基于 STDRATE 过程的标准化率及其置信区间估计的 SAS 宏*

湖北省疾病预防控制中心慢性病防治研究所(430079)　周梦格　张　岚　何田静　潘敬菊△

【提　要】　目的　探讨 STDRATE 过程在标准化率及其置信区间估计方面的应用,并通过编写 SAS 宏程序批量实现多维度分组数据的该过程。**方法**　利用死因监测数据实例介绍基于 STDRATE 过程实现多维度分组大样本数据标准化率及其置信区间估计的 SAS 宏命令(%Stdrate)。**结果**　%Stdrate 宏命令能够快速计算多维度分组大样本数据的标准化率及其置信区间并输出至指定数据集,可为日常及科研工作节约时间。**结论**　%Stdrate 宏命令为多维度分组数据直接标准化率及其置信区间估计提供了很大的方便,具有较强的实际应用价值。

【关键词】 STDRATE 过程　标准化率　置信区间　SAS 宏

【中图分类号】 R195.1　　**【文献标识码】** A　　**DOI** 10.11783/j.issn.1002-3674.2024.05.020

在流行病学研究中,我们通常用标准化率来描述不同时期或不同地区的率,如标准化患病率、发病率或死亡率等,以消除不同时期或地区某些因素(如人口年龄、性别等)构成不同造成的影响。同时,由于针对多数慢性非传染病等监测多采用选取监测点的方式进行监测,计算所得的率尚是监测点样本标准化率的点估计,若要估计其总体的标准化率,则需计算 95% 置信区间。目前,已有研究介绍如何在 SPSS 及 Stata 中运用 Bootstrap 法实现对总体标准化率的区间估计[1]。夏雷震等[2]也介绍了利用 SAS 宏程序计算多分组大样本数据标准化率及其置信区间的方法,但程序较为复杂且无法同时实现多维度分组数据率的标准化及其置信区间估计。随着我国公共卫生监测的不断发展,复杂的多维度分组数据日益增多,如死因监测工作中搜集的不同地区连续多年的多种疾病的死亡数据,如何快速计算不同年度不同地区多种疾病的标化率及其置信区间成为日常及科研工作中的重要问题。本研究通过编写基于 STDRATE 过程的 SAS 宏程序,实现多维度分组大样本数据标准化率的快速计算及其置信区间估计。

原理和方法

率的标准化目的是在共同的“标准”上比较两组或多组数据,以消除某影响因素的构成不同对结果的影响。常用的标准化方法有直接标准化法和间接标准化法,通常根据已有的资料条件选择不同的方法计算标准化率[3]。如对死亡率进行标准化时,若已知样本数据的年龄别死亡率,可采用直接标准化法,计算公式为 $P'=\frac{\sum_i N_i P_i}{N}$ 或 $P'=\sum_i(\frac{N_i}{N})P_i$,其中 N_i 为标准年龄别人口数,P_i 为实际年龄别死亡率,N 为标准人口总数。若只有样本数据的总死亡数和年龄别人口数,或者各年龄组人口数较小导致年龄别死亡率不稳定时,可选择间接标准化法,计算公式为 $P'=P\frac{r}{\sum n_i p_i}=P\times SMR$,其中 P 为标准总死亡率,r 为实际总死亡数,n_i 为实际年龄别人口数,p_i 为标准年龄别死亡率,SMR(standard mortality ratio)为标准化死亡比[4-5]。在公共卫生监测领域,常选用大样本的群体,如省、全国或世界的人口构成作为“标准”来计算标准化率[6]。鉴于目前多数监测数据及研究中可获得不同年龄组的率,因此,本研究主要介绍直接标准化法的 SAS 宏实现。

STDRATE 过程是一个强大的 SAS 过程步,可用于流行病学研究中直接标准化率或间接标准化率的计算,还可进行 Mantel-Haenszel 分层估计以及人群归因分数的估计[7-8]。若将 STDRATE 过程与 SAS 宏进行结合,可极大地减少针对多维度分组的大样本数据分析的工作量,对日常工作及科学研究均能提供极大的便利[9]。

