多元统计分析
——以SPSS为计算工具

DUOYUAN TONGJI FENXI

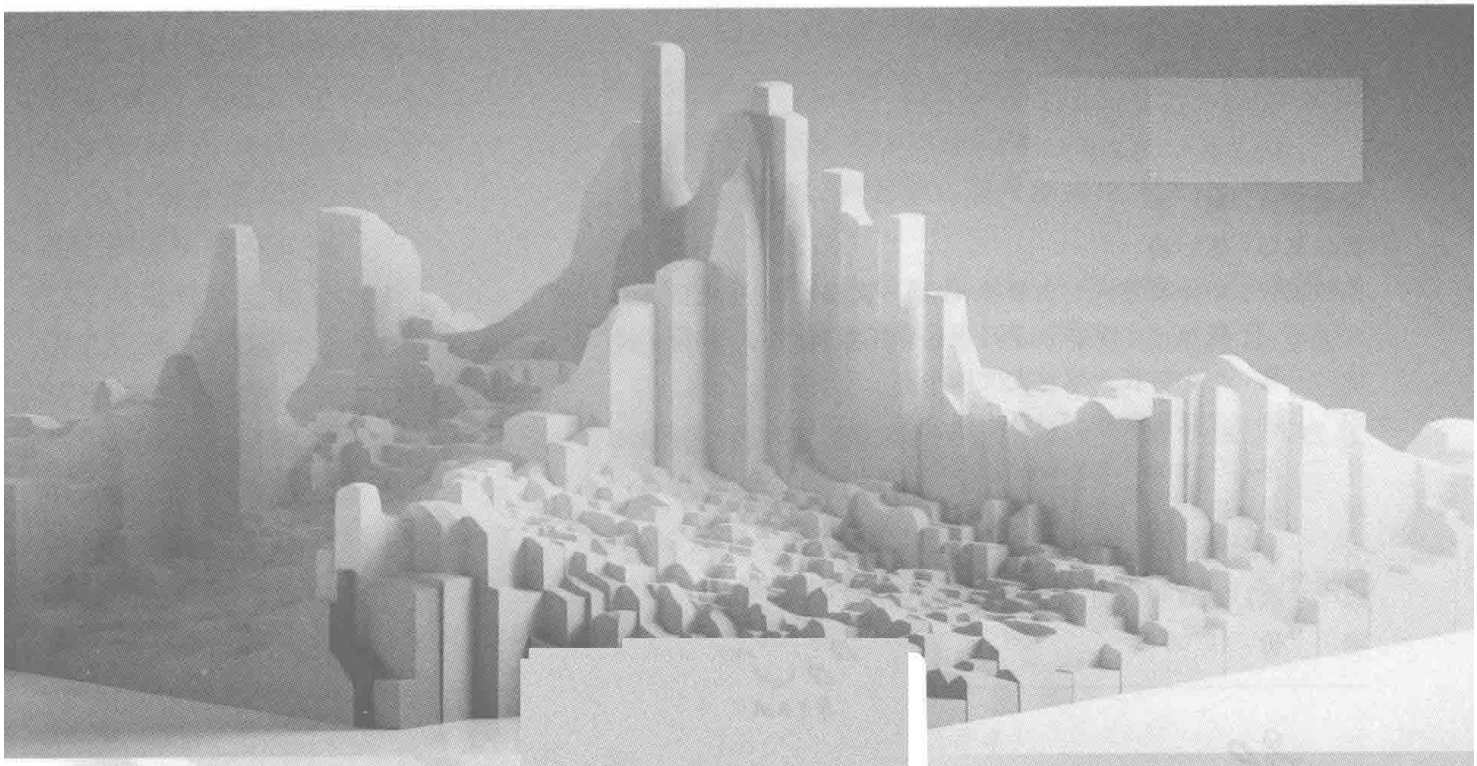
何国民　周峰利　编著

华中科技大学出版社
http://press.hust.edu.cn

多元统计分析
——以SPSS为计算工具

DUOYUAN TONGJI FENXI

何国民　周峰利　编著



华中科技大学出版社
http://press.hust.edu.cn
中国·武汉

内容简介

本书以SPSS(统计产品与服务解决方案)软件作为计算工具,将多元统计方法的基本理论、SPSS操作、多元统计方法应用案例有机结合,以多元统计方法应用为主线,以通俗易懂的语言对多元统计方法的基本理论进行深入浅出的介绍,以"操作示意图"的方式介绍各种多元统计方法的SPSS操作过程。通过案例的形式介绍多元统计方法的应用,可极大地降低读者学习多元统计方法的难度,使读者能够较轻松地学习与掌握各种多元统计方法的应用。本书克服了SPSS手册类教材只注重SPSS操作说明而忽略多元统计方法原理讲解的不足,同时弥补了多元统计方法专业教材只注重多元统计方法理论的论述而缺乏数据处理工具介绍的缺憾,是一本特色鲜明、具有广泛使用价值的教材。

本书包含以下内容:绪论,第一章"概率论基础",第二章"推断统计",第三章"相关与回归分析",第四章"聚类分析",第五章"判别分析",第六章"因子分析",附录A"文献阅读",附录B"常用统计表"。

