SURVEYMEANS 过程在抽样调查资料分析中的应用

李长平[1,2],胡良平[2,3*]

(1.天津医科大学公共卫生学院卫生统计学教研室,天津 300070;
2.世界中医药学会联合会临床科研统计学专业委员会,北京 100029;
3.军事医学科学院生物医学统计学咨询中心,北京 100850
*通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 传统的统计分析方法在进行差异性分析、线性与广义线性回归分析时,基本上都是基于样本来自无限总体、完全随机抽样的基础上估计抽样误差。而调查数据往往来自于分层、整群、多阶段或不等概率等复杂随机抽样方法,此时若采用前述提及的经典统计分析方法,则不能准确估计抽样误差。本文通过具体实例,介绍如何应用 SAS 软件中的 SURVEYMEANS过程,更好地实现对通过各种抽样方法获得的数据进行统计描述和简单的统计分析,以便达到准确估计抽样误差、对总体参数描述和估计的目的。

【关键词】 SAS 软件;SUVEYMEANS 过程;简单随机抽样;分层抽样;分层整群抽样
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2017.05.005

Application of the SURVEYMEANS procedure in the analysis of sampling survey data

Li Changping[1,2], *Hu Liangping*[2,3*]

(1. *Department of Health Statistics, School of Public Health, Tianjin Medical University, Tianjin* 300070, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*;
3. *Consulting Center of Biomedical Statistics, Academy of Military Medical Sciences, Beijing* 100850, *China*
** Corresponding author: Hu Liangping, E-mail: lphu812@sina.com*)

【**Abstract**】 When performing a difference analysis or a linear and generalized linear regression analysis , traditional statistical methods are basically based on the sample from the infinite population or completely random sampling to estimate the sampling error . However, the survey data are usually collected from complex random sampling methods , such as stratified, cluster, multi-stage or unequal probability. At this point, the sampling error cannot be accurately estimated if the classical statistical analysis methods mentioned above are adopted. Through specific examples, this article aimed to apply the SURVEYMEANS procedure in SAS software which can better implement the statistical description and analysis of the data obtained by various sampling methods , in order to estimate the sampling error and population parameters accurately .

【**Keywords**】 SAS software; SURVEYMEANS procedure; Simple random sampling; Stratified sampling; Stratified cluster sampling

1 调查资料统计分析方法概述

1.1 随机抽样方法简介

调查研究是医学科学研究常见的形式之一。而无论是观察性研究如横断面研究,还是分析性研究如病例对照研究、队列研究,绝大多数时候都会采用抽样调查的形式。那么,一旦采用抽样的形式选取研究对象,研究结果就会存在抽样误差。常用的概率抽样调查的方法有完全随机抽样、系统抽样、分层抽样、整群抽样等。不同的抽样方法,抽样误差大小的估计方法是不同的[1]。

1.2 调查资料统计描述与简单统计分析方法简介

传统统计分析软件(如 SAS,其 MEANS、GLM 过程等)中的算法,通常都是基于“无限总体、完全随机抽样”这样的假设基础上估计抽样误差的。而当抽样方法相对复杂,采用这些程序计算将不能得到正确的抽样误差估计值[2-3]。此时,SAS/STAT 中的 SURVEYMEANS 过程就能发挥其作用了。PROC SURVEYMEANS 利用 Taylor 扩展方法估计基于复杂抽样设计的统计量抽样误差。该方法获得统计量的一个线性近似值并用该近似值的方差估计来推断统计量本身的方差[4-5]。

1.3 调查资料回归分析方法简介

在 SAS 软件中,有 SASREG、SASLOGISTIC、SASPHREG 三个过程可被用来对各种复杂抽样调查资料进行建模。针对不同的随机抽样方法,采取相应的算法去估计方差-协方差矩阵,以便更好地估计回归系数;同时,还采用 Taylor 级数方法或重抽

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

样方法来估计抽样误差。

2 采用 SURVEYMEANS 过程实现统计描述与简单统计分析

2.1 SURVEYMEANS 过程简介

PROC SURVEYMEANS DATA = 用于指定要分析的输入数据集。当调查设计包括有限总体校正因子时,可以用 RATE = 或 TOTAL = 选项指定抽样率或抽样大小;

BY 指定分组单独分析变量;

CLASS 指定作为属性变量来分析的变量;

CLUSTER 指定整群抽样设计中的群识别变量;

DOMAIN 语句对亚总体或域进行分析的变量;

RATIO 计算分析变量均值或构成的比值,分子变量/分母变量;

STRATA 指定分层抽样设计中的分层变量;

