附光盘

医学科研数据的处理与分析方法

主编 陈大方
陈常中

北京大学医学出版社

医学科研数据的处理与分析方法

主　编

陈大方　北京大学公共卫生学院

陈常中　哈佛大学医学院

北京大学医学出版社

YIXUE KEYAN SHUJU DE CHULI YU FENXI FANGFA

图书在版编目（CIP）数据

医学科研数据的处理与分析方法/陈大方、陈常中主编．—北京：北京大学医学出版社，2006.3
ISBN 7-81071-903-3

Ⅰ.医… Ⅱ.陈… 陈… Ⅲ.医学统计—统计分析—应用软件，SAS Ⅳ.R195.1-39

中国版本图书馆 CIP 数据核字（2006）第 005494 号

医学科研数据的处理与分析方法

主　　编：陈大方　陈常中
出版发行：北京大学医学出版社（电话：010-82802230）
地　　址：（100083）北京市海淀区学院路 38 号　北京大学医学部院内
网　　址：http://www.pumpress.com.cn
E - mail：booksale@bjmu.edu.cn
印　　刷：莱芜市圣龙印务书刊有限责任公司
经　　销：新华书店
责任编辑：冯智勇　**责任校对**：金彤文　**责任印制**：张京生
开　　本：787mm×1092mm　1/16　**印张**：16.5　**字数**：412 千字
版　　次：2006 年 5 月第 1 版　2006 年 5 月第 1 次印刷　**印数**：1—2500 册
书　　号：ISBN 7-81071-903-3/R·903
定　　价：59.90 元

序

在数据处理和统计分析领域，SAS 系统被誉为国际上的标准软件系统。由于其强大的数据管理、统计功能与绘图功能，SAS 被广泛应用于包括医学研究在内的众多领域。SAS 系统是从大型机上的系统发展而来，在设计上也是完全针对专业用户进行设计，因此其操作至今仍以编程为主，人机对话界面不太友好，并且在编程操作时需要用户对所使用的统计方法有较清楚的了解。学习与掌握 SAS 软件过程是艰苦的，最初的阶段常会使人感到气馁。我本人是非统计学专业人员，对 SAS 常有着“爱恨交加”的情感，爱其功能之强大，恨其操作之繁琐。

近日，读陈大方与陈常中联袂之作《医学科研数据的处理与分析方法》。循书中要点，几经上机尝试，颇有感触。只需应用作者介绍的 SAS 宏程序，即可方便地进行数据的整理、检错、变量变换和分析，可省去复杂的 SAS 编程，运算结果简洁明了。此外，基于作者在处理医学数据方面的丰富经验，结合具体实例，为读者系统地介绍了从数据整理到统计分析技术与结果表达等方面的方法与技巧。

读罢此书我总的感觉是，SAS 这个统计软件界的“巨无霸”变得更易于操作了，“爱与恨”的天平倾向与爱多于恨了。书中各篇与章节排序，内容安排，所举实例都直接针对医药卫生领域科研人员的需求。实用性很强，是一部很有应用价值的参考书。

胡永华

2006 年 5 月 10 日

前　言

为了帮助广大医学工作者得心应手地处理和分析身边的医学科研数据，我们编写了《医学科研数据的处理与分析方法》一书，该书把医学统计学的理论方法与SAS统计软件功能有机巧妙地结合在一起，具有统计软件与数据处理分析方法指导的双重功能。因此，它不仅是一本具有数据处理与统计分析软件功能的应用型参考书，更是一本如何进行数据处理与统计分析的方法学指导型参考书，是作者十余年从事大量医学科研数据管理与分析的经验积累。本书最大的特点是把医学科研数据处理技巧和统计分析的原理、方法通过SAS软件平台所编写的SAS宏程序并应用丰富的实例进行介绍；其次，通过SAS宏程序对SAS原程序输出结果的加工处理，使最后的输出结果只保留关键的统计参数，避免了应用SAS原程序输出的冗长、复杂、难懂的统计参数，是SAS软件更高层次的应用；此外，本书所编写的SAS宏程序具有容易掌握使用、高效率处理和分析数据的特点，只需将宏命令中所设定的变量参数进行简单替换，不需要读者自己反复编写SAS程序就可以对新选定的变量进行统计分析。

