基于 SAS 软件实现随机抽样及应用

胡 完[1],胡良平[1,2*]

(1. 军事医学科学院生物医学统计学咨询中心,北京 100850;
2. 世界中医药联合会临床科研统计学专业委员会,北京 100029
* 通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 本文的目的是帮助读者方便快捷地运用 SAS 软件中的 SURVEYSELECT 过程实现随机抽样。首先,对 SURVEYSELECT过程及 SAS 数据集 Customers 进行了简单介绍。接着,结合简单随机抽样、分层随机抽样和控制排序分层随机抽样,介绍了随机抽样的 SAS 实现方法。读者只需要修改本文中所呈现的 SAS 程序中的少量参数,就可很方便地使用 SAS 软件实现随机抽样任务。事实说明,尽管 SAS 软件非常难学难用,但借助现成的 SAS 程序,可以轻松自如地解决很多具体问题。

【关键词】 SAS 软件;SAS 数据集;SAS 过程;简单随机抽样;分层随机抽样

中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2016.05.002

How to implement random sampling and application based on SAS software

HU Wan[1], *HU Liang-ping*[1,2*]

(1. *Consulting Center of Biomedical Statistics, Academy of Military Medical Sciences, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
* *Corresponding author: HU Liang-ping, E-mail: lphu812@sina.com*)

【Abstract】 The purpose of this article is to help readers to use SURVEYSELECT procedure in SAS software to implement random sampling fast and conveniently. Firstly, introducing SURVEYSELECT procedure and SAS data set customers. Then, introducing how to perform random sampling based on SAS combined with simple random sampling, stratified random sampling and control sorting stratified random sampling. The readers can finish their own random sampling task by using SAS software easily through modifying a few parameters in the SAS programs presented in this article. The fact is that despite the SAS software is very difficult to learn and use, but the users can solve many specific problems with a ready-made SAS program.

【Key words】 SAS software; SAS data set; SAS procedure; Simple random sampling; Stratified random sampling

1 在“科研方法专题”中为何要使用统计软件

本刊前两期(即 2016 年第 3 期和第 4 期)的“科研方法专题”[1-7]详细介绍了进行生物医学和临床科学研究时不可缺少的基础知识,即“三要素”和“四原则”。这些内容似乎仅仅与“文字描述”有关。然而,在确定受试对象的数量时,涉及到“样本含量估计”的技术问题;在从总体中选取受试对象时,涉及到“随机抽样”的技术问题;在将已选取的受试对象分配到不同的对比组中时,涉及到“随机分组”的技术问题。这里提及的三个“技术问题”,仅基于“文字描述”是得不到很好解决的,需要借助计算公式或查阅相应的表格方可实现。然而,在计算机已相当普及的今天,利用现成的统计软件方便快捷地实现前述的三个技术问题,则是大势所趋。本文在已介绍 SAS 软件基础知识和用法入门的基础上,介绍使用 SAS 软件实现随机抽样的基本方法。

2 与随机抽样有关问题的概述[8]

在 SAS 软件中,SURVEYSELECT 过程提供了各种各样的概率随机抽样方法[9]。这个过程既可以进行简单随机抽样,也可以进行包括分层抽样、整群抽样、不等概率抽样在内的复杂多阶段设计的抽样。在概率抽样中,调查总体中的每个抽样单位都有一个已知的共同的概率被抽取。概率抽样的这种属性避免了选择偏倚,同时可让研究者根据统计理论和样本信息对调查总体做出有效推断。

用 SURVEYSELECT 过程选择样品并构成样本,在所输入的 SAS 数据集中需要包括抽样框,抽样框是包含全部抽样单位的目录性清单。抽样单位是个体观测者或观测组(群组)。用户可以指定抽样方法、所需样本大小或抽样比例和其他一些选择参数。用 SURVEYSELECT 过程抽取样本后,产生包含抽样单位、抽样概率以及抽样权重的输出数据集。当进行多阶段抽样时,对每一个抽样阶段都需要调用这个过程来设定抽样框和选择参数。在等概率抽样中,抽样框或同一层内的每一抽样单位都有相同的概率被抽取。在等比例(PPS)抽样中,每一抽样单位被抽取的概率和它所在层的大小成正比。