VAR 指定分析变量;

WEIGHT 指定包含抽样权重的变量;

RUN;

2.2 基于完全随机抽样设计的统计分析

【例 1】假设从总体 4 000 名学生(七、八、九年级)中采用随机抽样方法抽取 40 名学生作为样本。研究者想通过对这 40 名学生的调查了解学生平均每周的冰淇淋花费,以及每周的冰淇淋花费超过 10 美元的学生的比例。具体数据如表 1 所示,表 1 数据存为 SAS 数据集,命名为 IceCream。

表 1 40 名学生每周冰激淋花费情况

Obs	grade	Spending($)	Group
1	7	7	less
2	7	7	less
3	8	12	more
4	9	10	more
5	7	1	less
6	7	10	more
……	……	……	……
37	9	6	less
38	9	11	more
39	7	2	less
40	7	9	less

注:Grade,年级;Spending,花费;less,<10 美元;more,≥10 美元

对应的 SAS 计算程序如下:

```
title1 'Analysis of Ice Cream Spending';
title2 'Simple Random Sample Design';
proc surveymeans data = IceCream total = 4000;
var Spending Group;
run;
```

【程序说明】proc surveymeans 调用 surveymeans 过程。TOTAL = 4 000,指进行一个样本量为 4 000 的有限总体校正的方差估计

【输出结果】

The SURVEYMEANS Procedure	
Data Summary	
Number of Observations	40

Statistics						
Variable	Level	N	Mean	Std Error of Mean	95% CL for Mean	
Spending		40	8.750000	0.845139	7.04054539	10.4594546
Group	less	23	0.575000	0.078761	0.41568994	0.7343101
	more	17	0.425000	0.078761	0.26568994	0.5843101

【结果说明】数值变量 Spending 的结果显示,学生总体中平均每周冰淇淋花费为 8.75 美元,95% 置信区间为(7.04,10.46)美元。属性变量 Group 的结果显示,学生总体中平均每周冰淇淋花费少于 10 美元的比例约为 57.5%,置信区间为(41.6%,73.4%),多于 10 美元的比例约为 42.5%,置信区间为(26.6%,58.45%)。

2.3 基于分层抽样设计的统计分析

【例 2】沿用例 1 的背景资料。假设上例中 4 000名学生是来自分层抽样设计,按年级分层,各年级抽取的学生人数见下表 2。试对数据进行分析。

表 2 各年级抽取的学生人数

Grade	Number of Students
7	1 824
8	1 025
9	1 151
Total	4 000

对应的 SAS 计算程序如下:

```
data StudentTotals;
input Grade _total_;
datalines;
7 1824
8 1025
9 1151
;
data IceCream;
set IceCream;
if Grade = 7 then Prob = 20/1824;
if Grade = 8 then Prob = 9/1025;
if Grade = 9 then Prob = 11/1151;
Weight = 1/Prob;
title1 'Analysis of Ice Cream Spending';
title2 'Stratified Simple Random Sample Design';
proc surveymeans data = IceCream total = StudentTotals;
stratum Grade/list;
var Spending Group;
weight weight;
run;
```

【程序说明】Grade 是分层变量,变量_total_表示各层总体大小,该名称为程序中固定格式。程序方差估计时利用每层总体观测量大小来校正有限总体抽样的影响。若不提供各总体的大小或抽样率,则系统假定样本中包含总体的率非常小,此时不做有限总体校正。在分层抽样设计中,当各层抽样概率不同时,需要定义样本的权重以便做到对均数无偏的估计。在本例中,采用按比例抽样的方式,每层抽样概率的倒数作为样本权重(即用 Weight 命令设置权重)。List 命令要求输出每层的信息。每个年级中抽取的样本数为事先按一定规则,如等比例抽样规定的样本数。

【输出结果】

(1)生成的 IceCream SAS 数据集截图如下:

	Grade	Spending	Group	Prob	Weight
1	7	7	less	0.0109649123	91.2
2	7	7	less	0.0109649123	91.2
3	8	12	more	0.0087804878	113.88888889
4	9	10	more	0.009556907	104.63636364
5	7	1	less	0.0109649123	91.2
6	7	10	more	0.0109649123	91.2
7	7	3	less	0.0109649123	91.2
8	8	20	more	0.0087804878	113.88888889
9	8	19	more	0.0087804878	113.88888889
10	7	2	less	0.0109649123	91.2
11	7	2	less	0.0109649123	91.2
12	9	15	more	0.009556907	104.63636364
13	8	16	more	0.0087804878	113.88888889
14	7	6	less	0.0109649123	91.2
15	7	6	less	0.0109649123	91.2
16	7	6	less	0.0109649123	91.2
17	9	15	more	0.009556907	104.63636364

