

• 马庆国 著 •

[管理]

统计

数据获取、统计原理
SPSS工具与应用研究



科学出版社

www.sciencep.com

管 理 统 计

——数据获取、统计原理、SPSS 工具与应用研究

马庆国 著

科 学 出 版 社

2002

内 容 简 介

本书依托人文社会科学领域使用最广泛的 SPSS 统计软件为基本工具, 结合实际案例讲述了人文社会科学特别是管理科学中常用的统计学原理、方法与技术。本书与注重数学证明的传统统计学著作和就操作谈操作的 SPSS 软件使用手册的不同之处在于, 略去了复杂的数学推导过程, 强调统计学在人文社会科学特别是管理学中的应用, 强调对 SPSS 软件计算原理的透彻理解与正确应用, 努力构建一个贯通统计学原理, SPSS 软件使用技巧以及应用统计学进行人文社会科学研究的的应用型统计学知识体系。读者学习本书后, 既能够根据实际工作需要设计统计调查方案、又能够利用 SPSS 软件整理、分析数据和解释数据分析结果, 从而大大提高分析和解决问题的综合能力。

本书可作为经济管理类专业以及其他人文社会科学领域的统计学教材, 也可供其他对应用统计学和 SPSS 软件有兴趣的实际工作者参考。

图书在版编目 (CIP) 数据

管理统计: 数据获取、统计原理、SPSS 工具与应用研究/马庆国著. —北京: 科学出版社, 2002

ISBN 7-03-010750-0

I. 管… II. 马… III. 经济统计学 IV. F222

中国版本图书馆 CIP 数据核字 (2002) 第 060807 号

科学出版社 出版

北京东黄城根北街16号

邮政编码: 100717

<http://www.sciencep.com>

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2002 年 8 月第 一 版 开本: B5 (720×1000)

2002 年 8 月第一次印刷 印张: 26

印数: 1—4 000 字数: 483 000

定价: 39.00 元 (含软盘)

(如有印装质量问题, 我社负责调换〈环伟〉)

前 言

本书目的

无论人们研究还是解决问题,前提都是正确地发现问题、分析问题,从而认识其本质。

如同研究其他问题一样,研究和解决管理科学以及其他社会科学领域的问题,也只有清楚地认识对象所处的状况、存在的问题之后,才能够提炼出对象的本质,从而得出对症的解决方案。

在社会科学领域特别是管理科学领域,最常用的研究方法之一是科学地设置问卷、通过正确地抽样调查获得数据,进而恰当地运用统计学的工具处理数据,并依据对统计结果的正确解释来导出结论、做出决策。在上述任何一个环节上出错,都会导致人们对问题的错误认识,导致错误的结论、决策或处理方案。

一个社会科学研究结论在付诸于实践之前,人们难以从结论本身判断其正误,只能从其研究的基础数据来源是否合理、问卷设计是否正确、抽样方法是否科学、调查过程是否能够保证数据的可靠性、数据处理是否正确等方面,来判断该结论是否建立在充分的论据基础之上,其论证过程在逻辑上是否一致,从而判断该结论是否成立。这种判断方法,正是国际上在研究结论付诸于实践之前判断管理学以及其他社会科学研究成果是否成立的基本方法。

改革开放以来,中国的管理科学研究和其他社会科学领域的研究都有了长足的进步,但依然存在着许多需要改进的问题。目前,最突出的是研究方法的规范化问题。迄今为止,在管理学以及其他社会科学领域中,我国大多数的研究论文的主流研究方法还限于总结性的、思辨性的方法,对于国际上通用的实证性方法、实验与对比实验方法、问卷调查和数据处理方法,以及长期跟踪观察的方法,还运用得比较少。而这些国际通用的研究方法正是管理学和其他社会科学领域的规范化研究方法,是国际学术交流的共同语言。研究方法不规范,研究成果就很难国际化。

让读者掌握具有重要实用价值的、常用的统计学原理、方法与工具,提高依据实际问题设计调查方案的能力,以及依托统计结果处理实际问题的能力,或者说,帮助读者掌握管理科学以及其他社会科学领域的规范的研究方法,掌握以事实证明或推翻他人的理论的方法,或用事实证明新的理论的方法,就是本书的目的。