SURVEYSELECT 过程提供的等概率抽样方法

有:①简单随机抽样(无放回);②无限制随机抽样(有放回);③系统随机抽样;④序贯随机抽样;⑤贝努利抽样。

SURVEYSELECT 过程提供的等比例(PPS)抽样方法有:①无放回 PPS 抽样;②有放回 PPS 抽样;③PPS 系统抽样;④基于 PPS 算法从每层抽取两个单元;⑤最小放回 PPS 序贯抽样。

该过程采用快速、高效的算法来实现这些抽样,它对于大的输入数据集或者抽样框表现很好。

SURVEYSELECT 过程通过在层内选择独立样本来进行分层抽样,层内个体不重叠出现在调查总体的亚组中。分层可控制各层样本大小,广泛应用于调查个体多样的实践中。例如通过分层可以保证感兴趣但样本量小的亚组有足够的样本量,或通过分层提高对总体估计的精度。在系统抽样或序贯抽样中,SURVEYSELECT 过程也可按控制变量的排序来额外控制隐性分层因素在层内的分布。

对于分层抽样,SURVEYSELECT 过程提供了分配各层样本大小的调查设计方法。可用的分配方法包括按比例分配、Neyman 分配和最优化分配。最优化分配在考虑层的大小、成本和方差情况下,在可用资源内使估计精度最大化。

SURVEYSELECT 过程提供重复抽样,当总样本是由一组相同个体组成时,可用相同的方法抽取每一个体。用户可以利用重复抽样来研究变量的非抽样误差,例如不同面试官面试结果的变异性。用户也可以用重复抽样结合样本大小来估计标准误以及执行各种重复采样和仿真任务。

3 实施抽样研究所需要数据集的概况[3]

使用统计软件实施随机抽样的前提是要创建由拟被抽取样本的全体(即抽样总体)构成的数据集。下面借用 SAS 软件 SURVEYSELECT 过程的帮助信息中介绍的一个例子,来讲解如何实施各种随机抽样方法。一个互联网服务提供商进行了一项客户满意度调查,这个调查的目标人群是该公司当前的用户。该公司计划从当前用户中选择一个样本,采访选中客户,然后根据样本数据推断整个被调查总体的情况。

SAS 数据集 Customers 包含抽样框,它是被调查总体的抽样单元目录。样本客户将从这个抽样框中抽取。数据集 Customers 是公司客户数据库的重要组成部分。它包括了每个客户的 4 项有关信息,即客户 ID 号(CustomerID)、客户来自的州名(State)、新老客户类型(Type)、服务使用量(Usage),共有 13 471个客户。读者或用户在 SAS 程序编辑窗口内输入以下 4 句 SAS 语句:

```
title1 'Customer Satisfaction Survey';
title2 'First 10 Observations';
proc print data = Customers(obs = 10);
run;
```

【程序说明】前两句将在输出结果的前两行产生标题,内容分别为"客户满意度调查(Customer Satisfaction Survey)"和"前 10 个观测(First 10 Observations)";第 3 句是调用 SAS 中的 print 过程打印数据集中的信息,其选择项"data = Customers(obs = 10)"的含义:采用的数据集名称为 Customers,且仅输出该数据集中的前 10 行信息(即前 10 个观测者在 4 个变量上对应的全部信息)。

将上述 4 句语句组成的一段简单 SAS 程序(只有过程步,没有数据步,因为调用 SAS 软件中已有的 SAS 数据集 Customers),便可显示 Customers 数据集中前 10 个观测数据,结果见表 1。

表 1 数据集 Customers 中前 10 位客户的有关信息

Obs	CustomerID	State	Type	Usage
1	416 - 87 - 4322	AL	New	839
2	288 - 13 - 9763	GA	Old	224
3	399 - 00 - 8654	GA	Old	2451
4	118 - 98 - 0542	GA	New	349
5	421 - 67 - 0342	FL	New	562
6	623 - 18 - 9201	SC	New	68
7	324 - 55 - 0324	FL	Old	137
8	832 - 90 - 2397	AL	Old	1563
9	586 - 45 - 0178	GA	New	615
10	801 - 24 - 5317	SC	New	728

