

李健宁 著

多元分析及其在 高等教育研究中的应用

DUOYUAN FENXI JIQI ZAI

GAODENG JIAOYU YANJIU ZHONG DE YINGYONG

安徽大学出版社

江苏省重点学科研究成果

多元分析及其在高等教育研究中的应用

李健宁 著

安徽大学出版社

图书在版编目(CIP)数据

多元分析及其在高等教育研究中的应用/李健宁著.

—合肥:安徽大学出版社,2009.10

ISBN 978-7-81110-687-9

I. 多... II. 李... III. 多元分析—应用—高等教育
—教育研究 IV. G640

中国版本图书馆 CIP 数据核字(2009)第 190191 号

多元分析及其在高等教育研究中的应用 李健宁 著

出版发行 安徽大学出版社(合肥市肥西路3号 邮编 230039)
联系电话 编辑部 0551-5108087 发行部 0551-5107784
电子信箱 ahdxxhps@mail.hf.ah.cn
网 址 www.ahupress.com.cn
书 号 ISBN 978-7-81110-687-9
经 销 全国新华书店
印 刷 合肥创新印务有限公司
开 本 710×1000 1/16
印 张 11.25
字 数 180 千
责任编辑 徐玲英 谈 菁
封面设计 孟献辉
定 价 22.00 元
版 次 2009 年 11 月第 1 版
印 次 2009 年 11 月第 1 次印刷

如有影响阅读的印装质量问题,请与出版社发行部联系调换

前 言

本书是作者多年从事多元统计教学及研究的结晶。本书的主要特点是将统计原理分析与实际研究案例及 SAS 软件操作有机结合,从而弥补了以往只注重统计原理阐述,缺少研究案例和软件操作的原理型教材或单纯介绍统计软件等应用手册的不足。本书内容既包括多元线性回归、因子分析等传统多元分析方法,又涉及结构方程模型、广义可加模型等前沿方法。本书献给那些缺乏专门的高等数学知识而又需要应用较为复杂的统计分析方法的读者,如教育学、心理学、社会学、经济学、管理学等学科的研究生和研究工作者。当然,高年级的社会科学领域的大学生通过自学本书,并借助于 SAS 软件或其他统计软件,也可以很好地在自己的研究中应用这些方法。

作者在介绍本人的研究成果的同时,也引用了部分参考文献中的研究案例,在此向有关作者深表感谢。

本书作为江苏省重点学科(高等教育学)的研究成果,得到了苏州大学教育学院领导和老师们的大力支持,安徽大学出版社的编辑谈菁老师为本书出版付出了辛勤劳动,我的妻子武小椿为我长期的工作和写作创造了一个优良的环境,我向他们表示衷心的感谢。

李健宁

2008 年于苏州大学

目 录

第 1 章 多元分析概述	1
1.1 引言	1
1.2 多元分析的应用	3
1.3 多元数据的组织	7
1.3.1 平均数	8
1.3.2 方差与标准差	8
1.3.3 随机变量	9
1.3.4 随机向量	10
1.3.5 多元数据	10
1.3.6 平均值向量	11
1.3.7 标准化数据矩阵	12
1.3.8 协方差矩阵	13
1.3.9 相关矩阵	14
第 2 章 多元线性回归模型(上)	16
2.1 多元线性回归分析	16
2.1.1 多元线性回归模型	16
2.1.2 多元线性回归方程的建立	17
2.1.3 回归效果的检验	18
2.1.4 高等学校生均教育经费影响因素分析	20
2.1.5 应用和解释多元回归分析时存在的潜在问题	21

2 多元分析及其在高等教育研究中的应用

2.2 逐步回归分析	23
2.2.1 自变量的选择	23
2.2.2 逐步回归分析的基本原理	24
2.2.3 区域高等教育发展水平影响因素分析	25
2.2.4 高职生职业价值观与自主性学习关系的实证研究	26
2.2.5 研究生学术忠诚影响因素实证研究	27
2.3 路径分析	30
2.3.1 路径分析的基本原理	30
2.3.2 路径分析在教育发展战略研究中的应用	34
2.4 多项式回归	36
2.5 趋势面分析	38
2.5.1 趋势面模型的建立	39
2.5.2 趋势面模型的参数估计	40
2.5.3 趋势面模型的适合度检验	40
2.5.4 普通高等学校本专研在校生比例区域分布模型	42
第3章 多元线性回归模型(下)	45
3.1 多对多回归	45
3.1.1 多对多线性回归的基本原理	45
3.1.2 区域经济水平与高等教育发展关系的回归分析	46
3.2 典型相关分析	49
3.2.1 典型相关分析的基本原理	49
3.2.2 区域经济发展与高等教育发展的典型相关分析	51
3.2.3 高等学校学科竞争潜力与学科竞争实力的典型相关分析	56
3.3 多重共线性	62
3.3.1 多重共线性的诊断方法	62
3.3.2 多重共线性的诊断案例分析	64
3.4 岭回归分析	66
3.4.1 岭回归的基本原理	67
3.4.2 区域高等教育发展影响因素的岭回归分析	67
3.5 偏最小二乘回归	68
3.5.1 偏最小二乘回归的基本原理	69

