主编 胡良平

现代统计学与SAS应用

军事医学科学出版社

现代统计学与SAS应用

主　编　胡良平
编　委　代炼忠　郭秀花
　　姚　晨　张学中
　　胡良平

军事医学科学出版社

内容提要

本书较全面地介绍了现代统计学的理论、方法及其应用技巧。针对多元统计分析方法计算量大和算法复杂的特点，以 SAS 软件包作为实现复杂统计计算的工具，本书着重介绍各种试验设计方法、统计分析方法及其适用条件、结合具体问题正确选用统计方法的技术以及对计算结果的正确解释和应用。在一切从实际出发的思想指导下，合理调整教材结构和编写形式，把处理同一类问题的统计方法集中到一起讲解，使貌似复杂的统计问题化繁为简，实用方便。本书具有以下独到之处：用计算器和计算机两种计算工具实现统计计算，便于读者选用；针对问题和资料阐述统计方法，有利于读者提高处理实际问题的综合能力；介绍的试验设计类型多，讲解详细，具有很强的可操作性；为读者成功地使用 SAS 软件提供了一条有效的捷径。

根据教学对象的层次和学时数适当取舍内容，本书可作为研究生、本科生、大中专生的统计学教材，高等院校和科研机构的教师、学者、科技人员、生物医学工作者、管理工作者等学习和应用统计方法的参考书；还可作为用 SAS 软件解决统计问题的实用手册。

* * *

图书在版编目(CIP)数据

现代统计学与 SAS 应用/胡良平主编. －北京：军事医学科学出版社，2000.8
ISBN 7－80121－274－6

Ⅰ.现… Ⅱ.胡… Ⅲ.系统分析-应用-统计分析 Ⅳ.C8

中国版本图书馆 CIP 数据核字(2000)第 64744 号

* *

军事医学科学出版社出版
(北京太平路 27 号　邮政编码：100850)
新华书店总店北京发行所发行
潮河印刷厂印刷

*

开本：787 mm×1092 mm　印张：27.75　字数：689 千字
2000 年 8 月第 1 版　2000 年 8 月第 1 次印刷
印数：1－3 000 册　定价：40.00 元

前　言

统计学是什么？统计学有何作用？统计学研究些什么内容？如何学好统计学？怎样正确运用统计学？这些问题是刚刚涉猎统计学的人们必然要思考的，甚至有些已学过多遍统计学的人，仍在苦苦地琢磨着这些问题，而且，大有百思不得其解之"困惑"。至于前三个问题，在本书的第一章绪论中已作了详细的讲解，此处不便赘述。在此，就后两个问题展开一些讨论，希望能与读者沟通思想，交流感情，从而有利于作者与读者之间尽早达成一致共识，为传播统计学知识、有效地发挥统计学的作用作出更大的贡献。

要想学好统计学，首先要对统计学所研究的内容有一个较为全面的了解。这种了解，不是停留在表面上的，而是对各部分内容从原理、方法、适用条件、计算到结果解释等各方面都认真思考和反复实践过。其次要从问题的原形入手去学统计学。统计学教科书一般都按"由浅入深"、"分门别类"等思路去编写，但是，用统计学时却面对的是错综复杂的实际资料，常使人感到无从下手。只有在学习的全过程中，都始终注意"从问题的原形入手去学统计学"，即弄清每一种统计方法所能解决的问题在实际工作中是如何表现的，又是如何将其转化成"与特定统计方法对应的统计学问题的"。例如：在选择统计分析方法之前，必须判定资料是定量的还是定性的。这似乎是一个不值得一提的问题，居然在某名牌期刊上还出现了因误判资料类型，而错误地选择了统计分析方法的实例；又例如：在分析定量资料时，必须先判定资料所取自的设计类型和资料本身所具备的前提条件之后，方可正确地选择统计分析方法，但很多人却不加思索地盲目套用 t 检验；再例如：在分析定性资料时，必须先判定列联表中定性变量的属性和分析目的之后，方可正确地选择统计分析方法，但很多人却糊里糊涂地生搬硬套 χ^2 检验。笔者主编的《医学统计应用错误的诊断与释疑》一书（于 1999 年 1 月由军事医学科学出版社出版），将有助于读者识别医学期刊论文中有关统计学的各种误用现象，同时，也有助于读者防止自己在运用统计学中犯类似错误。