图书在版编目(CIP)数据

多元统计分析:以SPSS为计算工具/何国民,周峰利编著. —武汉:华中科技大学出版社,2023.5
ISBN 978-7-5680-9331-6

Ⅰ. ①多… Ⅱ. ①何… ②周… Ⅲ. ①多元分析-统计分析-高等学校-教材 Ⅳ. ①O212.4

中国国家版本馆CIP数据核字(2023)第058480号

多元统计分析——以SPSS为计算工具

Duoyuan Tongji Fenxi——yi SPSS wei Jisuan Gongju

何国民 周峰利 编著

策划编辑:曾 光
责任编辑:白 慧
封面设计:孢 子
责任监印:朱 玢

出版发行:华中科技大学出版社(中国·武汉) 电话:(027)81321913
武汉市东湖新技术开发区华工科技园 邮编:430223

录 排:华中科技大学惠友文印中心
印 刷:武汉科源印刷设计有限公司
开 本:787mm×1092mm 1/16
印 张:11.5
字 数:294千字
版 次:2023年5月第1版第1次印刷
定 价:42.00元

前言

多元统计分析是以概率论为基础，应用线性代数的基本原理和方法，结合计算机软件对信息和资料进行收集、整理和分析的一门学科，是统计学中内容十分丰富、应用范围极其广泛的一个分支。

20世纪50年代以来，计算机技术的发展与普及，特别是各种统计分析软件的推出，有力地解决了多元统计方法在实际应用中“统计计算难”这一问题，多元统计方法得到了极为广泛的应用，当前已经成为社会科学领域不可或缺的强有力的研究工具。

本书以SPSS(统计产品与服务解决方案)软件作为计算工具，将多元统计方法的基本理论、SPSS操作、多元统计方法应用案例有机结合，以多元统计方法应用为主线，以通俗易懂的语言对多元统计方法的基本理论进行深入浅出的介绍，以“操作示意图”的方式介绍各种多元统计方法的SPSS操作过程。通过案例的形式介绍多元统计方法的应用，可极大地减轻读者学习多元统计方法的难度，使读者能够较轻松地学习与掌握各种多元统计方法的应用。

本书在编写过程中力求做到以下几点：

(1)对各种多元统计方法所涉及的基本概念、原理的介绍力求简单明了、深入浅出，并尽量少用统计专用符号、公式，目的是减少读者的阅读困难。

(2)对于SPSS的介绍，只是将其作为一个“超级计算器”，以“操作示意图”的方式介绍如何利用SPSS完成各种多元统计方法的计算任务，读者可以以“看图识字”的简单方式非常快速地学会用SPSS完成各种多元统计方法的计算工作，得到所需的计算结果。

(3)每种多元统计方法都配有具体应用案例，通过案例的形式学习多元统计方法，可极大地减轻读者学习多元统计方法的难度。

本书可作为高等院校管理、经济等专业本科生、研究生学习多元统计学的教材，以及统计、信息管理、市场调研、大数据分析等行业的从业人员的参考用书，还可作为多元统计分析与SPSS的自学读本。

本书由何国民、周峰利共同编写，其中绪论、第一章、第二章由何国民编写，第三章、第四章、第五章、第六章由周峰利编写。

由于作者水平有限，书中难免有不妥与错误之处，敬请统计界同仁及读者批评指正。

本书在编写中参阅了一些教材、著作和论文，在此向原作者们致以诚挚的谢意。本书的出版得到了武汉体育学院研究生院的资助，在此表示感谢！

目录

绪　论

一、统计学的产生与发展

Statistics(统计学)一词来源于法语 status(状态),大概出现于 17 世纪。1690 年英国经济学家威廉·配第《政治算术》一书的问世,标志着统计学的诞生。统计学的发展过程大致可划分为以下三个阶段。

第一阶段:统计学初创阶段(17 世纪中叶到 19 世纪末)。

国势学派、政治算术学派对统计学的产生的贡献。

17 世纪中叶,国势学派诞生于德国,代表人物是康令(H. Conring,1606—1681),该学派主张用记述的方法记录国家的重大事项,几乎不用数字资料。到 18 世纪,德国人阿亨瓦尔(G. Achenwall,1719—1772)首次在大学开讲"国势学"课程,将"统计"定义为"把国家的显著事项全部记录下来的学科",并称此学科为 statistik(德文:统计学)。国势学派对社会经济现象进行研究时,只注重文字分析,完全不用数据,对图形表格、数字式子十分蔑视,因而被称为有名无实的统计学。

17 世纪中叶,政治算术学派在英国兴起,代表人物是威廉·配第(W. Petty,1623—1687),他在代表作《政治算术》一书(图 0-1)中第一次用计量和比较的方法,将英国的国力与法国、意大利、荷兰等国的国力进行比较研究,以论证英国的国际地位。由于最早提出了一套较为系统的用于对社会经济现象进行数量性描述和分析比较的方法,威廉·配第被称为"统计学的创始人"。政治算术学派对社会经济现象进行分析时,注重数量分析,这与排斥数量只讲观念的国势学派不同。但由于在政治算术学派的所有著述中并没有提到"统计学",因此这个学派被称为有实无名的统计学。