本书的特点

统计研究,特别是结合实际问题的统计研究,往往涉及大量的数学运算。随着现代计算机技术的发展,已经开发了出许多面向用户、界面友好、功能强大的统计软件。SPSS 软件就是目前社会科学领域使用最广泛、功能最为强大的统计软件之一。SPSS 软件简化了统计计算和研究工作,降低了运用统计研究方法的数学门槛,大大提高了工作效率,为人文社会科学领域统计方法的普遍使用,提供了便利条件。

但由于学科分隔,现有的“数理统计学”、“应用统计学”著作几乎都特别注重数学推导和证明,忽视数据的获取,忽视软件的使用,而讲述 SPSS 软件的著作又往往就操作谈操作,无法深入到统计原理以及应用统计技术的层面。因此,在统计学原理、方法与研究实际问题以及计算机统计软件三者之间,依然存在着深深的知识鸿沟。

本书的特点在于面向实际工作者,强调应用。书中略去了复杂的数学推导过程,强调统计原理和技术在人文社会科学领域特别是管理科学中的应用,强调对 SPSS 软件计算原理的透彻理解和正确应用,努力构建一个贯通统计学原理、SPSS 软件使用技巧以及应用统计学于人文社会科学领域的完整的应用型统计学知识体系。读者学习本书后,既能够根据实际工作需要设计统计调研方案、将实际问题转化成统计问题,又能够利用 SPSS 的强大功能整理分析数据,并通过科学的数据解释,将统计结果运用到实践中去。

本书的特点可以概括为:**略去证明,讲清原理;
依托软件,突出应用。**

本书的使用范围

管理统计学是管理科学领域中应用最广泛的数量研究方法。对大多数管理问题的研究而言,正确掌握和应用管理统计方法,是非常重要的。此外,管理统计学的研究思路和技术方法也可以为其他社会科学研究,如经济学、社会学、行政学、科学学、民族学、人类学、考古学等研究领域的借鉴,同时在医学、生物、工学等领域也有非常重要和广泛的运用,是这些学科开展科学研究的重要而有用的研究工具。

本书适用于对于应用统计技术和 SPSS 软件有兴趣的实际工作者和研究者,也可兼作教材,供经济管理类专业以及其他人文社会科学领域的师生学习和参考。

本书的适用层次

本书是作者在长期的管理科学研究与教学过程中应用统计学解决实际问题的心得总结。为了推动统计学方法在人文社会科学研究中的应用,使更多的读者掌握管理统计学知识,本书在编排结构上,借鉴了国内外优秀教材的办法,努力作到条理清楚、举例精到、详略得当,对适用于不同层次读者的内容进行了标注。

其中,没有任何记号的章节适用于大专层次的读者;

没有任何记号的章节与标注一个星号的章节,适用于本科层次的读者;

标注两个星号的章节适用于研究生层次的读者;

在本科阶段没有学习过统计学的研究生,应从头学起;

不计划学习概率论的读者可以绕过第1章,直接从第2章开始。

以上划分仅供参考,教师可根据各学校的实际情况进行调整。

本书的创新

本书是作者多年研究与教学的成果,在借鉴前人工作的基础上,作者倾注了大量的心血,提出了一些比较新颖的看法,供读者批评和参考。

在第1章(概率论基础知识)中,作者对概率论做了扼要的概括,拎出了在实际工作中被频繁用到的基础知识(基本概念、特点和结论,但不包括数学证明)。并且强调对这些知识的直观理解,强调用简单、直观的事例,来理解和记忆严格的数学描述,例如,用“抛硬币”来记住构成随机实验的三个要素(1个关于实验条件,2个关于实验结果)。对于学过概率论的读者而言,这有利于他们从更高的角度概括性地理解概率论,通过对基础知识的透彻理解,产生对概率论认识的升华。

对于没有学过概率论的读者而言,依照如此扼要的归纳,为着应用的目的来学习概率论的主脉络,也不失为一个较好的选择。

在这一章中,作者设计了“条件概率计算表”。运用这张表,不仅有利于读者(特别是初学者)绕开条件概率公式、乘法公式、全概率公式、贝叶斯公式,完成有关计算,而且有利于深刻地理解上述公式,理解边缘分布的概念等。