怎样才能正确运用统计学？要在较全面地掌握了统计学知识的基础上，逐渐扩大统计学的应用范围。每用一种方法，都要做到"心中有数"，即此时共有哪几种可能的统计分析方法，其中最好的是哪一种；处理此问题时人们常犯的错误是什么。尤其是面对多因素多指标的资料时，思路必须十分清楚，欲达到什么分析目的，应该选用什么统计分析方法，这种方法涉及到的资料可以包括哪些类型的变量，这些资料是否满足拟选定的统计分析方法所要求的前提条件，怎样巧妙地实现复杂的统计计算，等等。要想正确运用统计学，首先必须老老实实地学好统计学。笔者编著的《医学统计学内容概要、考题精选与考题详解》一书（于 2000 年 1 月由军事医学科学出版社出版），将有助于读者突破"根据各种不同的实际资料，正确选用统计分析方法"的难关。

下面将结合实例，就定量资料统计分析中的一个难点——设计类型的辨析，谈一点"如何从问题原形入手学习和运用统计学"的体会，供广大读者参考。

人们在处理实际资料前，常按习惯将实验资料按"组别"或"药物"等单个标题来划分，从列出的表格看，使人很容易将其视为"单因素多水平设计"，因而，常错误地选用统计分析方法。医学期刊中，误用统计学的现象十分严重，其中分析定量资料的错误中，绝大部分错误都出在不能正确识别资料的设计类型上。下面举一个实例，说明如何"通过对资料原形的转换去辨析

其真正的设计类型”的方法。

【例】某人用 t 检验分析了如下的资料，这是很不妥当的！因为它不是多个单因素 2 水平的设计。资料的原形可概述如下：

为了研究“不同药物对小鼠迟发超敏反应的影响”，研究者在表中给出的分组标志是“药物”和“剂量”两项，写在“药物 剂量”两列之下的具体内容分别是“对照 －”、“补肾药 5”、“补肾药 10”、“补肾药 20”、“Cy 0.025”、“Cy＋补肾药 0.025＋5”、“Cy＋补肾药 0.025＋10”、“Cy＋补肾药 0.025＋20”；观测的指标是“耳肿重量”；每组均为 10 只小鼠。

【分析】原作者按“药物”和“剂量”两项并列的形式制表，使人不易看出实验设计的类型。像单因素 8 水平设计问题，又像是两个单因素 4 水平设计问题或是某种多因素设计问题。这是缺乏有关设计类型概念的人们习惯的列表方式，在选用统计分析方法时将起着严重的误导作用。

仔细看看以“药物”和“剂量”为总称的这两列。似乎该实验涉及了“药物”和“剂量”两个因素，事情是否果真如此？不妨试列出由它们组合成的表格，即把“药物种类”与“药物剂量”视为两个实验因素，分别放置在表的横行与纵列上。前者的 2 个水平分别为“补肾药”与“Cy 药”，后者的 5 个水平分别为“0、0.025、5、10、20”(g /kg)，它们之间共有 10 种组合，其中只有一半反映了原先的分组，另有两种组合是重复的(即原先的对照组)，还有 4 种组合不包括在原先的设计之中，而原先两药合用的 3 个组却无法表达出来。这说明从原表中抽象出“药物”和“剂量”这样两个因素是不够正确的转换方式。事实上，原表中所反映的是两种药具有各自的用药剂量，故将“补肾药的剂量”和“Cy 药的剂量”视为两个实验因素，问题就迎刃而解了。此时，“补肾药的剂量”有 4 个水平，即“0、5、10、20”(g /kg)；“Cy 药的剂量”有 2 个水平，即“0、0.025”(g / kg)。于是，将这两个因素分别放置在表的横行与纵列上，它们之间的 8 种水平组合正好就是原作者的实验所代表的真正含义。故其本质是分别具有 2 水平和 4 水平的两个因素的水平组合，即两因素(或称 4×2)析因设计，而不是单因素 8 水平设计，也不是两个单因素 4 水平设计问题。

概括地说，正确分析定量资料的关键是：明确观测指标；弄清因素、水平及其各因素之间的水平组合；找出与资料原形所对应的真正的设计类型；考察资料所具备的前提条件；正确运用统计分析软件实现统计计算；结合专业和统计学知识作出合理的解释。