图 0-1　威廉·配第的著作

第二阶段:推断统计方法体系基本确定阶段。

19 世纪中叶,以比利时统计学家阿道夫·凯特勒(A. Quetelet,1796—1874)为代表的数理统计学派将概率论等数学方法引入统计分析中,完成了统计学与概率论的结合,开创了推断统计的先河。在这个阶段,估计理论、样本分布理论、方差分析理论、假设检验理论等都获得了重大进展。

第三阶段:多元统计方法应用全面发展阶段。

从 20 世纪 50 年代起,受计算机科学、信息论等现代科学技术的影响,统计学的应用领域不断扩展,新的研究分支不断增加,如多元统计分析、探索性数据分析、数据挖掘技术、现代时间序列分析方法等。据有关学者统计,多元统计方法的应用是以指数式加速发展的。

二、统计学方法体系

统计学是一门收集、整理、分析数据的科学，要解决的根本问题是：从总体随机抽取样本，再由样本推断总体数量特征，如图 0-2 所示。

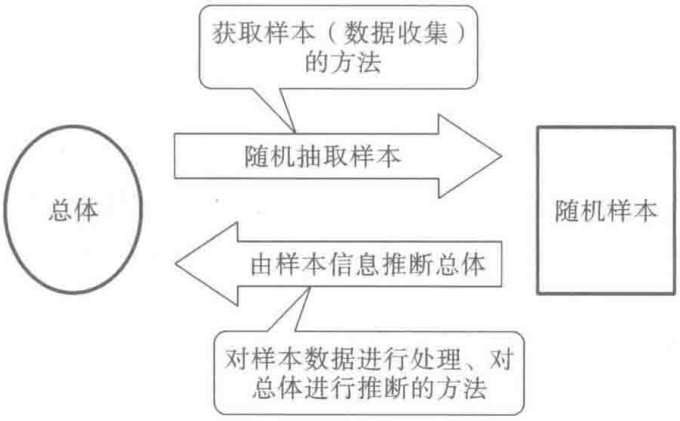


图 0-2　统计学要解决的根本问题

围绕“用样本推断总体”这一统计学要解决的根本问题，人们提出了各种统计方法，包括数据收集、描述统计、推断统计、多元统计四大类方法，如图 0-3 所示。

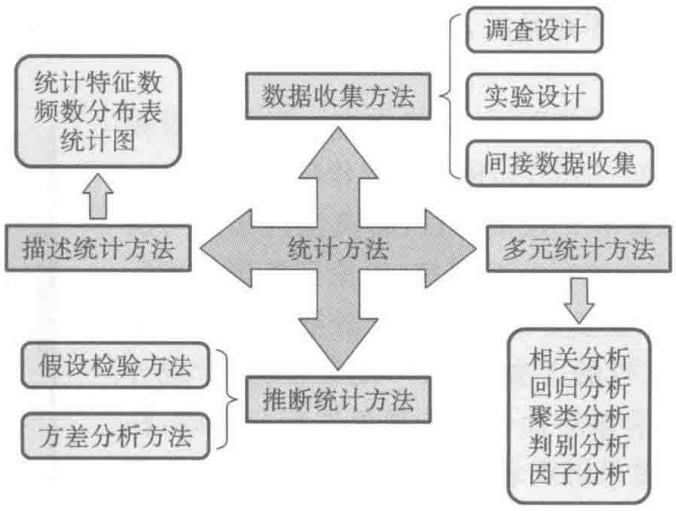


图 0-3　统计学方法体系

①数据收集——围绕研究问题，科学合理地收集样本数据。

主要方法：调查设计、实验设计、间接数据收集。

②描述统计——对抽样数据进行处理，得出一些特殊的图形、表格、数字，用它们来描述样本数据分布特点。

主要方法：数字特征描述、图表描述。

③推断统计——根据样本数据对总体的数量关系做出某种推断。

主要方法：参数估计、假设检验、方差分析。

④多元统计——分析多个变量间的数量关系。

主要方法：相关分析、回归分析、因子分析、判别分析、聚类分析。

三、统计学学科体系

