统计软件方法

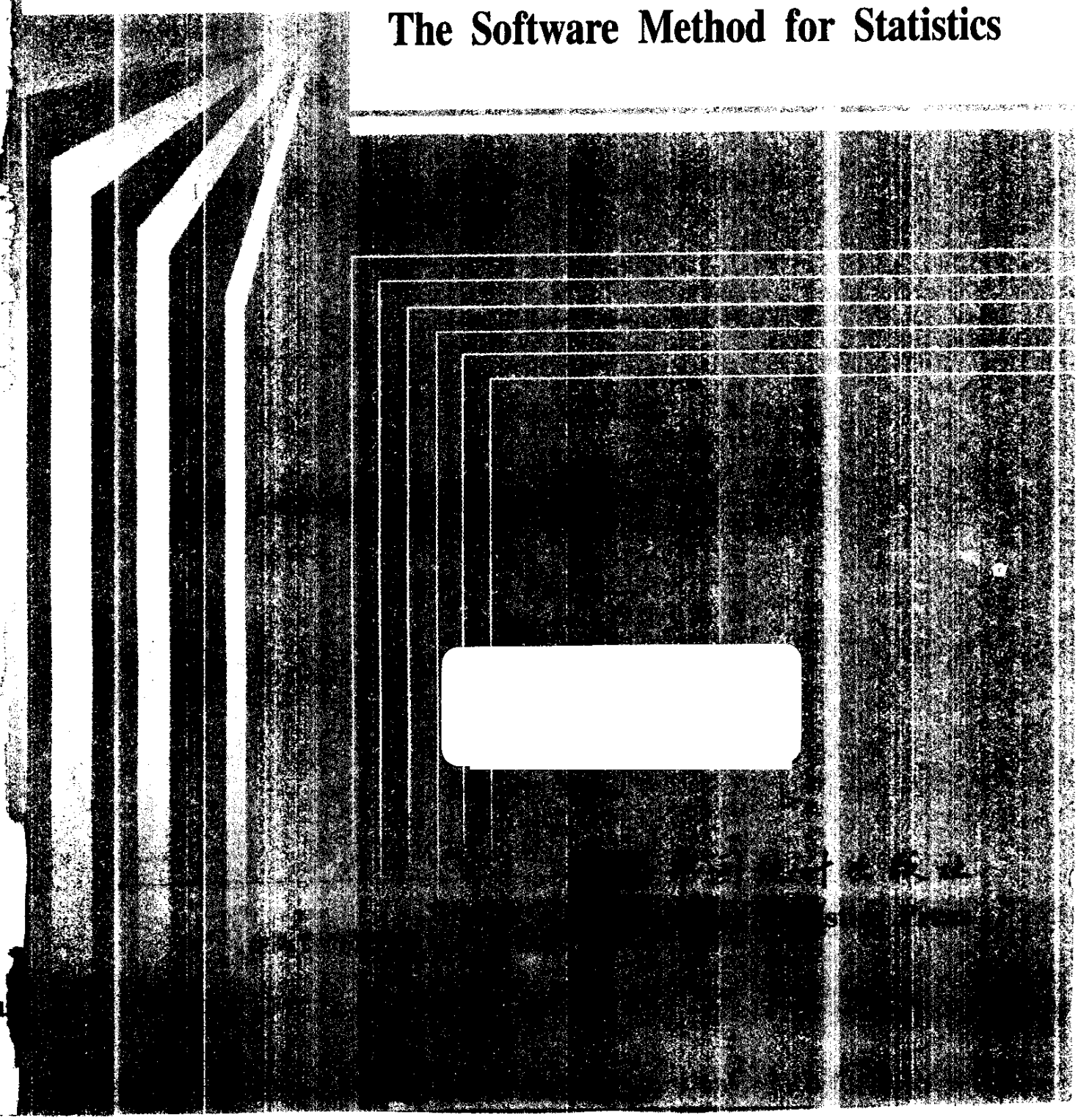
The Software Method for Statistics

米子川 著

中国统计出版社
China Statistics Press

统计软件方法

The Software Method for Statistics



前　言

统计学的生命力在于应用。

统计软件是寄生在统计领域的应用软件，还是游离于计算机软件边缘的统计工具？这是我们首先需要界定的问题。

在统计应用实践中，统计软件的功能在近30年中正变得越来越强大。随着计算机技术的飞速发展和统计应用方案的不断成熟，以计算机为主要工具的统计计算模型迅速成熟起来，并被广泛地应用到社会学、经济学、医学、体育及其它研究领域。目前，高校、科研部门、市场研究机构、医学研究及市场营销分析人员正在大量地使用统计软件进行大规模的数据分析和模型创建，这些应用必将极大地促进统计软件的普及和应用水平的提高。

