

医学统计方法 与 SAS 应用技巧

胡良平 周士波 编著

中国科学技术出版社

1

1
R195.1
HLP

11/221/2

医学统计方法 与 SAS 应用技巧

胡良平 周士波 编著

中国科学技术出版社出版

内 容 提 要

本书介绍了国际上通用统计软件包 SAS 的应用技巧和医学统计的基本方法。重点介绍如何简易、成功地使用 SAS 软件分析医学试验数据。内容不仅包括医学试验设计, 常用和多元统计分析方法的概况, 具体应用实例, 相应的 SAS 程序以及程序修改指导, 而且还有用 SAS 计算的输出结果及其解释, 统计和专业结论等。本书可作为广大医学院校师生, 卫生、生物、医学科研工作者和医务人员学习医学统计方法, 运用 SAS 解决实际问题的指南和参考书。

* * *

医学统计方法与 SAS 应用技巧

胡良平 周士波 编著

责任编辑 苗 芳 张 日

中国科学技术出版社出版(北京海淀区白石桥路 32 号)

新华书店北京发行所发行 各地新华书店经售

军事医学科学院情报研究所印刷厂印刷

开本: 850×1168 毫米 1/32 印张: 6.0 字数: 156.7 千字

1991 年 10 月第一版 1991 年 10 月第一次印刷

印数: 2,500 册 定价: 4.20 元

ISBN 7-5046-0558-1 / R · 115

前 言

到目前为止,国内介绍医学统计方法的书籍已相当多,但其中有许多是建立在手算或利用计算器计算基础之上的,虽然通俗易懂,但往往只能解决计算量较小的简单问题。近年来,随着微型电子计算机的日趋普及,关于“医学统计方法及计算机程序”的书籍也在不断涌现,为读者大大减轻了计算方面的沉重负担。因而,医用多元统计分析得到了较广泛的应用与发展。但是,因各书编写的计算机程序一般不具备“通用性”,故对读者使用和交流不便。随着改革开放向纵深发展,我国已引进了诸如“SAS”、“SPSS”、“BMDP”等国际通用的大型统计软件包,许多单位和个人利用它们解决了统计分析方面的不少问题。因这些软件包的开发、使用并不十分容易,所以往往只有少数从事统计或计算机工作的同志才会使用。如何使这些通用软件成为广大科研人员得心应手的计算工具,从而提高我国医学卫生界统计分析水平,除靠软件包的原版或译成中文的使用说明书外,还急需有一些介绍统计方法与实用程序相结合的应用性书籍,本书正是为满足这一客观需要、填补这一空白所做的一点尝试。

本书以 SAS 软件包为工具,节省了用于介绍计算方法的大量篇幅,通过具体实例,重点介绍如何简易、成功地使用 SAS 软件解决医学科研中的各类统计问题。尤其是书中“医学试验设计类型”、“一元方差分析”和“多元方差分析”等部分,集中反映了作者近几年来在医学统计教学、咨询和科研工作中如何应用 SAS 的经验和体会。全书虽然篇幅较少,但包括了常用统计分析和近代多元统计分析的绝大部分内容,具有较强的实用性。

本书共分 5 篇,第一篇 SAS 应用入门,扼要介绍 SAS 软件、SAS 程序的构成和在微机上运行 SAS 的方法,为初用 SAS 软件的读者提供了必备的概念、命令和操作方法。第二篇医学试验设

计,按各种设计类型的功能进行归纳讲述,易于理解和掌握。在介绍“常用设计类型”的基础上,增加了“具有重复测量的多因素设计类型”。此设计类型是医学科研工作者比较生疏、但又经常需要运用的。第三篇常用统计分析,包括单变量分析、线性相关和回归分析、t 检验、一元方差和协方差分析以及离散性资料的分析,其中“一元方差分析”是本篇的重点。第四篇多元统计分析,包括多元线性回归、多元方差和协方差分析、主成分和因子分析、聚类和判别分析、典型相关分析以及分类资料和失效时间资料的回归分析。每一个统计问题均按照下面几个环节进行介绍:其一,先对该统计问题作一扼要概述;其二,给出调用 SAS 软件子程序,它是把具体问题与 SAS 软件联系在一起的纽带,相当于打开 SAS 软件包的“钥匙”,使用 SAS 的捷径就在于作者已为读者配好了打开各种“锁”的“钥匙”;其三,给出子程序修改指导,读者可据此将自己的数据和有关变量代入子程序;其四,给出输出结果及其解释;其五,作出统计和专业结论。