3.5.2 区域经济水平与高等教育发展关系的偏最小二乘回归分析	71
第4章 协方差结构分析	73
4.1 主成分分析	73
4.1.1 主成分分析的基本模型	73
4.1.2 主成分的个数及解释	74
4.1.3 区域高等教育发展能力的综合分析评价	74
4.2 主成分回归	77
4.3 因子分析	79
4.3.1 因子分析的基本模型	80
4.3.2 因子分析模型中的主要统计量	81
4.3.3 因子分析的基本步骤	82
4.3.4 因子得分	83
4.3.5 我国区域高等教育发展水平的统计分析	86
4.3.6 因子分析与主成分分析的联系与区别	93
4.4 结构方程模型	93
4.4.1 结构方程模型的基本形式	95
4.4.2 从探索型因子分析到验证型因子分析	97
4.4.3 结构方程模型的基本步骤	98
4.4.4 结构方程模型的参数估计	99
4.4.5 结构方程模型的评价	100
4.4.6 男大学生无气力症调查分析	105
4.4.7 研究生学术努力行为的影响机制研究	109
第5章 分类与分组方法	116
5.1 聚类分析	116
5.1.1 聚类分析的方法	116
5.1.2 聚类统计量	117
5.1.3 系统聚类法	118
5.1.4 中国大学前100名聚类分析	119
5.2 判别分析	122
5.2.1 判别分析的种类	122

4 多元分析及其在高等教育研究中的应用

5.2.2 中国大学前 100 名判别分析	123
5.2.3 判别分析与因子分析的联系与区别	126
5.3 多维标度法	126
5.3.1 古典多维标度法的基本原理	127
5.3.2 高等学校科学研究发展水平的 MDS 分析	128

第 6 章 属性数据分析

131

6.1 Logistic(逻辑斯谛)回归分析	131
6.1.1 Logistic(逻辑斯谛)回归分析的基本原理	132
6.1.2 Logistic(逻辑斯谛)回归分析案例	133
6.1.3 多分类无序反应变量 Logistic 回归	134
6.1.4 多分类有序反应变量 Logistic 回归	136
6.2 对数线性模型	138
6.2.1 对数线性模型的基本原理	139
6.2.2 大学生闲暇活动分析的对数线性模型	142
6.3 对应分析	147
6.3.1 对应分析的基本原理	148
6.3.2 头发颜色与眼睛颜色的对应分析	149
6.3.3 美国大学竞争力的对应分析	153

第 7 章 混合线性模型与广义可加模型

156

7.1 混合线性模型	156
7.1.1 混合线性模型的结构	157
7.1.2 混合线性模型的 SAS 实现	158
7.1.3 高等学校工资增长趋势研究	159
7.2 广义可加模型	164
7.2.1 可加模型和广义可加模型	165
7.2.2 广义可加模型的 SAS 实现	166
7.2.3 女性生育一孩率影响因素分析	166

参考文献

169

第 1 章 多元分析概述

1.1 引 言

科学研究是一个反复学习的过程^{〔1〕}。对于高等教育研究来说,首先必须指定一些与某种教育现象有关的解释作为目标,然后通过收集数据和分析数据对这些目标进行检验。对通过实验或观察收集来的数据进行分析之后,我们通常会对现象提出一个改进的解释。在这个反复学习的过程中,往往会有一些教育、经济或社会变量被增加到研究中去,有些则会被剔除。正是由于教育或社会现象研究的复杂性,需要研究者去收集许多不同变量的观察值。为此,只有采用能充分地考察这些数据集的多变量技术(Multivariable Technique)或多元分析(Multivariate Analysis)方法,才有可能得到对教育现象更为全面、更接近实际的理解。正如一位研究者(Gatty R)所说的,“对于任何应用领域的研究,我们的大多数工具都是、或者都应该是多元的技术。对于一个问题的结论,除非该问题是按多元问题来对待的,否则其处理就只可能是表面的、肤浅的”^{〔2〕}。