全书共分三篇十三章。第一篇医学科研数据的处理方法，共设四章，主要介绍如何对医学科研数据进行整合、检错及变量处理。第二篇医学科研数据的分析方法，共设六章，主要介绍如何针对医学科研数据选择不同统计方法进行描述和分析。第三篇文章实例分析，共设三章，主要介绍我们这个学组近五年发表在SCI期刊（影响因子在2～35之间）的9篇文章的数据分析思路以及文章结果的产生过程。为了便于读者学习，书中使用的SAS宏程序以及书中介绍的例题全部刻写在附带的光盘上，以供练习使用。

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。

编　者

2006年3月于北京

目　录

第一篇　医学科研数据的处理方法

第二篇　医学科研数据的分析方法

第三篇　文章实例分析

第一篇　医学科研数据的处理方法

任何医学科研课题的实施都可以分为四步，包括课题的设计、科研资料的收集、科研资料的整理和对科研资料的最后统计分析，四个步骤相互联系，缺一不可。其中课题的设计是整个科研工作的基础，是保证课题能否成功的关键，在设计时应当对以后三个步骤进行周密的考虑。科研资料的收集则是依据科研设计的各项要求在整个研究过程中自始至终地认真贯彻执行，确保所收集的资料质量可靠，数据真实可信。而对科研资料的整理，一方面是对在研究过程中所收集到的资料的质量控制，可以及时发现资料中存在的问题，比如数据中是否出现空号、重码、极端值、缺失值以及逻辑错误，并将所发现的问题及时反馈查证，决定合理取舍；另一方面，根据不同统计分析模型对数据的不同要求，需要对数据中的变量和观察记录进行适当变换，以保证最后分析时所得到的结果更科学，更具有说服力。因此，数据处理是科研实施过程中和数据最后分析前非常重要的一个环节。而对科研资料的最后统计分析要求依据研究的目的和研究资料的属性选择适当的统计分析方法或模型对数据进行统计分析，既不夸大统计结果对研究事件本身所具有的影响，又不缩小统计结果对研究事件本身所起的作用，真正保证所得到的统计结果能够说明研究事件本身所需要说明的问题。在第一篇中，我们将主要介绍对数据的处理方法，包括如何熟悉数据，如何对不同多个数据进行连接和合并，如何处理数据中的错误，如何对数据中的变量进行分析前的变换，如何对数据中不同变量进行描述，以及如何对家系结构等特殊数据进行处理与描述。

第一章　多个数据集的横向检查

一个大型研究项目所收集的大量数据，无论是对于数据管理还是数据分析都是一个挑战。由于调查项目的广泛，数据一般被录入到多个数据文件中，如问卷数据是单独的一个文件（QUES），体检数据是单独的一个文件（EXAM），实验室数据又是单独的一个文件(LAB)。而且调查的时间或/和地点的跨度使得同一种数据可能分多批或多处录入而生成多个子文件，如问卷数据 QUES 文件可能分 QUES1、QUES2、QUES3 等等。再加上调查过程当中，调查方案可能会根据具体情况有相应的调整，因此不同时期或不同地点生成的数据文件结构可能存在差别。如何快速有效地了解、熟悉并处理多批次多种类的数据文件是数据管理工作的一个很大的挑战。本章介绍四个 SAS 宏程序专用于对所有数据文件的变量名和观测记录进行比较、统计和报告。

第一节　多个数据集变量的汇总报告

%dtsvchk（DATA=）宏程序的应用

一、程序说明

项目收集的研究对象的数据一般被录入放到多个数据文件中。项目数据管理中的变量名管理一般要求：①这些含不同信息的数据文件除研究对象编号变量名外，没有其他重名的变量，以免在数据横向合并时出现错误；②含相同信息的不同批次的数据文件所有变量名及变量类型均需相同，以免在数据纵向合并时造成不必要的错误。对多个数据集中的变量的分布情况进行检查是数据管理与分析中必不可少的一步。

