所提供的思路去证明, 也可参阅本书习题答案中此三题的证明概要.

有了上述关于正态总体的样本均值与样本方差的抽样分布的基础性定理,再结合上一节中关于常用统计分布的有关论述,便可容易地构造出单正态总体与双正态总体中样本的一些统计量(或枢轴量),并使之服从确定的已知分布,以下分别细述之.

1. 单正态总体的抽样分布

下述定理为讨论单正态总体参数的置信区间(见§ 5.5节)与假设检验(见§ 6.2节)提供了合适的统计量(或枢轴量).

定理 4.2 设 (X_1, X_2, \dots, X_n) 为正态总体 $X \sim N(\mu, \sigma^2)$ 的样本, \bar{X} 与 S^2 分别为该样本的样本均值与样本方差, 则有

$$(1) \quad U = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0, 1);$$

$$(2) \quad \frac{n-1}{2} S^2 \sim \chi^2(n-1);$$

$$(3) \quad T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

证明 结论(1)是定理4.1(1)的直接推论. 结论(2)已由定理4.1给出. 再因

$$T = \frac{U}{\sqrt{\frac{n-1}{2} S^2 / (n-1)}},$$

且由定理4.1知 U 与 $\frac{n-1}{2} S^2$ 相互独立, 故由本定理的结论(1)、(2)与命题4.3, 立知 $T \sim t(n-1)$.

作为概率论的推导结果, 定理4.2的结论无需涉及正态分布中的两个参数 μ 与 σ^2 是否已知. 换言之, 只要 μ 与 σ^2 分别表示该正态总体的期望与方差, 不论它们是否已知, 相应的结论都是成立的. 不过, 在统计学的应用中, 则需关心这两个参数是否已知. 只有当 μ 与 σ^2 都已知时, 定理4.2中提及的三个样本函数方可视为是样本的统计量. 退一步, 若 μ 未知, T 可视为是一枢轴量; 若 σ^2 未知, $\frac{n-1}{2} S^2$ 也可视为是一枢轴量, 因此它们也可用于对未知参数 μ 与 σ^2 作统计推断. 但在 μ 与 σ^2 都未知时, U 甚至不能作为枢轴量. 也正是基于这一原因, 在关于单正态总体参数的统计推断中, 当 σ^2 未知时, 需设计出枢轴量 T 来替换样本函数 U .

2. 双正态总体的抽样分布

在统计学的应用中, 有时要比较两个正态总体的参数. 下述定理为比较两个正态总体的参数提供了合适的统计量(或枢轴量), 它们在讨论双正态总体参数的置信区间(见§ 5.5节)与假设检验(见§ 6.3节)时要用到.

定理 4.3 设 $X \sim N(\mu_1, \sigma_1^2)$ 与 $Y \sim N(\mu_2, \sigma_2^2)$ 是两个相互独立的正态总体. 又设 $(X_1, X_2, \dots, X_{n_1})$ 是总体 X 的容量为 n_1 的样本, \bar{X} 与 S_1^2 分别为该样本的样本均值与样本方差. 再设 $(Y_1, Y_2, \dots, Y_{n_2})$ 是总体 Y 的容量为 n_2 的样本, \bar{Y} 与 S_2^2 分别为此样本的样本均值与样本方差. 另记 S^2 是 S_1^2 与 S_2^2 的加权平均, 即令

$$S^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2^2$$

则有

$$(1) U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1);$$

$$(2) F = \frac{\frac{\sigma_1^2}{n_1} S_1^2}{\frac{\sigma_2^2}{n_2} S_2^2} \sim F(n_1 - 1, n_2 - 1);$$

$$(3) \text{ 当 } \frac{\sigma_1^2}{n_1} = \frac{\sigma_2^2}{n_2} = \sigma^2 \text{ 时,}$$

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

证明 (1) 由定理 4.1 的结论 (1) 知

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

再因两个总体 X 与 Y 相互独立, 从而它们的样本均值 \bar{X} 与 \bar{Y} 也相互独立, 故

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

再记

$$\mu_0 = \mu_1 - \mu_2, \quad \sigma_0^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

则

$$U = \frac{(\bar{X} - \bar{Y}) - \mu_0}{\sigma_0} \sim N(0, 1).$$

(2) 由定理 4.1 的结论 (2) 知

$$\frac{n_1 - 1}{\sigma_1^2} S_1^2 \sim \chi^2(n_1 - 1), \quad \frac{n_2 - 1}{\sigma_2^2} S_2^2 \sim \chi^2(n_2 - 1).$$

再因两个总体 X 与 Y 相互独立, 从而它们的样本方差也相互独立, 故由命题 4.3 知

$$F = \frac{\frac{n_1 - 1}{\sigma_1^2} S_1^2 / (n_1 - 1)}{\frac{n_2 - 1}{\sigma_2^2} S_2^2 / (n_2 - 1)} \sim F(n_1 - 1, n_2 - 1).$$

(3) 记 $\frac{1}{n} = \frac{1}{n_1} + \frac{1}{n_2}$, 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时, 因 $\mu_0 = \mu / \bar{n}$, 由 (1) 中已证事实, 即知

$$U_1 = \frac{(\bar{X} - \bar{Y}) - \mu}{\frac{\sigma}{\bar{n}}} \sim N(0, 1).$$

再利用 (2) 中已证事实, 并结合命题 4.2 知

$$V = \frac{1}{2} \{ (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 \} \sim \sigma^2 (n_1 + n_2 - 2).$$

此外, 对任意 x, y , 与任意 $s_1 > 0, s_2 > 0$, 有

$$\begin{aligned} P\{\bar{X} = x, \bar{Y} = y, S_1^2 = s_1, S_2^2 = s_2\} \\ = P\{\bar{X} = x, S_1^2 = s_1\}P\{\bar{Y} = y, S_2^2 = s_2\} \quad (\text{因总体 } X \text{ 与 } Y \text{ 相互独立}) \\ = P\{\bar{X} = x\}P\{S_1^2 = s_1\}P\{\bar{Y} = y\}P\{S_2^2 = s_2\} \quad (\text{由定理 4.1}) \end{aligned}$$

这表明, \bar{X}, \bar{Y}, S_1^2 与 S_2^2 相互独立. 从而作为 \bar{X} 和 \bar{Y} 的函数 U_1 与作为 S_1^2 和 S_2^2 的函数 V 也相互独立.

综上所述, 由命题 4.4 即知

$$\frac{U_1}{\frac{V}{n_1 + n_2 - 2}} \sim t(n_1 + n_2 - 2).$$

再因

$$\frac{V}{n_1 + n_2 - 2} = \frac{1}{2}S^2$$

即知

$$T = \frac{(\bar{X} - \bar{Y}) - \mu}{S / \bar{n}} = \frac{U_1}{\frac{V}{n_1 + n_2 - 2}} \sim t(n_1 + n_2 - 2).$$

在定理 4.3 中, 仅第三个结论必须要求两个正态总体的方差相等. 自然, 当两个正态总体的方差相等, 即有 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时, 结论 (1) 将变得更简单, 可由下述结论替换之:

$$(a) \quad U_1 = \frac{(\bar{X} - \bar{Y}) - \mu}{\frac{\sigma}{\bar{n}}} \sim N(0, 1),$$

其中 $\mu = \mu_1 - \mu_2, \frac{1}{n} = \frac{1}{n_1} + \frac{1}{n_2}$.

结论 (2) 则可由下述结论替换之:

$$(b) \quad F_1 = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1).$$

如单正态总体情形中所述一样, 作为概率论推导的结果, 只要 μ_1 与 μ_2 分别表示总体 X 与 Y 的数学期望, σ_1^2 与 σ_2^2 分别表示总体 X 与 Y 的方差, 则不论它

们是否已知, 定理 4.3 的结论总是成立的. 又在统计应用中, 为比较两个总体的参数, 实际上仅对 $\mu = \mu_1 - \mu_2$ 与 $r = \sigma_1^2 / \sigma_2^2$ 发生兴趣, 其中 μ 表示两个总体数学期望之间的差异, r 则表示两个总体方差的比值. 这样, 当两正态总体的方差相等时, 即使 μ 未知, T 也可作为枢轴量; 而当 r 未知时, F 也可作枢轴量. 从而, 相应地可用于推断未知的期望差异值 μ 或方差的比值 r . 此外, 当两个正态总体的方差相等但未知时, 如期望差异值 μ 也未知, 样本函数 U_1 甚至连枢轴量也不是, 这也就是为什么要设计枢轴量 T 来替换样本函数 U_1 的原由.

二、一般总体抽样分布的极限分布

本分节将取消定理 4.2 中总体服从正态分布的条件, 并推导样本函数 U 与 T 的极限分布. 为此, 需引入随机变量依分布收敛的概念. 设 $F_n(x)$ 为随机变量 X_n 的分布函数, $F(x)$ 为随机变量 X 的分布函数, 并记 $C(F)$ 为由 $F(x)$ 的全体连续点组成的集合, 若

$$\lim_n F_n(x) = F(x), \quad x \in C(F),$$

则称随机变量 X_n 依分布收敛于 X , 简记为

$$X_n \xrightarrow{d} X \quad \text{或} \quad F_n(x) \xrightarrow{d} F(x).$$

我们有下述命题

* 命题 4.5 设随机变量 X 有连续的分布函数, 且有

$$X_n \xrightarrow{d} X, \quad Y_n \xrightarrow{P} 1,$$

则 $X_n Y_n \xrightarrow{d} X$.

证明 略.

上述命题的证明要用到上、下极限的概念. 为不使正文中的数学推导过长, 我们把上述命题的证明留作习题 (见习题四 (B) 的第 8 题). 有兴趣的读者可自证之, 或可参阅本书习题答案中关于此题的证明概要.

有了上述命题就不难证明下述定理.

定理 4.4 设 (X_1, X_2, \dots, X_n) 为总体 X 的样本, 并设总体 X 的数学期望与方差均存在, 分别记为 $EX = \mu$, $DX = \sigma^2$. 再记

$$U_n = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}, \quad T_n = \frac{\bar{X} - \mu}{S / \sqrt{n}},$$

其中 \bar{X} 与 S 分别表示上述样本的样本均值与样本方差, 则有

$$(1) \quad F_{U_n}(x) \xrightarrow{d} \Phi(x),$$

$$* (2) \quad F_{T_n}(x) \xrightarrow{d} \Phi(x),$$

以上 $F_{U_n}(x)$, $F_{T_n}(x)$ 与 $\phi(x)$ 分别表示 U_n 、 T_n 与标准正态分布的分布函数.

证明 (1) 由于

$$U_n = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{(X_1 + X_2 + \dots + X_n) - n\mu}{\frac{\sigma}{\sqrt{n}}}$$

在总体 X 的方差存在的前提下, 由 § 3.5 节中所述的 (林德伯格-勒维) 中心极限定理, 立知

$$\lim_{n \rightarrow \infty} F_{U_n}(x) = \lim_{n \rightarrow \infty} P(U_n \leq x) = \phi(x), \quad x \in \mathbb{R}^1.$$

(2) 由习题四(A)的第 6 题知

(1/S) $\xrightarrow{P} 1$. 再因

$$T_n = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = U_n \cdot \frac{1}{S}$$

故由本定理的结论(1)与命题 4.5, 便知

$$F_{T_n}(x) \xrightarrow{d} \phi(x).$$

定理 4.4 的适用范围很宽泛, 惟一的条件是总体的方差存在. 这样, 当样本容量 n 充分大时, U_n 与 T_n 都近似地服从标准正态分布. 因此, 如总体的方差 σ^2 已知, 便可利用枢轴量 U_n 近似地对总体未知的数学期望 μ 进行统计推断 (见 § 5.4 节与 § 6.4 节); 如 σ^2 未知, 则可选用枢轴量 T_n 近似地对 μ 作统计推断 (见 § 6.4 节).

最后, 我们指出, 前面介绍的关于正态总体抽样分布的精确分布与一般总体的若干抽样分布的极限分布, 都是用作关于单参数的统计推断的. 在第六章中, 我们还要讨论关于有限离散型总体的多参数假设检验, 为此需要引入一个以 χ^2 分布为极限分布的皮尔逊统计量, 这将在 § 6.4 节中予以详细介绍.

习 题 四

(A)

1. 设总体 X 服从以 λ ($\lambda > 0$) 为参数的指数分布, (X_1, X_2, \dots, X_n) 为其一个样本, 求该样本的样本密度.

2. 设总体 X 服从闭区间 $[0, 1]$ 上的均匀分布, (X_1, X_2, \dots, X_n) 为其一个样本, 求该样本的样本密度.

3. 设总体 X 服从以 p ($0 < p < 1$) 为参数的几何分布, (X_1, X_2, \dots, X_n) 为其一个样本, 求该样本的离散样本密度.

4. 设 (x_1, x_2, \dots, x_n) 与 (u_1, u_2, \dots, u_n) 为两组样本的样本值, 它们有下列关系:

$$u_i = \frac{x_i - a}{b} \quad (b \neq 0, a \text{ 为常数}).$$

求样本均值 \bar{u} , \bar{x} 及样本方差 s_u^2 与 s_x^2 之间的关系.

5. 设 (X_1, X_2, \dots, X_n) 为总体 X 的一个样本, A_k 为此样本的 k 阶原点矩. 若总体 X 的 k 阶原点矩 μ_k 存在, 利用大数定律证明

$$A_k \xrightarrow{P} \mu_k \quad (n \rightarrow \infty).$$

6. 设 (X_1, X_2, \dots, X_n) 为总体 X 的一个样本, S^2 为该样本的修正样本方差. 另设总体 X 的方差 $D[X] = \sigma^2$ 存在.

(1) 利用大数定律证明 $S^2 \xrightarrow{P} \sigma^2 \quad (n \rightarrow \infty)$;

(2) 利用结论 (1) 证明 $S^2/S^2 \xrightarrow{P} 1 \quad (n \rightarrow \infty)$.

7. 设总体 X 的分布函数为 $F(x)$, (X_1, \dots, X_n) 为其一个样本, 试以 $F(x)$ 表示该样本的极小值 $X_{(1)}$ 与极大值 $X_{(n)}$ 的分布函数. 特别地, 当总体 X 服从以 λ ($\lambda > 0$) 为参数的指数分布时, 试求 $X_{(1)}$ 与 $X_{(n)}$ 的分布函数.

8. 利用伽马函数与贝塔函数的定义与性质, 计算下列积分:

(1) $\int_0^{\infty} x^2 e^{-x} dx$;

(2) $\int_0^{\infty} x^6 e^{-x^2} dx$;

(3) $\int_0^1 x^2 \frac{y^4}{(1+y^2)^3} dy$.

9. 设 $X \sim N(0, 1)$. 利用伽马函数的定义与性质, 证明 $E[X^{2n}] = 1 \cdot 3 \cdot 5 \cdots (2n-1)$.

10. 设总体 n 服从自由度为 m 的 χ^2 分布, (X_1, X_2, \dots, X_n) 为其一个样本, 试求该样本的样本均值 \bar{X} 的密度函数.

11. 设总体 X 服从标准正态分布, 从此总体中取出一个容量为 6 的样本 (X_1, X_2, \dots, X_6) , 令

$$Y = (X_1 + X_2 + X_3)^2 + (X_4 + X_5 + X_6)^2.$$

试决定常数 c , 使得随机变量 cY 服从 χ^2 分布, 并求该 χ^2 分布的自由度.

12. 设总体 X 服从以 λ ($\lambda > 0$) 为参数的泊松分布, (X_1, X_2, \dots, X_n) 为其一个样本, 试求样本和 $S_n = X_1 + X_2 + \dots + X_n$ 的确切分布与渐近分布.

13. 查表求标准正态分布的下列上侧分位数:

$$u_{0.6}, u_{0.8}, u_{0.9} \text{ 与 } u_{0.95}.$$

14. 设总体 $X \sim N(\mu, \sigma^2)$, 假如我们要以 99.7% 的概率保证偏差 $|\bar{X} - \mu| \leq 0.1$, 试问在 $\sigma^2 = 0.5$ 时, 样本容量应取多大?

15. 查表求 χ^2 分布的下列上侧分位数:

$$\chi_{0.95}^2(5), \chi_{0.05}^2(5), \chi_{0.99}^2(10) \text{ 与 } \chi_{0.01}^2(10).$$

16. 证明 F 分布上侧分位数的关系式 (4.21) 式, 并查表求 F 分布的下列上侧分位数:

$$F_{0.95}(4, 6), F_{0.975}(3, 7) \text{ 与 } F_{0.99}(5, 5).$$

17. 查表求 t 分布的下列上侧分位数:

$$t_{0.05}(3), t_{0.01}(5), t_{0.10}(7) \text{ 与 } t_{0.005}(10).$$

习 题 四

(B)

1. 利用切比雪夫不等式求一枚均匀钱币需抛多少次才能使样本均值 \bar{X} 落在 0.4 到 0.6 之间的概率至少为 0.9? 这里 \bar{X} 是样本 (X_1, X_2, \dots, X_n) 的均值, 而 X_i 表示第 i 次抛均匀钱币时正面出现的次数.

2. 设总体 X 的方差为 $\sigma^2 = 4$, 而 \bar{X} 是容量为 100 的样本均值. 利用切比雪夫不等式求出一个下限和一个上限, 使得 $\bar{X} - \mu$ (μ 为总体 X 的数学期望) 落在这两个界限之间的概率至少为 0.90.

3. 设总体 X 的二阶原点矩存在, (X_1, X_2, \dots, X_n) 为其一个样本, \bar{X} 为该样本的样本均值, 试证 $X_i - \bar{X}$ 与 $X_j - \bar{X}$ ($i \neq j$) 的相关系数为 $-\frac{1}{n-1}$.

* 4. 设 X_1, X_2, \dots, X_n 为相互独立同分布的随机变量, 而且 $X_i \sim N(0, 1)$, $i = 1, 2, \dots, n$. 再设 A 为 n 阶正交矩阵, 即 $AA^T = E$, 其中 A^T 表示矩阵 A 之转置, E 为单位矩阵. 记 $X = (X_1, X_2, \dots, X_n)^T$, $Y = AX = (Y_1, Y_2, \dots, Y_n)^T$. 证明 Y_1, Y_2, \dots, Y_n 仍为相互独立同分布的随机变量, 且 $Y_i \sim N(0, 1)$, $i = 1, 2, \dots, n$.

* 5. 设 $X_1 + X_2 + \dots + X_n$ 为相互独立同分布的随机变量, 且 $X_i \sim N(0, 1)$, $i = 1, 2, \dots, n$. 记

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n), S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

证明 $\bar{X} \sim N(0, \frac{1}{n})$, $(n-1)S^2 \sim \chi^2(n-1)$, 且 \bar{X} 与 S^2 相互独立.

* 6. 证明定理 4.1.

7. 设随机变量 X 具有对称的密度函数 $f(x)$, $F(x)$ 为其分布函数. 再设 $g(x)$ 与 $G(x)$ 分别为 X 的密度函数与分布函数.

(1) 证明 $F(x) + F(-x) = 1$; $G(x) = 2F(x) - 1, x \geq 0$.

(2) 证明 $f(x) = \frac{1}{2}g(|x|)$.

* 8. 设随机变量 Y_n 依概率收敛至常数 1; 又设随机变量 X_n 依分布收敛至随机变量 X , 且随机变量 X 具有连续的分布函数. 证明

$$X_n Y_n \xrightarrow{d} X(n).$$

第 5 章

参 数 估 计

统计推断是统计学的重要内容，它大致可分为两类：估计问题与假设检验问题，其中估计问题又分为参数估计和非参数估计。本章将介绍参数估计的基本知识，并着重讨论求点估计的经典方法以及正态总体参数的区间估计。

§ 5.1 点估计概述

我们常常会面临这样一类问题：已知总体的分布类型，但不知道其中某些参数的真值。例如已知总体服从泊松分布，但不知其参数 λ 到底等于多少。这时我们希望通过所拥有的样本来对未知参数作出估计，这就是参数估计问题。其实，上述泊松分布总体 ($\lambda > 0$) 代表着一族总体，估计 λ 无非是要推断样本究竟来自这族总体中的哪一个。另外，总体分布中未知参数的实值函数通常也叫参数。因此，利用样本估计未知参数的实值函数也属于参数估计问题。参数估计有点估计与区间估计之分。我们首先讨论参数的点估计。

一、什么叫点估计

设 (X_1, \dots, X_n) 为来自总体 X 的样本， (x_1, \dots, x_n) 为相应的样本值。 θ 是总体分布中的未知参数， Θ 表示 θ 的取值范围，称为参数空间。尽管 θ 是未知的，但它的参数空间 Θ 是事先知道的。为了估计未知参数 θ ，我们构造一个统计量 $h(X_1, \dots, X_n)$ ，然后用 $h(X_1, \dots, X_n)$ 的值 $h(x_1, \dots, x_n)$ 来估计 θ 的真值。称 $h(X_1, \dots, X_n)$ 为 θ 的估计量，记作 $\hat{\theta}(X_1, \dots, X_n)$ ；称 $h(x_1, \dots, x_n)$ 为 θ 的估计值，记作 $\hat{\theta}(x_1, \dots, x_n)$ 。在不会引起误会的场合，估计量与估计值统称为点估计，简称为估计，并简记为 $\hat{\theta}$ 。事实上， $\hat{\theta}$ 的估计值是数轴上的一个点，用 $\hat{\theta}$ 的估计值作为 θ 的真值的近似值就相当于用一个点来估计 θ ，因此得名为点估计。

如果总体分布中含有多个未知参数 $\theta_1, \dots, \theta_r$ ， $(\theta_1, \dots, \theta_r) \in \Theta$ ，那么称统计量 $\hat{\theta}_i(X_1, \dots, X_n)$ 为 θ_i 的估计量，称相应的值为 $\hat{\theta}_i$ 的估计值， $i = 1, 2, \dots, r$ 。

至于待估参数为未知参数 θ 的实值函数 $g(\theta)$ 时，则称用来估计 $g(\theta)$ 的统计量 $\hat{g}(X_1, \dots, X_n)$ 为 $g(\theta)$ 的估计量，称相应的值为 \hat{g} 的估计值。

例 5.1 设某种型号的电子元件的寿命 X (以小时计) $\sim f(x; \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$, ($x > 0$). θ 为未知参数, $\theta > 0$. 现得样本值为

168, 130, 169, 143, 174, 198, 108, 212, 252,

试估计未知参数 θ .

由题意知总体 X 的均值为 θ , 即 $\theta = EX$, 因此用样本均值 \bar{X} 作为 θ 的估计量也许看起来是最自然的. 对给定的样本值计算得

$$\bar{x} = \frac{1}{9}(168 + 130 + \dots + 252) = 172.7,$$

故 $\theta = \bar{X}$ 与 $\theta = \bar{x} = 172.7$ 分别为 θ 的估计量与估计值.

我们注意到上例中除 $\theta = \bar{X}$ 以外 $\theta_1 = X_1$ 也可作为 θ 的估计量, 相应的估计值是 $\theta_1 = 168$; 若记 $X_{(1)} = \min\{X_1, \dots, X_n\}$, $X_{(n)} = \max\{X_1, \dots, X_n\}$, 则 $\theta_2 = \frac{1}{2}(X_{(1)} + X_{(n)})$ 同样有资格作 θ 的估计量, 相应的估计值 $\theta_2 = \frac{1}{2}(108 + 252) = 180$. 由此可见, 点估计的概念相当宽松. 对一个未知参数, 原则上可以随意构造其估计. 因此有必要建立一些评价估计量好坏的标准. 下面简单介绍三条最基本的标准: 无偏性, 有效性及相合性.

二、评价估计量的标准

1. 无偏性

我们希望估计量 $\hat{\theta}$ 的取值不要偏高也不要偏低, 即 $\hat{\theta}$ 的平均取值与 θ 的真值基本一致. 于是导出了无偏性标准.

定义 5.1 设 $\hat{\theta} = (X_1, \dots, X_n)$ 为参数 θ 的估计量, 若 $E\hat{\theta} = \theta$, 则称 $\hat{\theta}$ 是无偏估计量, 否则称 $\hat{\theta}$ 为 θ 的有偏估计量.

若 $\lim_n E\hat{\theta} = \theta$, 则称 $\hat{\theta}$ 是 θ 的渐近无偏估计量.

例 5.2 设 (X_1, \dots, X_n) 为取自总体 X 的样本, 总体 X 的均值为 μ 方差为 σ^2 . 则

(1) 样本均值 \bar{X} 是 μ 的无偏估计量;

(2) 样本方差 S^2 是 σ^2 的无偏估计量;

(3) 未修正的样本方差, 即样本二阶中心矩 $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 是 σ^2 的有偏估计量.

解 (1) 因为 $EX_i = EX = \mu$ $i = 1, 2, \dots, n$.

$E\bar{X} = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n EX_i = EX = \mu$ 故 $\mu = \bar{X}$ 是 μ 的一个无偏估计量.

(2) $DX_i = DX = \sigma^2$, $i = 1, 2, \dots, n$. $D\bar{X} = \frac{1}{n} DX = \frac{\sigma^2}{n}$,

于是

$$\begin{aligned}
 ES^2 &= E\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right] \\
 &= E\left\{\frac{1}{n-1} \left[\sum_{i=1}^n X_i^2 - n(\bar{X})^2\right]\right\} \\
 &= \frac{1}{n-1} \left[\sum_{i=1}^n EX_i^2 - nE(\bar{X})^2\right] \\
 &= \frac{1}{n-1} \left\{ \sum_{i=1}^n (\mu^2 + \sigma^2) - n[D\bar{X} + (E\bar{X})^2] \right\} \\
 &= \frac{1}{n-1} (n\sigma^2 - \sigma^2) \\
 &= \sigma^2
 \end{aligned}$$

故 S^2 是 σ^2 的一个无偏估计量.

$$\begin{aligned}
 (3) \quad E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right] &= E\left(\frac{n-1}{n} S^2\right) \\
 &= \frac{n-1}{n} ES^2 \\
 &= \frac{n-1}{n} \sigma^2
 \end{aligned}$$

故样本二阶中心矩是 σ^2 的有偏估计量. 但

$$\lim_{n \rightarrow \infty} E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right] = \sigma^2,$$

因此它是 σ^2 的一个渐近无偏估计量.

注意, 如果 \bar{X} 是 μ 的无偏估计量, $g(\cdot)$ 是 μ 的函数, 未必能推出 $g(\bar{X})$ 是 $g(\mu)$ 的无偏估计量. 例如总体 $X \sim N(\mu, \sigma^2)$, \bar{X} 是 μ 的无偏估计量, 但 $(\bar{X})^2$ 却不是 μ^2 的无偏估计量. 因为

$$E(\bar{X})^2 = D\bar{X} + (E\bar{X})^2 = \frac{\sigma^2}{n} + \mu^2,$$

而 $\frac{\sigma^2}{n} > 0$, 所以 $E(\bar{X})^2 > \mu^2$.

2. 有效性

有时一个参数存在许多无偏估计量, 选用哪个好呢? 显然应该看它们中间哪一个取值更集中, 即方差更小. 也就是说, 一个好的估计量应具有尽量小的方差. 由此引出了第二个标准——有效性.

定义 5.2 设 $\hat{\theta}_1$ 与 $\hat{\theta}_2$ 为参数 θ 的两个无偏估计量, 若 $D\hat{\theta}_1 < D\hat{\theta}_2$, 则称 $\hat{\theta}_1$ 较 $\hat{\theta}_2$ 有效.

例 5.3 设总体 X 的方差存在且大于零, $EX = \mu$ (X_1, X_2) 为 X 的一个

样本. 则 $\mu = \bar{X}$ 与 $\mu = X_1$ 都是 μ 的无偏估计量, 但 $D\mu = \frac{1}{2}DX < D\mu = DX$, 故 μ 比 μ 有效.

例 5.4 设总体 $X \sim N(1, \sigma^2)$, 其中参数 σ^2 未知, $\sigma^2 > 0$. (X_1, \dots, X_n) 为来自总体 X 的样本 ($n > 1$). 考虑 σ^2 的两个估计量:

$$\hat{\sigma}_1^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad \hat{\sigma}_2^2 = \frac{1}{n} \sum_{i=1}^n (X_i - 1)^2.$$

因为

$$E\hat{\sigma}_1^2 = ES^2 = \sigma^2,$$

$$\begin{aligned} E\hat{\sigma}_2^2 &= E\left[\frac{1}{n} \sum_{i=1}^n (X_i - 1)^2\right] = \frac{1}{n} \sum_{i=1}^n E(X_i - EX_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^n DX_i = \sigma^2 \end{aligned}$$

所以它们都是 σ^2 的无偏估计量. 下面来比较它们的方差. 由于

$$\frac{(n-1)S^2}{2} \sim \sigma^2(n-1),$$

$$\frac{1}{2} \sum_{i=1}^n (X_i - \mu)^2 \sim \sigma^2(n),$$

$$D\left[\frac{(n-1)S^2}{2}\right] = 2(n-1),$$

$$D\left[\frac{1}{2} \sum_{i=1}^n (X_i - \mu)^2\right] = 2n,$$

因此

$$D\hat{\sigma}_1^2 = DS^2 = \left(\frac{2}{n-1}\right)^2 D\left[\frac{(n-1)S^2}{2}\right] = \frac{2^4}{n-1},$$

$$\begin{aligned} D\hat{\sigma}_2^2 &= D\left[\frac{1}{n} \sum_{i=1}^n (X_i - 1)^2\right] = \left(\frac{2}{n}\right)^2 D\left[\frac{1}{2} \sum_{i=1}^n (X_i - 1)^2\right] \\ &= \frac{2^4}{n} \end{aligned}$$

后者小于前者, 故 $\hat{\sigma}_2^2$ 较 $\hat{\sigma}_1^2$ 有效.

3. 相合性

所谓相合性就是当样本容量无限增大时估计量 $\hat{\theta}$ 与 θ 的真值任意接近的概率趋于 1. 相合性也称一致性, 它反映了估计量的一种大样本性质.

定义 5.3 设 $\hat{\theta} = (X_1, \dots, X_n)$ 为未知参数 θ 的估计量, 若 $\hat{\theta}$ 依概率收敛于 θ , 即对任意 $\epsilon > 0$, 有

$$\lim_n P\{|\hat{\theta} - \theta| < \epsilon\} = 1$$

或

$$\lim_n P\{|\hat{\theta} - \theta| \geq \epsilon\} = 0,$$

则称 $\hat{\theta}$ 为 θ 的(弱)相合估计量.

例 5.5 设 (X_1, \dots, X_n) 是取自总体 X 的样本, 且 DX^k 存在, $k=1, 2, \dots, n$. 则 $\frac{1}{n} \sum_{i=1}^n X_i^k$ 为 EX^k 的相合估计量, $k=1, 2, \dots, n$.

事实上, 对指定的 k , 令 $Y = X^k$, $Y_i = X_i^k$, $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{i=1}^n X_i^k$
由大数定律知

$$P\text{-}\lim_{n \rightarrow \infty} \bar{Y} = EY = EX^k$$

从而 $\frac{1}{n} \sum_{i=1}^n X_i^k$ 是 EX^k 的相合估计.

作为特例, 样本均值 \bar{X} 是总体均值 EX 的相合估计量.

例 5.6 设总体 $X \sim N(\mu, \sigma^2)$, (X_1, \dots, X_n) 为其样本. 试证样本方差 S^2 是 σ^2 的相合估计量.

证 因 $ES^2 = \sigma^2$, $DS^2 = \frac{2\sigma^4}{n-1}$ (可参见例 5.4), 故由切比雪夫不等式推得, 对任意 $\varepsilon > 0$,

$$0 < P\{|\bar{S}^2 - ES^2| \geq \varepsilon\} = P\{|\bar{S}^2 - \sigma^2| \geq \varepsilon\} \\ \leq \frac{1}{\varepsilon^2} DS^2 = \frac{2\sigma^4}{\varepsilon^2(n-1)}$$

当 $n \rightarrow \infty$ 时上式左、右端均趋于 0, 根据相合性定义可知 S^2 是 σ^2 的相合估计量.

§ 5.2 极大似然法

无偏性、有效性以及相合性从各个不同角度描述了估计量的合理性. 那么如何求得比较理想的估计量呢? 本节将介绍一种既经典又很流行的求点估计的方法, 即极大似然法.

一、极大似然法的基本思想

极大似然法的思想很简单: 在已经得到试验结果的情况下, 我们应该寻找使这个结果出现的可能性最大的那个 θ 作为 θ 的估计. 下面我们仅就离散型总体和连续型总体这两种场合做进一步分析.

设 (X_1, \dots, X_n) 为来自总体 X 的样本, X 的分布类型已知, 但参数 θ 未知,

若 X 为离散型随机变量, 其概率分布的形式为 $P\{X=x\} = p(x; \theta)$, 则样本 (X_1, \dots, X_n) 的概率分布为 $P\{X_1=x_1, \dots, X_n=x_n\} = \prod_{i=1}^n p(x_i; \theta)$, 在 θ 固定时, 上式表示 (X_1, \dots, X_n) 取值 (x_1, \dots, x_n) 的概率; 反之, 当样本值 (x_1, \dots, x_n) 给定

时, 它可看作 θ 的函数, 我们把它记作 $L(\theta)$, 并称

$$L(\theta) = \prod_{i=1}^n p(x_i; \theta), \quad (4.1)$$

为似然函数. 似然函数 $L(\theta)$ 的值的大小意味着该样本值出现的可能性的. 既然已经得到了样本值 (x_1, \dots, x_n) , 那它出现的可能性应该是大的, 即似然函数的值应该是大的. 因而我们选择使 $L(\theta)$ 达到最大值的那个 θ 作为真 θ 的估计.

若 X 为连续型随机变量, 其密度函数为 $f(x; \theta)$, 则样本 (X_1, \dots, X_n) 的密度函数为 $\prod_{i=1}^n f(x_i; \theta)$. 在 θ 固定时, 它是 (X_1, X_2, \dots, X_n) 在 (x_1, x_2, \dots, x_n) 处的密度, 它的大小与 (X_1, \dots, X_n) 落在 (x_1, \dots, x_n) 附近的概率的大小成正比. 而当样本值 (x_1, \dots, x_n) 给定时, 它是 θ 的函数. 我们仍把它记为 $L(\theta)$ 并称

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta), \quad (4.2)$$

为似然函数. 类似于刚才的讨论, 我们选择使 $L(\theta)$ 最大的那个 θ 作为真 θ 的估计.

总之, 在有了试验结果即样本值 (x_1, \dots, x_n) 时, 似然函数 $L(\theta)$ 反映了 θ 的各个不同值导出这个结果的可能性的. 我们选择使 $L(\theta)$ 达到最大值的那个 θ 作为真 θ 的估计. 这种求点估计的方法就叫做极大似然法.

定义 5.4 若对任意给定的样本值 (x_1, \dots, x_n) 存在 $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$, 使

$$L(\hat{\theta}) = \max_{\theta} L(\theta), \quad (4.3)$$

则称 $\hat{\theta}(x_1, \dots, x_n)$ 为 θ 的极大似然估计值, 称相应的统计量 $\hat{\theta}(x_1, \dots, x_n)$ 为的极大似然估计量. 它们统称为 θ 的极大似然估计, 可简记为 MLE.

如果未知参数为 $\theta_1, \theta_2, \dots, \theta_r$, 那么似然函数是多元函数 $L(\theta_1, \dots, \theta_r)$. 若对任意给定的样本值 (x_1, \dots, x_n) 存在 $\hat{\theta}_i = \hat{\theta}_i(x_1, \dots, x_n)$, $i = 1, 2, \dots, r$, 使

$$L(\hat{\theta}_1, \dots, \hat{\theta}_r) = \max_{(\theta_1, \dots, \theta_r)} L(\theta_1, \dots, \theta_r), \quad (4.4)$$

则称 $\hat{\theta}_i$ 为 θ_i 的 MLE, $i = 1, 2, \dots, r$.

二、极大似然估计的一般求法

当似然函数关于未知参数可微时, 一般可通过求导数得到 MLE, 其主要步骤是:

写出似然函数 $L(\theta_1, \dots, \theta_r)$.

令 $\frac{\partial L}{\partial \theta_i} = 0$ 或 $\frac{\partial \ln L}{\partial \theta_i} = 0$, $i = 1, \dots, r$, 从中解得驻点. 注意, 函数 L 与

$\ln L$ 有相同的最值点, 而使用后者往往更方便.

判断驻点为最大值点.

求得各参数的 MLE.

例 5.7 设总体 $X \sim N(\mu, \sigma^2)$, μ 与 σ^2 均未知, $-\infty < \mu < +\infty$, $\sigma^2 > 0$. (X_1, \dots, X_n) 为来自 X 的样本, (x_1, \dots, x_n) 为样本值. 试求 μ 与 σ^2 的极大似然估计.

解 X 的密度为 $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$, 似然函数

$$L(\mu, \sigma^2) = \prod_{i=1}^n f(x_i; \mu, \sigma^2) \\ = (2\pi\sigma^2)^{-\frac{n}{2}} \exp -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2,$$

$$\text{于是 } \ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

$$\begin{aligned} \text{令 } \frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} &= 0 & \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu) &= 0 \\ \text{即 } \frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} &= 0 & -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 &= 0, \end{aligned}$$

$$\text{从中解得 } \mu = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}, \quad \sigma_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

分别计算

$$\begin{aligned} A &= \frac{\partial^2 \ln L(\mu, \sigma^2)}{\partial \mu^2} \bigg|_{(\mu_0, \sigma_0^2)} = -\frac{n}{\sigma_0^2} \\ B &= \frac{\partial^2 \ln L(\mu, \sigma^2)}{\partial \mu \partial \sigma^2} \bigg|_{(\mu_0, \sigma_0^2)} = 0 \\ C &= \frac{\partial^2 \ln L(\mu, \sigma^2)}{(\partial \sigma^2)^2} \bigg|_{(\mu_0, \sigma_0^2)} = -\frac{n}{2\sigma_0^4} \end{aligned}$$

$$\text{由于 } A < 0 \text{ 且 } AC - B^2 = \left(-\frac{n}{\sigma_0^2}\right) \left(-\frac{n}{2\sigma_0^4}\right) > 0$$

因此 μ_0, σ_0^2 是极大值点同时也是最大值点. 从而 $\mu = \bar{x}$ 与 $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

分别为 μ 与 σ^2 的极大似然估计量, $\mu = \bar{x}$ 与 $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ 分别为 μ 与 σ^2 的极大似然估计值.

正态总体均值与方差的 MLE 都是相合估计, 并且均值的 MLE 是无偏估计, 方差的 MLE 是渐近无偏估计.

这里有一点需要说明: 按照本课程的要求, 当似然函数的驻点惟一时, 不

必验证该驻点是否为最大值点. 可直接把驻点作为所求参数的极大似然估计.

例 5.8 设总体 X 服从泊松分布, 参数 λ 未知, $\lambda > 0$. (X_1, \dots, X_n) 为来自 X 的样本, (x_1, \dots, x_n) 为样本值. 试求 λ 的极大似然估计.

解 X 的概率分布为

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

似然函数为 $L(\lambda) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}, \quad x_i=0, 1, \dots,$

两边取对数得

$$\ln L(\lambda) = -n\lambda + \left(\sum_{i=1}^n x_i \right) \ln \lambda - \ln \left(\prod_{i=1}^n x_i! \right),$$

令 $\frac{d}{d\lambda} \ln L(\lambda) = 0$, 即 $-\lambda + \frac{1}{\lambda} \sum_{i=1}^n x_i = 0$,

从中解得 $\lambda_0 = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$.

故 $\lambda_0 = \bar{x}$ 是 λ 的极大似然估计量, $\lambda = \bar{x}$ 是 λ 的极大似然估计值.

显然泊松总体参数 λ 的极大似然估计是 λ 的无偏估计与相合估计.

极大似然估计有一个简单但很有用的性质: 如果 λ_0 是 λ 的极大似然估计, $u = g(\lambda)$ 是 λ 的函数且存在单值反函数 $\lambda = h(u)$. 那么 $g(\lambda_0)$ 是 $g(\lambda)$ 的极大似然估计. 这种性质叫做极大似然估计的不变性. 该性质还可以推广到多个参数的场合.

例 5.9 试求 5.1 例 5.1 中元件的平均寿命以及概率 $P\{X > 180\}$ 的极大似然估计值.

解 先求平均寿命 EX 即 λ 的极大似然估计量. 由于似然函数为

$$L(\lambda) = \prod_{i=1}^n \frac{1}{x_i!} e^{-\lambda} \lambda^{x_i}, \quad (x_i > 0, i=1, \dots, n)$$

$$\ln L(\lambda) = -n \ln \lambda - \frac{1}{\lambda} \sum_{i=1}^n x_i$$

令 $\frac{d}{d\lambda} \ln L = 0$ 即 $-\frac{n}{\lambda} + \frac{1}{\lambda^2} \sum_{i=1}^n x_i = 0$,

解得 $\lambda_0 = \bar{x}$, 故 $\lambda_0 = \bar{x}$ 是 λ 的极大似然估计量. 从而平均寿命 λ 的极大似然估计值为 $\bar{x} = 172.7$.

$P\{X > 180\} = e^{-\frac{180}{\lambda}}$ 是 λ 的函数. 由不变性可得 $P(X > 180)$ 的极大似然估计量为 $e^{-\frac{180}{\bar{x}}}$, 从而 $P(X > 180)$ 的极大似然估计值为 $e^{-\frac{180}{172.7}} = 0.353$.

下面举一个不能通过求导来获得极大似然估计的例子.

例 5.10 设总体 X 服从 $[0, \theta]$ 上的均匀分布, θ 未知, $\theta > 0$. $(X_1, \dots,$

X_n) 为 X 的样本, (x_1, \dots, x_n) 为样本值. 试求 θ 的极大似然估计.

$$\text{解} \quad \text{似然函数 } L(\theta) = \begin{cases} \frac{1}{n!} \theta^n, & 0 < x_1, \dots, x_n < \theta \\ 0, & \text{其他} \end{cases}$$

显然无法从 $L(\theta) = 0$ 得到 MLE. 我们考虑直接按极大似然法的基本思想来确定 θ 的 MLE.

欲使 $L(\theta)$ 最大, θ 应尽量小但又不能太小, 它必须同时满足 $x_i, i = 1, \dots, n$, 即

$$\theta \geq \max\{x_1, \dots, x_n\},$$

否则 $L(\theta) = 0$, 而 0 不可能是 $L(\theta)$ 的最大值. 因此, 当 $\theta = \max\{x_1, \dots, x_n\}$ 时, $L(\theta)$ 可达最大. $\hat{\theta} = \max\{x_1, \dots, x_n\}$ 即为 θ 的极大似然估计量, $\hat{\theta} = \max\{X_1, \dots, X_n\}$ 为 θ 的极大似然估计量.

例中的 $\hat{\theta} = \max\{X_1, \dots, X_n\}$ 并非无偏估计量. 为了证明这个结论, 需要求出它的分布. 记 $X_{(n)} = \max\{X_1, \dots, X_n\}$, 并设 $X_{(n)}$ 的分布函数与密度函数分别为 $G(x)$ 与 $g(x)$, 则

$$\begin{aligned} G(x) &= P\{X_{(n)} \leq x\} = P\{\max\{X_1, \dots, X_n\} \leq x\} \\ &= P\{X_1 \leq x, \dots, X_n \leq x\} \\ &= \prod_{i=1}^n P\{X_i \leq x\} \\ &= [F(x)]^n \end{aligned}$$

其中 $F(x)$ 为总体 X 的分布函数.

$$\text{又} \quad F(x) = \begin{cases} 0, & x < 0 \\ \frac{x}{\theta}, & 0 \leq x < \theta \\ 1, & x \geq \theta \end{cases}$$

$$\text{于是} \quad G(x) = [F(x)]^n = \begin{cases} 0, & x < 0 \\ \left(\frac{x}{\theta}\right)^n, & 0 \leq x < \theta \\ 1, & x \geq \theta \end{cases}$$

$$\text{从而} \quad g(x) = \begin{cases} \frac{nx^{n-1}}{\theta^n}, & 0 < x < \theta \\ 0, & \text{其他} \end{cases}$$

$$E X_{(n)} = \int_0^\theta x g(x) dx = \int_0^\theta \frac{nx^n}{\theta^n} dx = \frac{n}{n+1} \theta,$$

故 $\hat{\theta} = X_{(n)}$ 不是 θ 的无偏估计量而是 θ 的渐近无偏估计量. 略加修改后的

$\frac{n+1}{n}X_{(n)}$ 是 的无偏估计量.

综上所述, 求极大似然估计的要点在于掌握极大似然法的基本思想. 极大似然法充分利用了已知总体分布类型所蕴涵的信息, 因此极大似然估计一般具有较多的优良性质. 当然它的计算也往往比较麻烦, 在不少场合只能借助于计算机求得近似解.

§ 5.3 矩法

除极大似然法外, 矩法也是求点估计的常用方法. 它的基本思想是: 用相应的样本矩去估计总体矩; 用相应的样本矩的函数去估计总体矩的函数. 例如

$$\hat{E}X = \bar{X}, \quad \hat{D}X = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

一般地, 若记

$$\mu_k = EX^k, \quad \sigma_k = E(X - EX)^k, \quad A_k = \frac{1}{n} \sum_{i=1}^n X_i^k, \quad B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k,$$

则总体的 k 阶原点矩用相应的样本 k 阶原点矩来估计; 而总体的 k 阶中心矩用相应的样本 k 阶中心矩来估计, 即

$$\mu_k = A_k, \quad k = 1, 2, \dots,$$

$$\sigma_k = B_k, \quad k = 2, 3, \dots$$

这种求点估计的方法叫做矩法. 用矩法确定的估计量称为矩估计量, 相应的估计值称为矩估计值, 矩估计量与矩估计值统称为矩估计, 可简记为 ME.

矩法是一种古老的方法, 它的特点是并不要求已知总体分布的类型. 只要未知参数可以表示成总体矩的函数, 就能求出其矩估计. 当总体分布类型已知时, 由于矩法没有充分利用总体分布所提供的信息, 矩估计不一定是理想的估计. 但因矩法简便易行, 而且矩估计具有一定的优良性, 所以应用仍然十分广泛.

按照矩法的基本思想求矩估计的一般步骤为:

从总体矩入手将待估参数 表为总体矩的函数, 即

$$= g(\mu_1, \dots, \mu_1; \sigma_2, \dots, \sigma_s).$$

用 A_k, B_k 分别替换 g 中的 μ_k, σ_k .

$$= g(\mu_1, \dots, \mu_1; \sigma_2, \dots, \sigma_s) = g(A_1, \dots, A_1; B_2, \dots, B_s)$$

即为 的 ME.

例 5.11 设总体 $X \sim N(\mu, \sigma^2)$, (X_1, \dots, X_n) 为取自总体 X 的样本. 试求 μ, σ^2 的矩估计量.

解 $\mu = EX, \sigma^2 = DX$ 故 $\mu = \hat{E}X = \bar{X}$ 与 $\sigma^2 = \hat{D}X = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 分别为

μ 与 σ^2 的矩估计量.

由此可见, 正态总体 $N(\mu, \sigma^2)$ 中 μ 与 σ^2 的极大似然估计和矩估计是完全一样的.

例 5.12 试求例 5.10 中 θ 的矩估计.

解 由 $EX = \frac{1}{2}$, 得 $\theta = 2EX$, 从而 $\hat{\theta} = 2\hat{EX} = 2\bar{X}$ 即为 θ 的矩估计量, 而 $\hat{\theta} = 2\bar{X}$ 为 θ 的矩估计值. 此时 θ 的极大似然估计与矩估计是不同的.

例 5.13 设总体 X 服从参数为 m, p 的二项分布, m 已知, p 未知. (X_1, \dots, X_n) 为其样本, (x_1, \dots, x_n) 为样本值. 试求

- (1) p 的极大似然估计量及矩估计量;
- (2) p 与 q 之比的矩估计量, 其中 $q = 1 - p$.

解 (1) X 的分布列为 $P(X = x) = C_m^x p^x (1-p)^{m-x}$, $x = 0, 1, \dots, m$. 似然函数

$$L(p) = \prod_{i=1}^n C_m^{x_i} p^{x_i} (1-p)^{m-x_i}, \quad x_i = 0, 1, \dots, m,$$

$$\ln L(p) = \left(\sum_{i=1}^n x_i \right) \ln p + \left[\sum_{i=1}^n (m - x_i) \right] \ln(1-p) + \ln \left(\prod_{i=1}^n C_m^{x_i} \right),$$

$$\text{令 } \frac{d \ln L}{dp} = 0 \quad \text{即} \quad \frac{\sum_{i=1}^n x_i}{p} - \frac{\sum_{i=1}^n (m - x_i)}{1-p} = 0,$$

$$\text{解得 } p_0 = \frac{\sum_{i=1}^n x_i}{mn} = \frac{\bar{X}}{m}, \text{ 故 } p \text{ 的极大似然估计量为 } p = \frac{\bar{X}}{m}.$$

因 $EX = mp$, $p = \frac{EX}{m}$, 又 $\hat{EX} = \bar{X}$, 故 $p = \frac{\bar{X}}{m}$ 为 p 的矩估计量.

(2) 令 $g(p) = \frac{p}{q} = \frac{p}{1-p}$, $p = \frac{\bar{X}}{m}$, 故

$$\hat{g}(p) = \frac{p}{1-p} = \frac{\frac{\bar{X}}{m}}{1 - \frac{\bar{X}}{m}} = \frac{\bar{X}}{m - \bar{X}}$$

即为 $g(p)$ 的矩估计量.

但要注意, 因 $DX = mpq$, $g(p) = \frac{p}{q} = \frac{mp^2}{DX}$, 故

$$\hat{g}(p) = \frac{m \left(\frac{\bar{X}}{m} \right)^2}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} = \frac{(\bar{X})^2}{\frac{m}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$$

也是 $g(p)$ 的矩估计量.

例 5.14 设总体 X 的密度函数为

$$f(x; \theta, \mu) = \begin{cases} \frac{1}{\theta} e^{-\frac{x-\mu}{\theta}}, & x \geq \mu \\ 0, & x < \mu \end{cases}$$

其中参数 θ, μ 均未知, $\theta > 0$. (X_1, \dots, X_n) 为取自 X 的样本. 试求 θ, μ 的矩估计量.

解 令 $\frac{x-\mu}{\theta} = t$, 则

$$\begin{aligned} EX &= \int_{-\infty}^{+\infty} xf(x; \theta, \mu) dx = \int_{\mu}^{+\infty} \frac{x}{\theta} e^{-\frac{x-\mu}{\theta}} dx \\ &= \int_0^{+\infty} (t + \mu) e^{-t} dt = \theta(2) + \mu(1) = \theta + \mu \end{aligned}$$

$$\begin{aligned} DX &= \int_{-\infty}^{+\infty} (x - EX)^2 f(x; \theta, \mu) dx = \int_{\mu}^{+\infty} \frac{(x - \mu)^2}{\theta} e^{-\frac{x-\mu}{\theta}} dx \\ &= \theta^2(3) - 2\theta^2(2) + \theta^2(1) = 2\theta^2. \end{aligned}$$

于是从方程组 $\begin{cases} EX = \theta + \mu \\ DX = 2\theta^2 \end{cases}$, 得 $\begin{cases} \theta = \sqrt{DX} \\ \mu = EX - \theta \end{cases}$

又 $\hat{EX} = \bar{X}$, $\hat{DX} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$,

从而 θ 与 μ 的矩估计量分别为

$$\begin{aligned} \hat{\theta} &= \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}, \\ \hat{\mu} &= \bar{X} - \hat{\theta} = \bar{X} - \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}. \end{aligned}$$

§ 5.4 置信区间

前面讨论了参数的点估计, 即用一个点去估计未知参数. 比如正态总体 $X \sim N(\mu, \sigma^2)$, 我们可用随机点 \bar{X} 作为 μ 的估计量, 用相应的点 \bar{x} 作为 μ 的近似值, 而且 $\hat{\mu} = \bar{X}$ 是 μ 的优良估计. 但是任何一种估计, 如果不注明估计的误差, 这种估计恐怕是没有多大意义的. 点估计不能直接提供其估计的误差, 因此需要引入另一类估计——区间估计. 在区间估计理论中, 被广泛接受的一种观点是置信区间. 它是由奈曼 (Neymann) 于 1934 年提出的.

一、置信区间的概念

粗略地说, 置信区间就是一个随机区间, 它能以足够大的概率套住我们感兴趣的参数. 例如 μ 是未知参数 θ 的一个估计量, $(\hat{\theta} - \delta, \hat{\theta} + \delta)$ 为一个随机区

间, 其中 $\alpha > 0$. 若能使该区间套住 θ 的概率等于事先指定的数 $1-\alpha$, 即 $P\{-\alpha < \bar{X} - \theta < \bar{X} + \alpha\} = 1-\alpha$, 区间 $(\bar{X} - \alpha, \bar{X} + \alpha)$ 便是 θ 的一个置信区间. 下面给出置信区间的一般定义.

定义 5.5 设 θ 为总体分布的未知参数, (X_1, \dots, X_n) 为来自总体 X 的样本. 对给定的数 $1-\alpha$ ($0 < \alpha < 1$), 若存在两个统计量 $\underline{L} = \underline{L}(X_1, \dots, X_n)$ 与 $\bar{L} = \bar{L}(X_1, \dots, X_n)$, 使得

$$P\{\underline{L} < \theta < \bar{L}\} = 1-\alpha, \quad (4.5)$$

则称随机区间 (\underline{L}, \bar{L}) 为 θ 的 $1-\alpha$ 置信区间, 称 $1-\alpha$ 为置信度, 又分别称 \underline{L} 与 \bar{L} 为 θ 的置信下限与置信上限.

一旦有了样本值 (x_1, \dots, x_n) , 区间 (\underline{L}, \bar{L}) 的端点也随之确定. 称区间 $(\underline{L}(x_1, \dots, x_n), \bar{L}(x_1, \dots, x_n))$ 为置信区间 (\underline{L}, \bar{L}) 的一个实现, 它是一个普通的区间, 也简称为置信区间.

定义 5.5 中置信度 $1-\alpha$ 的含义是: 随机区间 (\underline{L}, \bar{L}) 套住 θ 的概率为 $1-\alpha$, 而不套住 θ 的概率为 α . 用频率来解释就是: 如果重复试验 100 次获得了样本 (X_1, \dots, X_n) 的 100 个样本值, 相应地得到了 (\underline{L}, \bar{L}) 的 100 个实现, 那么在这 100 个普通区间中, 大约有 $100(1-\alpha)$ 个区间套住了 θ , 而不套住 θ 的区间大约有 100 个. 若令 $1-\alpha = 0.95$, 则其中大约有 95 个区间套住 θ , 大约有 5 个区间不套住 θ .

事实上, 置信区间 (\underline{L}, \bar{L}) 也是对未知参数 θ 的一种估计. 区间的长度意味着误差, 因此可以说区间估计与点估计是互补的两种参数估计.

二、寻求置信区间的方法

通常采用基于点估计构造置信区间的方法来获得置信区间. 下面分小样本与大样本两种情况讨论.

1. 小样本情形

先来看一个简单的例子.

例 5.15 设总体 $X \sim N(\mu, \sigma^2)$, σ^2 已知, μ 未知, (X_1, \dots, X_n) 为来自 X 的样本. 试求 μ 的 $1-\alpha$ 置信区间.

样本均值 \bar{X} 是 μ 的极大似然估计, 从 \bar{X} 出发考虑一个样本与未知参数 μ 的函数 $U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$, 它不包含其他未知参数, 并且它的分布已知, $U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim$

$N(0, 1)$. 对事先指定的 $1-\alpha$, 确定 u_1, u_2 , 使 $P\{u_1 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < u_2\} = 1-\alpha$. 显然满足条件的 u_1, u_2 不止一对, 我们选取 $-u_{\frac{\alpha}{2}}$ 为 u_1 , $u_{\frac{\alpha}{2}}$ 为 u_2 , 其中 $u_{\frac{\alpha}{2}}$ 是水平的双侧分位数, 由

$$P\left\{-u_{\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < u_{\frac{\alpha}{2}}\right\} = 1 - \alpha.$$

经不等式变形得

$$P\left\{\bar{X} - u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha.$$

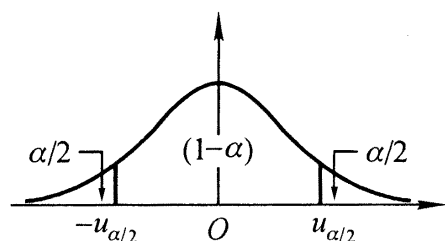


图 5.1 $N(0, 1)$ 分布的分位数

于是 $\bar{X} - u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, $\bar{X} + u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ 为 μ 的 $1 - \alpha$ 置信区间. 若令 $1 - \alpha = 95\%$, 则 μ 的 95% 置信区间为 $\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}$, $\bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$.