在这一章中,作者还概括了“离散分布和连续分布的对应关系”(离散型随机变量的概率与连续型随机变量的概率密度对应,离散型随机变量的累计概率与连续型随机变量的概率对应,见1.2.2小节)。这有利于读者正确地理解“概率论”的许多相关的结论,有利于读者创造性地处理好诸多的相关问题。

在第2章(数据与数据的获得)中,作者介绍了在实际应用中非常重要而国内有关“数理统计学”著作几乎不涉及的数据的测度类型问题,介绍了在管理科学中

(包括所有社会科学中)极为重要的实验数据的问题(见 2.2.3 小节和 2.4.1 至 2.4.3 小节)。

社会科学中的实验数据与自然科学中的实验数据同等重要。但是,由于社会科学中的不确定因素多,边界条件的可控性弱、获取实验数据的难度要更大一些。在社会科学领域,一个好的实验设计(哪怕仅仅是为了获得一些关键的数据的实验),本身就可以作为一项成果,在相应的学科领域流传。在这一章中,作为例子,还给出了作者在获取实验数据方面的某些研究成果(见 2.4.3 小节关于个人所得税问题的实验设计)。这些成果不一定是优秀的,但却是有参考意义的,可能会启发读者设计出更好的实验方案。

在这一章中,还给出了作者在科学研究和指导研究生中产生的有关问卷设计的原则、要点与关键。对这些知识成果的了解,将有利于读者在问卷设计中,少犯错误,少走弯路。

在第 3 章(样本数据特征的初步分析)中,作者把“数据的频次与频率”作为数据结构的最基本的特征来介绍。并且,针对数据的不同测度,介绍了众数、中位数、均值、极值、四分点、极差、四分位距、离差等,这些几乎所有统计学书籍都要介绍的概念(不同之处仅在于本书把它们与数据的不同测度联系起来了)。

这一章还介绍了表示数据集合基本结构特征的条形图、茎叶图、直方图和箱型图等图示方法。其中,有一些是我国学者常用的工具,如条形图、直方图。有一些是我国学者不常用的,如茎叶图、箱型图。希望本书的这些介绍,有利于这些工具的更广泛应用。

在制作直方图的数据分组方面,作者探讨了与 SPSS 相衔接的分组规则。这是摸索性的工作。

在第 4 章(SPSS 的简单应用)中,作者从“初学者应首先学会最重要的少数操作”的原则出发,挑选了能够使初学者尽快进入使用过程的几个关键操作,并扼要地介绍了这些关键操作。并且,从数据录入的实战出发,着重介绍了“数据并表”的方法和技巧。这些是作者在教学与研究中的体会。

在这一章,作者注入的新工作主要是“数据录入”的技巧、“核对录入正误”的技巧(为回收的调查表编号并作为变量录入),详见 4.1.7 与 4.1.10 小节。

对于第 5 章(总体分布、样本分布与参数估计),作者的主要意图是,让初学者明了样本与总体的关系,特别是从随机样本推断总体的关系。作者的主要工作是基于自己研究和应用的体会,筛选出应用面广的统计量、最重要的基础性定理、判断参数估计值优劣的准则等,并扼要地、分层次地介绍了它们。

在第 6 章(参数假设检验)中,作者注入的新工作主要是,在判断接受还是拒绝零假设的问题上,填补了国内传统的“数理统计学”查表判别与 SPSS 输出统计值的显著性概率之间的空挡,沟通了它们之间的衔接问题(国内的“数理统计学”是通

过查表并比较统计值与统计值的阈值,来判别拒绝还是接受零假设的。SPSS 是输出统计值的显著性概率,通过它与选取的显著性水平的比较,来判别拒绝还是接受零假设的),详见 6.2.2 小节。

鉴于“参数假设检验”在应用中的基础性、广泛性与重要性,本书较为宽泛地介绍了诸多参数的假设检验问题,特别是饶有趣味地介绍了“为什么 60 分及格”的统计学依据,着重介绍了在管理科学以及诸多社会科学领域中有着广泛应用的大样本任意分布的参数检验问题,特别是某些类个体占总体比例的检验问题(详见 6.3.1、6.3.2、6.5.1、6.5.2 小节)。