%dtsvchk（）宏程序用来观察多个数据集中变量的分布情况，了解哪些变量是哪个数据集单独所有，哪些变量是几个数据集共同拥有及不同数据文件中的同名变量类型是否相同。

DATA=	所要观察的SAS数据集名。

二、实例

【例一】 运用%dtsvchk（）宏程序，观察 BAT1 目录下SAS 数据集：REGIS1、EXAM1、QUES1、SPIR1、LABG1 中所包含的变量情况。

```
libname B1 'c:\SASMAC\BAT1';
%inc 'c:\SASMAC\EPI. MAC';
%dtsvchk (DATA=b1. regis1 b1. exam1 b1. qusel b1. spir1 b1. labg1);
```

【输出结果】 程序运行后，SAS OUTPUT 窗口显示：

结果一

Output of %dtsvchk（）: Variables Report for Datasets ===>

	# of Variables
A: B1. REGIS1	6
B: B1. EXAM1	5
C: B1. QUES1	11
D: B1. SPIR1	4
E: B1. LABG1	8

结果二

Variable name by datasets:

```
A11        - - - - E
A12        - - - - E
A21        - - - - e
A22        - - - - e
A31        - - - - e
A32        - - - - e
AGE        A - - - -
ALCOHOL    - - C - -
COUGH      - - C - -
DBP        - B - - -
EDU        - - C - -
FEV1       - - - D -
FMYID      a - - - -
FMYTYPE    A - - - -
FVC        - - - D -
HEIGHT     - B - - -
NID        A - - - -
OCCU       - - C - -
PHLEGM     - - C - -
PSMK       - - C - -
SBP        - B - - -
SEX        A - - - -
SMKAMT     - - C - -
SMOKE      - - C - -
SOB        - - C - -
SUBJ       a b c d E
TDATE      - - - d e
WEIGHT     - B - - -
WHEEZE     - - C - -
A B C ... index the dataset name
if UpperCase (e. g: A), means as a numeric in the index dataset (A)
if LowerCase (e. g: f), means as a character in the index dataset (F)
        ****** The END of %dtsvchk () *******
```

【结果解释】 在以上输出结果中，结果一表示每个数据集中所包含的变量的数目，A代表 REGIS1 数据集，该数据集里有 6 个变量。B 代表 EXAM1 数据集，该数据集里有 5 个变量，依此类推。结果二列出这些数据集所包含的所有变量名，并列出了每个变量出现在哪些数据集（A、B、C、D、E）中及变量类型是数据型还是字符型。大写 A、B、C、D、E 表示此变量在该数据文件中是数据型，小写表示字符型。如变量“SUBJ”出现在 a、b、c、d、E 中，表示在 E（数据文件 LABG1）中 SUBJ 是数据型，在其它数据文件中则都是字符型。

【例二】 运用%dtsvchk () 宏程序检查 BAT1 子目录下 LABG1 与 BAT2 子目录下 LABG2 两数据集中所包含的变量情况。

```
libname B1 'c:\SASMAC\BAT1';
libname B2 'c:\SASMAC\BAT2';
%inc 'c:\SASMAC\EPI. MAC';
%dtsvchk (DATA=b1. labg1 b2. labg2);
```

【输出结果】

```
Output of %dtsvchk (): Variables Report for Datasets ===>

                    # of Variables

A: B1. LABG1              8
B: B2. LABG2              8

Variable name by datasets:
A11      A   b
A12      A   b
A21      a   b
A22      a   b
A31      a   b
A32      a   b
SUBJ     A   b
TDATE    a   b
A B C... index the dataset name
if UpperCase (e. g: A), means as a numeric in the index dataset (A)
if LowerCase (e. g: f), means as a character in the index dataset (F)
        ****** The END of %dtsvchk ()  *******
```