实例分析和 SAS 宏实现

1.资料数据

某研究欲了解某省不同地区在不同时期的疾病死亡情况及其变化趋势,将对 2013—2019 年的死因监测数据进行分析,以 2010 年全国普查人口构成作为标准人口,采用直接标准化法计算不同年度不同地区及不同疾病的标化死亡率及其置信区间。死亡数据资料的整理形式见表 1。标准人口的数据整理形式见表 2。

在本例中,除疾病分类变量(Class)在 SAS 中为字符型变量外,其余变量均为数值型变量。在实际应用中,数据资料的整理形式如下:结局变量(Event)及人年变量(Pyear)为数值型变量且变量名称固定;其余变量既可整理成字符型变量,也可整理成数值型变量,

*基金项目:湖北省卫健委卫生健康科研项目(WJ2021M207)

△通信作者:潘敬菊,E-mail: panjjwang@ 163.com

但分析数据集(本例为 Death)及标准数据集(本例为 Standard)中的年龄组变量(本例为 Agegroup)类型需保持一致。

表 1 分年度分地区分疾病的年龄别死亡和人口资料形式

年份	地区	疾病分类	年龄组	死亡数	人年
2013	1	Disease1	0	3	48332
2013	1	Disease1	1	9	295342
2013	1	Disease1	5	9	349696
2013	1	Disease1	10	6	291008
2013	1	Disease1	15	18	360113
2013	1	Disease1	20	35	623965
2013	1	Disease1	25	52	657133
2013	1	Disease1	30	76	580663
2013	1	Disease1	35	139	541543
2013	1	Disease1	40	333	720023
2013	1	Disease1	45	656	728100
2013	1	Disease1	50	724	587896
2013	1	Disease1	55	1232	579623
2013	1	Disease1	60	1424	451329
2013	1	Disease1	65	1513	316504
2013	1	Disease1	70	1439	212410
2013	1	Disease1	75	1504	167014
2013	1	Disease1	80	1014	105506
2013	1	Disease1	85	548	74364
⋮	⋮	⋮	⋮	⋮	⋮
2019	2	Disease2	0	2	41888
2019	2	Disease2	1	12	240322
2019	2	Disease2	5	6	309316
2019	2	Disease2	10	6	281079
2019	2	Disease2	15	9	215238
2019	2	Disease2	20	8	239355
2019	2	Disease2	25	46	448929
2019	2	Disease2	30	52	534884
2019	2	Disease2	35	68	389321
2019	2	Disease2	40	147	386857
2019	2	Disease2	45	374	514165
2019	2	Disease2	50	614	567036
2019	2	Disease2	55	821	463415
2019	2	Disease2	60	1260	351659
2019	2	Disease2	65	1484	310875
2019	2	Disease2	70	1405	213737
2019	2	Disease2	75	1179	139061
2019	2	Disease2	80	724	85398
2019	2	Disease2	85	403	72347

*:SAS 中对应的数据集名称为 Death,变量名称分别为年份 Year,地区 Areas,疾病分类 Class,年龄组 Agegroup,死亡数 Event,人年 Pyear;其中年份有 7 组(2013—2019 年),地区有 2 组(1 和 2),疾病分类有 2 组(Disease1 和 Disease2)。

表 2 标准人口资料形式

年龄组	人年
0	13786434
1	61746176
5	70881549
10	74908462
15	99889114
20	127412518
25	101013852
30	97138203
35	118025959
40	124753964
45	105594553
50	78753171
55	81312474
60	58667282
65	41113282
70	32972397
75	23852133
80	13373198
85	7616148

*:SAS 中对应的数据集名称为 Standard,变量名称分别为年龄组 Agegroup,人年 Pyear。

2.SAS 程序及计算

(1)两组资料标准化率及其置信区间估计

当仅需计算两组资料的标准化率及其置信区间时,分析数据集可只保留一个维度的分组变量(如选取某年两个地区的某种疾病的年龄别死亡和人口资料,如本例中保留表 1 中 2015 年 Disease1 的地区、年龄组、死亡数和人年这 4 个变量),可使用表 3 中的程序直接进行计算,操作如下:

①首先利用 Libname 语句建立并命名一个永久性 SAS 数据逻辑库,本例中将逻辑库命名为 std,语句如下:

```
Libname std 'J:\Study\Stdrate'; /* 逻辑库名称及路径可自行定义 */
```

②将分析数据集(本例中表 1 的 Death)和标准人口数据集(本例中表 2 的 Standard)存储在该逻辑库中。若需从 excel 导入,可选择以下语句:

```
Proc import out = std.Death /* 输出的数据集名 */
datafile = "J:\Study\Stdrate\Death.xlsx" /* 要导入的 excel 文件的完整路径和数据名 */
dbms = Excel replace;
getnames = yes; /* 指出第一行是否有字段名 */
run;
```

③运行表 3 中的程序,输出标准化率及其置信区

间结果

临时逻辑库中的 Std_results 数据集即包含两组资料标准化率及其置信区间结果。更多详细信息可参考 SAS 帮助文档[8]。

3.多维度分组数据标准化率及其置信区间估计的 SAS 宏

本例中多维度指的是年份、地区和疾病分类三个维度;分组指的是年份有≥2 个组别,地区有≥2 个组别,疾病分类有≥2 个组别。

首先建立并命名一个 SAS 数据逻辑库并导入数据,步骤同两组资料标准化率计算的(1)(2)步;

然后在 SAS 中运行表 4 中的宏程序。最后调用宏程序即可,本例中调用宏程序时,宏参数设置如下:

%Stdrate(Libname = std, Filename = Death, Refdata = Standard, Level1 = Year, Level2 = Areas, Level3 = Class)

/ * 逻辑库名称 std,分析数据集 Death 和参考数据集 Standard 可自行定义;多维度分组宏变量 Level1、Level2 和 Level3 可自行定义并根据需要进行增减 * /

实际应用中可根据需要减少或增加维度变量,如分析两组资料的标准化率及其置信区间时,本例中可只保留 Level2 这一个宏变量,同时在 SAS 宏程序中减少宏变量的设置(具体更改方式见程序说明部分),所得结果与利用表 3 中程序计算所得结果一致。

表 3　两组资料标准化率及其置信区间估计的 SAS 程序[8]及说明

基于 STDRATE 过程的直接标化率及其置信区间估计	说明
Ods select Stdrate;	使用 ODS 选择需要输出的结果
Proc stdrate data = std.Death	Data = 指定分析数据集,本例为 Death,存储在 std 逻辑库中
refdata = std.Standard	Refdata = 指定参考数据集,本例中为 Standard,存储在 std 逻辑库中
method = direct	Method = 指定标化方法为直接标化法
stat = rate(mult = 100000)	Stat = 指定用于标准化的统计量,此处表示计算标准化率,并指定单位为/10 万人年
effect;	Effect 展示效应估计值和相关置信区间
population group = Areas event = event total = PYear;	Population = 指定分析数据集中的结局变量和人年变量
reference total = PYear;	Reference = 指定参考数据集中的人年变量
strata agegroup;	Strata = 指定标化过程中的层变量,本例对 agegroup 进行标化
Ods output Stdrate = _Std_rate;	使用 ODS 输出所需的结果至_Std_rate 数据集
Data Std_results;	使用 Data 步中的 Keep 语句保存所需变量
set _Std_rate;	
keep Areas CrudeRate Stdrate LowerCL UpperCL;	
run;	