此外,本书还通过例子(6.6.1 小节的例 6.6.1:以往某年级小学生 400 米平均成绩 100 秒,今年测 60 名同年级小学生的平均成绩是 105.385 秒,检验结果是接受零假设“平均成绩没变”),说明了常规“感觉”的不可靠性和参数假设检验的必要性。

在第 6 章的 6.6.1 小节,作者依据推算,详尽地解释了 SPSS(在有关模块中)与常规表示方法不同的、对置信区间的表示方法。在 6.6.2 与 6.6.3 小节,突出介绍了国内“数理统计学”很少涉及的独立样本与配对样本的检验问题。

在第 7 章(非参数假设检验)中,为了与 SPSS 衔接,作者尽可能多地介绍了相关的非参数检验的方法,如, Mann - Whitney 方法、Wilcoxon 方法、 χ^2 、Kolmogorov - Smirnov 方法等(在 SPSS 中, k - s 方法被设计为能够一次完成“一组样本是否服从正态分布、均匀分布、泊松分布、指数分布”的假设检验),但坚决地略去了证明,仅仅扼要地介绍原理、思路和结论,以便于读者使用 SPSS 软件,理解 SPSS 的输出结果。

在第 8 章(方差分析)中,作者用表格方式(见表 8.1.2,表 8.3.2,表 8.4.2),分别概括了单因素方差分析、双因素无重复实验方差分析和双因素有重复实验方差分析的理论模型及其主要的相关公式,扼要地介绍了数据处理的步骤,重点介绍了它们的应用,指出其实质是检验“不同因素不同水平的均值”有无显著差异(甚至检验有无交互作用)。

在这一章中,有些应用的构思是相当巧妙的。通过这些应用,能够筛选出作用显著的政策,找出政策的作用与企业类型的关系等(见 8.5.3 小节中例 8.5.2)。

在 SPSS 的应用方面,重点介绍了在“因素不同水平的均值”有显著差异的情况下,寻找“谁和谁有显著差异”的点击方法(即鼠标的点击过程)。

在第 9 章(普通相关分析)中,针对国内的绝大多数研究论文,在应用相关系数来说明变量之间的关系时,不做显著性检验的情况,作者特别介绍了“相关系数是否显著异于 0 的假设检验”的方法,并通过例题(9.5.3 小节中例 9.5.2:某班学生的数学成绩与语文成绩的等级相关系数高达 0.610,检验结果竟然是“相关系数 0.610 与 0 没有显著性差异”),说明“直觉”的不可靠性,相关系数显著性检验的必要

性。“不做显著性检验,就断言变量之间存在相关性”的论文其实是不科学的。而这种现象,在我国的相当多的学位论文,甚至是一些博士学位论文中,都是存在的。

在这一章中,针对国内的绝大多数(包括经济学和管理学在内的)社会科学研究的相关论文,在研究变量之间的相关关系时,普遍忽视偏相关分析的情况(只是简单地依据普通 Pearson 相关系数作结论的情况),作者特别介绍了偏相关系数的应用问题,通过例题(9.5.4 小节中例 9.5.3:旅游风景区的商业投资额与经济增长之间相关系数为 0.664, T 检验表明相关系数显著异于 0,但是两个变量之间的偏相关系数仅为 0.0825, T 检验表明偏相关系数与 0 没有显著性差异),说明仅仅从 Pearson 简单相关系数来做结论,是不恰当的,而剔除了其他变量影响的偏相关系数,是两个变量之间是否相关的真实反映。“不综合考虑一组相关联的变量,仅仅抓出其中的两个变量,通过普通相关系数来断定变量之间是否相关”的研究方式,在我国的有关的发表论文中包括一些学位论文中,也是普遍存在的。例如,仅仅从一个地区的税收优惠政策与该地区的经济增长之间的相关系数,断定优惠的税收政策导致了该地区的经济增长,就是这类错误的表现之一。正确的作法,应当是剔除该地区的其他经济政策(如货币政策、财政政策等)的影响之后,计算偏相关系数,并依据它来判定优惠的税收政策与该地区的经济增长,是否真正相关。