统计学经过三百多年的发展，已经成长为一棵枝繁叶茂的参天大树，目前已经形成了由若干分支组成的庞大学科体系，图 0-4 是一些常见的统计学教材。

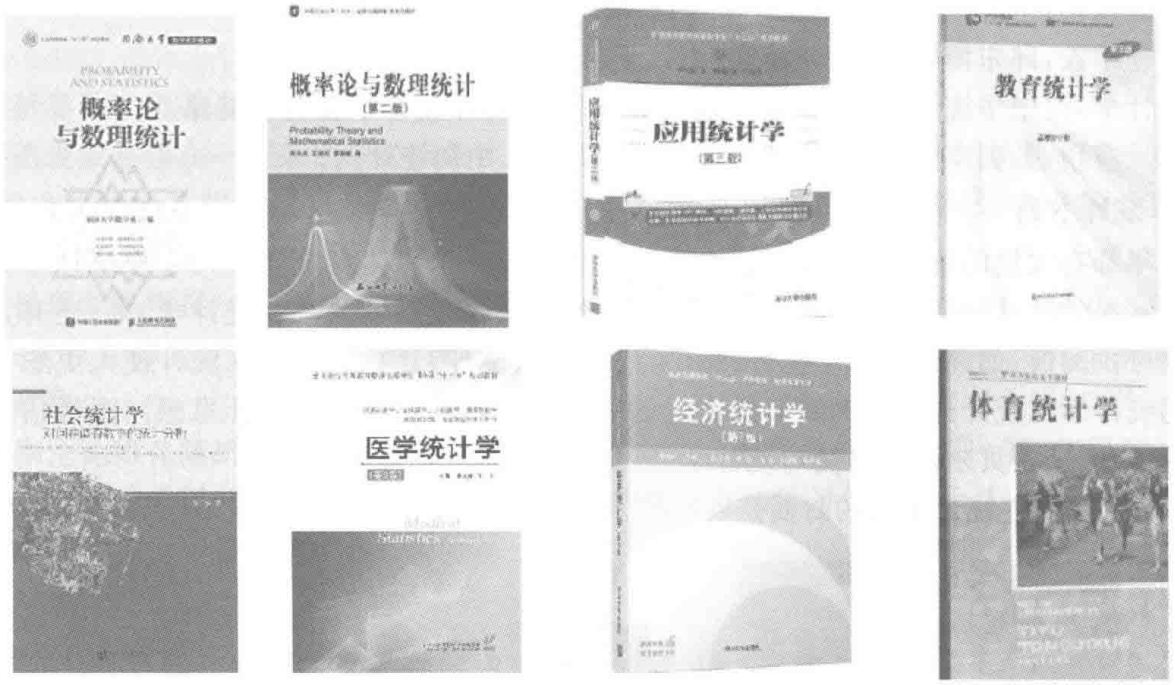


图 0-4　常见的统计学教材

根据研究内容的不同，我们大体上可将统计学分为数理统计学、应用统计学两大分支。

数理统计学：主要阐述统计方法的数学原理，其理论基础是概率论。

应用统计学：将统计方法应用于各个学科领域而形成的各学科统计学的总称。

统计学的学科体系如图 0-5 所示。

四、统计分析的过程

用统计方法进行科学研究的步骤可用图 0-6 表示。

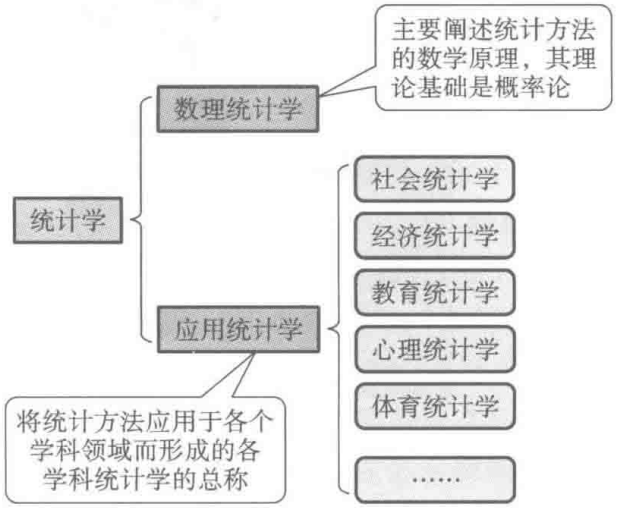


图 0-5　统计学学科体系

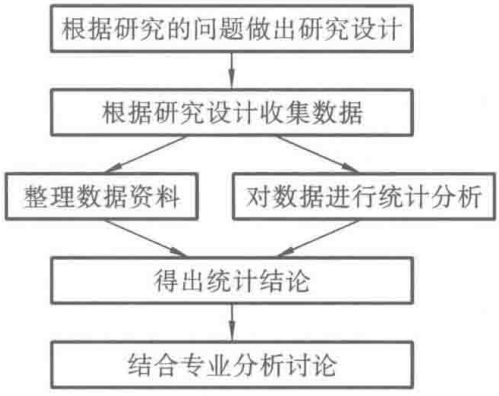


图 0-6　统计分析过程

五、统计学相关网站与书籍

以下是统计学相关网络信息与书籍[①]。

1. 网络信息类

在线网站：中国大学 MOOC(慕课)国家精品课程。

统计学人：邱东博客、郑来轶博客、小蚊子乐园。

统计学习：中华统计学习网、统计师考试网、中国统计教育学会教学资源。

统计应用：数据草堂、统计之都、统计菁英工作室、中国统计网。

2. 统计书籍

每年都有大量的统计学教材新鲜上市，如《经济统计学》《体育统计学》《心理统计学》《传播统计学》《医学统计学》《生物统计学》《旅游统计学》《教育统计学》《社会统计学》等；还有大量统计读物可供阅读，如《看漫画，学统计》《爱上统计学》《统计，让数字说话》《统计使人更聪明》《生活中的统计学》《统计连着你和我》《漫游数据王国》《统计学的世界》《统计思想》《漫话信息时代的统计学：兼话诺贝尔经济学奖与统计学》等。

图 0-7 是一些统计书籍的封面。

图 0-7　部分统计书籍的封面

① 邓力.《统计学原理》.北京：清华大学出版社，2012.