作者认为，统计软件方法不仅只是一种统计计算工具，而且是一种统计研究的方法。在数据的探索性分析中，在高维数据的模型创建中，在海量数据的挖掘中，统计软件正以独特的视角和研究切入点成为统计学发展的一个新的分支。

统计软件是统计学发展的一个独立的阶段，而不仅仅拘囿于一个独立的工具。目前，这个阶段正处于上升期，随着应用的逐渐深入，大量的应用经验正被整理出来。

统计软件模型不局限于数学模型和一般的统计数据推断模型，还应该被扩展到广义的应用模型中。这种模型将以更加贴近的方式取得对原始数据的近似。

谨以此书献给我的父母，他们一直生活在美丽的晋中农村，从

教多年，耕读不辍，经常提醒我注意生活中的细枝末节，正是这些微不足道的小事使我对工作和生活充满信心和乐趣。感谢我的妻子梁秀萍律师，我们经常讨论一些统计问题到深夜，她的活跃的思想和不拘一格的学术态度常使我兴奋不已。在我的专业上，她并不是一个外行。感谢李宝瑜教授和雷钦礼教授，没有他们的努力与关怀，就没有本书的出版；感谢刘建平教授、杭斌教授、王培勤教授、赵俊康教授，王天保教授、王桂花教授、马淑琴教授多年来对我的学术提携和研究指引，没有他们的教诲，我的进学之路将会更加漫长。

本书是讨论统计软件方法的一个开始，是系统探索统计软件应用规律的一次尝试，也是我创作的第一本书。它是不成熟的和有缺陷的，尽管殚尽竭虑，依然可能破绽重重。我衷心希望有更多的人来关注和重视统计软件的应用和探索性研究，为统计学的发展一尽绵薄之力，也使得本书更加成熟和完善。我恳切希望得到同行专家的批评指正，我的联系方式是 E-mail：mi _ zc@163.com。

米子川

2002 年 5 月 1 日于北京

目　录

第1章　统计软件引论

随着社会经济的发展，中国正逐渐融入世界。广泛而深入的市场数据研究和理论探索长期持续进行，大量深刻有序的数据呈现在不同媒体，管理方法日新月异但更加依赖统计分析，质量科学日渐走红，客户满意度测评及客户关系管理如火如荼，这些挺立在经济发展潮头的应用理念，无一不与统计学有着极大的关联。

统计学正在迎来一个应用的春天。

统计软件作为一种专门的应用软件，不仅作为工具，更作为一种方法在统计应用的过程中起到了无可替代的积极作用。在未来的日子里，这种作用还将不断地发扬光大。

§1.1　统计学的历史回顾

几乎所有的学术讨论都会从历史开始，对统计学的回顾也是如此。但是，从统计软件的角度来看待统计学历史，却是一个真正的开头。

统计技术是指运用统计学的方法原理，通过获取和提炼信息，高效率地解决实际问题的一门通用技术。目前，国际上科技、生产和市场的竞争无不与统计技术有关。统计学包含了两大方面内涵丰富的统计技术：一是获取信息的技术，即为了经济有效地获取数据资料，应该如何科学地进行观测、调查或试验；二是提炼信息的技术，即如何运用所获得的数据资料，用统计分析方法对实际问题

的规律性及其因果关系进行科学的分析和推断。

我们不妨一起回顾统计学的几个基本概念：

统计学 通过研究数据(资料)，包括数据的产生、收集、整理、描述、分析和推断，发现新知识和有用的信息，从而对所研究的问题给出解答和说明的一门学问。通常，这里的解答可以称之为统计结果，说明可以叫统计解释。如果把问题及其解答和说明整理成文章，则称之为统计报告。统计学是一门科学，也是一门艺术，其思想和方法博大精深。因此，统计学是现代科技和文明的一个宝库。

