SESUG 2022 Paper 230

Using SAS Macro and ODS Output to efficiently examine the descriptive and analytic statistics in epidemiology studies

Yue Pan, University of Miami Miller School of Medicine

ABSTRACT

The key feature of descriptive and analytic epidemiology is a comparison group, i.e. the exposure, and how it is associated with the study interest, i.e. the outcome. Researchers and investigators are required to present the descriptive and analytic statistics of both exposure and outcomes and test their relationship to describe their study sample. However, if there are many exposures and outcomes to examine, the syntax usually become repetitive and hard to navigate and identify useful results. This paper will present how to use SAS macro and ODS output to efficiently examine and generate descriptive and analytic statistics for epidemiology studies.

This presentation is aimed at beginner to intermediate SAS programmers and healthcare analysts who already have a basic understanding of SAS Macro and ODS Output and are looking to efficiently examine their data.

INTRODUCTION

There are three main types of Epidemiology studies (cohort studies, case-control studies, and cross-sectional studies). All required you to compare the relationship between at least two variables, i.e. primary outcomes, and primary exposures. Others also include multiple secondary outcomes and secondary exposures. In addition, it is commonly required to describe the study samples using descriptive statistics (i.e. mean or median, standard deviation, interquartile range, number, prevalence or proportion, etc) and basic statistical tests (Chi-square test, t-test, Wilcoxon-Mann-Whitney test, etc) for these variables.

I am using hypothetical data as an example to illustrate the steps to examine the descriptive and analytic statistics in epidemiology studies.

STEP 1. GENERATE TABLES FOR DESCRIPTIVE STATISTICS

We first generate the tables for descriptive statistics. We will use PROC Tabulate to generate the numbers and proportions for categorical variables while using mean or median, standard deviation, and interquartile range for continuous variables, depending on their normality test. This will set up the main structure of the descriptive table. In this example, we are interested in stratifying the sample statistics by the primary exposure: druguse_tri variable. We added `MISSING` to show the missingness. We also requested `N` and `COLPCTN` to show the numbers and proportions by druguse_tri. We added all after druguse_tri to also show the descriptive statistics for the overall sample.

The following code is for the descriptive statistics for categorical variables, overall, and stratified by druguse_tri:

```
PROC TABULATE DATA=PR_V2 MISSING ;
CLASS
DRUGUSE_TRI
CONTINOFCARE2
```

```
GENDER
EDU
INCOME
INSURANCE
EVERJAIL
HOMELESS6M
US1
;
TABLE
(CONTINOFCARE2
GENDER
EDU
INCOME
INSURANCE
EVERJAIL
HOMELESS6M
US1
), (DRUGUSE_TRI ALL)*(N COLPCTN);

RUN;
```

The following table is the output for descriptive statistics for categorical variables, overall, and stratified by druguse_tri from SAS:

	druguse_tri						All	
	0		1		2			
	N	ColPctN	N	ColPctN	N	ColPctN	N	ColPctN
continofcare2								
1	6	8.70	24	15.69	49	26.20	79	19.32
2	6	8.70	24	15.69	47	25.13	77	18.83
3	11	15.94	16	10.46	19	10.16	46	11.25
4	46	66.67	89	58.17	72	38.50	207	50.61
sex assigned at birth								
1	40	57.97	112	73.20	163	87.17	315	77.02
2	29	42.03	41	26.80	24	12.83	94	22.98
EDU								
0	23	33.33	64	41.83	59	31.55	146	35.70
1	46	66.67	89	58.17	128	68.45	263	64.30
INCOME								
0	55	79.71	141	92.16	164	87.70	360	88.02
1	14	20.29	12	7.84	23	12.30	49	11.98
Are you covered by health insurance or some other kind of health care plan?								
0	3	4.35	23	15.03	51	27.27	77	18.83
1	66	95.65	130	84.97	136	72.73	332	81.17
ever jail or prison								
0	38	55.07	27	17.65	20	10.70	85	20.78
1	31	44.93	126	82.35	167	89.30	324	79.22
Currently homeless								
.	.	.	1	0.65	.	.	1	0.24
0	54	78.26	105	68.63	78	41.71	237	57.95
1	15	21.74	47	30.72	109	58.29	171	41.81
Have you ever lived in the US?								
.	4	5.80	1	0.65	2	1.07	7	1.71
0	36	52.17	67	43.79	88	47.06	191	46.70
1	29	42.03	85	55.56	97	51.87	211	51.59

We then use PROC Means to generate the results for the continuous variables, stratified by the same variable druguse_tri, and overall respectively. Particularly, we have requested `N MEAN STD MIN MAX MEDIAN QRANGE Q1 Q3` as the results output.