第一章　概率论基础

自然界和人类社会中存在两类现象：必然现象与随机现象。

必然现象是指在一定条件下必然发生的现象。如在标准大气压下，水加热到100 ℃必然会沸腾；抛出的物体其初速度只要小于第一宇宙速度，必然会落回地面等，这些现象都是必然现象。

随机现象是指在一定条件下有时发生有时不发生的现象。如往地面掷一枚硬币，“出现正面”这一现象有时发生，有时不发生；投篮中“投中”这一现象有时发生，有时不发生，这些现象都是随机现象。

概率论是研究随机现象的数量规律的数学分支之一，本身具有丰富的内容，并得到了广泛的应用。而我们学习概率论的主要目的是为以后的统计方法学习打下理论基础，本章所涉及的内容也就围绕着这一目的展开。

第一节　随机事件及其概率

一、随机事件

对随机现象进行观察，会观察到不同的结果，如观察掷硬币这一随机现象就可能看到“出现正面”或“出现反面”这两种不同的结果。“出现正面”是掷硬币这一随机现象的一种观察结果，我们称之为随机事件，同样，“出现反面”也是随机事件。

随机事件：对随机现象进行观察，其观察结果叫随机事件，简称事件，用大写英文字母 A、B、C 等表示。

作为随机事件的特例，若某事件在每次试验中总是发生，则称该事件为必然事件，一般用字母 Ω 表示；反之，若某事件在每次试验中都不发生，则称该事件为不可能事件，一般用字母 Φ 表示。

对于一个随机事件（除必然事件和不可能事件以外）来说，它在一次试验中可能发生，也可能不发生，但通过大量观察可以发现，随机事件发生的可能性大小具有一定的规律性，这种规律性称为统计规律性。如掷硬币，虽然掷1次无法确定是“正面向上”还是“反面向上”，但通过大量的试验可以发现，“正面向上”与“反面向上”的次数比较接近，而且随着试验次数的增加，接近程度愈来愈高，各占50%。历史上曾有不少数学家做过这类试验，如表1-1所示。

对于随机事件，我们常常希望知道它们在一次试验中发生的可能性究竟有多大。例如，为了确定水坝的高度，就要知道在造水坝地段河流每年发生最大洪水达到某一高度这一事件发生的可能性大小；再如，为了评价一位射击运动员水平的高低，我们就需要知道该运动员命中各环的可能性大小。

表 1-1　历史上数学家们抛掷硬币的试验数据

实验者	掷硬币次数	出现正面次数	出现正面频率
蒲丰	4040	2048	0.5069
皮尔逊	12000	6019	0.5016
皮尔逊	24000	12012	0.5005

我们希望找到一个合适的数来表示随机事件在一次试验中发生的可能性大小，为此，首先引入描述事件发生频繁程度的量——频率。

二、频率

在相同的条件下，进行了 n 次试验，在这 n 次试验中事件 A 出现了 m 次，则称比值 m/n 为事件 A 的频率，记为 $F(A)=m/n$。

显然，对于任一事件 A 都有

$$0 \leqslant F(A) \leqslant 1$$

随机事件是否发生事先是不能确定的，但经过多次观察，随机事件发生的频率是有一定规律的。如民兵射击 100 次射中 95 次，我们说其命中率为 95%；抽查某工厂的产品 100 件，结果有 98 件合格品，则该工厂的产品合格率为 98%。

经验表明，随着试验次数的增多，随机事件频率的波动会越来越小，且会在一个固定的常数附近做微小的波动。以掷硬币为例，记“正面向上”为随机事件 A，抛掷总次数为 n，出现“正面向上”的次数为 m，比值 $F=m/n$ 为事件 A 的频率，所得结果如表 1-2 所示：

表 1-2　抛掷硬币试验的数据

$n=5$		$n=50$		$n=500$	
m	F	m	F	m	F
2	0.4	22	0.44	251	0.502
3	0.6	27	0.54	249	0.498
1	0.2	21	0.42	256	0.512
4	0.8	26	0.52	245	0.490
1	0.2	24	0.48	251	0.502

从表 1-2 可以看到，当 $n=5$ 时，事件 A 发生的频率波动相当剧烈，频率变动无规律可循；当 $n=50$ 时，频率波动明显减少，大致围绕 0.5 上下波动；当 $n=500$ 时，频率波动已相当小，可明显看出其波动中心为 0.5。

不难想象，当抛掷次数再增加时，频率会更加稳定在 0.5 附近，这些试验的结果是很有启发性的，它们表明虽然事件 A 在一次试验中可能发生也可能不发生，但在大量重复试验中，它出现的频率趋于稳定，而且试验次数越多，频率越接近某一常数（在上述掷硬币的实例中此常数为 0.5）。频率的这种随着试验次数增多而趋于稳定的情况称为频率的稳定性。