在变量值为名义级数据的情况下(名义级数据之间,不仅不能做加减乘除运算,而且不能比较大小,只知道不同的值是不同的,例如,红、白、黄不同的颜色,产品包装的不同类型等),判断变量之间是否存在相关关系(称为品质相关),原本是社会科学领域有着广泛应用基本问题,但在我国的相关研究论文中却很少出现。鉴于这种情况,作者在第 9 章着重介绍了检验是否存在品质相关的方法,给出了应用的示例。

在第 10 章(线性回归分析)中,作者注入的新工作主要是,对回归分析的几何解释(见 10.1.6、10.2.2、10.2.7、10.2.12、10.2.13、10.3.2 小节),通过几何解释,可以象征性地“推导”出多元线性回归系数的估计值 $\hat{\beta}$ 的计算公式,可以看到 F 统计量的几何含义、复相关系数的几何含义、偏相关系数的几何含义、偏解释变差(偏回归平方和)的几何含义。

在十多年前,我国学者在回归分析的应用中忽视对回归系数 T 检验的现象,到现在已经得到了基本改变。但是,在回归分析中重视高斯条件(高斯假设),恰当处理三大基本问题(多重共线性问题、异方差问题和序列相关问题)的论文,还相当少。为此,在第 10 章,作者强调了高斯条件,强调了三大基本问题的处理,特别是突出了依托 SPSS 软件,处理这三大基本问题的方法(只有这样,才能使得三大基本问题的处理走向实用。20 世纪 80 年代我国的不少相关著作,就介绍了回归分析的三大基本问题的处理,但是由于计算软件的不方便性,使得应用这些方法受到限制。现在软件发展了,对这三大基本问题的处理相对方便了,较为大面积地解决

这一问题的时机,已经基本成熟了。这也是本书这样安排这三大基本问题的依据)。

在第11章(主成分分析与因子分析)中,作者的主要工作是从实用出发,扼要地介绍主成分分析的作法以及其思路转换的脉络(略去几乎所有的数学证明),从全过程(数据的相关性检验、负载矩阵的求法、公共因子值的求法)扼要地介绍了因子分析的作法(略去所有的数学证明),并介绍了如何依托软件完成上述计算。

在第12章(聚类分析与判别分析)中,作者注入的新内容是,用简单的一维的例子、二维的例子来透彻理解聚类分析的思路与作法、来理解判别分析的概念、要素、性质与思路(见12.3.1小节)。为了读者便于理解,作者从“宏观把握”的角度,提炼出了聚类分析的思路与作法(跳出了陷于具体算法,而整体思路不清的状况),也基本类似地提炼了判别分析的思路与作法,并且,为了与SPSS软件衔接,做了大量的资料清理工作,以便读者更好地理解SPSS软件的输出结果。

聚类分析与判别分析,在管理科学、其他社会科学以及有关的自然科学中,有着非常广泛的应用前景,例如,聚类分析可以用于亚民族分类、方言分类、产业分类、企业分类(可偏重资产性指标,也可偏重企业文化类指标)、病种分类、菌体分类、病毒分类、动植物的物种分类等等,判别分析可以用于“依据财务数据对企业是否逃税”的判断、“依据有关指标对企业发展阶段”的判断、“依据消费者的特征对其购买与否”的判断、“依据病人的征兆指标对病人患病种类”的诊断、“依据地质的有关指标对是否将发生地震”的判断,“依据运动员的生理指标对其最适宜的运动项目”的判断等等。

鉴于此,作者选用了自己所主持的“企业税务稽查选案系统”的研究项目(见例12.6.2)作为例题,希望起到抛砖引玉的作用。

总而言之,本书做了大量的探索性、提炼性和创造性的工作,这些工作是作者长期从事科研工作和指导研究生时在统计学应用方面研究和思考的结晶,希望对我国管理科学以及其他社会科学领域在应用统计学方法进行基于数据的务实研究方面,能够有所裨益。当然,本书肯定还存在许多需要改进之处,也希望得到同行和各位读者的批评指正。

