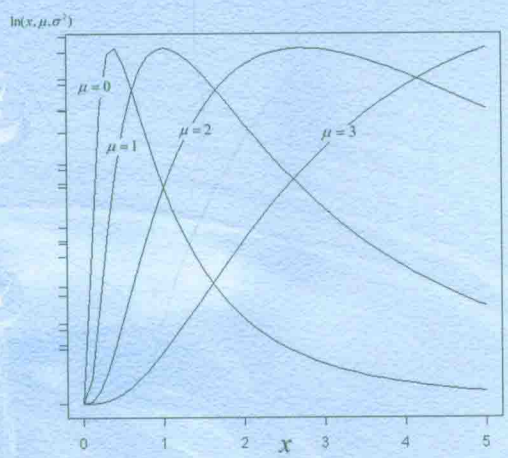
王志良 李淑贞 李立阳 李艳玲 著

水质统计理论及方法

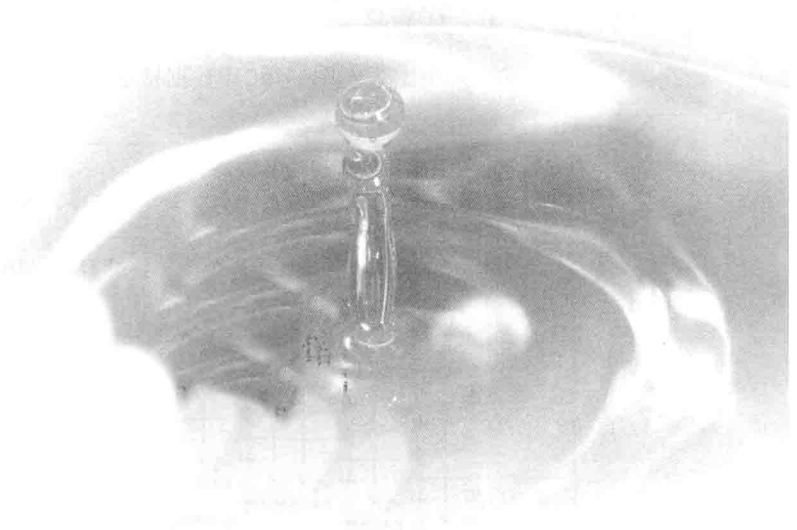
SHUIZHI TONGJI LILUN JI FANGFA



中国水利水电出版社
www.waterpub.com.cn

水质统计理论及方法

王志良 李淑贞 李立阳 李艳玲 著



中国水利水电出版社
www.waterpub.com.cn

内 容 提 要

全书共分8章，主要讲述：①国内外水质统计理论方法的现状；②应用于水质统计的概率统计基础；③水质统计数据的可视化处理；④应用R软件对水质统计数据进行计算和展示。

本书注重概率统计的基本概念，用具体的实例阐述其意义，并辅以数据的可视化加以说明，不再用繁杂的公式证明人们非常熟悉的定理和结论。本书的特色：①与一般的统计类图书比，这是一本研究对象仅限于水质的统计著作；②从本书内容及编排上看，增加了数据可视化及统计假设的内容；③水质分析的大部分内容采用R软件进行计算和展示。

本书有较强的实用性，可供水利系统各大流域机构的水质监测中心和省、直辖市、自治区水利厅水质监测管理部门的工作人员，以及高等院校水环境相关专业的本科生和从事水环境研究的工作者学习参考。

图书在版编目（CIP）数据

水质统计理论及方法 / 王志良等著. -- 北京 : 中国水利水电出版社, 2013.12
ISBN 978-7-5170-1626-7

Ⅰ. ①水… Ⅱ. ①王… Ⅲ. ①水质监测－数理统计 Ⅳ. ①X832

中国版本图书馆CIP数据核字(2013)第319682号

书　　名	**水质统计理论及方法**
作　　者	王志良　李淑贞　李立阳　李艳玲　著
出版发行	中国水利水电出版社 （北京市海淀区玉渊潭南路1号D座　100038） 网址：www.waterpub.com.cn E-mail：sales@waterpub.com.cn 电话：（010）68367658（发行部）
经　　售	北京科水图书销售中心（零售） 电话：（010）88383994、63202643、68545874 全国各地新华书店和相关出版物销售网点
排　　版	中国水利水电出版社微机排版中心
印　　刷	北京瑞斯通印务发展有限公司
规　　格	184mm×260mm　16开本　12.25印张　290千字
版　　次	2013年12月第1版　2013年12月第1次印刷
印　　数	0001—1000册
定　　价	**45.00**元