【结果解释】 以上结果显示 B1. LABG1 与 B2. LABG2 变量名均相同，但变量 SUBJ、A11、A12 的变量类型则不同。

%dtsvchk () 宏程序需列出各数据文件名。当对同一目录下的所有数据文件进行分析，而这些数据文件又都放在同一目录下时，可使用更简单的宏程序%sitevchk (SITE)，其中参数 "SITE" 即指向该子目录的 "libname"。

%sitevchk (SITE) 宏程序的应用

如果给定 "SITE" 即SAS的LIBNAME语句指向某一目录的SAS路径名，无需逐一列出每一个数据文件名，%sitevchk (SITE) 将会自动发现在该目录下的所有数据文件，并对其进行相应的处理。

【例三】 运用%sitevchk () 宏程序检测 "BAT1" 子目录下所有数据集中所包含的变量情况，程序调用如下：

```
libname B1 'c:\SASMAC \ BAT1';
%inc 'c:\SASMAC\EPI. MAC';
%sitevchk (B1);
```

三、分析与讨论

一般要检查一个数据集中的变量是采用 SAS 的 PROC CONTENTS 过程，SAS 没有专门的 PROC 过程来比较多个数据集变量名，只能用 PROC CONTENTS 程序分别对每个数据集进行分析，再对这些输出结果作相应的比较，整个过程十分复杂。%dtsvchk（）与%sitevchk（）宏程序大大简化了此过程，输出结果也十分简洁明了，能对多个数据集所有变量的名称和类型很快了如指掌。

第二节　多个数据集观测对象的汇总报告

%dtsrpt（DATA=，ID=，PRINTID=）宏程序的应用

一、程序说明

如果项目收集的研究对象的数据被放在多个数据文件中，那么数据分析前既需了解单个数据文件的记录数、变量数和唯一编号数（或称研究对象数），又需了解研究对象资料的完整性，即研究对象编号在各数据文件中的分布情况。%dtsrpt（）宏程序即为此目的而设计。现将该宏程序参数说明如下：

DATA=	所要分析的多个 SAS 数据集名。
ID=	各数据集里用于区分研究对象的编号变量名。
PRINTID=	为输出选择项，默认值=F，表示不打印每一研究对象（编号）数据完整情况；若置 PRNITID=T，则打印出每一编号在各数据文件中出现的情况。

二、实例

【例一】 试运用% dtsrpt（）宏程序统计 BAT1 子目录下 SAS 数据集 REGIS1、EXAM1、QUES1、SPIR1、LABG1 中观察对象资料完整性情况。

```
libname B1 'c:\SASMAC\BAT1';
%inc 'c:\SASMAC\EPI. MAC';
% dtsrpt (DATA = b1. regis1 b1. exam1 b1. ques1 b1. spir1 b1. labg1, ID = subj,
        PRINTID=t);
```

【输出结果】 程序运行后，SAS OUTPUT 窗口显示：

结果一

```
Datasets Report by %dtsrpt () ===>
```

	# Variables	# Observations	# Unique _ SUBJ
A: B1. REGIS1	6	435	428
B: B1. EXAM1	5	431	421
C: B1. QUES1	11	427	427
D: B1. SPIR1	4	384	377
E: B1. LABG1	8	428	428

结果二

# of Subjects in # of Datasets

428	1
2	2
55	3
371	4

结果三

Subjects in Datasets

428	----E
1	A-C--
6	A-C D-
1	A B---
49	A B C--
371	A B C D-

结果四

SUBJ In data

00001 A B C D-
00002 A B C D-

00008 A B C D-
00009 A B C D-
00010 A B C--
00011 A B C--
00012 A B C D-

00427 A B C D-
00428 A B C D-
1 ----E

```
10      ----E
100     ----E
101     ----E
-----
97      ----E
98      ----E
99      ----E
*** END of %dtsrpt ()  ***
```