在 SAS 数据集 Customers 中,变量CustomerID唯一地标识每个客户;变量 State 是客户所在州地址,该公司客户在 4 个州:格鲁吉亚(GA)、阿拉巴马州(AL)、佛罗里达州(FL)和南卡罗莱纳(SC);变量 Type 取值为"Old"表示该客户订购公司服务超过一年,与之相对应的取值为"New";变量 Usage 表示客户几分钟内平均每月服务使用量。

接下来的部分展示了在三种不同设计下采用 SURVEYSELECT 过程对客户满意度实施概率抽样调

查,给出所需要的 SAS 程序及抽样结果。这三种设计都是以每一个客户为一个抽样单位。第一种设计是不分层简单随机抽样;第二种设计是按 State 和 Type 分层,在层内采用简单随机抽样;第三种设计是按 Usage 分层,然后在层内排序,最后采用简单随机抽样。

4 简单随机抽样及 SAS 实现[8-9]

以下是 PROC SURVEYSELECT 语句采用简单随机抽样抽取 Customers 数据集中的一个概率样本。

```
title1 'Customer Satisfaction Survey';
title2 'Simple Random Sampling';
proc surveyselect data = customers method = srs n = 100
                  out = samplesrs;
run;
```

【程序说明】proc surveyselect 语句调用 surveyselect 过程。该语句包含了如下 4 个选项:第 1 个选项为"data = customers",指定 SAS 数据集 Customers 作为输入数据集来选择样本;第 2 个选项为"method = srs",指定抽样方法为简单随机抽样。在简单随机抽样中,每一个抽样单位都有同等的概率被抽取,样本是无放回抽取的,意味着每一个抽样单元不能被多次抽取;第 4 个选项为"n = 100",指定被抽取的样本大小为 100 个客户;第 4 个选项为"out = samplesrs 将抽取到的样本储存到名为 samplesrs 的 SAS 数据集中去。

上面的 SAS 程序可以产生如下的信息,见表 2。

表 2 对数据集 Customers 进行简单随机抽样的有关情况说明

Selection Method	Simple Random Sampling
Input Data Set	CUSTOMERS
Random Number Seed	39647
Simple Size	100
Selection Probability	0.007423
Sampling Weight	134.71
Output Data Set	samplesrs

表 2 概要地报告了使用 SURVEYSELECT 过程进行随机抽样的有关情况。采用简单随机抽样从 Customers 数据集中抽取了 100 个客户;随机种子数为 39647,SURVEYSELECT 过程用这个数字作为初始种子来产生随机数字。由于在 SURVEYSELECT 过程中没有指定 seed = option 选项,种子值采用的是计算机系统时间;每位客户被选中的概率为 0.007423,该概率等于样本大小(100)除以总体容量(13471)所得的商;样本中每一客户的抽样权重为 134.71,抽样权重是抽样概率的倒数;真正的抽样结果(即被抽取的 100 位客户在 4 个变量上的取值情况)被放置在输出数据集 samplesrs 中。

这 100 位样本客户被储存在 SAS 数据集 samplesrs 中。PROC SURVEYSELECT 并没有直接显示此输出数据集的内容。下面用 PROC PRINT 语句显示 samplesrs 中前 20 个观测。

```
title1 'Customer Satisfaction Survey';
title2 'Sample of 100 Customers, Selected by SRS';
title3 '(First 20 Observations)';
proc print data = samplesrs(obs = 20);
run;
```

【程序说明】参见前面产生表 1 的 SAS 程序后面的"程序说明",此处从略。

上面这段 SAS 程序产生的结果见表 3。

表 3 采用简单随机抽样从 customers 中随机抽取的 100 位客户中的前 20 位

Obs	CustomerID	State	Type	Usage
1	036-89-0212	FL	New	74
2	045-53-3676	AL	New	411
3	050-99-2380	GA	Old	167
4	066-93-5368	AL	Old	1232
5	082-99-9234	FL	New	90
……				