用该区间估计 μ 我们不仅直接可知估计成功的概率是 95%, 而且能够以 95% 的把握断言以 \bar{X} 代替 μ 的绝对误差小于 $1.96 \cdot \frac{\sigma}{\sqrt{n}}$.

由此看来, 基于点估计构造置信区间的关键在于寻找一个合适的函数 u , 在它的基础上诱导出所要的区间. 现在把求未知参数 μ 的置信区间的一般步骤归纳如下:

选取 μ 的一个较优的点估计 \bar{X} ,

围绕 \bar{X} 寻找一个依赖于样本与 μ 的函数 $u = u(X_1, \dots, X_n; \mu)$. u 的分布为已知分布. 像 u 这样的函数, 在前一章中被称为枢轴量.

对给定的置信度 $1 - \alpha$, 确定 u_1 与 u_2 , 使

$$P\{u_1 < u < u_2\} = 1 - \alpha.$$

一般可选取满足 $P\{u < u_1\} = P\{u > u_2\} = \frac{\alpha}{2}$ 的 u_1 与 u_2 .

利用不等式变形导出套住 μ 的置信区间 $(\underline{\mu}, \bar{\mu})$.

例 5.16 设总体 X 的密度为

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

未知参数 $\lambda > 0$, (X_1, \dots, X_n) 为取自 X 的样本.

(1) 试证 $\frac{2n\bar{X}}{\lambda} \sim \chi^2(2n)$;

(2) 试求 λ 的 $1 - \alpha$ 置信区间.

解 (1) 记 $Y = \frac{2n\bar{X}}{\lambda}$, 设 Y 的分布函数与密度函数分别为 $G(y)$ 与 $g(y)$, 则

$$\begin{aligned} G(y) &= P\{Y \leq y\} \\ &= P\left\{\frac{2n\bar{X}}{\lambda} \leq y\right\} \end{aligned}$$

$$= P\{X \leq \frac{\bar{y}}{2}\}$$

$$= F(\frac{\bar{y}}{2})$$

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

于是 $G(y) = \begin{cases} 1 - e^{-\frac{y}{2}}, & y > 0 \\ 0, & y \leq 0 \end{cases}$

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

即 $Y \sim \chi^2(2)$, 从而 $\frac{2}{n}X_i \sim \chi^2(2)$, $i = 1, \dots, n$.

又由 χ^2 分布的可加性得

$$\sum_{i=1}^n \frac{2}{n}X_i \sim \chi^2(2n).$$

而 $\sum_{i=1}^n \frac{2}{n}X_i = \frac{2}{n} \sum_{i=1}^n X_i = 2\bar{X}.$

故 $2\bar{X} \sim \chi^2(2n).$

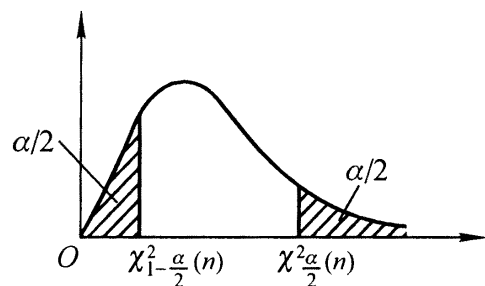


图 5.2 χ^2 分布的分位数

(2) \bar{X} 是 μ 的极大似然估计, 从 \bar{X} 出发考虑

枢轴量 $u = \frac{2n\bar{X}}{\sigma^2}$, 由(1)知 u 的分布只依赖于样本容量 n , 即 $u = \frac{2n\bar{X}}{\sigma^2} \sim \chi^2(2n)$

对给定的 $1-\alpha$, 由

$$P\left\{\chi^2_{1-\frac{\alpha}{2}}(2n) < \frac{2n\bar{X}}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}(2n)\right\} = 1-\alpha.$$

经不等式变形得

$$P\left\{\frac{2n\bar{X}}{\chi^2_{\frac{\alpha}{2}}(2n)} < \sigma^2 < \frac{2n\bar{X}}{\chi^2_{1-\frac{\alpha}{2}}(2n)}\right\} = 1-\alpha.$$

于是 $\left[\frac{2n\bar{X}}{\chi^2_{\frac{\alpha}{2}}(2n)}, \frac{2n\bar{X}}{\chi^2_{1-\frac{\alpha}{2}}(2n)}\right]$ 即为所求置信区间.

2. 大样本情形

如果枢轴量的分布不易确定, 有时可用极限分布来构造近似的置信区间, 当然此时要求样本容量足够大. 近似置信区间的求法与小样本情形类似, 不同的只是将枢轴量的精确分布改为极限分布. 以下仅举例说明.

例 5.17 设总体 X 服从参数为 p 的两点分布, p 未知, $0 < p < 1$. (X_1, \dots, X_n) 为其样本. 试求 p 的置信区间.

我们知道 p 的极大似然估计量为 \bar{X} . 基于 \bar{X} 考虑 $u = \frac{\bar{X} - EX}{\sqrt{DX/n}} = \frac{\bar{X} - p}{\sqrt{p(1-p)/n}}$. 根据定理 5.10 知当 n 足够大时 u 近似服从 $N(0, 1)$ 分布. 因此,

对给定的置信度 $1 - \alpha$, 由

$$P \left\{ \left| \frac{\bar{X} - p}{\sqrt{p(1-p)/n}} \right| < u_{\frac{1-\alpha}{2}} \right\} = 1 - \alpha$$

经不等式变形得

$$P \{ ap^2 + bp + c < 0 \} = 1 - \alpha,$$

其中 $a = n + (u_{\frac{1-\alpha}{2}})^2$, $b = -2n\bar{X} - (u_{\frac{1-\alpha}{2}})^2$, $c = n(\bar{X})^2$. 又由 $a > 0$ 知 $ap^2 + bp + c < 0$ 等价于 $p_1 < p < p_2$,

$$\begin{aligned} \text{其中 } p_1 &= \frac{1}{2a}(-b - \sqrt{b^2 - 4ac}) \\ p_2 &= \frac{1}{2a}(-b + \sqrt{b^2 - 4ac}). \end{aligned}$$

总之, 对给定的 $1 - \alpha$, 存在 p_1 与 p_2 使

$$P(p_1 < p < p_2) = 1 - \alpha,$$

于是 (p_1, p_2) 是 p 的一个置信度近似为 $1 - \alpha$ 的置信区间.

然而在实际问题中往往采用下面简化了的置信区间:

$$(\bar{X} - u_{\frac{1-\alpha}{2}} \sqrt{\bar{X}(1-\bar{X})/n}, \bar{X} + u_{\frac{1-\alpha}{2}} \sqrt{\bar{X}(1-\bar{X})/n}).$$

这是因为当 n 足够大时不仅 $\frac{\bar{X} - p}{\sqrt{p(1-p)/n}}$ 近似服从 $N(0, 1)$ 分布, 而且可以证

明 \bar{X} 代替分母中的 p 后, $\frac{\bar{X} - p}{\sqrt{\bar{X}(1-\bar{X})/n}}$ 仍然近似服从 $N(0, 1)$ 分布. 所以当 n

足够大时, 由 $P \left\{ \left| \frac{\bar{X} - p}{\sqrt{\bar{X}(1-\bar{X})/n}} \right| < u_{\frac{1-\alpha}{2}} \right\} = 1 - \alpha$, 得

$$P \left\{ \bar{X} - u_{\frac{1-\alpha}{2}} \sqrt{\bar{X}(1-\bar{X})/n} < p < \bar{X} + u_{\frac{1-\alpha}{2}} \sqrt{\bar{X}(1-\bar{X})/n} \right\} = 1 - \alpha$$

从而得到上述近似置信区间.

例 5.18 为了研究在一指定时间段内某地区的国际互联网用户所占的比例, 随机地调查了该地区的 400 名居民, 发现其中有 108 名居民为上网者. 试求该地区居民的上网率 p 的 95% 置信区间.

由题意知总体服从 0—1 分布, 参数 p 即上网率, 上网率是一种比率. 现在要求 p 的置信区间.

p 的置信度近似为 $1 - \alpha$ 的置信区间是

$$((\bar{X} - u_{\frac{\alpha}{2}} \sqrt{\bar{X}(1-\bar{X})/n}, \bar{X} + u_{\frac{\alpha}{2}} \sqrt{\bar{X}(1-\bar{X})/n})).$$

现已知 $n=400$, $\bar{X}=\frac{108}{400}=0.27$, $\alpha=0.05$, $u_{\frac{0.05}{2}}=1.96$, 故所求置信区间为 $(0.27 - 1.96 \times \sqrt{0.27 \times 0.73/400}, 0.27 + 1.96 \times \sqrt{0.27 \times 0.73/400})$ 即 $(0.23, 0.31)$.

三、单侧置信区间

上面所论的置信区间 $(-, +)$ 称为双侧置信区间. 在有些实际问题中要求形如 $(-, +)$ 或 $(-, -)$ 的置信区间, 称之为单侧置信区间. 例如, 在讨论产品的废品率时, 总希望它越小越好, 因此我们感兴趣的通常只是它的置信上限. 至于产品的平均寿命, 有时关心的是它的置信下限. 在这些场合寻找参数的单侧置信区间更有必要.

定义 5.6 设 μ 为总体 X 的未知参数, (X_1, \dots, X_n) 是来自 X 的样本. 对给定的数 $1-\alpha$, 若统计量 $\bar{X} = \bar{X}(X_1, \dots, X_n)$ 满足 $P(\bar{X} < \mu) = \alpha$, 则称 $(-, \bar{X})$ 为 μ 的置信度是 $1-\alpha$ 的单侧置信区间, 称 \bar{X} 为 μ 的单侧置信下限; 若统计量 $\bar{X} = \bar{X}(X_1, \dots, X_n)$ 满足 $P(\bar{X} > \mu) = \alpha$, 也称 $(\bar{X}, -)$ 为 μ 的置信度是 $1-\alpha$ 的单侧置信区间, 但称 \bar{X} 为 μ 的单侧置信上限.

例 5.19 (续例 5.15) 试求 μ 的单侧置信下(上)限, 其中置信度为 $1-\alpha$.

仍然考虑枢轴量 $U = \frac{\bar{X} - \mu}{\sqrt{\bar{X}(1-\bar{X})/n}}$. 对给定的 $1-\alpha$. 由 $P\left(\frac{\bar{X} - \mu}{\sqrt{\bar{X}(1-\bar{X})/n}} < u_{\alpha}\right) = 1-\alpha$ 经不等式变形得

$$P\left(\mu > \bar{X} - u_{\alpha} \sqrt{\bar{X}(1-\bar{X})/n}\right) = 1-\alpha$$

故 μ 的置信度是 $1-\alpha$ 的单侧置信下限为 $\bar{X} - u_{\alpha} \sqrt{\bar{X}(1-\bar{X})/n}$. 又由

$$P\left(\frac{\bar{X} - \mu}{\sqrt{\bar{X}(1-\bar{X})/n}} > -u_{\alpha}\right) = 1-\alpha$$

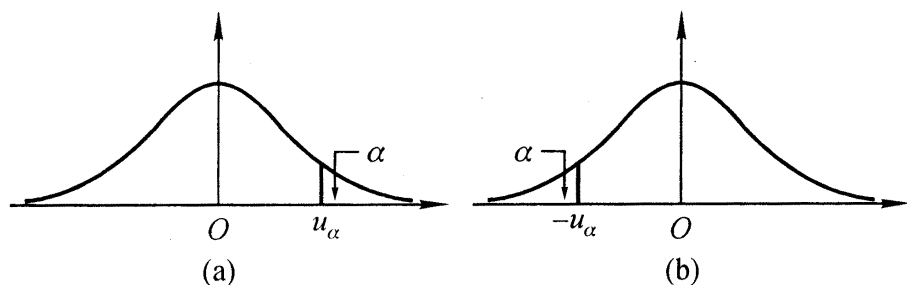


图 5.3 $N(0, 1)$ 分布的分位数

得
$$P\left\{\mu < \bar{X} + u_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

从而 μ 的置信度是 $1 - \alpha$ 的单侧置信上限为 $\bar{X} + u_{1-\alpha} \frac{\sigma}{\sqrt{n}}$.

例 5.20 为考虑某种香烟的尼古丁含量 (以 mg 计), 抽取了 8 支香烟并测得尼古丁的平均含量为 $\bar{x} = 0.26$. 侧设该香烟尼古丁含量 $X \sim N(\mu, 2.3)$. 试求 μ 的单侧置信上限, 置信度为 0.95.

解 μ 的单侧置信上限为 $\bar{X} + u_{1-\alpha} \frac{\sigma}{\sqrt{n}}$. $\alpha = 0.05$, $u_{0.95} = 1.65$, $\bar{x} = 0.26$, $\sigma^2 = 2.3$, $n = 8$, 故所求单侧置信上限为 $0.26 + 1.65 \times \frac{\sqrt{2.3}}{\sqrt{8}} = 1.14$.

§ 5.5 正态总体参数的置信区间

与其他总体相比, 正态总体参数的置信区间是最完善的, 应用也最广泛. 在构造正态总体参数的置信区间的过程中, t 分布、 χ^2 分布、 F 分布以及 $N(0, 1)$ 分布扮演了重要角色.

一、单正态总体参数的置信区间

设总体 $X \sim N(\mu, \sigma^2)$, $-\infty < \mu < +\infty$, $\sigma^2 > 0$. (X_1, \dots, X_n) 为来自 X 的样本.

1. 均值 μ 的置信区间

(1) 方差 σ^2 已知的情形

我们在上一节例 5.15 已讨论了在 σ^2 已知的条件下 μ 的 $1 - \alpha$ 置信区间为 $\bar{X} - u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, $\bar{X} + u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, 可简记为 $\bar{X} \pm u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$.

事实上, 在给定的置信度为 $1 - \alpha$ 时, 对任意的 $\alpha_1 > 0$, $\alpha_2 > 0$, $\alpha_1 + \alpha_2 = \alpha$, 凡是满足

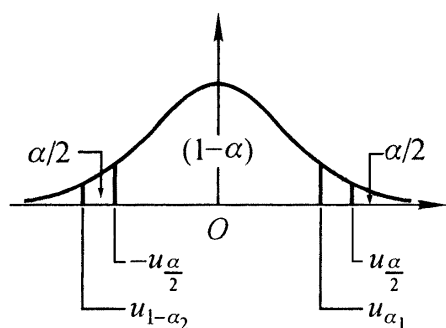


图 5.4 比较 $N(0, 1)$ 分布的分位数

$$P\left\{-u_{1-\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < u_{1-\frac{\alpha}{2}}\right\} = 1 - \alpha.$$

的区间 $\bar{X} - u_{\alpha_1} \cdot \frac{\sigma}{\sqrt{n}}$, $\bar{X} + u_{1-\alpha_2} \cdot \frac{\sigma}{\sqrt{n}}$ 都是 μ 的置信区间, 只不过在所有这类区间中就数 $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ 时的区间长度最短 (见图 5.4). 这个区间正是前面所推导出的 $\bar{X} \pm u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$.

例 5.21 设总体 $X \sim N(\mu, \sigma^2)$, 其中 μ 未知, $\sigma^2 = 4.50$. 现取得样本容量为 36 的一样本, 测得样本均值为 $\bar{x} = 54.40$. 试求 μ 的 95% 置信区间.

解 由题意知 $n = 36$, $\sigma^2 = 4.50$, $\alpha = 0.05$, $u_{\frac{0.05}{2}} = 1.96$, $u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{\sqrt{4.50}}{\sqrt{36}} = 1.47$.

于是 μ 的 95% 置信区间为

$$(\bar{x} - 1.47, \bar{x} + 1.47)$$

即 (52.93, 55.87)

这里要注意, (52.93, 55.87) 是一个普通区间, 因此不能说该区间套住 μ 的概率为 95%, 应理解成 “ μ 属于该区间” 这件事情的可信程度为 95%, 或该区间属于那些套住 μ 的区间类的概率为 95%.

例 5.22 设总体 $X \sim N(\mu, \sigma^2)$, 其中 μ 未知, $\sigma^2 = 4$. (X_1, \dots, X_n) 为其样本.

- (1) 当 $n = 16$ 时, 试求置信度分别为 0.9 及 0.95 的 μ 的置信区间的长度.
- (2) n 多大方能使 μ 的 90% 置信区间的长度不超过 1?
- (3) n 多大方能使 μ 的 95% 置信区间的长度不超过 1?

解 (1) 记 μ 的置信区间长度为 L , 则

$$L = \bar{X} + u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} - \bar{X} - u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 2u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

于是当 $1 - \alpha = 90\%$ 时, $L = 2 \times 1.65 \times \frac{2}{\sqrt{16}} = 1.65$, 当 $1 - \alpha = 95\%$ 时 $L = 2 \times 1.96$

$$\times \frac{2}{\sqrt{16}} = 1.96.$$

(2) 欲使 $L \leq 1$ 即 $2u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq 1$, 必须 $n \geq (2u_{\frac{\alpha}{2}})^2$, 于是当 $1 - \alpha = 90\%$ 时, $n \geq (2 \times 1.65)^2$ 即 $n \geq 44$. 也就是说, 样本容量 n 至少为 44 时, μ 的 90% 置信区间的长度才不超过 1.

(3) 当 $1 - \alpha = 95\%$ 时, 类似可得 $n \geq 62$.

在用置信区间对未知参数进行区间估计时, 我们总是希望估计的可靠度要高, 即置信区间包含未知参数的概率 $1 - \alpha$ 越大越好; 另一方面希望估计的精确度要高, 即置信区间的长度 L 越小越好. 但上面的例子给我们的启示是: 当样本容量 n 固定时可靠度与精确度是互相制约的. 如果提高可靠度即增大 $1 - \alpha$ 的值, 势必拉长区间, 从而精确度就减小; 反之, 缩短区间提高了精确度, 则可靠度降低. 总之, “鱼和熊掌不可兼得”, 若要可靠度与精确度两者都高除非增加样本容量.

(2) 方差 σ^2 未知的情形

用 σ^2 的无偏估计 S^2 代替 σ^2 , 选取枢轴量为 $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$, 由定理 4.2 知 $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$. 对给定的置信度 $1-\alpha$, 由

$$P\left\{-t_{\frac{\alpha}{2}}(n-1) < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\frac{\alpha}{2}}(n-1)\right\} = 1-\alpha,$$

经不等式变形得

$$P\left[\bar{X} - t_{\frac{\alpha}{2}}(n-1) \cdot \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2}}(n-1) \cdot \frac{S}{\sqrt{n}}\right] = 1-\alpha,$$

因此, μ 的 $1-\alpha$ 置信区间为

$$\bar{X} - t_{\frac{\alpha}{2}}(n-1) \cdot \frac{S}{\sqrt{n}}, \quad \bar{X} + t_{\frac{\alpha}{2}}(n-1) \cdot \frac{S}{\sqrt{n}}.$$

可简记为 $\bar{X} \pm t_{\frac{\alpha}{2}}(n-1) \cdot \frac{S}{\sqrt{n}}$.

2. 方差 σ^2 的置信区间

在实际问题中 μ 与 σ^2 往往都是未知的, 下面主要讨论在 μ 未知的场合方差 σ^2 的置信区间.

从 σ^2 的无偏估计 S^2 出发考虑枢轴量为 $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$,

由定理 4.2 知 $\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$. 对给定的 $1-\alpha$, 由

$$P\left\{\chi^2_{1-\frac{\alpha}{2}}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}(n-1)\right\} = 1-\alpha,$$

得
$$P\left\{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right\} = 1-\alpha,$$

因此 σ^2 的 $1-\alpha$ 置信区间为

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \quad \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}$$

也即

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \quad \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}$$

另外, 标准差 σ 的 $1-\alpha$ 置信区间为

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}}, \quad \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}}.$$

如果 μ 已知, 只须考虑枢轴量为 $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$. 按 χ^2 分布的定义知

$$\frac{1}{2} \sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n \frac{(X_i - \mu)^2}{2} \sim \chi^2(n).$$

类似可导出 σ^2 的 $1-\alpha$ 置信区间为

$$\frac{\sum_{i=1}^n (X_i - \mu)^2}{\frac{1}{2} \chi^2_{1-\frac{\alpha}{2}}(n)}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\frac{1}{2} \chi^2_{\frac{\alpha}{2}}(n)}.$$

例 5.23 为考察某大学成年男性的胆固醇水平, 现抽取了样本容量为 25 的一样本, 并测得样本均值 $\bar{x} = 186$, 样本标准差 $s = 12$. 假定所论胆固醇水平 $X \sim N(\mu, \sigma^2)$, μ 与 σ^2 均未知. 试分别求出 μ 以及 σ^2 的 90% 置信区间.

解 μ 的 $1-\alpha$ 置信区间为 $\bar{X} \pm t_{\frac{\alpha}{2}}(n-1) \cdot \frac{s}{n} = 0.1$, $s = 12$, $n = 25$, 查表得 $t_{\frac{0.1}{2}}(25-1) = 1.7109$ 于是 $t_{\frac{\alpha}{2}}(n-1) \cdot \frac{s}{n} = 1.7109 \times \frac{12}{25} = 4.106$, 从而 μ 的 90% 置信区间为 (186 ± 4.106) 即 $(181.89, 190.11)$.

σ^2 的 $1-\alpha$ 置信区间为 $\frac{(n-1)S^2}{\frac{1}{2}\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\frac{1}{2}\chi^2_{\frac{\alpha}{2}}(n-1)}$. 查表得

$$\frac{0.1}{2}(25-1) = 36.42, \quad \frac{1}{2} \cdot \frac{0.1}{2}(25-1) = 13.85,$$

于是置信下限为 $\frac{24 \times 12^2}{36.42} = 9.74$, 置信上限为 $\frac{24 \times 12^2}{13.85} = 15.80$

所求 σ^2 的 90% 置信区间为 $(9.74, 15.80)$

二、双正态总体参数的置信区间

在两个正态总体的场合, 通常被关注的是它们的均值差与方差比这两个参数. 设总体 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, $-\infty < \mu_i < +\infty$, $\sigma_i^2 > 0$, $i = 1, 2$. (X_1, \dots, X_{n_1}) 与 (Y_1, \dots, Y_{n_2}) 是分别取自总体 X 与 Y 的两个相互独立的样本, 记

$$\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i, \quad S_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2,$$

$$\bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i, \quad S_2^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2.$$

1. 均值差 $\mu_1 - \mu_2$ 的置信区间

(1) 方差 σ_1^2, σ_2^2 均已知的情形

\bar{X} 与 \bar{Y} 分别是 μ_1 与 μ_2 的无偏估计, 且

$$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}).$$

于是考虑枢轴量为 $\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$, 它服从 $N(0, 1)$ 分布. 对给定的 $1 - \alpha$, 由

$$P \left| \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| < u_{\frac{\alpha}{2}} = 1 - \alpha \quad (4.9)$$

推出 $\mu_1 - \mu_2$ 的 $1 - \alpha$ 置信区间为

$$(\bar{X} - \bar{Y}) - u_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{X} - \bar{Y}) + u_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

(2) 方差 σ_1^2 与 σ_2^2 未知但 $\sigma_1^2 = \sigma_2^2$ 的情形

记 $\sigma_1^2 = \sigma_2^2 = \sigma^2$, $S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$, 则 S^2 是 σ^2 的无偏估计即 $ES^2 =$

σ^2 . 于是用 S^2 代替 σ^2 , 令 $T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, 根据定理 4.3 知

$T \sim t(n_1 + n_2 - 2)$. 对给定的 $1 - \alpha$, 由

$$P\{-t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) \leq T \leq t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)\} = 1 - \alpha,$$

可推得 $\mu_1 - \mu_2$ 的 $1 - \alpha$ 置信区间为

$$\begin{aligned} & (\bar{X} - \bar{Y}) - t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) \cdot S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \\ & (\bar{X} - \bar{Y}) + t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) \cdot S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \end{aligned} \quad (4.10)$$

例 5.24 A, B 两个地区种植同一型号的小麦. 现抽取了 19 块面积相同的麦田, 其中 9 块属于地区 A, 另外 10 块属于地区 B, 测得它们的小麦产量 (以 kg 计) 分别如下:

地区 A: 100, 105, 110, 125, 110, 98, 105, 116, 112;

地区 B: 101, 100, 105, 115, 111, 107, 106, 121, 102, 92.

设地区 A 的小麦产量 $X \sim N(\mu_1, \sigma_1^2)$, 地区 B 的小麦产量 $Y \sim N(\mu_2, \sigma_2^2)$, μ_1, μ_2, σ_1^2 均未知. 试求这两个地区小麦的平均产量之差 $\mu_1 - \mu_2$ 的 90% 置信区间.

解 由题意知所求置信区间的两个端点分别为

$(\bar{X} - \bar{Y}) \pm t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) \cdot S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. 由 $\alpha = 0.1$, $n_1 = 9$, $n_2 = 10$ 查表得

$t_{0.05}^{17} = 1.7396$, 按已给数据计算得 $\bar{x} = 109$, $\bar{y} = 106$, $S_1^2 = \frac{550}{8}$, $S_2^2 = \frac{606}{9}$,

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = 68, \quad S = 8.246, \quad \text{于是置信下限为}$$

$$(109 - 106) - 1.7396 \times 8.246 \times \sqrt{\frac{1}{9} + \frac{1}{10}} = -3.59,$$

置信上限为

$$(109 - 106) + 1.7396 \times 8.246 \times \sqrt{\frac{1}{9} + \frac{1}{10}} = 9.59$$

故均值差 $\mu_1 - \mu_2$ 的 90% 置信区间为 $(-3.59, 9.59)$

2. 方差比 $\frac{\sigma_1^2}{\sigma_2^2}$ 的置信区间

我们只讨论 μ_1, μ_2 均未知的情形. S_1^2 与 S_2^2 分别是 σ_1^2 与 σ_2^2 的无偏估计, 于是考虑枢轴量 $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$. 根据定理 4.3 知 $F \sim F(n_1 - 1, n_2 - 1)$. 对给定的 $1 - \alpha$, $P(F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) < F < F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)) = 1 - \alpha$ 等价于

$$P\left(\frac{1}{F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)} \cdot \frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)} \cdot \frac{S_1^2}{S_2^2}\right) = 1 - \alpha.$$

故推得方差比 σ_1^2/σ_2^2 的 $1 - \alpha$ 置信区间为

$$\frac{1}{F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)} \cdot \frac{S_1^2}{S_2^2}, \quad \frac{1}{F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)} \cdot \frac{S_1^2}{S_2^2}, \quad (4.11)$$

其中分位数 $F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) = \frac{1}{F_{\frac{\alpha}{2}}(n_2 - 1, n_1 - 1)}$.

例 5.25 某钢铁公司的管理人员为比较新旧两个电炉的温度状况, 他们抽取了新电炉的 31 个温度数据及旧电炉的 25 个温度数据, 并计算得样本方差分别为 $S_1^2 = 75$ 及 $S_2^2 = 100$. 设新电炉的温度 $X \sim N(\mu_1, \sigma_1^2)$, 旧电炉的温度 $Y \sim N(\mu_2, \sigma_2^2)$. 试求 σ_1^2/σ_2^2 的 95% 置信区间.

解 σ_1^2/σ_2^2 的 $1 - \alpha$ 置信区间的两个端点分别是 $(F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1))^{-1} \cdot \frac{S_1^2}{S_2^2}$

与 $F_{\frac{\alpha}{2}}(n_2 - 1, n_1 - 1) \cdot \frac{S_1^2}{S_2^2}$. $\alpha = 0.05$, $n_1 = 31$, $n_2 = 25$, 查表得

$$F_{\frac{0.05}{2}}(30, 24) = 2.21, \quad F_{\frac{0.05}{2}}(24, 30) = 2.14$$

于是置信下限为 $\frac{1}{2.21} \times \frac{75}{100} = 0.34$, 置信上限为 $2.14 \times \frac{75}{100} = 1.61$, 所求置信区间为 $(0.34, 1.61)$.

以上我们根据建立置信区间的一般方法获得了正态总体参数的双侧置信区间, 类似可得相应的单侧置信区间. 最后用下面的表 5.1 及表 5.2 作为本节的

小结.

表 5.1 单正态总体参数的置信区间

待估参数	条件	枢轴量	双侧置信区间	单侧置信下(上)限
均值 μ	σ^2 已知	$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim N(0,1)$	$\bar{X} \pm u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$	$\bar{X}-u_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}$ $\bar{X}+u_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}$
	σ^2 未知	$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$	$\bar{X} \pm t_{\frac{\alpha}{2}}(n-1) \cdot \frac{S}{\sqrt{n}}$	$\bar{X}-t_{\alpha}(n-1) \cdot \frac{S}{\sqrt{n}}$ $\bar{X}+t_{\alpha}(n-1) \cdot \frac{S}{\sqrt{n}}$
方差 σ^2	μ 已知	$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i-\mu)^2 \sim \chi^2(n)$	$\frac{\sum_{i=1}^n (X_i-\mu)^2}{\frac{1}{2}(n-1)} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i-\mu)^2}{\frac{1}{2}(n)}$	$\sum_{i=1}^n (X_i-\mu)^2 / \chi^2_{1-\alpha/2}(n)$ $\sum_{i=1}^n (X_i-\mu)^2 / \chi^2_{\alpha/2}(n)$
	μ 未知	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}$	$(n-1)S^2 / \chi^2_{1-\alpha}(n-1)$ $(n-1)S^2 / \chi^2_{\alpha}(n-1)$

表 5.2 双正态总体参数的置信区间

待估参数	条件	枢轴量	双侧置信区间	单侧置信下(上)限
均值差 $\mu_1-\mu_2$	σ_1^2, σ_2^2 均已知	$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1)$	$(\bar{X}-\bar{Y}) \pm u_{\frac{\alpha}{2}} \sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}$	$(\bar{X}-\bar{Y})-\frac{u_{\alpha} \sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}{}$ $(\bar{X}-\bar{Y})+\frac{u_{\alpha} \sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}{}$
均值差 $\mu_1-\mu_2$	σ_1^2, σ_2^2 均未知 但 $\sigma_1^2=\sigma_2^2$	$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S \sqrt{1/n_1+1/n_2}} \sim t(n_1+n_2-2)$ $S^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$	$(\bar{X}-\bar{Y}) \pm t_{\frac{\alpha}{2}}(n_1+n_2-2) \cdot S \sqrt{1/n_1+1/n_2}$	$(\bar{X}-\bar{Y})-t_{\alpha}(n_1+n_2-2) \cdot S \sqrt{1/n_1+1/n_2}$ $(\bar{X}-\bar{Y})+t_{\alpha}(n_1+n_2-2) \cdot S \sqrt{1/n_1+1/n_2}$

续表

待估参数	条件	枢轴量	双侧置信区间	单侧置信下(上)限
方差比 $\frac{\sigma_1^2}{\sigma_2^2}$	μ, μ 均未知	$\frac{S_1^2/\frac{1}{2}}{S_2^2/\frac{1}{2}} \sim F(n_1-1, n_2-1)$	$\frac{1}{F_{\frac{\alpha}{2}}(n_1-1, n_2-1)} \cdot \frac{S_1^2}{S_2^2},$ $F_{\frac{\alpha}{2}}(n_2-1, n_1-1) \cdot \frac{S_1^2}{S_2^2}$	$\frac{1}{F_{\alpha}(n_1-1, n_2-1)} \cdot \frac{S_1^2}{S_2^2},$ $F_{\alpha}(n_2-1, n_1-1) \cdot \frac{S_1^2}{S_2^2}$

习 题 五

(A)

1. 设总体 X 的均值 μ 已知, 方差 σ^2 未知, (X_1, \dots, X_n) 为来自总体 X 的样本. 试证

$$s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

是总体方差 σ^2 的无偏估计量.

2. 设总体 $X \sim N(\mu^2)$, (X_1, \dots, X_n) 是来自 X 的样本, 试选择适当的常数 C 使

$$\frac{1}{C} \sum_{i=1}^{n-1} (X_{i+1} - X_i)^2$$

成为 σ^2 的无偏估计量.

3. 设总体 X 服从参数为 p 的 0—1 分布, (X_1, \dots, X_n) 是取自 X 的样本. 试证

$p^2 = (\bar{X})^2 - \frac{1}{n-1} B_2$ 是 p^2 的无偏估计量, 其中 $B_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$.

4. 设 (X_1, X_2) 是来自正态总体 $N(\mu, 1)$ 的样本. 试证下列两个估计量

(1) $\mu = \frac{1}{3} X_1 + \frac{2}{3} X_2$;

(2) $\mu = \frac{3}{4} X_1 + \frac{1}{4} X_2$.

都是 μ 的无偏估计量, 并分析哪一个估计量较有效.

5. 设 (X_1, \dots, X_n) 为取自总体 X 的样本. 试求下列总体分布中未知参数的极大似然估计量:

(1) X 的密度为 $f(x; \theta) = \begin{cases} \theta x^{\theta-1}, & 0 < x < 1, \\ 0, & \text{其他} \end{cases}$, 其中 θ 未知, $\theta > 0$;

(2) X 的分布列为 $P\{X = k\} = p(1-p)^{k-1}, k = 1, 2, \dots$, 其中 p 未知, $0 < p < 1$;

(3) X 的分布列为 $P\{X = 1\} = p, P\{X = 0\} = 1-p$, 其中 p 未知, $0 < p < 1$;

(4) X 的密度为

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

其中 λ 已知, θ 未知, $\theta > 0$.

6. 某地区去年每月因交通事故死亡的人数如下: 3, 2, 0, 5, 4, 3, 1, 0, 7, 2, 0, 2. 假设每月交通事故造成死亡的人数服从参数为 λ 的泊松分布, λ 未知, $\lambda > 0$. 试求

(1) λ 的极大似然估计值与矩估计值;

(2) $P(X=0)$ 的极大似然估计值.

7. 设总体 X 服从 $[a, b]$ 上的均匀分布, a, b 均未知, $b > a$, (X_1, \dots, X_n) 为其样本. 试求 a, b 的矩估计量.

8. 设 (X_1, \dots, X_n) 为来自总体 X 的样本, $EX = \mu$, $DX = \sigma^2$, μ 与 σ^2 均未知, $\mu > 0$, $\sigma^2 > 0$, 试求变异系数 $C = \frac{\sigma}{\mu}$ 的矩估计量.

9. 设总体 X 的密度为

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

其中 θ 是未知参数, $-\infty < \theta < +\infty$. (X_1, \dots, X_n) 为取自 X 的样本. 试求 θ 的矩估计量和极大似然估计量.

10. 设总体 X 在 $[0, \theta]$ 上服从均匀分布, θ 未知, $\theta > 0$. (X_1, \dots, X_n) 为取自该总体的样本. 试从 $U = \frac{X_{(n)}}{n}$ 导出 θ 的 $1-\alpha$ 置信区间, 其中 $X_{(n)} = \max\{X_1, \dots, X_n\}$.

11. 欲估计一批产品的次品率 p , 从中抽取容量为 100 的样本, 测得样本次品率为 0.1. 试求 p 的置信度近似为 95% 的单侧置信上限.

12. 已知某种金属丝的抗断强度 $X(\text{kg}) \sim N(\mu, 20^2)$. 今从一批产品中抽取 16 根, 测得其平均抗断强度 $\bar{x} = 281$. 试求未知参数 μ 的 95% 置信区间.

13. 设总体 $X \sim N(\mu, 9)$, (X_1, \dots, X_n) 为其样本. 欲使 μ 的 $1-\alpha$ 置信区间的长度不超过 2, 问在以下两种情况样本容量 n 至少应取多少:

(1) $\alpha = 0.1$ 时;

(2) $\alpha = 0.01$ 时.

14. 某工厂为治理废水抽取了 10 个水样, 测得水中所含某种有毒物质的浓度的平均值及标准差分别为 $\bar{x} = 17.10$, $s = 2.90$. 根据以往资料已知废水中含该种有毒物质的浓度 $X \sim N(\mu, \sigma^2)$. 试求 μ 的 98% 置信区间.

15. 设按某种工艺生产的金属纤维的长度 $X(\text{毫米}) \sim N(\mu, \sigma^2)$. 现有抽取的 15 根纤维, 测得其平均长度为 $\bar{x} = 5.4$, 样本方差 $S^2 = 0.16$. 试求 μ 的单侧置信下限, 置信度为 95%.

16. 设某纺织厂日产细纱支数 $X \sim N(\mu, \sigma^2)$, 今从一天纺出的产品中抽取 15 缕, 测得细纱支数的标准差 $s = 2.1$. 试求细纱支数的均匀度 (方差) σ^2 的 95% 置信区间.

17. 为比较甲、乙两种品牌灯泡的寿命状况, 抽取了 10 只甲种灯泡和 8 只乙种灯泡. 测得平均寿命分别为 $\bar{x} = 1400$ (小时) 和 $\bar{y} = 1250$ (小时), 样本标准差分别为 $S_1 = 52$ (小时) 和 $S_2 = 64$ (小时). 设两种灯泡的寿命分别服从正态分布 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$, 其中 $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ 均未知. 求这两种灯泡的平均寿命之差 $\mu_1 - \mu_2$ 的 95% 置信区间.

18. 在上题中假定甲、乙两种灯泡的寿命分别服从正态分布 $N(\mu_1, \sigma_1^2)$ 与 $N(\mu_2, \sigma_2^2)$, $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ 均未知. 试求方差比 σ_1^2 / σ_2^2 的 95% 的置信区间.

习 题 五

(B)

1. 设 (X_1, \dots, X_n) 是来自总体 X 的样本, $EX = \mu$ $DX = \sigma^2$, $\mu > 0$, $\sigma^2 > 0$. 试证

$$\mu = \frac{2}{n(n+1)} \sum_{k=1}^n kX_k$$

是 μ 的无偏估计量与相合估计量.

2. 设总体 X 服从 $[-\theta, +\theta]$ 上的均匀分布, θ 是未知参数. (X_1, \dots, X_n) 为取自 X 的样本. 试求 θ 的极大似然估计量, 并说一说你对所得结果的看法.

3. 设 X 为指数分布总体, 其密度为

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

其中 θ 是未知参数, $\theta > 0$. (X_1, \dots, X_n) 为取自 X 的一个大样本, $n = 50$. 已知样本均值 $\bar{x} = 2410$. 试用极限分布构造 θ 的置信度近似为 99% 的置信区间.

第 6 章

假 设 检 验

前一章讨论了统计推断中的估计问题，本章将讨论另一类统计推断问题——假设检验。假设检验也可区分为参数假设检验与非参数假设检验两大类，参数假设检验又可区分为单参数假设检验与多参数假设检验。本章重点讨论单参数假设检验。单参数假设检验仅检验总体的某一未知参数，如总体的数学期望、总体方差，或当已知总体分布所属的类型时，检验总体分布中某一未知参数。处理单参数假设检验问题的关键是寻求一个仅含待检验参数的枢轴量，并使之服从或渐近地服从一已知的确定分布。第四章的§ 4.4 节已为此作了必要的准备。鉴于统计应用中最常见的总体为正态总体，我们首先在§ 6.2 节与§ 6.3 节分别讨论了单正态总体与双正态总体的参数假设检验。§ 6.4 节介绍了一般总体的单参数假设检验。这一节中重点讨论了伯努利总体的参数假设检验，作为这一统计模型的引申，在§ 6.5 节的第一分节中讨论了检验有限离散型总体的多项分布²检验法，这已属多参数假设检验。然后，作为多项分布²检验法的应用，又在此节中讨论了两种非参数检验法：拟合优度²检验与独立性检验。

作为本章的前导，我们在§ 6.1 节以单参数假设检验问题为背景，介绍了假设检验的基本统计思想，有关的重要概念与检验的一般步骤。

§ 6.1 假设检验概述

鉴于本章主要讨论单参数假设检验问题，故本节以此作为背景来探讨一般的假设检验问题。将分成四个问题细述之。

一、假设检验问题的提法

现设总体 X 的分布函数为 $F(x; \theta)$ ，其中分布函数的形式是已知的， θ 是一个待检验的参数。需要强调一点的是， θ 仅是一个待检验的参数，而分布函数本身可能还含有其他的参数。例如，当总体 $X \sim N(\mu, \sigma^2)$ 时，若取 $\theta = \mu$ 为待检验的参数，则分布还含有另一参数 σ^2 ；若取 $\theta = \sigma^2$ 为待检验的参数，分布便含有另一参数 μ 。

同前一章一样, 仍记 Θ 为待检验参数 θ 的可能取值范围, 即参数空间. 所谓单参数假设检验问题, 便是要对参数 θ 先作出某种假设, 然后再检验该假设是否成立. 至于对参数 θ 所作的假设, 实际上是假定该参数 θ 属于参数空间 Θ 的某一预先指定的子集 Θ_0 , 通常记为:

$$H_0: \theta \in \Theta_0.$$

以下恒称上述假设为零假设、原假设或基本假设. 为了检验该假设是否成立, 通常需取总体 X 的一个样本 (X_1, X_2, \dots, X_n) , 然后根据该样本提供的信息, 判断上述假设是否成立. 判断的结果不外乎两个. 一种可能是接受该假设, 即认为此假设成立; 另一种可能是否定上述假设, 即认为该假设不成立. 如果出现后一种情况, 有时还需提供另一个假设作备择用:

$$H_1: \theta \in \Theta_1, \quad \text{其中 } \Theta_1 = \Theta - \Theta_0.$$

以下恒称此假设为零假设的备择假设或对立假设. 一旦否定原假设时, 便接受此假设.

综上所述, 通常可把关于总体分布中某一选定的参数 θ 的假设检验问题简记为:

$$H_0: \theta \in \Theta_0 \quad H_1: \theta \in \Theta_1.$$

假设检验的任务便是根据样本提供的信息, 作出否定原假设 (从而接受备择假设) 或接受原假设的决断.

例 6.1 糖厂用自动包装机将糖装箱, 以利外运. 每箱的标准重量规定为 100kg. 每天开工时, 需要先检验一下包装机工作是否正常. 根据以往的经验知道, 用自动包装机装箱, 其重量的起伏是服从正态分布的, 且已知各箱重量的标准差 $\sigma = 1.15\text{kg}$. 某日开工后, 抽测了九箱, 其重量如下 (单位为 kg):

99.3, 98.7, 100.5, 101.2, 98.3, 99.7, 99.5, 102.1, 100.5.

试问这天包装机工作是否正常?

就上例而言, 总体 $X \sim N(\mu, \sigma^2)$, 其中 $\sigma^2 = (1.15)^2$ 是已知的, μ 是待检验的参数. 此例的参数空间为 $\Theta = R^1$, 假设检验问题可简记为:

$$H_0: \mu = 100 \quad H_1: \mu \neq 100.$$

如零假设或备择假设中指定的参数空间的子集 Θ_0 或 Θ_1 仅包含一个值, 便称相应的假设为简单假设; 否则, 称为复合假设. 上例中零假设为简单假设, 备择假设则为复合假设.

二、假设检验的基本统计思想

以上我们给出了假设检验问题的表述. 接下来自然会问: 如何根据样本提供的信息, 在原假设与备择假设之间作抉择呢? 基本的统计思想是小概率原理, 即认为“小概率事件在一次试验中是不太可能发生的”那么, 在假设检验中是

如何运用小概率原理的呢？通常可分为两个步骤.

首先，给定一个很小的正数 α ，并定义一个可由样本 (X_1, X_2, \dots, X_n) 描述的事件 A ，使得在“零假设 H_0 成立”的前提下，该事件发生的概率不会超过 α ，即

$$P(A|H_0) \leq \alpha.$$

其次，根据样本的一次观测结果，即样本值 (x_1, x_2, \dots, x_n) ，来判断事件 A 是否发生. 如果这一事件果真在一次试验中发生了，那么假定“零假设 H_0 成立”，便和小概率原理矛盾. 但小概率原理本身是无可置疑的；那么，唯一可质疑的便是“零假设 H_0 成立”这一前提，因为事件 A 是在“零假设 H_0 成立”的前提下方成为一个小概率事件. 这样，便有力地否定了零假设. 相反，若由样本的观测值 (x_1, x_2, \dots, x_n) 判断出事件 A 没有发生，和小概率原理就不矛盾，因此没有充足的理由否定零假设 H_0 ，从而“只得”接受零假设 H_0 .

综上所述，简单地说，假设检验的基本思路是一种基于小概率原理的反证法.

三、检验的否定域和显著性水平

在明确了假设检验的基本步骤后，接下来一个颇具技术性的问题便是，如何根据给定的正数 α ，定义一个可由样本 (X_1, X_2, \dots, X_n) 描述的事件 A ，使得在“零假设 H_0 成立”的前提下，该事件发生的概率不会超过 α ？我们以下述检验问题为例来探讨解决这一技术性问题的一般途径.

假设总体 $X \sim N(\mu, \sigma_0^2)$ ，其中 σ_0^2 为一已知正数. 待检验的假设为：

$$H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0.$$

上述假设中 μ_0 为一指定的已知常数. 现设 (X_1, X_2, \dots, X_n) 是总体 X 的一个容量为 n 的样本. 由第四章 § 4.4 节的定理 4.2 知，

$$U = \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} \sim N(0, 1),$$

其中 \bar{X} 为样本均值.

上述样本函数 U 含待检验的参数 μ 从而是一枢轴量. 由于它服从标准正态分布，根据标准正态分布双侧分位数的定义有下列关系式：

$$P \left| \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} \right| > u_{\frac{\alpha}{2}} = \alpha, \quad (6.1)$$

其中 $u_{\frac{\alpha}{2}}$ 为标准正态分布的水平 α 的双侧分位数.

现在，构造一个样本空间的子集：

$$C = (X_1, X_2, \dots, X_n) : \left| \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} \right| > u_{\frac{\alpha}{2}}, \quad (6.2)$$

其中 \bar{x} 为样本值 (x_1, x_2, \dots, x_n) 的均值. 注意, 由于 μ_0 与 σ_0 皆为已知常数, C 便是样本空间一个完全确定的子集. 然后, 可把事件 A 定义为:

$$A = \{(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n) \in C\}. \quad (6.3)$$

显然, 这是一个完全由样本描述的事件, 且满足

$$\begin{aligned} P(A|H_0) &= P\{(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n) \in C|H_0\} \\ &= P\left\{\left|\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}\right| > u_{\frac{\alpha}{2}}|H_0\right\} \\ &= \alpha. \end{aligned}$$

上述分析表明, 只要正数 α 取得充分小, 在“零假设 H_0 成立”的前提下, 由 (6.2) 与 (6.3) 式定义的事件 A 发生的概率恰为 α , 是一个小概率事件. 这样, 若样本 (x_1, x_2, \dots, x_n) 的某一观测值 $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$ 满足:

$$\left|\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right| > u_{\frac{\alpha}{2}},$$

说明小概率事件 A 发生了, 便拒绝零假设 H_0 , 从而接受备择假设 H_1 ; 否则, 便接受零假设 H_0 .

例 6.1 的解: 显然, 例 6.1 是上述检验问题的特例, 相应于 $\mu_0 = 1.15$, $\mu = 100$, $n = 9$. 现取 $\alpha = 0.05$, 查附表 2 知 $u_{\frac{\alpha}{2}} = 1.96$, 再由例 6.1 中所列的样本值可算出 $\bar{x} = 99.98$. 这样, 便可计算

$$u = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} = \frac{99.98 - 100}{1.15/\sqrt{9}} = -0.052.$$

由于 $|-0.052| < u_{\frac{\alpha}{2}} = 1.96$, 便没有充分的理由否定零假设 H_0 , 从而认为这天自动包装机的工作是正常的.

由前述假设检验问题可启发我们对于一般的单参数假设检验问题的思考. 设总体 X 的分布函数为 $F(x; \theta)$, 分布函数所属的类型已知, θ 为待检验的参数. 待检验的假设为:

$$H_0: \theta = \theta_0 \quad H_1: \theta = \theta_1 \quad (\theta_0 \neq \theta_1). \quad (6.4)$$

现设 (X_1, X_2, \dots, X_n) 是总体的一个容量为 n 的样本. 由上述特定的假设检验问题看出, 为定义一个可由样本 (X_1, X_2, \dots, X_n) 描述的事件 A , 首先要指定样本空间中一个完全确定的子集 C (对于前述假设检验问题, 子集 C 由 (6.2) 式给出). 然后, 可把事件 A 定义为:

$$A = \{(X_1, X_2, \dots, X_n) \in C\}. \quad (6.5)$$

如前所述, 为在假设检验问题中运用小概率原理, 关键的一点是要求在“零假设 H_0 成立”的前提下, 由 (6.5) 式定义的事件 A 是一小概率事件, 即应有

$$P(A|H_0) = P\{(X_1, X_2, \dots, X_n) \in C|H_0\} = \alpha, \quad (6.6)$$

其中 α 是一个预先指定的很小的正数, 以下通称其为检验的显著性水平. 一般说

来,它是根据实际问题的需要由检验者预先指定的,常取的值为 0.01, 0.05 与 0.10.

如果样本空间的子集 C 满足(6.6)式中的要求,由小概率原理即可推知,一旦样本的观测值 (x_1, x_2, \dots, x_n) 属于子集 C ,便应否定零假设 H_0 ,从而接受备择假设 H_1 ;而当样本值 (x_1, x_2, \dots, x_n) 不属于 C 时,则接受零假设 H_0 .今后称具有上述性质的样本空间的子集 C 为零假设 H_0 的否定域,简称为否定域或拒绝域;类似地,称 C 的补集 \bar{C} 为零假设 H_0 的接受域,简称为接受域.

综上所述,解决单参数假设检验问题的关键是构造一个合适的 C 并保证(6.6)式成立.由前述假设检验问题知,解决这一难点的基本思路是构造一个仅含待检验参数 μ 的枢轴量 $U(X_1, X_2, \dots, X_n; \mu)$ (在前述假设检验问题中取为 $U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$),并使得这一枢轴量服从一个已知的确定分布(前述假设检验问题中 $U \sim N(0, 1)$).因枢轴量服从已知的确定分布,便可利用该分布的上侧或双侧分位数,再结合零假设 H_0 构造否定域 C (前述假设检验问题中在构造由(6.2)式定义的否定域 C 时,以零假设 H_0 中的已知常数 μ 替换了待检验的参数 μ).

至此,我们可把解决单参数假设检验问题(6.4)的一般步骤归纳如下:

1. 当给定总体 X 的样本 (X_1, X_2, \dots, X_n) 后,首先构造一个仅含待检验参数 μ 的枢轴量 $U(X_1, X_2, \dots, X_n; \mu)$,它服从一个已知的确定的分布;
2. 对于给定的检验的显著性水平 α ,由上述枢轴量与其所服从的已知的确定分布的上侧分位数或双侧分位数,再结合零假设 H_0 ,确定检验的否定域 C ,使得

$$P((X_1, X_2, \dots, X_n) \in C | H_0) = \alpha;$$

3. 对于样本的某一具体的观测值 (x_1, x_2, \dots, x_n) ,判断它是否属于否定域 C .若属于否定域 C ,便拒绝零假设 H_0 (从而接受备择假设 H_1);若属于接受域 \bar{C} ,便接受零假设 H_0 .

由此可见,一旦确定了检验的否定域 C ,事实上也就给出了相应的检验法则.这里我们需强调一点,作为上述假设检验步骤的基点是,枢轴量 $U(X_1, X_2, \dots, X_n; \mu)$ 需服从已知的确定分布.对于有些总体,如正态总体,这一要求能满足,这属小样本统计范畴.不过,对于一般总体来说,上述要求较难满足,这时便可让样本容量 n 趋于无穷,转而考虑枢轴量的极限分布.如相应的极限分布是一个已知的确定分布,那么只要样本容量 n 充分大,该极限分布也可替代枢轴量的精确分布,近似地用作关于参数 μ 的假设检验问题,这属大样本统计范畴.

四、检验的两类错误

以上我们总结了单参数假设检验的一般步骤,现着重说明一下检验的显著

性水平在假设检验问题中所起的作用，同时也引出了关于检验的两类错误的概念。

一般说来，统计推断的特点是由样本提供的信息来推断总体。这样，推断时所下的结论未必总是正确的。就假设检验问题而言，如零假设 H_0 事实上是成立的，但却因样本观察值落入否定域而拒绝了零假设 H_0 ，这便犯了弃真错误，通常称为第一类错误；相反，如零假设 H_0 不成立，却因样本的观测值落入接受域，而接受了零假设 H_0 时，便犯了纳伪错误，通常称为犯第二类错误。根据检验法则和 (6.6) 式，我们可以断言，当零假设 H_0 成立时，拒绝零假设的概率至多为显著性水平 α ，这表明犯第一类错误的概率至多为 α ，从而说明检验的显著性水平是用以控制犯第一类错误的概率的。由此可能会产生一种错觉，以为只要把显著性水平 α 取得越小，假设检验的准确程度就越高。事实上不然，因为显著性水平只是用来控制犯第一类错误的，而在假设检验中还存在着犯第二类错误的可能性。一般说来，当样本容量给定时，在降低显著性水平的同时，往往会增大犯第二类错误的可能性。要同时减少犯两种错误的可能性，只有通过增加样本的容量才能达到目的。

五、多参数与非参数假设检验问题

以上我们对单参数假设检验问题作了较详尽的介绍。原则上说来，以上介绍的所有内容也适用于多参数假设检验或非参数假设检验问题，只需在若干细节上作适当调整即可。为此，仅说明两点：

1. 对于多参数假设检验问题，可寻求一个包含所有待检验参数的枢轴量，并使之服从或渐近地服从一个已知的确定分布；
2. 非参数假设检验问题可近似地化为一个多参数假设检验问题来解决。

鉴于正态总体是统计应用中最常见的总体，我们将首先在以下两节分别讨论单正态总体与双正态总体的参数假设检验。

§ 6.2 单正态总体的参数假设检验

本节考虑总体 $X \sim N(\mu, \sigma^2)$ 的参数假设检验问题。鉴于正态总体含两个参数 μ 与 σ^2 ，我们将分别考虑对于总体数学期望 μ 与总体方差 σ^2 的参数假设检验。

一、关于总体数学期望 μ 的假设检验

当检验关于总体数学期望 μ 的假设时，另一参数，即方差 σ^2 是否已知，会影响到对于枢轴量的选择，故分两种情形进行讨论。

1. 方差 σ^2 已知的情形

设总体 $X \sim N(\mu, \sigma_0^2)$, 其中 σ_0^2 已知, μ 是待检验的参数. 又设 (X_1, X_2, \dots, X_n) 是总体的一个容量为 n 的样本. 首先, 如上节所示, 可构造一个仅含待检验参数 μ 的枢轴量:

$$U = \frac{\bar{X} - \mu}{\sigma_0 / \sqrt{n}} \quad (6.7)$$

其中 \bar{X} 为样本均值. 如前所述, 上述枢轴量 U 服从标准正态分布.

关于参数 μ 可提出三种不同类型的假设检验问题, 分别对应于三种不同的检验法则.

1.1 双侧检验法

双侧检验法适用于下述假设检验问题:

$$H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0.$$

上述假设中, μ_0 为一指定的常数. 这一假设检验问题已在上节中讨论过, 对给定显著性水平 α , 假设的否定域 C 为

$$C = (X_1, X_2, \dots, X_n); \left| \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} \right| > u_{\alpha/2} \quad (6.8)$$

1.2 右侧检验法

右侧检验法适用于下述假设检验问题:

$$H_0: \mu \leq \mu_0 \quad H_1: \mu > \mu_0.$$

不同于上述双侧检验法, 这里的零假设 H_0 为复合假设. 因由 (6.7) 式定义的枢轴量服从标准正态分布, 于是, 当给定显著性水平 α 后, 由标准正态分布的上侧分位数的定义可建立下述关系式:

$$P \left(\frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} > u_{\alpha} \right) = \alpha \quad (6.9)$$

其中 u_{α} 为标准正态分布的水平 α 的上侧分位数.