统计方法 指统计学中关于数据的产生、收集、整理、描述、分析和推断等方面所采用的方法。

数理统计 指统计学的数学基础部分，也是统计方法的重要基础。数理统计的一些研究领域包括参数估计，假设检验，非参数统计，大样本统计，回归分析，多元分析，时间序列分析等。

应用统计 指统计学的应用方面，主要研究统计方法及其在实际中的应用，以及结合其它学科的情况和特点如何来应用统计学。应用统计的一些研究领域包括抽样调查，试验设计，抽样检验，质量控制，可靠性工程，生物医学统计，社会经济统计，地质统计学，等等。

统计方法 一种数据收集和分析处理的工具，其应用的好坏和水平高低，要看使用者的素质，这跟工匠使用工具或剑客使用刀剑的道理是一样的。因此，统计方法的应用是一门技术，我们称它为统计技术，它是统计方法成功实践的经验积累。由于统计学是一门比较深奥的学问，在应用中，统计方法被误用或使用不当的情况是很常见的。统计技术要求运用统计学的原理和方法，科学且经济有效地解决实际问题，追求高效益。

从本质上看，统计学的历史就是应用的历史。没有应用的推动，就不会有统计学繁荣发展的今天。恩格斯说，社会一旦产生需求，就会比办十所大学更能推动社会生产力的进步。从统计学最初的发端来看，无论是社会管理、人口研究，还是农业生产、科学试验

以及工厂管理，几乎所有的应用都是以数据做媒介的。统计应用不仅为统计学带来活力，也为各个不同行业的进步提供了积极的方法支持，这就是统计学的价值所在。今天，我们可以看到，大量的统计技术被应用在质量管理、市场研究、数据挖掘、经济分析、金融和政策研究、工厂现场管理、科学试验等场合，由此而带来的价值，难以估量。这也是统计软件得以迅速发端的重大背景。

统计学的历史是方法的历史。我们在回顾统计方法的时候，与许多真正有价值的方法不期而遇。在我的理解中，统计学的本质在于对潜藏在大量的随机现象中的统计规律的发掘与探索，这种大浪淘沙般的艰苦工作与科学的统计方法有着天然的联系。只有科学有效的方法才可以得出正确和有价值的结论。比如方差分析技术的应用，我们可以通过方差分析清楚地分辨出数据变化的来源究竟是随机的波动，还是因子之间的巨大差异，甚至可以搞清楚单个因子不同水平之间的差异有多大。随着应用的不断深入，我们看到统计方法依旧在不断更新，针对新的应用数据的分析方法依旧层出不穷。在 SPSS、SAS 等软件的最新版本中，我们经常可以看到一些应用前沿的方法开始通过软件的传播发挥其巨大的作用，为社会创造价值。我们知道，方法的进步就意味着统计的进步。

统计学的历史是计算的历史。统计学的应用对计算能力的要求较高，在计算一些统计模型时，这种要求显得尤为重要。比如，我们要做一个多元的线性回归模型，就必然会遇到一个求解多元方程的问题，如果数据比较多，这个计算就会十分困难。现在的常用方法是借助计算机软件，手工的计算能力已经难以达到了。事实上，许多统计方法在杰出的研究成果之后，并没有及时付诸应用，恐怕更多的阻碍还是在于计算能力的桎梏。这也是早期的统计软件一直集中精力在计算能力上大力发展的一个缘由。现代统计软件已经不仅仅是作为计算工具了，更多超乎想象的数据处理工作都可以通过统计软件做得更好。

统计学的历史就是数据分析的历史。数据是真正统计的开始，

就像盖房子一样，数据就是建筑材料。我们经常遇到这样的情况，同样的数据在不同的分析者手里，有时会得到不同的甚至是大相径庭的结论，这是为什么？在统计软件的应用中，选择不同的拟合方法，得到不同计算结果的情形会更多，如何解释这些结论就显得尤为重要了。从数据分析的历史上看，统计模型的计算结论不是统计应用的结束，而是统计应用的开始。统计软件是数据分析史上一个新时代的开始，是真正数据挖掘的起点。

统计学的意义在于发现蕴藏深远的规律，而不是发明深奥的统计模型。因此，统计学的历史也是淘汰和保留的历史，是一个对数据规律进行筛选和过滤的历史。我们了解从前的统计应用案例，许多方案已经不再具有价值了，比如利用平均差来评价数据的变异程度。而留下来的方法必然是实践的选择，像标准差。在这个历史中，我们无法断定哪些方法是一无是处的，哪些方法是奉若至宝的，只有通过实践的选择才可以得出真正的结论。