注:因篇幅所限,仅显示出了前 5 位

表 3 显示了包含样本客户的输出数据集 samplesrs 的前 20 个观测。这个数据集包含了输入数据集 Customers 中的所有变量。

5 分层随机抽样及 SAS 实现[8-9]

在 Customers 数据集中,抽样框是按 State 和 Type 分层后的所有客户目录性清单。这就把抽样框按 State 和 Type 取值分成了互不重叠的亚组,其亚组的数目等于 State 和 Type 两个变量或因素的水平数之乘积。然后,SAS 软件将在每一层中独立选择样本。

PROC SURVEYSELECT 要求输入数据集为按分层变量排序后的数据集。下面 PROC SORT 语句使 Customers 数据集按分层变量 State 和 Type 进行排序。

```
proc sort data = Customers;
by State Type;
run;
```

下面 PROC FREQ 语句显示 customers 数据集中 State 和 Type 两个变量所形成的交叉频数表。

```
title1 'Customer Satisfaction Survey';
title2 'Strata of Customers';
proc freq data = customers;
    tables State* Type;
run;
```

上面这段 SAS 程序被执行后，输出结果见表 4。

表 4 数据集 Customers 按 State 和 Type 两变量形成交叉表后形成的网格及各网格内的频数分布情况

Frequency
Percent
Row Pct
Col Pct

Table of State by Type			
State	Type		
	New	Old	Total
AL	1238	706	1944
	9.19	5.24	14.43
	63.68	36.32	
	14.43	14.43	
FL	2170	1370	3540
	16.11	10.17	26.28
	61.30	38.70	
	25.29	28.01	
GA	3488	1940	5428
	25.89	14.40	40.29
	64.26	35.74	
	40.65	39.66	
SC	1684	875	2559
	12.50	6.50	19.00
	65.81	34.19	
	19.63	17.89	
Total	8580	4891	13471
	63.69	36.31	100.00

表 4 给出了 13471 个客户按 Type 分组后再按 State 分组所形成的频数分布表。四个州和两类客户共形成 8 个层，每层中计算出 4 个数值，从上到下分别代表“频数”、“占全部客户数的百分比”、“占行合计的百分比”和“占列合计的百分比”。

下面 PROC SURVEYSELECT 语句根据 State 和 Type 两个变量进行分层随机抽样设计从 Customers 数据集按概率抽取一个客户样本。

```
title1 'Customer Satisfaction Survey';
title2 'Stratified Sampling';
proc surveyselect data = customers method = srs n = 15 seed =
1953 out = samplestrata;
    strata state type;
run;
```

【程序说明】strata 语句声明分层变量为 State 和 Type。在 PROC SURVEYSELECT 语句中，“method = srs”选项指定抽样方法为简单随机抽样，“n = 15”选项指定每层抽取的样本大小为 15 位客户。如果想不同层指定不同样本大小，可以使用“n = SAS - data - set (即一个新数据集名)”选项来声明一个新数据集，该数据集包含每层样本大小。“seed = 1953”选项指定“1953”为产生随机数的初始种子数。

上面这段 SAS 程序仅显示抽样情况的总结信息，一共有 120 位客户被抽取，因篇幅所限，有关分层随机抽样的总结信息从略。

下面 PROC PRINT 语句将显示分层随机抽样结果数据集 samplestrata 中前 30 个观测。

```
title1 'Customer Satisfaction Survey';
title2 'Sample Selected by Stratified Design';
title3 '(First 30 Observations)';
proc print data = samplestrata(obs = 30);
run;
```