由此, 可如下确定检验的否定域:

$$C = (X_1, X_2, \dots, X_n): \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} > u_{\alpha} \quad (6.10)$$

再因在 $H_0: \mu \leq \mu_0$ 成立时

$$\frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} \leq \frac{\bar{X} - \mu}{\sigma_0 / \sqrt{n}}$$

此时

$$\frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} > u_{\alpha} \Rightarrow \frac{\bar{X} - \mu}{\sigma_0 / \sqrt{n}} > u_{\alpha} \quad (6.11)$$

从而有

$$P\{(X_1, X_2, \dots, X_n) \in C | H_0\} = P\left\{\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} > u_{\alpha} | H_0\right\} \\ = P\left\{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > u_{\alpha} | H_0\right\} \\ = \dots$$

例 6.2 有一工厂生产一种灯管, 已知灯管的寿命 X 服从正态分布 $N(\mu, 40000)$, 根据以往的生产经验, 知道灯管的平均寿命不会超过 1500 小时. 为了提高灯管的平均寿命, 工厂采用了新的工艺. 进一步为了弄清楚新工艺是否真的能提高灯管的平均寿命, 他们测试了新工艺生产的 25 只灯管的寿命, 其平均值是 1575 小时. 尽管样本的平均值大于 1500 小时, 试问: 可否由此判定这恰是新工艺的效应, 而非偶然的原因使得抽出的这 25 只灯管的平均寿命较长呢?

解 可把上述问题提成下述假设检验问题:

$$H_0: \mu \leq 1500 \quad H_1: \mu > 1500.$$

从而可利用右侧检验法来解, 相应于 $\mu = 1500$, $\sigma_0 = 200$, $n = 25$. 取显著性水平 $\alpha = 0.05$, 查附表已知 $u_{\alpha} = 1.645$, 因已测出 $\bar{x} = 1575$, 从而可算得

$$u = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} = \frac{1575 - 1500}{200/\sqrt{25}} = 1.875.$$

由于 $u = 1.875 > u_{\alpha} = 1.645$, 从而否定零假设 H_0 , 接受备择假设 H_1 , 即认为新工艺事实上提高了灯管的平均寿命.

1.3 左侧检验法

左侧检验法适用于下述假设检验问题:

$$H_0: \mu \geq \mu_0 \quad H_1: \mu < \mu_0.$$

类似于右侧检验法, 当给定显著性水平 α 后, 可将否定域取为

$$C = \{(X_1, X_2, \dots, X_n): \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} < -u_{\alpha}\} \quad (6.12)$$

区别于双侧检验法, 今后统称右侧检验法与左侧检验法为单侧检验法.

至此, 当总体方差 σ^2 已知时, 关于总体数学期望 μ 的假设检验已全部介绍完毕. 今后, 在考虑假设检验问题时, 当最终是借助枢轴量服从或渐近地服从标准正态分布作为出发点的, 便称相应的检验法为 U 检验法. 显然, 当总体方差已知, 上述关于正态总体数学期望的三类假设检验问题皆采用了 U 检验法.

2. 方差 σ^2 未知的情形

设总体 $X \sim N(\mu, \sigma^2)$, 其中总体方差 σ^2 未知, μ 是待检验的参数. (X_1, X_2, \dots, X_n) 是总体 X 的容量为 n 的一个样本. 类似于总体方差 σ^2 为已知的情形, 我们完全可以平行地讨论关于总体数学期望 μ 的假设检验问题. 仅有的差别在于, 我们不能再以由 (6.7) 式定义的枢轴量作为出发点来考虑问题. 这是

因为现在总体的方差 σ^2 未知, 而枢轴量除含待检验的参数 μ 外, 不能再含其他未知的参数. 为此改取

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (6.13)$$

作为枢轴量, 其中 \bar{X} 为样本均值, S 为样本标准差 (它替换了 (6.7) 式中的 σ). 由第四章 4.4 节的定理 4.2 知, 上述枢轴量 T 服从自由度为 $n-1$ 的 t 分布. 再由于 t 分布与标准正态分布一样, 也具有对称密度, 因此方差 σ^2 为已知情形时的讨论方法完全适用于现在的情形, 以下仅列出三类假设检验问题的否定域, 并恒以 α 表示给定的检验的显著性水平, μ_0 表示指定的已知常数.

2.1 双侧检验法

待检验的假设为:

$$H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0.$$

对于上述假设检验问题, 否定域可取为

$$C = \{(x_1, x_2, \dots, x_n) : \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| > t_{\alpha/2}(n-1)\} \quad (6.14)$$

其中 $t_{\alpha/2}(n-1)$ 为水平 α 的自由度为 $n-1$ 的 t 分布的双侧分位数, \bar{x} 与 s 分别为样本值的均值与标准差.

2.2 右侧检验法

待检验的假设为:

$$H_0: \mu \leq \mu_0 \quad H_1: \mu > \mu_0.$$

对于上述假设检验问题, 否定域可取为:

$$C = \{(x_1, x_2, \dots, x_n) : \frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{\alpha}(n-1)\} \quad (6.15)$$

其中 $t_{\alpha}(n-1)$ 为水平 α 的自由度为 $n-1$ 的 t 分布的上侧分位数.

2.3 左侧检验法

待检验的假设为:

$$H_0: \mu \geq \mu_0 \quad H_1: \mu < \mu_0.$$

对于上述假设检验问题, 否定域可取为:

$$C = \{(x_1, x_2, \dots, x_n) : \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t_{\alpha}(n-1)\} \quad (6.16)$$

今后, 在考虑假设检验问题时, 当最终是借助枢轴量服从或渐近地服从 t 分布作为出发点的, 便称相应的检验方法为 T 检验法. 显然, 当总体方差 σ^2 未知时, 关于正态总体数学期望 μ 的三类假设检验问题, 皆采用了 T 检验法.

例 6.3 (续例 6.1) 同例 6.1, 但假定标准差 σ 未知. 这时可采用 T 检验法的双侧检验法来解. 本例中 $\mu_0 = 100$, $n = 9$. 显著性水平 α 取为 0.05. 查附

表 5 知, $t_{/2} (n-1) = t_{0.025} (8) = 2.306$. 在例 6.1 中已算出 $\bar{x} = 99.98$, 现求

$$s^2 = \frac{1}{8} \sum_{i=1}^9 (x_i - \bar{x})^2 = \frac{1}{8} \sum_{i=1}^9 x_i^2 - 9\bar{x}^2 = 1.469,$$

从而 $s = 1.212$. 由此, 可计算

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{99.98 - 100}{1.212/\sqrt{9}} = -0.495$$

这样, 因 $-0.495 < 2.306 = t_{0.025} (8)$, 便不能否认零假设 H_0 , 从而认为这天包装机的工作是正常的.

例 6.4 一公司声称某种类型的电池的平均寿命至少为 2.15 小时. 有一实验室检验了该公司制造的 6 套, 得到如下的寿命小时数:

$$19, 18, 22, 20, 16, 25$$

试问: 这些结果是否表明, 这种类型的电池低于该公司所声称的寿命? (显著性水平 $\alpha = 0.05$)

解 可把上述问题归纳为下述假设检验问题:

$$H_0: \mu \geq 2.15 \quad H_1: \mu < 2.15.$$

这可利用 T 检验法的左侧检验法来解. 本例中 $\mu = 2.15$, $n = 6$. 对于给定的显著性水平 $\alpha = 0.05$, 查附表 5 知 $t_{\alpha} (n-1) = t_{0.05} (5) = 2.015$. 再据测得的 6 个寿命小时数算得:

$$\bar{x} = 20, \quad s^2 = 10.$$

由此计算

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{20 - 2.15}{\sqrt{10}/\sqrt{6}} = -1.162.$$

这样, 因 $t = -1.162 > -2.015 = -t_{0.05} (5)$, 所以不能否定零假设 H_0 , 从而认为这种类型电池的寿命并不比公司宣称的寿命为短.

二、关于总体方差 σ^2 的假设检验

以上我们详尽地讨论了关于正态总体数学期望 μ 的假设检验问题. 以下将考虑关于正态总体方差的假设检验.

设总体 $X \sim N(\mu, \sigma^2)$, σ^2 是待检验的参数. 显然, 在构造含待检验参数 σ^2 的枢轴量时, 也应根据数学期望 μ 已知或未知的情形分别予以考虑. 但由于所用的方法基本上是相似的. 我们只介绍数学期望 μ 未知的情形, 而把数学期望 μ 已知的情形留待读者自己考虑. 现设 (X_1, X_2, \dots, X_n) 是总体的容量为 n 的一个样本, 令

$$W = \frac{n-1}{2} S^2 \quad (6.17)$$

其中 S^2 是样本的方差. 由第四章 § 4.4 节的定理 4.2 知, 上述枢轴量 W 服从自由度为 $n-1$ 的 χ^2 分布. 关于总体方差 σ^2 也可提出三种不同类型的假设检验问题, 下面我们较详尽地叙述了右侧检验法. 对于双侧检验法与左侧检验法的叙述较简略, 读者可自行补出需要说明的有关细节. 以下恒设 α 是给定的显著性水平, σ_0^2 是一指定的已知正数.

1. 双侧检验法

双侧检验法适用于下述检验问题:

$$H_0: \sigma^2 = \sigma_0^2 \quad H_1: \sigma^2 \neq \sigma_0^2.$$

首先, 由于根据 (6.17) 式定义的枢轴量服从自由度为 $n-1$ 的 χ^2 分布, 由 χ^2 分布上侧分位数的定义可建立下述关系式:

$$P\left(\frac{n-1}{2}S^2 < \chi_{1-\frac{\alpha}{2}}^2(n-1) \quad \frac{n-1}{2}S^2 > \chi_{\frac{\alpha}{2}}^2(n-1)\right) = \alpha, \quad (6.18)$$

其中 $\chi_{1-\frac{\alpha}{2}}^2(n-1)$ 与 $\chi_{\frac{\alpha}{2}}^2(n-1)$ 分别为水平 $1-\frac{\alpha}{2}$ 与水平 $\frac{\alpha}{2}$ 的自由度为 $n-1$ 的 χ^2 分布的上侧分位数. 由此可如下确定检验的否定域:

$$C = (X_1, X_2, \dots, X_n): \frac{n-1}{2}S^2 < \chi_{1-\frac{\alpha}{2}}^2(n-1) \text{ 或 } \frac{n-1}{2}S^2 > \chi_{\frac{\alpha}{2}}^2(n-1). \quad (6.19)$$

其中 s^2 是样本方差 S^2 的观察值.

2. 右侧检验法

右侧检验法适用于下述假设检验问题:

$$H_0: \sigma^2 \leq \sigma_0^2 \quad H_1: \sigma^2 > \sigma_0^2.$$

鉴于上述零假设 H_0 为一复合假设, 可由 χ^2 分布的上侧分位数的定义建立下述关系式:

$$P\left(\frac{n-1}{2}S^2 > \chi_{\alpha}^2(n-1)\right) = \alpha, \quad (6.20)$$

其中 $\chi_{\alpha}^2(n-1)$ 是水平 α 的自由度为 $(n-1)$ 的 χ^2 分布的上侧分位数. 由此, 可如下确定检验的否定域:

$$C = (X_1, X_2, \dots, X_n): \frac{n-1}{2}S^2 > \chi_{\alpha}^2(n-1). \quad (6.21)$$

这是因为当 $H_0: \sigma^2 \leq \sigma_0^2$ 成立时,

$$\frac{n-1}{2}S^2 \leq \frac{n-1}{2}\sigma_0^2$$

此时

$$\frac{n-1}{2}S^2 > \chi_{\alpha}^2(n-1) \quad \frac{n-1}{2}S^2 > \chi_{\alpha}^2(n-1) \quad (6.22)$$

从而, 有

$$P\{(X_1, X_2, \dots, X_n) \in C | H_0\} = P\left\{\frac{n-1}{2} S^2 > \chi^2_{1-\alpha}(n-1) | H_0\right\} \\ = P\left\{\frac{n-1}{2} S^2 > \chi^2_{1-\alpha}(n-1)\right\} \\ =$$

3. 左侧检验法

左侧检验法适用于下述假设检验问题:

$$H_0: \sigma^2 = \sigma_0^2 \quad H_1: \sigma^2 < \sigma_0^2.$$

上述零假设 H_0 也是一个复合假设, 故类似地可取下述形式的否定域:

$$C = \{(X_1, X_2, \dots, X_n): \frac{n-1}{2} S^2 < \chi^2_{1-\alpha}(n-1)\} \quad (6.23)$$

其中 $\chi^2_{1-\alpha}(n-1)$ 是水平 $(1-\alpha)$ 的自由度为 $n-1$ 的 χ^2 分布的上侧分位数.

为叙述时方便起见, 类似于前述的命名法, 今后在考虑假设检验问题时, 当最终是借助枢轴量服从或渐近地服从 χ^2 分布作为出发点的, 便称相应的检验方法为 χ^2 检验法. 显然, 上述关于总体方差的三类假设检验问题皆采用了 χ^2 检验法. 有兴趣的读者可以考虑一下, 若总体数学期望 μ 已知, 记为 $\mu = \mu_0$, μ_0 为一已知常数, 这时可选择怎样一个更为简便的枢轴量来替代由 (6.17) 式定义的枢轴量 W ? 此外, 可对给定的显著性水平 α , 试着写出相应的三类假设检验问题的否定域 (此已留作习题六 (B) 的第 1 题, 读者可自行解之).

例 6.5 某炼铁厂铁水的含碳量 X , 在正常情况下服从正态分布. 现对操作工艺进行了某些改变, 从中抽取了 7 炉铁水的试样, 测得含碳量数据 (单位为千克) 如下:

4.421, 4.052, 4.357, 4.394, 4.326, 4.287, 4.683

试问: 是否可以认为新工艺炼出的铁水含碳量的方差仍为 0.112^2 ? (显著性水平 $\alpha = 0.05$)

解 可把上述问题归纳为下述假设检验问题:

$$H_0: \sigma^2 = 0.112^2 \quad H_1: \sigma^2 < 0.112^2.$$

这样, 便可采用 χ^2 检验法的双侧检验法来解. 对于给定的显著性水平 $\alpha = 0.05$, 查附表 3

$$\text{知} \quad \chi^2_{1-\frac{\alpha}{2}}(n-1) = \chi^2_{0.975}(6) = 1.237, \\ \chi^2_{\frac{\alpha}{2}}(n-1) = \chi^2_{0.025}(6) = 14.449.$$

再由 7 个含碳量数据可算得

$$\bar{x} = 4.36, (n-1)s^2 = \sum_{i=1}^7 (x_i - \bar{x})^2 = 0.2106.$$

由此, 可算出

$$w = \frac{(n-1) s^2}{\sigma_0^2} = \frac{0.2106}{0.112^2} = 16.789.$$

这样，因 $w = 16.789 > 14.449 = \chi_{0.025}^2(6)$ ，所以拒绝零假设，从而认为采用新工艺后铁水含碳量的方差发生了变化。

例 6.6 某工厂生产金属丝，产品指标为折断力。折断力的方差被用作工厂生产精度的表征。方差越小，表明精度越高。以往工厂一直把该方差保持在 64 (kg^2) 与 64 以下。最近从一批产品中抽取 10 根作折断力试验，测得的结果（单位为千克）如下：

578, 572, 570, 568, 572, 570, 572, 596, 584, 570.

由上述样本数据算得：

$$\bar{x} = 575.2, \quad s^2 = 75.74.$$

为此，厂方怀疑金属丝折断力的方差是否变大了。如确实增大了，表明生产精度不如以前，就需对生产流程作一番检验，以发现生产环节中存在的问题。为确认上述疑虑是否为真，假定金属丝折断力服从正态分布，并作下述假设检验：

$$H_0: \sigma^2 \leq 64 \quad H_1: \sigma^2 > 64.$$

上述假设检验问题可利用 χ^2 检验法的右侧检验法来解。就本例而言，相应于 $\sigma_0^2 = 64$ ， $n = 10$ 。对于给定的显著性水平 $\alpha = 0.05$ ，查附表 3 知，

$$\chi_{0.05}^2(n-1) = \chi_{0.05}^2(9) = 16.919. \quad \text{这样，有}$$

$$w = \frac{(n-1) s^2}{\sigma_0^2} = \frac{9 \times 75.74}{64} = 10.65 < 16.919 = \chi_{0.05}^2(9).$$

于是，不能拒绝零假设 H_0 ，从而认为样本方差的偏大系偶然因素，生产流程正常，故不需再作进一步的检查。

迄今我们已较为系统地介绍了单正态总体的参数假设检验，现把已提及的所有假设检验问题的要点总结于表 6-1 中供读者参考。

§ 6.3 双正态总体的参数假设检验

上一节我们详尽地讨论了单正态总体的参数假设检验，本节将考虑双正态总体的参数假设检验。确切地说，设有两个相互独立的正态总体： $X \sim N(\mu_1, \sigma_1^2)$ ， $Y \sim N(\mu_2, \sigma_2^2)$ 。现我们关心的不是逐一对每个参数的值作假设检验，而是着重考虑两个总体之间的差异。为此可令 $\mu = \mu_1 - \mu_2$ ， $r = \sigma_1^2 / \sigma_2^2$ ，其中 μ 表示两个总体数学期望的差异， r 则表示两个总体方差的比值。当引入这两个新的参数后，便可将假设检验问题化为对上述两个新引入的参数的假设检验问题。以下恒设 $(X_1, X_2, \dots, X_{n_1})$ 是总体 X 的容量为 n_1 的一个样本， $(Y_1, Y_2, \dots,$

Y_{n_2}) 是总体 Y 的容量为 n_2 的一个样本, 并以 X 和 Y 分别表示这两个样本的样本均值, 再以

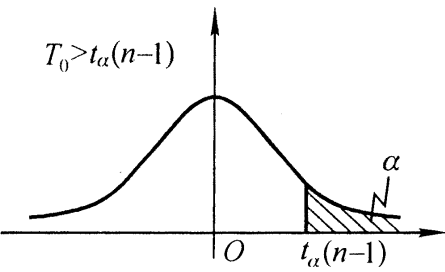
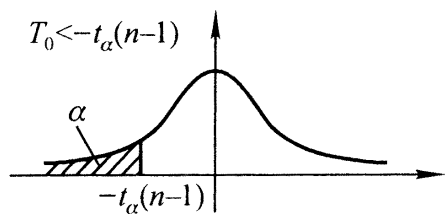
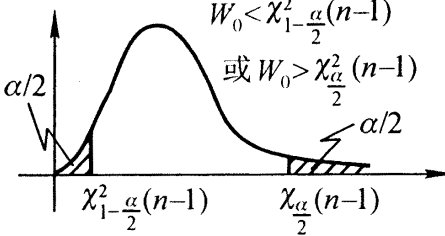
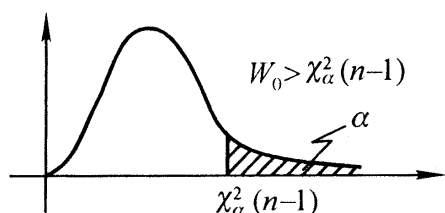
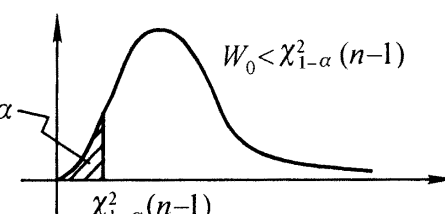
$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2, S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 \tag{6.24}$$

分别表示这两个样本的样本方差. 此外, 以 S^2 表示上述两个样本方差的加权平均, 即记

表 6-1 单正态总体的假设检验一览表

已知条件	零假设与备择假设	枢轴量 (统计量)	应查分布表	否定域
$X \sim N(\mu, \sigma^2)$, μ 未知, σ^2 已知, μ_0 是指定的数.	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$U = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$ $U_0 = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$	标准正态分布表	
	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$			
	$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$			
$X \sim N(\mu, \sigma^2)$, μ 与 σ^2 未知, μ_0 是指定的数.	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$ $T_0 = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$	自由度为 $n - 1$ 的 t 分布表	

续表

已知条件	零假设与备择假设	枢轴量 (统计量)	应查分布表	否定域
$X \sim N(\mu, \sigma^2)$, μ 与 σ^2 未知, μ 是指定的数.	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ $T_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	自由度为 $n-1$ 的 t 分布表	 <p>A graph of the t-distribution curve centered at O. The horizontal axis is marked with $t_{\alpha}(n-1)$. The area under the curve to the right of this point is shaded and labeled α. The text $T_0 > t_{\alpha}(n-1)$ is written above the curve.</p>
	$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$			 <p>A graph of the t-distribution curve centered at O. The horizontal axis is marked with $-t_{\alpha}(n-1)$. The area under the curve to the left of this point is shaded and labeled α. The text $T_0 < -t_{\alpha}(n-1)$ is written above the curve.</p>
$X \sim N(\mu, \sigma^2)$, μ 与 σ^2 未知, σ^2 ($\sigma^2 > 0$) 是指定的数.	$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$	$W = \frac{n-1}{2} S^2$ $W_0 = \frac{n-1}{2} s^2$	自由度为 $n-1$ 的 χ^2 分布表	 <p>A graph of the chi-square distribution curve. The horizontal axis is marked with $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ and $\chi^2_{\frac{\alpha}{2}}(n-1)$. The area under the curve to the right of $\chi^2_{\frac{\alpha}{2}}(n-1)$ is shaded and labeled $\alpha/2$. The text $W_0 < \chi^2_{1-\frac{\alpha}{2}}(n-1)$ or $W_0 > \chi^2_{\frac{\alpha}{2}}(n-1)$ is written above the curve.</p>
	$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$			 <p>A graph of the chi-square distribution curve. The horizontal axis is marked with $\chi^2_{\alpha}(n-1)$. The area under the curve to the right of this point is shaded and labeled α. The text $W_0 > \chi^2_{\alpha}(n-1)$ is written above the curve.</p>
	$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$			 <p>A graph of the chi-square distribution curve. The horizontal axis is marked with $\chi^2_{1-\alpha}(n-1)$. The area under the curve to the left of this point is shaded and labeled α. The text $W_0 < \chi^2_{1-\alpha}(n-1)$ is written above the curve.</p>

$$S^2 = \frac{1}{n_1 + n_2 - 2} (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 \quad (6.25)$$

对于样本值 $(x_1, x_2, \dots, x_{n_1})$ 与 $(y_1, y_2, \dots, y_{n_2})$, 以下分别以小写字母 x, y, s_1^2, s_2^2 与 s^2 表示上述诸统计量相应的观测值. 本节将统一地采用上述诸记法, 不另作赘述.

本节将假设检验问题分为两组. 一组可归结为对于期望差异值 μ 的假设检验, 另一组可归结为对于方差比值 r 的假设检验.

一、可归结为对于两总体期望差异值 μ 的假设检验

当检验参数 μ 时, 需构造一个仅含待检验参数 μ 的枢轴量, 但由于两总体还含其他的参数 σ_1^2 和 σ_2^2 , 它们的已知与未知将影响到对于枢轴量的选择, 故再区分两种情形.

1. 两总体方差 σ_1^2 与 σ_2^2 为已知的情形

由于 σ_1^2 与 σ_2^2 为已知常数, 可令

$$\sigma_0^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}, \quad (6.26)$$

这也是一个确定的已知值. 于是, 可构造下述枢轴量

$$U = \frac{(\bar{X} - \bar{Y}) - \mu}{\sigma_0} \quad (6.27)$$

由第四章 § 4.4 节的定理 4.3 知, 枢轴量 U 服从标准正态分布. 这样, 如前一节所述, 我们便可利用 U 检验法对参数 μ 作假设检验. 进一步又可区分为三类假设检验问题. 鉴于 U 检验法已在上一节中有详尽介绍, 我们仅给出三类假设检验问题的否定域, 读者可自行补出需要说明的有关细节. 以下恒设 α 为给定的检验的显著性水平.

1.1 双侧检验法

双侧检验法适用于下述假设检验问题:

$$H_0: \mu = 0 \quad H_1: \mu \neq 0.$$

该假设检验问题等价于下述关于两个正态总体数学期望的假设检验问题:

$$H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0.$$

显然, 上述假设检验问题可采用 U 检验法中(相应于 $\mu = 0$ 时的)双侧检验法来解, 故类似于(6.8)式, 可将否定域取为:

$$C = (x_1, x_2, \dots, x_{n_1}; y_1, y_2, \dots, y_{n_2}): \left| \frac{\bar{x} - \bar{y}}{\sigma_0} \right| > u_{\frac{\alpha}{2}}$$

1.2 右侧检验法

右侧检验法适用于下述假设检验问题:

$$H_0: \mu = 0 \quad H_1: \mu > 0 .$$

该假设检验问题等价于下述关于两个正态总体数学期望的假设检验问题:

$$H_0: \mu = \mu_0 \quad H_1: \mu > \mu_0 .$$

显然, 上述假设检验问题可采用 U 检验中(相应于 $\mu_0 = 0$ 时的)右侧检验法来解, 故类似于(6.10)式, 可将否定域取为:

$$C = (x_1, x_2, \dots, x_{n_1}; y_1, y_2, \dots, y_{n_2}) : \frac{\bar{x} - \bar{y}}{\sigma_0} > u_{\alpha}$$

1.3 左侧检验法

左侧检验法适用于下述假设检验问题:

$$H_0: \mu = 0 \quad H_1: \mu < 0 .$$

该假设检验问题等价于下述关于两个正态总体数学期望的假设检验问题:

$$H_0: \mu = \mu_0 \quad H_1: \mu < \mu_0 .$$

显然, 上述假设检验问题可采用 U 检验中(相应于 $\mu_0 = 0$ 时的)左侧检验法来解, 故类似于(6.12)式, 可将否定域取为:

$$C = (x_1, x_2, \dots, x_{n_1}; y_1, y_2, \dots, y_{n_2}) : \frac{\bar{x} - \bar{y}}{\sigma_0} < -u_{\alpha} .$$

例 6.7 假设 A 厂生产的灯泡的使用寿命 $X \sim N(\mu, 95^2)$, B 厂生产的灯泡的使用寿命 $Y \sim N(\mu, 120^2)$, 现从两厂产品中分别抽取了 100 只和 75 只, 测得灯泡的平均寿命相应为 1180 小时和 1220 小时. 问在显著性水平 $\alpha = 0.05$ 下, 这两个厂家生产的灯泡的平均使用寿命有无显著性差异?

解 本例可采用双侧检验法来解, 相应于 $\sigma_1^2 = 95^2$, $\sigma_2^2 = 120^2$, $n_1 = 100$, $n_2 = 75$. 由此, 据(6.26)式可算出 $\sigma_0 = 16.8$. 再因已知 $\bar{x} = 1180$, $\bar{y} = 1220$, 便可算出

$$u = \frac{\bar{x} - \bar{y}}{\sigma_0} = \frac{1180 - 1220}{16.8} = -2.381.$$

对于给定的显著性水平 $\alpha = 0.05$, 查附表 2 知, $u_{\frac{\alpha}{2}} = u_{0.025} = 1.96$. 这样, 因 $|u| = 2.381 > 1.96$, 从而拒绝零假设 H_0 (或 H_0^*), 即认为两厂生产的灯泡的平均使用寿命有显著差异.

例 6.8 在某学院中, 从比较喜欢参加体育运动的男生中随意选出 50 名, 测得平均身高为 174.3 厘米, 在不愿参加运动的男生中随意选 50 名, 测得其平均身高为 170.4 厘米. 假设两种情形下, 男生的身高都服从正态分布, 其标准差相应为 5.3 厘米与 6.1 厘米. 问该学院中参加体育运动的男生是否比不参加体育运动的男生长得要高些? (显著性水平 $\alpha = 0.05$)

解 以 X 与 Y 分别表示喜欢与不喜欢参加体育运动的男生的身高, 由题设知 $X \sim N(\mu, 5.3^2)$, $Y \sim N(\mu, 6.1^2)$. 可采用右侧检验法来解. 就本例而言, $\sigma_1^2 = 5.3^2$, $\sigma_2^2 = 6.1^2$, $n_1 = n_2 = 50$. 由此, 据(6.26)式可算出 $\sigma_0 = 1.143$. 再因已知 $\bar{x} =$

174.3, $y = 170.4$, 便可算出:

$$u = \frac{\bar{x} - \bar{y}}{s_0} = \frac{174.3 - 170.4}{1.143} = 3.412.$$

对于给定的显著性水平 $\alpha = 0.05$, 查附表 2 知 $u = u_{0.05} = 1.645$. 这样, 因 $u = 3.412 > 1.645 = u_{0.05}$, 便拒绝零假设 H_0 (式 H_0^*), 从而认为喜欢体育运动男生的身高平均说来比一般男生明显偏高.

2. 两总体方差相等但为未知的情形

现设 $\sigma_1^2 = \sigma_2^2 = \sigma^2$, 但 σ^2 未知. 在此前提下, 解决关于两个正态总体数学期望的假设检验问题的基本思路和前面所述的两总体方差 σ_1^2 与 σ_2^2 为已知的情形相似. 仅有的差别在于枢轴量的选择. 现因 $\sigma_1^2 = \sigma_2^2 = \sigma^2$, 但 σ^2 未知, 由 (6.26) 式定义的 s_0 便是未知的, 这样, 由 (6.27) 式定义的样本函数 U 中, 除含待检验的参数 μ 外, 还含有另一未知参数 σ , 便不再是一枢轴量. 考虑到, 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时, 由 (6.24) 与 (6.25) 两式定义的统计量 S^2 恰是 σ^2 的无偏估计量 (见题六(A) 的第 12 题, 读者可自行证明), 便可改取下述枢轴量:

$$T = \frac{(\bar{X} - \bar{Y}) - \mu}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \quad (6.28)$$

由第四章 § 4.4 节的定理 4.3 知, 上述枢轴量 T 服从自由度为 $n_1 + n_2 - 2$ 的 t 分布, 简记为: $T \sim t(n_1 + n_2 - 2)$. 这样, 便可利用 T 检验法对参数 μ 作假设检验. 类似于两总体方差 σ_1^2 与 σ_2^2 为已知的情形, 也可进一步区分为三类不同的假设检验问题. 以下恒设 α 是给定的检验的显著性水平. 我们仅列出各类假设检验问题的否定域. 有兴趣的读者可自行补出需说明的有关细节.

2.1 双侧检验法

假设检验问题的提法完全和 1.1 中所述相同, 只是现在需采用 T 检验法中 (相应于 $\mu = 0$ 时的) 双侧检验法来解, 故类似于 (6.14) 式, 可将否定域取为:

$$C = (x_1, x_2, \dots, x_{n_1}; y_1, y_2, \dots, y_{n_2}) : \left| \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| > t_{\frac{\alpha}{2}}(n_1 + n_2 - 2).$$

2.2 右侧检验法

假设检验问题的提法完全和 1.2 中所述相同, 只是现在需采用 T 检验法中 (相应于 $\mu = 0$ 时的) 右侧检验法来解, 故类似于 (6.15) 式, 可将否定域取为:

$$C = (x_1, x_2, \dots, x_{n_1}; y_1, y_2, \dots, y_{n_2}) : \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha}(n_1 + n_2 - 2).$$

2.3 左侧检验法

假设检验问题的提法完全和 1.3 中所述相同, 只是现在需采用 T 检验法中 (相应于 $\mu = 0$ 时的) 左侧检验法来解, 故类似于 (6.16) 式, 可将否定域取为:

$$C = (x_1, x_2, \dots, x_{n_1}; y_1, y_2, \dots, y_{n_2}) : \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > -t(n_1 + n_2 - 2).$$

例 6.9 某地某年高考后随机抽得 15 名男生、12 名女生的物理考试成绩如下:

男生: 49 48 47 53 51 43 39 57 56 46 42 44 55 44 40

女生: 46 40 47 51 43 36 43 38 48 54 48 34

从这 27 名学生的成绩能说明这个地区男女生的物理考试成绩不相上下吗? (显著性水平 $= 0.05$)

解 若把这地区男生和女生物理考试的成绩分别近似地看作是服从正态分布的随机变量 $X \sim N(\mu, \sigma^2)$ 与 $Y \sim N(\mu, \sigma^2)$, 即可把本例归结为双侧检验法中的假设检验问题. 就本例而言, $n_1 = 15, n_2 = 12$, 从而 $n = n_1 + n_2 = 27$. 再由本例中提供的数据可算出 $\bar{x} = 47.6, \bar{y} = 44$;

$$(n_1 - 1)s_1^2 = \sum_{i=1}^{15} (x_i - \bar{x})^2 = 469.6, (n_2 - 1)s_2^2 = \sum_{i=1}^{12} (y_i - \bar{y})^2 = 412$$

$$s = \frac{1}{n_1 + n_2 - 2} \{ (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 \} = \frac{1}{25} (469.6 + 412) = 5.94.$$

由此可算出:

$$t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{47.6 - 44}{5.94 \sqrt{\frac{1}{15} + \frac{1}{12}}} = 1.566.$$

取显著性水平 $= 0.05$, 查附表 5 知, $t_{\frac{\alpha}{2}}(n - 2) = t_{0.025}(25) = 2.060$. 因 $|1.566| < 2.060 = t_{0.025}(25)$, 从而没有充分理由否认零假设 H_0 (或 H_0^*), 即认为这一地区男女生的物理考试成绩不相上下.

例 6.10 某纺织厂生产纱线. 为提高纱线强力, 想采用新原料. 为确认新原料是否真能提高纱线强度, 先用新原料试生产了一批纱线, 并从中测得了如下 7 个强力数据 (单位为 kg):

1.55, 1.47, 1.52, 1.60, 1.43, 1.53, 1.54.

此外, 又从利用原有材料生产的纱线中测得下述 8 个强力数据:

1.42, 1.49, 1.46, 1.34, 1.38, 1.54, 1.38, 1.51.

现假定按新、旧两种原料生产的纱线强力可分别看作为服从正态分布的总体 $X \sim N(\mu, \sigma^2)$ 与 $Y \sim N(\mu, \sigma^2)$, 这里隐含地假定了工艺的生产精度是一样的. 根据检验目的, 可把本例归结为右侧检验法中的假设检验问题. 就本例而言,

$n_1 = 7, n_2 = 8$, 从而 $n = n_1 + n_2 = 15$, 再由两组样本数据可分别算出: $\bar{x} = 1.52, s_1^2 = 0.0031; \bar{y} = 1.44, s_2^2 = 0.0051$. 由此利用 (6.25) 式可进一步算出 $s = 0.0646$. 这样, 可算出

$$t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{1.52 - 1.44}{0.0646 \sqrt{\frac{1}{7} + \frac{1}{8}}} = 2.395.$$

对于给定的显著性水平 $\alpha = 0.05$, 查附表 5 知, $t_{\alpha/2}(n-2) = t_{0.025}(13) = 1.771$. 因 $t = 2.395 > 1.771 = t_{0.025}(13)$, 故拒绝零假设 H_0 (或 H_0^*), 即认为使用新原料生产的纱线的强力确实增加了.

二、可归结为对于两总体方差比值 r 的假设检验

以上我们讨论了可归结为对于两总体期望差异值 μ 的假设检验问题. 以下考虑可归结为对于两总体方差比值 r 的假设检验问题. 为此需引入一种新的 F 检验法. 注意, 如前所述, 方差比值 $r = \sigma_1^2 / \sigma_2^2$. 当考虑关于单参数 r 的假设检验问题时, 首先要构造一个仅含待检验参数 r 的枢轴量. 这时, 两总体的数学期望 μ 与 μ 是否已知, 会影响到枢轴量的选择. 但由于统计方法的基本思路是一样的, 我们仅介绍两总体数学期望 μ 与 μ 为未知的情形, 而把 μ 与 μ 为已知的情形留给读者自己考虑.

首先, 构造下述枢轴量:

$$F = \frac{1}{r} \frac{S_1^2}{S_2^2} \quad (6.29)$$

其中 S_1^2 与 S_2^2 分别为两独立正态总体 X 与 Y 的样本方差 (见 (6.24) 式). 由第四章 § 4.4 节的定理 4.3 知, 上述枢轴量服从第一自由度为 $n_1 - 1$, 第二自由度为 $n_2 - 1$ 的 F 分布, 简记为 $F \sim F(n_1 - 1, n_2 - 1)$.

以下分别考虑关于单参数 r 的三类假设检验问题.

1. 双侧检验法

双侧检验法适用于下述假设检验问题:

$$H_0: r = 1 \quad H_1: r \neq 1.$$

该假设检验问题等价于下述关于两正态总体方差的假设检验问题:

$$H_0^*: \frac{\sigma_1^2}{\sigma_2^2} = \frac{\sigma_1^2}{\sigma_2^2} \quad H_1^*: \frac{\sigma_1^2}{\sigma_2^2} \neq \frac{\sigma_1^2}{\sigma_2^2}.$$

首先, 因由 (6.29) 式定义的枢轴量 $F \sim F(n_1 - 1, n_2 - 1)$, 故可由 F 分布上侧分位数的定义 (见 § 4.3 节的第三分节) 建立下述关系式:

$$P \left(\frac{1}{r} \frac{S_1^2}{S_2^2} < F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) \text{ 或 } \frac{1}{r} \frac{S_1^2}{S_2^2} > F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) \right) = \alpha, \quad (6.30)$$

其中 $F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1)$ 与 $F_{\frac{\alpha}{2}}(n_1-1, n_2-1)$ 分别为水平 $1-\frac{\alpha}{2}$ 与 $\frac{\alpha}{2}$ 的自由度为 n_1-1 与 n_2-1 的 F 分布的上侧分位数. 于是, 可将否定域取为

$$C = \{(x, y) : \frac{S_1^2}{S_2^2} < F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1) \text{ 或 } \frac{S_1^2}{S_2^2} > F_{\frac{\alpha}{2}}(n_1-1, n_2-1)\}. \quad (6.31)$$

以上简记 $x = (x_1, x_2, \dots, x_{n_1}), y = (y_1, y_2, \dots, y_{n_2})$. 这是因为

$$\begin{aligned} & P\{(X_1, X_2, \dots, X_{n_1}; Y_1, Y_2, \dots, Y_{n_2}) \in C | H_0\} \\ &= P\left\{\frac{S_1^2}{S_2^2} < F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1) \text{ 或 } \frac{S_1^2}{S_2^2} > F_{\frac{\alpha}{2}}(n_1-1, n_2-1) | H_0\right\} \quad (\text{由(6.31)式}) \\ &= \end{aligned}$$

2. 右侧检验法

右侧检验法适用于下述假设检验问题:

$$H_0: r = 1 \quad H_1: r > 1.$$

该假设检验问题等价于关于两正态总体方差的假设检验问题:

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = \frac{\sigma_1^2}{\sigma_2^2} \quad H_1: \frac{\sigma_1^2}{\sigma_2^2} > \frac{\sigma_1^2}{\sigma_2^2}.$$

不同于双侧检验法, 上述零假设 H_0 是一复合假设. 这样, 首先由 F 分布上侧分位数的定义建立下述关系式:

$$P\left\{\frac{1}{r} \frac{S_1^2}{S_2^2} > F(n_1-1, n_2-1)\right\} = \quad (6.32)$$

于是可取否定域为:

$$C = \{(x_1, x_2, \dots, x_{n_1}; y_1, y_2, y_{n_2}) : \frac{S_1^2}{S_2^2} > F(n_1-1, n_2-1)\} \quad (6.33)$$

这是因为在 $H_0: r = 1$ 成立时

$$\frac{1}{r} \frac{S_1^2}{S_2^2} = \frac{S_1^2}{S_2^2},$$

故

$$\frac{S_1^2}{S_2^2} > F(n_1-1, n_2-1) \quad \frac{1}{r} \frac{S_1^2}{S_2^2} > F(n_1-1, n_2-1), \quad (6.34)$$

从而有

$$\begin{aligned} & P\{(X_1, X_2, \dots, X_{n_1}; Y_1, Y_2, \dots, Y_{n_2}) \in C | H_0\} \\ &= P\left\{\frac{S_1^2}{S_2^2} > F(n_1-1, n_2-1) | H_0\right\} \\ &= P\left\{\frac{1}{r} \frac{S_1^2}{S_2^2} > F(n_1-1, n_2-1) | H_0\right\} \\ &= \end{aligned}$$

3. 左侧检验法

左侧检验法适用于下述假设检验问题:

$$H_0: r = 1 \quad H_1: r < 1.$$

该假设检验问题等价于下述关于两正态总体方差的假设检验问题:

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_1: \frac{\sigma_1^2}{\sigma_2^2} < 1.$$

类似于以上关于右侧检验法的论证, 否定域可取为

$$C = (x_1, x_2, \dots, x_{n_1}; y_1, y_2, \dots, y_{n_2}): \frac{S_1^2}{S_2^2} < F_{1-\alpha}(n_1-1, n_2-1)$$

其中 $F_{1-\alpha}(n_1-1, n_2-1)$ 为水平 $1-\alpha$ 的自由度为 (n_1-1) 与 (n_2-1) 的 F 分布的上侧分位数.

今后, 在考虑假设检验问题时, 当最终是借助枢轴量服从或渐近地服从 F 分布作为出发点的, 便称相应的检验方法为 F 检验法. 显然, 上述关于两总体方差比值 r 的三类假设检验问题皆采用了 F 检验法.

例 6.11 (续例 6.10) 在例 6.10 中, 曾隐含地假定了总体 X 与 Y 有相同的方差. 如果对这一点持有怀疑, 即问: 有无可能是 $X \sim N(\mu, \sigma_1^2)$, $Y \sim N(\mu, \sigma_2^2)$ 但 $\sigma_1^2 \neq \sigma_2^2$ 呢? 为打消这一疑虑, 可提出双侧检验法中的假设检验问题. 利用例 6.10 中已算出的数据, 可计算:

$$f = \frac{S_1^2}{S_2^2} = \frac{0.0031}{0.0051} = 0.6078.$$

就本例而言, $n_1 = 7$, $n_2 = 8$. 现取检验的显著性水平 $\alpha = 0.05$. 查附表 4 知, $F_{\frac{\alpha}{2}}(n_1-1, n_2-1) = F_{0.025}(6, 7) = 5.12$,

$$F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1) = \frac{1}{F_{\frac{\alpha}{2}}(n_2-1, n_1-1)} = \frac{1}{F_{0.025}(7, 6)} = \frac{1}{5.70} = 0.1754$$

由于

$$F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1) = 0.1754 < f = 0.6078 < F_{\frac{\alpha}{2}}(n_1-1, n_2-1) = 5.12,$$

故没有充分理由拒绝零假设 H_0 (或 H_0^*), 从而认为两总体的方差是相同的.

例 6.12 有两台机床生产同一型号的滚珠. 根据已有经验知, 这两台机床生产的滚珠直径都服从正态分布. 现分别从这两台机床生产的滚珠中抽取 7 个和 9 个滚珠, 并测得它们的直径如下所示 (单位为毫米):

机床甲: 15.2, 14.5, 15.5, 14.8, 15.1, 15.6, 14.7.

机床乙: 15.2, 15.0, 14.8, 15.2, 15.0, 14.9, 15.1, 14.8, 15.3.

试问: 机床乙生产的滚珠直径的方差是否比机床甲生产的滚珠直径的方差小? (检验的显著性水平 $\alpha = 0.05$).

解 以 X 和 Y 分别表示机床甲与机床乙所生产的滚珠直径, 已知 $X \sim N(\mu, \sigma_1^2)$, $Y \sim N(\mu, \sigma_2^2)$. 就题意可把本例归结为右侧检验法中的假设检验问题.

现有 $n_1=7, n_2=9$, 另由两组样本数据算出: $\bar{x}=15.057, s_1^2=0.1745; \bar{y}=15.033, s_2^2=0.0438$

于是

$$f = \frac{s_1^2}{s_2^2} = \frac{0.1745}{0.0438} = 3.984.$$

对于给定的显著性水平 $\alpha=0.05$, 查附表 4 知 $F(\alpha, n_1-1, n_2-1) = F_{0.05}(6, 8) = 3.58$. 这样, 便有

$$f = 3.984 > 3.58 = F_{0.05}(6, 8).$$

故否定零假设 H_0 (或 H_0^*), 即认为机床乙生产的滚珠直径的方差明显的比机床甲生产的滚珠直径的方差小.

至此, 我们较为系统地介绍了双正态总体的参数假设检验. 现把已提及的所有检验问题的要点总结于表 6-2 中供读者参考. 有兴趣的读者可考虑一下, 当总体 X 与 Y 的数学期望 μ 与 μ 已知时, 可选择怎样一个略为简便些的枢轴量来替换由 (6.29) 式定义的枢轴量 F ? 此外, 可对给定的显著性水平 α , 试着写出三类假设检验问题的否定域, 并把它们补充到表 6-2 中(见习题六(B)的第 2 题).

表 6-2 双正态总体的假设检验一览表

已知条件	零假设与备择假设	统计量	应查分布表	否定域
$X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2), \mu_1$ 与 μ_2 未知, σ_1^2 与 σ_2^2 已知.	$H_0: \mu = \mu$ $H_1: \mu \neq \mu$	$U = \frac{\bar{X} - \bar{Y}}{0},$ 其中 $0 = \frac{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}{}$	标准正态分布表	
	$H_0: \mu \leq \mu$ $H_1: \mu > \mu$			
	$H_0: \mu \geq \mu$ $H_1: \mu < \mu$			

续表

已知条件	零假设与备择假设	统计量	应查分布表	否定域
$X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$, $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ 皆未知.	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$T = \frac{\bar{X} - \bar{Y}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ 其中 $S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$	$t(n_1 + n_2 - 2)$ 分布	
	$H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$			
	$H_0: \mu_1 \geq \mu_2$ $H_1: \mu_1 < \mu_2$			
$X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$, $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ 皆未知.	$H_0: \frac{\sigma_1^2}{\sigma_2^2} = \frac{\sigma_1^2}{\sigma_2^2}$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq \frac{\sigma_1^2}{\sigma_2^2}$	$F = \frac{S_1^2}{S_2^2}$	$F(n_1 - 1, n_2 - 1)$ 分布	
	$H_0: \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2}$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} > \frac{\sigma_1^2}{\sigma_2^2}$			
	$H_0: \frac{\sigma_1^2}{\sigma_2^2} \geq \frac{\sigma_1^2}{\sigma_2^2}$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} < \frac{\sigma_1^2}{\sigma_2^2}$			

§ 6.4 关于一般总体数学期望的假设检验

前两节详尽地讨论了正态总体的参数假设检验问题，由于可以精确地求出相应枢轴量所服从的分布，故属小样本统计范畴。本节首先讨论贝努里总体的参数假设检验。然后，也将讨论关于一般总体数学期望的假设检验。前者存在小样本统计推断方法。此外，关于这两类假设检验问题还都可借助一些统计量（或枢轴量）的极限分布近似地进行假设检验，这属大样本统计范畴。

一、伯努利总体的参数假设检验

考虑伯努利总体的参数假设检验，一方面有其实际应用背景，这早已在第四章§ 4.1 节的例 4.2 中指出，因为伯努利分布中的参数 p 可视为有些实际总体中具有某一统计特征的个体所占的比率。另一方面，在理论上也有一定意义。这是因为下一节介绍的关于有限离散型总体的假设检验问题恰可视为关于伯努利总体的参数检验问题的推广。这样，如熟悉伯努利总体参数检验的大样本统计推断方法，还将有助于理解下一节介绍的检验有限离散型总体的多项分布²检验法。以下为简化问题的叙述，在假设检验问题中将不再列出备样假设。

现设总体 X 服从以 p ($0 < p < 1$) 为参数的伯努利分布，即有：

$$P(X=1)=p=1-P(X=0).$$

又设 (X_1, X_2, \dots, X_n) 是总体 X 的容量为 n 的一个样本，记

$$N = X_1 + X_2 + \dots + X_n \quad (6.35)$$

显然， N 恰等于样本中取值为 1 的诸分量的个数。不难证明统计量 N 服从以 n 与 p 为参数的二项分布，即有

$$P(N=k) = C_n^k p^k (1-p)^{n-k} = b(k; n, p), \quad k=0, 1, 2, \dots, n. \quad (6.36)$$

考虑关于参数 p 的假设检验问题的途径有两个，以下分别细述之。在下面的讨论中，恒设 p_0 是一指定正数 ($0 < p_0 < 1$)。

1. 小样本途径

对于参数 p 也可提出三类假设检验问题。

1.1 双侧检验法— $H_0: p = p_0$

在由 (6.36) 式确定的二项分布中，仅含一个未知参数 p ，一旦确定了参数 p 的值，二项分布也就完全确定了。如零假设 H_0 成立，则统计量 N 便服从以 n 与 p_0 为参数的二项分布。这是一个完全确定的分布。这样，对于给定的显著性水平 α ，我们不难确定两个正整数 k_1 与 k_2 ，它们分别满足：

$$k_1 = \max_{j=0}^j: \sum_{i=0}^j b(i; n, p_0) \leq \frac{\alpha}{2} \quad (6.37)$$

与

$$k_2 = \min_{i=j}^n j: \sum_{i=j}^n b(i; n, p_0) \geq \frac{\alpha}{2}. \quad (6.38)$$

然后, 可将否定域取为

$$C = \{(X_1, X_2, \dots, X_n): (X_1 + \dots + X_n) \geq k_1 \text{ 或 } (X_1 + \dots + X_n) \leq k_2\} \quad (6.39)$$

这是因为

$$\begin{aligned} & P\{(X_1, X_2, \dots, X_n) \in C | H_0\} \\ &= P\{N \geq k_1 \text{ 或 } N \leq k_2 | H_0\} \\ &= \sum_{i=0}^{k_1} b(i; n, p_0) + \sum_{i=k_2}^n b(i; n, p_0) \\ &= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha \end{aligned}$$

以上我们较为详尽地讨论了关于参数 p 的双侧检验法. 类似地, 也可考虑关于参数 p 的右侧检验法与左侧检验法. 以下对于给定的显著性水平 α 仅列出检验的否定域. 有兴趣的读者可自行补出需说明的有关细节.

1.2 右侧检验法— $H_0: p = p_0$

对于上述零假设 H_0 , 可将否定域取为

$$C^* = \{(X_1, X_2, \dots, X_n): X_1 + X_2 + \dots + X_n \geq k^*\},$$

其中正整数 k^* 由下式确定:

$$k^* = \min_{i=j}^n j: \sum_{i=j}^n b(i; n, p_0) \geq \alpha. \quad (6.40)$$

1.3 左侧检验法— $H_0: p = p_0$

对于上述零假设 H_0 , 可将否定域取为

$$C^* = \{(X_1, X_2, \dots, X_n): X_1 + X_2 + \dots + X_n \leq k^*\},$$

其中正整数 k^* 由下式确定:

$$k^* = \max_{i=0}^j j: \sum_{i=0}^j b(i; n, p_0) \geq \alpha. \quad (6.41)$$

例 6.13 设已知某产品的废品率不超过 $p_0 = 0.07$. 现进行 20 次还原抽样, 试确定零假设 $H_0: p = p_0$ 的显著性水平为 $\alpha = 0.05$ 的否定域.

解 进行 20 次还原抽样, 可认为得到一个容量为 $n = 20$ 的样本. 本例属右侧检验法.

$$\text{因 } \sum_{i=4}^{20} b(i; 20, 0.07) = 0.0471 < 0.05,$$

即知由 (6.40) 式确定的临界值 $k^* = 4$, 从而否定域为

$C^* = \{(X_1, X_2, \dots, X_n): X_1 + X_2 + \dots + X_n \geq 4\}$. 这表明在 20 次还原抽样中, 若废品出现次数大于或等于 4, 即认为废品率超过 0.07.

基于二项分布计算上的复杂性,借助(6.36)——(6.38)、(6.40)与(6.41)诸式来确定相应否定域中诸临界值 k_1, k_2, k^* 与 k^* 并不是很方便的. 通常需查阅二项分布的累计分布表. 为了克服计算上的复杂性与查表时的不方便, 对于伯努利总体的参数假设检验, 更常采用大样本的统计推断方法.

2. 大样本途径

当总体 X 服从以 $p(0 < p < 1)$ 为参数的伯努利分布时, 不难算出 $E[X] = p$, $D[X] = p(1 - p)$. 现设 (X_1, X_2, \dots, X_n) 是总体 X 的一个容量为 n 的样本, 并记

$$U_n = \frac{\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}}{\sqrt{\frac{p(1-p)}{np(1-p)}}} = \frac{N - np}{np(1-p)}$$

(6.42)

其中 $N = X_1 + X_2 + \dots + X_n$, \bar{X} 为样本均值.

由第四章§ 4.4 节的定理 4.4 知, 当样本容量 n 充分大时, 枢轴量 U_n 渐近地服从标准正态分布, 因此便可利用§ 6.2 节中所述的 U 检验法近似地对参数 p 作假设检验. 鉴于在实际应用中对比率 p 作参数假设检验时, 一般说来, 所取的样本容量都很大, 故上述检验法是可行的. 现把相应的检验法则列于表 6-3 中, 其中 p_0 为指定的正数, α 为给定的显著性水平, $u_{\frac{\alpha}{2}}$ 与 u 分别为标准正态分布的水平 $\frac{\alpha}{2}$ 的双侧分位数与上侧分位数.

表 6-3 伯努利总体参数 p 的近似 U 检验法

H_0 与 H_1	统 计 量	检验法则
$H_0: p = p_0$ $H_1: p \neq p_0$	$U_n = \frac{N - np_0}{\sqrt{np_0(1 - p_0)}}$	$ U_n > u_{\frac{\alpha}{2}}$, 否定 H_0 ; $ U_n \leq u_{\frac{\alpha}{2}}$, 接受 H_0 .
$H_0: p = p_0$ $H_1: p > p_0$	同上	$U_n > u$, 否定 H_0 ; $U_n \leq u$, 接受 H_0 .
$H_0: p = p_0$ $H_1: p < p_0$	同上	$U_n < -u$, 否定 H_0 ; $U_n \geq -u$, 接受 H_0 .

注意, 上表中由于以指定的正数 p_0 替换了由 (6.42) 式给出的枢轴量 U_n 中的未知参数 p , 从而得到了一统计量, 故以 U_n 记之, 以示区别.

例 6.14 某市对成年人查体, 随机抽取 100 人的样本, 发现有 59 人患有不同程度的牙疾, 问以 0.05 的显著性水平是否说明该市 50% 以上的人患有牙疾.

解 待检验的假设为 $H_0: p = 0.5$. 样本容量 $n = 100$. 本例可采用近似 U 检验法的右侧检验法来解. 鉴于统计量 U_n 的观测值为

$$u = \frac{59 - 50}{\sqrt{100 \times 0.5 \times 0.5}} = 1.8 > u_{0.05} = 1.64.$$

故否定零假设 H_0 ，表明该市有 50% 以上的成年人患有不同程度的牙疾。

二、一般总体数学期望的大样本假设检验

我们在§ 6.2 节讨论了关于正态总体数学期望的假设检验；在上一分节中讨论了关于伯努利总体参数 p 的假设检验，基于此参数 p 恰为伯努利总体的数学期望，因此，迄今为止，我们讨论了两种特殊总体的数学期望的假设检验。本分节将讨论关于一般总体数学期望的假设检验。确切地说，设 X 为任一总体，记其数学期望为 $E[X] = \mu$ 现需对总体 X 的数学期望 μ 作假设检验。由于对总体 X 无任何其他知识，故一般说来，假定总体的方差 $D[X] = \sigma^2$ 也是未知的。现设 (X_1, X_2, \dots, X_n) 是总体 X 的容量为 n 的一个样本，令：

$$T_n = \frac{\bar{X} - \mu}{S / \sqrt{n}}, \tag{6.43}$$

其中 \bar{X} 与 S 分别为上述样本的样本均值与样本标准差。由第四章§ 4.4 节的定理 4.4 知，当样本容量 n 充分大时， T_n 渐近地服从标准正态分布，尽管 T_n 不是严格意义上的枢轴量，但由于已知其渐近分布，我们仍将它当作枢轴量来对待，从而也可利用 U 检验法，借助上述枢轴量 T_n 对总体的未知数学期望 μ 作假设检验。现把相应的假设检验问题与检验法则列于表 6-4 中，其中 μ_0 为指定的常数， α 为给定的显著性水平， $u_{\frac{\alpha}{2}}$ 与 $u_{1-\frac{\alpha}{2}}$ 分别为水平 α 的标准正态分布的双侧分位数与上侧分位数。

表 6-4 一般总体数学期望 μ 的近似 U 检验法

H_0 与 H_1	统 计 量	检验法则
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$T_n = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$	$ T_n > u_{\frac{\alpha}{2}}$ ，否定 H_0 ； $ T_n \leq u_{\frac{\alpha}{2}}$ ，接受 H_0 ；
$H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$	同上	$T_n > u_{1-\alpha}$ ，否定 H_0 ； $T_n \leq u_{1-\alpha}$ ，接受 H_0 ；
$H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$	同上	$T_n < -u_{1-\alpha}$ ，否定 H_0 ； $T_n \geq -u_{1-\alpha}$ ，接受 H_0 。

注意，上表中由于以指定常数 μ_0 替换了由 (6.43) 式给出的枢轴量 T_n 中的

未知参数 μ 从而得一统计量, 故以 T_n 记之, 以示区别.