【结果解释】 结果一列出了每个数据集中分别所包含的变量数、记录数和研究对象数，并给定一字符（A、B、C、D、E）分别代表每个数据集，如 A 代表 REGIS1，B 代表 EXAM1。研究对象数不同于记录数。研究对象数指的是编号或样本数，一个编号代表一个研究对象，如一个编号在一个数据库中有多条记录，则记录数大于编号数。数据集 REGIS1 有 435 条记录但只有 428 个编号，说明有记录编号重复。

结果二说明有 428 个研究对象存在于 1 个数据集中，55 个研究对象存在于 3 个数据集中，371 个研究对象存在于 4 个数据集中。

结果三进一步列出有多少研究对象具体存在于哪些数据集中。有 428 个研究对象只存在于数据集 E 中；1 个研究对象只存在于数据集 A 和 C 中；371 个研究对象存在于数据集 A、B、C、D 中。

结果四详细列出每个研究对象具体存在于那些数据集中。如果 PRINTID=F，则不输出结果四。

【例二】 同种数据库分批录入时，往往会因为多种原因造成后面一批数据文件中也含有前一批数据文件所录入的一些数据，可用%dtsrpt () 检查两批数据调查对象编号是否有重叠情况。试运用%dtsrpt () 宏程序检查 BAT1 子目录下数据集 REGIS1 与 BAT2 子目录下数据集 REGIS2 观察对象重叠情况。

```
libname B1 'c:\SASMAC\BAT1';
libname B2 'c:\SASMAC\BAT2';
%inc 'c:\SASMAC\EPI. MAC';
%dtsrpt (DATA=b1. regis1 b2. regis2, ID=subj);
```

【输出结果】 程序运行后，SAS OUTPUT 窗口显示：

```
Datasets Report by %dtsrpt () ===>
                    # Variables  # Observations   # Unique _ SUBJ
A: B1. REGIS1          6            435              428
B: B2. REGIS2          6            397              390
# Subjects in # Datasets

        818          1
# Subjects in Datasets

        390    - B
```

```
        428   A -
*** END of %dtsrpt ()   ***
```

【结果解释】由以上结果可知，没有任何编号同时出现在数据集 A（表示 B1. REGIS1）与 B（表示 B2. REGIS2）中。

%dtsrpt () 宏程序需列出各数据文件名。当对同一目录下的所有数据文件进行分析，而这些数据文件又都放在同一目录下时，可使用更简单的宏程序：%siterpt (SITE, ID=, PRINTID =)。

%siterpt (SITE, ID=, PRINTID=) 宏程序的应用

给定“SITE”即 SAS 的 LIBNAME 语句指向一目录的 SAS 路径名，无需逐一列出每一数据文件名，%siterpt () 自动发现在该目录下的所有数据文件。

【例三】 运用%siterpt () 宏程序观察“BAT1”子目录下所有数据集观测对象的记录数情况，程序调用如下：

```
libname B1 'c:\SASMAC\BAT1';
%inc 'c:\SASMAC\EPI. MAC';
%siterpt (B1, ID=id);
```

三、分析与讨论

一般要检查一批数据研究对象资料完整性的方法是：从每个数据文件中取出编号，并生成一个变量对每个数据文件进行标记，再将它们按编号合并，对每个数据文件的标记变量及其组合进行统计。如数据中有重复观察的记录，则还需要做进一步的处理。%dtsrpt () 与%siterpt () 大大地简化了上述操作，而且功能十分强大，既可以用来检查同一批数据中研究对象资料的完整情况，也可以帮助我们检查多批数据间编号的重码情况。通过设置 PRINTID=T，可打印出每个编号在各数据文件中的分布，从而帮助我们找出需核查的编号。如例一，从结果四可看出数据集 E 的编号变量（SUBJ）是数据型，而其它数据集 A、B、C、D 的编号变量（SUBJ）是字符型，取值不同，造成它们间编号无重叠，结果四列出每一编号缺哪些数据，从而可依此核对原始记录。