凡购买我社图书，如有缺页、倒页、脱页的，本社发行部负责调换

前　言

社会的发展与进步对水环境的要求越来越高，国家在水环境监测方面的投入越来越多，水质监测的数据量在不断增加，对监测数据进行统计分析将有利于水环境的保护与管理，水利部门和环境部门在水质监测的相关规范中也要求对水质数据进行统计分析。不过，就作者所知，国内将数理统计的理论方法完整地应用于水质数据分析的著作不多，有的著作中大量介绍概率论与数理统计的理论与方法，选用的例子甚至与“水”没多大关系；有的著作中仅有一至两章介绍水质的趋势分析。鉴于此，作者在近年的教学科研与环境监测工作中，产生了编写一本较为完整的介绍水质统计理论与方法的著作的想法。

在对本书内容进行规划时，我们的想法是，除了介绍必要的数理统计知识，以及如何应用这些基础知识于水质数据分析外，在书中要体现一些统计学最新的理论或方法，比如数据的可视化方法、非参数统计理论方法等。在选择统计分析工具时，我们的目标是选择最新的容易使用的有较好应用前景的计算机软件，R 软件因为其是开源程序，在国外从事统计教学和科研的学者中使用广泛，在一些环境管理或服务的政府部门及相关公司都有应用。考虑到国家正在实施创新战略，一些发达国家经常拿知识产权“说事儿”，我们极力推荐 R 软件在我国水利、环境等部门及相关高校中的科技人员中普及使用，若然，将是国之幸事。

全书内容共分 8 章，其中第 1 章绪论部分重点讲述水质统计分析方法国内外现状。第 2 章与第 4 章介绍数理统计的基本概念、原理及方法，这些内容在一般的概率统计教材或水文统计教材中都会看到，本书的不同之处在于书中给出的例子大多都是黄河中下游监测断面的常见监测指标。第 3 章介绍了水质数据分析中常见的一元变量和二元变量的图形分析方法，这部分内容是本书的特色之一。水质统计分析相关中文图书中，一般很少见到将数据的图示化分析单独列为一章的。在章节序号的安排时，这一章在书中的顺序斟酌再三，将其置于数理统计的基础与假设检验两章之间。第 5 章趋势分析是环境监测规范中要求做的“规定动作”，是本书的重点内容。第 6 章、第 7 章讲述变量之间的关系。第 6 章除了讲常见的皮尔逊相关系数外，还介绍了秩相关的内容。

线性回归模型在水质统计分析中是常见的模型，第7章较为完整地讨论了水质统计中线性回归模型的建模及求解的原理及过程。第8章是作者重点推介的内容，较为系统地介绍了R软件的前生今世，介绍了R软件在水质统计分析中的使用方法，分析的数据对象是来自黄河中下游9个断面8个指标30年的数据集。

全书的写作由华北水利水电大学王志良、李艳玲和黄河水利委员会李淑贞、李立阳共同完成，由王志良统稿。感谢研究生梁洪运和周其龙，他们在本书的写作初期整理了很多相关资料，书中的例题由他二人进行了验算。感谢也送给其他为本书的出版做过工作的人们。

本书的写作参考了很多文献资料，其中的大部分文献都在书尾参考文献中列出。也许由于我们的疏忽，有些文献，特别是从网络上看到的文献，没有在本书列出。如果您能直率地告诉我们，我们将不胜感激。缘于作者水平有限，书中可能存在不妥甚或错误之处，敬请读者不吝赐教。

本书可供水利及环保部门从事水质监测及水环境保护的科技人员或管理人员阅读参考，也可供高等院校应用统计专业或水环境相关专业的教师、学生在教学科研中参考学习。

作者

2013年11月

目　录

第1章 绪 论

1.1 水质分析背景

水是生态系统维持及健康发展的物质基础之一，是21世纪可持续发展战略实施的重要保障。随着社会经济发展，水在国民经济和社会发展中的地位和作用日益突出。水环境问题伴随着人类社会经济活动的非理性发展而逐渐出现。水环境的污染造成的水资源危机已成为一个国家在政策、经济和技术上所面临的复杂问题，成为发展经济的主要制约因素之一，引起了国际上的广泛关注。水的问题主要分两种：水量问题和水质问题，前者表现为水量的匮乏，需要国家通过宏观的水量调动与分配进行解决；后者则表现为水质无法满足正常生活生产的要求，需要通过水质评价对不合格水体进行确定，并进而消除污染。

《中华人民共和国水法》、《中华人民共和国水污染防治》规定：水行政主管部门应当对水功能区进行水质监测，核定水域的纳污能力，提出水域限制排污总量控制的意见。这一规定为加强水质监测管理，提高水质监测分析与评价水平提出了明确要求。