当我们需要了解许多变量之间的关系时,多元分析必然成为一个困难的问题。因为我们的大脑常常被一大堆数据弄得不知所措;另一方面,多元统计

〔1〕 Richard A. Johnson, Dean W. Wichern. 实用多元统计分析[M]. 陆璇译. 北京:清华大学出版社,2001:3。

〔2〕 引自柯惠新,祝建华,孙江华. 传播统计学[M]. 北京:北京传播大学出版社,2003:247。

2 多元分析及其在高等教育研究中的应用

方法的推导和计算比在一元情形下需要更多的数学知识,这对于人文、社会科学领域的学生和研究工作者来说确实是一个较大的挑战。令人欣慰的是,计算机技术的发展和普及为多元分析的实际应用提供了重要的条件。20 多年以前,社会科学学科的研究生若能利用因子分析处理调查数据,也许就有可能形成一篇博士论文;而现在,从事人文科学、社会科学或行为科学的研究者,能借助计算机统计软件(如 SAS、SPSS、BMDP 等)对数据进行多元分析的已大有人在。由于统计软件的发展,因子分析等昔日十分高深的统计技术,现在已经达到了十分普及的程度。

不过,为了能在高等教育研究中正确地应用多元分析技术,必须系统地学习这些方法的基本原理、掌握其适用的场合以及学会如何解释分析的结果。否则一知半解地去使用统计软件,就可能得到一堆毫无意义的甚至是完全错误的数据,这对科学研究只会是有害而无益的。多元分析的基本原理涉及较为复杂的数学推演。这里,我们只是从应用的角度出发,力图能以通俗的方式表达其中的统计原理和思想,尽量回避那些过于复杂的数学演算和证明,以方便缺乏数学基础的读者学习。然而,多元分析模型的基础毕竟是概率论和数理统计,不可能完全绕过某些高等数学知识,这就要求读者具有进行定量思考的能力。

统计分析系统(Statistical Analysis System, SAS)已有 30 多年的历史,它是美国 SAS 软件研究所研制的一套大型集成应用软件系统,具有完备的数据存取、数据管理、数据分析和数据展现功能,它与其他软件包(如 SPSS 和 BMDP)在设计目标上有根本的区别。SPSS 和 BMDP 着眼于提供独立的分析过程,而 SAS 的所有分析过程(Procedures)只是作为一个可编程序母体语言(Programmable Language)的一部分,事实上它相当于一种具有较多过程和函数的语言。遇上所有现成过程都不能直接应用的情况, SAS 用户可应用母体语言直接进行计算,也可融合一些现成过程并补充成为一个新过程。从这个角度讲, SAS 作为专业统计软件,现在还很难有其他统计软件能与之抗衡。然而,由于 SAS 系统是从大型机上的系统发展而来的,其操作至今仍以编程为主,人机对话界面不太友好,系统地学习与掌握需要花费一定的时间和精力。对大多数研究工作者而言,需要掌握的仅是如何利用统计软件来解决自己的实际问题。

本书在有限的篇幅内涉及 21 种多元统计的方法,包括多元线性回归、路径分析、逐步回归、多项式回归、趋势面分析、多对多回归、典型相关分析、岭回归、偏最小二乘回归、主成分分析、主成分回归、因子分析、结构方程模型、聚类分析、判别分析、多维标度法、Logistic 回归、对数线性模型、对应分析、混合线性模型、广义可加模型。这其中的许多方法如果展开论述都可以形成一本书。笔者只是将每种方法的基本思想、适用条件、SAS 程序语言和研究案例及结论展示给读者,如果要了解各种方法的原理和 SAS 输出的详细结果需要参考专门的教材或应用手册。

1.2 多元分析的应用

Kotz 和 Johnson 主编的 13 卷“统计科学百科全书”(Encyclopedia of Statistical Sciences)是迄今最完整的关于统计的权威性的百科全书。它列举了 40 多个运用统计的领域,它们包括精算、农业、动物学、人类学、考古学、审计学、晶体学、人口统计学、牙医学、生态学、经济计量学、教育学、选举预测和策划、工程、流行病学、金融、水产渔业研究、遗传学、地理学、地质学、历史学、人类遗传学、水文学、工业、法律、语言学、文学、劳动力计划、管理科学、市场营销学、医学诊断、气象学、军事科学、核材料安全管理、眼科学、制药学、物理学、政治学、心理学、心理物理学、质量控制、宗教研究、社会学、调查抽样、分类学和气象改善^[1]。

我们通过一些实际的问题,解释统计方法的选择和研究目的之间的关系,这些问题以及本书中的大量案例能够使读者对多元统计分析方法在各个领域中的广泛应用有一定的了解。多元分析方法对不同的研究问题具有相应的具体解决问题的方法,这可参看表 1-1。

[1] 吴喜之. 非参数统计[M]. 北京: 中国统计出版社, 1999: 1.