The following code is for the descriptive statistics for continuous variables, overall and stratified by druguse_tri:

```
PROC MEANS DATA=PR_V2 N MEAN STD MIN MAX MEDIAN QRANGE Q1 Q3;
/*CLASS DRUGUSE_TRI;*/
VAR
AGE
VL
CD4
;
RUN;
```

```
PROC MEANS DATA=PR_V2 N MEAN STD MIN MAX MEDIAN QRANGE Q1 Q3 ;
CLASS DRUGUSE_TRI;
VAR
AGE
VL
CD4
;
RUN;
```

The following table is the output for descriptive statistics for continuous variables, overall, and stratified by druguse_tri from SAS:

druguse_tri	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum	Median	Quartile Range	Lower Quartile	Upper Quartile
0	69	AGE	age	69	45.7391304	10.0581430	22.0000000	72.0000000	47.0000000	13.0000000	39.0000000	52.0000000
		VL	Viral Load	68	11555.79	27721.66	2.6600000	158765.00	20.0000000	4172.50	20.0000000	4192.50
		cd4	CD4 count	69	636.0869565	393.4920486	9.0000000	1629.00	587.0000000	560.0000000	335.0000000	895.0000000
1	153	AGE	age	153	47.9934641	8.3599112	24.0000000	69.0000000	49.0000000	12.0000000	42.0000000	54.0000000
		VL	Viral Load	153	20662.47	136027.55	20.0000000	1667569.00	47.0000000	8359.00	20.0000000	8379.00
		cd4	CD4 count	153	569.1176471	357.8642410	6.0000000	1707.00	533.0000000	449.0000000	321.0000000	770.0000000
2	187	AGE	age	187	45.2085561	8.6019710	23.0000000	64.0000000	46.0000000	12.0000000	39.0000000	51.0000000
		VL	Viral Load	187	36743.32	140500.66	20.0000000	1266850.00	2346.00	27767.00	20.0000000	27787.00
		cd4	CD4 count	187	478.4973262	304.0018105	43.0000000	1972.00	422.0000000	366.0000000	253.0000000	619.0000000

Variable	Label	N	Mean	Std Dev	Minimum	Maximum	Median	Quartile Range	Lower Quartile	Upper Quartile
AGE	age	409	46.3398533	8.8489171	22.0000000	72.0000000	47.0000000	12.0000000	40.0000000	52.0000000
VL	Viral Load	408	26515.08	127108.87	2.6600000	1667569.00	164.5000000	14638.00	20.0000000	14658.00
cd4	CD4 count	409	538.9828851	345.3411175	6.0000000	1972.00	472.0000000	449.0000000	276.0000000	725.0000000

STEP 2. ANALYTIC STATISTICS AND UNIVARIATE ANALYSIS

The second step is to test each of the variables' relationship with the primary exposure. In our example, it is druguse_tri. The normal approach is to run a separate model for each of the variables and then gather each of the results together. The syntax usually become repetitive and hard to navigate and identify the useful results. A better way is to use SAS macro and ODS output to efficiently run the analytic statistics and output the univariate analysis results into one single file.

Continued with our previous example. Our primary exposure druguse_tri was a categorical variable. We also have Continofcare2, Gender, Edu, Income, Insurance, Everjail, Homeless6m, Us1 as categorical variables. The Chi-square goodness of fit test allows us to test whether the observed proportions for a categorical variable differ from hypothesized proportions. We then used a Chi-square test to examine the association between druguse_tri and each of the categorical variables.

In this step, we first create an empty dataset CHISQ. The purpose is to use it later to save and hold the results output from the macro. We then created a macro. For the Chi-square test, only one macro variable &VAR. is needed to loop through all the categorical variables. Noticed that we added an "output out" statement to particularly export the Chi-square test results (i.e. PCHI) to a temporary dataset: STATS. Finally, we used a data step to append the result to the CHISQ dataset. Therefore, every time the macro change to a different &VAR, STATS dataset will be overwritten to reflect the new one. While after each macro, it is appended and saved to the CHISQ dataset.

The following code is for the analytic statistics for categorical variables and druguse_tri:

```
DATA CHISQ;
RUN;

%MACRO AA(VAR);
TITLE"CHISQ FOR &VAR.";
PROC FREQ DATA=PR_V2 ;
TABLES (&VAR.)*DRUGUSE_TRI /CHISQ CMH NOCOL NOPERCENT ;

OUTPUT OUT=STATS PCHI;
RUN;

DATA CHISQ;
SET CHISQ STATS;
RUN;

%MEND;
%AA( CONTINOFCARE2   );
%AA( GENDER     );
%AA( EDU   );
%AA( INCOME     );
%AA( INSURANCE  );
%AA( EVERJAIL   );
%AA( HOMELESS6M );
%AA( US1   );
```

The following table is the output for the PROC Freq statement, as an example, we showed the results of CONTINOFCARE2.