水质监测是水质评价与水污染防治的主要依据，随着水体污染问题的日渐严重，水质监测成为社会经济可持续发展必须解决的重大问题，尤其是内陆水体，如大型湖泊与大江大河，其水质影响到国民生产和人们的生活用水，准确、快捷的水质监测显得尤为重要。近年来，各省市、各大流域机构均按照《中华人民共和国水法》的要求，加强了水质监测工作的领导。目前水质监测范围基本覆盖了各大流域，并逐步扩展到各大水源地、地下水及入河排污口，全国重点一、二级水功能区均已建立起比较完善的水质监测体系，并积累了大量可供查询、分析的水资源质量信息。这种比较完善的水质监测体系，为建设水质统计分析与评价系统提供了比较有利的条件。随着社会对水质监测信息的广泛重视，各级水行政主管部门对水质监测的工作要求也不断提高。目前，水功能区常规监测评价、供水资源地监测评价等水质分析与评价工作正逐步深化。建设科学完善的水质统计分析方法与评价系统已经成为各级水行政主管部门的迫切需求。

1.2 国内外研究概况

在资源日益紧张的今天，日常生活所必需的水资源广泛引起人们的关注，在全世界都在提倡节约水资源、保护水资源的时代，人们希望有关部门对水资源状态有一个实时的信息公布。国外学者对水质模型的研究起步较早，主要以欧美等发达国家的研究为代表，1925年Streeter和Phelps提出的第一个水质模型，即河流S-P模型，主要进行氧平衡的研究，也涉及一些非耗氧物质，属于一维稳态模型。1980年Wood提出基于稳态水力模

型的水质模型后，1986 年 Clark 等提出了一个能够在时变条件下模拟水质变化的模型，Grayman 等在 1988 年提出了一个类似的水质模型，大部分模型都使用了“扩展时段模拟”（Extended Period Simulation，简称 EPS）方法，因为它们没有模拟由于流速变化造成的惯性影响，故实际应称作准动态模型。作为试验与应用基地的美国环保署（Environment Protection Agency，简称 EAP）建立了两套输配水模拟器 DSS，用于水质模型的研究，并发布了供水模型 EAPNET，该模型已被国外公共事业领域以及私营领域广泛使用，同时进行了大范围的模型参数校验工作。由美国环境保护局提出的水质模型系统 WASP（The Water Quality Analysis Simulation Program，水质分析模拟程序）可以对河流水质模型中能进行一维、二维、三维动态水质模拟，实现了多维模拟、多介质模拟、形态模拟、动态模拟等特点。针对水质统计分析方法，美国农业部 2003 年颁布了《国家水质手册》（Handbook of Water Quality）。国内的研究学者在水质模型研究方面也作了不少工作。河海大学开发了河网水量、水质统一的 Hwqnow 模型。郭磊、高学平等建立了水动力、水体污染物输运及底泥污染物输运数值模型，采用有限差分与有限体积相结合的方法，对北大港水库氯离子进行动态数值模拟。清华大学申满斌、陈永灿等针对三峡库区主要污染物，建立了考虑泥沙吸附污染物和泥沙冲淤对污染物输移扩散影响的岸边排放污染物浓度场计算的三维浑水水质模型。重庆市环境科学研究院和重庆大学针对长江嘉陵江重庆段干流和城区江段，分别开发了一维和二维水质数学模型，取得了较好的模拟效果。

近几年来，人工神经网络在水质模型方面的应用取得了飞速的发展。人工神经网络（Artificial Neural Network，简称 ANN）是 20 世纪 80 年代中后期世界范围内迅速发展起来的一个前沿领域，因其良好的预测性和实用性被广泛应用于各个领域，譬如人工神经网络除可以直接应用于对水质进行模拟预测外，还可以被嵌入到水质模型模拟中，如通过人工神经网络率定水质模型中的各参数，使其对水质的分析和模拟过程更趋于合理化，同时增强处理非线性问题的能力，提高预报精度等。目前已发展了多种神经网络，例如 Hopfield 模型、Feldmann 等的连接型网络模型、Hinton 等的玻尔茨曼机模型，以及 Rumelhart 等的多层感知器模型和 kohomen 的自组织网络模型等。在这众多神经网络模型中，目前较为广泛应用的 BP 神经网络，它通常是指基于误差反向传播算法（BP 算法）的多层前向神经网络，它是 D. E. Rumelhart 和 J. L. McCelland 及其研究小组在 1986 年研究并设计出来的。BP 算法已经成为目前最为广泛的神经网络学习算法，据统计有近 90% 的神经网络应用是基于 BP 算法的。BP 神经网络具有并行处理、非线性、容错性、自适应和自学习的特点，在数据拟合与模拟中有着无比的优越性。T. R. Neelakantan 等用人工神经网络建立了水库运行的模拟—优化模型；Marina campolo 等用 ANN 来预测河流枯水期的流量并得出结论：当它与水质模型相结合时对河流的水质管理非常有用；V. chanramouli 等用动态规划和 ANN 来模拟多水库水系的运行方案；一些学者尝试用遗传算法优化 BP 神经网络，将优化后的 BP 神经网络用于水质评价，取得较好的效果。ANN 还可应用于水系模型的误差更新，随着 ANN 的不断发展和完善，相信 ANN 在水质模型方面的应用将会更深入、更全面、更系统。