关于统计学中其他具体的学术问题，请读者仔细阅读本书各篇内容，慢慢去领悟统计学的真谛。下面让我们一同来看看本书与其他类似书籍相比，所具备的几个独到之处：

其一，用计算器和计算机两种计算工具实现统计计算，便于读者选用；

其二，针对实际问题和具体资料讲授统计方法，有利于读者提高处理实际问题的综合能力；

其三，介绍的试验设计类型多，讲解详细，具有很强的可操作性；

其四，为读者方便、成功地学会使用 SAS 软件提供了一条有效的捷径；

其五，与常用统计分析方法对应的 SAS 引导程序比较齐全，并附有程序修改指导和输出结果的解释；

其六，内容丰富、实用，使用方便。

这些特点是一般介绍 SAS 软件的书籍和以计算器为计算工具的统计学教科书所缺少的，正因如此，预计本书将产生可喜的社会效益。

本书虽然是以DOS环境下的SAS软件为计算工具，但其中的全部SAS引导程序在WINDOWS版SAS系统中仍可不加修改地被调用。所有用过WINDOWS版SAS软件的用户都知道：其中的非编程法能解决的统计计算问题所占的比例很小，绝大部分统计计算问题仍需要借助编程法来实现，即仍需要SAS引导程序。不仅如此，根据笔者的经验，要想灵活、方便地对付各种复杂的统计问题，非得采取编程法不可！

本书共分6篇，第1篇统计学基础知识与SAS软件应用技巧，介绍了统计学的基本概念和学习方法、试验设计入门、统计描述、SAS软件应用入门、编写SAS实用程序的技巧、单变量统计分析等。第2篇试验设计与定量资料的统计分析，介绍了与t检验、非参数检验、各种方差分析有关的试验设计和数据处理方法。第3篇试验设计与定性资料的统计分析，介绍了处理二维和高维列联表资料的各种统计分析方法，包括$\chi 2$检验、Fisher的精确检验、Ridit分析、秩和检验、定性资料的相关分析、线性趋势检验、Kappa检验、对数线性模型和Logistic回归模型等。第4篇试验设计与回归分析，介绍了回归分析的种类和选用方法、简单直线回归、多项式回归、简单曲线回归、非线性曲线拟合、多元线性回归、协方差分析、直接试验设计及其资料的回归分析等。第5篇生存分析，介绍了生存资料的特点、生存时间函数和生存分析方法的分类等基本概念；生存资料的非参数分析方法、COX模型和参数模型的回归分析方法与应用。第6篇多元统计分析，介绍了主成分分析、因子分析、对应分析、聚类分析、判别分析、典型相关分析。

根据教学对象的层次和学时数适当取舍内容，本书可作为研究生、本科生、大中专生的统计学教材；可作为高等院校和科研机构的教师、学者、科技人员、生物医学工作者、管理工作者等学习和应用统计方法的参考书；还可作为用SAS软件解决实际问题的实用手册。

在本书即将出版之际，谨向参加编写的全体同仁表示衷心的感谢！

由于我们水平有限，缺点和错误在所难免，敬请读者批评指正。

主　编　胡良平

2000－06－15于北京

目　录

第1篇　统计学基础知识与SAS软件应用技巧

第 2 篇　试验设计与定量资料的统计分析

第 3 篇　试验设计与定性资料的统计分析

第 4 篇　试验设计与回归分析

第 5 篇　生存分析

第 6 篇　多元统计分析

第1篇 统计学基础知识与SAS软件应用技巧

第1章 绪 论

第1节 统计学的理论基础和研究对象

统计学是运用概率论和数理统计的原理、方法研究统计研究设计，数字资料的搜集、整理、分析和推断，从而掌握事物内在客观规律的一门学科。马克思主义认为：世界是物质的，物质是运动的，运动是有规律的，规律是可以认识的。运动的形式随着时间、空间、条件的变化而变化，一切运动都是由量变到质变，反映事物变化的规律离不开统计这个工具。因此，统计学广泛运用于各行各业，从事医学研究和疾病防治工作也不例外。