例 6.15 在可靠性理论与应用中, 常根据设备或部件不同的失效性质, 以指数分布, 韦布尔 (Weibull) 分布, 伽马分布, 对数正态分布等多种寿命分布类来描述设备或部件的使用寿命. 某厂新研究并开发了某类设备所需的关键部件. 由于尚缺乏足够的经验数据, 还无法判定此部件的使用寿命所服从的分布类型. 现通过加速失效试验法, 测得了 100 个新生产部件的使用寿命, 并算出了它们样本均值的观测值为 $\bar{x} = 17.84$ (kh), 样本标准差的观测值为 $s = 1.25$ (kh), 试问: 由这些数据能否判定此部件的连续使用寿命至少为 2 年? (给定显著性水平 $\alpha = 0.01$)

解 以每年 365 天计算, 一部件若可连续使用二年, 则使用的小时数至少应为 $2 \times 365 \times 24 = 17520$, 折合为 17.52kh. 为此可考虑下述假设检验问题:

$$H_0: \mu \leq 17.52 \quad H_1: \mu > 17.52$$

这可利用近似 U 检验法的右侧检验来解. 本例中 $\mu_0 = 17.52$, $n = 100$. 由测出的数据可计算

$$t_n = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{17.84 - 17.52}{1.25/\sqrt{100}} = 2.56$$

对于给定的显著性水平 $\alpha = 0.01$, 查附表 2 知, $u_{\alpha} = u_{0.01} = 2.33$. 由 $t_n = 2.56 > u_{0.01} = 2.33$, 故否定零假设 H_0 , 即认为部件至少可以连续使用两年.

* § 6.5 拟合优度 χ^2 检验法

本章前四节集中讨论了参数假设检验, 而且限于讨论单参数假设检验. 本节首先讨论了多项分布的 χ^2 检验法. 这虽仍属参数假设检验, 但已是多参数假设检验. 然后再借助多项分布的 χ^2 检验法, 讨论了两类非参数假设检验: 分布的拟合优度检验与独立性检验.

一、多项分布的 χ^2 检验法

多项分布的 χ^2 检验法可视为关于伯努利总体的二项分布检验法的推广, 它是用以检验仅取有限个值的离散型总体的. 现设总体 X 是取 r 个可能值的离散型随机变量. 不失一般性, 设总体 X 的 r 个可能值为 $1, 2, \dots, r$, 再记

$$p_i = P\{X = i\}, \quad i = 1, 2, \dots, r. \quad (6.44)$$

显然, $p_1 + p_2 + \dots + p_r = 1$, 现考虑下述假设检验问题:

$$H_0: p_i = p_{i,0}, \quad i = 1, 2, \dots, r. \quad (6.45)$$

其中 $p_{i,0}$, $i = 1, 2, \dots, r$, 是 r 个指定的正数, 且满足 $p_{1,0} + p_{2,0} + \dots + p_{r,0} = 1$.

显然, 当 $r=2$ 时, 上述假设检验问题即为前一节第一分节中已讨论过的关于伯努利总体的参数假设检验. 自然, 那里假定总体 X 的两个可能值为 0 与 1, 而非 1 与 2, 不过, 这不是本质的. 当 $r=3$ 时, 上述假设检验虽仍属参数假设检验, 但因其零假设 H_0 是针对多个参数提出的, 故已为多参数假设检验问题. 同样, 为简化问题的叙述, 本节也不列出备择假设.

现设 (X_1, X_2, \dots, X_n) 是上述总体 X 的一个容量为 n 的样本. 如 § 6.1 节第五分节中已指出的那样, 对于多参数假设检验问题, 关键是寻求一个包含所有待检验参数的枢轴量, 并使之服从或渐近地服从一个不依赖所有待检验参数的一个已知的确定分布. 类似于伯努利总体的参数假设检验, 先以 N_i 表示该样本中取值为 i 的诸分量的个数, 并称其为可能值 i 的频数. 显然每一频数 N_i 皆可视为是样本的统计量, 且满足:

$$N_1 + N_2 + \dots + N_r = n. \quad (6.46)$$

不难证明, 由诸频数组成的随机向量 (N_1, N_2, \dots, N_r) 服从以 n, p_1, p_2, \dots, p_r 为参数的多项分布, 即有

$$P\{N_1 = n_1, N_2 = n_2, \dots, N_r = n_r\} = \frac{n!}{n_1! n_2! \dots n_r!} p_1^{n_1} p_2^{n_2} \dots p_r^{n_r} \quad (6.47)$$

其中 $n_i, i=1, 2, \dots, r$, 为非负整数, 且满足 $n_1 + n_2 + \dots + n_r = n$.

上述多项分布中含 r 个参数: p_1, p_2, \dots, p_r . 这样, 前面关于有限离散型总体提出的零假设 H_0 (见 (6.45)) 也可视为是对上述多项分布中所含的 r 个未知参数提出的零假设. 特别是, 当 $r=2$ 时, 上述多项分布即为二项分布. 这表明, 关于伯努利总体的参数假设检验问题实际上也是关于二项分布的参数假设检验. 同时也说明了, 本分节讨论的问题事实上可视为 § 6.4 节第一分节中关于伯努利总体参数假设检验问题的推广. 既然对于伯努利总体来说, 我们尚认为采用大样本统计推断方法更为合适; 那么, 处理关于有限离散型总体的多参数假设检验时, 自然更倾向于采用大样本的统计推断方法. 采用这一途径的直观思路如下所述.

首先, 不难看出, 当样本容量 n 充分大时对每一频数 N_i ($i=1, 2, \dots, r$) 而言, 下述枢轴量:

$$\frac{(N_i - np_i)^2}{np_i(1-p_i)}, i=1, 2, \dots, r$$

皆渐近地服从自由度为 1 的 χ^2 分布.

这样, 以此作为切入点, 皮尔逊 (K. Pearson) 将上述诸量略作修改后再相加, 引入了下述样本函数:

$$\chi^2 = \sum_{i=1}^r \frac{(N_i - np_i)^2}{np_i}. \quad (6.48)$$

严格地说,上述样本函数由于包含了所有待检验的 r 个未知参数,从而不是严格意义下的统计量;确切地说,应称其为枢轴量.但统计学文献现已习惯地称其为 χ^2 统计量,或皮尔逊统计量,故本书以下也沿用这一约定俗成的称呼.强调一下,由 (6.48) 式定义的样本函数自身并不服从 χ^2 分布,但皮尔逊证明了它的极限分布为 χ^2 分布.这也是称其为 χ^2 统计量的缘由.

定理 6.1 (K. Pearson) 当样本容量 n 趋于无穷时,由 (6.48) 式定义的 χ^2 统计量的极限分布是自由度为 $r-1$ 的 χ^2 分布.

上述定理的证明已超出本书范围,故略之.有兴趣的读者可参阅文献 [3], p. 113—114 或 [4], p. 69—71. 定理前的一段陈述,可视为对这一定理的直观解释,对自由度则可作如下说明:

由 (6.46) 式可推知, χ^2 统计量中的 r 个基本变量:

$$N_1 - np_1, N_2 - np_2, \dots, N_r - np_r$$

满足一个制约关系式,即有

$$\sum_{i=1}^r (N_i - np_i) = \sum_{i=1}^r N_i - n \sum_{i=1}^r p_i = n - n = 0.$$

于是,极限分布的自由度便为 $r-1$.

定理 6.1 是检验 (6.45) 中的零假设 H_0 的理论基础.具体的检验步骤如下所述:

(1) 构造渐近服从已知的确定分布的统计量

首先注意到,定理 6.1 的结论并不依赖了 χ^2 统计量中的诸参数 p_1, p_2, \dots, p_r 是否已知.在“零假设 H_0 成立”的前提下, χ^2 统计量中的诸参数 p_i 便可分别以指定的正数 $p_{i,0}$ 替换,从而成为真正意义下的统计量.我们以 χ^2_0 记之,以示区别:

$$\chi^2_0 = \sum_{i=1}^r \frac{(N_i - np_{i,0})^2}{np_{i,0}}. \quad (6.49)$$

这样,当样本容量 n 充分大时,在“零假设 H_0 成立”的前提下,由定理 6.1 即知统计量 χ^2_0 . 渐近地服从自由度为 $r-1$ 的 χ^2 分布,因此便可利用 χ^2 检验法来检验零假设 H_0 .

(2) 利用 χ^2 分布的上侧分位数确定检验的否定域

由于 χ^2 统计量是用以刻画实际频数 N_i 与理论频数 np_i 的差异的,故在检验零假设 H_0 时,更侧重于注意有无大的差异发生.为此可采用 χ^2 检验法的右侧检验法来解.首先,对于给定的显著性水平 α ,当样本容量 n 充分大时,由 χ^2 分布上侧分位数的定义,可建立如下近似关系式:

$$P\{\chi^2_0 > \chi^2_{(r-1)}(\alpha) | H_0\} \quad (6.50)$$

其中 $\chi^2_{(r-1)}$ 为水平 α 的自由度为 $r-1$ 的 χ^2 分布的上侧分位数.

现以 n_1, n_2, \dots, n_r 分别表示诸频数统计量 N_1, N_2, \dots, N_r 的观测值, 并仍以 χ_0^2 表示由 (6.49) 式定义的统计量之观测值, 便可将否定域取为:

$$C = \{(n_1, n_2, \dots, n_r) : \chi_0^2 > \chi^2(r-1)\}. \quad (6.51)$$

这样就有

$$P\{(N_1, N_2, \dots, N_r) \in C | H_0\} = P(\chi_0^2 > \chi^2(r-1) | H_0)$$

这表明下述事件:

$$A = \{(N_1, N_2, \dots, N_r) \in C\}$$

在“零假设 H_0 成立”的前提下发生的概率近似地等于检验的显著性水平, 从而是一小概率事件.

(3) 检验法则

由上所述, 一旦诸频数统计量的观测值 (n_1, n_2, \dots, n_r) 果真落入否定域 C (即如有 $\chi_0^2 > \chi^2(r-1)$), 且假定零假设 H_0 (见 (6.45) 式) 成立, 便表明有一小概率事件在一次试验中就发生了. 这有悖于小概率原理, 从而就有充分的理由否定零假设 H_0 ; 否则, 便接受零假设 H_0 .

以下我们把上述检验法称为关于多项分布的 χ^2 检验法.

例 6.16 一家工厂分早中晚三班, 每班 8 小时. 近期发生了一些事故. 在近期记录的 15 次事故中, 有 6 次发生在早班, 3 次发生在中班, 6 次发生在晚班, 从而怀疑班次不同与事故发生率有关. 试利用上述记录数据来判断这一猜测是否成立. (显著性水平 $\alpha = 0.05$)

解 如下定义一离散型随机变量 X : 若事故在早班发生, 令 $X=1$; 若事故在中班或晚班发生, 分别令 $X=2$ 或 $X=3$. 再记 $p_i \stackrel{\text{def}}{=} P(X=i)$, $i=1, 2, 3$. 显然, “事故的发生与班次无关”等介于“事故发生在早、中或晚班的可能性都是一样的”. 为此可检验下述零假设:

$$H_0: p_i = \frac{1}{3}, \quad i=1, 2, 3.$$

这可利用多项分布的 χ^2 检验法来解. 本例中 $r=3$; $p_{i,0} = \frac{1}{3}$, $i=1, 2, 3$; $n_1=6$, $n_2=3$, $n_3=6$, $n=15$. 由此即可算出

$$\chi_0^2 = \sum_{i=1}^3 \frac{(n_i - np_{i,0})^2}{np_{i,0}} = \frac{1}{5} [(6-5)^2 + (3-5)^2 + (6-5)^2] = 1.2.$$

对于给定的显著性水平 $\alpha = 0.05$, 查附表 3 知 $\chi^2(r-1) = \chi^2_{0.05}(2) = 5.991$. 由于 $\chi_0^2 = 1.2 < \chi^2(r-1) = 5.991$, 便知无充分理由拒绝零假设, 从而认为事故的发生与班次无关.

从上例中已记录的事故数据看, 6 : 3 : 6 的比例似乎应当说明中班事故率要低于平均值 $1/3$. 但上述分析表明, 有时直观的判断未必可靠.

二、一般总体分布拟合优度的 χ^2 检验

本分节将对总体的分布作假设检验. 这里限定总体是连续型随机变量, 或是取无限个可能值的离散型随机变量. 这是因为, 取有限个可能值的离散型总体的假设检验已可由多项分布的 χ^2 检验法解决. 这样, 上述假设检验问题便是非参数假设检验. 解决该问题的方法是, 借助多项分布的 χ^2 检验法, 将这一非参数假设检验问题转化为一个多参数的假设检验问题. 以下分理论分布完全已知和含未知参数两种情形讨论.

1. 理论分布完全已知的情形

设总体 X 的分布函数为 $F(x)$. (X_1, X_2, \dots, X_n) 是总体 X 的一个容量为 n 的样本. 再设 $F_0(x)$ 为一不含未知参数的已知的分布函数. 现考虑下述零假设:

$$H_0: F(x) = F_0(x). \quad (6.52)$$

处理上述假设检验问题的一个途径是将总体 X 有限地离散化, 从而归为上一分节中已解决的问题. 具体步骤如下所述.

首先, 任取 $r-1$ 个实数: $a_1 < a_2 < \dots < a_{r-1}$, 并依序将实轴 $(-\infty, +\infty)$ 划分为 r 个互不相交的子区间:

$$I_1 = (-\infty, a_1], I_2 = (a_1, a_2], \dots, I_r = (a_{r-1}, +\infty).$$

其次, 定义一个仅取 r 个可能值的离散型随机变量 Y :

$$Y = i \quad X \in I_i, \quad i = 1, 2, \dots, r.$$

由此即可借助总体 X 的分布函数 $F(x)$ 来描述离散型随机变量 Y 的概率分布列, 确切地说, 记

$$\begin{aligned} p_1 &= P(Y=1) = P(X \in I_1) = F(a_1), \\ p_i &= P(Y=i) = P(X \in I_i) = F(a_i) - F(a_{i-1}), \quad i=2, \dots, r-1 \\ p_r &= P(Y=r) = P(X \in I_r) = 1 - F(a_{r-1}). \end{aligned} \quad (6.53)$$

最后, 利用给定的已知分布函数 $F_0(x)$, 类似(6.53)式确定 r 个正数 $p_{i,0}$, $i=1, 2, \dots, r$, 并由零假 H_0 过渡到对于有限离散型总体 Y 的零假设. 确切地如下所述:

$$\begin{aligned} p_{1,0} &= F_0(a_1), \\ \text{记 } p_{i,0} &= F_0(a_i) - F_0(a_{i-1}), \quad i=2, \dots, r-1 \\ p_{r,0} &= 1 - F_0(a_{r-1}). \end{aligned} \quad (6.54)$$

显然, 由零假设 H_0 (见 6.52) 与 (6.53) 及 (6.54) 两式, 便可导出关于有限离散型总体 Y 的零假设:

$$H_0^*: p_i = p_{i,0}, \quad i=1, 2, \dots, r. \quad (6.55)$$

这样,就把原来的假设检验问题转化成一个有限离散型总体 Y 的假设检验问题,从而便可利用多项分布的 χ^2 检验法来解. 如前所述,用以检验的统计量可取为

$$\chi^2_0 = \sum_{i=1}^r \frac{(N_i - np_{i,0})^2}{np_{i,0}} \quad (6.56)$$

其中 N_i 表示样本中诸分量落入子区间 I_i 内的频数, r 个常数 $p_{i,0}$ ($i=1, \dots, r$) 则由(6.54)确定.

上述统计量 χ^2_0 可视为以一个已知的有限离散型总体去拟合一个未知的连续型总体(或取无限个可能值的离散型总体)而产生的差异的某种度量,再鉴于上述统计量的极限分布又是 χ^2 分布,故称上述检验法为总体分布拟合优度的 χ^2 检验法.

必须指出原有的零假设 H_0 与经过有限离散化处理后得到的零假设 H_0^* 并不等价. 零假设 H_0^* 仅和下述零假设 H_0 等价:

$$H_0: F(a_i) = F_0(a_i), i=1, 2, \dots, r$$

显然,上述零假设 H_0 是一个较原有零假设 H_0 (见(6.52))弱的假设. 因为若 H_0 成立, H_0 自然成立;反之,则不真. 这样,当采取多项分布的 χ^2 检验法而否定零假设 H_0 ,进而否定原来的零假设 H_0 时,理由是充分的;反之,如接受零假设 H_0 ,则不能保证原来的零假设 H_0 一定成立,于是由此作出接受 H_0 的判断就显得有些牵强. 自然,也有更为严谨的检验法,如柯尔莫哥洛夫() 检验法可以克服这一缺陷. 但由于其深度已超出本书范围,故不予介绍. 有兴趣的读者可参阅文献 [3], p114—116 或 [5], p221—224.

此外,上述总体分布拟合优度的 χ^2 检验法和 r 的选择与诸分点 a_1, a_2, \dots, a_{r-1} 的取法有关. 有一种经验法则认为,应保证由(6.54)式确定的诸值 $p_{i,0}$ 满足 $np_{i,0} \geq 5$ ($i=1, 2, \dots, r$).

另需指出的是,如总体 X 是取无限个可能值的离散型总体,相应的操作步骤会更简单些. 这时由于总体已是一离散型总体,故只需将与尾概率值相对应的无限多个可能值合并成一组即可,这可由下例解释之.

例 6.17 某黑盒中有偶数个球,它们除颜色外,其它方面完全一样,其中一种球着黑色,另一种球着白色. 现作如下试验:以有放回抽取方式从此黑盒中摸球,直到摸取的是白球为止,并记录下抽取的次数. 重复执行上述试验 100 次,并将试验结果汇总成下表,此表中将首次摸到白球时抽取次数大于或等于 5 的频数合并成一组:

首次摸到白球的抽取次数的频数

抽取次数	1	2	3	4	5	总计
频数	43	31	15	6	5	100

试由上表中数据判断黑盒中白球与黑球的个数是否相等（显著性水平 $\alpha = 0.05$ ）.

解 假设黑盒中白球所占比例为 p . 以 X 表示有放回抽取时, 首次摸到白球的抽取次数, 显然, X 服从以 p 为参数的几何分布, 即有

$$p_i = P\{X = i\} = p(1 - p)^{i-1}, i = 1, 2, \dots$$

黑盒中白球与黑球个数相等, 当且仅当 $p = \frac{1}{2}$. 为此可考虑下述零假设:

$$H_0: p_i = \frac{1}{2}^i, i = 1, 2, \dots$$

再因 $p_{i+1}^* = P\{X > i\} = (1 - p)^i, i = 0, 1, 2, \dots$, 当将首次摸到白球的抽取次数大于或等于 5 的频数合并成一组后, 便可转而考虑下述零假设:

$$H_0^*: p_i = \frac{1}{2}^i, i = 1, 2, 3, 4; p_5^* = \frac{1}{2}^4.$$

这样, 便可采用多项分布的 χ^2 检验法来解. 本例中 $r = 5; p_{i,0} = \frac{1}{2}^i, i = 1, 2, 3, 4, p_{5,0} = \frac{1}{2}^4; n_1 = 43, n_2 = 31, n_3 = 15, n_4 = 6, n_5 = 5, n = 100$. 由此可算出

$$\chi^2_0 = \sum_{i=1}^r \frac{(n_i - np_{i,0})^2}{np_{i,0}} = 3.2.$$

对于给定的显著性水平 $\alpha = 0.05$, 查附表 3 知 $\chi^2_{\alpha}(r-1) = \chi^2_{0.05}(4) = 9.488$. 因 $\chi^2_0 = 3.2 < \chi^2_{0.05}(4) = 9.488$, 便知无充分理由拒绝零假设, 从而认为黑盒中白球与黑球个数相等.

关于连续型总体的例子, 我们将在理论分布中含未知参数的情形中给出 (见例 6.18).

2. 理论分布含未知参数的情形

以上讨论中假定 $F_0(x)$ 是不含未知参数的完全已知的确定分布, 现进一步假定这一分布尚依赖于若干未知参数, 从而需考虑下述假设检验问题:

$$H_0: F(x) = F_0(x; \theta_1, \theta_2, \dots, \theta_t), \tag{6.57}$$

其中 F_0 的分布类型已知, 但含 t 个未知参数: $\theta_1, \theta_2, \dots, \theta_t$.

上述假设检验最常遇见的例子是要检验“总体服从正态分布”这一假设, 这时零假设 H_0 给出一个包含两个未知参数 μ 与 σ^2 的分布簇:

$$\{N(\mu^2): -\infty < \mu < +\infty, \mu^2 > 0\}.$$

如零假设 H_0 由(6.57)给出, 就不能直接套用前述的检验程序. 这是因为, 由(6.54)式给出的诸概率 $p_{i,0}$ 就不再是已知正数了. 事实上,

$$p_{i,0}(\theta_1, \theta_2, \dots, \theta_t) = F_0(a_i; \theta_1, \theta_2, \dots, \theta_t) - F_0(a_{i-1}; \theta_1, \theta_2, \dots, \theta_t). \quad (6.58)$$

鉴于诸参数 $\theta_1, \theta_2, \dots, \theta_t$ 未知, 即使已知 F_0 的分布类型, 也无法依上式算出诸概率 $p_{i,0}(\theta_1, \theta_2, \dots, \theta_t)$ 的确切值, 从而也就不能算出由(6.49)式给出的统计量 χ_0^2 之值. 为此, 需对原有检验程序作适当修改, 具体做法如下:

(1) 以诸未知参数的极大似然估计量 $\hat{\theta}_i, i=1, 2, \dots, t$, 来替换(6.58)式中的诸未知参数 θ_i .

在由(6.57)给出的零假设成立的前提下, 首先求出诸未知参数 θ_i 的极大似然估计量 $\hat{\theta}_i$. 然后再将它们代入(6.58)式, 求出诸概率 $p_{i,0}$ 的估计量:

$$\hat{p}_{i,0} = F_0(a_i; \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_t) - F_0(a_{i-1}; \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_t), \quad i=1, 2, \dots, r. \quad (6.59)$$

以上约定 $F_0(a_0; \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_t) = 0, F_0(a_r; \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_t) = 1$.

(2) 构造原有统计量 χ_0^2 的估计量:

$$\chi_0^2 = \sum_{i=1}^r \frac{(N_i - np_{i,0})^2}{np_{i,0}} \quad (6.60)$$

费希尔(R. A. Fisher)证明了, 如零假设 H_0 成立, 且分布 $F_0(x; \theta_1, \theta_2, \dots, \theta_t)$ 满足一定的正则条件, 则当样本容量 n 趋于无穷时, 上述统计量 χ_0^2 的极限分布是自由度为 $r-t-1$ 的 χ^2 分布. 这一事实的证明已超出本书范围, 有兴趣的读者可参阅文献[8], p 298, p. 302—305.

(3) 确定否定域

以 n_1, n_2, \dots, n_r 表示诸频数统计量 N_1, N_2, \dots, N_r 的观测值, 并仍以 χ_0^2 表示由(6.60)式确定的统计量的观测值, 则类似于理论分布完全已知的情形, 当样本容量 n 充分大时, 可将否定域取为:

$$C = \{(n_1, n_2, \dots, n_r): \chi_0^2 > \chi^2_{(r-t-1)}\}.$$

其中 $\chi^2_{(r-t-1)}$ 是水平 α 的自由度为 $r-t-1$ 的 χ^2 分布的上侧分位数.

显然, 由上述叙述知, 在采用上述检验法时, 至少应将实轴划分成 $t+2$ 个互不相交的子区间, 即要求 $r \geq t+2$.

例 6.18 随机地抽取了 2000 年 1 月出生的 50 名男婴, 分别测出他们的体重, 并算出样本均值为 $\bar{x} = 3160$ (g), 未修正样本方差为 $s_0^2 = 465.5^2$ (g²). 再按体重分组, 统计出体重不超过 2450g 的男婴有 2 名, 体重超过 3700g 的有 3 名. 体重数分属于下述区间(2450, 2700], (2700, 2950], (2950, 3200], (3200, 3450], (3450, 3700] 的男婴各有 5、7、12、10 与 11 名. 试以这些观察数据判断新生男婴

的体重是否服从正态分布(显著性水平 = 0.05).

解 以 X 表示所生男婴的体重. 待检验的零假设为:

$$H_0: X \sim N(\mu, \sigma^2).$$

由给出的统计数据知: $n=50, r=7; n_1=2, n_2=5, n_3=7, n_4=12, n_5=10, n_6=11$ 与 $n_7=3$.

另在第五章中已求出正态分布两个参数 μ 与 σ^2 的极大似然估计量分别为样本均值与样本的未修正方差, 即有 $\mu=\bar{X}, \sigma^2=S_0^2$.

这样, 在零假设 H_0 成立的前提下, 可先利用标准正态分布函数计算:

$$F_0(a_i; \mu, \sigma^2) = \Phi\left(\frac{a_i - 3160}{465.5}\right), \quad i=1, 2, \dots, 6.$$

其中 $a_i = 2450 + (i-1) \times 250, i=1, 2, \dots, 6$.

再由上述数据计算

$$p_{i,0} = F_0(a_i; \mu, \sigma^2) - F_0(a_{i-1}; \mu, \sigma^2), \quad i=1, 2, \dots, 7.$$

以上约定 $F_0(a_0; \mu, \sigma^2) = 0, F_0(a_7, \mu, \sigma^2) = 1$.

最后, 将通过上述计算程序得到的数据列表计算出 χ^2_0 , 如下表所示:

表 6-5 计算 χ^2_0 所需的中间数据

i	1	2	3	4	5	6	7
$p_{i,0}$	0.063	0.098	0.165	0.210	0.196	0.145	0.123
$np_{i,0}$	3.15	4.90	8.25	10.50	9.80	7.25	6.15
n_i	2	5	7	12	10	11	3
$(n_i - np_{i,0})^2$	1.323	0.010	1.563	2.250	0.040	14.063	9.923
$(n_i - np_{i,0})^2 / np_{i,0}$	0.420	0.002	0.189	0.214	0.004	1.940	1.613

将上表中最后一行数据相加, 即可算出

$$\chi^2_0 = \sum_{i=1}^7 \frac{(n_i - np_{i,0})^2}{np_{i,0}} = 4.382$$

对于给定的显著性水平 α , 查附表 3 即知

$$\chi^2_{\alpha}(r-t-1) = \chi^2_{0.05}(4) = 9.488 \quad (r=7, t=2).$$

因 $\chi^2_0 = 4.382 < \chi^2_{\alpha}(r-t-1) = 9.488$, 故没有充分理由拒绝零假设 H_0 , 从而认为男婴的体重服从正态分布.

三、独立性检验

本分节将检验关于两总体的独立性的假设. 确切地说, 设有两个总体 X, Y . 待检验的零假设为:

$$H_0: X \text{ 与 } Y \text{ 独立.} \quad (6.61)$$

一般说来, 这是一个非参数假设检验问题. 解决该问题的思路也是将两总体有限地离散化, 从而将原来的非参数假设检验问题转化为一个可借助多项分布的²检验法处理的多参数假设检验问题. 具体步骤如下所述:

(1) 将两个总体有限地离散化

首先, 将总体 X 与 Y 的取值范围分别划分成 r 和 q 个互不相交的子区间 A_1, A_2, \dots, A_r 与 B_1, B_2, \dots, B_q . 并记

$$p_{ij} = P\{X \in A_i, Y \in B_j\}, \quad i = 1, 2, \dots, r; \quad j = 1, 2, \dots, q;$$

$$p_{i0} = P\{X \in A_i\} = \sum_{j=1}^q p_{ij}, \quad i = 1, 2, \dots, r;$$

$$p_{0j} = P\{Y \in B_j\} = \sum_{i=1}^r p_{ij}, \quad j = 1, 2, \dots, q;$$

其次, 根据总体 X 与 Y 取值范围的上述分割, 可将关于总体 X 与 Y 的独立性假设转化成下述零假设:

$$H_0^*: p_{ij} = p_{i0} \cdot p_{0j}, \quad i = 1, 2, \dots, r; \quad j = 1, 2, \dots, q. \quad (6.62)$$

这是一个含 $(r+q)$ 个未知参数 p_{i0} ($i = 1, 2, \dots, r$) 与 p_{0j} ($j = 1, 2, \dots, q$) 的关于 $(r \times q)$ 个参数 p_{ij} ($i = 1, 2, \dots, r; j = 1, 2, \dots, q$) 的零假设, 从而便可采用上一分节介绍的方法来检验.

(2) 取样并由样本确定诸频数统计量

取两元总体 (X, Y) 的一个容量为 n 的样本: $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, 并记 N_{ij} 为上述样本中诸分量落入矩形区域 $A_i \times B_j$ 的频数. 再令

$$N_{i\cdot} = \sum_{j=1}^q N_{ij}, \quad i = 1, 2, \dots, r; \quad N_{\cdot j} = \sum_{i=1}^r N_{ij}, \quad j = 1, 2, \dots, q.$$

显然有

$$n = \sum_{i=1}^r N_{i\cdot} = \sum_{j=1}^q N_{\cdot j} = \sum_{i=1}^r \sum_{j=1}^q N_{ij}.$$

可把上述诸统计量列成下表, 通常称为联列表.

(3) 求 $(r+q)$ 个未知参数 p_{i0} ($i = 1, 2, \dots, r$) 与 p_{0j} ($j = 1, 2, \dots, q$) 的极大似然估计量, 并构造统计量²₀.

鉴于 p_{i0} 与 p_{0j} 为未知参数, 我们需替之以它们的极大似然估计量 \hat{p}_{i0} 与 \hat{p}_{0j} . 可以证明:

表 6-6 两元联列表

		B _j				(N _{i0})
		1	2	...	q	
A _i	1	N ₁₁	N ₁₂	...	N _{1q}	N _{1.}
	2	N ₂₁	N ₂₂	...	N _{2q}	N _{2.}
	r	N _{r1}	N _{r2}	...	N _{rq}	N _{r.}
(N _{0j})		N _{.1}	N _{.2}	...	N _{.q}	n

$$p_{i.} = \frac{N_{i.}}{n}, \quad i= 1, 2, \dots, r; \quad p_{.j} = \frac{N_{.j}}{n}, \quad j= 1, 2, \dots, q.$$

这样，由（6.60）式，零假设 H₀^{*} 与上述诸式便可构造下述统计量：

$$\chi^2_0 = n \sum_{i=1}^r \sum_{j=1}^q \frac{(N_{ij} - N_{i.}N_{.j}/n)^2}{N_{i.}N_{.j}} \tag{6.63}$$

(4) 确定零假设 H₀^{*} 成立时上述统计量 χ^2_0 的极限分布

鉴于 $\sum_{i=1}^r p_{i.} = \sum_{j=1}^q p_{.j} = 1$ ，便知 (r+ q) 个未知参数中仅有 (r+ q- 2) 个独立参数. 再由于 r(q- 1) = (r+ q- 2) - 1 = (r- 1) (q- 1)，由费希尔证明的事实即知，若零假设 H₀^{*} 成立，则当样本容量 n 趋于无穷时，由（6.63）式定义的统计量 χ^2_0 的极限分布是自由度为 (r- 1) (q- 1) 的 χ^2 分布. 这样，借助上述统计量 χ^2_0 ，便可利用 χ^2 检验法的上侧检验法来检验零假设 H₀^{*}. 之后的细节是熟知的，故不再赘述. 通常称上述假设检验为两总体的独立性检验.

例 6.19 为了解吸烟习惯与患慢性气管炎病的关系，对 339 名 50 岁以上的人作了调查. 详细情况见下表：

吸烟习惯与患慢性气管炎病的关系调查表

	患慢性气管炎者	未患慢性气管炎者	合 计	患病率
吸烟	43	162	205	21%
不吸烟	13	121	134	9.7%
合计	56	283	339	16.5%

试由上表提供的数据判断吸烟者与不吸烟者的慢性气管炎的患病率是否有所不同（显著性水平 = 0.01）.

解 以 A₁ 与 A₂ 区分吸烟与不吸烟；另以 B₁ 与 B₂ 区分患与不患慢性气管炎，故有 r= q= 2. 另由上表中所列数据知，n₁₁= 43，n₁₂= 162，n₂₁= 13，n₂₂=

121; $n_{1.} = 205$, $n_{2.} = 134$; $n_{.1} = 56$, $n_{.2} = 283$, $n = 339$. 将上述数据代入 (6.63) 式即可算出:

$$\chi^2_{0.01} = 339 \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - n_{i.}n_{.j}/339)^2}{n_{i.}n_{.j}} = 7.48$$

对给定的显著性水平 $\alpha = 0.01$, 查附表可知,

$$\chi^2_{0.01}((r-1)(q-1)) = \chi^2_{0.01}(1) = 6.635, \text{ 因}$$

$$\chi^2_{0.01} = 7.48 > \chi^2_{0.01}((r-1)(q-1)) = 6.635$$

故拒绝零假设 H_0 , 即认为慢性气管炎的患病率与吸烟有关.

习 题 六

(A)

1. 设总体 X 服从参数为 $(\lambda > 0)$ 的泊松分布, 参数 λ 未知, $(X_1, X_2, \dots, X_{20})$ 为其一个样本. 对下述假设检验问题:

$$H_0: \lambda = 0.2 \quad H_1: \lambda = 0.1,$$

取否定域为: $C = \{(X_1, X_2, \dots, X_{20}): X_1 + X_2 + \dots + X_{20} = 0\}$. 求犯第一类错误与第二类错误的概率.

2. 设总体 $X \sim N(\mu, 9)$, μ 为未知参数, $(X_1, X_2, \dots, X_{25})$ 为其一个样本. 对下述假设检验问题:

$$H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0,$$

取否定域 $C = \{(X_1, X_2, \dots, X_{25}): |\bar{X} - \mu_0| > c\}$. 试求常数 c , 使得该检验的显著性水平为 0.05.

3. 某百货商场的日销售额服从正态分布, 去年的日均销售额为 53.6 (万元), 方差为 6^2 , 今年随机抽查了 10 个日销售额, 分别是

57.2, 57.8, 58.4, 59.3, 60.7, 71.3, 56.4, 58.9, 47.5, 49.5.

根据经验, 方差没有变化, 问今年的日均销售额与去年相比有无显著变化? ($\alpha = 0.05$)

4. 有一种安眠剂, 据说, 在一定剂量下能比某种旧安眠剂平均增加睡眠 3h. 已知使用旧安眠剂的睡眠时间 (单位: h) 服从正态分布 $N(20.8, 1.8^2)$. 为了检验对新安眠剂的这种说法是否正确, 收集了一组使用新安眠剂的睡眠时间 (单位: h):

26.7, 22.0, 24.1, 21.0, 27.2, 25.0, 23.4

试问这组数据能否说明新安眠剂确有新的疗效? ($\alpha = 0.05$)

5. 以往一台机器生产的垫圈的平均厚度为 0.050cm, 为了检查这台机器是否处于正常工作状态, 现抽取 10 个垫圈的一组样本, 测得其平均厚度为 0.053, 样本方差为 0.0032², 在显著水平 (1) $\alpha = 0.05$, (2) $\alpha = 0.01$ 下, 检验机器是否处于正常工作状态.

6. 某厂生产的缆绳, 其抗拉强度的均值为 10 600 (kg/cm^2). 今改进工艺后, 生产一批缆绳, 抽取 10 根, 测得抗拉强度为

10533, 10641, 10688, 10572, 10793, 10729, 10600, 10683, 10721, 10570.

认为抗拉强度服从正态分布. 当显著水平 $\alpha = 0.05$ 时, 问新生产的缆绳的抗拉强度是否

比过去生产缆绳的抗拉强度要高.

7. 某厂生产一种轴承, 在正常情况下强度检验显示轴承能够承受的压强服从正态分布 $N(80000, 4000^2)$ (单位为 kg/cm^2). 在生产过程中管理人员要经常抽样进行检验, 以判断生产情况是否正常. 今有 100 个样品, 测得承受压强的均值为 79600. 问生产是否正常 (假设方差不变). 该管理人员取定显著水平 $\alpha = 0.05$.

8. 某炼铁厂铁水的含碳量 X 在正常情况下服从正态分布, 现对操作工艺进行了某些改变, 从中抽取 7 炉铁水的试样, 测得含碳量数据如下: 4.421, 4.052, 4.357, 4.394, 4.326, 4.287, 4.683. 问是否可以认为新工艺炼出的铁水含碳量的方差仍为 0.112^2 ($\alpha = 0.05$).

9. 某洗衣粉包装机, 在正常工作情况下, 每袋标准重量为 1000g, 标准差不能超过 15g. 假设每袋洗衣粉的净重服从正态分布. 某天为检查机器工作是否正常, 从已装好的袋中, 随机抽查 10 袋, 测其净重 (克) 为:

1020, 1030, 968, 994, 1014, 998, 976, 982, 950, 1048.

问这天机器工作是否正常 ($\alpha = 0.05$)?

10. 为研究正常成年男、女血液红细胞的平均数之差别, 检查某地正常成年男子 156 名, 正常成年女子 74 名, 计算得男性红细胞平均数为 465.13 万/ mm^3 , 样本标准差为 54.80 万/ mm^3 ; 女性红细胞平均数为 422.16 万/ mm^3 , 样本标准差为 49.20 万/ mm^3 . 由经验知道正常成年男性与女性的红细胞数均服从正态分布, 且方差相同. 试检验该地正常成年人的红细胞平均数是否与性别有关. ($\alpha = 0.01$)

11. 在上题中假设男、女性红细胞数的分布的方差相等. 仍以上题中数据检验这一假设是否成立. ($\alpha = 0.10$)

12. 设双正态总体假设检验中, 两正态总体的方差相等, 但未知. 确切地说, 记 $\sigma_1^2 = \sigma_2^2 = \sigma^2$, 但 σ^2 未知. 试证, 这时由 (6.24) 式与 (6.25) 式表述的统计量 S^2 是未知方差 σ^2 的无偏估计量.

13. 一种特殊药品的生产厂家声称, 这种药能在 8h 内解除一种过敏的效率有 90%, 在有这种过敏的 200 人中, 使用药品后, 有 160 人, 在 8h 内解除了过敏, 试问生产厂家的说法是否真实 ($\alpha = 0.01$)?

14. 在某地抽查了 27 个家庭, 其中有 6 家使用 H 牌洗衣粉, 问 H 牌洗衣粉在该地的占有率是否大于 1/6 ($\alpha = 0.05$)?

15. 在某公路上, 50min 之间, 观察每 15 秒内过路的汽车的辆数, 得到频数分布如下:

过路的车辆数	0	1	2	3	4	5
频 数	92	68	28	11	1	0

问这个分布能否认为是泊松分布 ($\alpha = 0.10$)?

16. 在一正四面体的 20 个面上, 分别标以数字 0, 1, 2, ..., 9, 每个数字在两个面上标出. 为检验其匀称性, 共作 800 次投掷试验, 数字 0, 1, 2, ..., 9 朝正上方的次数如下表所示:

数 字	0	1	2	3	4	5	6	7	8	9
频数	74	92	83	79	80	73	77	75	76	91

问该正 20 面体是否匀称? ($\alpha = 0.05$)

17. 1992 年调查郊区某桑场采桑员和辅助工的桑毛虫皮炎发病情况, 结果如下表:

	采 桑	不采桑	合 计
患者人数	18	12	30
健康人数	4	78	82
合 计	22	90	112

试问发生皮炎是否与工种有关? ($\alpha = 0.05$)

18. 假设要对从 3 名候选人 A、B、C 中选出一名学生代表的投票进行分析. 设有 4 个系的学生参加选举, 每张选票上都标明投票者所在系. 现随机抽出 200 张选票, 投票情况如下表所示:

系 别	候 选 人			合 计
	A	B	C	
一	24	23	12	59
二	24	14	10	48
三	17	8	13	38
四	27	19	9	55
合 计	92	64	44	200

试问投票结果与投票者在什么系是否相互独立 ($\alpha = 0.10$)?

习 题 六

(B)

1. 设总体 $X \sim N(\mu, \sigma^2)$, μ 为已知常数, σ^2 未知. 又设 (X_1, X_2, \dots, X_n) 是总体 X 的容量为 n 的一个样本. 对总体方差 σ^2 作假设检验时, 试构造一个有别于 (6.17) 式的枢轴量 W , 并使之服从一已知的确定分布. 再对给定的显著性水平 α , 写出三类假设检验的否定域.
2. 设两正态总体 $X \sim N(\mu_1, \sigma_1^2)$ 与 $Y \sim N(\mu_2, \sigma_2^2)$ 相互独立, μ_1 与 μ_2 为已知常数, 两总体方差的比值 $r = \sigma_1^2 / \sigma_2^2$ 未知. 又设 $(X_1, X_2, \dots, X_{n_1})$ 与 $(Y_1, Y_2, \dots, Y_{n_2})$ 分别为总

体 X 与 Y 的两个样本. 试构造一个有别于 (6.29) 式的枢轴量 F , 并使之服从一已知的确定分布, 以对下述零假设作假设检验:

(1) $H_0: r = 1$; (2) $H_0: r > 1$; (3) $H_0: r < 1$. 再对给定的显著性水平 α , 写出上述三类假设检验的否定域.

3. 保险学中一般假定非寿险保单的持有人在保单固定的有效期内, 因事故要求索赔的次数服从泊松分布. 设泊松分布的参数为 λ , 通常该参数值是未知的. 保险公司为合理地厘定保险费, 需对此未知参数作统计推断. 为此, 先抽取同一非寿险险种的 4000 张保单, 假定每张保单具有相同的有效期, 且从同一日生效. 这 4000 张保单的索赔经验如下表所示:

产生 0, 1, 2 或 3 次索赔的保单数	
索赔次数	保 单 数
0	3288
1	642
2	66
3	4
保单数总计	4000

试利用上表中提供的数据:

- (1) 求未知参数 λ 的极大似然估计量的估计值;
- (2) 运用大样本的统计推断方法检验下述假设:

$H_0: \lambda = 0.2 \qquad H_1: \lambda > 0.02.$

4. 设计如下的试验, 用以检验受试者是否具有特异功能: 在三张卡片上分别写上小写字母 a, b, c , 并把三张卡片字母朝下按任意次序放好. 再给受试者三个分别写上大写字母 A, B, C 的信封, 让他把卡片与信封配对, 即把他认为写着字母 a (或 b, c) 的卡片放入信封 A (或 B, C) 中. 然后检查配对正确的个数. 如此重复 50 次. 现将某一受试者的试验结果列成下表:

正确配对个数的频数表	
正确配对个数	出现频数
0	14
1	24
3	12
试验总次数	50

试根据上表中数据检验受试者是否有特异功能 (显著性水平 $\alpha = 0.05$).

5. 设总体 X 服从以 p ($0 < p < 1$) 为参数的伯努利总体, 参数 p 未知. (X_1, X_2, \dots, X_n) 是总体 X 的一个容量为 n 的样本. 对下述假设检验问题:

$H_0: p = p_0 \qquad H_1: p \neq p_0 \quad (p_0 \text{ 为指定正数})$

可构造枢轴量:

$$Q=\frac{n(\overline{X}-p)^2}{p(1-p)}$$

- (1) 证明, 当样本容量 n 时, 枢轴量的极限分布为自由度等于 1 的 χ^2 分布;
- (2) 证明, 上述枢轴量 Q 即为皮尔逊为检验多项分布引入的 (相应于 $r=2$ 时的) χ^2 统计量.
- (3) 蒲封投掷钱币 4040 次, 得到图案向上的频数为 $n_1=2048$. 试以枢轴量 Q 为出发点检验该钱币是否均匀 (显著性水平 $\alpha=0.05$).

6. 设 $A_i=\{x:\frac{i-1}{4}<x<\frac{i}{4}\}$, $i=1,2,3$; $A_4=\{x:\frac{3}{4}<x<1\}$, 取 80 个观察值, 其中落入区间 A_i ($i=1,2,3,4$) 的频数分别为 6, 18, 20 与 36. 试问: 在 0.1 的显著性水平下, 假设总体 X 的密度函数为

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

是否可信?

第 7 章

方 差 分 析

类似于我们在前面一章中讲过的对等方差两个正态总体均值之间差异进行的 t 检验，方差分析（Analysis of Variance，简记作 ANOVA）是一种用来对于两个以上等方差的正态总体均值之间的差异进行检验的统计方法。方差分析的主要思想和理论是由费希尔（R. A. Fisher）在 1920 年左右提出来的。

本章主要介绍单因素方差分析和双因素方差分析模型的结构、分析的思路及步骤。

§ 7.1 问题的提出

我们先看下面的例子。

例 7.1 某管理学院对自己培养出来的 MBA 学生毕业之后的工作情况进行跟踪调查，希望了解四个不同专业毕业的 MBA 学生在第一年工作中所获得的平均收入是否有显著的差别。学院从已经毕业的学生当中按不同专业分别随机抽取 10 名同学进行调查。表 7-1 中列出了调查的结果。

表 7-1 某学院 MBA 毕业生工作第一年收入调查表（单位：万元）

专业	调查结果										平均
A ₁	9.6	8.3	5.2	13.3	8.1	13	10.2	4.6	11.4	10.1	9.38
A ₂	7.8	12.1	11.2	3.6	7.9	4.1	10.5	8.7	16	9.1	9.10
A ₃	11.3	14	6.2	8.3	10.8	6.3	9.7	11.3	12.7	8.9	9.95
A ₄	9.5	10.6	8.2	17.5	7.2	11	7.1	21	4.5	10.2	10.68

表 7-1 的最右边一列给出了调查得到的四个专业毕业生第一年的平均收入情况。可以看到，这四个值是不同的，A₄ 专业毕业的学生的平均收入显得比其它专业的学生收入要高一些，那么，我们是否可以由此断定专业的选择对该学院的 MBA 学生毕业后第一年的平均收入是有影响的呢？

回答这一问题之前，我们先分析一下调查得到的样本数据。事实上，不仅不同专业毕业的学生的收入之间是不同的，即便同一专业的学生之间收入也是

有差别的. 影响个人收入的因素是多方面的, 除了学历、工作时间、性别等方面之外, 还有个人的经历、能力、运气等偶然性的因素起作用. 因此, 分析以上四个专业学生平均收入的差异性不仅要考虑到专业不同的影响, 还要考虑到影响到每一个人收入的偶然性因素的作用. 如果这种差异性主要是由后者造成的, 那么我们就无法得到刚才提出的问题中的结论, 从而认为专业不同对学生第一年的工作收入影响并不显著; 反之, 如果相对这些偶然性因素的影响, 专业的区别对造成这种收入的差异性起到了更加主要的作用, 那么我们就认为专业这一因素对学生第一年的工作收入影响是显著的.

这种通过分析对比导致变量取值差异性的主要原因来确定某些因素对该变量的影响是否显著的方法就是方差分析的主要思想. 在例 7.1 中只考虑了专业一个因素的影响, 这类问题因而称为单因素方差分析, 其中专业的四个不同值通常称为四个水平.

类似地, 也可以同时考虑多个因素对某一变量的影响问题. 与单个因素的方差分析问题不同的是, 导致各个样本之间的差异性的原因除了给定的几个因素和偶然性的误差之外, 还可能是几个因素对变量的“交互作用”, 因素间可能存在的“交互作用”使得采集样本和进行方差分析等问题都变得更加复杂了.

在本书中, 我们仅考虑两个因素的方差分析问题, 又称双因素方差分析.

例 7.2 为了叙述方便我们仍然以例 7.1 中考虑的问题作背景. 假设该学院现在不仅希望了解专业选择对 MBA 毕业生工作第一年收入的影响, 还希望弄清楚与其他四所大学里的管理学院相同专业毕业的 MBA 学生相比, 该学院毕业的学生是否具有优势. 我们可以把后一点理解成 MBM 毕业生的收入是否与选择的学校这一因素有关.

与例 7.1 中的情况不同, 现在考虑的是五所大学、四种专业的 MBA 毕业生第一年的工作收入情况, 我们仍然以 A_i , $i = 1, 2, 3, 4$, 表示四种专业, 以 B_j , $j = 1, 2, 3, 4, 5$, 分别表示五所大学. 为了考察大学和专业的选择两种因素对毕业学生工作收入的影响, 必须分别从每一所大学里的每一个专业毕业的学生中随机抽取一定数目的样本单位并调查他们工作第一年的收入情况.

如果我们不考虑大学和专业两个因素对 MBA 收入的“交互作用”, 这意味着假设某所大学的一个专业毕业生比另一所大学该专业的毕业生平均收入高出一千元的话, 那么这两所大学其它专业的毕业生的平均收入也将保持相同方向相同程度的差距, 换句话说, 该大学毕业生比另一所大学毕业生在收入上的优势不会随专业的不同而改变, 与专业无关; 同理, 假如某一专业毕业生相比另一专业毕业生在收入上多了一千元, 那么对每个大学这种专业上的优势都是相同的, 与大学无关. 这就叫做大学与专业两种对 MBA 收入没有“交互作用”. 如果我们假设这种“交互作用”不存在, 那么在抽取样本时就可以只对每所大学

每个专业随机抽取一个毕业生做调查就可以了, 这样只需调查 20 名毕业生的收入数据, 每个专业对应了 5 个人, 每个大学对应了 4 个人, 导致这 20 人收入差异性的原因可能是专业的因素, 也可能是大学的因素或者偶然性的因素.

反过来, 如果我们需要考虑大学和专业两个因素的可能的“交互作用”, 就必须对每所大学里的每个专业抽取至少两个的毕业生做调查, 只有这样才能够反映出可能的“交互作用”的影响而不是偶然因素的作用.

以下我们规定对每所大学里的每个专业抽取相等数目的毕业生, 即对应每一个组合 (B_j, A_i) , $j = 1, 2, \dots, 5$, $i = 1, 2, \dots, 4$, 调查得到的样本容量相同. 这样得到的一组数据叫做均衡数据, 否则称之为非均衡数据. 本章中我们只考虑均衡数据的情况.

表 7-2 给出了对应每个大学和专业的组合各自调查了三个学生所得的结果.

表 7-2 五所大学 MBA 毕业生工作第一年收入调查表 (单位: 万元)

<div>大学 专业</div>	B ₁	B ₂	B ₃	B ₄	B ₅
A ₁	9.6	6.8	11.0	7.5	4.5
	8.3	10.2	7.3	6.3	6.8
	5.2	4.6	5.2	9.0	8.0
A ₂	7.8	8.1	5.7	9.8	3.8
	12.1	18.0	4.9	7.0	5.2
	11.2	6.5	13.0	3.6	1.8
A ₃	11.3	4.2	8.2	7.4	6.3
	14.0	9.5	7.3	16.0	2.6
	6.2	8.0	15.0	8.5	12.0
A ₄	9.5	6.4	7.5	12	5.2
	10.6	3.6	18.0	15.0	2.1
	8.2	12.0	9.4	8.8	7.8

在下面的两节中我们将分别对例 7.1 和例 7.2 给出方差分析的过程.

§ 7.2 单因素方差分析

一、模型的结构

首先我们可以把例 8.1 推广到一般的形式. 设某个因素 A 具有 r 个水平, 分别记为 A_1, A_2, \dots, A_r . 我们的问题是要研究因素 A 是否对某个数量变量 X 产生影响.

为此, 我们设水平 A_i 对应的样本来自总体 X_i ($i = 1, 2, \dots, r$). 在经典统

计分析中，通常还假设 X_i 服从正态分布 $N(\mu, \sigma^2)$ ($i=1, 2, \dots, r$)，即 r 个总体服从方差相等的正态分布，这里的等方差性是方差分析中对模型做出的重要的假设. 容易看出，如果 r 个总体的均值也都相等，此时 r 个总体将服从相同的正态分布，这说明了因素 A 的不同水平对应的变量在统计意义上讲是没有差异的，因此因素 A 对标志值 X 没有显著影响；反之，如果 r 个总体的均值不全相等，则可以认为因素 A 对 X 有显著影响.

表 7-3 单因素方差分析数据结构

水 平	抽样结果						平均
A_1	X_{11}	X_{12}	X_{13}	X_{14}	X_{1k}	$\overline{X_1}$
A_2	X_{21}	X_{22}	X_{23}	X_{24}	X_{2k}	$\overline{X_2}$
A_r	X_{r1}	X_{r2}	X_{r3}	X_{r4}	X_{rk}	$\overline{X_r}$

这样一来，刚才提出的问题就变成了要对以下的假设进行检验：

$$H_0: \mu_1 = \mu_2 = \dots = \mu_r = \mu \tag{7.1}$$

为了检验该假设，在 r 个总体中各自独立地抽取容量为 k 的样本，得到 X_{ij} ， $i=1, 2, \dots, r$ ； $j=1, 2, \dots, k$ ，其中 X_{ij} ， $j=1, 2, \dots, k$ 表示水平 A_i 对应的一个样本. 见表 7-3 所示. 因此，按照前面对模型做出的假设，有 $X_{ij} \sim N(\mu_i, \sigma^2)$ ， $j=1, 2, \dots, k$. 于是，对其中的每个 i 令

$$X_{ij} = \mu_i + \varepsilon_{ij}, \quad j=1, 2, \dots, k. \tag{7.2}$$

则水平 A_i 对应的样本中第 j 个个体被分解成相应均值 μ_i 和第 j 次抽样的随机因素的影响 ε_{ij} ，其中

$$\varepsilon_{ij} \sim N(0, \sigma^2), \text{ 且相互独立} \tag{7.3}$$

这里的(7.2), (7.3)即是我们所采用的模型形式.

为了讨论问题的方便，通常还将 (7.2) 式做进一步的分解，为此，令

$$\mu_i = \frac{1}{r} \sum_{i=1}^r \mu_i, \tag{7.4}$$

$$\alpha_i = \mu_i - \mu \quad i=1, 2, \dots, r, \tag{7.5}$$

则 (7.2) 式变成了

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i=1, 2, \dots, r; \quad j=1, 2, \dots, k. \tag{7.2'}$$

这里的 μ 是 r 个总体均值 μ_i 的平均值. 如果原假设 (7.1) 成立，则每个均值 μ_i 都与 μ 相等； α_i 则体现了水平 A_i 对变量的特殊影响，称之为水平 A_i 的效应 (effect)，根据它的定义易知

$$\sum_{i=1}^r \alpha_i = 0 \tag{7.6}$$

这样一来, 检验假设(7.1)的问题就变成了要基于模型(7.2), (7.3), (7.6)对以下零假设进行检验:

$$H_0: \mu_i = 0, \quad i = 1, 2, \dots, r \quad (7.7)$$

即要检验 r 个水平的效应是否全部为零; 相应的备择假设为

H_1 : 存在某一个 $i, 1 \leq i \leq r$, 使得 $\mu_i \neq 0$.