模糊数学在水环境方面有很多应用。基于模糊水质模型的研究多属于水质评价模型，针对预测性研究的模糊模型相对较少，这类研究多以模糊判别、随机模型方法或灰色模型

方法作分析预测联合使用，主要以宏观性整体性预报为主。众所周知，水文环境条件有很大的随机性，要定量分析相关关系有很大的困难。此外，水质的变化是连续的，而水质标准中的污染物浓度的表示却是不连续的。为了解决这一矛盾，1965 年美国控制论专家 Zadeh L. A. 提出模糊集合的概念，模糊数学得到了前所未有的发展，同时被广泛运用于生产实践中。Y. Y. Yin 等运用模糊关系分析（Fuzzy Relation Analysis，简称 FRA）模型来分析大量的不同的备选方案，同其他的在不确知情况下影响评价的多准则方法相比，FRA 法在数据的可获得性、需求的计算能力和结果说明上有优势。将熵和熵权赋值法引入到模糊数学水质评价模型中来，可以有效避免主观确定权重的随意性，在实际的应用中计算简单、结果合理。

由于环境的水文条件具有很大的随机性，这就导致了水环境数学模型输出的不确定性。Andrews K. T. 等分析了模拟—优化模型中不确定性的来源有：污染物的排放量和河流背景值的随机性；估计模型参数所需的河流和水质资料的不充分；对污染物传输过程和水质管理系统的简化缺乏充分的认识。为了提高模型的精确度和结果的可靠性，有必要对模型不确定性进行研究。模型不确定性分析（Uncertainty Analysis）的核心是水质参数的不确定性。徐一剑、曾思育等基于不确定性分析的框架，开发了动态环状河网水质模型，用 HSY（Homberger Spear Young）算法作水质参数的不确定性分析，求得模型参数的空间分布，从而提高模型使用的可靠性，降低决策的风险度，有效解决了我国平原地区环状河网水文条件复杂及监测数据稀缺的问题。

遥感（Remote Sensing，简称 RS）技术最早是由美国海军研究局的艾弗林·普鲁伊特（Evelny L L. Puritt）提出的。遥感即遥远的感知，遥感技术是一种利用物体反射或辐射电磁波的固有特性，远距离不直接接触物体而识别、测量并分析目标物性质的技术。近几年来遥感技术的发展使得河流、湖泊水体污染的监测更加便利。通过遥感技术，可以获得大尺度的水质信息，并且具有高时间分辨率、连续的特点，充分弥补了在水质监测中的不足。

另外，灰色聚类法、物元分析法和主成分分析法也广泛应用于水质分析。

整体来说，国外对水质模型的研究比国内的研究要早，已经建立了多种水质模型，广泛应用于水质规划及环境治理方面，并将其软件化，提高了模型的通用性。国内针对河流水质及其模型的研究方面做了大量工作，对水质模型的研究已由初期探索阶段逐渐深入到对河流水质的变化规律进行监测和趋势研究，上升到了一定的理论高度，起到了指导实践的作用。

1.3 水质统计分析目的

水质统计分析对于现今的人类生存和发展来讲，有着至关重要的作用和意义，一方面，通过水质分析，可以明确水质的主要情况，了解污染物时间和空间发展变化规律与特征，并在数学建模的基础上，对水质趋势进行分析和预测，及时掌握水功能区水质状况，为保护水资源和防治水污染提供科学依据。另外一个方面，对于自然资源的开发和有效的利用来讲，也是不可缺少的重要环节。