The FREQ Procedure

Frequency Row Pct	Table of continofcare2 by druguse_tri			
	druguse_tri			
continofcare2	0	1	2	Total
1	6 7.59	24 30.38	49 62.03	79
2	6 7.79	24 31.17	47 61.04	77
3	11 23.91	16 34.78	19 41.30	46
4	46 22.22	89 43.00	72 34.78	207
Total	69	153	187	409

Statistics for Table of continofcare2 by druguse_tri

Statistic	DF	Value	Prob
Chi-Square	6	30.4717	<.0001
Likelihood Ratio Chi-Square	6	31.6894	<.0001
Mantel-Haenszel Chi-Square	1	26.5379	<.0001
Phi Coefficient		0.2730	
Contingency Coefficient		0.2633	
Cramer's V		0.1930	

Sample Size = 409

Summary Statistics for continofcare2 by druguse_tri

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	26.5379	<.0001
2	Row Mean Scores Differ	3	29.1341	<.0001
3	General Association	6	30.3972	<.0001

The following table is the output for the temporary STATS dataset, which includes _PCHI_ for Chi-square test value, DF_PCHI for degrees of freedom, and P_PCHI for p-value.

PCHI	DF_PCHI	P_PCHI
30.471679272	6	0.0000319685

The following table is the output for the temporary CHISQ dataset

PCHI	DF_PCHI	P_PCHI
.	.	.
30.471679272	6	0.0000319685
26.279307411	2	1.9657166E-6
4.0756058516	2	0.1303147077
7.0195198413	2	0.0299040929
19.63549085	2	0.0000544763
61.750881408	2	3.899119E-14
39.67585468	2	2.4238047E-9
2.3340151857	2	0.3112970765

Similarly, we can also change the code for the continuous variables. Depending on their normality test, we choose to use parametric or non-nonparametric tests for the continuous variables.

Here, I am using a non-parametric test as an example, based on the normality test (not included in the paper) for the variables. Particularly, I am interested in reporting the Kruskal Wallis test. The Kruskal Wallis test is used when you have one independent variable with two or more levels and an ordinal dependent variable. In other words, it is the non-parametric version of ANOVA. It is also a generalized form of the Mann-Whitney test method, as it permits two or more groups.

In this step, we first create an empty dataset WIL to store the results output later. We then created a macro, for the Kruskal Wallis test, with only one macro variable &VAR. to loop through all the continuous variables. Noticed that we added an "output out" statement to particularly export the Kruskal Wallis test (i.e. WILCOXON) to a temporary dataset: STATS. Finally, we used a data step to append the result to the WIL dataset. Therefore, every time the macro change to a different &VAR, STATS dataset will be overwritten to reflect the new one, but after each step, it is appended and saved to the WIL dataset.

The following code is for the analytic statistics for continuous variables and druguse_tri:

```
DATA WIL;
RUN;
%MACRO BB(VAR);
PROC NPAR1WAY DATA=PR_V2;
CLASS DRUGUSE_TRI;
VAR &VAR.;

OUTPUT OUT=STATS WILCOXON;
RUN;

DATA WIL;
SET WIL STATS;
RUN;

%MEND;

%BB( AGE  );
%BB( VL   );
%BB( CD4  );
```

The following table is the output for the temporary STATS dataset (results for CD4), which includes _VAR_ for the variable name, _KW_for Kruskal Wallis test value, DF_KW for degrees of freedom, and P_KW for p-value.

VAR	_KW_	DF_KW	P_KW
cd4	11.093015012	2	0.0039010579

The following table is the output for the temporary WIL dataset

VAR	_KW_	DF_KW	P_KW
	.	.	.
AGE	9.5337193481	2	0.0085070532
VL	21.392371635	2	0.0000226311
cd4	11.093015012	2	0.0039010579

STEP 3. COMBINE RESULTS FROM STEP1 AND STEP2

The last step is to combine the results from STEP1 and STEP2, by adding the analytic results to the descriptive statistics. After editing the title and labels, this usually concludes the Table 1 results in an epidemiological paper.

CONCLUSION

This paper presents how to use SAS macro and ODS output to efficiently examine the descriptive and analytic statistics of epidemiology studies. The macro and the ods output is an efficient tool to help researchers and investigators to select the analytic statistical results and create the descriptive statistics table for an epidemiology study.

RECOMMENDED READING

- *https://support.sas.com/resources/papers/proceedings/proceedings/sugi29/243-29.pdf*
- *https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/procstat/procstat_freq_details124.htm*

ACKNOWLEDGMENTS

The author would like to thank all supporters and organizers of this conference. Special thanks to Barbara Okerson, SESUG 2022 Academic Chair for her support and advice.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Yue Pan
University of Miami Miller School of Medicine
panyue@miami.edu
https://www.linkedin.com/in/yue-pan-67283522/