在探讨如何对以上假设进行检验之前, 我们不妨先来看一下模型 (7.2), (7.3), (7.6) 中各参数的估计问题. 以求极大似然估计为例, 对于任给的 $i, 1 \leq i \leq r$, X_{ij} 服从分布 $N(\mu + \mu_i, \sigma^2)$, $j = 1, 2, \dots, k$, 其密度函数为

$$f(x_{ij}) = (2\pi\sigma^2)^{-\frac{1}{2}} \cdot \exp\left[-\frac{1}{2\sigma^2}(x_{ij} - \mu - \mu_i)^2\right] \quad (7.8)$$

由此可得似然函数(未知参数 $\mu, \mu_1, \mu_2, \dots, \mu_r; \sigma^2$ 的函数)为 $L(\mu, \mu_1, \mu_2, \dots, \mu_r; \sigma^2) = (2\pi\sigma^2)^{-\frac{rk}{2}} \cdot \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^r \sum_{j=1}^k (x_{ij} - \mu - \mu_i)^2\right]$. (7.9)

对于每一个给定的 σ^2 , 求 L 的最大值即相当于求以下函数的最大值.

$$V(\mu, \mu_1, \mu_2, \dots, \mu_r) = \sum_{i=1}^r \sum_{j=1}^k (x_{ij} - \mu - \mu_i)^2. \quad (7.10)$$

根据多元函数微分学的知识, 在函数 V 的极大值点上其一阶偏导数为零, 因此, 求 V 的最大值点可以通过求解以下线性方程组获得.

$$\begin{aligned} \frac{\partial V}{\partial \mu} &= -2 \sum_{i=1}^r \sum_{j=1}^k (x_{ij} - \mu - \mu_i) = 0, \\ \frac{\partial V}{\partial \mu_i} &= -2 \sum_{j=1}^k (x_{ij} - \mu - \mu_i) = 0, \quad i = 1, 2, \dots, r \end{aligned} \quad (7.11)$$

求解该方程组并利用(7.6)式就可以得到 μ 和水平 A_i 的效应值 μ_i 的最大似然估计分别为

$$\mu = \frac{1}{rk} \sum_{i=1}^r \sum_{j=1}^k X_{ij} = \bar{X}, \quad (7.12)$$

$$\mu_i = \frac{1}{k} \sum_{j=1}^k X_{ij} = \bar{X}_i, \quad (7.13)$$

$$\mu_i = \mu - \mu = \bar{X}_i - \bar{X}. \quad (7.14)$$

可以证明, 以上估计量分别是 $\mu, \mu_i (i = 1, 2, \dots, r)$ 的无偏估计. 这里的 \bar{X} 表示所有观测的平均值, \bar{X}_i 表示水平 A_i 对应的样本的平均值. 类似地, 可以求得 σ^2 的极大似然估计量为

$$\sigma^2 = \frac{1}{rk} \sum_{i=1}^r \sum_{j=1}^k (X_{ij} - \mu - \mu_i)^2 = \frac{1}{rk} \sum_{i=1}^r \sum_{j=1}^k (X_{ij} - \bar{X}_i)^2 \quad (7.15)$$

具体过程留做章后习题.

对于例 7.1, 我们可以求得 A_1, A_2, A_3, A_4 4 个专业对毕业生工作第一年平

均收入的效应的估计值. 容易算出 $\mu_i, i=1, 2, 3, 4$ 和 μ 的估计值分别为 $\mu_1=9.38, \mu_2=9.22, \mu_3=9.95, \mu_4=10.28, \mu=9.7075$, 所以四个效应的估计值分别为 $\mu_1=-0.3275, \mu_2=-0.4875, \mu_3=-0.2425, \mu_4=0.5725$. 相比之下, 似乎专业 A_4 对学生的平均收入的效应最大, 但是, 这里的四个效应值与零都相差不大, 为了获得较为可信的结论, 需要对零假设(7.1)或(7.7)进行检验.

二、检验统计量

正如我们在上一节对例 7.1 的分析中所指出的, 可以通过分析对比导致样本 $X_{ij}, i=1, 2, \dots, r, j=1, 2, \dots, k$ 之间的差异性的原因来确定因素 A 的影响是否显著. 为此, 我们引入离差平方和来度量各个体间的差异程度:

$$Q = \sum_{i=1}^r \sum_{j=1}^k (X_{ij} - \bar{X})^2 \quad (7.16)$$

导致个体间的差异被认为可能有两个原因, 一是因素 A 处于不同水平, 一是来自因素 A 以外的因素. 我们需要分析究竟因素 A 是否真的导致了个体间的明显差异, 进而说明因素 A 是否对我们所考虑的变量有显著影响, 为此, 我们将个体间的总差异按原因进行分解. 事实上 Q 可分解为:

$$\begin{aligned} Q &= \sum_{i=1}^r \sum_{j=1}^k [(X_{ij} - \bar{X}_{i\cdot}) + (\bar{X}_{i\cdot} - \bar{X})]^2 \\ &= \sum_{i=1}^r \sum_{j=1}^k (X_{ij} - \bar{X}_{i\cdot})^2 + 2 \sum_{i=1}^r \sum_{j=1}^k (X_{ij} - \bar{X}_{i\cdot}) (\bar{X}_{i\cdot} - \bar{X}) \\ &\quad + \sum_{i=1}^r k (\bar{X}_{i\cdot} - \bar{X})^2 \end{aligned} \quad (7.17)$$

根据前面(7.12)、(7.13)中对 $\bar{X}_{i\cdot}$ 和 \bar{X} 的定义可知

$$\sum_{i=1}^r \sum_{j=1}^k (X_{ij} - \bar{X}_{i\cdot}) (\bar{X}_{i\cdot} - \bar{X}) = 0 \quad (7.18)$$

因此, (7.17)式可改写为:

$$\begin{aligned} Q &= \sum_{i=1}^r \sum_{j=1}^k (X_{ij} - \bar{X}_{i\cdot})^2 + \sum_{i=1}^r k (\bar{X}_{i\cdot} - \bar{X})^2 \\ &= Q_1 + Q_2 \end{aligned} \quad (7.19)$$

式 (7.19) 通常被称为离差平方和的分解. Q_1 是对每一组内部样本观测值与该组样本平均之间的离差的平方和, 因此可以称之为组内平方和, 它衡量了每个样本单位受 A 以外的因素影响的程度, 也可以称之为误差平方和; Q_2 则是各个组的样本平均值与全部样本平均值离差的平方和, 它度量了组与组之间的系统性的差异, 因此可以称之为组间平方和或系统平方和.

Q_2 的大小在一定程度上反映了因素 A 对 X 的影响, 从而也反映了零假设(7.1)或(7.7)能否被拒绝. 事实上, 如果 Q_2 相对比较大, 它解释了总离差

平方和 Q 的主要部分, 则说明了组与组之间的系统性的差异是导致个体 X_{ij} 之间差异性的主要原因, 此时 A 的影响是显著的; 反之 Q_2 相对比较小时, A 的影响是不显著的. 因此, 直观上看出我们可以针对原假设 (7.1) 或 (7.7) 寻找形如满足条件 $Q_2/Q_1 > c$ (其中 c 是某一正的常数) 的否定域.

注意到 $X_{ij} - \mu \sim N(0, \sigma^2)$, 令 $\bar{X}_i = \frac{1}{k} \sum_{j=1}^k X_{ij}$, $\bar{X}_i = \frac{1}{r} \sum_{i=1}^r \bar{X}_i$, 记

$$Q = \sum_{i=1}^r \sum_{j=1}^k (X_{ij} - \bar{X}_i)^2$$

$$Q_1 = \sum_{i=1}^r \sum_{j=1}^k (X_{ij} - \bar{X}_i)^2$$

$$Q_2 = \sum_{j=1}^k k(\bar{X}_i - \bar{X})^2$$

容易证明: $\frac{1}{2}Q \sim \sigma^2(rk-1)$, $\frac{1}{2}Q_1 \sim \sigma^2(r(k-1))$, $\frac{1}{2}Q_2 \sim \sigma^2(r-1)$, 此外可以证明 Q_1 与 Q_2 相互独立(略去证明), 于是, 由命题 4.8, 可构造枢轴量

$$F = \frac{Q_2/(r-1)}{Q_1/r(k-1)} \sim F(r-1, r(k-1)). \quad (7.21)$$

另一方面, 容易验证:

$$Q = \sum_{i=1}^r \sum_{j=1}^k (X_{ij} - \bar{X}_i)^2$$

$$Q_1 = \sum_{i=1}^r \sum_{j=1}^k (X_{ij} - \bar{X}_i)^2$$

$$Q_2 = \sum_{i=1}^r k(\bar{X}_i - \bar{X})^2$$

于是在 H_0 成立时,

$$F = \frac{Q_2/(r-1)}{Q_1/r(k-1)} = \frac{Q_2/(r-1)}{Q_1/r(k-1)} \sim F(r-1, r(k-1)) \quad (7.22)$$

从而对给定的显著性水平 α , 可以构造下列否定域:

$$C = \{x: F = \frac{Q_2/(r-1)}{Q_1/r(k-1)} > F(r-1, r(k-1))\} \quad (7.23)$$

或表为:

$$C = \{x: \frac{Q_2}{Q_1} > \frac{r-1}{r(k-1)} F(r-1, r(k-1))\}$$

其中 $F(r-1, r(k-1))$ 是水平 α 的自由度为 $(r-1)$ 和 $r(k-1)$ 的 F 分布的上侧分位数.

三、方差分析表

根据前面的分析, 我们可以将解决本节开始提出的问题的思路小结如下:

- 1. 将问题转化成对零假设(7.1)或(7.7) 进行检验的问题;
- 2. 根据得到的样本, 依照公式(7.19)计算出 Q_1 和 Q_2 ;
- 3. 按照公式(7.22) 计算检验统计量 F 的值;
- 4. 给定显著水平 α , 查 F 分布表得到临界值 $F_{\alpha}(r-1, r(k-1))$, 与 F 进行比较并得出结论.

这就是进行单因素方差分析的步骤. 通常为了表达的方便和直观, 使用如表 7-4 所示的方差分析表.

表 7-4 单因素方差分析表

方差来源	平方和	自由度	均方和	F 值
组间	$Q_2 = \sum_{i=1}^r k(\bar{X}_{i\cdot} - \bar{X})^2$	$r-1$	$Q_2/(r-1)$	$\frac{Q_2/(r-1)}{Q_1/r(k-1)}$
组内	$Q_1 = \sum_{i=1}^r \sum_{j=1}^k (X_{ij} - \bar{X}_{i\cdot})^2$	$r(k-1)$	$Q_1/r(k-1)$	
总和	$Q = \sum_{i=1}^r \sum_{j=1}^k (X_{ij} - \bar{X})^2$	$rk-1$		

对于实际问题, 运用一些统计软件可以非常方便地得到形如表 7-4 所示的方差分析表. 下面的表 7-5 是针对例 8.1 运用 Microsoft Excel 97 中文版提供的“数据分析”功能得到的输出结果

表 7-5 例 7.1 方差分析表

差异源	SS	df	MS	F	P -Value	F _{crit}
组间	7.31475	3	2.43825	0.2105	0.88466	2.866265
组内	416.993	36	11.58314			
总计	424.30775	39				

表 7-5 中前五列的结构完全对应表 7-4, 其中 SS 表示平方和、df 表示自由度、MS 表示均方和, F 则对应表 7-4 中的 F 值, 表 7-5 中的 P -Value 表示 F 分

参见 Excel 97 中文版的有关技术说明书. 例如王其文主编, 经济管理计算机基础教程. 北京: 高等教育出版社, 1999

布的以 F 值为上侧分位数的上侧概率值, 易看出对设定的显著性水平 α , $F > F_{\alpha}$ 等价于 $P\text{-Value} < \alpha$. 该例中 $P\text{-Value}$ 等于 0.88466 即自由度为 3 和 36 的 F 分布, 水平为 0.88466 上侧分位数为 0.2105. 假如给定显著水平 α 等于 0.05, 则该 P 值远远大于 α . 表示在给定的显著水平下不能拒绝原假设, 从而认为专业的选择对毕业生工作第一年的收入没有显著影响. 表 7-5 最后一列也给出了临界值, 即 F_{crit} , 它是事先设定的显著性水平 α 对应的上侧分位数, 此例中事先设定了 $\alpha = 0.05$; 算得 $F_{\text{crit}} = F_{0.05}(3, 36) = 2.866265$, 由 $F < F_{\text{crit}}$, 同样也可以获得不能拒绝零假设的结论.

在结束本节之前, 需要说明两点: 首先在使用方差分析之前一定要注意该方法对模型做出的假定, 尤其是等方差的假定是一个比较严格的条件, 在这些条件不满足的情况下使用方差分析可能会导致错误的结论. 其次, 尽管本节仅对均衡数据的情况介绍了单因素方差分析的步骤, 但是可以毫无困难地将其推广到非均衡数据时的情况, 即对不同组抽取不同容量的样本. 在此不复赘述.

§ 7.3 双因素方差分析

现在研究两个因素对某一变量的影响问题. 我们用 A 、 B 分别表示两个因素, 并设因素 A 具有 r 个水平, 分别记作 A_1, A_2, \dots, A_r ; 因素 B 具有 s 个水平, 分别记作 B_1, B_2, \dots, B_s . 因素 A, B 不同水平的联合称为一个组合, 记作 (A_i, B_j) ($i = 1, 2, \dots, r; j = 1, 2, \dots, s$). 这样的组合一共有 rs 个. 类似地, 我们设对应于组合 (A_i, B_j) 的一组变量值来自总体 X_{ij} , 并且假定该总体服从正态分布, 即 $X_{ij} \sim N(\mu_{ij}, \sigma^2)$ ($i = 1, 2, \dots, r; j = 1, 2, \dots, s$). 这里 rs 个总体的方差是相等的. 容易看出, 这是对例 7.2 进行的一个自然的推广.

正如我们在前面例 7.2 的分析中所看到的, 研究 A 和 B 对 X_{ij} 的影响, 除了需要了解 A, B 两个因素各自的影响之外, 还可能存在着来自 A 和 B 对 X_{ij} 的“交互作用”的影响. 因此, 必须将其区别对待.

为此, 令

$$\mu_{i\cdot} = \frac{1}{s} \sum_{j=1}^s \mu_{ij} \quad (i = 1, 2, \dots, r) \quad (7.24)$$

$$\mu_{\cdot j} = \frac{1}{r} \sum_{i=1}^r \mu_{ij} \quad (j = 1, 2, \dots, s) \quad (7.25)$$

$$\mu = \frac{1}{rs} \sum_{i=1}^r \sum_{j=1}^s \mu_{ij} = \frac{1}{r} \sum_{i=1}^r \mu_{i\cdot} = \frac{1}{s} \sum_{j=1}^s \mu_{\cdot j} \quad (7.26)$$

则 $\mu_{i\cdot}$ 表示水平 A_i 对应的各组变量的平均, 所以如果 r 个 $\mu_{i\cdot}$ ($i = 1, 2, \dots, r$) 不全相等, 即表明因素 A 的 r 个水平所对应的变量的均值之间是有差异的,

从而意味着因素 A 的影响是显著的; 类似地, $\mu_{\cdot j}$ 代表了与水平 B_j 对应的各组总体的均值, s 个 $\mu_{\cdot j}$ ($j = 1, 2, \dots, s$) 不全相等同样意味着 B 的影响是显著的.

根据定义, μ 表示了各组总体的总平均值, 因此是变量 X 一般水平的反映. 如果记

$$\alpha_i = \mu_{\cdot i} - \mu \quad (i = 1, 2, \dots, r) \tag{7.27}$$

$$\beta_j = \mu_{\cdot j} - \mu \quad (j = 1, 2, \dots, s) \tag{7.28}$$

那么, 同样可以称 α_i 是水平 A_i 的效应, β_j 是水平 B_j 的效应, 而且, 由定义易知

$$\sum_{i=1}^r \alpha_i = 0 \tag{7.29}$$

$$\sum_{j=1}^s \beta_j = 0 \tag{7.30}$$

这样一来, 研究因素 A 的影响是否显著即是要检验假设

$$H_{01}: \alpha_i = 0 \quad (i = 1, 2, \dots, r) \tag{7.31}$$

而研究因素 B 的影响是否显著即要检验假设

$$H_{02}: \beta_j = 0, \quad (j = 1, 2, \dots, s) \tag{7.32}$$

那么什么是 A, B 两个因素的交互作用的影响呢?

令

$$r_{ij} = \mu_{ij} - \mu - \alpha_i - \beta_j, \quad (i = 1, 2, \dots, r; j = 1, 2, \dots, s) \tag{7.33}$$

则如果所有 $r_{ij} = 0$, 就有

$$\mu_{ij} = \mu + \alpha_i + \beta_j, \quad (i = 1, 2, \dots, r; j = 1, 2, \dots, s) \tag{7.34}$$

此时容易看出, 因素 A 和 B 的影响是相互独立的, 即如果在 B 的某一水平 B_{j_0} 下, A 的一个水平 A_{i_0} 的效应 α_{i_0} 为正的常数, 说明此时 A_{i_0} 对变量 X 的影响标志值高于总体平均, 那么在 B 的其它水平 B_j ($j \neq j_0$) 下, A_{i_0} 影响的程度和方向是相同的. 这种 A、B 两个因素的影响的可分离性就是我们在例 7.2 中提到的无“交互作用”, 因此, 我们将 r_{ij} ($i = 1, 2, \dots, r; j = 1, 2, \dots, s$) 称为因素 A 的水平 A_i 与因素 B 的水平 B_j 的“交互作用”的效应.

如果 $r_{ij} = 0$ ($i = 1, 2, \dots, r; j = 1, 2, \dots, s$), 则称为无交互作用, 此时(7.34)式成立; 反之, 如果存在某 i, j 使得 $r_{ij} \neq 0$; 则称为 A, B 对变量 X 有交互作用.

正如对例 7.2 的分析, 是否考虑两个因素的交互作用对样本容量的要求也是不同的, 因此, 我们下面分两种情况来分别予以讨论.

一、无交互作用的情况

表 7-6 是不考虑因素间的“交互作用”的双因素方差分析的典型的样本数据结构. 其中在每一个组合 (A_i, B_j) $i = 1, 2, \dots, r \quad j = 1, 2, \dots, s$ 对应的总体中只随机选取一个观测即可, 也叫无重复数据. 在不引起混淆的前提下, 我

们不妨仍以 X_{ij} 来标记该样本.

根据前面的论述, 此时模型的结构是:

$$X_{ij} = \mu + \mu_i + \mu_j + \epsilon_{ij}, (i = 1, 2, \dots, r; j = 1, 2, \dots, s);$$
$$\epsilon_{ij} \sim N(0, \sigma^2) \text{ 且相互独立};$$
$$\sum_{i=1}^r \mu_i = \sum_{j=1}^s \mu_j = 0$$

(7.35)

这样, 在不考虑“交互作用”时研究两因素 A 和 B 对 X_{ij} 的影响问题就变成了基于模型(7.35)对假设(7.31)和(7.32)进行检验的问题.

与前面一节相似, 在进行检验之前, 也可以对(7.35)进行参数估计. 在此省略有关步骤, 只写出部分参数的极大似然估计量. 其中 $\mu, \mu_i, \mu_j, \epsilon_{ij} (i = 1, 2, \dots, r; j = 1, 2, \dots, s)$ 的最大似然估计分别是

$$\mu = \frac{1}{rs} \sum_{i=1}^r \sum_{j=1}^s X_{ij} \stackrel{\text{def}}{=} \bar{X},$$

(7.36)

表 7-6 无交互作用时双因素方差分析样本数据典型结构

B 的水平 A 的水平	B ₁	B ₂	...	B _s	平均
A ₁	X ₁₁	X ₁₂	...	X _{1s}	$\bar{X}_{1\cdot}$
A ₂	X ₂₁	X ₂₂	...	X _{2s}	$\bar{X}_{2\cdot}$
...
A _r	X _{r1}	X _{r2}	...	X _{rs}	$\bar{X}_{r\cdot}$
平均	$\bar{X}_{\cdot 1}$	$\bar{X}_{\cdot 2}$...	$\bar{X}_{\cdot s}$	\bar{X}

$$\mu_{i\cdot} = \frac{1}{s} \sum_{j=1}^s X_{ij} \stackrel{\text{def}}{=} \bar{X}_{i\cdot},$$

(7.37)

$$\mu_{\cdot j} = \frac{1}{r} \sum_{i=1}^r X_{ij} \stackrel{\text{def}}{=} \bar{X}_{\cdot j},$$

(7.38)

$$\mu_i = \bar{X}_{i\cdot} - \bar{X},$$

(7.39)

$$\mu_j = \bar{X}_{\cdot j} - \bar{X}.$$

(7.40)

同样地, 寻求对零假设 (7.31) 或 (7.32) 进行检验的统计量依赖于对样本方差或离差平方和的分解, 以及由此进行的各个因素对总的离差的贡献程度的分析.

总的离差平方和 Q 可以进行如下形式的分解:

$$Q = \sum_{i=1}^r \sum_{j=1}^s (X_{ij} - \bar{X})^2$$
$$= \sum_{i=1}^r \sum_{j=1}^s [(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})$$
$$+ (\bar{X}_{i\cdot} - \bar{X}) + (\bar{X}_{\cdot j} - \bar{X})]^2$$

$$\begin{aligned}
 &= \sum_{i=1}^r \sum_{j=1}^s (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2 + \\
 &\quad \sum_{i=1}^r s \cdot (\bar{X}_{i\cdot} - \bar{X})^2 + \sum_{j=1}^s r \cdot (\bar{X}_{\cdot j} - \bar{X})^2 \\
 &= Q_1 + Q_2 + Q_3,
 \end{aligned} \tag{7.41}$$

其中 Q_1, Q_2, Q_3 分别表示等号右端的三个平方和. Q_2 反映的是因素 A 的影响对总离差的贡献, 称之为 A 的离差平方和; Q_3 反映了因素 B 的影响对总离差的贡献, 称之为 B 的离差平方和; Q_1 则称为误差平方和, 它反映了 A、B 因素之外的随机性因素导致的样本之间差异性.

因此, 针对原假设 H_{01} (7.31) 和 H_{02} (7.32) 可以寻求分别满足条件

$$Q_2/Q_1 \leq c_1 \tag{7.42}$$

和

$$Q_3/Q_1 \leq c_2 \tag{7.43}$$

的样本集合作为各自的否定域 C_A, C_B , 其中 c_1, c_2 为待定的正常数.

类似于单因素情形, 可以证明 $\frac{Q_1}{2} \sim \chi^2((r-1)(s-1))$; 在 H_{01} 成立时, $Q_2/\frac{Q_1}{2} \sim \chi^2(r-1)$; H_{02} 成立时, $Q_3/\frac{Q_1}{2} \sim \chi^2(s-1)$, 如果 H_{01} 与 H_{02} 同时成立, 则 $Q/\frac{Q_1}{2} \sim \chi^2(rs-1)$.

同样, 由柯赫伦分解定理可知 H_{01}, H_{02} 同时成立时, Q_1, Q_2, Q_3 是相互独立的, 在 H_{01} 成立时, Q_2 与 Q_1 独立; 在 H_{02} 成立时, Q_3 与 Q_1 独立.

所以, 在 H_{01} 成立时统计量

$$F_A = \frac{Q_2/(r-1)}{Q_1/((r-1)(s-1))} = \frac{Q_2}{Q_1} \cdot (s-1) \tag{7.44}$$

服从自由度为 $r-1$ 和 $(r-1) \cdot (s-1)$ 的 F 分布; 在 H_{02} 成立时, 统计量

$$F_B = \frac{Q_3/(s-1)}{Q_1/((r-1)(s-1))} = \frac{Q_3}{Q_1} \cdot (r-1) \tag{7.45}$$

服从自由度为 $s-1$ 和 $(r-1) \cdot (s-1)$ 的 F 分布.

这样, 在给定显著水平 α 时, 可以通过查表分别获得水平 α 的上侧分位数 $F_A(r-1, (r-1)(s-1))$ 和水平 α 的上侧分位数 $F_B(s-1, (r-1)(s-1))$. 然后, 构造 H_{01} 的否定域

$$C_A = \{x: F_A = \frac{Q_2/(r-1)}{Q_1/((r-1)(s-1))} \geq F_A(r-1, (r-1)(s-1))\}$$

或表示为:

$$C_A = \{x: Q_2/Q_1 \geq \frac{1}{s-1} \cdot F_A(r-1, (r-1)(s-1))\}$$

构造 H_{02} 的否定域

$$C_B = \{x: F_B = \frac{Q_3/(s-1)}{Q_1/(r-1)(s-1)} \cdot F_B(s-1, (r-1)(s-1))\}$$

或表为

$$C_B = \{x: Q_3/Q_1 \leq \frac{1}{r-1} \cdot F_B(s-1, (r-1)(s-1))\}$$

即可得到显著水平为 α 的检验.

与单因素方差分析相似, 也可以通过形如表 7-7 所示的方差分析表将上面的思路表达得更为直观和简洁.

表 7-7 双因素方差分析表(无交互作用)

方差来源	平方和	自由度	均方和	F 值
因素 A	$Q_2 = \sum_{i=1}^r s \cdot (\bar{X}_{i \cdot} - \bar{X})^2$	$r-1$	$\frac{Q_2}{r-1}$	$F_A = \frac{Q_2/(r-1)}{Q_1/(r-1)(s-1)}$
因素 B	$Q_3 = \sum_{j=1}^s r \cdot (\bar{X}_{\cdot j} - \bar{X})^2$	$s-1$	$\frac{Q_3}{s-1}$	$F_B = \frac{Q_3/(s-1)}{Q_1/(r-1)(s-1)}$
误差	$Q_1 = \sum_{i=1}^r \sum_{j=1}^s (\bar{X}_{ij} - \bar{X}_{i \cdot} - \bar{X}_{\cdot j} + \bar{X})^2$	$(r-1)(s-1)$	$\frac{Q_1}{(r-1)(s-1)}$	
总和	$Q = \sum_{i=1}^r \sum_{j=1}^s (\bar{X}_{ij} - \bar{X})^2$	$rs-1$		

作为以上方法的一个应用, 我们考察例 7.2 的一个特殊情况, 即不考虑学校与专业两个因素之间可能的“交互作用”, 此时, 可以简化样本的采集过程, 使用如表 7-6 所示的无重复样本结构. 具体来说, 就是对应每个学校的每个专业各自抽取抽取一个学生进行调查即可. 假设得到的结果如表 7-8 所示.

我们可以运用表 7-7 中列举的公式分别计算出各个平方和、均方和以及 F_A 和 F_B , 然后按给定的显著水平 α 通过查表得到上侧分位数 F_{α} 和 F_{α} , 对比 F_A 与 F_{α} 以及 F_B 与 F_{α} 即得结论.

应用统计软件可以更为方便地得到上述结果, 比如使用 Excel97 提供的“数据分析”工具中的“方差分析: 无重复双因素分析”功能, 针对表 7-8 提供的数据可以直接输出方差分析表, 见表 7-9, 设定显著水平为 0.05.

表 7-8 MBA 毕业生工作第一年收入调查表 (无重复样本)

<div>大学</div> <div>专业</div>	B ₁	B ₂	B ₃	B ₄	B ₅	平均
A ₁	9.6	6.8	11	7.5	4.5	7.88
A ₂	7.8	8.1	5.7	9.8	3.8	7.04
A ₃	11.3	4.2	8.2	7.4	6.3	7.48
A ₄	9.5	6.4	7.5	12	5.2	8.12
平均	9.55	6.375	8.1	9.175	4.95	7.63

表 7-9 中的“行”即因素 A,“列”即因素 B (与数据输入方式有关,这里按表 7-8 相同的方式输入.); 其余各项与表 7-5 的说明相同.

表 7-9 表-8 中数据的方差分析表

差异源	SS	df	MS	F	P—Value	F Crit
行	3.366	3	1.122	0.314	0.815261	3.4903
列	60.207	4	15.052	4.207	0.023416	3.25916
误差	42.929	12	3.577			
总计	106.502	19				

在表 7-9 中,无论是通过 F 值与分位数值进行比较,还是通过 F 的 P 值直接与显著水平 对比,都可以得到结论:在给定了显著水平为 0.05 的情况下,可以拒绝零假设 H_{02} ,即认为大学的选择这一因素对 MBA 毕业生工作第一年的平均收入是有影响的,但是不能拒绝零假设 H_{01} ,从而认为专业的影响并不显著.

二、有交互作用的情况

如果考虑两个因素之间的“交互作用”,或者需要对这种交互作用的影响的存在性进行考察时,就必须针对每一个组合 (A_i, B_j) ($i=1, 2, \dots, r; j=1, 2, \dots, s$) 所对应的总体中抽取多于一个的样本,通常又称之为有重复样本,只有这样才能够正确地说明组合 (A_i, B_j) 对应总体均值同其它组合的差异是否是源于“交互作用”这种系统的影响,而不至于仅仅被解释成随机误差的作用. 下面假设对应每一组合重复抽取 m ($m \geq 2$) 次,即数据结构是均衡的,记所得的样本为 X_{ijk} , ($i=1, 2, \dots, r; j=1, 2, \dots, s; k=1, 2, \dots, m$).

我们不仅要检验因素 A、B 各自是否影响到 X 的取值,即对原假设 (7.31) 和 (7.32) 进行检验,而且还要检验两个因素之间对于 X 是否存在着交

互作用, 即对以下零假设进行检验:

$$H_{03}: r_{ij} = 0. \quad (i = 1, 2, \dots, r; j = 1, 2, \dots, s.) \quad (7.46)$$

其中 r_{ij} 由 (7.33) 定义.

检验的思路依然是通过离差平方和的分解来寻找各自的检验统计量, 理论与前两种情形类似, 故略去.

为了叙述的方便, 先引入以下记号, 其中 i, j 分别取 $i = 1, 2, \dots, r$ 和 $j = 1, 2, \dots, s$ 中的任一值.

$$\begin{aligned} \bar{X}_{ij \cdot} &= \frac{1}{m} \sum_{k=1}^m X_{ijk}, \\ \bar{X}_{i \cdot \cdot} &= \frac{1}{s} \sum_{j=1}^s \bar{X}_{ij \cdot}, \\ \bar{X}_{\cdot j \cdot} &= \frac{1}{r} \sum_{i=1}^r \bar{X}_{ij \cdot}, \\ \bar{X} &= \frac{1}{r} \sum_{i=1}^r \bar{X}_{i \cdot \cdot} = \frac{1}{s} \sum_{j=1}^s \bar{X}_{\cdot j \cdot} = \frac{1}{rs} \sum_{i=1}^r \sum_{j=1}^s \bar{X}_{ij \cdot} = \frac{1}{rsm} \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^m X_{ijk}. \end{aligned} \quad (7.47)$$

相应的因素 A, B 的各个水平的效应以及其交互作用的 $\alpha_i, \beta_j, \gamma_{ij}, i = 1, 2, \dots, r; j = 1, 2, \dots, s$ 的极大似然估计运用上述记号可分别表示为:

$$\begin{aligned} \alpha_i &= \bar{X}_{i \cdot \cdot} - \bar{X}, \quad i = 1, 2, \dots, r, \\ \beta_j &= \bar{X}_{\cdot j \cdot} - \bar{X}, \quad j = 1, 2, \dots, s, \\ \gamma_{ij} &= \bar{X}_{ij \cdot} - \bar{X}_{i \cdot \cdot} - \bar{X}_{\cdot j \cdot} + \bar{X}, \quad i = 1, 2, \dots, r, j = 1, 2, \dots, s. \end{aligned} \quad (7.48)$$

将总的离差平方和进行如下形式的分解:

$$\begin{aligned} Q &= \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^m (X_{ijk} - \bar{X})^2 \\ &= \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^m [(X_{ijk} - \bar{X}_{ij \cdot}) \\ &\quad + (\bar{X}_{ij \cdot} - \bar{X}_{i \cdot \cdot} - \bar{X}_{\cdot j \cdot} + \bar{X}) \\ &\quad + (\bar{X}_{i \cdot \cdot} - \bar{X}) + (\bar{X}_{\cdot j \cdot} - \bar{X})]^2 \\ &= \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^m (X_{ijk} - \bar{X}_{ij \cdot})^2 \\ &\quad + \sum_{i=1}^r \sum_{j=1}^s m [\bar{X}_{ij \cdot} - \bar{X}_{i \cdot \cdot} - \bar{X}_{\cdot j \cdot} + \bar{X}]^2 \\ &\quad + \sum_{i=1}^r sm [\bar{X}_{i \cdot \cdot} - \bar{X}]^2 \end{aligned}$$

$$\begin{aligned}
 & + \sum_{j=1}^s r m_j (\bar{X}_{i\cdot j\cdot} - \bar{X})^2 \\
 & = Q_1 + Q_2 + Q_3 + Q_4,
 \end{aligned} \tag{7.49}$$

即将上式第三个等号右边的四个平方和分别记作 Q_1 、 Q_2 、 Q_3 、 Q_4 。其中 Q_3 、 Q_4 和 Q_2 分别体现了因素 A、B 以及它们之间的交互作用的影响对总离差的贡献，而 Q_1 则是随机误差的影响程度的反映，所以，针对原假设 H_{01} 、 H_{02} 、 H_{03} 的否定域 W_1 、 W_2 、 W_3 可以各自设置为分别满足条件

$$\begin{aligned}
 Q_3/Q_1 & \leq c_1, \\
 Q_4/Q_1 & \leq c_2, \\
 Q_2/Q_1 & \leq c_3
 \end{aligned}$$

的样本集合，其中 c_1 、 c_2 、 c_3 是待定的正的常数。

可以证明无论原假设是否成立，总有

$$\frac{Q_1}{2} \sim \chi^2(rs(m-1)) \tag{7.50}$$

如果 H_{01} （即（7.31）式）成立，则有

$$\frac{Q_3}{2} \sim \chi^2(r-1); \tag{7.51}$$

如果 H_{02} （即（7.32）式）成立，则有

$$\frac{Q_4}{2} \sim \chi^2(s-1);$$

如果 H_{03} （即（7.46）式）成立，则有

$$\frac{Q_2}{2} \sim \chi^2((r-1)(s-1)).$$

而在 H_{01} 、 H_{02} 、 H_{03} 同时成立时，

$$\frac{Q}{2} \sim \chi^2(rsm-1)$$

并且此时 Q_1 、 Q_2 、 Q_3 、 Q_4 是独立的。

进一步地，在 H_{01} 成立时，以下统计量

$$F_A = \frac{Q_3/(r-1)}{Q_1/rs(m-1)}$$

服从自由度为 $r-1$ 和 $rs(m-1)$ 的 F 分布，在给定的显著水平 α 下，查表获得上侧分位数 $F_A(r-1, rs(m-1))$ ，则 H_{01} 的否定域 C_A 可设置

$$C_A = \{x: F_A = \frac{Q_3/(r-1)}{Q_1/rs(m-1)} \geq F_A(r-1, rs(m-1))\},$$

或表为

$$C_A = \{x: Q_3/Q_1 \geq \frac{r-1}{rs(m-1)} \cdot F_A(r-1, rs(m-1))\}$$

这样就得到了 H_{01} 的一个显著水平为 α 的检验.

类似地, 在 H_{01} 成立时, 定义检验统计量

$$F_B = \frac{Q_4/(s-1)}{Q_1/rs(m-1)},$$

则 F_B 服从自由度为 $s-1$ 和 $rs(m-1)$ 的 F 分布, 从而, 可以将 H_{02} 的否定域 C_B 设置为

$$C_B = \{x: F_B = \frac{Q_4/(s-1)}{Q_1/rs(m-1)} \geq F_B(s-1, rs(m-1))\}$$

或表为:

$$C_B = \{x: Q_4/Q_1 \geq \frac{s-1}{rs(m-1)} \cdot F_B(s-1, rs(m-1))\}$$

在 H_{03} 成立时, 统计量

$$F_{AB} = \frac{Q_2/(r-1)(s-1)}{Q_1/rs(m-1)}$$

服从自由度为 $(r-1) \cdot (s-1)$ 和 $rs(m-1)$ 的 F 分布, 令 F_{AB} 表示该分布的上侧分位数, 同样地, 可以将 H_{03} 的否定域 C_{AB} 设置为

$$C_{AB} = \{x: F_{AB} = \frac{Q_2/(r-1)(s-1)}{Q_1/rs(m-1)} \geq F_{AB}((r-1)(s-1), rs(m-1))\}$$

或表为:

$$C_{AB} = \{x: Q_2/Q_1 \geq \frac{(r-1)(s-1)}{rs(m-1)} \cdot F_{AB}((r-1)(s-1), rs(m-1))\}.$$

这样就得到了 H_{03} 的一个显著水平为 α 的检验.

与前面相似, 我们使用方差分析表将上述步骤总结整理一下, 采用的数据为对应每个组合都有 $m(m-2)$ 个重复观测的均衡结构, 方差分析表的形式如表 7-10 所示.

根据上面各式逐项计算出方差分析表中相应的值, 查表获得临界值 F_A 、 F_B 、 F_{AB} , 再与三个 F 值进行对比即可获得结论.

做为例子, 我们使用 Excel97 提供的“数据分析”工具中的“方差分析: 可重复双因素分析”功能对表 7-2 给出的数据进行了分析. 表 7-11 是输出的方差分析表, 其中设定的显著水平为 0.05.

表 7-11 中第一列各项目分别对应了表 7-10 第一列中相应的项目, 即“样本”此处就表示“因素 A”, “列”表示“因素 B”, 以此类推, 其余各列的内容在前面已有介绍.

表 7-10 双因素方差分析表(m 次重复)

方差来源	平方和	自由度	均方和	F 值
因素 A	$Q_3 = \sum_{i=1}^r sm(\bar{X}_{i..} - \bar{X})^2$	$r - 1$	$\frac{Q_3}{r - 1}$	$F_A = \frac{Q_3 / (r - 1)}{Q_1 / rs(m - 1)}$
因素 B	$Q_4 = \sum_{j=1}^s rm(\bar{X}_{.j.} - \bar{X})^2$	$s - 1$	$\frac{Q_4}{s - 1}$	$F_B = \frac{Q_4 / (s - 1)}{Q_1 / rs(m - 1)}$
A 与 B 的交互作用	$Q_2 = \sum_{i=1}^r \sum_{j=1}^s m(\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X})^2$	$(r - 1)(s - 1)$	$\frac{Q_2}{(r - 1)(s - 1)}$	$F_{AB} = \frac{Q_2 / (r - 1)(s - 1)}{Q_1 / rs(m - 1)}$
误差	$Q_1 = \sum_{i=1}^r \sum_{j=1}^s \sum_{l=1}^m (\bar{X}_{ijk} - \bar{X}_{ij.})^2$	$rs(m - 1)$	$\frac{Q_1}{rs(m - 1)}$	
总和	$Q = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^m (\bar{X}_{ijk} - \bar{X})^2$	$rs m - 1$		

表 7-11 表 7-2 中数据的方差分析表

差异源	SS	df	MS	F	P -Value	F -Crit
行	34.22067	3	11.40689	0.887293	0.455956	2.838746
列	135.3573	4	33.83933	2.632216	0.048271	2.605972
交互	115.336	12	9.61133	0.747624	0.69756	2.003461
内部	514.2333	40	12.8558			
总计	799.1473	59				

根据表 7-11 中 F 值与临界值或者相应地 F 的 P 值与显著水平 0.05 之间的对比, 易知在给定显著水平为 0.05 的情况下不能够拒绝原假设 H_{01} 及 H_{03} , 即不能认为专业对 MBA 毕业生第一年的收入有显著影响, 不能认为专业和大学两个因素对学生收入有“交互作用”的影响. 但是, 由于 $F_B = 2.632216 > F_B = 2.605972$, 或者由于 F_B 的 P 值 $0.048271 <$ 给定的显著水平 0.05, 因此, 在 0.05 的显著水平下能够拒绝原假设 H_{02} , 从而认为大学的选择对 MBA 毕业学生第一年的工作收入有显著影响.

由公式(7.54)可以求得五所大学各自的效应的估计值为:

$$_1 = \overline{X}_{.1.} - \overline{X} = 9.50 - 8.36 = 1.14$$
$$_2 = \overline{X}_{.2.} - \overline{X} = 8.16 - 8.36 = -0.20$$
$$_3 = \overline{X}_{.3.} - \overline{X} = 9.38 - 8.36 = 1.02$$
$$_4 = \overline{X}_{.4.} - \overline{X} = 9.24 - 8.36 = 0.88$$
$$_5 = \overline{X}_{.5.} - \overline{X} = 5.51 - 8.36 = -2.85$$

这样不难看出,五所大学中第一所 B₁ 对 MBA 毕业生工作第一年的收入效应最大,因此,仅从这一方面来看,报考 B₁ 大学的 MBA 是考生的一个较好的选择.

习 题 七

(A)

1. 在三所小学的五年级男生中随机抽取了 6 名学生,测得他们的身高数据如下表所示:

小学	身高数据(cm)					
甲	128.1	134.1	133.1	138.9	140.8	127.4
乙	150.3	147.9	136.8	126.0	150.7	155.8
丙	140.6	143.1	144.5	143.7	148.5	146.4

对上述数据进行方差分析,并判断三所小学五年级男生的平均身高是否有显著差异(取显著水平 $\alpha = 0.05$)

2. 为了对比三种不同类型汽车的耗油量,分别在每种汽车中抽取 5 辆并各自行驶 500 英里,对每辆汽车计算得到每加仑汽油能够行驶的英里数如下表所示

A 型	19	21	20	19	21
B 型	19	20	22	21	23
C 型	24	26	23	25	27

试分析三种类型的汽车平均每加仑汽油所能够行驶的里程是否有显著不同($\alpha = 0.05$).

3. 某公司销售部经理想要确定 5 种推销方法中哪一种对它们的产品最有效,为此他随机选择了 35 个地区,每 7 个实施同一种推销方法,结果每种方法的平均销售水平及方差如下表所示:

推销方法	平均销售额	销售额的方差
1	80	28
2	86	30
3	72	30
4	76	25
5	85	20

给定显著水平为 0.05,该经理能否得出这五种方法的平均销售水平一致的结论?

4. 下表中给出了三位操作工人分别在四台不同机器上操作三天的日产量.

工人 机器	甲			乙			丙		
A ₁	15	15	17	19	19	16	16	18	21
A ₂	17	17	17	15	15	15	19	22	22
A ₃	15	17	16	18	17	16	18	18	18
A ₄	18	20	22	15	16	17	17	17	17

试在显著性水平 $\alpha = 0.05$ 下检验:

- (1) 操作工之间的差异是否显著?
- (2) 机器之间的差异是否显著?
- (3) 操作工与机器之间的交互作用是否显著?

习 题 七

(B)

1. 某同学考虑对例 7.1 中的问题检验如下假设

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

他采取的措施是对四个专业分组两两进行两总体等方差假设的 t 检验. 如果他进行 t 检验时采用的显著水平是 $\alpha = 0.05$, 试求根据他的方式作为对 H_0 的检验时犯第一类错误的概率是多少?

2. 试证明按照文中(7.14)式定义的 $\bar{y}_{i\cdot}$ $i= 1, 2, \dots, r$ 分别是因素 A 的各个水平的效应 α_i $i= 1, 2, \dots, r$ 的无偏估计.

3. 验证文中(7.15)式成立.

4. 在单因素方差分析问题中, 如果因素 A 只含有两个水平, 试证明此时使用 F 检验与前面一章中介绍的两总体等方差假设的 t 检验是等价的.

第 8 章

回 归 分 析

回归分析是研究两个或两个以上变量之间的相互关系的一种重要的统计方法。与变量之间的函数关系不同，这种关系描述的是变量之间相互依存相伴发生的性质，它不是一种确定性的关系。比如不能根据对某种商品的给定的一种广告投入来完全确定该种商品的销售额，但是确实增加广告投入往往可以提高销售额，因此这两者之间确实存在一定的关联性。回归分析通过建立统计模型来研究这种关系，并由此对相应的变量进行预测和控制。回归分析具有广泛的应用基础，在经济学中，回归分析是进行经济计量分析的主要工具。

本章介绍线性回归模型的估计、检验以及相应的预测和控制等问题，我们将重点讨论一元线性回归模型。

§ 8.1 一元线性回归模型及其参数估计

一、回归模型

无论是在物理、生物等自然科学还是在经济、管理等社会学科中，都要研究变量与变量的关系问题，通常这种关系有两种形式：

一种是我们熟知的函数关系，比如匀速运动物体运动的距离 s 、速度 v 和运动的时间 t 之间具有明确的关系

$$s = v \cdot t,$$

任意给定 v 和 t 的值， s 也被完全确定。这种确定性是函数关系的重要特征。

另一种关系则无法通过明确的函数关系来表达，我们可以先看下面的例子。

例 8.1 某广告公司为了研究某一类产品的广告费用与其销售额之间的关系，对多个厂家进行了调查，获得数据资料如表 8-1 所示。

表 8-1 广告投入与销售额资料（单位：万元）

厂 家	1	2	3	4	5	6	7	8	9	10
广告费	35	60	25	30	35	40	25	20	50	45
销售额	440	520	380	475	385	525	450	365	540	50

以下为了分析的方便我们用 X 表示广告费用、用 Y 表示销售额. 从表 8-1 中不难看出, X 与 Y 之间不可能存在一个明确的函数关系式, 事实上, 即便不同厂家投入了相同的广告费用, 其各自的销售额也不会是完全相同的, 影响销售额的因素是多种多样的, 除了广告投放的影响, 还与厂家产品的特色、定价、销售渠道、售后服务以及其它一些偶然性因素的影响, 因此尽管厂家 1 和 5 以及厂家 3 和 7 广告投入分别相同, 其销售额却是不同的.

但是, 也不能就此认为这两者之间没有什么关系. 为了更加清楚地看清其中的规律, 通常采用“散点图”的方式将各个厂家的广告费用和销售额成对数据 $(X_i, Y_i), i=1, 2, \dots, 10$ 在平面直角坐标系中用点表示出来, 见图 8-1 所示.

在图 8-1 中可以看出, 随着广告投入费用的增加, 销售额基本上也呈上升趋势, 图中的点大致上分布在一条向右上方倾斜的直线附近, 高投入的广告费用对应的是一组比较高的销售额, 伴随比较低的广告投入结果也是一组比较低的销售额.

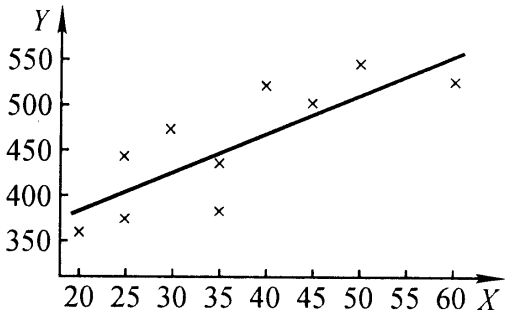


图 8-1 广告投入与销售额的散点图

统计学上将类似于上述广告费用与销售额之间这种不具有确定函数关系的两个变量之间的统计关系称为相关关系.

研究两个变量之间的相关关系主要从两个方向进行: 一个方向是进行相关分析, 即通过引入一定的统计指标来量化变量之间相关的程度, 此时两个变量的地位是对等的, 没有方向上的差异. 比如相关系数就是这样一个指标.

人们往往通过对 (X, Y) 的一个观察样本 $(X_i, Y_i), i=1, 2, \dots, n$ 来对 X 与 Y 的相关系数作出估计, 常用的估计是所谓的样本相关系数

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2 \right]^{1/2}} \tag{8.1}$$

可以证明, r 是相关系数 的渐近无偏估计.

另一个研究变量间相关关系的方向就是回归分析.

回归分析是如何对变量间的相关关系进行研究的呢? 我们回到前面关于广告费用与销售额的例子. 从图 8-1 和前面的分析, 可以将销售额的观察结果 y 看成是两部分叠加而成, 一部分由广告投入的线性函数引起, 记作 $\beta_0 + \beta_1 X$, 另一部分是随机因素引起, 记作 ϵ . 即:

$$y = \beta_0 + \beta_1 X + \epsilon$$

相应地, 变量 Y 与变量 X 之间的关系可表示为:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (8.2)$$

其中 Y 是随机变量, X 是一个普通变量, 对它可加以控制. ε 是一随机变量, 它可看成随机误差, 因而可以假定 $E[\varepsilon] = 0$. 于是, 在给定 X 的取值时, 有

$$E[Y|X] = \beta_0 + \beta_1 X$$

对本例而言回归分析的主要任务是确定 (8.2) 中的线性函数 $\beta_0 + \beta_1 X$ 的系数 β_0 和 β_1 . 一旦通过估计和检验得到一个合适的线性函数 $\beta_0 + \beta_1 X$, 便可利用 (8.2) 对 Y 进行预测和实施控制. 例 8-1 所简述的变量 X 和 Y 是一种特别简单的关系, 但它在实际中却具有广泛的应用, 在对它进行深入讨论之前, 我们先对回归分析作出一般的阐述.

正如前面所描述的, 回归分析是寻求一个随机变量 Y 对另一个或一组 (随机或非随机) 变量 X_1, \dots, X_m 的相关性的一种统计方法, 它主要通过对变量的观察所获得的统计数据来确定反映变量间关系的经验公式, 并通过所得公式进行统计描述, 分析和推断, 进而解决预测、控制和优化问题.

用来进行回归分析的数学模型通常有如下一般形式:

$$Y = f(X_1, X_2, \dots, X_m) + \varepsilon \quad (8.3)$$

其中 $f(X_1, \dots, X_m)$ 是 X_1, \dots, X_m 的一个确定的函数, 这一函数, 通常称为 Y 对 X_1, \dots, X_m 的回归函数. ε 是数学期望为 0 的随机变量, 称为随机误差. 方程 $y = f(x_1, x_2, \dots, x_m)$ 称作回归方程, 自变量 X_1, X_2, \dots, X_m 称为“回归变量”或“解释变量”, 在预测问题中, 也称之为“预报因子”; 因变量 Y 称作“响应变量”, “被解释变量”, 在预测问题中也称之为“预测量”.

用来进行回归分析的数学模型 (含有关假设) 称为回归模型. 只含有一个回归变量的回归模型称为一元回归模型, 否则称为多元回归模型. 在应用回归分析时, 一般假设回归函数 f 的形式是已知的, 但含有未知参数. 最简单、最重要的且最便于处理, 因而也最常用的情形是函数 f 关于未知参数 $\beta_0, \beta_1, \dots, \beta_k$ 是线性的情形, 这样的回归模型称为线性回归模型, 否则称为非线性回归模型. 许多非线性回归模型经过变换可化为线性回归模型. 因此, 我们重点介绍线性回归模型.

二、一元线性回归模型

在所有回归模型中, 最简单的情形是两个变量间的线性回归模型. 这一模型中回归函数是单个解释变量的、系数未知的线性函数, 即有形如:

$$f(X) = \beta_0 + \beta_1 X \quad (8.4)$$

的形式, 其中 β_0, β_1 为待定的参数, 此时的回归方程表现为一条直线.

除此之外，为了处理问题的方便，通常我们只考虑 X 为可控制变量，即 X 不是一个随机变量。这一假定在大多数实际情况下是合理的，比如例 8.1 中的广告费用的确是一个人们可以控制的确定的量。在今后我们用 x 代替 X 以示它为一确定的量。 n 组样本则以 $(x_i, Y_i) \ i=1, 2, \dots, n$ 来表示。

根据样本抽取原则，可以设 Y_1, Y_2, \dots, Y_n 是相互独立的，这意味着对任意 i, Y_i 的取值不依赖于其它 Y_j 的值 ($j \neq i, i=1, 2, \dots, n$)。由 (8.4) 式易知

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ i=1, 2, \dots, n, \tag{8.5}$$

其中 $\varepsilon_i \ (i=1, 2, \dots, n)$ 是 n 个独立的随机变量，期望 $E(\varepsilon_i) = 0 \ (i=1, 2, \dots, n)$ 。

这里的 $\varepsilon_i \ (i=1, 2, \dots, n)$ 又可以理解为 Y_i 偏离回归直线的大小，通常还假定这种偏离的程度不会随 x_i 的不同而发生改变，即假定

$$E(\varepsilon_i^2) = D(\varepsilon_i) = \sigma^2, \ i=1, 2, \dots, n \tag{8.6}$$

其中 σ^2 是待定的正常数，上式又被称为同方差 (Homoscedasticity) 假设。

至此，我们给出了在经典统计中对一元线性回归模型做出的几个基本假设。现在总结如下：

- (H₁) 回归函数是自变量 x 的一次线性函数；
- (H₂) 自变量 x 被看作确定变量；
- (H₃) n 个样本 Y_1, Y_2, \dots, Y_n 是独立的；
- (H₄) (8.5) 式中的 ε_i 满足同方差条件，即 $D(\varepsilon_i) = \sigma^2 \ i=1, 2, \dots, n$ ，从而 Y_1, Y_2, \dots, Y_n 的方差也相等，即

$$D(Y_i) = \sigma^2, \ i=1, 2, \dots, n \tag{8.7}$$

一元线性回归分析所考虑的主要问题就是在以上基本假设下对模型 (8.5) 中的参数 β_0, β_1 和 σ^2 进行统计推断。

三、最小二乘估计

对 β_0, β_1 的估计实际上就是在平面直角坐标系中“估计”一条直线

$$Y = \beta_0 + \beta_1 X \tag{8.8}$$

使得它尽可能地接近回归直线 $Y = \beta_0 + \beta_1 X$ 。如图 8-2 所示。

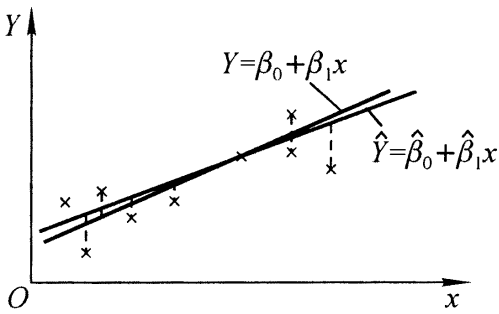


图 8-2 估计的回归直线

一种比较直观的思路是希望观测得到的样本点 (x_i, Y_i) 与估计直线上相应的点 (x_i, \hat{Y}_i) , $i=1, 2, \dots, n$ 尽可能地接近。或者说使它们每一对点之间的距离都尽可能地短，从而要求直线 (8.8) 使得以下平方和达到最小值

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 \quad (8.9)$$

即

$$Q(\beta_0, \beta_1) = \min_{\beta_0, \beta_1} Q(\beta_0, \beta_1) \quad (8.10)$$

这里的 $Q(\beta_0, \beta_1)$ 是一个非负的二元函数, 根据多元函数微分学的知识, β_0 、 β_1 应该满足

$$\left. \frac{\partial Q}{\partial \beta_0} \right|_{\beta_0 = \hat{\beta}_0, \beta_1 = \hat{\beta}_1} = \left. \frac{\partial Q}{\partial \beta_1} \right|_{\beta_0 = \hat{\beta}_0, \beta_1 = \hat{\beta}_1} = 0, \quad (8.11)$$

将上式展开, 即得

$$\begin{aligned} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) &= 0; \\ \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) X_i &= 0. \end{aligned} \quad (8.12)$$

经过整理, 得到关于 β_0 和 β_1 的一个方程组

$$\begin{aligned} Y &= \beta_0 + \beta_1 X; \\ \sum_{i=1}^n X_i Y_i &= n \bar{X} \beta_0 + \sum_{i=1}^n X_i^2 \beta_1. \end{aligned} \quad (8.13)$$

其中 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$.

求解上述方程组, 即可得到

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}, \quad (8.14)$$

$$\beta_1 = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2}. \quad (8.15)$$

容易验证, 这样得到的 β_0 、 β_1 的确是 (8.9) 式定义的二元函数 Q 的最小值点. 我们称 β_0 、 β_1 是 β_0 、 β_1 的最小二乘估计 (Least Squares Estimators, 简记作 LSE). 将 β_0 、 β_1 代入 (8.8) 式即可得到估计出来的回归直线方程为

$$\hat{Y} = \bar{Y} + \beta_1 (\hat{X} - \bar{X}). \quad (8.16)$$

这是一条过点 (\bar{X}, \bar{Y}) 和 $(0, \beta_0)$ 的直线.

为了书写简洁, 我们引入记号

$$\begin{aligned} l_{XX} &= \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X}) = \sum_{i=1}^n X_i^2 - n \bar{X}^2, \\ l_{XY} &= \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}, \end{aligned}$$

则 β_1 可以表示成

$$b_1 = \frac{l_{xy}}{l_{xx}} \quad (8.17)$$

对于每一个 x_i ($i = 1, 2, \dots, n$), 称

$$\hat{Y}_i = b_0 + b_1 x_i \quad (8.18)$$

为相应的真实值 Y_i 的回归值或者拟合值, 记

$$e_i = Y_i - \hat{Y}_i, \quad i = 1, 2, \dots, n, \quad (8.19)$$

称 e_i ($i = 1, 2, \dots, n$) 为拟合残差或残差.

任给常数 a 和 b , 由 (8.12) 式容易验证下列简单的性质

$$\sum_{i=1}^n e_i (a x_i + b) = 0. \quad (8.20)$$

特别地, 有 $\sum_{i=1}^n e_i = 0$ 成立.

对于例 8.1 中给出的样本数据, 根据 (8.14) 和 (8.15) 式可以求得 b_0 、 b_1 的最小二乘估计值分别为 $b_0 = 309.5276$, $b_1 = 4.067736$, 由此得到回归直线的方程的估计是

$$\hat{Y} = 309.5276 + 4.067736x.$$

四、 b_0 和 b_1 的性质

最小二乘法得到的 b_0 和 b_1 的估计量是否是无偏的, 其有效性如何? 这些问题的解答将有助于我们对最小二乘估计进行评价.