第 2 章　概率与数理统计基础

2.1　随　机　事　件

2.1.1　随机试验

概率论与数理统计研究的对象是随机现象。

在一定条件下，并不总是出现相同结果的现象称为随机现象，最简单的如：抛一枚硬币。随机现象有两个特点：

(1) 结果不止一个；

(2) 哪一个结果出现，人们事先并不知道。

通过试验来研究随机现象，但这种试验是广义的，除各种各样的科学试验外，对随机现象的观察，也被认为是一种试验。下面是一些水质随机试验的例子。

例 1：观察一条河流是否出现大型无脊椎动物。

例 2：观察一条河流溶解氧的等级。

例 3：观察一些河流藻类存在的区域。

例 4：记录一条河流不同时间段内的含氮量。

例 5：在某流域任选一条河流测试其污染的程度。

例 6：观察一天中某河流溶解氧的最高值和最低值。

不难发现，上面这些试验都具有以下几个特点：

(1) 试验在相同条件下可以重复进行；

(2) 试验的可能结果不止一个，但试验前能明确所有可能的结果；

(3) 试验前不可预知哪个结果会出现。

把具有上面三个特点的试验称为随机试验，简称为试验，记为 E。

2.1.2　样本空间

为了描述随机现象，可以通过随机试验来考察随机现象各种可能出现的结果，并引入样本空间的概念来记录它们。

随机试验 E 的所有可能的试验结果所组成的集合，称为 E 的样本空间，记为 Ω。样本空间的每个元素，即试验的每个结果，称为样本点，记为 ω。

例如，测量一条河流中重金属的含量指标，这就是一个样本空间，其中单个的砷的含量指标、汞的含量指标就是样本点。

注意：从上面的样本空间可看出：①样本空间的元素个数可以是有限个，也可以是可列个，还可以是不可列个；②样本空间的元素可以是一维的，也可以是多维的；③对一个

随机现象观察的角度、目的不同，样本空间可能不一样。

2.1.3 随机事件

在试验的过程中，有很多情况可能发生也可能不发生。例如，在测定一条河流中重金属是否超标中，“超标”、“不超标”都是可能发生也可能不发生的情况；在记录某河流断面在某段时间间隔内溶解氧含量是否达标的试验中，“溶解氧正常”、“溶解氧低于标准100倍”，都是可能发生也可能不发生的情况，而往往人们对这些情况感兴趣。为便于研究，我们把试验中可能发生也可能不发生的情况，称为随机事件，简称为事件。通常用大写字母 A，B，…表示。

任何事件总对应样本空间的某一个子集。例如：在测量一条河流中汞含量是否超标的试验中，事件 A：“汞含量正常”，即 $A=\{$含量少于 0.001mg/L$\}$；在记录某河流某一断面在某段时间间隔内溶解氧含量是否达标的试验中，事件 B：“溶解氧含量达标”，即 $B=\{$大于 7mg/L$\}$。

于是，数学上可以这样定义随机事件。试验 E 的样本空间 Ω 的子集称为 E 的随机事件，简称事件。事件发生指的是对应的子集中某个样本点出现。每次试验，样本空间中必然有一个样本点出现，即每次试验样本空间作为事件总是发生，此时称 Ω 为必然事件。空集 ϕ 也是样本空间的子集，但不包含任何样本点，每次试验空集作为事件不可能发生，故称 ϕ 为不可能事件。另外，称由样本空间中单个样本点组成的集合为基本事件。

2.1.4 事件间的关系及运算

同一试验的众多事件往往存在着一定的关系。研究这些关系，有助于我们认识、分析更加复杂的事件。因为事件是一个集合，故事件间的关系、运算可按集合间的关系、运算来处理。根据“事件发生”的含义，不难理解事件间的关系、运算的概率含义。

设 Ω 为试验 E 的样本空间，A，B，…都是事件，即 Ω 的子集。

1. 包含关系

如果 $A\subset B$，则称事件 A 包含于事件 B，或事件 B 包含事件 A，它是指事件 A 发生必然导致事件 B 发生。

如果 $A\subset B$，且 $B\subset A$，则称事件 A 与事件 B 相等，记为 $A=B$。

2. 事件的和

两个事件 A，B 至少有一个发生也是一个事件，称为事件 A 与事件 B 的和事件。记作 $A\cup B$。显然 $A\cup B=\{\omega\in A$ 或 $\omega\in B\}$。

类似的称 $\bigcup\limits_{k=1}^{n}A_k$ 为 n 个事件 A_1，A_2，…，A_n 的和事件；称 $\bigcup\limits_{k=1}^{+\infty}A_k$ 为可列个事件 A_1，A_2，…，A_n，…的和事件。

3. 事件的积

称 $A\cap B=\{\omega\in A$ 且 $\omega\in B\}$ 为事件 A 与事件 B 的积事件。$A\cap B$ 发生当且仅当事件 A 与事件 B 同时发生。$A\cap B$ 也简记为 AB。

类似地，称 $\bigcap\limits_{k=1}^{n}A_k$ 为 n 个事件 A_1，A_2，…，A_n 的积事件；称 $\bigcap\limits_{k=1}^{+\infty}A_k$ 为可列个事件 A_1，A_2，…，A_n，…的积事件。

4. 事件的差

称 $A-B=\{\omega\in A$ 且 $\omega\notin B\}$ 为事件 A 与事件 B 的差事件。$A-B$ 发生当且仅当事件 A 发生且事件 B 不发生。

5. 互不相容事件

如果 $A\cap B=\Phi$，则称事件 A 与事件 B 互不相容，或称事件 A 与事件 B 互斥。这意味着事件 A 与事件 B 不能同时发生。

6. 事件的互逆

如果 $A\cap B=\Phi$，且 $A\cup B=\Omega$，则称事件 A 与事件 B 互逆，或称事件 A 与事件 B 互为对立事件。这意味着事件 A 与事件 B 不能同时发生，但必有一个发生。一般地，事件 A 的对立事件记作 $\overline{A}$，不难得到 $A-B=A\overline{B}$。