4 多元分析及其在高等教育研究中的应用

表 1-1 统计方法和研究内容之间的关系

问 题	内 容	方 法
数据或结构性简化	尽可能简单地表示所研究的现象,但不损失很多有用的信息,并希望这种表示能够很容易解释	多元回归、逐步回归、多对多回归、聚类分析、主成分回归、因子分析、对应分析、多维标度法、典型相关分析
分类和组合	基于所测量到的一些特征,给出好的分组方法,对相似的对象或变量分组	聚类分析、判别分析、主成分分析、多维标度法
变量之间的相关关系	变量之间是否存在相关关系,相关关系又是怎样体现的	多元回归、逐步回归、路径分析、主成分回归、多项式回归、趋势面分析、多对多回归、典型相关分析、岭回归、偏最小二乘回归、主成分分析、因子分析、结构方程模型、多层线性模型、广义可加模型、Logistic 回归、对应分析、对数线性模型
预测与决策	通过统计模型或最优准则,对未来进行预见或判断	多元回归、聚类分析、判别分析
假设的提出及检验	检验由多元总体参数表示的某种统计假设,能够证实某种假设条件的合理性	多元总体参数估计、假设检验

下面,我们从教育学、经济学、管理学及文学四个学科的部分文献看看多元分析广阔的应用领域与适用范围。

教育学方面

Logistic 分布在教育测量理论中的应用^{〔1〕}、主成分分析在数学建模中的应用^{〔2〕}、高等学校经费支出的对应分析方法^{〔3〕}、路径分析在教育发展战略研究中的应用^{〔4〕}、我国普通高等教育发展水平的全局主成分分析^{〔5〕}、大学本科

〔1〕 黄小平,黄斌,黄本利. Logistic 分布在教育测量理论中的应用[J]. 湖北工学院学报, 1998(1):88—87。

〔2〕 朱宁,陈克西,李竹梅. 主成分分析在数学建模中的应用[J]. 工科数学,1999(4): 106—110。

〔3〕 刘喜华,吴育华. 高等学校经费支出的对应分析方法[J]. 教育与经济,2000(2):34—37。

〔4〕 李健宁. 路径分析在教育发展战略研究中的应用[J]. 系统工程理论与实践,2001(8): 142—143。

〔5〕 段仁军,张伟. 我国普通高等教育发展水平的全局主成分分析[J]. 数理统计与管理, 2002(6):27—31。

基础课和专业课关系的典型相关分析法研究^[1]、Q聚类分析在分类指导型教学中的应用^[2]、主成分分析法在高校共青团工作评价中的运用^[3]、高等学校工资增长趋势研究^[4]、结构方程模型及其在语言测试中的应用^[5]、变量聚类—全局主成分分析在我国普通高等教育发展水平评价中的应用^[6]等。

经济学方面

中国经济增长的均衡路径分析^[7]、中国粮食生产系统影响因素的主成分分析^[8]、房地产业预警预报系统影响因素的主成分分析^[9]、对应分析法在居民消费结构中的应用^[10]、长江三角洲各城市综合实力的主成分分析^[11]、聚类分析在品牌市场定位研究中的应用^[12]、基于 Poisson 对数线性模型的居民点