作者声明

作者对本书中所有的创新拥有知识产权。其他书籍凡是大量使用本书中作者创新成果的,均需获得本书作者的书面同意。仅引用本书观点和结论时必须注明出处。

目 录

第 1 章 概率论基础知识	(1)
1.1 随机实验、样本空间、概率与条件概率	(1)
1.2 随机变量与概率分布的基本概念	(12)
1.3 几个典型的概率分布	(24)
第 2 章 数据与数据的获得	(27)
2.1 总体、个体、特征与数据	(27)
2.2 数据类型	(28)
2.3 获得数据的调查方法与问卷设计	(32)
2.4 获得数据的实验方法	(43)
第 3 章 样本数据特征的初步分析	(49)
3.1 样本数据结构的基本特征:频次与频率	(49)
3.2 观察刻度级样本数据结构的茎叶图与直方图方法	(57)
3.3 样本数据的位置特征:对数据中心的描述	(67)
3.4 样本数据的离散特征	(73)
3.5 样本数据特征的综合表达:箱形图	(78)
第 4 章 SPSS 的简单应用	(83)
4.1 使用 SPSS 的基础知识	(83)
4.2 SPSS 的简单应用	(99)
第 5 章 总体分布、样本分布与参数估计	(121)
5.1 总体分布与样本分布	(121)
5.2 统计量与统计量的分布	(126)
5.3 点估计	(131)
5.4 判断点估计的优劣标准	(134)
5.5* 区间估计	(136)
5.6 SPSS 在参数估计中的应用	(139)
第 6 章 参数假设检验	(144)
6.1 假设检验的基本概念	(144)
6.2 一个正态总体下的参数假设检验	(145)
6.3 一个 0-1 总体分布下的参数假设检验	(155)
6.4 两个正态总体下的参数假设检验	(160)

6.5*	大样本下两个任意总体的均值检验	(164)
6.6	用 SPSS 作假设检验	(167)
第 7 章*	非参数假设检验	(182)
7.1*	两个总体分布的非参数假设检验	(182)
7.2*	一个总体分布的非参数假设检验	(189)
7.3*	用 SPSS 作非参数检验	(193)
第 8 章	方差分析	(206)
8.1	单因素方差分析	(206)
8.2	用 SPSS 作单因素方差分析	(210)
8.3*	无重复实验的双因素方差分析	(217)
8.4*	重复实验的双因素方差分析	(221)
8.5*	用 SPSS 作双因素方差分析	(226)
第 9 章	普通相关分析	(240)
9.1	两个随机变量的总体相关与样本相关	(240)
9.2	Spearman 等级相关	(242)
9.3*	偏相关	(242)
9.4	相关系数异于零的显著性检验	(244)
9.5	SPSS 对普通相关分析的处理	(245)
9.6**	品质相关与 SPSS 的处理	(254)
第 10 章	线性回归分析	(261)
10.1	一元线性回归	(261)
10.2*	多元线性回归	(267)
10.3*	逐步回归	(274)
10.4	用 SPSS 处理经典回归问题	(277)
10.5**	多元线性回归的三大基本问题	(288)
10.6**	用 SPSS 处理线性回归的三大基本问题	(298)
第 11 章**	主成分分析与因子分析	(308)
11.1**	主成分分析	(308)
11.2**	因子分析	(315)
11.3**	用 SPSS 进行因子分析	(327)
第 12 章*	聚类分析与判别分析	(336)
12.1*	聚类分析的作法	(336)
12.2*	用 SPSS 做聚类分析	(339)
12.3**	判别分析的作法	(346)
12.4**	Fisher 判别法	(356)

12.5** Bayes 判别法	(358)
12.6** 用 SPSS 作判别分析	(361)
附录 1 常用数理统计表	(373)
附录 2 各章思考题	(381)
主要参考文献	(395)

第 1 章 概率论基础知识

1.1 随机实验、样本空间、概率与条件概率

1.1.1 一些基本概念

1. 随机实验(Random Trial, or Random Experiment)

“抛硬币”就是一个简单的随机实验,可用此例子来理解、记忆如下概念:

- (1)在同一条件下可无限次重复的实验;
- (2)实验结果有多个,且不确定;
- (3)事前不知实验结果(Outcome)。

2. 基本事件(Elementary Event)