从 (8.15) 式不难将 b_1 写成如下形式

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})}{l_{xx}} \cdot Y_i, \quad (8.21)$$

即 b_1 是 Y_1, Y_2, \dots, Y_n 的线性组合.

于是,

$$\begin{aligned} E(b_1) &= \sum_{i=1}^n \frac{(x_i - \bar{x})}{l_{xx}} \cdot E(Y_i) \\ &= \sum_{i=1}^n \frac{(x_i - \bar{x})}{l_{xx}} \cdot (b_0 + b_1 \cdot x_i) \\ &= b_0 \cdot \sum_{i=1}^n \frac{(x_i - \bar{x})}{l_{xx}} + b_1 \cdot \sum_{i=1}^n \frac{(x_i - \bar{x}) \cdot x_i}{l_{xx}} \\ &= b_1, \end{aligned} \quad (8.22)$$

因此, b_1 是 b_1 的无偏估计.

考虑到 Y_1, Y_2, \dots, Y_n 之间的独立性以及同方差性, 易知

$$D(b_1) = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{l_{xx}^2} \cdot D(Y_i) = \frac{\sigma^2}{l_{xx}}. \quad (8.23)$$

类似地, 利用 (8.14) 式以及 (8.21) 式也可以将 $\hat{\beta}_0$ 写成 Y_1, Y_2, \dots, Y_n 的线性组合的形式:

$$\hat{\beta}_0 = \sum_{i=1}^n \frac{1}{n} - \frac{(\bar{X}_i - \bar{X})\bar{X}}{l_{xx}} \cdot Y_i. \quad (8.24)$$

同样地可以验证

$$E(\hat{\beta}_0) = \beta_0 \quad (8.25)$$

以及

$$D(\hat{\beta}_0) = \frac{1}{n} + \frac{\bar{X}^2}{l_{xx}} \cdot \sigma^2 \quad (8.26)$$

因此, $\hat{\beta}_0$ 也是 β_0 的无偏估计.

由于 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 都可以表示成样本 Y_1, Y_2, \dots, Y_n 的线性组合的形式, 通常称这种形式的估计为线性估计.

可以证明, 对于最小二乘估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 来说有下面的定理成立.

高斯-马尔可夫定理: 在假设 $(H_1) \sim (H_4)$ 满足的条件下, 模型 (8.5) 中参数 β_0 和 β_1 的最小二乘估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ (分别由 (8.14) 和 (8.15) 给出) 是最佳线性无偏估计量 (Best Linear Unbiased Estimators, 简记作 BLUE).

上述定理中的“最佳”一词的含义为在所有关于 β_0 和 β_1 的线性无偏估计量中 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别是其中方差最小的一个.

另外, 由 (8.8) 式定义的 Y 可以作为 $E(Y|X)$ 的一个估计, 而且由于

$$E(Y) = E(\hat{\beta}_0) + E(\hat{\beta}_1) \cdot X = \beta_0 + \beta_1 X = E(Y|X),$$

可知 Y 是 $E(Y|X)$ 的无偏估计.

由于

$$\begin{aligned} \text{cov}(\hat{\beta}_0, \hat{\beta}_1) &= \text{cov}(\bar{Y} - \bar{X}\hat{\beta}_1, \hat{\beta}_1) \\ &= \text{cov}(\bar{Y}, \hat{\beta}_1) - \text{cov}(\bar{X}\hat{\beta}_1, \hat{\beta}_1) \\ &= \text{cov}\left(\frac{1}{n} \sum_{i=1}^n Y_i, \sum_{i=1}^n \frac{(X_i - \bar{X})}{l_{xx}} \cdot Y_i - \bar{X} \cdot \hat{\beta}_1\right) \\ &= \sum_{i=1}^n \frac{(X_i - \bar{X})}{nl_{xx}} \cdot D(Y_i) - \bar{X} \cdot D(\hat{\beta}_1) \\ &= \sum_{i=1}^n \frac{(X_i - \bar{X})}{nl_{xx}} \cdot \sigma^2 - \bar{X} \cdot \frac{\sigma^2}{l_{xx}} \\ &= -\frac{\bar{X}}{l_{xx}} \cdot \sigma^2, \end{aligned} \quad (8.27)$$

于是, Y 的方差为

$$D(Y) = D(\hat{\beta}_0) + D(\hat{\beta}_1) \cdot X^2 + 2 \cdot \text{cov}(\hat{\beta}_0, \hat{\beta}_1) \cdot X = \frac{1}{n} + \frac{(X - \bar{X})^2}{l_{xx}} \cdot \sigma^2. \quad (8.28)$$

§ 8.2 一元线性回归模型的检验

不难看出, 对于一元线性回归模型而言, 无论变量 x 是否对变量 Y 有无影响, 总可以利用上一节中给出的最小二乘估计公式 (8.14) 和 (8.15), 由给定的一组样本值求出 β_0 和 β_1 的估计值 $\hat{\beta}_0$ 、 $\hat{\beta}_1$ 并得到相应的回归方程. 正如我们对例 9.1 所做的一样. 但是, 是否广告费用真正对销售额有影响和解释作用呢? 如果答案是否定的, 那么上面求解出来的回归方程就是没有意义的. 因此对回归模型的显著性进行检验, 也就是对 x 的变化是否影响到 Y 进行检验, 是非常有必要的.

在一元线性回归模型中, 回归函数 $E(Y|x)$ 是 x 的线性函数, 如果 x 的变化与 Y 无关, 则说明 $E(Y|x)$ 与 x 无关, 即有 $\beta_1 = 0$; 反之, 若 $\beta_1 = 0$, 则回归函数是一个常数, 从而 x 变化对 Y 不产生影响. 因此此时检验变量 x 是否对 Y 有解释作用等价于检验下面假设

$$H_0: \beta_1 = 0. \quad (8.29)$$

能否被拒绝.

为了寻求上述假设的检验统计量, 我们常要对一元线性回归模型 (8.5) 作出另外一个假设条件

(H₅) ϵ_i 服从正态分布, 即 $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \dots, n$.

易知, 在 (H₅) 成立时, Y_1, Y_2, \dots, Y_n 也服从正态分布, 且 $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1, 2, \dots, n$. 另外, 显然 β_0, β_1, Y 也都服从正态分布, 且

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{1}{n} + \frac{\bar{x}^2}{l_{xx}} \cdot \sigma^2\right), \quad (8.30)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{l_{xx}}\right) \quad (8.31)$$

$$Y \sim N\left(\beta_0 + \beta_1 \cdot x, \frac{1}{n} + \frac{(x - \bar{x})^2}{l_{xx}} \cdot \sigma^2\right). \quad (8.32)$$

本章后面的分析都是基于假设条件 (H₁) ~ (H₅) 成立而作出的.

一、方差分析

正如我们在前面一章中介绍的, 分析变量 x 是否对 Y 有影响, 可以通过将 Y 的总离差平方和进行分解来确定诸因素对 Y 的各个样本之间的差异做出的贡献, 事实上, 导致这种差异性的原因无非是变量 x 的变化引起的或者是其他一些随机性因素的作用. 对比这两种因素对 Y 的样本之间差异的贡献是寻找假设 (8.29) 的检验统计量的一个直观思路.

此时, 总的离差平方和为

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2 = l_{YY}. \quad (8.33)$$

可以将其做如下分解.

$$\begin{aligned} SST &= \sum_{i=1}^n [(Y_i - Y_i) + (Y_i - \bar{Y})]^2 \\ &= \sum_{i=1}^n (Y_i - Y_i)^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - Y_i) \cdot (Y_i - \bar{Y}). \end{aligned}$$

由于 (8.20), 上式最后一项

$$\begin{aligned} &\sum_{i=1}^n (Y_i - Y_i) \cdot (Y_i - \bar{Y}) \\ &= \sum_{i=1}^n e_i \cdot 1 \cdot (X_i - \bar{X}) \\ &= 1 \cdot \sum_{i=1}^n e_i (X_i - \bar{X}) \\ &= 0, \end{aligned}$$

因此,

$$\begin{aligned} SST &= \sum_{i=1}^n (Y_i - Y_i)^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2 \\ &= SSE + SSR \end{aligned} \quad (8.34)$$

其中 $SSE = \sum_{i=1}^n (Y_i - Y_i)^2 = \sum_{i=1}^n e_i^2$ 称为残差平方和, $SSR = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 = \hat{\beta}_1^2 \cdot l_{xx}$ 称为回归平方和.

为了能够更加清楚地了解这两个平方和各自的含义, 可以先看一下它们各自的期望值. 其中,

$$\begin{aligned} E(SSR) &= E(\hat{\beta}_1^2 \cdot l_{xx}) = l_{xx} \cdot E(\hat{\beta}_1^2) \\ &= l_{xx} \cdot [D(\hat{\beta}_1) + (E(\hat{\beta}_1))^2] \\ &= l_{xx} \cdot \frac{\sigma^2}{l_{xx}} + \frac{\sigma^2}{1} \\ &= \sigma^2 + \frac{\sigma^2}{1} \cdot l_{xx} \end{aligned} \quad (8.35)$$

可见, 如果 H_0 成立, 则 $E(SSR) = \sigma^2$, 此时 SSR 仅体现了随机误差所引起的差异; 如果 H_0 不成立, 则 $\beta_1 \neq 0$, 此时 $E(SSR) > \sigma^2$, 它同时反映了 x 变化所引起的差异. 所以称之为回归平方和.

对于 SSE 而言, 由于

$$SSE = \sum_{i=1}^n (Y_i - Y_i)^2$$

$$\begin{aligned}
&= \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 X_i - Y_i)^2 \\
&= \sum_{i=1}^n [(\hat{\beta}_0 - Y_i) + (\hat{\beta}_1 - 1)X_i]^2 \\
&= \sum_{i=1}^n [(\hat{\beta}_0 - Y_i)^2 + X_i^2 \cdot (\hat{\beta}_1 - 1)^2 + \hat{\beta}_1^2 + 2 \cdot (\hat{\beta}_0 - Y_i) \cdot X_i + 2X_i(\hat{\beta}_1 - 1) \cdot X_i + 2 \cdot X_i \cdot (\hat{\beta}_0 - Y_i) \cdot (\hat{\beta}_1 - 1)] \\
&= n(\hat{\beta}_0 - \bar{Y})^2 + (\hat{\beta}_1 - 1)^2 \cdot \sum_{i=1}^n X_i^2 + \sum_{i=1}^n \hat{\beta}_1^2 + 2 \sum_{i=1}^n (\hat{\beta}_0 - Y_i) \cdot X_i \\
&\quad + 2 \sum_{i=1}^n X_i \cdot (\hat{\beta}_1 - 1) \cdot X_i + 2 \sum_{i=1}^n X_i \cdot (\hat{\beta}_0 - Y_i) \cdot (\hat{\beta}_1 - 1),
\end{aligned}$$

而且, 由 (8.24) 和 (8.21) 式, 得

$$\begin{aligned}
E[(\hat{\beta}_0 - Y_i) \cdot X_i] &= -E(\hat{\beta}_0 \cdot X_i) \\
&= -\sum_{i=1}^n \frac{1}{n} \cdot \frac{(X_i - \bar{X}) \cdot \bar{X}}{l_{xx}} R(Y_i - \bar{Y}) \\
&= -\frac{1}{n} \cdot \frac{(\bar{X} - \bar{X}) \cdot \bar{X}}{l_{xx}} = 0; \\
E[X_i \cdot (\hat{\beta}_1 - 1) \cdot X_i] &= -X_i \cdot E(\hat{\beta}_1 - 1) \\
&= -X_i \cdot \sum_{j=1}^n \frac{(X_j - \bar{X})}{l_{xx}} \cdot E(Y_j - \bar{Y}) \\
&= -\frac{(\bar{X} - \bar{X})X_i}{l_{xx}} = 0.
\end{aligned}$$

故有

$$\begin{aligned}
&E(SSE) \\
&= nD(\hat{\beta}_0) + D(\hat{\beta}_1) \cdot \sum_{i=1}^n X_i^2 + \sum_{i=1}^n D(Y_i) \\
&\quad - 2 \sum_{i=1}^n \frac{1}{n} \cdot \frac{(X_i - \bar{X}) \cdot \bar{X}}{l_{xx}} \cdot 2 - 2 \sum_{i=1}^n \frac{(X_i - \bar{X}) \cdot X_i}{l_{xx}} \cdot 2 \\
&\quad + 2 \sum_{i=1}^n X_i \cdot \text{cov}(\hat{\beta}_0, \hat{\beta}_1). \\
&= n \frac{1}{n} + \frac{\bar{X}^2}{l_{xx}} + \sum_{i=1}^n \frac{X_i^2}{l_{xx}} + n - 2 + 2 \sum_{i=1}^n \frac{(X_i - \bar{X}) \cdot \bar{X}}{l_{xx}} \\
&\quad - 2 \sum_{i=1}^n \frac{(X_i - \bar{X}) \cdot X_i}{l_{xx}} - 2 \sum_{i=1}^n \frac{X_i \cdot \bar{X}}{l_{xx}} = n - 1 + \frac{\bar{X}^2}{n} + \sum_{i=1}^n \frac{X_i^2}{l_{xx}} + 2 \sum_{i=1}^n \frac{(X_i - \bar{X}) \cdot \bar{X}}{l_{xx}}
\end{aligned}$$

$$= 2 \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2 = 2 \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2 \frac{1}{l_{xx}} = (n-2) \sigma^2. \quad (8.36)$$

可见, SSE 仅反映了随机误差因素所导致的变异, 因此称之为残差平方和. 由上述结果还可以知道

$$E \frac{SSE}{n-2} = \sigma^2, \quad (8.37)$$

即 $\frac{SSE}{n-2}$ 是 σ^2 的无偏估计量.

了解了 SSR 和 SSE 各自的含义后, 可以设想对零假设 H_0 (8.29) 寻找满足条件

$$\frac{SSR}{SSE} \geq c \quad (8.38)$$

的样本集合作为其否定域, 这里的 c 是待定的正常数. 因为 $\beta_1 = 0$ 时, SSR 显然有增大的趋势.

由于 $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ ($i = 1, 2, \dots, n$), 所以在 H_0 成立时 $Y_i \sim N(\beta_0, \sigma^2)$, $i = 1, 2, \dots, n$. 易知此时

$$\frac{SST}{2} \sim \sigma^2(n-1) \quad (8.39)$$

又因为 $\beta_1 \sim N\left(\beta_1, \frac{\sigma^2}{l_{xx}}\right)$, 故在 H_0 成立时 $\beta_1 \sim N\left(0, \frac{\sigma^2}{l_{xx}}\right)$, 于是

$$\frac{SSR}{2} = \frac{\beta_1^2 \cdot l_{xx}}{2} \sim \sigma^2(1). \quad (8.40)$$

另外, $e_i = Y_i - \hat{Y}_i$ ($i = 1, 2, \dots, n$) 显然可以写作 Y_1, Y_2, \dots, Y_n 的线性组合形式. 因此 e_i ($i = 1, 2, \dots, n$) 服从正态分布, SSE 是 n 个正态随机变量的平方和, 考虑到 (8.12) 中的两个约束等式, 可知 $\frac{SSE}{2}$ 服从自由度为 $n-2$ 的 σ^2 分布.

这样, 根据柯赫伦分解定理可知, 在 H_0 成立的条件下, SSR 与 SSE 独立. 因此, 如下定义的统计量 F 在 H_0 成立时服从自由度为 1 和 $n-2$ 的 F 分布

$$F = \frac{SSR}{SSE/(n-2)} \quad (8.41)$$

于是, 给定显著水平 α , 查表获得临界值 $F(1, n-2)$, 即 $F(1, n-2)$ 分布的 α 上侧分位数, H_0 的否定域可取为

$$C = \{(x, y) : F \geq F(1, n-2)\} \quad (8.42)$$

或表为:

$$C= \{(x, y): \frac{SSR}{SSE} \leq \frac{F(1, n-2)}{n-2}\}$$

(8.43)

这样就得到了 H_0 的一个显著水平为 α 的检验.

类似于前面一章, 上述方程也可以用“方差分析表”表达出来. 见表 8-2 所示.

表 8-2 回归方程显著性检验的方差分析表

方差来源	平方和	自由度	均方和	F 值
回归	$SSR = \sum_{i=1}^n (Y_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
残差	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$n - 2$	$MSE = \frac{SSE}{n - 2}$	
总计	$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$n - 1$		

作为例子, 对例 8.1 中给出的广告费用及销售额的数据 (表 8-1) 可以按照上面给出的有关公式分别求得各个平方和, 从而求出 F 值, 与临界值对比即可获得有关结论.

使用统计软件可以方便地获得类似表 8-2 的方差分析表. 表 8-3 是我们针对例 8.1 中的数据使用 Excel97 提供的“数据分析”工具中的“回归”功能项得到的方差分析表, 其中显著水平设为 0.05.

表 8-3 例 8.1 数据的方差分析

	df	SS	MS	F	Significance F
回归分析	1	23206.43	23206.43	12.93417	0.007018793
残 差	8	14353.57	1794.196		
总 计	9	37560			

表 8-3 在结构上同前一章 Excel97 输出的方差分析表略有区别, 但是内容和分析方法是一致的. 表 8-3 中的“Significance F”一项即是前面所谓的 F 的 P 值, 由于这一值 0.007018793 远远小于给定的显著水平 0.05, 故可以在给定显著水平为 0.05 的前提下拒绝原假设 $H_0: \beta_1 = 0$, 从而可以认为广告费用的确对销售额有影响, 回归方程是显著的. 另外, 通过查表也可以得到自由度为 1 和 8 的 F 分布的 0.05 上侧分位数为 $F_{0.05}(1, 8) = 5.32$, 将这一临界值同 F 值相对比, 由 $F = 12.93417 > F_{0.05}(1, 8)$, 因此可以获得相同的结论.

二、可决系数

根据前面离差平方和的分解公式 (8.34) 以及对 SSR 和 SSE 的各自含义的分析, 不难看出, 回归平方和 SSR 在总的离差平方和 SST 中所占比重越大, 从而残差平方和 SSE 所占比重越小, 则越发说明了变量 x 对变量 Y 有一定的影响作用. 因此, 我们引入如下的指标 R^2 来度量变量 x 对变量 Y 的解释能力.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}. \quad (8.44)$$

R^2 被称为可决系数, 由定义知

$$R^2 \leq 1. \quad (8.45)$$

R^2 作为一个相对的指标, 测度了拟合的回归直线所导致离差平方和占样本的总离差平方和的百分比, 因此它也是对回归方程拟合优度的一种测度. R^2 越接近于 1, 则回归方程对样本点拟合得越好, x 对 Y 的解释能力越强, 特别地, 如果 $R^2 = 1$, 则 $SSE = 0$, 说明每个样本点都落在了回归直线上, 即被充分拟合了, 此时由 x 的变化完全可以解释 Y 的变化.

利用前面得到的 $SSR = \sum_{i=1}^n \hat{y}_i^2$ 以及 \hat{y}_i 的计算公式 (8.17), 可得

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \end{aligned} \quad (8.46)$$

容易看出可决系数 R^2 就是变量 x 和 Y 之间的样本相关系数 r (此时将 x 看成随机变量) 的平方, 即

$$R^2 = r^2 \quad (8.47)$$

这样一来, 样本相关系数 r 和 R^2 尽管定义的角度不同, 但是对于一元线性回归模型而言, 它们反映的性质是等价的, 即 x 与 Y 的相关程度越强, x 的回归方程对 Y 的解释能力也越强.

基于上述认识, 我们可以对零假设 $H_0: \beta_1 = 0$ 设置满足条件

$$|F| \geq d \quad (8.48)$$

的样本集合作为共否定域, 其中 d 为某一待定正常数.

事实上, 由于

$$\begin{aligned} r^2 = R^2 &= \frac{SSR}{SST} = \frac{1}{1 + \frac{SSE}{SSR}} \\ &= \frac{1}{1 + \frac{n-2}{F}} = \frac{F}{F + (n-2)} \end{aligned} \quad (8.49)$$

在 $n > 2$ 时是 F 的增函数, 所以在 F 大于上侧分位数 $F(1, n-2)$ 时,

$$r^2 > \frac{F(1, n-2)}{F(1, n-2) + (n-2)}.$$

反之也是对的.

因此令 (8.48) 式中的 d 为

$$d = \frac{F(1, n-2)}{F(1, n-2) + (n-2)} \quad (9.50)$$

这样设置否定域为满足条件 $|r| \geq d$ 的样本集合即可得到 H_0 的一个显著水平为 α 的检验. 为此, 有时也称 d 为 r 的显著水平为 α 的临界值, 记为 $r_{\frac{\alpha}{2}}(n-2) = d$, 直接从书后所附的样本相关系数临界值表可以对给定的显著水平和样本容量查找相应的临界值, 从而大大方便了假设检验的步骤.

对例 8.1 而言, 可以算得广告费用与销售额之间的样本相关系数 $r = 0.786034$, 给定显著水平 $\alpha = 0.05$, 查表得临界值 $r_{0.025}(8) = 0.632$, 由 $|r| > r_{0.025}(8)$, 所以拒绝原假设 $H_0: \beta_1 = 0$, 得到的结论跟前面 F 检验是一致的.

三、t 检验

如果单纯地把 β_0 和 β_1 理解为模型中的参数, 由于前面已经得到了 β_0 和 β_1 的点估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 及其分布, 所以利用点估计可以方便地构造关于相应参数的假设检验问题的检验统计量.

在前面已经得到

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{l_{xx}}\right),$$

故

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{l_{xx}}} \sim N(0, 1). \quad (8.51)$$

又由于

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2),$$

而且 SSE 与 $SSR = \hat{\beta}_1^2 \cdot l_{xx}$ 独立, 从而 SSE 与 $\hat{\beta}_1$ 独立, 根据 t 分布的定义, 则

$$\begin{aligned} & \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{l_{xx}}}}{\sqrt{\frac{SSE}{\sigma^2 \cdot (n-2)}}} \\ &= \frac{\hat{\beta}_1 - \beta_1}{\frac{1}{\sqrt{l_{xx}}} \sqrt{\frac{SSE}{n-2}}} \sim t(n-2) \end{aligned} \quad (8.52)$$

其中

$$s_1 = \frac{\overline{SSE}}{n-2} \cdot \frac{1}{l_{xx}} \quad (8.53)$$

称为 s_1 的标准误差, 它是 s_1 的标准差——的估计值.

l_{xx}

所以, 在零假设 $H_0: \beta_1 = 0$ 成立的条件下

$$t_1 = \frac{\hat{\beta}_1}{s_1} \sim t(n-2) \quad (8.54)$$

t_1 可以作为该假设的一个检验统计量, 给定显著水平 α , 令 $t_{\frac{\alpha}{2}}(n-2)$ 表示自由度为 $n-2$ 的 t 分布的 $\frac{\alpha}{2}$ 上侧分位数, 那么显然设定否定域

$$C = \{(x, y): |\hat{\beta}_1| \geq t_{\frac{\alpha}{2}}(n-2) \cdot s_1\} \quad (8.55)$$

即可得到 $H_0: \beta_1 = 0$ 的一个显著水平为 α 的检验.

类似地, 可以证明

$$\frac{\hat{\beta}_0}{s_0} \sim t(n-2) \quad (8.56)$$

其中

$$s_0 = \frac{\overline{SSE}}{n-2} \cdot \frac{1}{n + \frac{\bar{x}^2}{l_{xx}}} \quad (8.57)$$

称为 s_0 的标准误差, 它是 s_0 的标准差 $\cdot \frac{1}{n + \frac{\bar{x}^2}{l_{xx}}}$ 的估计值.

同样, 在原假设 $H_0: \beta_0 = 0$ 成立时,

$$t_0 = \frac{\hat{\beta}_0}{s_0} \sim t(n-2). \quad (8.58)$$

从而设定否定域为满足条件

$$|\hat{\beta}_0| \geq t_{\frac{\alpha}{2}}(n-2) \cdot s_0 \quad (8.59)$$

的样本集合, 即可得到 $H_0: \beta_0 = 0$ 的一个显著水平为 α 的检验.

对于例 8.1 中给出的数据, 我们已经求得回归方程的两个系数的估计值分别为 $\hat{\beta}_0 = 309.5276$, $\hat{\beta}_1 = 4.067736$, 由公式 (8.57) 和 (8.53) 可以分别求得两个估计值的标准误差分别为 $s_0 = 43.40214$, $s_1 = 1.131054$. 从而相应的两个 t 值分别为 $t_0 = 7.131622$ 和 $t_1 = 3.596411$, 查表可得自由度为 8 的 t 分布的 0.025 上侧分位数为 2.306, t_0 、 t_1 都大于该临界值, 从而在给定显著水平为 0.05 的前提下 $H_0: \beta_0 = 0$ 和 $H_0: \beta_1 = 0$ 都可以被拒绝.

如果使用诸如 Excel97 等统计软件, 除了给出上述各值的计算结果之外, 通常还会给出 t_0 和 t_1 的 P 值, 比如相应刚才算出的 t_0 和 t_1 其 P 值分别为 9.89e

10^{-5} 和 0.007018793, 两个 P 值都远小于显著水平 0.05, 因此利用这一点也可以得出拒绝原假设的结论.

也许有人已经注意到, 上面对 $H_0: \beta_1 = 0$ 的 F 检验和 t 检验中 F 值和 t_1 值的 P 值都是 0.007018793, 实际上, 对于一元线性回归模型来说, 上述两个检验是等价的.

§ 8.3 一元线性回归的预测和控制

如果根据样本估计出来的回归方程通过了显著性的检验, 那么就可以利用这一回归方程进行预测和控制问题的研究.

一、预测问题

所谓的预测, 就是给定自变量 x 的一个值 x_0 , 求 x_0 对应的因变量 Y_0 的过程.

譬如对例 8.1 来说, 广告公司希望根据对 10 个厂家的广告费用与销售额的资料, 进一步了解如果另外一个厂家对同一类产品投入了广告费用为 55 万元, 该厂家的销售额是多少. 这就是一个典型的预测问题.

根据在前面两节中对模型的设置, 易知 Y_0 为一随机变量, 而且

$$Y_0 = \beta_0 + \beta_1 X_0 + \varepsilon_0 \quad (8.60)$$

其中 $\varepsilon_0 \sim N(0, \sigma^2)$.

如果 $\beta_0, \beta_1, \sigma^2$ 都是已知常数, 那么此时 Y_0 的分布也是已知的, 即 $Y_0 \sim N(\beta_0 + \beta_1 X_0, \sigma^2)$.

进一步, 如果此时需要找出一个数值作为 Y_0 的代表值, 显然在均方误差最小的原则下, $E(Y_0 | X_0) = \beta_0 + \beta_1 X_0$ 是唯一的选择, 即

$$E([Y_0 - E(Y_0 | X_0)]^2) = \min_c E([Y_0 - c]^2). \quad (8.61)$$

但是, 在实际问题当中, $\beta_0, \beta_1, \sigma^2$ 都是未知的, 正如前面两节中所做的, 只能利用样本对其进行估计, 因此实际上的预测问题就是利用拟合得到的回归方程 $Y = \beta_0 + \beta_1 X$ 进行以下两个方面的工作:

对 $E(Y_0 | X_0) = \beta_0 + \beta_1 X_0$ 进行估计;

对 Y_0 进行直接的估计.

下面我们分别来看.

1. 对 $E(Y_0 | X_0)$ 的估计

设 $Y_0 = \beta_0 + \beta_1 X_0$, 则类似于 (8.32) 式可知

$$Y_0 \sim N\left(\beta_0 + \beta_1 X_0, \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{l_{xx}}\right) \cdot \sigma^2. \quad (8.62)$$

所以, Y_0 是 $E(Y_0|x_0)$ 的无偏估计量.

为了求 $E(Y_0|x_0)$ 的区间估计, 由 (8.62) 式, 易知

$$\frac{Y_0 - (\beta_0 + \beta_1 x_0)}{\sigma_{Y_0}} \sim N(0, 1), \quad (8.63)$$

其中 σ_{Y_0} 表示 Y_0 的标准差, 即

$$\sigma_{Y_0} = \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}. \quad (8.64)$$

又由于 SSE 与 β_0 、 β_1 都独立以及 $\frac{SSE}{n-2} \sim \sigma^2$ ($n-2$) 可知 SSE 与 Y_0 独立, 而且

$$\frac{Y_0 - (\beta_0 + \beta_1 x_0)}{\sigma_{Y_0}} \sim t(n-2) \quad (8.65)$$

其中

$$\sigma_{Y_0} = \sqrt{\frac{SSE}{n-2} \cdot \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}} \right)} \quad (8.66)$$

是 σ_{Y_0} 的估计.

由 (8.65) 式, 设自由度为 $n-2$ 的 t 分布的 $\frac{1-\alpha}{2}$ 上侧分位数为 $t_{\frac{1-\alpha}{2}}(n-2)$, 那么 $E(Y_0|x_0) = \beta_0 + \beta_1 x_0$ 的置信度为 $1-\alpha$ 的置信区间为

$$Y_0 - \sigma_{Y_0} \cdot t_{\frac{1-\alpha}{2}}(n-2) \leq E(Y_0|x_0) \leq Y_0 + \sigma_{Y_0} \cdot t_{\frac{1-\alpha}{2}}(n-2). \quad (8.67)$$

容易看出, 这一置信区间的长度是 $2 \cdot \sigma_{Y_0} \cdot t_{\frac{1-\alpha}{2}}(n-2)$, 它是 x_0 的函数, 特别地, 当 $x_0 = \bar{x}$ 时, 即 x_0 取所有样本点的算术平均值时, 该置信区间的长度最短.

2. 对 Y_0 的预测

正如前面提到的, Y_0 是一个随机变量, 服从分布 $N(\beta_0 + \beta_1 x_0, \sigma^2)$; $Y_0 = \beta_0 + \beta_1 x_0$ 是由样本值构成的统计量, 其分布满足 (8.62) 式.

那么用 Y_0 作为对 Y_0 的预测其预测精度是多少? 这里的预测精度通常按照以下的方式衡量. 给定 α 为一个小正数, 求 δ 使之满足

$$P(|Y_0 - \hat{Y}_0| \leq \delta) = 1 - \alpha,$$

则 δ 越小就说明 Y_0 的预测精度越高, 并且称区间

$$(\hat{Y}_0 - \delta, \hat{Y}_0 + \delta)$$

为 Y_0 的概率为 $1-\alpha$ 的预测区间. 考虑到 Y_0 与已有样本 Y_1, Y_2, \dots, Y_n 之间的独立性, 易知 Y_0 与 \hat{Y}_0 相互独立, 所以 $Y_0 - \hat{Y}_0$ 服从正态分布, 而且

$$E(Y_0 - \hat{Y}_0) = 0;$$

$$D(Y_0 - \hat{Y}_0) = D(Y_0) + D(\hat{Y}_0)$$

$$= \sigma^2 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}} \cdot \sigma^2 = \sigma_{Y_0}^2 + \sigma_{\hat{Y}_0}^2$$

故有

$$\frac{Y_0 - \bar{Y}_0}{\sqrt{Y_0 - Y_0}} \sim N(0, 1). \quad (8.68)$$

同样地, 利用 SSE 与 Y_0 及 Y_0 之间的独立性可知

$$\frac{Y_0 - \bar{Y}_0}{\sqrt{Y_0 - Y_0}} \sim t(n-2), \quad (8.69)$$

其中

$$\sqrt{Y_0 - Y_0} = \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}} \cdot \sqrt{\frac{SSE}{n-2}} \quad (8.70)$$

是 $\sqrt{Y_0 - Y_0}$ 的估计.

由 (8.69) 式可知

$$P\{|\hat{Y}_0 - Y_0| \leq \sqrt{Y_0 - Y_0} \cdot t_{\frac{\alpha}{2}}(n-2)\} = 1 - \alpha. \quad (8.71)$$

因此可以得到 Y_0 的概率为 $1 - \alpha$ 的预测区间是

$$(\hat{Y}_0 - \sqrt{Y_0 - Y_0} \cdot t_{\frac{\alpha}{2}}(n-2), \hat{Y}_0 + \sqrt{Y_0 - Y_0} \cdot t_{\frac{\alpha}{2}}(n-2)), \quad (8.72)$$

其预测精度可以用

$$\sqrt{Y_0 - Y_0} \cdot t_{\frac{\alpha}{2}}(n-2)$$

来反映.

由 (8.70) 式可以看出, 在样本给定的前提下, $\sqrt{Y_0 - Y_0}$ 是 x_0 的函数, 不妨记之为 $\delta(x_0)$, x_0 越靠近样本的算术平均值 \bar{x} , $\delta(x_0)$ 越小, 表示预测精度越高, 在 $x_0 = \bar{x}$ 时, 预测精度达到最高. 见图 8-3.

另外, 增加样本的容量 n 或者扩大 x_1, x_2, \dots, x_n 的分散程度以致使 l_{xx} 增加, 都会导致 $\delta(x_0)$ 减小, 从而提高预测的精度.

特别地, 如果 n 非常大, 而且 x_0 在 \bar{x} 附近取值时, 近似地有 $\sqrt{Y_0 - Y_0} \approx \sqrt{\frac{SSE}{n-2}}$, 此时自

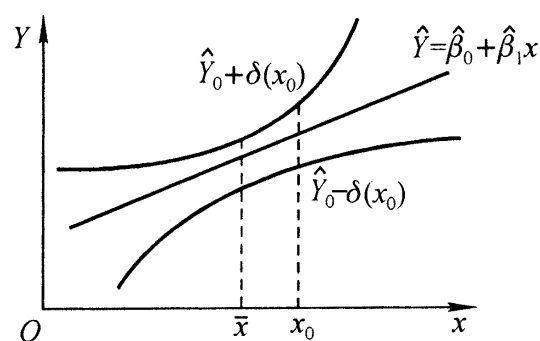


图 8-3 Y_0 的预测区间图示

由度为 $n-2$ 的 t 分布与 $N(0, 1)$ 也会十分接近, 因此可以用标准正态分布的 $\frac{\alpha}{2}$ 上侧分位数 $u_{\frac{\alpha}{2}}$ 来取代 $t_{\frac{\alpha}{2}}(n-2)$, 这样一来,

Y_0 的概率为 $1 - \alpha$ 的预测区间就可以近似地记成:

$$(\hat{Y}_0 - \sqrt{Y_0 - Y_0} \cdot u_{\frac{\alpha}{2}}, \hat{Y}_0 + \sqrt{Y_0 - Y_0} \cdot u_{\frac{\alpha}{2}}) \quad (8.73)$$

其中 $\sqrt{\frac{SSE}{n-2}}$ 是 $\sqrt{Y_0 - Y_0}$ 的估计.

下面我们利用以上的讨论结果来分析在例 8.1 中如果一个厂家投入了 55

万元的广告费用时其相应的销售额的预测问题. 令 $x_0 = 55$, 设 Y_0 为所求的销售额. 则由拟合的回归方程可知

$$Y_0 = 309.5276 + 4.067736x = 533.253,$$

因此 Y_0 的均值 $E(Y_0)$ 的估计值是 533.253 万元, 它是对 Y_0 的平均水平的预测. $E(Y_0)$ 的 95% 的置信区间是 $(533.253 - 57.29166, 533.253 + 57.29166) = (475.9613, 590.5447)$. 而且, 由

$$(55) = 113.2396,$$

可知 Y_0 的概率为 95% 的预测区间是 $(420.0134, 646.4926)$.

在此例中, 如果取 $x_0 = 36.5 = \bar{x}$, 则此时对应的销售额的概率为 95% 的预测区间是 $(458 - 102.4449, 458 + 102.4449) = (355.5551, 560.4449)$, 它是所有 95% 的预测区间中最短的一个. 如果使用近似公式 (8.73), 则可以得到这一预测区间的近似表示是

$$(458 - 2 \times 42.358, 458 + 2 \times 42.358) = (373.284, 542.716)$$

其中采用了 $u_{0.025} = 2$.

二、控制问题

对于例 8.1 中考虑的问题, 如果某个厂家感兴趣的是为了有 95% 的把握使得它的销售额高于 380 万元, 那么该厂家至少应该投入多少广告费用. 这就是一个控制问题.

控制问题可以看作是预测的反问题. 它的较为一般的提法是, 给定某一小正数 $(0 < \alpha < 1)$, 如何控制自变量 x 的取值, 才能以 $1 - \alpha$ 的概率保证因变量 Y 的落在某个预先指定的区间 (y_L, y_U) 之内.

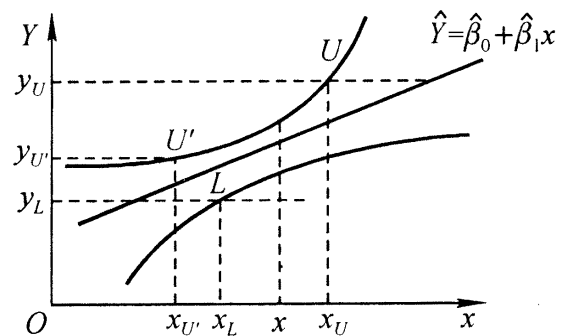


图 8-4 控制问题的直观求解

利用图形可以直观地求解上述问题. 在图 8-4 上, 通过 Y 轴上的 $(0, y_L)$ 、 $(0, y_U)$ 两点各画一条平行于 x 轴的直线, 并设它们分别交 Y^- 及 Y^+ 于 L 、 U 两点, 设 L 、 Q 两点对应的横坐标分别为 x_L 、 x_U , 则容易看出, 如果自变量 x 取 x_L 和 x_U 之间的

任一值, 相应的因变量 Y 的概率为 $1 - \alpha$ 的预测区间都将包含在指定的区间 (y_L, y_U) 之内, 因此, x_L 与 x_U 之间的任一 x 都符合控制问题的要求.

当然从这里也可以看出, 预先指定的区间 (y_L, y_U) 的长度不能太短, 至少应该不少于 $2 \sigma(x_L)$, 即 $x = x_L$ 时对应的 Y 的概率为 $1 - \alpha$ 的预测区间的长度, 否则, 比如取 (y_L, y_U) 小于该长度, 则正如图 8-4 所示, x_L 与 x_U 之间的任一 x 值对应的 Y 的 $1 - \alpha$ 预测区间都比 (y_L, y_U) 更大, 都不是控制问题的解. 此时

按照上述回归模型无法求解该控制问题.

在实际问题当中, 通常利用 (8.73) 式表示的预测区间的近似表达式来求解 x , 即通过方程

$$Y(x_L) - \frac{1}{2} \cdot u_2 = y_L$$

$$Y(x_U) + \frac{1}{2} \cdot u_2 = y_U$$

可以解出

$$x_L = \frac{y_L + \frac{1}{2} \cdot u_2 - y_0}{f_1}, \tag{8.74}$$

$$x_U = \frac{y_U + \frac{1}{2} \cdot u_2 - y_0}{f_1}, \tag{8.75}$$

那么 x_L 与 x_0 之间的 x 即所求.

利用以上公式可以十分方便地求解本小节开始提出的广告费用的控制问题, 即至少投入多少广告费用才能以 95% 的概率保证销售额不低于 380 万元, 由公式 (8.74) 易得

$$x_L = \frac{380 + \frac{1}{2} \cdot 42.358 - 309.5276}{4.067736} = 38.15,$$

因此至少要投入 38.15 万元的广告费用.

§ 8.4 一元非线性问题的线性化

在一元线性回归模型中, 假设因变量 Y 对自变量 x 的回归函数是一个线性函数 (见 (8.4) 式), 这意味着在样本数据的散点图上这些点将在一条直线的附近变动. 但是在许多实际问题当中, Y 与 x 的关系不一定是线性的, 它们之间可能存在着某种复杂的非线性的联系, 表现为散点图上的点围绕着某条曲线波动. 请看下面的例子.

例 9.2 经过调查得到 8 个厂家对同种类型的产品年新增加投资额和年利润额的数据资料, 如表 8-4 所示.

图 8-5 给出了年利润额 Y 占年新增加投资额 x 的散点图, 从图中可以清楚地看出来, 随着 x 的增大 Y 也有明显的增加的趋势, 因此两者之间存在着相关关系, 但是这种相关关系与其用一条直线来描述倒不如用曲线描述更加合适, 因此 Y 与 x 之间更加倾向于被认为是一种非线性关系. 回归方程也需要用一些非线性函数来刻画, 比如

使用 Y 的概率为 $1-\alpha$ 的单侧预测区间求解该问题更为直接和合理. 参见有关文献.

表 8-4 八个厂家年投资额与利润数据资料

厂家	1	2	3	4	5	6	7	8
年新增投资额 x (万元)	4	6	10	11	15	17	18	20
利润额 Y (万元)	6	7	9	10	17	24	23	26
$\ln Y$	1.79	1.95	2.20	2.30	2.83	3.18	3.14	3.26

$$Y = a_0 \cdot e^{a_1 x} \tag{8.76}$$

或者

$$Y = a_0 + a_1 \cdot x^2 \tag{8.77}$$

等等. 图 8-6 给出的是变量 $\ln Y$ 与变量 x 的散点图, 从中可以看出这些点基本上是围绕一条直线波动, 说明变量 $\ln Y$ 与 x 之间近似是一种线性关系, 从而也印证了回归方程取 (8.76) 形式的合理性.

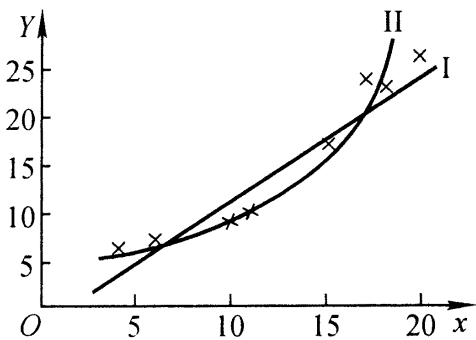


图 8-5 年投资额与利润数据的散点图

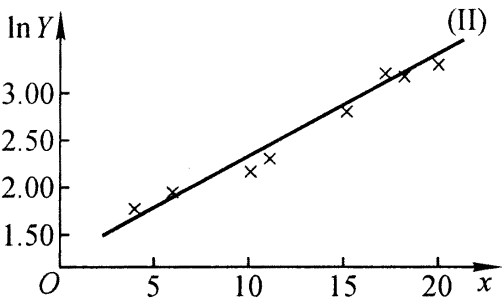


图 8-6 经过对数变换后的散点图

同时, 图 8-6 也提示给我们一种求解回归方程 (8.76) 的思路, 即通过求解变量 $\ln Y$ 对 x 的线性回归方程即可得到相应的 (8.76) 式所表示的 Y 对 x 的回归方程, 即在图 8-6 中的回归直线同图 8-5 中的曲线 () 是一致的. 具体来说, 首先对样本数据 $(x_i, Y_i), i=1, 2, \dots, n)$ 作对数变换

$$Z_i = \ln Y_i, i=1, 2, \dots, n; \tag{8.78}$$

然后利用最小二乘法求出变量 Z 对 x 的回归方程

$$Z = a_0 + a_1 \cdot x, \tag{8.79}$$

即图 8-6 中的直线方程, 则相应的形如 (8.76) 式的 Y 对 x 的回归方程是

$$Y = e^Z = e^{a_0} \cdot e^{a_1 x} \tag{8.80}$$

即 $a_0 = e^{a_0}, a_1 = a_1$.

利用表 8-4 中给出的数据, 可以得到 $\ln Y$ 对 x 的线性回归方程是

$$Z = 1.3139 + 0.1003 x$$

(14.234) (14.934)

其中下面括号中的数是对应系数的 t 值. 由此可得 Y 对 x 的回归方程是

$$Y = 3.7208 \cdot e^{0.1003x} \quad (8.81)$$

如果采用形如 (8.77) 式的抛物线型回归方程, 容易看出, 令 $w = x^2$, (8.77) 式就是表示了变量 Y 对 w 的线性回归方程

$$Y = \beta_0 + \beta_1 \cdot w. \quad (8.82)$$

所以, 对样本数据做变换 $w_i = x_i^2$ ($i = 1, 2, \dots, n$), 利用 (w_i, Y_i) ($i = 1, 2, \dots, n$) 求解出 (8.82) 中的系数估计 β_0 、 β_1 代入 (8.77) 式即得到 Y 对 x 的回归方程.

对表 9-4 中的数据计算结果为

$$Y = \underset{(4.547)}{4.413} + \underset{(13.630)}{0.057} x^2 \quad (8.83)$$

通过上面的例子可以看出, 对于一些常见的非线性回归方程, 可以通过对样本进行适当的变换将其化成新的变量之间的线性回归方程, 这种求解思路除了适用于 (8.76)、(8.77) 两种类型的非线性函数形式外, 还适用于以下常见的函数:

1. 双曲线函数

函数形式为

$$\frac{1}{Y} = a + b \cdot \frac{1}{x}$$

或

$$Y = \frac{x}{ax + b}$$

令 $Z = \frac{1}{Y}$, $w = \frac{1}{x}$, 则

$$Z = a + bw$$

因此, 需对原始样本数据 (x_i, Y_i) 进行变换 $\frac{1}{x_i}, \frac{1}{Y_i}$ ($i = 1, 2, \dots, n$), 利用后者求解 $\frac{1}{Y}$ 对 $\frac{1}{x}$ 之间的线性回归方程即可得到 Y 对 x 的非线性回归方程的形式. 以下对其他函数形式的回归方程求解类同, 因此只给出相应变换的形式.

2. 对数曲线函数

函数形式为

$$Y = a + b \cdot \ln x,$$

令 $w = \ln x$, 则

$$Y = a + b \cdot w$$

3. 幂函数形式

函数形式为

$$Y = a \cdot x^b,$$

令 $Z = \ln Y$, $a = \ln a$, $w = \ln x$, 则

$$Z = a + b \cdot w$$

4. S 型曲线函数

函数形式为

$$Y = \frac{1}{a + be^{-x}},$$

令 $Z = \frac{1}{Y}$, $w = e^{-x}$ 则

$$z = a + b \cdot w$$

对回归方程选择一种合适的函数形式, 必须事先对散点图进行认真的分析, 此时对上述各种函数类型所描述的曲线形状有充分的了解是十分必要的. 我们在图 8-7 中给出了上述几种函数的图示, 可供参考.

正如对例 8.2 的分析中所看到的, 对同一种散点图所呈现的 Y 与 x 之间的关系, 可以选择不同的函数形式来描述回归方程, 那么如何判断并比较不同回归方程的拟合优度呢?

通常使用的比较准则有下面两个:

1. 相关指数 R

定义 R 是

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \tag{8.84}$$

的平方根, 上式与线性回归模型中定义的可决系数是一致的, 等号右边的第二项表示了样本点对回归曲线的偏离占样本点总的离散程度的比重, 如果回归曲线对样本点拟合得很好, 这一项自然就会比较小, 从而 R^2 较大, 因此对于不同的曲线回归方程通常选择 R^2 较大的一个.

2. 剩余标准差 S

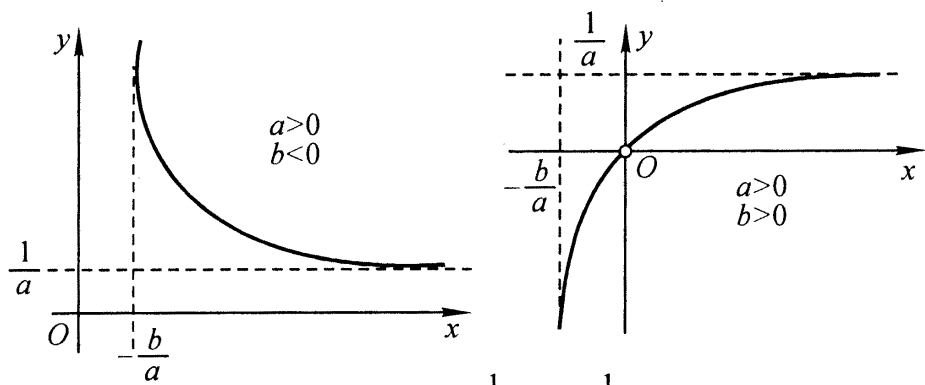
定义 S 为

$$S = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}} \tag{8.85}$$

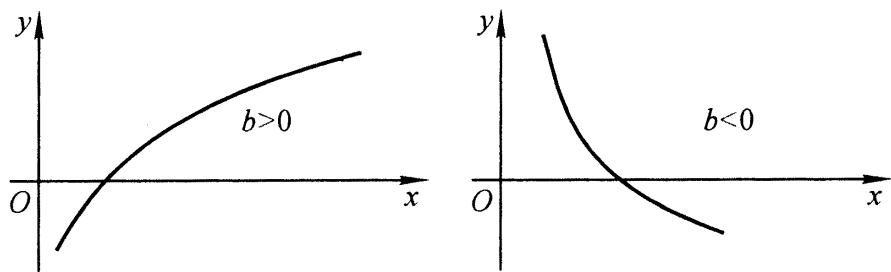
它反映了样本点偏离回归曲线的平均大小, 当然 S 小一点比较好.

从(8.84)、(8.85)两个式子中不难发现上述两个准则是一致的, 即 R 越大则 S 越小, 反之亦然.

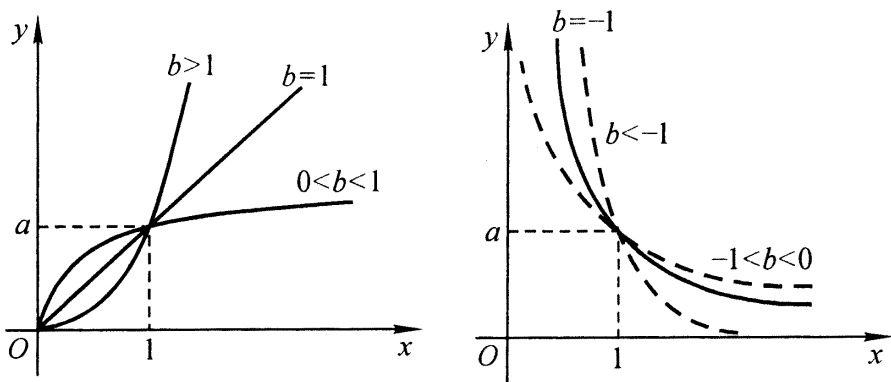
对于前面利用例 8.2 中的数据得到的两个曲线回归方程即指数形式(8.81)和抛物线形式(8.83)分别计算它们的相关指数 R 和剩余标准差 S , 计算结果列



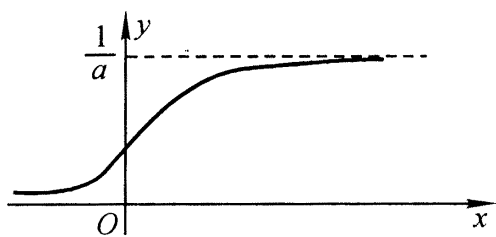
(a) 双曲函数 $\frac{1}{y}=a+b\cdot\frac{1}{x}$



(b) 对数函数 $y=a+b\cdot\ln x$



(c) 幂函数 $y=a\cdot x^b$



(d) 双曲函数 $y=\frac{1}{a+b\cdot e^{-x}}$

图 8-7 常用的曲线形式

在了表 8-5 中. 为了便于比较, 同时还列出取回归方程为线性函数时所对应的 R 与 S 的值.

从表 8-5 中可以清楚地看出, 三种不同形式的回归方程当中抛物线型方程对应的相关指数最大, 而剩余标准差最小, 因此相比之下抛物线型方程是最可取的. 直线回归方程的拟合效果在三者之中是最差的, 这也进一步说明了选择适当的非线性回归模型的必要性.

表 8-5 三种不同回归方程的比较

曲线形式	回归方程	相关指数 R	剩余标准差 S
直线	$Y = -2.02 + 1.37x$	0.9637	2.3762
指数曲线	$Y = 3.7208 \cdot e^{0.1003x}$	0.9804	1.7522
抛物线	$Y = 4.4132 + 0.0574x^2$	0.9842	1.5746

§ 8.5 多元线性回归分析

在许多实际问题当中, 往往需求研究多个变量之间的相关关系, 比如某种产品的销售额不仅受到投入的广告费用的影响, 通常还与产品的价格、消费者的收入状况、社会保有量以及其它可替代产品的价格等诸多因素有关系, 研究这样一个变量同其它多个变量之间的关系的主要方法是运用多元回归分析、多元线性回归分析是一元线性回归分析的自然推广形式, 两者在参数估计、显著性检验等技术方面是非常相似的.

一、多元线性回归模型

设影响因变量 Y 的自变量个数为 p , 并分别以 x_1, x_2, \dots, x_p 记之, 所谓多元线性模型是指这些自变量对 Y 的影响是线性的, 即

$$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_p \cdot x_p + \epsilon$$
 (8.86)

其中,

$$f(x_1, x_2, \dots, x_p) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_p \cdot x_p$$
 (8.87)

称 Y 对 x_1, x_2, \dots, x_p p 个自变量的线性回归函数. 类似于一元线性回归, 在这里我们已假定 p 个自变量都是确定的量.

记 n 组样本分别是 $(x_{i1}, x_{i2}, \dots, x_{ip}; Y_i) \quad (i = 1, 2, \dots, n)$. 那么由 (8.86) 式得到

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \dots, n$$
 (8.88)

类似地于一元情形, 通常假设

$$E(\epsilon_i) = 0, D(\epsilon_i) = \sigma^2, i = 1, 2, \dots, n,$$
 (8.89)

即 Y_1, Y_2, \dots, Y_n 是等方差的, 而且 $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ 之间相互独立. (8.90)

为了检验回归模型及其系数的显著性, 通常还假设 $\epsilon_i (i = 1, 2, \dots, n)$ 都服从正态分布, 在这一条件下, (8.89)、(8.90) 等价于

$$\epsilon_i \sim N(0, \sigma^2) \text{ 且 } cov(\epsilon_i, \epsilon_j) = 0, (i \neq j; i, j = 1, 2, \dots, n).$$
 (8.91)

这样一来,多元线性回归分析的主要问题就是基于模型 (8.88)、(8.91) 对于其中的未知参数 $\beta_0, \beta_1, \dots, \beta_p$ 和 σ^2 进行估计、检验以及相应地利用得到的回归模型进行预测等.

为了表达得简洁起见,通常使用矩阵将上述模型 (8.88)、(8.91) 统一记成

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ \varepsilon &\sim N_n(O, \sigma^2 I_n). \end{aligned} \quad (8.92)$$

其中

$$\begin{aligned} Y &= (Y_1, Y_2, \dots, Y_n), \\ &= (\beta_1, \beta_2, \dots, \beta_n), \\ X &= \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}, \\ &= (\beta_0, \beta_1, \beta_2, \dots, \beta_p), \end{aligned}$$

I_n 是 n 阶单位阵, O 表示 n 维零向量.