概率论和数理统计是统计学的理论基础，它在不同领域的应用形成了不同的统计学科，如在医学中应用，就有医学统计学。无论在基础医学、临床医学和预防医学各个方面的科学研究以及防治工作计划的拟订和结果的正确评价，都必须进行周密的试验（或调查）设计、有计划地收集资料并进行合理的统计分析。

统计学所要研究的对象是有变异的事物，自然界的一切事物有着不同的内在规律，但由于受着许多偶然因素的影响，以致在相同的条件下，同一类事物之间会存在着差异，这种差异统计上称为变异。例如同为健康人，即使是同性别、同年龄，他们的身长、体重、血压等指标的取值都是不同的。由于事物之间有变异，研究者必须在观察一定数量的基础上进行统计分析才有价值。统计研究不是孤立地研究各种现象，而是通过一定数量的观察，从这些现象里研究事物间的相互关系，阐明事物客观存在的规律。由于统计研究对象之间存在着变异，变异的出现是由于许多内外因素偶然性的配合所致，因此，统计研究的各种现象的表现是一种随机事件。随机事件是指一次试验结果不确定，而在一定数量重复试验的条件下呈现出统计规律性的事件。科学研究的目的就是在于阐明客观存在的规律，从而通过它们对同类事物加以估计和预测，以便应用于实际，所以统计须在一定数量观察的基础上进行研究。

第2节 统计学的任务和作用

统计学的任务可概述为：①结合专业知识和具体要求进行统计研究设计（包括调查设计和试验设计），收集和整理资料；②对所收集的资料进行统计描述和处理；③对统计处理的结果进行分析和解释，根据样本资料所提供的信息推断总体的规律性，从而作出科学的结论，并用它来指导今后的实践。

统计学的作用就在于它能帮助人们有计划、有目的地进行调查研究或试验研究，合理地分析和解释试验数据，科学地揭示数据之间隐含的内在规律性。

必须强调指出的是：统计学只能帮助人们发现规律而不能创造规律。至今仍有一些人不能正确地看待统计学的作用，尤其是对试验设计的重要性认识模糊。他们不善于在试验研究开始之前就从统计学的角度去考虑应当如何确定试验因素、观测指标、受试对象（包括种类和数量），如何合理地安排试验，以便用最少的人力、物力和时间，有效地控制和估计试验误差，获得准确可靠的试验结果；而是等试验结束后，急需发表文章或参加会议时，才想到要用统计学来为他的试验数据进行“修饰”。此时，常常会出现这样的现象：由于试验缺乏完善的设计方案的指导，要么数据量不够，要么严重地违背了试验设计的基本原则，导致所收集的资料无法处理，或统计结论与专业知识自相矛盾，或结论模棱两可。更有甚者，不是根据指标的性质、试验设计的类型和研究目的有针对性地来选用统计分析方法，而是将各种统计方法一一试用，看哪一种方法计算出的结果与他所预期的结果一致，就认定哪种方法。由此而得到的科研成果或学术论文的科学性是值得怀疑的，所有尊重科学的人都决不会容忍这种现象继续蔓延下去。

我们应当清醒地认识到，运用统计方法推导出来的结论是否可靠，关键取决于以下几个方面：调查或试验设计是否周密完善、是否按设计要求进行实施；所选用的指标是否特异性和客观性强、灵敏度和精确度高；数据是否真实可靠、样本含量是否足够大；所选用的统计方法是否妥当；结果的解释是否正确。因为在运用统计学的全过程中，稍有不慎就有可能犯统计学上的四型错误（参见第 2 篇第 1 章），它们分别产生于试验设计、数据处理、统计推断和结果解释阶段。

第 3 节　统计学的主要内容

1. 统计研究设计

调查设计：指调查研究工作全过程的计划*。

试验设计：指对试验因素作合理、有效的安排，最大限度地减少试验误差，使之达到高效、快速、准确、可靠和经济的目的。

两者的区别：在调查中，研究者较被动地进行观察，只希望干扰因素的影响尽可能地减少；在试验中，研究者能较主动地安排试验因素，控制试验条件，尽可能排除或抵消非试验因素的干扰和影响。