【程序说明】参见前面产生表 1 的 SAS 程序后面的“程序说明”，此处从略。

上面这段 SAS 程序被执行后，输出结果见表 5。

表 5 采用分层随机抽样从 customers 中随机抽取的 120 位客户中的前 30 位

Obs	State	Type	CustomerID	Usage	SelectionProb	SamplingWeight
1	AL	New	002 - 26 - 1498	1189	0.012116	82.5333
2	AL	New	070 - 86 - 8494	106	0.012116	82.5333
3	AL	New	121 - 28 - 6895	76	0.012116	82.5333
4	AL	New	131 - 79 - 7630	265	0.012116	82.5333
5	AL	New	211 - 88 - 4991	108	0.012116	82.5333
……						

注：因篇幅所限，仅显示出了前 5 位

表 5 显示了输出数据集 samplestrata 的前 30 个观测数据，samplestrata 含有 8 个层，每层 15 位客户，一共有 120 位客户。变量 SelectionProb 指样本中每个客户被抽中的概率。由于在同一层中每位客户被抽中的概率相同，所以抽样概率等于层样本大小(15)除以该层的总样本含量之商。由于层间大小不同，所以抽样概率在不同层中是不一样的。变量 SamplingWeight 为抽样权重，抽样权重为抽样概率的倒数。

6 控制排序分层随机抽样及 SAS 实现[8]

下一个客户满意度调查抽样设计是按 State 分层，同时在层内按 Type 和 Usage 排序。在分层和控制排序后，在每一层中按系统随机抽样方法抽取客户。系统抽样加上抽样前控制排序使得样本的 Type 和 Usage 取值在每层(State)内是均匀分布的。下面 PROC SURVEYSELECT 语句根据这种设计从 Customers 数据集中按概率抽取客户样本。

```
title1 'Customer Satisfaction Survey';
title2 'Stratified Sampling with Control Sorting';
proc surveyselect data = Customers method = sys rate = .02
        seed = 1234 out = SampleControl;
    strata State;
    control Type Usage;
run;
```

【程序说明】STRATA 语句声明分层变量为 State。CONTROL 语句声明控制变量为 Type 和 Usage。在 PROC SURVEYSELECT 语句中，"METHOD = SYS"选项指定抽样方法为系统随机抽样。"RATE = 0.02"选项指定每层抽样率为 0.02。"SEED = 1234"选项指定产生随机数的初始种子数。

上面这段 SAS 程序仅显示抽样情况的总结信息，一共有 271 位客户被抽取，因篇幅所限，按 State 分层同时在层内按 Type 和 Usage 排序的总结信息和抽样结果(即采用 PRINT 过程输出数据集 SampleControl 的内容)均省略。

参考文献

[1] 郭春雪，胡良平. 正确把握精神卫生临床试验设计三要素的要领(Ⅰ)—受试对象[J]. 四川精神卫生，2016，28(3)：197－201.

[2] 胡完，胡良平. 正确把握精神卫生临床试验设计三要素的要领(Ⅱ)—影响因素[J]. 四川精神卫生，2016，28(3)：202－206.

[3] 谷恒明，胡良平. 正确把握精神卫生临床试验设计三要素的要领(Ⅲ)—观测指标[J]. 四川精神卫生，2016，28(3)：207－210.

[4] 杨孟渊，胡良平. 精神卫生科研如何严格遵守试验设计四原则之随机原则[J]. 四川精神卫生，2016，29(4)：289－294.

[5] 沈宁，胡良平. 精神卫生科研如何严格遵守试验设计四原则之对照原则[J]. 四川精神卫生，2016，29(4)：295－302.

[6] 张效嘉，胡良平. 精神卫生科研如何严格遵守试验设计四原则之重复原则[J]. 四川精神卫生，2016，29(4)：303－306.

[7] 张效嘉，胡良平. 精神卫生科研如何严格遵守试验设计四原则之均衡原则[J]. 四川精神卫生，2016，29(4)：307－310.

[8] SAS Institute Inc. SAS/STAT 9.3 User's Guide[M]. Cary，NC：SAS Institute Inc，2011：7633－7704.

[9] 胡良平. 科研设计与统计分析[M]. 北京：军事医学科学出版社，2012：206－227.

(收稿日期：2016－10－11)

(本文编辑：陈 霞)