频率的稳定性说明随机事件发生可能性的大小是随机事件本身固有的一种客观属性，而不是人的主观意志可以随意改变的，因此可以对它进行度量，我们可以用一个数来描述随机事件在一次试验中发生的可能性大小，该数就是概率。

三、概率

直观地说，概率是描述随机事件发生可能性大小的一个度量。根据频率的稳定性，下面给出概率的定义，一般称其为概率的统计定义。

随机事件的概率：在 n 次重复试验中随机事件 A 发生的次数记为 m，当 n 很大时，频率 m/n 会稳定地在某一数值 p 的附近摆动，而且随着试验次数 n 的增加，其摆动的幅度越来越小，称 p 为随机事件 A 的概率，记为

$$P(A) = p$$

例如，在投硬币的试验中，“出现正面”这一随机事件发生的频率在 0.5 附近摆动，且随着试验次数的增多，摆动的幅度会越来越小，因此，可以认为“出现正面”这一随机事件的概率为 0.5。

频率与概率之间的关系是非常密切的，也正因为如此，它们具有一些相同的性质。

对于频率，由于事件发生的次数总满足 $0 \leqslant m \leqslant n$，因此有

$$0 \leqslant m/n \leqslant 1$$

而对于不可能事件 Φ 必有 $m=0$，对于必然事件 Ω 一定有 $m=n$，可知它们的频率为

$$F(\Omega) = 1, F(\Phi) = 0$$

对于概率类似地有

$$0 \leqslant P(A) \leqslant 1$$

以及

$$P(\Omega) = 1, P(\Phi) = 0$$

四、小概率事件原则

一般若 $P(A) \leqslant 0.05$，则称事件 A 为小概率事件。小概率事件在一次试验中几乎不可能发生，这一原则称为小概率事件原则。

小概率事件原则是统计学中由样本推断总体的重要原则，在以后的学习中将会多次用到此原则，图 1-1 所示的摸球模型就是典型的由样本推断总体的问题。

图 1-1　摸球模型

试根据所取 5 球的结果推断：盒中最多只有 10 个白球这一说法能否成立？

上述问题的解决就需要用到小概率事件原则，该问题将在第二章给予回答。

第二节　随机变量及其概率分布

一、随机变量的定义

随机事件可能是与数量有关的，如某学生的一次跳远成绩、某人的身高等随机事件都可用数表示；随机事件也可能是与数量无关的，如投篮中的“投中”与“没有投中”，掷硬币中的“出现正面”与“出现反面”等随机事件都与数量无直接关系，不能直接用数来表示。

为了能更好地对随机现象进行研究，应把随机事件数量化，这就需引入随机变量的概念。

例 1-1　考查掷硬币这一试验，它有两种可能结果：“出现正面”或“出现反面”。为了便于研究，我们用一个数来代表试验的一个结果。例如，用数字“0”代表“出现反面”，用数字“1”代表“出现正面”，这样，当我们讨论试验结果时，就可以简单地说随机试验的结果是 1 或是 0。这种将随机事件数量化的方法，实际上就相当于引入一个变量 X，变量 X 的取值与试验的结果有关，当试验结果为“出现正面”时 $X=1$，当试验结果为“出现反面”时 $X=0$，这里变量 X 随着试验的不同结果而取不同的值，由于试验的结果是随机的，因而 X 的取值也是随机的，因此称 X 为随机变量。

例 1-2　考虑测试学生跳远成绩这一试验，试验的结果（学生跳远成绩）本身就是一个数，我们以 Y 记学生的跳远成绩（以厘米计），某学生跳了 470 cm，则 $Y=470$，这里变量 Y 的取值由试验的结果确定，Y 随着试验的不同结果而取不同的值，它也是随机变量。

随机变量的数学定义是很严格的，下面只给出其描述性定义。

随机变量：当我们用一个变量的取值来表示随机试验的结果时，该变量的取值随着试验的不同结果而不同，也就是说变量的取值是随机的，称此变量为随机变量。随机变量一般用 X、Y、Z 等大写英文字母表示。

引入随机变量后，我们就可用随机变量的取值来表示随机事件。

比如在例 1-1 中，$X=0$ 表示“掷硬币出现反面”这一随机事件；$X=1$ 表示“掷硬币出现正面”这一随机事件。

又如在例 1-2 中，$Y=450$ 表示“跳远成绩为 450 cm”这一随机事件；$420<Y<500$ 表示“跳远成绩在 420 cm 到 500 cm 之间”这一随机事件。

如上所述，随机变量随着试验的不同结果而取不同的值，因此，在试验之前只能知道随机变量可能取值的范围，而不能确切地知道它取什么值。此外，随机变量取各个值有一定的概率，这一性质显示了随机变量与普通的变量有着本质的区别。

二、随机变量的分类

随机变量按其取值情况可分为离散型随机变量、连续型随机变量两种类型。

离散型随机变量：如果随机变量所有可能取到的值是有限多个或可列多个，这种随机变量称为离散型随机变量。

比如，在例 1-1 中，用 X 的取值表示掷硬币这一试验的结果，X 的所有可能取值为 0 或 1，是有限多个，因此 X 为离散型随机变量。

又如，某一不透明的盒中装有 10 个外形一样的球，其中 5 个黑球，5 个白球，现从中不放回地任取 4 球，用 Y 表示所取 4 球中白球的个数，则 Y 的所有可能取值为 0、1、2、3、4，Y 的取值为有限多个，因此 Y 是离散型随机变量。