本书几乎对所有需要进行统计分析的读者都会有所帮助,尤其适合在医学、卫生、生物等领域内从事医疗、教学和科研工作的同志了解医学统计方法和利用 SAS 分析试验数据时参考。

本书的问世,与军事医学科学院计算中心全体同志的关心和帮助是分不开的,全书由张丽媛同志录入电子计算机,为反复修改带来了很大方便;本院程海云、陈兴国、赵春华、高沛永、唐仲雄等热情地提供了许多医学试验设计类型和宝贵的试验资料,这也是本书具有独到之处的重要基础。在此,对关心、支持和帮助本书的所有同志,表示衷心的感谢!

编著者

1991 年 5 月于北京 军事医学科学院

目 录

第一篇 SAS 应用入门

第一章 SAS 简介	(1)
第一节 概述	(1)
第二节 应用 SAS 的捷径	(2)
第二章 SAS 程序的构成	(2)
第一节 概述	(2)
第二节 建立 SAS 数据集的几种方法	(4)
第三章 在微机上运行 SAS 的方法	(7)
第一节 非交互方式	(7)
第二节 交互地显示管理方式	(7)

第二篇 医学试验设计

第一章 医学试验设计概述	(11)
第一节 试验设计的要素	(11)
第二节 试验设计的原则	(11)
第二章 医学试验设计类型	(14)
第一节 考虑 1 个试验因素, 不考虑区组因素	(14)
第二节 考虑 1 个试验因素, 1 个区组因素	(15)
第三节 考虑 1 个试验因素, 2 个区组因素	(16)
第四节 考虑 $K(K \geq 2)$ 个试验因素, 不考虑区组 因素	(16)
第五节 考虑 $K(K \geq 2)$ 个试验因素, 1 个区组 因素	(21)

第三篇 常用统计分析

第一章 连续性资料的统计分析	(23)
第一节 单变量分析	(23)
第二节 线性相关和回归分析	(25)
第三节 t 检验	(31)
第四节 一元方差分析	(35)
第五节 一元协方差分析	(84)
第二章 离散性资料的统计分析	(88)
第一节 $R \times C$ 表资料的分析	(88)
第二节 分层计数资料的分析	(90)

第四篇 多元统计分析

第一章 多元线性回归分析	(97)
第一节 不筛选自变量的多元线性回归分析	(97)
第二节 筛选自变量的多元线性回归分析	(100)
第二章 多元方差分析和协方差分析	(104)
第一节 多元方差分析	(104)
第二节 多元协方差分析	(113)
第三章 主成分分析和因子分析	(117)
第一节 主成分分析	(117)
第二节 因子分析	(121)
第四章 聚类分析和判别分析	(127)
第一节 聚类分析	(127)
第二节 判别分析	(135)
第五章 典型相关分析	(143)
第六章 分类资料的回归分析	(148)

第一节	<i>LOGISTIC</i> 回归分析	(148)
第二节	对数线性模型	(155)
第七章	失效时间资料的回归分析	(159)

第五篇 VAX SAS 应用入门

第一章	VAX SAS 的基本用法	(168)
第一节	概述	(168)
第二节	VAX SAS 的运行方式	(168)
第二章	VAX SAS 设计技巧	(173)
第一节	分析指定单位、指定疾病全年各月发病情况 ...	(173)
第二节	分析任意单位、疾病全年各月发病情况	(174)
第三节	怎样构造汉字报表表头	(177)
第四节	任意两个单位疾病信息的比较	(181)
参考文献		(184)

第一篇 SAS 应用入门

第一章 SAS 简介

第一节 概述

SAS 是 statistical analysis system(统计分析系统)的缩写,美国于 1966 年开始研制 SAS,并且一直在不断更新版本,本书所使用的是最新推出的版本——SAS 6.03 版。SAS 设计的目标是使该系统在世界上几乎所有主要的计算机系统下都可以运行。它是属于第 4 代语言的软件产品,只要掌握了 SAS 应用方法,在绝大多数场合下,可以代替学习计算机高级语言和其它统计软件。SAS 已成为国际上几个最著名的通用统计软件之一,它功能齐全,便于学习和使用,是一个值得推广的统计软件。

SAS 提供的基本功能是:

- 信息的存贮和提取 SAS 可以从磁盘、磁带、终端上读取数据,建成 SAS 数据集存入 SAS 数据库中。在该库中可以方便地存取文件和数据。

- 数据的修改和处理 利用 SAS 语言和函数可以对数据进行修改,并可根据需要对数据进行剪裁建成新的 SAS 数据集。

- 书写报告功能 SAS 可以以几乎任何格式读入数据,也可以任何方式输出数据。除了固定的报表方式, SAS 支持用户设计任何格式的报表。

- 统计分析功能 对数据进行统计分析是 SAS 系统最基本和最主要的功能。它的范围包括从最简单的描述性统计到最复杂的多元统计分析。本书重点就在于介绍如何方便地应用 SAS 对医学试验数据进行统计分析,并对其输出结果作出统计学和专业方面的解释。

· 文件管理功能 SAS 系统提供一些实用程序,可以对数据集进行编辑、修改、合并,同时能对多个文件进行处理,并生成多个报表。

此外, SAS 还提供 SAS / ETS 软件,可用于商业分析、预测和制定财政计划;提供 SAS / GRAPH,可为商业和科学研究绘制各种彩色图形等。

第二节 应用 SAS 的捷径

从统计分析的角度来看,应用 SAS 的关键是要编写出一段 SAS 程序,用户通过它把数据提交给 SAS 系统,并激活 SAS 中的某个过程使之处理数据。要想编写出符合各种情况的 SAS 程序,就必须对 SAS 软件包的全貌和 SAS 语言有较详细的了解,但这决非一日之功!

为解决上述问题,本书为广大用户提供了一条应用 SAS 的捷径,即针对各种常见的统计学问题和实例,编好了一系列相应的 SAS 程序。用户只需“对号入座”,对现成的 SAS 程序略加修改,便可成为用户自己的一把钥匙,用它可以很方便地指导用户使用 SAS。书中不仅有各种实例、SAS 程序和 SAS 程序修改指导(主要修改内容是把 SAS 程序中的旧数据删除,换上用户的新数据),还有 SAS 程序的运行结果及其解释,它将成为读者迅速成功地应用 SAS 的得力助手。

第二章 SAS 程序的构成

第一节 概述

为了应用 SAS,必须编写出由 SAS 语言组成的程序。SAS 程序一般由 SAS 数据步 (SAS data step) 和 SAS 过程步 (SAS procedure step) 组成。

SAS 数据步的主要功能是,按照 SAS 系统的规定,把要分析、处理的数据建成 SAS 数据集 (SAS data set),以便能够存取由一条条记录组成的特定的数据文件。

SAS 过程步的功能是根据对数据分析的要求,调用相应的过程来处理已建立的 SAS 数据集中的数据。

下面以建立 1 个因变量 Y 与 4 个自变量 X1~X4 之间的多元线性回归方程为例,引入对 SAS 程序的基本概念。设试验中取得的数据如表 1.2.1。

表 1.2.1 多元线性回归分析的数据格式

编号	X1	X2	X3	X4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
...	...	...	...	...	...
13	10	68	8	12	109.4

首先要建立适于 SAS 分析的 SAS 数据集,其程序如下:

```
DATA A;
  INPUT X1-X4 Y;
  CARDS;
    (输入全部数据)
    7    26    6    60    78.5
    1    29    15   52    74.3
    ...  ...   ...   ...   ...
    10   68    8    12   109.4
  ;
RUN;
```

数据步第 1 个语句必须为 DATA 语句,它是 SAS 数据步开始的标志。语句 DATA A 表示要建立 1 个名字为 A 的 SAS 数据集。

第 2 个语句是 INPUT,它用来描述所生成的 SAS 数据集中每个记录的定义,即指出记录中每个字段的名称、类型、长度及位置。在本语句中将每行数值的前 4 个依次送给变量 X1、X2、

X3、X4, 将第 5 个值送给 Y。在将建立的 SAS 数据集 A 中, 每行前 4 列数代表 4 个自变量的观测值, 第 5 列是因变量 Y 的观测值。

第 3 个语句是 CARDS 表示下行将通过终端输入的试验数据。即将上述数据一行一行从终端输入, 最后一行输完后, 加“;”号表示数据输入终止。

RUN 语句执行后, 即生成我们所要求的 SAS 数据集 A。

下面开始 SAS 过程步, 它的程序为:

```
PROC REG;
```

```
MODEL Y=X1-X4;
```

```
RUN;
```

SAS 过程步的第 1 句必须是 PROC 开头, 后面紧跟所要调用的 SAS 过程名, 这里 REG 为进行多元线性回归分析的过程名。下一个 MODEL 语句是限定 REG 功能的, 表明是建立 Y 与 X1~X4 之间的多元线性回归。RUN 语句, 则执行“用 SAS 数据集 A 中的数据进行多元回归分析”这项任务。