*:运行程序后,不仅可获得两组资料标准化率及其置信区间估计值,还可获得两组资料标准化率比较的统计检验结果,具体内容可参考 SAS 帮助文档[8]。

表 4　多维度分组数据标准化率及其置信区间估计的 SAS 宏程序及说明

基于 STDRATE 过程的直接标化法的 SAS 宏	说明
%Macro Stdrate(Libname = , Filename = , Refdata = , Level1 = , Level2 = , Level3 =);	定义一个名为 Stdrate 的宏程序,并使用关键字参数创建宏变量,如需增加或减少维度因素(Level1、Level2、Level3),可增加或删除维度宏变量的创建
Proc sort data = &libname..&Filename;	对维度变量(年份、地区和疾病分类)及年龄组进行排序,若维度有所增加或减少,可相应增加或删除宏变量(&Level1、&Level2、&Level3)
by &Level1 &Level2 &Level3 agegroup;	
run;	
Ods select Stdrate;	使用 ODS 选择需要输出的结果
Proc stdrate data = &libname..&Filename	Data = 指定分析数据集,本例为 Death,存储在 std 逻辑库中
refdata = &libname..&Refdata	Refdata = 指定参考数据集,本例中为 Standard,存储在 std 逻辑库中
method = direct	Method = 指定标化方法为直接标化法
stat = rate(mult = 100000);	Stat = 指定用于标准化的统计量,此处表示计算标准化率,并指定单位为/10 万人年
population event = event total = PYear;	Population = 指定分析数据集中的结局变量和人年变量
reference total = PYear;	Reference = 指定参考数据集中的人年变量
strata agegroup;	Strata = 指定标化过程中的层变量,本例对年龄组(agegroup)进行标化
by &Level1 &Level2 &Level3;	By = 指定需进行单独分析的维度,本例中有三个维度(年份、地区和疾病分类),每个维度有多个分组(7 个年份,2 个地区,2 个疾病),若有更改可相应增加或删除维度宏变量(&Level1、&Level2、&Level3)
Ods output Stdrate = _Std_rate;	使用 ODS 输出所需的结果输出至数据集并命名为_std_rate
Proc sql;	使用 sql 语句创建新的数据集 Stdrate,并选择所需的变量
create table Stdrate as	
select &Level1, &Level2, &Level3, CrudeRate, Stdrate, LowerCL, UpperCL	
from _Std_rate;	
quit;	结束宏的定义
%Mend;	

结　果

1.两组资料标准化率及其置信区间估计结果

运行表 3 中的程序后,本例可得到 2015 年 Disease1 的地区 1 和地区 2 的标准化率及其置信区间估计结果(数据集 Stdrate)。结果见表 5。

表 5　两组资料标化死亡率及其置信区间

地区	粗死亡率	标化死亡率	95%下限	95%上限
1	142.50	107.85	105.80	109.91
2	138.48	112.18	109.74	114.63

*:SAS 中对应的数据集名称为 Stdrate,变量名称分别为地区 Areas,粗死亡率 CrudeRate,标化死亡率 StdRate,95%下限 LowerCL,95%上限 UpperCL;率的单位为:1/10 万。

2.多维度分组资料标准化率及其置信区间估计结果

调用宏程序后,Work 临时逻辑库中的数据集 Stdrate 中即包含不同年份不同地区不同疾病的粗死亡率、标化死亡率及置信区间。本例中的部分结果见表 6。

表 6　分年度分地区分疾病的直接标化死亡率及其置信区间

年份	地区	疾病分类	粗死亡率	标化死亡率	95%下限	95%上限
2013	1	Disease1	139.57	112.27	110.13	114.41
2013	1	Disease2	1.26	1.02	0.82	1.23
2013	2	Disease1	107.58	92.02	89.80	94.24
2013	2	Disease2	0.56	0.48	0.32	0.64
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019	1	Disease1	167.84	109.79	107.84	111.74
2019	1	Disease2	1.85	1.18	0.98	1.38
2019	2	Disease1	148.50	104.32	102.06	106.57
2019	2	Disease2	1.60	1.08	0.85	1.30

*:SAS 中对应的数据集名称为 Stdrate,变量名称分别为年份 Year,地区 Areas,疾病分类 Class,粗死亡率 CrudeRate,标化死亡率 StdRate,95%下限 LowerCL,95%上限 UpperCL;率的单位为:1/10 万。