再如，某一不透明的盒中装有 100 个外形一样的球，其中 99 个黑球，1 个白球，现从中任取 1 球，若是黑球则放回盒中，并再取 1 球；若取出的是白球则停止取球。用 Z 表示第 1 次取到白球时的取球次数，则 Z 的所有可能取值为 $1,2,3,\cdots,n,\cdots$，此时 Z 的取值即为可列多个，此处 Z 是离散型随机变量。

连续型随机变量：如果随机变量取值不只是可列个，而是可取区间$[a,b]$或$(-\infty,+\infty)$上的一切值，这种随机变量称为连续型随机变量[①]。

比如在例 1-2 中，用 Y 记学生的跳远成绩，Y 的取值范围为一区间（比如$[0,10]$）上的一切可能值，因此 Y 为连续型随机变量。

三、概率分布的概念

对于一个随机变量，我们不仅需要知道它可以取哪些值，更重要的是要知道该随机变量取这些值的概率大小。如要了解一名运动员的投篮水平，可以用随机变量 $X=1$ 表示投中，$X=0$ 表示没有投中，但只知道 X 的取值还不能描述该运动员的投篮水平，我们还必须知道 $X=0$、$X=1$ 的概率，这样才能清楚、全面地描述该运动员的投篮水平。

概率分布：随机变量的取值及取值的概率称为随机变量的概率分布。

例 1-3　随机抛掷 2 枚均匀的骰子，用 X 表示 2 枚骰子落地后朝上一面的点数之和，则 X 的所有可能取值及取各值的概率如下表所示：

X 的取值	2	3	4	5	6	7	8	9	10	11	12
X 取值的概率	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

上表就清楚地说明了随机变量 X 可以取哪些值及取这些值的概率，我们称上表为 X 的概率分布。

例 1-4　用 $Y=n$ 表示射击中命中 n 环（$n=0$、6、7、8、9、10），则某运动员在一段时间内的射击水平可用下表描述：

Y 的取值	0	6	7	8	9	10
Y 取值的概率	0.01	0.14	0.3	0.35	0.15	0.05

上表就清楚地说明了随机变量 Y 可以取哪些值及取这些值的概率，我们称上表为 Y 的概率分布。

根据随机变量的概率分布，可容易地得到随机变量在某一范围内取值的概率。

---

① 概率论中连续型随机变量的定义：对于随机变量 X，如果存在非负可积函数 $f(x)(-\infty<x<+\infty)$，对于任意实数 $a,b(a<b)$ 都有 $P(a\leqslant x\leqslant b)\int_a^b f(x)dx$，则称 X 为连续型随机变量。这一定义告诉我们，对于可取区间$[a,b]$或$(-\infty,+\infty)$上一切值的随机变量，并不一定是连续型随机变量，只有满足上述条件的随机变量才算连续型随机变量，不满足上述条件的则是非连续型随机变量。

严格来说，随机变量可分为三种类型：离散型、连续型、非连续型。不过在实践中，由于非连续型随机变量的概率分布无法描述，无法对其进行研究，所以我们以后只对离散型、连续型两类随机变量进行研究。

例如，根据例 1-3 中 X 的概率分布，可计算出 X 取偶数的概率为：

$$P(X=2、4、6、8、10、12)=\frac{1}{36}+\frac{3}{36}+\frac{5}{36}+\frac{5}{36}+\frac{3}{36}+\frac{1}{36}=0.5$$

再如，根据例 1-4 中 Y 的概率分布，可计算出 $6\leqslant Y\leqslant 8$ 的概率（即该运动员命中 6 至 8 环的概率）为：

$$P(6\leqslant Y\leqslant 8)=0.14+0.3+0.35=0.79$$

四、离散型随机变量的概率分布

对于离散型随机变量，由于它的所有可能取值为有限个或可列个，我们一般用分布列来描述其概率分布。

分布列：设离散型随机变量 X 可能取到的值为 $x_1,x_2,\cdots,x_n,\cdots$，$X$ 取到各个值的概率为 $p_1,p_2,\cdots,p_n,\cdots$，以如下表格表示 X 的取值及取值的概率情况。

X 的取值	x_1	x_2	…	x_n	…
X 取值的概率	p_1	p_2	…	p_n	…

称上述表格为随机变量 X 的概率分布列，简称分布列。

分布列满足性质：①$\sum_{i=1}^{\infty}P_i=1$；②$0\leqslant P_i\leqslant 1$。

例 1-4 中就是用概率分布列来描述某运动员在一段时间内的射击水平。对于离散型随机变量来说，概率分布列是了解它的“窗口”，从中我们可以一目了然地看出随机变量 X 的取值范围和取这些值的概率，从而全面地掌握这一随机变量。