上面关于事件的各种关系可用图 2.1 直观地表示出来。

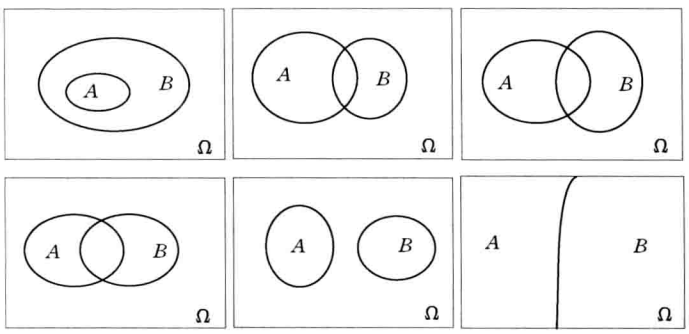


图 2.1　事件关系示意图

事件的运算与集合的运算是一致的，在进行事件的运算时，经常会用到以下规则：

(1) 交换律：$A\cup B=B\cup A$，$A\cap B=B\cap A$；

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

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

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

(4) 德·摩根 (De Morgen) 律：$\overline{A\cup B}=\overline{A}\cap\overline{B}$；

$\overline{A\cap B}=\overline{A}\cup\overline{B}$。

德·摩根律又称为对偶律，它可利用事件的关系来解释。因为 $A\cup B$ 表示事件 A 与事件 B 至少有一个发生，它的对立事件 $\overline{A\cup B}$ 当然是两个都不发生，而 A 与 B 均不发生的另一种表示是 $\overline{A}\cap\overline{B}$，即 $\overline{A\cup B}=\overline{A}\cap\overline{B}$ 成立。类似地可解释等式 $\overline{A\cap B}=\overline{A}\cup\overline{B}$ 成立。

2.2　随　机　变　量

为了更方便、更系统地研究随机试验，揭示随机现象的统计规律性，首先要做的是按照研究的需要，将随机试验的结果数量化，即把试验的每个结果用一个实数与之对应。例如，水质检验结果是“达标”与“不达标”的一个随机试验，可以用实数“1”对应“达标”，用实数“0”对应“不达标”，这样，本来由“达标”和“不达标”组成的样本空间

对应于由 0 与 1 组成的集合。又如，观察一条河流大型无脊椎动物出现的种类，其样本空间为={1，2，3，4，5，6}。如果只关心 6 点是否出现，当试验结果是 6 时，让它对应实数 1，当试验结果是其他情况时，让它们都对应实数 0。此时，该试验的样本空间 Ω 也对应于由 0 与 1 组成的集合。如果引入变量 X 来表示试验结果对应的实数，显然，它的取值依赖于样本点，它是定义在样本空间 Ω 上的实函数。因为样本空间中的样本点哪个会出现，试验前无法确定，即样本点的出现是随机的。所以，变量 X 的取值也是随机的。且它的取值会以一定的概率出现，故称 X 为随机变量。

(1) 研究随机变量及其分布以及各种性质构成了概率论的核心内容。它的重要性除以上所述外，还对试验各种结果的概率特性进行了更广泛意义上的研究。令 A 为任意事件，现定义一个随机变量 X，当 A 发生时，$X=1$；当 A 不发生时，$X=0$，则求概率 $P(A)$ 完全等价于研究该随机变量取值为 1 的概率，而这个概率完全由该随机变量的分布所决定，因此研究随机变量的概率分布极其重要，后面我们将一一展开。

下面给出随机变量的具体定义。

定义 2.1 设 E 是随机试验，Ω 是其样本空间。如果对于 Ω 中的每个样本点 ω，总有唯一确定的实数值 $X(\omega)$ 与之对应，则称 $X(\omega)$ 为随机变量，简记为 X。

通常用大写字母 X，Y，Z，…表示随机变量，用小写字母 x，y，z，…表示它们的可能取值。

引入随机变量的概念后，可以用随机变量的取值来表示随机事件。

例如，测量某一流域内河流重金属含量是否超标，其样本空间为 Ω={超标，正常}。定义随机变量 X，当含量超标时，$X=1$；当含量正常时，$X=0$，则 $\{X=1\}$ 表示事件“重金属超标”，对应于集合 {1}；$\{X=0\}$ 表示事件“重金属正常”，对应于集合 {0}。

(2) 一般地，对于实数集的某个子集 L，随机变量在 L 内取值记为 $\{X\in L\}$，其表示事件$\{\omega|X(\omega)\in L\}$，即由使随机变量 X 的取值落在 L 内的样本点组成的事件，也就是有 $\{X\in L\}=\{\omega|X(\omega)\in L\}$。