(2)Output 结果:

Stratum Information

Stratum Index	Grade	Population Total	Sampling Rate	N Obs	Variable	Level	N
1	7	1824	1.10%	20	Spending		20
					Group	less	17
						more	3
2	8	1025	0.88%	9	Spending		9
					Group	less	0
						more	9
3	9	1151	0.96%	11	Spending		11
					Group	less	6
						more	5

上表显示按三个年级分层、各层的总人数、抽样率、抽取的各层总样本数、对应的变量及细分组样本数信息。

Statistics

Variable	Level	N	Mean	Std Error of Mean	95% CL for Mean	
Spending		40	9.141298	0.531799	8.06377052	10.2188254
Group	less	23	0.544555	0.058424	0.42617678	0.6629323
	more	17	0.455445	0.058424	0.33706769	0.5738232

对 Spending 分析的结果显示,学生总体中平均每周冰淇淋花费为 9.14 美元,标准误为 0.53,95%置信区间为(8.06,10.22)美元。对 Group 分析的结果显示,学生总体中平均每周冰淇淋花费少于 10 美

元的约为 54.5%,置信区间为(42.6%,66.3%),多于 10 美元的约为 45.5%,置信区间为(33.7%,57.4%),标准误为 5.8%。

2.4 基于分层整群抽样设计的统计分析

【例 3】沿用例 1 的背景资料。假设从总体 4 000名学生中采用分层整群抽样获取 40 例样本。4 000 名学生的总体情况如下表 3 所示:

表 3 4 000 名学生各年级及学习小组构成情况

Grade	Number of Study Groups	Number of Students
7	608	1 824
8	252	1 025
9	403	1 151
Total	617	4 000

4 000 名学生来自七、八、九年级。各年级有对应的人数(表 3 第 3 列)和若干学习小组(表 3 第 2 列)。每个学习小组中有 2 ~4 个学生。

在本例中,抽样单位(或"群")是学习小组。以年级为分层单位,对学习小组进行随机抽选,选中的学习小组中的所有学生作为样本。假定从七、八、九年级分别抽取了 8、3、5 个学习小组。

对应的 SAS 程序如下:

```
①data IceCreamStudy;
input Grade StudyGroup Spending @@;
if (Spending<10) then Group='less';
else Group='more';
datalines;
7 34 7 7 34 7 7 412 4 9 27 14
7 34 2 9 230 15 9 27 15 7 501 2
9 230 8 9 230 7 7 501 3 8 59 20
7 403 4 7 403 11 8 59 13 8 59 17
8 143 12 8 143 16 8 59 18 9 235 9
8 143 10 9 312 8 9 235 6 9 235 11
9 312 10 7 321 6 8 156 19 8 156 14
7 321 3 7 321 12 7 489 2 7 489 9
7 78 1 7 78 10 7 489 2 7 156 1
7 78 6 7 412 6 7 156 2 9 301 8
;
②data StudentGroups;
input Grade _total_;
datalines;
7 608
8 252
9 403
;
③data IceCreamStudy;
set IceCreamStudy;
if Grade=7 then Prob=8/608;
if Grade=8 then Prob=3/252;
if Grade=9 then Prob=5/403;
Weight=1/Prob;

title1 'Analysis of Ice Cream Spending';
title2 'Stratified Clustered Sample Design';
④proc surveymeans data=IceCreamStudy total=StudentGroups;
strata Grade/list;
cluster StudyGroup;
var Spending Group;
weight weight;
run;
```

【程序说明】数据步①中 Group 表示年级,StudyGroup 表示学习小组,不同年级的小组编号可以相同,因为小组编号是按年级和其小组数排的顺序编号。Spending 表示冰淇淋花费,Group 是根据冰淇淋花费进行分组。数据步②中 Grade 是分层变量,变量_total_表示各层学习小组数,该名称为程序中固定格式,用于表达主要抽样单位。数据步③中,定义主要抽样单位的权重。权重为群抽样概率的倒数。过程步④中,strata 定义分层变量,cluster 定义群抽样单位变量。

【输出结果】

(1)生成的 IceCreamstudy SAS 数据集截图如下:

	Grade	StudyGroup	Spending	Group	Prob	Weight
1	7	34	7	less	0.0131578947	76
2	7	34	7	less	0.0131578947	76
3	7	412	4	less	0.0131578947	76
4	9	27	14	more	0.0124069479	80.6
5	7	34	2	less	0.0131578947	76
6	9	230	15	more	0.0124069479	80.6
7	9	27	15	more	0.0124069479	80.6
8	7	501	2	less	0.0131578947	76
9	9	230	8	less	0.0124069479	80.6
10	9	230	7	less	0.0124069479	80.6
11	7	501	3	less	0.0131578947	76
12	8	59	20	more	0.0119047619	84
13	7	403	4	less	0.0131578947	76
14	7	403	11	more	0.0131578947	76
15	8	59	13	more	0.0119047619	84
16	8	59	17	more	0.0119047619	84
17	8	143	12	more	0.0119047619	84
18	8	143	16	more	0.0119047619	84
19	8	59	18	more	0.0119047619	84

(2)Output 输出结果:

Data Summary	
Number of Strata	3
Number of Clusters	16
Number of Observations	40
Sum of Weights	3162.6

Class Level Information		
CLASS Variable	Levels	Values
Group	2	less more

Stratum Information								
Stratum Index	Grade	Population Total	Sampling Rate	N Obs	Variable	Level	N	Clusters
1	7	608	1.32%	20	Spending		20	8
					Group	less	17	8
						more	3	3
2	8	252	1.19%	9	Spending		9	3
					Group	less	0	0
						more	9	3
3	9	403	1.24%	11	Spending		11	5
					Group	less	6	4
						more	5	4

上表中给出了按三个年级分层,各层的总群数、抽样率、抽取的各层总样本数、对应的变量水平、细分组样本数信息和群数。

Statistics						
Variable	Level	N	Mean	Std Error of Mean	95% CL for Mean	
Spending		40	8.923860	0.650859	7.51776370	10.3299565
Group	less	23	0.561437	0.056368	0.43966057	0.6832130
	more	17	0.438563	0.056368	0.31678698	0.5603394

对 Spending 分析的结果显示,学生总体中平均每周冰淇淋花费为 8.92 美元,标准误为 0.53,95% 置信区间为(7.52,10.33)美元。对 Group 分析的结果显示,学生总体中平均每周冰淇淋花费少于 10 美元的比例约为 56.1%,置信区间为(44.0%,68.3%),多于 10 美元的比例约为 43.9%,置信区间为(31.7%,56.0%),标准误为 5.6%。

2.5 SURVEYMEANS 在域分析(Domain analysis)中的应用

域分析是指对亚群或域的统计计算,进行分析的亚组可以与样本抽样设计无关,该分析也称为亚组分析、亚群分析或子域分析。如下例所示:

【例 4】欲对前 800 家公司情况进行分析,了解其概况及经济相关状况,同时了解不同公司市场类型特征下的经济情况。现有其中 66 家公司的样本,但该 66 家公司的抽取并没有考虑到市场类型这一因素,即为市场类型的非概率抽样,样本中每个市场类型中含有多少个公司是一个随机变量。此时,要对每一个市场类型作相应的分析,可采用域分析。

【SAS 程序如下】

```
data Company;
    length Type  $14;
    input Type $ Asset Sale Value Profit Employee Weight;
datalines;
    Other          2764.0   1828.0   1850.3   144.0   18.7   9.6
    Energy         13246.2  4633.5   4387.7   462.9   24.3   42.6
    Finance        3597.7   377.8    93.0     14.0    1.1    12.2
    Transportation 6646.1   6414.2   2377.5   348.2   47.1   21.8
    HiTech         1068.4   1689.8   1430.2   72.9    4.6    4.3
    Manufacturing  1125.0   1719.4   1057.5   98.1    20.4   4.5
    Other          1459.0   1241.4   452.7    24.5    20.1   5.5
```

```
Finance          2672.3    262.5    296.2    23.1    2.2    9.3
Finance          311.0     566.2    932.0    52.8    2.7    1.9
Energy           1148.6    1014.6   485.1    60.6    4.0    4.5
Finance          5327.0    572.4    372.9    25.2    4.2    17.7
Energy           1602.7    678.4    653.0    75.6    2.8    6.0
Energy           5808.8    1288.4   2007.0   318.8   5.9    19.2
Medical          268.8     204.4    820.9    45.6    3.7    1.8
Transportation   5222.6    2627.8   1910.0   245.6   22.8   17.4
Other            872.7     1419.4   939.3    69.7    12.2   3.7
……
;
title1 'Top Companies Profile Study';
proc surveymeans data = Company total = 800 mean sum;
    var Asset Sale Value Profit Employee;
    weight Weight;
    domain Type;
run;
```