二、回归系数的最小二乘估计

利用最小二乘法估计回归系数即是求解一组 $\beta_0, \beta_1, \dots, \beta_p$ 使之满足如下定义的平方和 Q 达到最小,

$$Q(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_p X_{ip})^2 \quad (8.93)$$

即

$$Q(\beta_0, \beta_1, \dots, \beta_p) = \min_{\beta_0, \beta_1, \dots, \beta_p} Q(\beta_0, \beta_1, \dots, \beta_p)$$

由多元函数微分学的知识可知 Q 在 $(\beta_0, \beta_1, \dots, \beta_p)$ 点的各个一阶偏导数均为零, 于是 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 满足以下 $p+1$ 个方程:

$$\begin{aligned} -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_p X_{ip}) &= 0 \\ -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_p X_{ip}) X_{i1} &= 0 \\ &\dots\dots\dots \\ -2 \sum_{j=1}^p \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_p X_{ip}) X_{ij} &= 0 \end{aligned}$$

整理该方程组可得

$$\begin{aligned} n\beta_0 + \sum_{i=1}^n X_{i1}\beta_1 + \dots + \sum_{i=1}^n X_{ip}\beta_p &= \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_{i1}^2\beta_1 + \sum_{i=1}^n X_{i1}X_{ip}\beta_p &= \sum_{i=1}^n X_{i1}Y_i \\ &\dots\dots\dots \\ \sum_{i=1}^n X_{ip}^2\beta_p &= \sum_{i=1}^n X_{ip}Y_i \end{aligned} \quad (8.94)$$

上述方程组又被称为正规方程组，用矩阵形式可以表示为：

$$(X'X)\beta = X'Y \quad (8.95)$$

其中 $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ ，如果矩阵 $X'X$ 可逆，则正规方程组的解是

$$\beta = (X'X)^{-1}X'Y \quad (8.96)$$

经常还用到正规方程的另一种形式，考虑到 (8.94) 中的第一式可以写成

$$\beta_0 = \bar{Y} - X_1\beta_1 - X_2\beta_2 - \dots - X_p\beta_p, \quad (8.97)$$

其中 $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ ， $X_k = \frac{1}{n} \sum_{i=1}^n X_{ik}$ ($k=1, 2, \dots, p$)，将其代入其余 p 个式子可以得到

$$\begin{aligned} l_{11}\beta_1 + l_{12}\beta_2 + \dots + l_{1p}\beta_p &= l_{1Y} \\ l_{21}\beta_1 + l_{22}\beta_2 + \dots + l_{2p}\beta_p &= l_{2Y} \\ &\dots\dots\dots \\ l_{p1}\beta_1 + l_{p2}\beta_2 + \dots + l_{pp}\beta_p &= l_{pY}, \end{aligned} \quad (8.98)$$

其中

$$\begin{aligned} l_{kj} &= \sum_{i=1}^n (X_{ik} - \bar{X}_k)(X_{ij} - \bar{X}_j), \quad k, j = 1, 2, \dots, p; \\ l_{kY} &= \sum_{i=1}^n (X_{ik} - \bar{X}_k)(Y_i - \bar{Y}), \quad k = 1, 2, \dots, p. \end{aligned}$$

如果以 L 表示 (8.98) 的系数矩阵，即 $L = (l_{kj})_{p \times p}$ ，以 l_Y 表示 (8.98) 式等号右边的常数项， $l_Y = (l_{1Y}, l_{2Y}, \dots, l_{pY})$ ，则 (8.98) 可以简记为

$$L\beta^* = l_Y, \quad (8.99)$$

在这里以 β^* 表示向量 $(\beta_1, \beta_2, \dots, \beta_p)$ 。如果 L 可逆，则

$$\beta^* = L^{-1}l_Y \quad (8.100)$$

将其代入 (8.97) 式即可得到 β_0 的值。

易知这样得到的 $\beta_0, \beta_1, \dots, \beta_p$ 与 (8.96) 是一致的，而且的确是 Q 的最小值点，我们称这样得到的 $\beta_0, \beta_1, \dots, \beta_p$ 是回归系数 $\beta_0, \beta_1, \dots, \beta_p$ 的最小

二乘估计. 而

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \tag{8.101}$$

就是拟合得到的回归方程.

类似地也可以证明所谓的高斯——马尔可夫定理, 即在模型 (8.92) 的假设条件下, 上面由最小二乘法得到的 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ 分别是 $\beta_0, \beta_1, \dots, \beta_p$ 的最佳线性无偏估计 (BLUE).

三、回归方程的显著性检验

回归方程的显著性检验用来考察所选用的自变量 x_1, x_2, \dots, x_p 是否的确对因变量 Y 起了解释作用, 即要检验假设

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0 \tag{8.102}$$

如果 H_0 为真, 说明 Y 没有受到这些自变量中任何一个的影响, 此时回归方程是没有意义的.

以 $Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, (i = 1, 2, \dots, n)$ 表示第 i 组样本对应的拟合值, 类似于§ 8.2中对一元线性回归问题分析的思路, 进行如下的离差平方和分解:

$$\begin{aligned} SST &= \sum_{i=1}^n (Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\ &\stackrel{\text{def}}{=} SSE + SSR, \end{aligned}$$

其中利用到等式

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0,$$

该等式易由正规方程 (8.94) 推知.

这里的 SSR 、 SSE 分别称为回归平方和、残差平方和, 各自的含义也与 § 9.2 中的分析相似, 而且, 可以证明, 在 H_0 成立时, 如下定义的统计量

$$F = \frac{SSR/p}{SSE/(n-p-1)} \tag{8.103}$$

服从自由度为 p 和 $n-p-1$ 的 F 分布.

给定显著水平 α , 令 $F_{\alpha}(p, n-p-1)$ 为自由度为 p 和 $n-p-1$ 的 F 分布的上侧分位数, 则将拒绝域设置为满足条件

$$F > F_{\alpha}(p, n-p-1)$$

的样本集合即可得到 H_0 的一个显著水平为 α 的检验.

表8-6给出的方差分析表是对上述分析过程的一个总结.

同样地，称

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

表8-6 多元线性回归方程显著性检验的方差分析表

方差来源	平方和	自由度	均方和	F 值
回归	$SSR = \sum_{i=1}^n (Y_i - \bar{Y})^2$	P	$MSR = \frac{SSR}{P}$	$F = \frac{MSR}{MSE}$
残差	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	n - p - 1	$MSE = \frac{SSE}{n - p - 1}$	
总计	$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$	n - 1		

为可决系数，易知 $0 \leq R^2 \leq 1$ ， R^2 越接近于1则说明拟合值的离差平方和占样本总的离差平方和的比重越大，或者拟合的残差越小，因此 R^2 是拟合优度的反映。称 R^2 的平方根R为复相关系数，类似于度量两个变量之间的线性相关程度的直线相关系数，R可以看作是Y与 $x_1、x_2、\dots、x_p$ 之间相关关系及其密切程度的一种反映。

四、回归系数的检验

对于多元线性回归模型来说，回归方程的显著性检验被通过（即 H_0 （8.102）被拒绝）只能说明p个自变量在整体上对Y是有影响的，但是这并不意味着每一个自变量对Y是必需的，因此，还需要进一步对每一个自变量前面的系数进行检验。

譬如为了考察变量 x_j 对Y的影响是否显著，就需要对如下的原假设进行检验。

$$H_0: \beta_j = 0 \tag{8.104}$$

如果在给定的显著水平下不能拒绝 H_0 ，说明 x_i 对Y影响并不显著，可以考虑在回归方程中去掉 x_j 。

由于 β_j 的最小二乘估计量为 b_j ，故可以选择 H_0 的否定域为满足条件 $|b_j| > c$ 的样本集合，其中c为某一待定的正常数。可以证明，在 H_0 （8.104）式成立的条件下，构造统计量

$$F_j = \frac{b_j^2 / 1}{SSE / (n - p - 1)}, \tag{8.105}$$

和

$$t_j = \frac{\bar{y}_j / \sqrt{l^{jj}}}{\sqrt{SSE / (n - p - 1)}}, \quad (8.106)$$

其中, l^{jj} 是矩阵 L^{-1} 对角线上的第 j 个元素, 则

$$F_j \sim F(1, n - p - 1),$$

$$t_j \sim t(n - p - 1).$$

因此, 给定显著水平 α , 令 $F_{\alpha}(1, n - p - 1)$ 表示 $F(1, n - p - 1)$ 分布的 α 上侧分位数, 令 $t_{\frac{\alpha}{2}}(n - p - 1)$ 表示 $t(n - p - 1)$ 分布的 $\frac{\alpha}{2}$ 上侧分位数, 所以可以将 H_0 的否定域设定为满足条件

$$F_j > F_{\alpha}(1, n - p - 1) \quad (8.107)$$

或者满足条件

$$|Q_j| > t_{\frac{\alpha}{2}}(n - p - 1) \quad (8.108)$$

的样本集合, 都可以得到 H_0 的一个显著水平为 α 的检验. 通常称满足 (8.107) 式的否定域为 F 检验, 满足 (8.108) 式的否定域称为 t 检验. 对于 H_0 (8.104) 而言, F 检验和 t 检验是等价的.

在实际问题中, 如果有多个回归系数不显著, 则不能同时将相应的自变量在回归方程中去除掉, 而只能先将 F_j 值 (或者 $|Q_j|$ 值) 最小的一个自变量删除, 再对剩余的 $p - 1$ 个自变量重新建立回归方程, 并进行系数的显著性检验, 如此直至所有变量的系数都显著为止.

五、多元线性回归模型的预测

假设经过回归方程和回归系数的显著性检验之后最终得到的拟合方程是 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p$, 那么任意给定一组自变量的值 $(x_{01}, x_{02}, \dots, x_{0p})$, 并令 Y_0 是该组值对应的因变量的值, 即

$$Y_0 = \beta_0 + \beta_1 x_{01} + \dots + \beta_p x_{0p} + \varepsilon_0,$$

则

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \dots + \hat{\beta}_p x_{0p} \quad (8.109)$$

是 $E(Y_0)$ 的无偏估计.

给定 $0 < \alpha < 1$, $E(Y_0)$ 的置信度为 $1 - \alpha$ 的置信区间为

$$Y_0 - \hat{\beta}_0 \pm t_{\frac{\alpha}{2}}(n - p - 1), Y_0 + \hat{\beta}_0 \pm t_{\frac{\alpha}{2}}(n - p - 1) \quad (8.110)$$

其中

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} + \frac{\sum_{k=1}^n \sum_{j=1}^n l^{kj} (x_{0k} - \bar{x}_k)(x_{0j} - \bar{x}_j)}{\sum_{k=1}^n \sum_{j=1}^n l^{kj}} \\ &= \frac{SSE}{n - p - 1} \text{ 是 } \sigma^2 \text{ 的估计,} \end{aligned}$$

而 l^{kj} 是矩阵 L^{-1} 中第 k 行第 j 列元素, $k, j = 1, 2, \dots, p$.

类似地, 对于给定的 $0 < \alpha < 1$, 还可以考虑 Y_0 的概率为 $1-\alpha$ 的预测区间, 利用

$$\frac{Y_0 - \hat{Y}_0}{\hat{\sigma}} \sim t(n - p - 1), \tag{8.111}$$

其中

$$= \sqrt{1 + \frac{1}{n} + \sum_{k=1}^n \sum_{j=1}^n l^{kj} (\bar{X}_{0k} - \bar{X}_k)(\bar{X}_{0j} - \bar{X}_j)},$$

可得 Y_0 的概率为 $1-\alpha$ 的预测区间是

$$(\bar{Y}_0 - \hat{\sigma} t_{\frac{\alpha}{2}}(n - p - 1), \bar{Y}_0 + \hat{\sigma} t_{\frac{\alpha}{2}}(n - p - 1)), \tag{8.112}$$

它要比 $E(Y_0)$ 的 $1-\alpha$ 的置信区间大一些.

如果 n 很大, 而且给定的自变量的值 $(X_{01}, X_{02}, \dots, X_{0p})$ 非常接近于样本的中间位置, 即 $(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)$, 那么 Y_0 的 $1-\alpha$ 的预测区间可以近似地表示成

$$(\bar{Y}_0 - \hat{\sigma} u_{\frac{\alpha}{2}}, \bar{Y}_0 + \hat{\sigma} u_{\frac{\alpha}{2}}) \tag{8.113}$$

其中 $u_{\frac{\alpha}{2}}$ 是标准正态分布的 $\frac{\alpha}{2}$ 上侧分位数.

习 题 八

(A)

1. 在某个地区随机抽取9个家庭, 调查得到每个家庭各自每月人均收入与食品消费支出的数据如下表所示 (单位: 元):

编 号	1	2	3	4	5	6	7	8	9
人均收入	102	124	156	182	194	229	272	342	495
食品支出	87	92	90	124	150	160	182	216	220

根据上面数据, 试求

- (1) 人均收入与食品消费支出的相关系数;
- (2) 建立食品支出 Y 对人均收入 x 的一元线性回归模型, 并对模型进行检验;
- (3) 根据以上建立的模型, 试求该地区一个人均收入为每月300元的家庭人均食品支出的概率为95% 的预测区间.

2. 为了考察某班学生的学习成绩, 随机抽取了七名同学, 他们两个学期各门课程的平均成绩记录在下表中:

利用上述数据建立第二学期各科平均成绩对第一学期各科年均成绩的一元线性回归模型, 并对模型的显著性进行检验 (取显著水平 $= 0.05$), 据此你可以得出什么结论?

学生编号	1	2	3	4	5	6	7
第一学期	76	69	82	92	88	70	84
第二学期	80	63	84	90	91	76	71

3. 假设某种钢材的硬度与含铜量百分比以及温度之间服从线性关系，下面给出六次试验的数据资料.

钢材硬度 Y	含铜量 (%) x_1	温度 x_2
78.9	0.02	1000
55.2	0.02	1200
80.9	0.10	1000
57.4	0.10	1200
85.3	0.18	1000
60.7	0.18	1200

根据上面数据

- (1) 建立硬度对含铜量及温度的二元线性回归方程;
- (2) 对回归方程的显著性进行检验 ($\alpha = 0.05$);
- (3) 对回归方程的系数进行显著性检验 ($\alpha = 0.05$), 并对系数的含义进行解释.
4. 随机选择了10个地区对某种产品一周之内的销售额 Y、广告费用 x_1 进行调查, 数据列在下面的表中, 同时列出的还有每个地区的人口密度 x_2 , 试建立 Y 对 x_1 、 x_2 的线性回归模型.

地区编号	销售额 Y (千元)	广告费 x_1 (千元)	人口密度 x_2 (人/ 平方公里)
1	20	0.2	50
2	25	0.2	50
3	24	0.2	50
4	30	0.3	60
5	32	0.3	60
6	40	0.4	70
7	28	0.3	50
8	50	0.5	75
9	40	0.4	70
10	50	0.5	74

5. 下面的表中列出了分别对变量 x 与变量 y 独立进行8次观测的结果:

x	3	5	9	13	16	21	27	30
y	102	97	110	128	172	213	326	450

- 根据以上数据,
- (1) 画出散点图;
 - (2) 选择两种曲线形式建立 y 对 x 的曲线回归模型, 并对两种模型的似合优度进行比较.

习 题 八

(B)

1. 对于一元线性回归模型

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \dots, n;$$

在条件 $(H_1) \sim (H_5)$ 都满足的情况下, 试求 β_0, β_1 以及 σ^2 的最大似然估计.

- 2. 试求回归系数 β_1 与相关系数 r 之间的关系并由此给出 r 的一种解释.
- 3. 如果存在 δ 使得

$$P(Y_0 - Y_0 - \delta) = 1 - \alpha$$

成立, 则称 $(Y_0 - \delta, Y_0 + \delta)$ 为 Y_0 的概率为 $1 - \alpha$ 的单(上)侧预测区间; 类似地可以定义其下侧预测区间. 给定 α , 试求 Y_0 的上下两个单侧预测区间的表达式.

4. 对于(8.94)和(8.98)式, 试证明矩阵 $X'X$ 可逆当且仅当 L 可逆.

第 9 章

主成分分析与典型相关分析

同时考察多个（通常指两个以上）随机变量之间的相互关系是多元统计学研究的内容，前面一章介绍的多元线性回归分析就是一个例子。本章简要介绍另外两种在实际应用当中比较常见的用来研究多变量间相关关系的统计方法，即主成分分析和典型相关分析，前者研究一组变量内部的相关关系，而后者则用于两组变量之间相关性的分析。

§ 9.1 主成分分析

一、问题的提出

在许多实际问题当中，为了更加详尽地把握研究对象的信息，充分地反映个体之间的差异性，往往需要尽可能多地使用一些变量来分别测度个体不同方面的属性和特征。在描述统计中称这样的变量为标志。由于每种标志变量侧重不同方面的信息，标志变量的多样化就是全面展示个体特性的一种自然要求。但是，变量的增加势必会增加对问题进行统计分析时的复杂程度，因此，在力争保留原有变量所包含的信息的前提下尽可能地减少变量的数目就是十分必要的。

举例来说，大家知道为了全面考察一个学生的学习成绩，仅仅根据一门课的分数的来进行评判显然是有失公平的，为此通常需要选择多门课来对其进行综合评价。表 9-1 中列出了某所大学一个班的学生在第一学年全部必修课的考试成绩，这 13 门课程的分数分别测度了一个学生对不同方面的基础知识的学习能力及掌握程度，将其做为学生学习成绩的一种信息反映是合理的也是具有代表意义的。但是，直接使用这 13 门课的成绩将这些学生按照学习成绩的优劣编排一个次序却是比较困难的，它实际上是一个对这 32 个 13 维向量排序的问题。由于在二维以上的欧氏空间中不存在类似实数轴上的完备序结构，这一问题是非常困难的。通常的作法是综合这 13 门课程的成绩产生一个新的变量，并利用这个新的变量取值的大小来对学生排序。产生新变量的方式通常有求和、

表 9-1 某班学生第一学年全部必修课成绩

编号	K 1	K 2	K 3	K 4	K 5	K 6	K 7	K 8	K 9	K 10	K 11	K 12	K 13	平均	第一主分量得分
001	82	89	97	92	90	99	94	92	84	86	90	91	93	90.69	3.6679
002	85	90	92	94	88	98	95	94	87	76	94	83	84	89.23	2.8732
003	79	78	92	97	91	92	94	78	87	86	87	87	92	87.69	2.6345
004	89	92	92	89	86	88	88	73	86	81	95	90	90	87.62	2.2763
005	80	87	90	90	87	92	88	84	81	84	92	79	86	86.15	1.6120
006	83	74	88	97	91	82	92	77	82	84	87	82	99	86.00	1.9945
007	87	91	87	80	87	80	92	80	79	90	92	82	85	85.54	1.3732
008	92	84	81	88	85	82	94	83	84	87	89	84	78	85.46	1.4205
009	85	77	86	96	77	71	95	84	81	88	86	86	98	85.38	1.4559
010	80	90	92	78	82	91	84	87	86	73	91	85	85	84.92	1.4291
011	89	91	92	79	83	69	95	80	87	82	94	86	76	84.85	1.3099
012	91	81	94	92	78	82	86	77	87	70	85	84	86	84.08	1.2141
013	85	89	83	85	84	79	89	73	77	90	92	84	82	84.00	0.6399
014	85	90	88	93	88	85	95	64	77	60	92	75	96	83.69	0.6367
015	80	78	92	85	84	90	86	80	71	89	77	86	87	83.46	1.0227
016	83	92	86	93	85	66	80	69	84	89	79	89	89	83.38	0.8656
017	83	82	80	69	81	96	95	86	80	86	88	82	74	83.23	0.5451
018	80	86	85	96	81	80	80	76	82	74	81	86	91	82.92	0.6473
019	81	91	93	71	83	84	87	78	85	67	87	82	78	82.08	0.4798
020	78	95	81	80	82	92	80	74	68	75	96	78	78	81.31	- 0.6832
021	83	89	88	73	81	79	87	72	85	68	96	74	81	81.23	- 0.2417
022	74	89	79	87	86	78	92	73	74	72	88	87	74	81.00	- 0.3004
023	69	89	75	96	90	48	95	70	68	90	90	84	84	80.62	- 0.6462
024	81	93	86	66	71	77	89	81	80	64	91	90	72	80.08	- 0.5931
025	73	84	77	91	73	63	85	68	75	68	87	83	91	78.31	- 1.4761
026	80	90	91	48	79	57	90	67	87	70	91	84	70	77.23	- 1.2066
027	73	96	78	88	79	54	89	60	69	63	91	72	83	76.54	- 2.4788
028	67	90	80	86	74	57	83	60	60	78	86	69	81	74.69	- 3.3274
029	79	85	88	65	75	61	88	60	68	70	87	78	64	74.46	- 2.7511
030	68	81	70	73	67	57	81	62	50	60	87	78	63	69.00	- 5.2961
031	68	0	80	82	80	75	92	81	68	66	0	76	82	65.38	- 3.0483
032	30	0	87	60	68	55	94	66	65	86	0	82	61	57.92	- 6.0493

说明: (1) K 1~ K 7 分别表示第一学期 7 门必修课程成绩, K 8~ K 13 分别表示第二学期 6 门必修课程成绩;

(2) K 2 和 K 11 分别表示两个学期的体育课程成绩.

(3) 其中“ 平均 ” 项是按照算术平均求得的每个学生 13 门课程的平均成绩, “ 编号 ” 是按照 “ 平均 ” 项的大小排序产生的.

$$= \mathbf{L} \cdot \quad \cdot \mathbf{L} \quad (9.5)$$

如果 $Y_1 = l_1 \cdot X$ 满足 $l_1 \cdot l_1 = 1$ 且使 $D(Y_1)$ 取得最大值, 则称 Y_1 为第一主成分或者第一主分量, 这表明消除常数的作用后 Y_1 是 X 的所有线性组合中方差最大的一个.

对于 $Y_2 = l_2 \cdot X$, 如果 $l_2 \cdot l_2 = 1$ 且在所有与 Y_1 不相关的 X 的线性组合中 Y_2 是方差最大的一个, 则称 Y_2 是 X 的第二主成分.

类似地称 Y_i 是 X 的第 i 个主成分, 如果满足 $l_i \cdot l_i = 1$ 而且在与前 $i-1$ 个主成份 Y_1, Y_2, \dots, Y_{i-1} 都不相关的所有 X 的线性组合当中 Y_i 是方差最大的一个.

按照以上定义, 如果 Y_1, Y_2, \dots, Y_p 构成 X 的 p 个主成分, 则 Y 的协方差矩阵 Σ_Y 将具有非常特殊的结构, 它首先是一个对角阵, 即 Σ_Y 对角线以外的元素均为零; 另外, Σ_Y 的对角线上的 p 个元素满足 $\lambda_1, \lambda_2, \dots, \lambda_p > 0$; 而且可以证明有如下更为深入的结论成立:

定理 9.1 设 X 的协方差矩阵 Σ_X 的特征值为 $\lambda_1, \lambda_2, \dots, \lambda_p, 0, e_1, e_2, \dots, e_p$ 是与之相应的标准正交特征向量, 则 X 的第 i 个主成分为

$$Y_i = e_i \cdot X, \quad i = 1, 2, \dots, p \quad (9.6)$$

这样一来,

$$\Sigma_Y = D(Y_i) = e_i \cdot X \cdot e_i = \lambda_i, \quad i = 1, 2, \dots, p \quad (9.7)$$

即 Σ_Y 的对角线上的元素正好是 Σ_X 的 p 个特征根 (重根按重数计入) 而 L 正是依次以 Σ_X 的 p 个标准正交特征向量为列向量组的正交矩阵, 记 $L = (e_1, e_2, \dots, e_p)$. 因此, 从线性代数的角度来看, 可以说求 X 的主成分实际上就是对 X 的协方差矩阵 Σ_X 实施正交相似对角化的过程.

有关定理 9.1 的证明可以参照书后所附的有关文献, 在此略去.

由于相似矩阵具有相同的迹, 所以 Σ_X 与 Σ_Y 的迹是相等的, 即 $\text{tr}(\Sigma_X) = \text{tr}(\Sigma_Y)$, 从而

$$\begin{aligned} \text{tr}(\Sigma_X) &= \lambda_1 + \lambda_2 + \dots + \lambda_p \\ &= D(X_1) + D(X_2) + \dots + D(X_p) \\ &= \text{tr}(\Sigma_X) \\ &= \lambda_1 + \lambda_2 + \dots + \lambda_p \\ &= D(Y_1) + D(Y_2) + \dots + D(Y_p) \end{aligned} \quad (9.8)$$

因此, p 个主成分变量所反映的总变异与原来的 p 个变量 X_1, X_2, \dots, X_p 所反映的总变异是相同的. 如果矩阵 Σ_X 的最大的几个特征值占了全部特征值的较大比重, 则相应的前面几个主成分变量就概括了总变异中的大部分, 使用满足这种条件的少数主成分变量来取代原有的 p 个变量就不至于损失太多的信息. 为此, 引入以下定义:

定义 9.1 称 $\lambda_i / \sum_{k=1}^p \lambda_k$ 为第 i 个主成分 Y_i 的方差贡献率, 称 $\sum_{k=1}^i \lambda_k / \sum_{k=1}^p \lambda_k$ 为前 i 个主成分的累计方差贡献率 ($i=1, 2, \dots, p$).

通常情况下, 选取前面少数几个主成分使其累计方差贡献率达到 70% ~ 90%, 用其取代原有的 p 个变量就不会有太多的信息损失并可以达到压缩变量的目的.

例 9.1 假设已经求得三个原始变量 X_1, X_2, X_3 的协方差矩阵是

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 7 \end{pmatrix}$$

易知它的三个特征根和相应的特征向量分别为

$$\begin{aligned} \lambda_1 &= 7.1985, & e_1 &= (0.0314, 0.1946, 0.9804); \\ \lambda_2 &= 2.4631, & e_2 &= (-0.5552, -0.8122, 0.1790); \\ \lambda_3 &= 0.3384, & e_3 &= (0.8311, -0.5499, 0.0825); \end{aligned}$$

由定理 10.1 可知三个主成分分别是

$$\begin{aligned} Y_1 &= 0.0314X_1 + 0.1946X_2 + 0.9804X_3; \\ Y_2 &= -0.5552X_1 - 0.8122X_2 + 0.1790X_3; \\ Y_3 &= 0.8311X_1 - 0.5499X_2 + 0.0825X_3; \end{aligned}$$

又因为 $\lambda_1 + \lambda_2 + \lambda_3 = \text{tr}(\Sigma) = 10$, 这样可以方便地求得 Y_1 的方差贡献率为 71.985%, Y_1, Y_2 的累计方差贡献率为 96.616%, 因此使用 Y_1, Y_2 两个主成分取代原来的三个变量 X_1, X_2, X_3 基本不会损失太多的信息. 在有些情况下甚至可以直接使用第一主成分 Y_1 来取代原来的三个变量, 这可以视问题的具体要求而定.

在该例中, 不难看出第一主成分 Y_1 主要体现的是变量 X_3 的作用, 这决不是偶然的, 事实上由于三个原始变量中 X_3 的方差最大, 它已经概括了三个变量总变差的大部分, 因此在第一主成分中它占居主导作用是必然的.

在许多实际问题中, 由于不同变量采用的量纲不同, 它们之间的方差可能相差很大, 类似于上面的例子, 这样直接使用它们得到的主成分可能会表现出明显的倾向性, 不利于对问题的分析, 为此, 对于方差变化很大的原始变量 X_1, X_2, \dots, X_p 可以事先进行如下的标准化处理:

$$Z_i = \frac{X_i - E(X_i)}{S_{td}(X_i)}, \quad i=1, 2, \dots, p \quad (9.9)$$

这里 $S_{td}(X_i)$ 是变量 X_i 的标准差, $i=1, 2, \dots, p$.

由于

$$\text{cov}(Z_i, Z_j) = \frac{\text{cov}(X_i, X_j)}{\sqrt{s_{ii}} \cdot \sqrt{s_{jj}}} = r_{ij}, i, j = 1, 2, \dots, p \tag{9.10}$$

其中 r_{ij} 是变量 X_i 与 X_j 的相关系数. 这样一来, $Z = (Z_1, Z_2, \dots, Z_p)$ 的协方差矩阵就是 X 的相关矩阵 R .

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix} \tag{9.11}$$

完全类似地, 根据 Z 的协方差阵 R 的特征值 $\lambda_1, \lambda_2, \dots, \lambda_p \geq 0$ 和相应的特征向量 e_1, e_2, \dots, e_p 可以得到 p 个主成分为

$$Y_i = e_i' Z = \sum_{k=1}^p e_{ki} Z_k = \sum_{k=1}^p \frac{e_{ki}(X_k - E(X_k))}{\sqrt{s_{kk}}}, i = 1, 2, \dots, p. \tag{9.12}$$

例 9.2 由例 9.1 中的 X_1, X_2, X_3 的协方差矩阵 可以求得它们的相关矩阵是

$$R = \begin{pmatrix} 1 & 0.7071 & 0 \\ 0.7071 & 1 & 0.2673 \\ 0 & 0.2673 & 1 \end{pmatrix}$$

其特征值及对应的特征向量分别为

$$\begin{aligned} \lambda_1 &= 1.7559, & e_1 &= (0.6614, 0.7071, 0.2500); \\ \lambda_2 &= 1.0000, & e_2 &= (-0.3536, 0.0000, 0.9354); \\ \lambda_3 &= 0.2441, & e_3 &= (0.6614, -0.7071, 0.2500); \end{aligned}$$

由上述结果及(9.12)式可以得到三个主成分为

$$\begin{aligned} Y_1 &= 0.6614(X_1 - \mu_1) + 0.5(X_2 - \mu_2) + 0.0945(X_3 - \mu_3); \\ Y_2 &= -0.3536(X_1 - \mu_1) + 0.3535(X_3 - \mu_3); \\ Y_3 &= 0.6614(X_1 - \mu_1) - 0.5(X_2 - \mu_2) + 0.0945(X_3 - \mu_3); \end{aligned}$$

其中的 μ_1, μ_2, μ_3 分别代表变量 X_1, X_2, X_3 的期望.

此时, 由于 Z_1, Z_2, Z_3 三个标准化后的变量的方差均为 1, 所以 Y_1 的方差贡献率为 $\lambda_1/3 = 58.53\%$, Y_1 与 Y_2 的累积方差贡献率为 $(\lambda_1 + \lambda_2)/3 = 91.86\%$.

与例 9.1 中使用协方差矩阵求得的主成分相比, 使用相关矩阵得到的第一主成分更加强调了 X_1 与 X_2 的作用, 由此可见, 使用两种方法得到的主成分变量往往具有较大差异, 在实用过程中应该对此加以注意.

三、样本的主成分分析

假设从总体中随机地抽取 n 个个体, 并对每一个个体都分别测定其 p 个变

量 $X = (X_1, X_2, \dots, X_p)$ 的取值, 得到的观测数据可用下面的矩阵表示

$$Q = \begin{pmatrix} X_{11} & X_{21} & \dots & X_{p1} \\ X_{12} & X_{22} & \dots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{2n} & \dots & X_{pn} \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}.$$

这里以 $X_k = (X_{1k}, X_{2k}, \dots, X_{pk})$ 表示第 k 个个体相应的 p 个变量的观测值.

由上述样本观测值, 可以得到 p 个变量 X_1, X_2, \dots, X_p 的样本均值和样本方差分别为

$$\bar{X}_i = \frac{1}{n} \sum_{k=1}^n X_{ik} \quad (9.13)$$

$$S_i^2 = \frac{1}{n-1} \sum_{k=1}^n (X_{ik} - \bar{X}_i)^2, \quad i = 1, 2, \dots, p \quad (9.14)$$

变量 X_i 与 X_j 的样本协方差为

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^n (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j), \quad i, j = 1, 2, \dots, p \quad (9.15)$$

于是 X 的样本协方差矩阵可以表示为

$$= (s_{ij})_{p \times p} = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})(X_k - \bar{X})'. \quad (9.16)$$

其中 $\bar{X} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)$ 表示 X 的样本均值向量.

类似地, X 的样本主成分也是 X_1, X_2, \dots, X_p 的一些特殊的线性组合, 为此, 设

$$y_j = l_j' X = \sum_{i=1}^p l_{ij} X_i, \quad j = 1, 2, \dots, p \quad (9.17)$$

是 p 个线性组合, 则 y_j 对应 n 个个体的观测值分别为

$$y_{jk} = l_j' X_k, \quad k = 1, 2, \dots, n; \quad j = 1, 2, \dots, p \quad (9.18)$$

于是 y_j 的样本均值是

$$\bar{y}_j = \frac{1}{n} \sum_{k=1}^n y_{jk} = l_j' \bar{X}, \quad j = 1, 2, \dots, p. \quad (9.19)$$

而 y_i 与 y_j 的样本协方差为

$$\begin{aligned} s_{ij} &= \frac{1}{n-1} \sum_{k=1}^n (y_{ik} - \bar{y}_i)(y_{jk} - \bar{y}_j) \\ &= \frac{1}{n-1} \sum_{k=1}^n l_i' (X_k - \bar{X})(X_k - \bar{X})' l_j \\ &= l_i' s_{ij} l_j, \quad i, j = 1, 2, \dots, p \end{aligned} \quad (9.20)$$

因此, y_1, y_2, \dots, y_p 的样本协方差矩阵为

$$= L' s_{ij} L, \quad (9.21)$$

其中 $L = (l_1, l_2, \dots, l_p)$ 是 p 个线性组合(9.17)的系数矩阵.

完全类似前面定义总体主成分的过程可以定义样本的主成分, 在此将其综合成如下的定义.

定义 9.2 称线性组合 $y_1 = l_1 X$ 是 X 的第一个样本主成分, 如果满足 $l_1 \cdot l_1 = 1$ 且 y_1 的样本方差最大; 假设 y_1, y_2, \dots, y_{i-1} 是 X 的前 $i-1$ 个样本主成分, 则称 y_i 是 X 的第 i 个样本主成分, 如果满足下列条件

$$l_i \cdot l_i = 1;$$

y_i 与 y_1, y_2, \dots, y_{i-1} 的样本协方差均为 0;

在 , 条件下, y_i 的样本方差取得最大值.

同样地, 对上面定义的样本主成分, 也成立类似于定理 10.1 的结论.

定理 9.2 设 X 的样本协方差矩阵 的特征值为 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, e_1, e_2, \dots, e_p 是与之对应的标准正交特征向量, 则 X 的第 i 个样本主成分为

$$y_i = e_i \cdot X, \quad i = 1, 2, \dots, p. \quad (9.22)$$

不难看出, 此时 y_i 的样本方差是 λ_i , $i = 1, 2, \dots, p$, y_i 与 y_j 的样本协方差等于 0 ($i \neq j$ 时), 而且

$$\text{tr}(\Sigma) = \sum_{i=1}^p \lambda_i = \sum_{i=1}^p \text{tr}(\lambda_i e_i e_i^T) = \text{tr}(\Sigma),$$

即总的样本方差相等. 因此, 对于样本主成分, 同样地可以称 $\lambda_i / \sum_{j=1}^p \lambda_j$ 是第

i 个样本主成分的 (样本) 方差贡献率, 称 $\sum_{j=1}^i \lambda_j / \sum_{j=1}^p \lambda_j$ 为前 i 个样本主成分的累计 (样本) 方差贡献率.

与总体主成分的情况一样, 在 X_1, X_2, \dots, X_p 的样本方差变化很大时, 可以事先对其进行标准化, 即对原始的观测数据作变换

$$Z_{ik} = \frac{X_{ik} - \bar{X}_i}{S_i}, \quad k = 1, 2, \dots, n; \quad i = 1, 2, \dots, p \quad (9.23)$$

为了讨论问题的方便, 可以将上述变换之后的数据看成 $Z = (Z_1, Z_2, \dots, Z_p)$ 对 n 个个体的观测数据, 其中

$$Z_i = \frac{X_i - \bar{X}_i}{S_i}, \quad i = 1, 2, \dots, p \quad (9.24)$$

易知 Z 的样本协方差矩阵就是 X 的样本相关矩阵 R , 利用 R 的特征值和特征向量可以求得 Z 从而求得 X 的样本主成分.

同样, 使用样本相关矩阵 R 和使用样本协方差矩阵 求得的 X 的样本主成分通常会有一定的差别.

无论对于通过上述哪种方式得到的主成分 y_1, y_2, \dots, y_p , 均可以定义其主分量得分.

定义 9.3 设 $y_i = e_i X$, $i = 1, 2, \dots, p$ 是样本主分量, 将第 k 次观测值 X_k 代入, 得到

$$y_{ik} = e_i \cdot X_k, \quad i = 1, 2, \dots, p; \quad k = 1, 2, \dots, n.$$

则称 $y_k = (y_{1k}, y_{2k}, \dots, y_{pk})$ ($k = 1, 2, \dots, n$) 为主分量得分.

换向话说, 主分量得分即是以主分量变量取代了原有变量之后对 n 个个体的观测结果.

至于根据样本主成分进行的统计推断等问题超出了本书的范围. 可另参考有关文献.

例 9.3 考虑表 9-1 中给出的学生成绩的观测数据. 试求一下其中的 13 门课程成绩 (分别以 k_1, k_2, \dots, k_{13} 表示) 的主成分.

经过计算可以看出, 这 13 门课程成绩的样本方差相差很大, 比如两门体育课成绩 k_2 和 k_{11} 的样本方差分别是 $S_2^2 = 487.7167$, $S_{11}^2 = 497.9194$, 而 k_6 的样本方差是 $S_6^2 = 24.67339$, 因此利用样本相关矩阵来求主成分.

表 9-2 给出了样本相关矩阵的最大的 11 个特征值即前面 11 个主成分的样本方差, 以及各个主成分的方差贡献率和累积方差贡献率.

表 9-2 13 门课程成绩的前八个主成分及其方差情况

样本主成分	方 差	方差贡献率	累计方差贡献率
y_1	5.12768	0.39444	0.39444
y_2	2.30022	0.17694	0.57138
y_3	1.58259	0.12174	0.69312
y_4	0.99401	0.07646	0.76958
y_5	0.85855	0.06604	0.83562
y_6	0.67087	0.05161	0.88723
y_7	0.52022	0.04002	0.92725
y_8	0.29858	0.02297	0.95022

由此可见, 前面 8 个样本主成分的累次方差贡献率超过了 95%, 使用其取代原来的 13 门课程成绩变量不会造成太多信息损失, 事实上, 由于前面三个主成分的累积方差贡献率也已接近 70%, 在具体问题中可以采用更少的主成分变量来取代原来的变量.

从最大特征根 $\lambda_1 = 5.12768$ 所对应的标准正交化特征向量可以写出第一个样本主分量的表达式为

$$y_1 = 0.3419z_1 + 0.2135z_2 + 0.2712z_3 + 0.2307z_4 + 0.3492z_5 + 0.3281z_6 + 0.1134z_7 + 0.3126z_8 + 0.3602z_9 + 0.1914z_{10} + 0.2343z_{11} + 0.2424z_{12} + 0.3009z_{13} \tag{9.25}$$

可以看到, 在 y_1 的表达式中 k_1, k_2, \dots, k_{13} 的系数均为正数且相差不是太大, 由此说明 y_1 基本上体现了各门课程的一种平均水平的测度. 在表 9-1 的最后一列, 我们给出对应每个学生第一样本主分量的得分, 将其与学生十三门课程的平均成绩相对比, 不难发现, 按两者大小对学生排出的次序是比较吻合的.

事实上, 可以证明, 在所有不同的两门课程成绩的样本相关系数都相等的特殊情况下, 按第一主成分得分对学生排出的名次与按平均成绩排出的名次是完全一致的. 参见课后习题 3.

§ 9.2 典型相关分析

一、问题的提出

在前面一章中我们已经知道, 利用相关系数可以衡量两个随机变量之间的(直线)相关关系, 对于一个随机变量同一组随机变量之间的相关关系, 可以应用在多元线性回归模型中引入的多重相关系数来衡量. 除此之外, 在很多实际问题中, 通常还需要考虑一组随机变量同另一组随机变量之间相关关系的测度问题.

譬如, 面对表 9-1 提供的两个学期的两组成绩, 有人可能会问, 第一学期的 7 门课程成绩与第二学期的 6 门课程成绩之间具有怎样的相关性? 一般性的常识会告诉我们, 这两组成绩之间应该具有较强的相关性, 通过两组成绩的样本协方差矩阵只能揭示它们彼此两两之间的相关程度, 而不能将其中的任何一个作为两组整体变量(向量)之间相关性的度量.

基于主成分的思路, 可以分别对两组变量进行综合, 选取它们各自的一个特殊的线性组合作为代表该组变量的一个新变量, 并且使得这一对新变量之间呈现出最大的相关性, 这样通过一对新变量之间的相关系数就基本刻画了两组变量之间相关的程度. 这就是典型相关分析的思想, 它是由 H. Hotelling (1936) 引入的, 通常称这样的一对新变量为典型相关变量. 如果一对典型相关变量不能完全刻画两组变量之间的相关性, 还可以继续寻找第二对、第三对... 典型相关变量, 直至新一对变量之间的相关性为 0 时为止.

二、总体的典型相关分析

考虑两组随机变量 $X = (x_1, x_2, \dots, x_p)$, $Y = (y_1, y_2, \dots, y_q)$, 不妨假设 $p \geq q$, 而且 $E(x_i) = 0, E(y_j) = 0, i = 1, 2, \dots, p; j = 1, 2, \dots, q$, 否则只需考虑每个变量减去其期望值之后得到的新变量即可, 这样并不影响它们之间的相关性.

记 X 的协方差矩阵为 Σ_{xx} , Y 的协方差矩阵为 Σ_{yy} , X 与 Y 的(互)协方差

矩阵为 Σ_{xy} ; 令 $Z = (x_1, x_2, \dots, x_p; y_1, y_2, \dots, y_q)$, 则 Z 的协方差矩阵可以表示为

$$\Sigma_Z = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \quad (9.26)$$

其中 $\Sigma_{yx} = \Sigma_{xy}$ 是 y 与 x 的协方差矩阵.

如果 $\Sigma_{xy} = 0$, 说明 x 与 y 的各分量之间两两不相关, 对于这种特殊情况没有研究的必要. 所以, 以下均假设 $\Sigma_{xy} \neq 0$.

为了对每一组寻找一种特殊的线性组合作为代表变量, 设

$$\begin{aligned} u_1 &= l_1' x \\ v_1 &= m_1' y \end{aligned} \quad (9.27)$$

我们希望寻找合适的系数向量 l_1 和 m_1 使得 u_1, v_1 之间具有最大的相关性, 即 (u_1, v_1) 达到最大.

考虑到对任意常数 a, b, c, d , $ac > 0$ 时, 总有

$$(au_1 + b, cv_1 + d) = (u_1, v_1) \quad (9.28)$$

因此不妨设 u_1, v_1 都满足方差等于 1 的条件, 也就是说,

$$\begin{aligned} D(u_1) &= l_1' \Sigma_{xx} l_1 = 1, \\ D(v_1) &= m_1' \Sigma_{yy} m_1 = 1, \end{aligned} \quad (9.29)$$

此时,

$$(u_1, v_1) = \text{cov}(u_1, v_1) = l_1' \Sigma_{xy} m_1, \quad (9.30)$$

这样在满足 (9.29) 的约束条件下求 (9.30) 式的最大值即得到一组 l_1 和 m_1 , 从而得到 u_1, v_1 , 它们即是前面提到的两组变量的一对综合变量, 称之为 x 和 y 的第一对典型相关变量, 它们之间的相关系数 (u_1, v_1) 是 x 和 y 各自的线性组合之间最大的相关系数, 称为第一典型相关系数.

有些情况下, 仅用第一对典型相关变量还不足以完全描述两组变量 x 和 y 之间的相关性, 这表现在存在 x 与 y 的各自的一个线性组合, 尽管它们与 u_1, v_1 都不相关, 但是它们之间是相关的. 此时, 需要引入第二对典型相关变量 u_2, v_2 , 不妨设

$$u_2 = l_2' x, \quad v_2 = m_2' y \quad (9.31)$$

对于 u_2, v_2 , 不仅需要满足 $D(u_2) = D(v_2) = 1$, 而且需要 $\text{cov}(u_2, u_1) = \text{cov}(u_2, v_1) = \text{cov}(v_2, u_1) = \text{cov}(v_2, v_1) = 0$ 以保证它们与第一对典型变量之间不存在相关性, 在上述约束条件下求解 (u_2, v_2) 的最大值即可得到 l_2 及 m_2 , 从而得到第二对典型相关变量 u_2, v_2 .

类似地, 在前面两对典型变量仍然不足以完全地反映 x 与 y 的相关关系时, 可以继续定义第三对, 第四对...直至第 p 对典型相关变量.

综上所述, 可以将各对典型相关变量的定义归结如下:

定义 9.4 由 (9.27) 式定义的一对新变量 (u_1, v_1) 在方差为 1 的条件下如果使得它们的相关系数最大, 则称其为 x 和 y 的第一对典型相关变量; 假设已经定义了前面 $k-1$ 对典型相关变量 $(u_1, v_1), (u_2, v_2), \dots, (u_{k-1}, v_{k-1})$. 那么第 k 对典型相关变量 (u_k, v_k) 定义为

$$u_k = l_k \cdot x, \quad v_k = m_k \cdot y$$

而且满足下面三个条件

$$(i) D(u_k) = l_k' x x l_k = 1, D(v_k) = m_k' y y m_k = 1;$$

$$(ii) \text{cov}(u_k, u_j) = 0, j = 1, 2, \dots, k-1;$$

$$\text{cov}(u_k, v_j) = 0, j = 1, 2, \dots, k-1;$$

$$\text{cov}(v_k, u_j) = 0, j = 1, 2, \dots, k-1;$$

$$\text{cov}(v_k, v_j) = 0, j = 1, 2, \dots, k-1;$$

$$(iii) \text{ 在上面两个条件的约束下, } (u_k, v_k) = l_k' x y m_k \text{ 取得最大值.}$$

根据上述定义, 求解典型相关变量的问题实质上可以化归成一个条件极值问题, 下面以对 (u_1, v_1) 的求解为例.

由 (9.27) 式, 求 (u_1, v_1) 即是求其系数向量 l_1 及 m_1 . 利用拉格朗日乘数法, 构造如下的目标函数

$$(l_1, m_1) = l_1' x y m_1 - \frac{\mu}{2} (l_1' x x l_1 - 1) - \frac{\mu}{2} (m_1' y y m_1 - 1) \quad (9.32)$$

对关于 l_1 和 m_1 分别求得, 并令相应的一阶偏导数为 0, 则得到方程组

$$\begin{aligned} x y m_1 - x x l_1 &= 0 \\ y x l_1 - \mu y y m_1 &= 0 \end{aligned} \quad (9.33)$$

对上述第一个方程左乘 l_1 , 第二个方程左乘 m_1 , 并利用 (9.29) 则有

$$= l_1' x y m_1 = m_1' y x l_1 = \mu \quad (9.34)$$

即两个拉格朗日乘子是相等的.

下面假设 $x x$ 、 $y y$ 均为正定矩阵, 这样从 (9.33) 中的第二个方程可得

$$m_1 = \frac{1}{y y^{-1}} y x l_1 \quad (9.35)$$

将其代入第一个方程, 得到

$$x y \frac{1}{y y^{-1}} y x l_1 - x x l_1 = 0,$$

即

$$l_1 = \frac{1}{\lambda} l_1 \quad (9.36)$$

其中 $\lambda = \frac{1}{x x^{-1}} x y \frac{1}{y y^{-1}} y x$.

(9.36) 式说明 λ 是 $\frac{1}{x x^{-1}} x y \frac{1}{y y^{-1}} y x$ 的特征根, l_1 是相应的特征向量.

由于

$$\frac{1}{x x} \frac{1}{x x^{-1}} = \frac{1}{x x^{-1}} \frac{1}{x y} \frac{1}{y y^{-1}} \frac{1}{y x} \frac{1}{y y} \frac{1}{y x^{-1}} \quad T T.$$

其中 $T = \begin{pmatrix} -\frac{1}{xx^2} & x \\ -\frac{1}{yy^2} & y \end{pmatrix}$, 所以 λ 的特征根都是非负实根. 将这些特征根由大至小排列次序并记为

$$\lambda_1^2, \lambda_2^2, \dots, \lambda_p^2, 0 \quad (9.37)$$

考虑到 (9.30) 以及 (9.34), 易见 λ_1 就是 u_1, v_1 的相关系数, 所以应将 λ_1^2 选为 λ 的最大的一个特征根, l_1 即是最大特征根所对应的特征向量. m_1 可由 (9.35) 式求得.

将 l_1, m_1 代入 (9.27) 式, 即可得到第一对典型相关变量 u_1, v_1 , 它们的相关系数 λ_1 是 λ_1^2 的算术平方根.

类似地, 对于一般情况, 可以证明下面的定理成立.

定理 9.3 x 与 y 的第 k 对典型相关变量 (u_k, v_k) 之间的相关系数 λ_k 是矩阵 λ 的第 k 个特征根的算术平方根, 对应的特征向量即是 u_k 的系数向量 l_k , v_k 的系数向量 m_k 可由下式求得

$$m_k = \frac{1}{\lambda_k} \begin{pmatrix} -1 \\ yy^{-1} \end{pmatrix} yx l_k \quad (9.38)$$

根据定理 9.3, 如果 λ 的特征根中只有 r 个非零 ($r < p$), 那么 x 与 y 之间只存在 r 对典型相关变量.

三、样本的典型相关分析

在实际问题中, 由于通常 μ 未知, 只能根据抽取的值对其进行估计得到样本协方差矩阵 S , 基于 S 求得的典型相关变量和典型相关系数分别称为样本典型相关变量和样本典型相关系数.

设 $X_k = (X_{1k}, X_{2k}, \dots, X_{pk})$, $Y_k = (Y_{1k}, Y_{2k}, \dots, Y_{qk})$, ($k = 1, 2, \dots, n$) 是两组变量的样本观测值. 类似于 (9.16) 式, 定义样本协方差矩阵为

$$\begin{aligned} S_{xx} &= \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})(X_k - \bar{X})'; \\ S_{yy} &= \frac{1}{n-1} \sum_{k=1}^n (Y_k - \bar{Y})(Y_k - \bar{Y})'; \\ S_{xy} &= \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})(Y_k - \bar{Y})'; \\ S_{yx} &= S_{xy}' \end{aligned}$$

其中, \bar{X}, \bar{Y} 分别是 X, Y 的样本均值向量.

分别将 $S_{xx}, S_{xy}, S_{yx}, S_{yy}$ 代替 $\Sigma_{xx}, \Sigma_{xy}, \Sigma_{yx}, \Sigma_{yy}$ 代入 (9.38) 得到其样本值

$$m_i = \frac{1}{\lambda_i} \begin{pmatrix} -1 \\ yy^{-1} \end{pmatrix} yx \cdot l_i, \quad i = 1, 2, \dots, p \quad (9.39)$$

求 λ 的特征值 $\lambda_1^2, \lambda_2^2, \dots, \lambda_p^2, 0$ 及相应的特征向量 l_1, l_2, \dots, l_p , 并利用

$$m_i = \frac{1}{\lambda_i} \begin{pmatrix} -1 \\ yy^{-1} \end{pmatrix} yx \cdot l_i, \quad i = 1, 2, \dots, p \quad (9.40)$$

求得相应的 m_1, m_2, \dots, m_p , 其中 i 是 λ_i^2 的算术平方根.
令

$$u_i = l_i x, v_i = m_i y, i = 1, 2, \dots, p, \tag{9.41}$$

则 (u_i, v_i) 即 x, y 的第 i 对样本典型相关变量, 它们的相关系数 ρ_i 即第 i 个样本典型相关系数, $i = 1, 2, \dots, p$.

实际的计算中往往先对样本数据进行标准化处理, 见(9.23)式, 这样一来实际上是以样本相关矩阵 R 代替样本协方差矩阵

$$\begin{matrix} & \begin{matrix} x & x \\ x & y \end{matrix} \\ = & \begin{matrix} y & x \\ y & y \end{matrix} \end{matrix} \tag{9.42}$$

其中计算过程是相似的, 在此不再重复了.

例 9.4 下面我们讨论一下本节开始所提出的问题, 即对表 9-1 中给出的学生成绩, 分析其中第一学期 7 门课程成绩与第二学期 6 门课程成绩之间的相关性.

为方便起见, 记 $x = (k_1, k_2, \dots, k_7)$ $y = (k_8, k_9, \dots, k_{13})$. 表 9-3 给出了十三门课程样本相关矩阵 R 的一部分, 即 x 与 y 的样本 (互) 相关矩阵 R_{xy} , 从中可以看到, 两门体育课程 k_2 与 k_{11} 之间的相关系数最大.

表 9-3 两个学期成绩的样本互相关矩阵

	k8	k9	k10	k11	k12	k13
k1	0.4128	0.6412	0.0827	0.7128	0.2754	0.4587
k2	0.0103	0.3100	- 0.0191	0.9756	0.1316	0.2251
k3	0.4663	0.7219	0.1928	0.1089	0.4006	0.2909
k4	0.2292	0.1617	0.2992	0.2146	0.1349	0.8151
k5	0.4456	0.5355	0.4563	0.3378	0.2730	0.6080
k6	0.7584	0.5476	0.1901	0.2407	0.3148	0.4020
k7	0.3501	0.2576	0.2947	- 0.1337	0.1535	0.0740

根据样本相关矩阵 R 求得的各对典型相关变量的系数向量以及各个相应的典型相关系数列在表 9-4 中.

可以看出, 第一对样本典型相关变量之间显示出很强的相关性, 其相关系数为 0.985366, 它们反映了两个学期的各门课程的成绩之间相关的程度. 值得注意的是, 从 l_1 和 m_1 的取值中可以看出, 第一对典型变量主要是体现了两门体育课程成绩的作用, 在 u_1 中 k_2 的系数是 0.9648, 在 v_1 中 k_{11} 的系数是 0.9794, 明显大于其它课程成绩的系数, 事实上, 这也是 k_2 和 k_{11} 之间的相关系数最大的

一种体现.

表 9-4 两个学期成绩的典型相关分析结果

序号 i	典型相关 系数 λ_i	典型相关变量的系数向量 l_i 及 m_i
1	0.985366	$l_1 = (0.0863, 0.9648, - 0.0073, 0.0177, - 0.0925, 0.1094, 0.1326)$ $m_1 = (0.0427, 0.0396, - 0.0378, 0.9794, 0.0337, - 0.0128)$
2	0.897357	$l_2 = (0.2732, - 0.4856, 0.3223, 0.6428, 0.0865, 0.2002, 0.0435)$ $m_2 = (0.3714, 0.1529, 0.0664, - 0.3012, - 0.042, 0.7382)$
3	0.807288	$l_3 = (- 0.3549, 0.1988, - 0.1820, 0.7802, 0.1039, - 0.4606, - 0.2779)$ $m_3 = (- 0.521, - 0.7858, 0.1112, 0.1290, 0.2270, 0.8095)$
4	0.701973	$l_4 = (0.3058, - 0.2314, 0.8161, - 0.1764, 0.6633, - 1.24, - 0.2463)$ $m_4 = (- 1.1737, 1.1512, 0.3780, - 0.2943, 0.0198, - 0.065)$
5	0.395791	$l_5 = (- 0.675, 0.3213, - 0.4821, - 0.464, 1.075, 0.074, 0.3688)$ $m_5 = (0.1469, - 0.0581, 1.0991, 0.1938, - 0.5376, - 0.3820)$
6	0.137279	$l_6 = (- 1.3418, 0.9988, 0.7534, 0.4939, - 0.6763, 0.3289, 0.4531)$ $m_6 = (- 0.087, - 0.5066, 0.027, 0.0233, 1.1859, - 0.0891)$

习 题 九

1. 设随机变量 X_1, X_2, X_3 的协方差矩阵是
- $$= \begin{pmatrix} 1 & -3 & 0 \\ -3 & 7 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$
- 试求变量 X_1, X_2, X_3 的三个主成分, 并计算各个主成分的方差贡献率.
2. 设随机变量 X_1, X_2 的期望分别为 μ_1 和 μ_2 , 其标准差分别为 2 和 25, 又知 X_1 与 X_2 的相关系数是 0.32, 试分别根据 X_1 与 X_2 的协方差矩阵和相关矩阵来求它们的主成分及其方差贡献率.
3. 设随机变量 X_1, X_2, \dots, X_p 的相关矩阵为
- $$R = \begin{pmatrix} 1 & & \dots \\ & 1 & \dots \\ & & \dots & 1 \end{pmatrix}$$
- 即 X_1, X_2, \dots, X_p 之间是等相关的. 试写出它们的第一主成分和方差贡献率.
4. X_1, X_2, X_3, X_4 四个随机变量的样本相关矩阵为

$$R = \begin{pmatrix} 1 & & & \\ 0.7346 & 1 & & \\ 0.7108 & 0.6972 & 1 & \\ 0.7040 & 0.7086 & 0.8392 & 1 \end{pmatrix}$$

试求 (X_1, X_2) 与 (X_3, X_4) 的典型相关变量以及典型相关系数.

习题参考答案

习题一

(A)

9. 0.15, 0.5, 0.10, 0.5.