一次随机实验的可能结果,称为基本事件或基本随机事件。

若随机实验 E 是“抛 2 次(或先后 2 枚)硬币”,其基本事件就是“正、反”,“正、正”,“反、正”,“反、反”。

3. 样本空间(Sample Space)

所有基本事件所组成的集合,称为样本空间或基本空间。

例如,随机实验 E 是“抛 2 次(或先后 2 枚)硬币”,其样本空间就是集合: {“正、反”,“正、正”,“反、正”,“反、反”}。

4. 随机事件(Random Event)

随机事件简称事件,指基本事件所组成的集合。

例如,随机实验 E 是“抛 2 次(或先后 2 枚)硬币”,事件“两枚出现相同面”,就由两个基本事件组成:“正、正”、“反、反”。事件“至少出现 1 个正面”,就由 3 个基本事件“正、正”、“正、反”、“反、正”组成。

5. 相容事件 (Mutually Inclusive Events) 与不相容事件 (Mutually Exclusive Events)

在随机实验中,不能同时发生或其交集为空集的几个事件,称为不相容事件,反之就称为相容事件。

例如,随机实验 E 为“先后投 2 枚硬币”, A 事件“两枚出现相同面”(由事件“正、正”和事件“反、反”构成),与 B 事件“两枚出现不同的面”(由事件“正、反”和“反、正”构成),就是两个不相容的事件。

C 事件“至少出现 1 枚正面”(由“正、正”、“正、反”、“反、正”组成),与 D 事件“至少出现 1 枚反面”(由“反、反”、“正、反”、“反、正”组成),就是两个相容的事件。

若形象地把事件看成平面上的点集,那么,若 A 与 B 没有共同的点,则 A 与 B 就是不相容事件;若 A 与 B 有共同的点,则 A 与 B 就是相容事件。

6. 概率 (Probability)

用通俗的语言说,概率指在随机实验中,对事件出现的可能性大小的一种严格的度量。所谓严格,是指从无限次重复的角度看,度量结果具有惟一性。

随机实验 E “抛 1 枚均匀的骰子”,做 60 次实验,1 点(俗称大点)朝上的次数,可能是 9、10 或 11,那么 1 点朝上的频率分别是 $9/60, 1/6, 11/60$ 。所以,频率不是随机实验中事件出现的可能性的严格度量,不是概率。但是,随着实验的次数的增加,频率约等于概率。所以,可以通俗地将频率理解为“概率的模糊的影子”。

概率的定义:设 E 是随机实验, S 是其样本空间,给 E 的每一个事件 A 赋一个实数 $P(A)$,若 $P(A)$ 满足如下条件(Postulate),就称为 A 的概率:

(1) 对每一个事件 A , 都有 $0 \leq P(A) \leq 1$;

(2) $P(S) = 1$;

(3) 对于两两互不相容的事件 $A_k (k = 1, 2, \dots)$, 有: $A_k (k = 1, 2, \dots)$ 的并集的概率, 等于各个 A_k 的概率之和:

$$P\left\{\bigcup_k A_k\right\} = P\{A_1 \cup A_2 \cup \dots\} = P(A_1) + P(A_2) + \dots = \sum_k P(A_k)$$

7. 概率运算的主要性质 (Properties of Probability)

(1) 设 $\bar{A}$ 是 A 的对立事件(若把事件看成平面上的点集, 则 $\bar{A}$ 是样本空间 S 中、除 A 以外的所有的点), 则

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

当 $\bar{A}$ 为样本空间 S 时, A 为空集 $\emptyset$, 从而可以得出空集 $\emptyset$ 的概率为 0:

$$P(A) = 1 - P(\bar{A}) = 1 - P(S) = 0$$

(2) 对任意两个事件 A 与 B , 有

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

其中, AB 是 A 与 B 的交集 $A \cap B$ 的缩写。

关于交集概率 $P\{A \cap B\}$ 的计算公式, 我们将在“概率的乘法定理”中给出。

为了便于理解上式, 可参见图 1.1.1。

若 A 与 B 的交集 $A \cap B$ 为空集(即 A 与 B 不相容或不相交), 则 A 与 B 的并集 $A \cup B$ 的概率, 就是 A 、 B 两个事件的概率之和:

$$P\{A \cup B\} = P(A) + P(B)$$

(3) 若事件 $A \subset B$, 即 A 中的点都包含于 B 中, 都是 B 中的点, 则

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

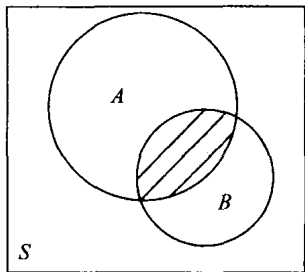


图 1.1.1 事件并集的概率计算

8. 等概率随机实验(Equally Likely Outcomes)