讨　论

近年来,随着我国疾病监测系统的不断发展以及流行病学研究中的大样本队列的不断壮大,可产生几万甚至几十万的复杂数据,如全国或各省的多年死因监测数据,全国或各省大样本的高血压或糖尿病患病率的调查等[10]。在报告疾病死亡率、患病率时,一般都要进行年龄的标准化,以便各地区间进行比较。当研究群体足够大,从而可提供稳定的分层疾病死亡/发生率时(如全国或全省或各县市区),可以使用直接标准化法。此外,当多个研究人群采用同一个参考总体时(如选取全省或全国或世界人口),直接标准化率可在研究人群之间进行比较。面对多维度分组的大样本复杂数据率的标准化及其置信区间的估计,采用传统的统计公式或 Excel 进行计算既复杂又容易出错,此时可借助 SAS 统计分析软件达到快捷地实现率的标准化及置信区间估计的目的。

以往有研究介绍率的标化的 SAS 宏,但未直接给出宏的定义,且输出结果未给出标化率的置信区间[11]。此外,夏雷震等[2]在 2018 年编写了实现大样本数据率的标准化的 SAS 宏程序,但宏的定义过程稍显复杂,且无法同时实现多维度分组资料的率的标准化及其置信区间估计。例如,针对死因监测数据中连续多年、多地区及多种疾病的死亡率的标准化及其置信区间估计,其程序无法完成。在本研究中,可直接在 SAS 中运行表 4 中的宏定义后,再调用%Stdrate 宏命令,即可实现多维度分组(如:某年某地某疾病)大样本数据标准化率及其置信区间的估计。此外,科研工作者或基层工作人员也可根据自己的数据情况和需求,通过增加或减少维度变量(对宏定义中的维度宏变量,做相应的增加或减少)来批量计算不同维度及多个分组数据的标准化率及其置信区间,极大地节约了计算时间并提高了计算的准确性,从而为下一步的科学研究和工作报告的撰写奠定了基础,在实际科研和工作中具有较强的应用价值。

参　考　文　献

[1] 朱焱, 赵耐青. Bootstrap 估计标准化率及其可信区间[J]. 复旦学报:医学版, 2009(3):4.

[2] 夏雷震, Hezi Fu, 金城,等. 大样本数据标准化率的 SAS 宏实现[J]. 中国卫生统计, 2018(4):618-621.

[3] 孙振球,徐勇勇主编. 医学统计学. 第 3 版[M]. 北京:人民卫生出版社,2010,81-85.

[4] Naing NN. Easy way to learn standardization: direct and indirect methods[J]. Malays J Med Sci, 2000, 7(1):10-15.

[5] Victor JS. Standardization of rates and ratios *: Concepts and basic methods for deriving measures that are comparable across populations that differ in age and other demographic variables[EB/OL].(2003-06-12)[2022-11-03]. http://www.epidemiolog.net/evolving/Standardization.pdf.

[6] 李晓松主编. 卫生统计学. 第 8 版[M]. 北京:人民卫生出版社,2017,328-332.

[7] Yuan Y. Paper423-2013:Computing Direct and Indirect Standardized Rates and Risks with the STDRATE Procedure[EB/OL].(2013-07-08)[2022-11-03]. https://support.sas.com/resources/papers/proceedings13/423-2013.pdf.

[8] SAS Institute Inc. 2020. SAS/STAT © 15.2 User's Guide. Cary, NC: SAS Institute Inc.

[9] SAS Institute Inc. 2019. SAS © Certified Professional Prep Guide: Advanced Programming Using SAS © 9.4. Cary, NC: SAS Institute Inc.

[10] 李立明,詹思延主编. 流行病学. 第 8 版[M]. 北京:人民卫生出版社,2017,177-188.

[11] 谷鸿秋, 李卫, 王杨. 流行病学研究中"率"的标化和"率"的校正:方法探讨及 SAS 宏实现[J]. 中华疾病控制杂志, 2014(6):557-560.

(责任编辑:张　悦)