例 1-5 某一不透明的盒中装有 10 个外形一样的球，其中 5 个黑球、5 个白球，现从中任取 3 球，用 Y 表示取到的白球数，求 Y 的概率分布列。

分析：求 Y 的概率分布列，就是求 Y 能取哪些值及取这些值的概率。

【解】由于取出的 3 个球中可能有 0 个白球、1 个白球、2 个白球、3 个白球，因此 Y 的取值范围为 0、1、2、3。

从 10 个球中任取 3 球的总取法数 $n=C_{10}^3=\frac{10\times 9\times 8}{3\times 2\times 1}=120$

其中所取的 3 个球全是黑球（此时 $Y=0$）的取法数 $m=C_5^3=\frac{5\times 4\times 3}{3\times 2\times 1}=10$

所以

$$P(Y=0)=\frac{10}{120}=0.08$$

类似求出

$$P(Y=1)=0.42,P(Y=2)=0.42,P(Y=3)=0.08$$

因此，Y 的概率分布列如下：

Y 的取值	0	1	2	3
Y 取值的概率	0.08	0.42	0.42	0.08

五、连续型随机变量的概率分布

对于连续型随机变量，由于其取值是无限不可列的，显然不能用分布列的形式来描述它的

概率分布。另外，连续型随机变量取某一具体值的概率皆为 0，即当 X 为连续型随机变量时，$P(X=a)=0$，因此，用分布列来描述连续型随机变量的概率分布毫无意义。

对于连续型随机变量，一般用概率密度函数、概率密度曲线来描述其概率分布。下面用一个例子来说明什么是概率密度函数、概率密度曲线。

例 1-6　测得某地区 100 名 12 岁男孩的身高数据，经初步整理，得到各身高段的人数，如表 1-3 所示。

表 1-3　某地区 100 名 12 岁男孩的身高分布状况

身高/cm	人数	身高/cm	人数
(0,130]	5	(145,150]	18
(130,135]	10	(150,155]	14
(135,140]	15	(155,160]	10
(140,145]	20	160 以上	8

用 x 轴表示身高，y 轴表示人数，用条形图描述 100 名 12 岁男孩的身高分布状况，见图 1-2。

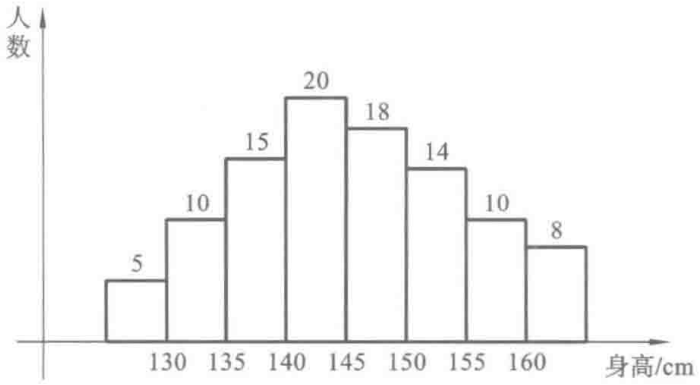


图 1-2　100 名 12 岁男孩的身高分布图

当测量的人数不断增多，统计时各身高段的范围越来越小时，可以想象，身高的频数分布图会趋于一条光滑的曲线，如图 1-3 所示。

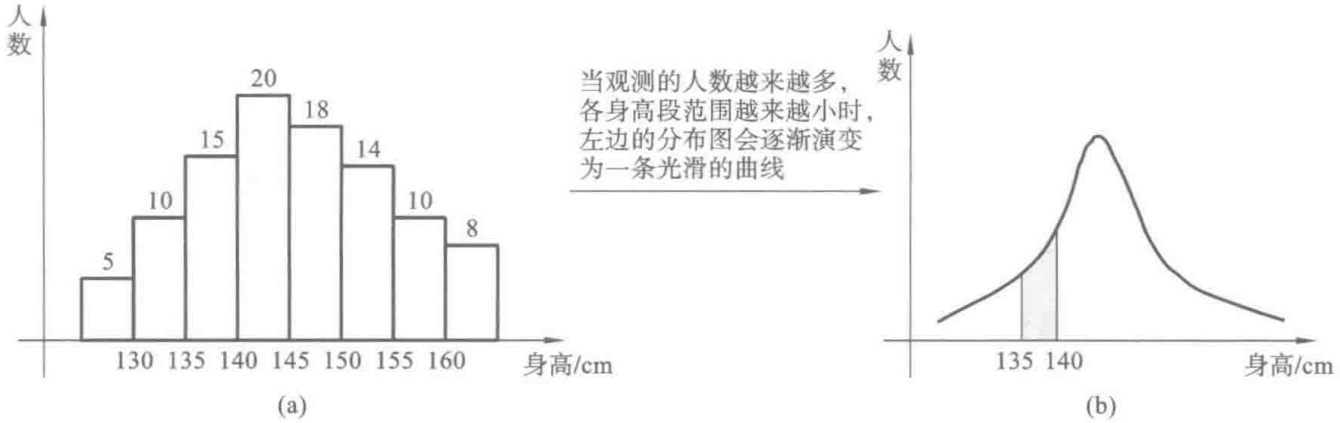


图 1-3　概率密度曲线