[1] 佟磊,姚俭.大学本科基础课和专业课关系的典型相关分析法研究[J].上海理工大学学报(社会科学版),2003(2):70—73。

[2] 田振清,周越,王春霞,陈梅.Q聚类分析在分类指导型教学中的应用[J].内蒙古师范大学学报(自然科学汉文版),2004(2):140—144。

[3] 郑畅,莫丽丽.主成分分析法在高校共青团工作评价中的运用[J].武汉理工大学学报(社会科学版),2004(3):393—395。

[4] 刘勇,张贵生.高等学校工资增长趋势研究[J].山西大学学报(哲学社会科学版),2004(2):115—118。

[5] 韩宝成.结构方程模型及其在语言测试中的应用[J].现代汉语,2006(1):76—86。

[6] 徐雅静,汪远征.变量聚类—全局主成分分析在我国普通高等教育发展水平评价中的应用[J].数理统计与管理,2006(5):566—573。

[7] 雷钦礼.中国经济增长的均衡路径分析[J].统计研究,2002(6):10—14。

[8] 张志强.中国粮食生产系统影响因素的主成分分析[J].北京农学院学报,2001(1):65—68。

[9] 周忠学,李永江.房地产业预警预报系统影响因素的主成分分析[J].企业经济,2003(12):130—132。

[10] 刘坚,赵振业,刘金国.对应分析法在居民消费结构中的应用[J].济南大学学报(自然科学版),2004(1):12—15。

[11] 叶依广,周蕾.长江三角洲各城市综合实力的主成分分析[J].长江资源流域与环境,2004(3):197—202。

[12] 黄劲松,赵平.聚类分析在品牌市场定位研究中的应用[J].数理统计与管理,2005(1):21—26。

6 多元分析及其在高等教育研究中的应用

与地理因子的相关性研究^[1]、安全生产与经济社会发展多元回归分析^[2]、农民增收的主成分回归分析^[3]、基于岭回归分析法的中国区域经济差异影响因素分析^[4]等。

管理学方面

管理人员个性的因子分析及其科学评价问题的研究^[5]、典型相关分析模型在企业并购中的应用^[6]、顾客满意度线性建模中多重共线性处理方法的模拟研究^[7]、顾客满意度指标重要性测量的主成分分析与多元回归方法^[8]、累积 Logistic 回归在企业竞争力评价中的应用^[9]、基于最小二乘法的产品满意度预测模型^[10]、结构方程模型理论及其在管理研究中的应用^[11]、结构方程模型用于顾客满意度测评之实际应用^[12]等。

[1] 董春,罗玉波,刘纪平.基于 Poisson 对数线性模型的居民点与地理因子的相关性研究[J].中国人口·资源与环境,2005(4):79-83。

[2] 黄盛初,周心权,张斌川.安全生产与经济社会发展多元回归分析[J].煤炭学报,2005(5):580-584。

[3] 吴航,黄恒君.农民增收的主成分回归分析[J].中国统计,2005(12):20-30。

[4] 李政,钟永红.基于岭回归分析法的中国区域经济差异影响因素分析[J].统计与决策,2006(2下):103-105。

[5] 翟洪昌,许铎.管理人员个性的因子分析及其科学评价问题的研究[J].心理科学,2000(2):211-213。

[6] 张天平.典型相关分析模型在企业并购中的应用[J].吉首大学学报(自然科学版),2002(3):88-93。

[7] 张新安,田澎.顾客满意度线性建模中多重共线性处理方法的模拟研究[J].工业工程与管理,2004(4):73-77。

[8] 刘金兰,朱晓.顾客满意度指标重要性测量的主成分分析与多元回归方法[J].天津大学学报(社会科学版),2004(2):159-163。

[9] 熊巍,赵海娟,程红莉.累积 Logistic 回归在企业竞争力评价中的应用[J].统计与信息论坛,2004(1):85-88。

[10] 华尔天,李国富,毛明杰.基于最小二乘法的产品满意度预测模型[J].中国机械工程,2005(20):1831-1834。

[11] 林嵩,姜彦福.结构方程模型理论及其在管理研究中的应用[J].科学学与科学技术管理,2006(2):38-41。

[12] 程开明.结构方程模型用于顾客满意度测评之实际应用[J].市场研究,2006(5):33-39。

文学方面

1985年至1986年间,复旦大学李贤平运用聚类分析、主成分分析、典型相关分析法对《红楼梦》的著作权进行了研究,形成了关于《红楼梦》作者和成书过程的新学说,在红学界引起轰动。李贤平的主要结果是^[1]:(1)前80回和后40回截然地分为两类,证实了前80回和后40回不是出于一人的手笔;(2)前80回是曹雪芹所写。通过与曹雪芹的另一著作做类似的分析,结果证实了用词手法完全相同,据此断定为曹雪芹一人手笔;(3)后40回非高鹗一人所写。分析结果发现后40回依回目的先后可分为几类,得出的结论推翻了后40回是高鹗一人所写的论断。后40回成书比较复杂,既有残稿也有外人笔墨,不是高鹗一人所续。