10. 0.7

11. (1) $\frac{C_{N-1}^K C_{N-N_1}^{n-K}}{C_N^n}$, (2) $1 - \frac{C_{N-N_1}^n}{C_N^n}$,

(3) $1 - \frac{C_{N-N_1}^n}{C_N^n} - \frac{C_{N_1}^1 C_{N-N_1}^{n-1}}{C_N^n}$

12. $\frac{24!}{P_{30}^{24}}$ (或 $\frac{1}{C_{30}^{24}}$), $\frac{1}{2}$

13. $1 - \frac{P_n^{365}}{365^n}$

14. (1) $1 - (\frac{8}{9})^{25}$, (2) $1 - (\frac{7}{9})^{25}$ (3) $1 - 2 \cdot (\frac{8}{9})^{25} + (\frac{7}{9})^{25}$

(4) $C_{25}^3 (\frac{1}{9})^3 (\frac{8}{9})^{22}$

15. (1) $\frac{1}{4}$ (2) $\frac{3}{8}$

16. $\frac{C_a^K C_b^{i-K}}{C_{a+b}^i}$

17. $\frac{4}{10}, \frac{1}{6}, \frac{5}{6}$

18. $C_{10}^3 \cdot \frac{1}{2^{10}}$

19. 0.25

20. $\frac{48}{13!}$

21. $\frac{3}{10}$

22. 0.121

23. $\frac{a-1}{a+b-1}, \frac{a-1}{a+b-1}, \frac{a(a-1)}{a(a+b)+a(b-1)}$

25. $\frac{1}{3}, \frac{1}{2}$

26. $\frac{1}{3}$

27. (1) $\frac{4}{10}$, (2) $\frac{24}{90}$, (3) $\frac{24}{720}$
28. 0.59
29. 0.056, $\frac{1}{18}$
30. 0.455, 0.14
31. 0.087, 0.49
38. 0.316
39. 0.63
40. 0.2286, 0.0497

习题一

(B)

1. $\frac{9}{19}$, $\frac{10}{19}$
2. $\frac{2(n-r-1)}{n(n-1)}$, $\frac{1}{n-1}$
3. $C_{2N-K}^N (\frac{1}{2})^{2N-K}$
4. $\frac{m^K - (m-1)^K}{n^K}$
5. 0.25
6. $\frac{p_1}{1-q_1q_2}$, $\frac{q_1p_2}{1-q_1q_2}$
7. $p = \frac{29}{90}$ $q = \frac{20}{61}$
8. 0.08, 0.6

习题二

(A)

- | | | | | | | | |
|------------|---------------|------------------|---------------|------------|---------------|------------------|---------------|
| 1. X = | 40 | 出现正面 | $\frac{1}{2}$ | Y = | 10 | 出现正面 | $\frac{1}{2}$ |
| | 20 | 出现反面 | $\frac{1}{2}$ | | 30 | 出现反面 | $\frac{1}{2}$ |
| | 0 | $x < 20$ | | | 0 | $x < 10$ | |
| $F_X(x) =$ | $\frac{1}{2}$ | $20 \leq x < 40$ | | $F_Y(x) =$ | $\frac{1}{2}$ | $10 \leq x < 30$ | |
| | 1 | $x \geq 40$ | | | 1 | $x \geq 30$ | |
2. (1) $\frac{1}{2(2^{100}-1)}$ (2) $\frac{1}{3}$

$$4. (1) P\{X = k\} = 0.3^{k-1} \cdot 0.7 \quad k = 1, 2, \dots$$

$$(2) \quad p_i = \frac{7}{10}, \frac{7}{30}, \frac{7}{120}, \frac{1}{120}$$

$$5. \frac{4}{5}, \frac{4}{5}, \frac{7}{10}$$

$$6. 1, 0.39$$

$$8. (1) \text{是}, f(x) = \begin{cases} 0 & x < 0 \\ 2x & 0 \leq x < 1 \\ 0 & x \geq 1 \end{cases} \quad (2) \text{不是}$$

$$9. (1) \frac{1}{2}, F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2}e^x & 0 \leq x < 1 \\ 1 - \frac{1}{2}e^{-x} & x \geq 1 \end{cases}, \quad 1 - \frac{1}{2}(e^{-\frac{x}{2}} - e^{-1})$$

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

$$11. \frac{3}{2}, -\frac{1}{5}$$

$$12. 10.37, 0.0291$$

$$13. 1073.6$$

$$17. (1) P\{X = k\} = C_n^k (0.94)^k (0.06)^{n-k} \quad (2) (0.94)^n$$

$$(3) 1 - (0.94)^n - n \cdot (0.94)^{n-1} \cdot 0.06 \quad (4) 0.06n, 0.0564n$$

$$18. pq^k$$

$$19. 1, 1, 2, \frac{1}{6e}$$

$$20. 0.352$$

$$21. 320000$$

$$24. 3e^{-2} - 2e^{-3}$$

$$26. 1 - 18.5e^{-5}$$

$$27. 0.06931, 0.083$$

$$28. 86.45$$

$$29. \quad \begin{matrix} 20 & 22 & 24 & 26 \\ p_i & 0.1 & 0.4 & 0.3 & 0.2 \\ & 100 & 121 & 144 & 169 \end{matrix}$$

$$p_i \quad 0.1 \quad 0.4 \quad 0.3 \quad 0.2$$

31.
$$F_{X^2}(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{x} & 0 < x < 1 \\ 1 & x \geq 1 \end{cases} \quad f_{X^2}(x) = \begin{cases} \frac{1}{x^2} & 0 < x < 1 \\ 0 & \text{其他} \end{cases}$$

32.
$$F_Y(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x} & x \geq 0 \end{cases}$$

$$f_Y(x) = \begin{cases} 0 & x < 0 \\ e^{-x} & x \geq 0 \end{cases}$$

33. $\frac{1}{24}(a+b)(a^2+b^2)$

34. 1000

习题二

(B)

2. (1)
$$f(x) = \begin{cases} \frac{1}{1000}e^{-\frac{x}{1000}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

(2)
$$f_x(x) = \begin{cases} \frac{n}{1000}[1 - e^{-\frac{x}{1000}}]^{n-1}e^{-\frac{x}{1000}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

(3)
$$f_Y(x) = \begin{cases} \frac{n}{1000}e^{-\frac{nx}{1000}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

3. 37

习题三

(A)

2. (1)

$X_2 \backslash X_1$	0	1	2	p_i^x
0	$\frac{1}{56}$	$\frac{5}{28}$	$\frac{5}{28}$	$\frac{21}{56}$
1	$\frac{5}{56}$	$\frac{5}{14}$	$\frac{5}{28}$	$\frac{35}{56}$
p_j^y	$\frac{3}{28}$	$\frac{15}{28}$	$\frac{5}{14}$	

(2) $\frac{5}{14}, \frac{21}{56}, \frac{13}{28}$

3. (1) $k_1 = 12, k_2 = 21$

$$(2) f_{X_1}(x) = \begin{cases} 3e^{-3x} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad f_{Y_1}(y) = \begin{cases} 4e^{-4y} & y \geq 0 \\ 0 & y < 0 \end{cases}$$

$$g_{X_2}(x) = \begin{cases} \frac{21}{4}e^{-3x}(1-e^{-4x}) & x \geq 0 \\ 0 & x < 0 \end{cases} \quad g_{Y_2}(y) = \begin{cases} 7e^{-7y} & y \geq 0 \\ 0 & y < 0 \end{cases}$$

$$4. (1) f(x, y) = \begin{cases} \frac{1}{2} & 0 \leq x \leq 2, 0 \leq y \leq 1 \\ 0 & \text{其他} \\ 0 & x < 0 \text{ 或 } y < 0 \end{cases}$$

$$F(x, y) = \begin{cases} \frac{xy}{2} & 0 \leq x \leq 2, 0 \leq y \leq 1 \\ \frac{x}{2} & 0 \leq x \leq 2, y > 1 \\ y & x > 2, 0 \leq y \leq 1 \\ 1 & x > 2, y > 1 \end{cases}$$

$$(2) f_X(x) = \begin{cases} \frac{1}{2} & 0 \leq x \leq 2 \\ 0 & \text{其他} \\ 0 & x < 0 \end{cases} \quad f_Y(y) = \begin{cases} 1 & 0 \leq y \leq 1 \\ 0 & \text{其他} \\ 0 & y < 0 \end{cases}$$

$$F_X(x) = \begin{cases} \frac{x}{2} & 0 \leq x \leq 2 \\ 1 & x > 2 \end{cases} \quad F_Y(y) = \begin{cases} y & 0 \leq y \leq 1 \\ 1 & y > 1 \end{cases}$$

$$(3) \frac{2}{3}$$

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

$$(2) f_X(x) = \begin{cases} 2(1-x) & 0 \leq x \leq 1 \\ 0 & \text{其他} \\ 0 & x < 0 \text{ 或 } y < 1 \end{cases} \quad f_Y(y) = \begin{cases} y & 0 \leq y \leq 1 \\ 0 & \text{其他} \end{cases}$$

$$\frac{x(y-1)}{2} \quad 0 \leq x \leq 2, 1 \leq y \leq 2$$

$$6. (1) F_{(X,Y)} = \begin{cases} \frac{x}{2} & 0 \leq x \leq 2, y \geq 2 \\ y-1 & x > 2, 1 \leq y \leq 2 \\ 1 & x > 2, y > 2 \\ 0 & x < 0 \text{ 或 } y < 1 \end{cases}$$

$$\frac{x(y-1)}{2} \quad 0 \leq x \leq 4, 1 \leq y \leq 4$$

$$(2) G_{(X,Y)} = \begin{cases} \frac{x}{2} & 0 \leq x \leq 4, y < 4 \\ y-1 & x > 4, 1 \leq y \leq 4 \\ 1 & x > 4, y < 4 \end{cases}$$

$$(3) \frac{1}{4}$$

$$8. P\{X_1=0, X_2=1\}=\frac{1}{3} \quad P\{X_1=1, X_2=1\}=\frac{2}{3}$$

$$9. p_{ij} = p_j^X p_{j|i}^Y \quad p_j^Y = \sum_i p_{ij} \quad p_{ij} = p_i^X p_{j|i}^Y$$

$$p_{i|i}^Y = \frac{p_{ij}}{p_j^Y} = \frac{p_i^X p_{j|i}^Y}{p_i^X p_{j|i}^Y}$$

$$10. p_{11} = \frac{1}{4}, p_{12} = 0, p_{13} = \frac{1}{4}, p_{22} = \frac{1}{2}; \text{不独立.}$$

11.

$\begin{matrix} Y \\ X \end{matrix}$	y_1	y_2	y_3	p_i^X
x_1	$\frac{1}{24}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{4}$
x_2	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{4}$
p_j^Y	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$	1

$$13. \frac{19}{36}, \frac{2}{36}.$$

$$14. f_{X_1 Y_1}(x, y) = \begin{cases} 3e^{-3x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$f_{Y_1 X_1}(y, x) = \begin{cases} 4e^{-4y} & y \geq 0 \\ 0 & y < 0 \end{cases}$$

$$f_{X_2 Y_2}(x, y) = \begin{cases} 3e^{-3x-3y} & x \geq y \\ 0 & x < y \end{cases} \quad (y \geq 0)$$

$$f_{Y_2 X_2}(y, x) = \begin{cases} 4 \cdot \frac{e^{-4y}}{1-e^{-4x}} & 0 < y < x \\ 0 & \text{其他} \end{cases} \quad (x > 0)$$

$$15. f_{Y X}(y, x) = \begin{cases} \frac{1}{x+1} & 0 < y < x+1 \\ 0 & \text{其他} \end{cases} \quad (0 < x < 1)$$

$$f_{Y X}(y, x) = \begin{cases} \frac{1}{2} & x-1 < y < x+1 \\ 0 & \text{其他} \end{cases} \quad (1 \leq x \leq 2)$$

$$f_{X Y}(x, y) = \begin{cases} \frac{1}{y+1} & 0 < x < y+1 \\ 0 & \text{其他} \end{cases} \quad (0 \leq y \leq 1)$$

$$f_{X Y}(x, y) = \begin{cases} \frac{1}{3-y} & y-1 < x < 2 \\ 0 & \text{其他} \end{cases} \quad (1 \leq y < 3)$$

18. X_1 与 Y_1 独立, X_2 与 Y_2 不独立

19. 不独立

$$20. f(x, y) = \begin{cases} \frac{1}{2} e^{-\frac{y^2}{2}} & 0 \leq x \leq 1, -\infty < y < +\infty \\ 0 & \text{其他} \end{cases}$$

23.

	0	1	2		- 2	- 1	0	1	2	
p_i	$\frac{1}{9}$	$\frac{4}{9}$	$\frac{4}{9}$,	p_i	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{2}{9}$	$\frac{1}{9}$
		- 2			- 1			0	1	2
0		0			0			$\frac{1}{9}$	0	0
1		$\frac{2}{9}$			0			0	0	0
2		$\frac{1}{9}$			0			$\frac{2}{9}$	0	$\frac{1}{9}$

24.

p_i	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
-------	----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------	----------------

26.

$f_{X_1+X_2}(z)=\begin{cases} 0 & z \leq 0 \\ 2ze^{-z} & z > 0 \end{cases}$

27.

$f_X(x)=\frac{1}{2}e^{-\frac{(x-3)^2}{2}} \quad f_Y(y)=\frac{1}{2}e^{-\frac{(y-1)^2}{2}}$

29.

$F(s)=\begin{cases} \frac{1}{4}s & 0 < s \leq 2 \\ 1-\frac{1}{s} & s > 2 \end{cases} \quad f(s)=\begin{cases} \frac{1}{4} & 0 < s \leq 2 \\ \frac{1}{s^2} & s > 2 \end{cases}$

30.

0, 0.4

31.

$\frac{5}{3}$

33.

8

34.

$9[1-(\frac{8}{9})^{25}]$

36.

$-\frac{15}{224}$

37.

$0, \frac{1}{49}$

38.

1, 0

40.

$-\frac{4}{66}$

41.

$(1) \frac{1}{2} \quad (2) 13x_A^2-6x_A+9$

$(3) x_A=\frac{3}{13}, 0 \leq x_A \leq \frac{6}{13}$

42.

若 $\frac{B}{A} \leq 1$, $x = \frac{B}{A}$ 时, P 无风险

$\frac{B}{A} < \frac{\min(x_A, x_B)}{\max(x_A, x_B)}$

43.

$(1) 2.755\% \quad (3) 3\%$

44. $E[Y|X=x]=\frac{x+1}{2} \quad 0 < x < 1$

$E[Y|X]=\frac{X+1}{2} \quad 0 < X < 1$

45. $E[X|Y]=1 \quad E[Y|X]=0$

46. 0.0465

47. 2265

48. 0.0027, 440.

习题三

(B)

1. (1)

$X_2 \backslash X_1$	0	1	2	3
0	21/792	105/792	105/792	21/792
1	70/792	210/792	126/792	14/792
2	35/792	63/792	21/792	1/792

$X_2 \quad 0 \quad 1 \quad 2 \quad 3$

(2)

$p_{j \cdot}^{X_2} \quad 126/792 \quad 378/792 \quad 252/792 \quad 36/792$

(3)

X_1	0	1	2
$p_{i \cdot 0}$	6/36	20/36	10/36
$p_{i \cdot 1}$	10/36	20/36	6/36
$p_{i \cdot 2}$	15/36	18/36	3/36
$p_{i \cdot 3}$	21/36	14/36	1/36

$Y \quad 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5$

(4)

$p_{i \cdot} \quad 21/792 \quad 175/792 \quad 350/792 \quad 210/792 \quad 35/792 \quad 1/792$

(5)

$Y_1 \backslash X_1$	0	1	2
0	3/66	12/66	6/66
1	15/66	20/66	0
2	10/66	0	0

$\begin{matrix} Y_2 \\ X_1 \end{matrix}$	0	1	2	3
0	102/990	159/990	51/990	3/990
1	125/990	270/990	120/990	10/990
2	25/990	75/990	45/990	5/990

$$2. (1) P\{Y_i = k, Y_j = s\} = \frac{C_{25}^k C_{25-k}^s 7^{25-k-s}}{9^{25}} \quad (k+s \leq 25)$$

$$P\{Y_i = k\} = \frac{C_{25}^k 8^{25-k}}{9^{25}}$$

Y_i 与 Y_j 不独立

(2)

$\begin{matrix} X_j \\ X_i \end{matrix}$	0	1	$P_{k^i}^{x_i}$
0	$(\frac{7}{9})^{25}$	$(\frac{8}{9})^{25} - (\frac{7}{9})^{25}$	$(\frac{8}{9})^{25}$
1	$(\frac{8}{9})^{25} - (\frac{7}{9})^{25}$	$1 - 2(\frac{8}{9})^{25} + (\frac{7}{9})^{25}$	$1 - (\frac{8}{9})^{25}$
$p_{s^i}^{x_i}$	$(\frac{8}{9})^{25}$	$1 - (\frac{8}{9})^{25}$	

X_i 与 X_j 不独立.

$$(3) \text{cov}(X_i, X_j) = (\frac{7}{9})^{25} - (\frac{8}{9})^{50} \quad (i \neq j)$$

$$DX_i = (\frac{8}{9})^{25} - (\frac{8}{9})^{50}$$

$$x_i x_j = [(\frac{7}{9})^{25} - (\frac{8}{9})^{50}] / [(\frac{8}{9})^{25} - (\frac{8}{9})^{50}]$$

$$(4) DX = 9 [(\frac{8}{9})^{25} - (\frac{8}{9})^{50}] + 72 [(\frac{7}{9})^{25} - (\frac{8}{9})^{50}]$$

$$3. (1) P\{Y = s, X = k\} = C_{25-k}^s p_2^s (1-p_2)^{25-k-s} \quad 0 \leq s \leq 25-k$$

$$(2) P\{X = k, Y = s\} = C_{25}^s p_1^k (1-p_1)^{25-k} C_{25-k}^s p_2^s (1-p_2)^{25-k-s}$$

$$= \frac{25!}{k! s! (25-k-s)!} p_1^k [p_2 (1-p_1)]^s [1-p_1-p_2(1-p_1)]^{25-k-s}$$

$$(3) E[Y|X] = (25-X)p_2$$

$$E[Y^2|X] = (25-X)p_2(1-p_2) + (25-X)^2 p_2^2$$

$$(4) EY = E[E(Y|X)] = (25-25p_1)p_2 = 25(1-p_1)p_2$$

$$DY = EY^2 - (EY)^2 = E(E[Y^2|X]) - (EY)^2$$

$$= 25(1-p_1)p_2[1-p_2(1-p_1)]$$

4. (3) $Y \sim b(25, (1-p_1)p_2)$, 在 $Y = s$ 的条件下 X 的条件分布 $X|Y=s \sim b(25-s,$

$$\frac{p_1}{(1-p_1)p_2})$$

$$5. (1) f(x, y) = \begin{cases} xe^{-xy} & 1 \leq x \leq 2, y \geq 0 \\ 0 & \text{其他} \end{cases}$$

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

$$(3) f_{X|Y}(x|y) = \begin{cases} \frac{x e^{xy}}{y(e^{-y} - 2e^{-2y}) + e^{-y} - e^{-2y}} & 1 < x < 2 \quad (y > 0) \\ 0 & \text{其他} \end{cases}$$

$$(4) 1.474$$

$$(5) 0.8655$$

$$6. (1) f_X(x) = \begin{cases} \frac{1}{2}x & 0 < x < 1 \\ \frac{1}{2} & 1 < x < 2 \\ \frac{3-x}{2} & 2 < x < 3 \\ 0 & \text{其他} \end{cases}$$

$$f_Y(y) = \begin{cases} \frac{3-2y}{2} & 0 < y < 1 \\ 0 & \text{其他} \end{cases}$$

X 与 Y 不独立

$$(2) \frac{5}{12}, 0, \frac{11}{144}, \text{X 与 Y 不相关}$$

$$(3) f_{Y|X}(y|x) = \begin{cases} \frac{1}{x} & 0 < y < x \quad (0 < x < 1) \\ 0 & \text{其他} \end{cases}$$

$$f_{Y|X}(y|x) = \begin{cases} 1 & 0 < y < 1 \quad (1 < x < 2) \\ 0 & \text{其他} \end{cases}$$

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{3-x} & 0 < y < 3-x \quad (2 < x < 3) \\ 0 & \text{其他} \end{cases}$$

$$f_{X|Y}(x|y) = \begin{cases} \frac{1}{3-2y} & y < x < 3-y \quad (0 < y < 1) \\ 0 & \text{其他} \end{cases}$$

$$(4) E[Y|X] = \begin{cases} \frac{x}{2} & 0 < x < 1 \\ \frac{1}{2} & 1 < x < 2 \end{cases}$$

$$\frac{3-x}{2} \quad 2 < x < 3$$

$$E[X|Y] = \frac{3}{2} \quad (0 < Y < 1)$$

8.

X ₁ \ X ₂	0	1
0	$1 - e^{-1}$	0
1	$e^{-1} - e^{-2}$	e^{-2}

$$\cos(X_1, X_2) = e^{-2} - e^{-3}$$

$$x_1, x_2 = \frac{e^{-1}}{e^2 - 1}$$

9.

U \ V	0	1
0	$\frac{1}{4}$	0
1	$\frac{1}{4}$	$\frac{1}{2}$

$$P_{UV} = \frac{1}{3}$$

13. (1) $\frac{2}{3}$

$$(2) f_Z(z) = \begin{cases} \frac{3}{2} - z & 0 < z < 1 \\ 0 & \text{其他} \end{cases}$$

习题四

(A)

1. $f(x_1, x_2, \dots, x_n) = n e^{-(x_1 + \dots + x_n)}, x_i > 0, i = 1, 2, \dots, n.$

2. $f(x_1, x_2, \dots, x_n) = \begin{cases} 1, & 0 < x_i < 1, i = 1, 2, \dots, n; \\ 0 & \text{其他.} \end{cases}$

3. $P(X_1 = k_1, X_2 = k_2, \dots, X_n = k_n) = \frac{p^n}{1-p} (1-p)^{k_1 + k_2 + \dots + k_n},$

其中 k_i 为正整数, $i = 1, 2, \dots, n.$

4. $\bar{u} = \frac{\bar{x} - a}{b}, s_u^2 = \frac{s_x^2}{b^2}.$

7. $P(X_{(1)} > x) = 1 - (1 - F(x))^n,$

$P(X_{(n)} > x) = (F(x))^n;$

特别地, 若 X 服从以 $(\lambda > 0)$ 为参数的指数分布, 则 $P\{X_{(1)} > x\} = 1 - e^{-\lambda x}, x > 0;$

$P\{X_{(n)} > x\} = (1 - e^{-\lambda x})^n, x > 0.$

8. (1) 240; (2) $15 \sqrt{16};$ (3) $3 \sqrt{16}.$

$$\begin{aligned} 9. E[X^{2n}] &= \int_0^\infty x^{2n} e^{-\frac{1}{2}x^2} dx = \int_0^\infty t^{n-\frac{1}{2}} e^{-t} dt \quad (\text{令 } t = \frac{1}{2}x^2) \\ &= \frac{2^n}{\sqrt{\pi}} \left(n + \frac{1}{2}\right) = \frac{2^n}{\sqrt{\pi}} \cdot \frac{2n-1}{2} \cdot \frac{2n-3}{2} \cdots \frac{1}{2} \left(\frac{1}{2}\right) \\ &= 1 \cdot 3 \cdot 5 \cdots (2n-1). \end{aligned}$$

10. $f(x) = \frac{n^{\frac{m}{2}}}{2^{\frac{m}{2}} \frac{\Gamma(n)}{2}} x^{\frac{m}{2}-1} e^{-\frac{nx}{2}}, x > 0.$

11. $c = \frac{1}{3},$ 自由度为 2.

12. $P(S_n = k) = \frac{(n)^k}{k!} e^{-n}, k = 0, 1, 2, \dots;$

又当样本容量 n 充分大时, S_n 渐近地服从正态分布 $N(n, n),$ 即有

$$P(S_n = x) = \frac{1}{\sqrt{n}} \exp\left(-\frac{x^2}{2n}\right).$$

$$13. u_{0.6} = 0.253, u_{0.8} = 0.8416, u_{0.9} = 1.28,$$

$$u_{0.95} = 1.65.$$

$$14. 442.$$

$$15. 1.145, 11.071, 2.558, 23.209$$

$$16. 0.1623, 0.0684, 0.0912$$

$$17. 2.353, 3.365, 1.415, 3.169$$

习题四

(B)

$$1. 250$$

$$2. (-0.63, 0.63)$$

* 4. 证明概要: 由例 4.1 知, n 维随机向量 X 的联合密度函数为

$$f(x) = \frac{1}{2^n} \exp\left(-\frac{1}{2}x^T x\right)$$

其中 $x = (x_1, x_2, \dots, x_n)^T$ 这样, n 维随机向量 Y 的分布函数为

$$F(y_1, y_2, \dots, y_n) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} \dots \int_{-\infty}^{y_n} \frac{1}{2^n} \exp\left(-\frac{1}{2}x^T x\right) dx_1 dx_2 \dots dx_n$$

在上述 n 重积分中作变量替换 $t = Ax$. 由于 A 为正交矩阵, 便知 $x = A^{-1}t = A^T t$, 而变换的雅可比行列式为

$$J = \left| \frac{\partial x_i}{\partial t_j} \right| = |A^T|.$$

再由于 A 为正交矩阵, 即知行列式 $|A^T|$ 的值为 1 或 -1, 从而雅可比行列式的绝对值 $|J| = 1$.

1. 于是

$$\begin{aligned} F(y_1, y_2, \dots, y_n) &= \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} \dots \int_{-\infty}^{y_n} \frac{1}{2^n} \exp\left[-\frac{1}{2}(A^T t)^T (A^T t)\right] dt_1 dt_2 \dots dt_n \\ &= \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} \dots \int_{-\infty}^{y_n} \frac{1}{2^n} \exp\left[-\frac{1}{2}t^T t\right] dt_1 dt_2 \dots dt_n \end{aligned}$$

这表明 Y_1, Y_2, \dots, Y_n 相互独立同分布, 且 $Y_i \sim N(0, 1)$, $i = 1, 2, \dots, n$.

* 5. 证明概要: 首先, 由命题 4.3 知 $\bar{X} \sim N\left(0, \frac{1}{n}\right)$. 其次, 不难验证

$$(n-1) S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n(\bar{X})^2.$$

$$\text{现令 } Y_k = \frac{1}{\sqrt{k(k+1)}} (X_1 + X_2 + \dots + X_k - X_{k+1}), \quad k = 1, 2, \dots, n-1;$$

$$Y_n = \frac{1}{\sqrt{n}} (X_1 + X_2 + \dots + X_n) = \bar{X}.$$

可以验证这是由 $(X_1, X_2, \dots, X_n)^T$ 至 $(Y_1, Y_2, \dots, Y_n)^T$ 的正交变换, 即若记 $Y = (Y_1, Y_2, \dots, Y_n)^T$, $X = (X_1, X_2, \dots, X_n)^T$, 当把上述诸式统写成矩阵等式 $Y = AX$ 后, 可以验证 A 为一正交矩阵. 这样, 由上题知 Y_1, Y_2, \dots, Y_n 相互独立同分布, 且 $Y_i \sim N(0, 1)$, $i = 1, 2, \dots, n$.

再因 A 为正交矩阵, 即知

$$Y_1^2 + Y_2^2 + \dots + Y_n^2 = Y^T Y = (AX)^T AX = X^T (A^T A) X = X^T X \\ = X_1^2 + X_2^2 + \dots + X_n^2,$$

从而

$$(n-1)S^2 = \sum_{i=1}^n X_i^2 - n(\bar{X})^2 = \sum_{i=1}^n Y_i^2 - Y_n^2 = Y_1^2 + Y_2^2 + \dots + Y_{n-1}^2.$$

这样, 由命题4.1与定义4.6即知 $(n-1)S^2 \sim \chi^2(n-1)$.

此外, $(n-1)S^2$ 与 Y_n 独立, 从而也和 \bar{X} 独立.

* 6. 提示: 记

$$X_i^* = \frac{X_i - \mu}{\sigma}, \quad i = 1, 2, \dots, n.$$

则 $X_1^*, X_2^*, \dots, X_n^*$ 相互独立同分布, 且由 $X_i^* \sim N(0, 1)$, $i = 1, 2, \dots, n$. 这样, 利用上题结论即可证明定理4.1.

* 8. 因 $Y_n \xrightarrow{P} 1$, 任取 $0 < \epsilon < 1$, 恒有

$$\lim_n P\{1 - \epsilon < Y_n < 1 + \epsilon\} = 1.$$

记 $A_n = \{1 - \epsilon < Y_n < 1 + \epsilon\}$, 便有 $\lim_n P(\bar{A}_n) = 0$.

因

$$P(X_n Y_n < x) = P(X_n Y_n < x, 1 - \epsilon < Y_n < 1 + \epsilon) + P(X_n Y_n < x, \bar{A}_n) \\ = P(X_n < \frac{x}{1 - \epsilon}) + P(\bar{A}_n)$$

再由于 $X_n \xrightarrow{d} X$, 若记 $F(x)$ 为 X 的分布函数,

由上式可推知 $\lim_n P(X_n Y_n < x) = \lim_n P(X_n < \frac{x}{1 - \epsilon}) = F(\frac{x}{1 - \epsilon})$.

再在上式中令 $\epsilon \rightarrow 0$, 因 $F(x)$ 是连续函数即得

$$\lim_n P(X_n Y_n < x) = F(x).$$

类似地, 有

$$P(X_n Y_n < x) = P(X_n < \frac{x}{Y_n}, 1 - \epsilon < Y_n < 1 + \epsilon) + P(X_n Y_n < x, \bar{A}_n) \\ = P(X_n < \frac{x}{1 + \epsilon}, 1 - \epsilon < Y_n < 1 + \epsilon) + P(X_n Y_n < x, \bar{A}_n) \\ = P(X_n < \frac{x}{1 + \epsilon}) - P(X_n < \frac{x}{1 + \epsilon}, \bar{A}_n) + P(X_n Y_n < x, \bar{A}_n)$$

这样, 在上式中令 $\epsilon \rightarrow 0$, 由于 $\lim_n P(\bar{A}_n) = 0$, 即知

$$\lim_n P(X_n Y_n < x) = \lim_n P(X_n < \frac{x}{1 + \epsilon}) = F(\frac{x}{1 + \epsilon})$$

再在上式中令 $\epsilon \rightarrow 0$, 可得

$$\lim_{n \rightarrow \infty} P(X_n Y_n \leq x) = F(x).$$

这样,

$$F(x) = \lim_{n \rightarrow \infty} P(X_n Y_n \leq x) = \overline{\lim_{n \rightarrow \infty} P(X_n Y_n \leq x)} = F(x)$$

故

$$\lim_{n \rightarrow \infty} P(X_n Y_n \leq x) = F(x), \quad \forall x \in \mathbb{R}.$$

习题五

(A)

2. $C = 2(n-1)$

4. μ 较有效

$$5. \quad \text{MLE} = \frac{n}{\sum_{i=1}^n \ln X_i}$$

$$p_{\text{MLE}} = \frac{1}{\bar{X}}$$

$$p_{\text{MLE}} = \bar{X}$$

$$\text{MLE} = \bar{\bar{X}}$$

6. 都是2.5

12.18

$$7. \quad a_{\text{ME}} = \bar{X} - \sqrt{\frac{1}{3} B_2}, \quad b_{\text{ME}} = \bar{X} + \sqrt{\frac{1}{3} B_2}, \quad B_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$8. \quad \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}{\sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}}$$

$$9. \quad \text{ME} = \bar{X} - 1, \quad \text{MLE} = X_{(1)}, \quad X_{(1)} = \min \{x_1, \dots, x_n\}$$

$$10. \quad (\frac{X_{(n)}}{2}, \frac{X_{(n)}}{1}), \quad \text{其中 } \alpha_1, \alpha_2 \text{ 满足方程 } \frac{\alpha_2}{2} - \frac{\alpha_1}{1} = 1 -$$

11. 0.15

12. (271, 291)

13. $n = 25, \quad n = 60$

14. (14.51, 19.69)

15. 5.22

16. (2.36, 10.97)

17. (92.12, 207.88)

18. (0.14, 2.77)

习题五

(B)

2. 满足 $X_{(n)} - 1$ $X_{(1)}$ 的任何一个
3. (1766, 3795)

习题六

(A)

1. 犯第一类错误的概率为 e^{-2} , 犯第二类错误的概率为 $1 - e^{-1}$.
2. $c = 1.176$
3. 今年的日均销售额与去年不同.
4. 能说明新安眠剂已达到新的疗效.
5. (1) 在显著水平 $= 0.05$ 下, 认为机器工作状态不正常; (2) 在显著水平 $= 0.01$ 下, 认为机器工作是正常的.
6. 认为新生产的缆绳的抗拉强度有明显提高.
7. 认为生产情况是正常的.
8. 不能认为方差仍是 0.112^2 .
9. 待检验的零假设有两个: (1) $H_0: \mu = 1000$; (2) $H_0: \sigma^2 = 15^2$. 检验结果认为第一个零假设是相容的, 但否定了第二个零假设, 即认为方差超过了 15^2 , 故此日包装机工作不正常, 应停机检修.
10. 认为正常成年男、女性红细胞数有显著差别.
11. 不能拒绝方差相等的假设.
13. 不真实.
14. 不大于 $1/6$.
15. 可以认为15秒内通过此公路的汽车辆数服从泊松分布.
16. 可以认为是匀称的.
17. 发生皮炎与工种有关.
18. 可以认为投票结果与投票者在什么系是相互独立的.

习题六

(B)

$$1. \text{ 令 } W^* = \frac{1}{2} \sum_{i=1}^n (X_i - \mu)^2$$

则 W^* 服从自由度为 n 的 χ^2 分布.

$$2. \text{ 令 } F^* = \frac{1}{r} \frac{(S_1^*)^2}{(S_2^*)^2},$$

其中 $(S_1^*)^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (X_i - \mu)^2$, $(S_2^*)^2 = \frac{1}{n_2} \sum_{j=1}^{n_2} (Y_j - \mu)^2$, 则 F^* 从第一自由度为 n_1 , 第二自由度为 n_2 的 F 分布.

$$3. (1) \bar{x} = 0.1965$$

(2) 可运用近似 U 检验法的双侧检验法来解. 没有充分理由拒绝零假设, 从而认为泊松分布的参数 $\lambda = 0.2$.

4. 以 X 表示一次试验中正确配对个数, 并记 $p_i = P(X = i)$, $i = 0, 1, 3$. 根据古典概率模型可验证: “受试者无特异功能” $H_0: p_0 = \frac{1}{3}, p_1 = \frac{1}{2}, p_3 = \frac{1}{6}$. 采用多项分布的 χ^2 检验法可算出 $\chi_0^2 = 2.11$. 对于给定的显著性水平 $\alpha = 0.05$, 查附表3知, $\chi_{0.05}^2(r-1) = \chi_{0.05}^2(2) = 5.991$. 因 $\chi_0^2 = 2.11 < \chi_{0.05}^2(r-1) = 5.991$, 可认为受试者无特异功能.

5- (3) 钱币是均匀的.

6. 可信.

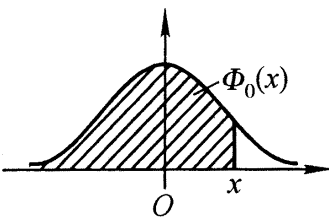
续表

<div>m</div> <div></div>	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0
0	0.011109	0.006738	0.004087	0.002479	0.001503	0.0000912	0.000553	0.000335
1	0.049990	0.033690	0.022477	0.014873	0.009773	0.006383	0.004148	0.002684
2	0.112479	0.084224	0.061812	0.044618	0.031760	0.022341	0.015556	0.010735
3	0.168718	0.140374	0.113323	0.089235	0.068814	0.052129	0.038888	0.028626
4	0.189808	0.175467	0.155819	0.133853	0.111822	0.091226	0.072917	0.057252
5	0.170827	0.175467	0.171001	0.160623	0.145369	0.127717	0.109374	0.091604
6	0.128120	0.146223	0.157117	0.160623	0.157483	0.149003	0.136719	0.122138
7	0.082363	0.104445	0.123449	0.137677	0.146234	0.149003	0.146484	0.139587
8	0.046329	0.065278	0.084872	0.103258	0.118815	0.130377	0.137328	0.139587
9	0.023165	0.036266	0.051866	0.068838	0.085811	0.101405	0.114441	0.124077
10	0.010424	0.018133	0.028526	0.041303	0.055777	0.070983	0.085830	0.099262
11	0.004264	0.008242	0.014263	0.022529	0.032959	0.045171	0.058521	0.072190
12	0.001599	0.003434	0.006537	0.011264	0.017853	0.026350	0.036575	0.048127
13	0.000554	0.001321	0.002766	0.005199	0.008927	0.014188	0.02101	0.029616
14	0.000178	0.000472	0.001086	0.002228	0.004144	0.007094	0.011305	0.016924
15	0.000053	0.000157	0.000399	0.000891	0.001796	0.003311	0.005652	0.009026
16	0.000015	0.000049	0.000137	0.000334	0.000730	0.001448	0.002649	0.004513
17	0.000004	0.000014	0.000044	0.000118	0.000279	0.000596	0.001169	0.002124
18	0.000001	0.000004	0.000014	0.000039	0.000100	0.000232	0.000487	0.000944
19		0.00001	0.000004	0.000012	0.000035	0.000085	0.000192	0.000397
20			0.00001	0.000004	0.000011	0.000030	0.000072	0.000159
21				0.000001	0.000004	0.000010	0.000026	0.000061
22					0.000001	0.000003	0.000009	0.000022
23						0.000001	0.000003	0.000008
24							0.000001	0.000003
25								0.000001
26								
27								
28								
29								

续表

<div>m</div>	8.5	9.0	9.5	10.0	<div>m</div>	20	<div>m</div>	30
0	0.000203	0.000123	0.000075	0.000045	5	0.0001	12	0.0001
1	0.001730	0.001111	0.000711	0.000454	6	0.0002	13	0.0002
2	0.007350	0.004998	0.003378	0.002270	7	0.0005	14	0.0005
3	0.020826	0.014994	0.010696	0.007567	8	0.0013	15	0.0010
4	0.44255	0.033737	0.025403	0.018917	9	0.0029	16	0.0019
5	0.075233	0.060727	0.048265	0.037833	10	0.0058	17	0.0034
6	0.106581	0.091090	0.076421	0.063055	11	0.0106	18	0.0057
7	0.129419	0.117116	0.103714	0.090079	12	0.0176	19	0.0089
8	0.137508	0.131756	0.123160	0.112599	13	0.0271	20	0.0134
9	0.129869	0.131756	0.130003	0.125110	14	0.0382	21	0.0192
10	0.110303	0.118580	0.122502	0.125110	15	0.0517	22	0.0261
11	0.085300	0.097020	0.106662	0.113736	16	0.0646	23	0.0341
12	0.060421	0.072765	0.084440	0.094780	17	0.0760	24	0.0426
13	0.039506	0.050376	0.061706	0.072908	18	0.0814	25	0.0571
14	0.023986	0.032384	0.041872	0.052077	19	0.0888	26	0.0590
15	0.013592	0.019431	0.026519	0.034718	20	0.0888	27	0.0655
16	0.007220	0.010930	0.015746	0.021699	21	0.0846	28	0.0702
17	0.003611	0.005786	0.008799	0.012764	22	0.0767	29	0.0726
18	0.001705	0.002893	0.004644	0.007091	23	0.0669	30	0.0726
19	0.000762	0.001370	0.002322	0.003732	24	0.0557	31	0.703
20	0.000324	0.000617	0.001103	0.001866	25	0.0446	32	0.0659
21	0.000132	0.000264	0.000433	0.000898	26	0.0343	33	0.0599
22	0.000050	0.000108	0.000216	0.000404	27	0.0254	34	0.0529
23	0.000019	0.000042	0.000089	0.000176	28	0.0182	35	0.0453
24	0.000007	0.000016	0.000025	0.000073	29	0.0125	36	0.0378
25	0.000002	0.000006	0.000014	0.000029	30	0.0083	37	0.0306
26	0.000001	0.000002	0.000004	0.000011	31	0.0054	38	0.0242
27		0.000001	0.000002	0.000004	32	0.0034	39	0.0186
28			0.000001	0.000001	33	0.0020	40	0.0139
29				0.000001	34	0.0012	41	0.0102
							42	0.0073
							43	0.0501
					35	0.0007	44	0.0035
					36	0.0004	45	0.0023
					37	0.0002	46	0.0015
					38	0.0001	47	0.0010
					39	0.0001	48	0.0006

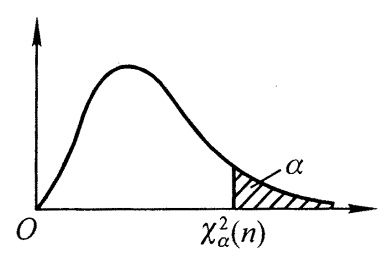
附表 2 标准正态分布函数值表 $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt \quad (x \geq 0)$



x	0.00	0.01	0.02	0.03	0.04
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0.6554	0.6591	0.6628	0.6664	0.6700
0.5	0.6915	0.6950	0.6985	0.7019	0.7054
0.6	0.7257	0.7291	0.7324	0.7357	0.7389
0.7	0.7580	0.7611	0.7642	0.7673	0.7703
0.8	0.7881	0.7910	0.7939	0.7967	0.7995
0.9	0.8159	0.8186	0.8212	0.8238	0.8264
1.0	0.8413	0.8438	0.8461	0.8485	0.8508
1.1	0.8643	0.8665	0.8686	0.8708	0.8729
1.2	0.8849	0.8869	0.8888	0.8907	0.8925
1.3	0.90320	0.90490	0.90678	0.90824	0.90988
1.4	0.91924	0.92073	0.92220	0.92364	0.92507
1.5	0.93319	0.93448	0.93574	0.93699	0.93822
1.6	0.94520	0.94630	0.94738	0.94845	0.94950
1.7	0.95543	0.95637	0.95728	0.95818	0.95907
1.8	0.96407	0.96485	0.96562	0.96638	0.96712
1.9	0.97128	0.97193	0.97257	0.97320	0.97381
2.0	0.97725	0.97778	0.97831	0.97882	0.97932
2.1	0.98214	0.98257	0.98300	0.98341	0.98382
2.2	0.98610	0.98645	0.98679	0.98713	0.98745
2.3	0.98928	0.98956	0.98983	0.99010	0.99036
2.4	0.99180	0.99202	0.99224	0.99245	0.99266
2.5	0.99379	0.99396	0.99413	0.99430	0.99446
2.6	0.99534	0.99547	0.99560	0.99573	0.99586
2.7	0.99653	0.99664	0.99674	0.99683	0.99693
2.8	0.99745	0.99752	0.99760	0.99767	0.99774
2.9	0.99813	0.99819	0.99825	0.99831	0.99836
3.0	0.99865	0.99869	0.99874	0.99878	0.99882
3.1	0.99903	0.99906	0.99910	0.99913	0.99916
3.2	0.99931	0.99934	0.99936	0.99938	0.99940
3.3	0.99952	0.99953	0.99955	0.99957	0.99958
3.4	0.99966	0.99968	0.99969	0.99970	0.99971
3.5	0.99977	0.99978	0.99978	0.99979	0.99980
3.6	0.99984	0.99985	0.99985	0.99986	0.99986
3.7	0.99989	0.99990	0.99990	0.99990	0.99991
3.8	0.99993	0.99993	0.99993	0.99994	0.99994
3.9	0.99995	0.99995	0.99996	0.99996	0.99996

续表					
x	0.00	0.01	0.02	0.03	0.04
4.0	0.99997	0.99997	0.99997	0.99997	0.99997
4.1	0.99998	0.99998	0.99998	0.99998	0.99998
4.2	0.99999	0.99999	0.99999	0.99999	0.99999
4.3	0.99999	0.99999	0.99999	0.99999	0.99999
4.4	0.99999	0.99999	1.00000	1.00000	1.00000
x	0.05	0.06	0.07	0.08	0.09
0.0	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8944	0.8962	0.8980	0.8997	0.90147
1.3	0.91140	0.91309	0.91466	0.91621	0.91774
1.4	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99702	0.99711	0.99720	0.99728	0.99737
2.8	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99886	0.99889	0.99893	0.99897	0.99900
3.1	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99996	0.99996	0.99996	0.99997	0.99997
4.0	0.99997	0.99998	0.99998	0.99998	0.99998
4.1	0.99998	0.99998	0.99998	0.99999	0.99999
4.2	0.99999	0.99999	0.99999	0.99999	0.99999
4.3	0.99999	0.99999	0.99999	0.99999	0.99999
4.4	1.00000	1.00000	1.00000	1.00000	1.00000

附表 3 χ^2 分布上侧分位数表 $P\{\chi^2(n) > \chi^2_\alpha(n)\} = \alpha$

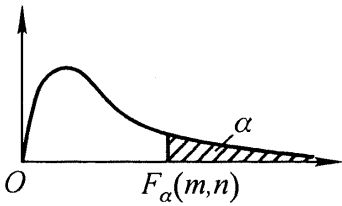


$n \backslash \alpha$	0.995	0.99	0.975	0.95	0.90	0.75
1	—	—	0.001	0.004	0.016	0.102
2	0.010	0.020	0.051	0.103	0.211	0.575
3	0.072	0.115	0.216	0.352	0.584	1.213
4	0.207	0.297	0.484	0.711	1.064	1.923
5	0.412	0.554	0.831	1.145	1.610	2.675
6	0.676	0.872	1.237	1.635	2.204	3.455
7	0.989	1.239	1.690	2.167	2.833	4.255
8	1.344	1.646	2.180	2.733	3.490	5.071
9	1.735	2.088	2.700	3.325	4.168	5.899
10	2.156	2.558	3.247	3.940	4.865	6.737
11	2.603	3.053	3.816	4.575	5.578	7.584
12	3.074	3.571	4.404	5.226	6.304	8.438
13	3.565	4.107	5.009	5.892	7.042	9.299
14	4.075	4.660	5.629	6.571	7.790	10.165
15	4.601	5.229	6.262	7.261	8.547	11.037
16	5.142	5.812	6.908	7.962	9.312	11.912
17	5.697	6.408	7.564	8.672	10.085	12.792
18	6.265	7.015	8.231	9.390	10.865	13.675
19	6.844	7.633	8.907	10.117	11.651	14.562
20	7.434	8.260	9.591	10.851	12.443	15.452
21	8.034	8.897	10.283	11.591	13.240	16.344
22	8.643	9.542	10.982	12.338	14.042	17.240
23	9.260	10.196	11.689	13.091	14.848	18.137
24	9.886	10.856	12.401	13.848	15.659	19.037
25	10.520	11.524	13.120	14.611	16.473	19.939
26	11.160	12.198	13.844	15.379	17.292	20.843
27	11.808	12.879	14.573	16.151	18.114	21.749
28	12.461	13.565	15.308	16.928	18.939	22.657
29	13.121	14.257	16.047	17.708	19.768	23.567
30	13.787	14.954	16.791	18.493	20.599	24.478
31	14.458	15.655	17.539	19.281	21.434	25.390
32	15.134	16.362	18.291	20.072	22.271	26.304
33	15.815	17.074	19.047	20.867	23.110	27.219
34	16.501	17.789	19.806	21.664	23.952	28.136
35	17.192	18.509	20.569	22.465	24.797	29.054
36	17.887	19.233	21.336	23.269	25.643	29.973
37	18.586	19.960	22.106	24.075	26.492	30.893
38	19.289	20.691	22.878	24.884	27.343	31.815
39	19.996	21.426	23.654	25.695	28.196	32.737
40	20.707	22.164	24.433	26.509	29.051	33.660

续表

<div><div></div><div>n</div></div>	= 0.995	0.99	0.975	0.95	0.90	0.75
41	21.421	22.906	25.215	27.326	29.907	34.585
42	22.138	23.650	25.999	28.144	30.765	35.510
43	22.859	24.398	26.785	28.965	31.625	36.436
44	23.584	25.148	27.575	29.787	32.487	37.363
45	24.311	25.901	28.366	30.612	33.350	38.291
<div><div></div><div>n</div></div>	= 0.25	0.10	0.05	0.025	0.01	0.005
1	1.323	2.706	3.841	5.024	6.635	7.879
2	2.773	4.605	5.991	7.378	9.210	10.597
3	4.108	6.251	7.815	9.348	11.345	12.838
4	5.385	7.779	9.488	11.143	13.277	14.860
5	6.626	9.236	11.071	12.833	15.086	16.750
6	7.841	10.45	12.592	14.449	16.812	18.548
7	9.037	12.017	14.067	16.013	18.475	20.278
8	10.219	13.362	15.507	17.535	20.090	21.955
9	11.389	14.684	16.919	19.023	21.666	23.589
10	12.549	15.987	18.307	20.483	23.209	25.188
11	13.701	17.275	19.675	21.920	24.725	26.756
12	14.845	18.549	21.026	23.337	26.217	28.299
13	15.984	19.812	22.362	24.736	27.688	29.819
14	17.117	21.064	23.685	26.119	29.141	31.319
15	18.245	22.307	24.996	27.488	30.578	32.801
16	19.369	23.542	26.296	28.845	32.000	34.267
17	20.489	24.769	27.587	30.191	33.409	35.718
18	21.605	25.989	28.869	31.526	34.805	37.156
19	22.718	27.204	30.144	32.852	36.191	38.582
20	23.828	28.412	31.410	34.170	37.566	39.997
21	24.935	29.615	32.671	35.479	38.932	41.401
22	26.039	30.813	33.924	36.781	40.289	42.796
23	27.141	32.007	35.172	38.076	41.638	44.181
24	28.241	33.196	36.415	39.364	42.980	45.559
25	29.339	34.382	37.652	40.646	44.314	46.928
26	30.435	35.563	38.885	41.923	45.642	48.290
27	31.528	36.741	40.113	43.194	46.963	49.645
28	32.620	37.916	41.337	44.461	48.278	50.993
29	33.711	39.087	42.557	45.722	49.588	52.336
30	34.800	40.256	43.773	46.979	50.892	53.672
31	35.887	41.422	44.985	48.232	52.191	55.003
32	36.973	42.585	46.194	49.480	53.486	56.328
33	38.058	43.745	47.400	50.725	54.776	57.648
34	39.141	44.903	48.602	51.966	56.061	58.964
35	40.223	46.059	49.802	53.203	57.342	60.275
36	41.304	47.212	50.998	54.437	58.619	61.581
37	42.383	48.363	52.192	55.668	59.892	62.883
38	43.462	49.513	53.384	56.896	61.162	64.181
39	44.539	50.660	54.572	58.120	62.428	65.476
40	45.616	51.805	55.758	59.342	63.691	66.766
41	56.692	52.949	56.942	60.561	64.950	68.053
42	47.766	54.090	58.124	61.777	66.206	69.336
43	48.840	55.230	59.304	62.990	67.459	70.616
44	49.913	56.369	60.481	64.201	68.710	71.893
45	50.985	57.505	61.656	65.410	69.957	73.166

附表 4 F 分布上侧分位数表 $P\{F(m, n) > F_\alpha(m, n)\} =$



= 0.10

<div>m n</div>	1	2	3	4	5	6	7	8	9
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
27	2.92	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68
	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63

续表

= 0.10

<div><div>n</div><div>m</div></div>	10	12	15	20	24	30	40	60	120	
1	60.19	60.17	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.70	3.78	3.76
5	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.56
24	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.100

续表

= 0.05

<div><div>n</div><div>m</div></div>	1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.06	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

续表

= 0.05										
<div>m \ n</div>	10	12	15	20	24	30	40	60	120	
1	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

续表

= 0.025

<div><div>n</div><div>m</div></div>	1	2	3	4	5	6	7	8	9
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
4	12.22	10.65	8.98	9.60	9.36	9.20	9.07	8.98	8.90
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
7	8.07	6.54	5.89	5.52	5.52	5.12	4.99	4.90	4.82
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
9	7.21	5.71	5.03	4.72	4.48	4.32	4.20	4.10	4.03
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
12	6.55	5.10	4.42	4.12	3.89	3.73	3.61	3.51	3.44
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05
17	6.01	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
23	5.76	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22
	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11

续表

= 0.025

<div>m n</div>	10	12	15	20	24	30	40	60	120	
1	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

续表

= 0.01

<div>m n</div>	1	2	3	4	5	6	7	8	9
1	4652	4999.5	5403	5626	5764	5859	5928	5982	6022
2	98.50	90.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.53	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.45	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	6.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.83	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.30	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	7.72	5.52	4.64	4.14	3.82	3.59	3.42	3.29	3.18
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.61	2.41

续表

= 0.01										
<div><div>n</div><div>m</div></div>	10	12	15	20	24	30	40	60	120	
1	6056	6106	6157	6200	6235	6261	6287	6313	6339	6336
2	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.26	3.12	2.98	2.83	2.75	2.67	2.53	2.50	2.40	2.31
23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

续表

= 0.005

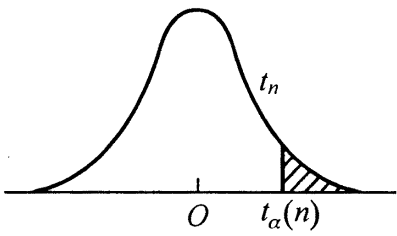
<div><div>n</div><div>m</div></div>	1	2	3	4	5	6	7	8	9
1	16211	20000	21615	22500	23056	23437	23715	23925	24091
2	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4
3	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88
4	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51
8	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54
10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97
11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54
12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.03	4.94
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88
22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64
26	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60
27	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.68	3.56
28	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52
29	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48
30	9.18	6.35	5.24	4.62	4.32	3.95	3.74	3.58	3.45
40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22
60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01
120	8.18	5.54	4.50	3.92	3.55	3.28	3.00	2.93	2.81
	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62

续表

= 0.005

<div><div>n</div><div>m</div></div>	10	12	15	20	24	30	40	60	120	
1	24224	24426	24630	24836	24940	25044	25148	25253	25359	25465
2	199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5
3	43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83
4	20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32
5	13.62	13.38	13.15	12.90	12.78	12.60	12.53	12.40	12.27	12.14
6	10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88
7	8.38	8.18	7.97	7.75	7.65	7.53	7.42	7.31	7.19	7.08
8	7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95
9	6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19
10	5.85	5.66	5.47	5.27	5.17	5.67	4.97	4.86	4.75	4.64
11	5.42	5.24	5.05	4.86	4.76	4.65	4.55	4.44	4.34	4.23
12	5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90
13	4.82	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.65
14	4.60	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.44
15	4.42	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.26
16	4.27	4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.11
17	4.14	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	2.98
18	4.03	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.87
19	3.93	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.78
20	3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69
21	3.77	3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.61
22	3.70	3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.55
23	3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.48
24	3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43
25	3.54	3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.38
26	3.49	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.33
27	3.45	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.29
28	3.41	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.25
29	3.38	3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.21
30	3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18
40	3.12	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	1.93
60	2.90	2.74	2.57	2.39	2.29	2.19	2.09	1.96	1.83	1.69
120	2.71	2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.43
	2.52	2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.00

附表 5 t-分布上侧分位数表 $P(t_n > t(n)) =$



<div>n \</div>	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
	1.282	1.645	1.960	2.326	2.576

主要参考文献

- [1] 洛哈吉 V K. 概率论及数理统计导论. 高尚华译. 北京: 高等教育出版社, 1983
- [2] 华东师范大学数学系. 概率论与数理统计教程. 北京: 高等教育出版社, 1983
- [3] 陈希孺, 倪国熙. 数理统计学教程. 上海: 上海科学技术出版社, 1988
- [4] 复旦大学. 概率论 第 2 册: 数理统计 第 1 分册. 北京: 高等教育出版社, 1979
- [5] 成世学, 严颖, 张诒兰, 等. 经济数学基础 (3): 概率统计. 北京: 中国人民大学出版社, 1994
- [6] 范培华, 袁荫棠等. 经济数学基础, 第 3 分册: 概率统计. 修订本. 成都: 四川人民出版社, 1998
- [7] 严士健, 刘秀芳, 徐承彝. 概率论与数理统计. 北京: 高等教育出版社, 1990
- [8] 陈希孺. 数理统计引论. 北京: 科学出版社, 1981
- [9] 王成名, 余鑫晖, 等. 应用概率统计. 桂林: 广西师范大学出版社, 1994
- [10] Alder H L, Roessler E B. 概率与统计导论. 胡崇能, 李隆章校译. 北京: 北京大学出版社, 1984
- [11] 华东师范大学数学系. 概率论与数理统计习题集. 北京: 高等教育出版社, 1982
- [12] 盛骤, 等. 概率论与数理统计. 北京: 高等教育出版社, 1989
- [13] 陆璇. 数理统计基础. 北京: 清华大学出版社, 1998
- [14] 茆诗松等. 高等数理统计. 北京: 高等教育出版社, 1998
- [15] 王玲玲, 周纪芄. 常用统计方法. 上海: 华东师范大学出版社, 1998
- [16] 王学仁等. 实用多元统计分析. 上海: 上海科学技术出版社, 1990
- [17] 王其文. 经济管理计算机基础教程. 北京: 高等教育出版社, 1999
- [18] 周概容. 应用统计方法辞典. 北京: 中国统计出版社, 1993