若一个随机实验的基本事件的个数有限, 且基本事件出现的概率相等, 则该随机实验称为等概率随机实验或等可能概型。抛均匀的硬币、抛均匀的骰子, 都是等概率随机实验的例子。

在等概率随机实验中, 事件 A 的概率计算公式是

$$P(A) = A \text{ 包含的基本事件个数} / \text{该实验中基本事件的总个数}$$

例如, 随机实验 E “先后抛 2 枚均匀的硬币”, 共有 4 个基本事件: “正、正”、“正、反”、“反、正”、“反、反”。若我们要考察的事件 A 是“至少有 1 枚正面朝上”, 那么, 事件 A 包括 3 个基本事件“正、正”、“正、反”、“反、正”, 所以 $P(A) = 3/4$ 。

1.1.2 条件概率与概率乘法定理

1. 条件概率(Conditional Probability)

我们用以下实例来介绍条件概率:

例 1.1.1 一个包装箱里有 6 个产品。假设其中 4 个为一级品, 2 个为二级品。若随机实验 E 是“从该包装箱中抽取 1 个产品”, 那么, 显然每次抽取到二级品的概率是 $1/3$ 。

设事件 A 是“第一次抽取并抽到二级品”, 事件 B 是“第二次抽取并抽到二级品”, 那么, 在事件 A 发生的条件下, 再从剩余的 5 个产品中抽出 1 个产品, 事件 B “第二次抽到二级品”发生的概率就是 $1/5$ 。

我们称这样的概率为“事件 A 发生条件下, 事件 B 发生的概率”, 简称为“事

件 B 的条件概率”, 记为 $P\{B|A\}$ 。

在例 1.1.1 中, $P\{B|A\} = 1/5$ 。为对比条件概率与非条件概率的区别, 我们看 $P(B)$ 等于多少。首先, 事件 B 可以表达为不相容的 2 个事件 AB 与 $\overline{A}B$ 之和, 即 $B = AB + \overline{A}B$ (若将其看成 S 平面上的点集, 相当于用 A 与 $\overline{A}$ 把 B 分成不相交的两块), 于是, 由不相交事件的概率的性质, 有

$$P(B) = P(AB) + P(\overline{A}B) = \frac{2}{6} \times \frac{1}{5} + \frac{4}{6} \times \frac{1}{5} = \frac{1}{3}$$

其实, 对于原来的随机实验 (也是原来的样本空间) 而言, 很显然, 无论是事件 A (第一次抽取且抽到二级品), 还是事件 B (第二次抽取且抽到二级品), 出现的概率都是 $\frac{1}{3}$, 即 $P(A) = P(B) = \frac{1}{3}$ 。

由上例, 不难理解如下公式:

$$P\{B|A\} = \frac{P(AB)}{P(A)} = \frac{\frac{1}{15}}{\frac{1}{3}} = \frac{1}{5}, \quad P(A) > 0$$

一般都把这个式子, 作为条件概率的定义式。

条件概率的定义

对样本空间 S 中的两个事件 A, B , 若 $P(A) > 0$, 则条件概率 $P\{B|A\}$ 由如下定义式给出:

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

2. 概率的乘法公式 (定理) (Multiplication Theorem)

由条件概率定义式, 易得:

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

这就是所谓概率的乘法公式或乘法定理。

概率的乘法公式

对样本空间中任意两个事件 A, B , 有

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

条件概率的本质是, 事件 A 的出现, 改变了产生事件 B 的范围条件, 即改变了样本空间, 也就是改变了随机实验 (注意: 产品的个数不是样本空间)。事件 A, B 之间也不一定要有特定的时间先后的关系。