按照随机变量的可能取值情况，将随机变量分为两种不同的类型：

(1) 离散型随机变量。这类随机变量的可能取值为有限个或可列个。如掷一颗骰子，X 表示掷出的点数，它可取 6 个值，为离散型随机变量。又如 Y 表示某公交站台等车的乘客人数，则 Y 可取 0，1，2，…的任一非负整数值，即可取可列个值，也为离散型随机变量。

(2) 非离散型随机变量。这类随机变量的可能取值为无穷不可列个。如 X 表示某水利工程的寿命，理论上 X 可取区间 $[0，+\infty)$ 内任一实数值，故其为非离散型随机变量。

在水质监测项目中收集到的随机变量包含两种类型：连续型和离散型。数据类型的选择会影响统计的应用。我们根据收集到的信息判断选择数据的类型。

用测量设备测量得到的可以连续取值的数据，称为连续型数据，也称计量型数据。连续型数据意味着在一些范围内的所有数值都是可能取到的，它可以在连续坐标上表示。这类数据的特点是数据能够比较敏感地反映被测量的变化，包含的信息比较丰富，在进行统计分析时，可以用较少的样本数量获得准确的结论。但是，有时对测量手段要求较高，有

时需要花费比较大的成本去获取数据。

离散（或非连续）型数据，也称计数型数据，如合格/不合格、通过/不通过、是/否、好/坏等都具有分明界限的测量数据。这类数据不如计量型数据包含的信息那样敏感丰富，在进行统计分析时，所需样本量往往比较大，但其对测量手段和成本的要求不高。随着现在对计数型数据的统计分析方法的不断应用，计数型数据越来越受到重视。

实际测量时，确定被测数据的类型十分重要，它决定了数据的统计分析方法以及统计结论的可靠性。通常，在不考虑成本因素时应尽量获得计量型数据以便提供尽可能多的有用信息。

通常，被测量的真值是未知的。测量的目的是使测量结果数据尽量逼近真值，数据所包含的信息的多少不但取决于数据的类型，更重要的是依赖于所选择的测量尺度，而测量尺度才真正决定研究这些数据时应使用的统计分析方法，也只有确定了测量尺度，才能真正了解这些统计分析方法是否适用和有效。

在统计学上，将测量尺度分为四类：分类、定序、定距和定比。

分类数据包括一些类别但不将类别进行排序，这类数据并不是测量得到的，而是按现象的性质差异进行辨别与分类的，是可以以文字表述或以数字形式表示的名义值，但只起到标签的作用。通常，分类数据是二进制的，例如存在或不存在。一条溪流中是否出现大型无脊椎动物的分类群可以作为分类数据的例子。

分类尺度是测量形式中最简单、最弱的一种，把分类变量看成一种分类形式比看成一种测量尺度更确切。在分类尺度中对分类的划分是有要求的，那就是样本中的所有项均应属于一类且仅属于一类。以分类尺度收集的数据称为属性数据。对于分类尺度，允许进行的运算只有“=”（表示物体具有某属性）或者“≠”。另外，分类尺度也可包含两个或以上的水平，但这些水平并没有自然顺序。

定序数据意味着数据是有排序的。顺序变量用来测量一些东西的等级。定序尺度可以对可能的取值进行排序，但是不对数值之间的间距进行定义。定序数据是按现象的顺序差异进行辨别与分类的结果，可以文字表述或以数字形式表示，但也只起到标签的作用。例如营养等级——贫营养的、中营养的和富含营养的，就是由顺序尺度计量。定序数据可以具有3个或以上的水平，这些水平均具有自然的有序性，这类数据也属于属性数据。

对于定序尺度，可以使用的运算有“=”（相等）、“≠”（不等）、“>”（大于）和“<”（小于）。

定距尺度是一种连续型数据尺度，它按现象的绝对数量差异进行辨别与区分，以数字表示，有计量单位。这种数据中“0”是没有意义的，且没有倍数的概念。用定距数据所做的统计分析可以使用算术平均等各种统计量，但没有“比值”、“比率”的概念。

定距数据也用到排序，但不同类别间的间隔是相等的，用这些间隔和类别来描述数据。例如，温度区间可以是小于25℉，25～50℉，50～75℉和大于75℉；区间也可以用来描述鱼的大小类别，例如小于10cm，10～20cm，20～30cm和大于30cm。