统计软件的发展在统计学的发展中起到了积极的推动作用，从某种意义上讲，统计软件的应用是统计学发展中一个不可割裂的重要阶段。我们不要简单地把统计软件作为一件工具，因为工具是实现劳动者操作目标的载体，工具使得劳动者的手臂延长了，速度加快了，但是不会替劳动者想到什么。工具能够做到的，劳动者当然什么都知道。而统计软件却不一样，你需要计算的指标，它可以为你一一罗列，你根本无法从数据中看到的指标、规律、关系、动态、分布等，它都可以精确计算出来，给你一个超乎需求的计算结论，这就不是一件简单的工具可以完成的。在数据挖掘场合，统计软件的巨大潜能再一次地爆发了。通过挖掘，我们不但得到了自己从未想象过的统计关系，而且不断地从中挖掘出能够产生巨大价值的决策依据，为应用环节提供动态的支持。

简单的评论

统计软件是统计应用环节中的一项专门技术，是高等统计方法的重要载

体，在统计数据分析中具有重要地位。统计软件的价值与使用者对统计学的理解有着高度相关。我们假定，研习统计软件者应该具有统计学、计算机应用、计算机软件等的基本知识，以及相应数据背景的专业知识，比如研究电信数据，则希望能够懂得电信机构的一些管理、运营、销售及服务的知识。

离开统计学的数据分析软件是没有价值的，而统计软件的核心技术正满足统计技术广泛而深入的应用需求。

§1.2 统计软件在应用与研究中的作用

统计研究的目的是希望获得更多隐性的数据规律，而统计软件是达成应用最好的方法。一般地，我们把统计软件作为一个应用软件。就软件而言，统计软件似乎与别的应用软件没有什么大的区别，比如财务软件。事实上，像财务软件这样的应用软件，是严格依照软件设计者的意图清晰而精确地工作的，是一种流程性的人工替代的计算机软件。而统计软件的独特性就在于，它的应用远远超过了程序设计人员的想象。如果统计软件的计算结论可以想象，那么统计软件的价值也就不复存在了。因此，在这里，我们把统计软件作为一种统计应用的方法，而不是简单的计算或人工替代工具。

统计软件在应用研究中的作用主要在对数据的全方位分析和处理上。这种应用主要表现在以下几个方面，我们将在本书中系统地讨论这些卓有价值的应用方法。

- ◆ 数据呈现和数字特征描述
- ◆ 统计分布和数值分析
- ◆ 变量统计关系
- ◆ 统计模型与检验
- ◆ 多元数据分析
- ◆ 数据挖掘

§1.3 统计软件方法的渐进过程

统计应用的层次有着清晰的分野。在统计软件的应用能力分析中，我们把这些过程划分为四个层次，这些层次之间是渐进的和逐渐深入的，反映了不同层次的数据与样本的关系，标志着统计学应用水平的不断提升。

1.3.1 描述统计

描述统计是统计学的一部分，是统计分析初始阶段各种方法的统称，主要研究如何对搜集到的数据进行既能描述数据全貌，又能反映所要研究现象的内容和本质的各种缩简数据的方法。描述统计主要包括统计分组，编制统计表，绘制统计图，计算各种统计量，如算术平均数、中位数、众数、几何平均数、调和平均数等表示集中趋势或代表值的一些统计量；标准差、平均差、四分位差、相对变异系数等一些表示离散趋势或分散程度的一些统计量；表示两列或多列变量之间相关程度的各种相关系数；表示变量之间变化的数量关系的回归系数，以及一些表示分布状态的特征值等。实现这一功能的统计软件比较多，是统计软件的基础功能模块，几乎所有的统计软件都可以完成统计数据的描述，像 EXCEL、LOTUS 等 Office 软件也可以高水平地完成描述统计。

1.3.2 变量关系分析

从统计意义上考虑，分析变量之间的关系有多种方法。从变量关系的类型看，包括相关关系与回归关系。从关系的密切程度看，两个以上的变量可以表现为多种不同的关系，包括相似、相关、关联、相合等，各自所依托的计算方法不同，所代表的变量类型不同，所表现的关系的强度也不尽相同。统计软件在分析和表现变量关系上有着很丰富的经验，可以快速计算和分析变量的各种关系，从