从上面给出的一个最简单最基本的 SAS 程序, 可以发现每个 SAS 语句都必须以“;”号结束。每个程序行中可以写多条 SAS 语句, 只要语句之间用“;”号隔开即可。此外在每个 SAS 程序中可以有多个 SAS 数据步和多个 SAS 过程步。如果有多个 SAS 数据步, 则意味着建立了多个 SAS 数据集, 因此在过程步调用某过程时, 要指明处理哪个数据集。如“PROC REG DATA=OLD;”, 用 DATA=OLD 指明对名为 OLD 的 SAS 数据集进行多元线性回归分析。

第二节 建立 SAS 数据集的几种方法

上面介绍的是从终端上输入数据建立 SAS 数据集的方法。如果遇到数据量较大的情况, 用这种方式显然不够方便。此时, 可以用 SAS 自身的编辑功能或用 EDLIN、WORDSTAR 等编辑软件将数据事先写入一个数据文件, 如将上述数据存入

C:ABC.DAT 中,然后再用下述语句建立 SAS 数据集 A。

```
DATA A;  
  INFILE 'C:ABC.DAT';  
  INPUT X1-X4 Y;  
RUN;
```

这里用 INFILE 语句,告诉 SAS 系统从 C:ABC.DAT 文件中取数据,然后按 INPUT 语句的描述建立 SAS 数据集 A。

如果数据是存放在 DBASE 数据库 DDB 中,那么可由下述命令:

```
.USE DDB  
.COPY TO ABC.DAT SDF
```

将数据写入 ABC.DAT 中。再用前述过程,建成 SAS 数据集 A,即可用 SAS 程序对它进行处理。

如果想建立只含 X1、X2、X3 与 Y 的数据集,那么可以在 SAS 数据集 A 的基础上来进行,应用下述语句:

```
DATA NEWA;  
  SET A;  
  KEEP X1-X3 Y;  
RUN;
```

SET 语句说明是从 SAS 数据集 A 中生成新的 SAS 数据集 NEWA,KEEP 语句表明在 NEWA 中只保留 X1、X2、X3、Y 这 4 个变量,这样就生成了一个新的 SAS 数据集 NEWA,数据直接从 A 中取来,不必再重新输入。

如果需要对因变量(也可对某自变量)取一个对数变换,可以由下述方法实现。

```
DATA NEWB;                                PROC REG;  
  SET A;                                  MODEL Z=X1-X4;  
  Z=LOG(Y);                               RUN;  
RUN;
```

用赋值语句产生新变量 Z, 然后再求 Z 与 X1~X4 之间的多元线性回归方程, 这样就建立了 LOG(Y) 与 X1~X4 之间的多元线性回归方程。

上面各种方式建立的 SAS 数据集都属于临时数据集, 即它在 SAS 运行过程中产生, 退出 SAS 系统时则不再存在。

如果建立的数据集, 包含数据量大, 且要多次调用进行多种统计分析, 那么建立一个永久的 SAS 数据集是有意义的。

建立永久 SAS 数据集的基本程序如下。

```
LIBNAME A1 'C:\SAS\';
```

```
DATA A1.ABC;
```

```
INPUT X1-X4 Y;
```

```
CARDS;
```

```
(输入全部数据)
```

```
;
```

```
RUN
```

这里用 LIBNAME 语句将库关联名 A1 与目录 C:\SAS\ 联接起来, 再用 DATA 语句表明在目录 C:\SAS\ 下建立 SAS 数据集 ABC。当本程序段执行后, 1 个永久的文件 ABC.SSD 就写入 C:\SAS\ 目录下, 此文件只供在 SAS 系统下调用。

建成了永久 SAS 数据集后, 要对它进行多元线性回归分析, 可用下述程序。

```
LIBNAME A1 'C:\SAS\';
```

```
PROC REG DATA=A1.ABC;
```

```
MODEL Y=X1-X4;
```

```
RUN;
```