定比尺度是另外一种连续型数据尺度，比定距尺度数据更高了一个层次。这种数据中的“0”是有意义的。定比尺度数据不但可以识别数据间差距的大小，而且可以识别和比

较比值的大小；不但可以对定比尺度数据进行算术平均运算，使用统计量，而且可以进行除法运算求得倍数。

对于各种测量尺度以及可使用的统计方法的比较，如下表 2.1 所示。

表 2.1　　测量尺度类型和可使用的统计方法

尺度	定　义	例　子	统计方法
定类	具备/不具备某属性，只能计算属于某类别的个数		百分比；比例；χ^2 检验
定序	可以说某项所包含的属性比另一项多/少；可以给一些项目排序		排序；相关性
定距	任意两个相邻点之间等距；即使等距假设不正确；常常被当做定比尺度；可以加减、排序		相关性；t 检验；F 检验；多元回归
定比	零点表示不具有属性；可以加、减、乘、除		t 检验；F 检验；相关性；多元回归

2.3 概　　率

2.3.1 概率的定义

一个事件在一次实验中可能发生，也可能不发生；不同的事件在同样的实验中发生的可能性有大有小。简单说来，在一次试验中，用来衡量事件 A 可能性大小的度量（数值），称为事件 A 发生的概率，记作 $P(A)$。为了研究概率的大小与性质，首先引入频率的概念，用以描述事件发生的频繁程度。

2.3.2 频率

在相同条件下做 n 次试验，如果事件 A 在这 n 次重复中出现了 n_A 次，则称 $\frac{n_A}{n}$ 为事件 A 发生的频率，记为 $f_n(A)$，即

$$f_n(A)=\frac{n_A}{n} \tag{2.1}$$

不难证明，频率有如下性质：

(1) 非负性：$0\leqslant f_n(A)\leqslant 1$；

(2) 规范性：$f_n(\Omega)=1$；

(3) 可加性：对于任意有限多个两两互不相容的事件 A_1，A_2，…，A_k，有

$$f_n\left(\bigcup_{i=1}^{k} A_i\right)=\sum_{i=1}^{k} f_n(A_i)\text{。} \tag{2.2}$$

试验次数 n 不同，事件 A 的频率一般来说是不同的，也就是说频率具有随机性；但当 n 比较大时，一般而言，能呈现某种规律性，即一个随机事件 A 发生的频率 $f_n(A)$ 通常在某个常数 $P(A)$ 附近摆动，而且当实验次数 n 很大时，频率 $f_n(A)$ 便会很接近常数

$P(A)$，这就是频率的稳定性或随机现象的统计规律性。

必须指出，尽管 n 很大时频率 $f_n(A)$ 会很接近常数 $P(A)$，但这并不意味着 $n_1<n_2$ 时，$f_{n_2}(x)$ 会比 $f_{n_1}(x)$ 更接近 $P(A)$。

2.3.3　概率的公理化定义

定 2.2　设 E 是随机试验，Ω 是它的样本空间，对 E 的每一个事件 A，赋予一个实数 $P(A)$ 与之对应，如果 $P(A)$ 满足如下条件：

(1) 非负性：对于每个事件 A，有 $P(A)\geqslant 0$；

(2) 规范性：$P(\Omega)=1$；

(3) 可列可加性：对于任意可列个两两互不相容的事件 A_1，A_2，…，A_n，…，即 $A_iA_j=\Phi$ $(i\neq j,\ i,\ j=1,\ 2,\ 3\cdots)$，有

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

则称 $P(A)$ 为事件 A 的概率。

该定义一般称为概率的公理化定义。由概率的定义，可以得到概率的一些重要性质。

性质 1　$$P(\Phi)=0$$

性质 2　$$P(\bigcup_{i=1}^{k}A_i)=\sum_{i=1}^{k}P(A_i)$$

性质 3　设 A，B 是两个事件，如果 $A\subset B$，则有

$$P(A)\leqslant P(B),P(B-A)=P(B)-P(A)$$

性质 4　对于任意事件 A，有

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

性质 5　对任意两事件 A，B 有

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

2.3.4　古典概型

如果一个随机试验的样本空间是由有限个样本点构成，且每个样本点在试验中是等可能出现的，那么，事件 A 发生的概率就可以用下列公式来计算

$$P(A)=\frac{m}{n}$$

式中：m 为事件 A 所含的样本点的个数；n 为样本空间中全部样本点数。

2.3.5　几何概型

如果样本空间 Ω 可以用一个集合区域 G 来表示，样本点 w 落在 G 的任意子区间 A 中的可能性与区域 A 的几何测度（曲线的长度，曲面的面积，立体的体积等）成正比，但与 A 的形状以及 A 在 G 中所处的位置无关，这时事件 A 发生的概率 $P(A)=\dfrac{A\text{ 的测度}}{G\text{ 的测度}}$。

2.3.6　条件概率与事件的独立性

1. 条件概率的定义

在随机试验中，有时除了需要知道事件 B 发生的概率 $P(B)$ 外，还需要知道在事件 A 已经发生的条件下事件 B 的概率，把这个概率记作 $P(B|A)$，我们称之为条件概率。其