这部分内容将在第 1 篇第 2 章中详述。

2. 统计描述（含单变量统计分析）

统计表和统计图：这是表达统计资料常用的两种方法。用统计表表达资料，简练、准确；用统计图表达资料，形象、直观。

定量资料集中趋势的度量：常用下列平均指标来描述，即算术均数、几何均数、调和均数、中位数和众数。

定量资料离散趋势的度量：常用下列变异指标来描述，即标准差、标准误差、变异系数、极差和四分位数间距。

随机变量及其概率分布：包括离散型随机变量的概率分布（如：二项分布、普阿松分布、几何分布、超几何分布等）和连续型随机变量的概率分布（如：正态分布、t 分布、χ^2 分布、F 分布、

---

* 标 1 个星号的内容未作介绍

对数正态分布、指数分布、威布尔分布等)**。

定量资料分布趋势的度量:常用的指标有分位数、偏度系数和峰度系数。

定性资料的统计描述——相对指标(包括率和比)。

以上内容将在第1篇第3章中介绍。

3. 统计分析

(1) 假设检验

关于定量资料分布类型的假设检验、定量资料方差(或方差阵)的假设检验、定量资料均数(或均值向量)的假设检验、定性资料分布情况或位置的假设检验、两种属性之间的独立性检验以及两种方法判断结果的一致性检验等。

这部分内容将在第2、第3两篇中介绍。

(2) 区间估计

置信区间的估计,即对总体参数(均数、率、方差等)进行区间估计;容许区间的估计,即对总体中一定比例的个体某指标取值范围的估计。

这部分内容将在第1篇第3章中介绍。

(3) 研究变量之间的关系

①各指标之间无自变量与因变量之分:

A. 研究变量之间的相互关系有直线相关分析、典型相关分析等;

B. 研究多个变量内部的从属关系,并寻找综合指标,降低变量的维数,其常用的方法有主成分分析、因子分析、对应分析;

C. 研究多个变量内部或多个样品之间的亲疏关系有聚类分析;

D. 研究多个变量内部的各种复杂关系有线性结构方程的协方差分析*。

②各指标之间有自变量与因变量之分:研究变量之间的依存关系有直线回归分析、曲线回归分析、多项式回归分析、多元线性回归分析、logistic 概率模型回归分析、生存资料的参数模型回归分析、COX 模型回归分析和对数线性模型分析。

(4) 判别分析

根据一些明确分类的总体所提供的信息,对未知个体的归属进行分类的判别分析。内容(3)、(4)将在第4~6篇中介绍。

第4节 学习统计方法的捷径

学习和使用统计方法的全过程可划分为以下三部分:其一,对统计学的概念和方法有一个大概的了解,以便根据具体情况正确选用统计方法;其二,正确运用统计方法处理实际资料;其三,把专业与统计知识紧密结合起来,对计算结果给出合理的解释,从而作出科学的结论。对于非统计工作者来说,第二部分是最大的障碍,常因统计计算公式复杂,计算过程繁琐而望而生畏。现在,随着计算机技术的发展,统计计算已能通过统计分析软件加以完成,统计计算不再是困扰科技工作者的难题了。

学习统计方法的捷径是:学习并掌握一个现成的统计分析软件,以便能在电子计算机上实现各种复杂的统计计算,将主要精力和时间用于学习第一、三两部分内容。本书借助国际上著

** 标2个星号的内容作简略介绍

名的统计分析系统——SAS 软件包，作为计算工具，讲述统计学的理论、方法及其应用技巧。

第 5 节　统计资料的类型

正确区分统计资料的类型是正确选用统计分析方法的首要前提。在科学研究中，统计指标常常分为定量和定性指标两大类，所谓定量指标是指对每个观察单位用计量方法测量某项指标数值大小；而定性指标是指记录每个观察单位的某一方面的特征和性质。两类指标进一步又可细分为计量、计数、名义和有序资料四类。严格地说，一谈到资料的类型，就应该是对某个具体的指标而言，因为一个较复杂的统计资料可能包括上述四种类型的资料，笼统地说，只能称之为混合型资料。现举例（表 1.1.1）说明如下。

表 1.1.1　资料类型

No.	定量指标			定性指标			
	计量资料		计数资料	名义资料		有序资料	
	X_1（年龄）	X_2（胆固醇：mmol/L）	X_3（脉搏：次/min）	X_4（职业）	X_5（血型）	X_6（疗效）	X_7（尿糖）
1	38	5.77	72	工人	A	治愈	−
2	49	4.30	69	农民	O	好转	+
3	26	7.07	81	商人	AB	无效	++
4	57	4.73	75	军人	B	恶化	+++
…	…	…	…	…	…	…	…