1.3 多元数据的组织

对两个以上变量进行观察,得到 N 个(组)数据,这种数据称为多元数据(Multi-Dimensional Data)。例如,表1-2是10名大学生性格测量数据(部分数据)。

表 1-2 10 名大学生性格测量数据(部分)

编号	社交性	活动性	进取性	持久性	规律性	神经质
1	12	3	6	2	1	0
2	14	14	19	12	20	16
3	7	9	14	12	15	8
4	4	8	16	10	8	2
5	5	7	16	18	11	4
6	11	8	13	10	15	11
7	12	9	14	2	0	8
8	10	10	15	18	18	11
9	4	1	6	11	9	2
10	10	10	12	15	13	16

根据表1-2的数据,我们讨论多元数据统计量的计算以及矩阵表达形式。

[1] 高惠璇. 应用多元统计分析[M]. 北京:北京大学出版社,2005:8-9。

1.3.1 平均数

对数据的量化描述主要有两个方面的指标：一是描述数据的中心位置；另一是描述数据的分散(变异)程度。这里介绍一下平均数(算术平均数)。它是从一组数据中抽象出来的一个数，是一组数据的代表值，表明事物的数量方面在一定时间、地点和条件下的共同性质的一般水平。对于单一变量可以用向量 x 表示。例如，对 10 名大学生性格测量数据的社交性用向量 x 表示，其中，第 i 个测量值表示为 x_i 。

$$x = (x_1 \ x_2 \ x_3 \cdots x_8 \ x_9 \ x_{10}) = (12 \ 14 \ 7 \cdots 10 \ 4 \ 10) \tag{1.1}$$

这时平均数为

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \tag{1.2}$$

其中， N 是预测对象的数目，称为样本容量。计算出的 6 个观测变量的平均数见表 1-3。

表 1-3 10 名大学生性格测量数据的平均数

社交性	活动性	进取性	持久性	规律性	神经质
8.900	7.900	13.100	11.000	11.000	7.800

1.3.2 方差与标准差

描述差异程度的指标有很多，这里我们介绍一下方差与标准差。方差有样本方差(sample variance)与无偏样本方差(unbiased sample variance)两种。分别定义为

$$s^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \tag{1.3}$$

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \tag{1.4}$$

方差的平方根定义为标准差，有样本标准差 s 与无偏样本标准差 $\hat{\sigma}$ 。

$$s = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \tag{1.5}$$

$$\hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2} \tag{1.6}$$

观察表 1-4, 规律性的离散程度最大, 其次是持久性, 离散程度最小的是社交性。

表 1-4 10 名大学生性格测量数据的离散度

	社交性	活动性	进取性	持久性	规律性	神经质
s^2	1.890	12.090	15.890	36.100	40.000	29.760
$\hat{\sigma}^2$	13.211	13.433	17.656	40.111	44.444	33.067
s	3.448	3.477	3.989	6.008	6.325	5.455
$\hat{\sigma}$	3.635	3.665	4.202	6.333	6.667	5.750

1.3.3 随机变量

我们定义一个变量 X , 所求随机事件可以表示为 X 的取值。这种把随机事件“数字化”的思想导致了“随机变量”概念的诞生, 极大地推动了概率论的发展。

在一些随机试验中, 试验的结果本身就是由数量来表示的。例如, 投掷一枚骰子, 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; 观察一个灯泡的使用寿命, 实际使用寿命可能是 $[0, +\infty]$ 中任何一个实数。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一元钱, 出现反面时输一元钱, 则可对“出现正面”指定一个数“1”, 对出现反面指定一个数“-1”。无论哪种情况, 其共同点是: 对每一个可能结果, 有唯一的一个实数与之对应。这种对应关系实际定义了样本空间 Ω 上的函数, 通常记作

$$X = X(\omega) \quad (\omega \in \Omega)$$

随机变量 X 的值与其概率 p_k 组合形成概率分布。随机变量的期望值为

$$E[X] = \sum_{k=1}^l p_k x_k \quad (1.7)$$

利用 p_k 的估计值 $\hat{p}_k$

$$\sum_{k=1}^l \hat{p}_k x_k = \sum_{k=1}^l \frac{N_k}{N} x_k = \frac{1}{N} \sum_{k=1}^l N_k x_k = \frac{1}{N} \sum_{i=1}^N x_i = \bar{x} \quad (1.8)$$

可见, 随机变量 X 的期望值与平均数是一致的。

对于期望值, 下面三个公式是很重要的。

$$E[a] = a \text{ (常量的期望值是常量)}$$