【程序说明】数据步中 Type 表示市场类型，Asset表示资产（百万美元），Sale 表示销售额（百万美元），Value 表示公司的市场价值（百万美元），Profit 表示利润（百万美元），Employee 表示员工数（千），weight 代表权重，共 66 行即 66 家公司的数据。为节省篇幅，仅列出部分数据。

【输出结果】：

The SURVEYMEANS Procedure	
Data Summary	
Number of Observations	66
Sum of Weights	799.8

Statistics				
Variable	Mean	Std Error of Mean	Sum	Std Dev
Asset	6523.488510	720.557075	5217486	1073829
Sale	4215.995799	839.132506	3371953	847885
Value	2145.935121	342.531720	1716319	359609
Profit	188.788210	25.057876	150993	30144
Employee	36.874869	7.787857	29493	7148.003298

Domain Statistics in Type					
Type	Variable	Mean	Std Error of Mean	Sum	Std Dev
Energy	Asset	7868.302932	1941.699163	1449341	785962
	Sale	5419.679099	2416.214417	998305	673373
	Value	2249.297177	520.295162	414321	213580
	Profit	289.564658	52.512141	53338	25927
	Employee	14.151194	3.974697	2606.650000	1481.777769
Finance	Asset	7890.190264	1057.185336	1855773	704506
	Sale	829.210502	115.762531	195030	74436
	Value	565.068197	76.964547	132904	48156
	Profit	63.716837	10.099341	14986	5801.108513

	Employee	5.806293	0.811555	1365.640000	519.658410
HiTech	Asset	5031.959781	732.436967	321542	183302
	Sale	5464.292019	731.296997	349168	196013
	Value	6707.828482	1194.160584	428630	249154
	Profit	346.407042	42.299004	22135	12223
	Employee	70.766980	8.683595	4522.010000	2524.778281
Manufacturing	Asset	7403.004250	1454.921083	888361	492577
	Sale	7207.638833	2112.444703	864917	501679
	Value	2986.442750	799.121544	358373	196979
	Profit	211.933583	39.993255	25432	13322
	Employee	83.314333	31.089019	9997.720000	6294.309490
Medical	Asset	5046.570609	1218.444638	140799	131942
	Sale	3313.219713	758.216303	92439	85655
	Value	2561.614695	530.802245	71469	64663
	Profit	218.682796	44.051447	6101.250000	5509.560969
	Employee	46.518996	11.135955	1297.880000	1213.651734
Other	Asset	1850.250000	338.128984	58838	31375
	Sale	1620.784906	168.686773	51541	24593
	Value	1432.820755	297.869828	45564	24204
	Profit	115.089937	27.970560	3659.860000	2018.201371
	Employee	14.306604	2.313733	454.950000	216.327710
Retail	Asset	2939.845750	393.692369	235188	94605
	Sale	7395.453500	1746.187580	591636	263263
	Value	2103.863125	529.756409	168309	78304
	Profit	157.171875	31.734253	12574	5478.281027
	Employee	93.624000	15.726743	7489.920000	3093.832061
Transportation	Asset	4712.047359	888.954411	267644	163516
	Sale	4030.233275	1015.555708	228917	142669
	Value	1703.330282	313.841326	96749	58947
	Profit	224.762324	56.168925	12767	8287.585418
	Employee	30.946303	6.786270	1757.750000	1066.586615

以上结果是给出总的和各市场类型对应的各指标的均数、标准误及置信区间。

另外,在 SURVEYMEANS 过程中,还能对抽样调查数据存在缺失值、其它抽样方法如有放回分层整群抽样[6]等进行处理,在此不再一一介绍。对于本文中涉及的置信区间的计算方法在此不再赘述,具体计算公式参考相关文献[7]。

参考文献

[1] LehtonenR, Pahkinen E. Practical methods for design and analysis of complex survey[M]. New York: Wiley, 2004: 22 - 37.

[2] 刘建华, 金水高. 复杂抽样调查总体特征量及其方差的估计[J]. 中国卫生统计,2008, 25(4): 377 - 379.

[3] Brick JM, Kalton G. Handling missing data in survey research [J]. Stat Methods Med Res, 1996, 5(3): 215 - 238.

[4] Woodruff RS. A simple method for approximating the variance of a complicated estimate [J]. J Am Stat Assoc, 1971, 66 (334): 411 - 414.

[5] Fuller WA. Regression analysis for sample survey[J]. Sankhya, 1975, 37(3): 117 - 132.

[6] Francisco CA,Fuller WA. Quantile estimation with a complex survey design[J]. AnnStat, 1991,19(1): 454 - 469.

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

(收稿日期:2017 - 08 - 17)

(本文编辑:陈 霞)