【说明】 计量资料的具体取值通常是正实数（零、正整数和小数），即可以取某区间内所有的值；计数资料的具体取值通常是零和正整数；名义资料的取值通常是文字、字母或代号，即使是用数字表示，也只是一种分组的标志，并不代表数量的大小；有序资料的取值与名义资料相同，只是不同取值之间有半定量的关系，可以按数量的相对大小或程度的高低排出顺序，故这种资料又称为等级资料。

第 6 节　数据结构与统计方法的匹配

在进行统计处理时，人们所面临的资料是混合型的，为便于讨论问题，不妨把任何一个完整的资料称之为数据结构。一般来说，对不同的数据结构，有相应的统计分析方法与之相匹配。下面将根据统计学中的主要内容，展示与之对应的数据结构，以便使用者在处理数据时参考。

1. Ⅰ型数据结构——只含定量资料

（1）数据结构见表 1.1.2

表 1.1.2　103 例冠心病患者的部分资料

编号	X_1（年龄）	X_2（胆固醇，mmol/L）	X_3（甘油三酯，mmol/L）	X_4（低密度脂蛋白，%）	X_5（高密度脂蛋白，g/L）
1	60	5.77	2.32	122	0.30
2	46	4.30	0.58	84	0.57
3	55	7.07	1.75	197	0.34
…	…	…	…	…	…
103	76	5.05	1.14	135	0.42

(2)统计方法的选择　如果每次只分析一个指标，可进行统计描述或单变量统计分析，也可进行区间估计或假设检验(需给定总体均数或公认的标准值)。如果每次要分析两个或两个以上指标，则可选用上述"研究变量之间的关系"中所介绍的某些方法。

2. Ⅱ型数据结构——只含定性资料

(1) 数据结构见表 1.1.3

表 1.1.3　103 例冠心病患者的部分资料

编号	X_1(性别)	X_2(高血压史)	X_3(吸烟史)	X_4(基因型 XbaI)	X_5(基因型 EcoRI)
1	男	无	无	−/−	−/−
2	女	无	无	−/−	+/−
3	男	有	无	+/−	+/+
…	…	…	…	…	…
103	男	有	有	−/−	+/+

(2) 统计分析方法的选择

表 1.1.3 的资料无法直接进行统计分析，常将它整理成列联表(见表 1.3.3 和表 1.3.4)的形式之后，再用定性资料的统计分析方法分析(如：定性资料的假设检验、logistic 概率模型回归分析、对数线性模型分析、对应分析)。

3. Ⅲ型数据结构——同时含有定量和定性资料

(1)数据结构

该结构是表 1.1.2 与表 1.1.3 的混合。

(2)统计分析方法的选择

根据研究者的需要，可分别对Ⅰ、Ⅱ和Ⅲ型数据结构进行分析。

把定性指标作为分组标志，定量指标作为观测结果(或称反应变量)，可选用定量资料的假设检验、判别分析等方法。

把定性资料数量化后看作定量资料，就将Ⅲ型数据结构转变成Ⅰ型数据结构，可选用相应的统计分析方法。如：对性别而言，可用 0 表示男性，用 1 表示女性，使之量化；水平数≥3 的定性变量的数量化方法，参见第 4 篇第 2 章第 1 节。

把定量资料离散化后看作定性资料，就将Ⅲ型数据结构转变成Ⅱ型数据结构，可选用相应的统计分析方法。如：对年龄而言，可分别按＜35 岁、35～50 岁、＞50 岁划分成青年组、中年组、老年组，使之离散化。

第 7 节　几个重要的统计名词

1. 必然事件与随机事件

在一定条件下必然发生的事件称为必然事件；而可以发生也可以不发生、可以这样发生也可以那样发生的事件称为随机事件。在医学上有很多事件都是随机事件，如病人来医院就诊，其最终转归可以是治愈，也可以是无效。因此，病人的疗效是随机事件。

2. 同质与变异

客观事物总是千差万别而各不相同，即使是性质相同的事物，就同一观察指标来看，各观