显然, 用永久数据集的方法对大量数据多次进行调用分析是十分方便的。

第三章 在微机上运行 SAS 的方法

在微机上运行 SAS 主要有两种方式,交互地显示管理方式和非交互方式。

第一节 非交互方式

首先将 1 段写好的 SAS 程序以 1 个文件的方式存到你的目录下,设该文件为 AAA.PRG, 存在 C:\SAS\下,则运行该程序的步骤如下:

```
C>CD SAS (或 C>CD\SAS)
```

```
C:\SAS>SAS AAA.PRG (或 C>SAS AAA.PRG)
```

回车后,SAS 系统执行 AAA.PRG 程序。执行完毕后,屏幕上显示两个 NOTES,说明 SAS 软件的版权及属性。同时告知源程序语句来自 AAA.PRG,运行过程记录存在 AAA.LOG 中,该程序计算结果存在 AAA.LST 中。

使用这种方式运行 SAS 作业的优点是比较方便,运行过程和计算结果都由 SAS 系统自动存盘。一旦 SAS 程序编写得不正确时,就得不到正确结果。为此须将 AAA.LOG 文件调出,看错误出在哪里,然后修改程序。再按上述步骤重复工作,直至获得正确结果。

第二节 交互地显示管理方式

所谓交互地显示管理方式是指用户在显示管理系统 (简称 SAS DMS)支持下,与 SAS 系统之间通过发送命令来运行 SAS 任务。我们仍以前面介绍的多元线性回归分析的数据为例,介绍应用 SAS DMS 交互运行 SAS 程序的一般方法。假定 SAS 软件被安装在硬盘 C 的子目录 SAS 下,发命令:

```
C>CD SAS (或 C>CD\SAS)
```

回车后进入 SAS 子目录,在 C:\SAS>(或 C>)提示下键入 SAS。

```
C:\SAS>SAS (或 C>SAS)
```

回车后稍等片刻,便进入 SAS DMS 状态。此时显示屏上出现 3 个最常用窗口,从上到下依次为输出窗口(OUTPUT),记录窗口(LOG)和程序编辑窗口(PROGRAM EDITOR,缩写为 PGM)。如下所示:

OUTPUT

Command === >

LOG

Command === >

PROGRAM EDITOR

Command === >

00001

00002

00003

此时用户可在 PGM 窗口输入、编辑上述 SAS 程序,为扩大视野,可先按 F7 功能键使光标所在的编辑窗口充满整个屏幕(若再次按 F7 键,窗口恢复原来的大小),然后依次输入程序如下:

Program Editor

Command === >

00001 DATA A;

00002 INPUT X1-X4 Y;

00003 CARDS;

00004 7 26 6 60 78.5

00005 1 29 15 52 74.3

·
·
·

00017;

```
00018 RUN;  
00019 PROC REG;  
00020 MODEL Y=X1-X4;  
00021 RUN;  
00022
```

|

当程序输入完毕后,要检查一下有无错误。此时要用4个光标移动键,根据要求上下左右移动光标,每用PgUp键1次向上卷动屏幕1页,每用PgDn键1次向下卷1页。若程序中某一句有错误字符,可先移动光标到该字符之后,然后用←键删除;若某处有遗漏字符,移动光标到某处之后,先按INS键进入插入状态,然后输入所缺字符,若需插入行,在行号中键入I,回车后显示空行。检查完毕后,欲执行该程序,按F10键就将上述程序提交SAS系统执行。执行时程序行从PGM窗口消失,运行中的有关情况和出错信息立即在LOG窗口显示,而多元线性回归分析的结果则在OUTPUT窗口显示。

如果程序中存在错误,系统不能正常执行,得不到输出结果。为改正错误,先按F3键进入LOG窗口,再按F7使LOG窗口充满全屏,移动光标显示全部信息,找到错误原因。然后按F6键回到PGM窗口,按照前述方法移动光标改正程序。完成后,按F10键提交SAS执行。如果仍存在错误,重复上述步骤,直至正确执行为止。

要想看到输出结果,先按F4键进入OUTPUT窗口,再按F7键使OUTPUT窗口充满全屏。运用移动光标键和翻页键,可以查看全部输出信息。

要得到输出结果的硬拷贝,可采用下述方法:

(1)将输出结果存入文件,为此在OUTPUT窗口的命令行上键入: