Paper 062-2012

A Macro to Summarize and Generate a Report of Variables on a SAS® Data Set

Greg Grandits, Division of Biostatistics, University of Minnesota, Minneapolis, MN
Greg Thompson, Division of Biostatistics, University of Minnesota, Minneapolis, MN

ABSTRACT

Before starting analysis of data from a SAS® data set it is useful to have documentation and summary information of each variable before you begin working with the data set. We have developed a macro that generates a report summarizing each variable on a data set, one line per variable, which includes both numeric and character variables, and formats date variable statistics with date formats. A nicely formatted report is generated that gives the user a good understanding of the data set and is a useful reference when beginning to write SAS programs using the data set.

INTRODUCTION

Before starting analysis of data from a SAS data set it is useful to have documentation and summary information of each variable on the data set. The CONTENTS procedure provides documentation of each variable, listing the name, label, format, and whether the variable is character or numeric. However, it gives no information about statistics of the variable, like counts, means, standard deviations, and minimums and maximum values. The MEANS procedure provides a good summary of numeric variables but does not summarize character variables. In addition, the MEANS procedure does not summarize date variables very well since the statistics displayed are not formatted. For many studies dates are important variables to summarize; formatted mean, minimum and maximum values for dates can be very helpful.

In the Division of Biostatistics at the University of Minnesota we have developed a macro that generates a report summarizing each variable on a data set, one line per variable, which includes both numeric and character variables, and formats date variable statistics with date formats. Numeric variables are summarized with number of non-missing values, minimum and maximum values, and the mean and standard deviation. For numeric date variables, the mean, minimum and maximum values are formatted with date formats. Character variables are summarized with number of non-missing values, and the minimum and maximum values. For all variables, the format associated with the variable (if present) is displayed, and the length of the variable. A nicely formatted report is generated that gives the user a good understanding of the data set and is a useful reference when beginning to write SAS programs using the data set. The user can print this report or view it on their computer as they begin to work with the data set.

MACRO USE

The call to the macro is simple, the user simply specifies the data set and library reference, with some optional parameters related to the type of report files generated. The macro uses the CONTENTS, MEANS, and TRANSPOSE procedures and a data step to generate and gather the needed statistics. The REPORT procedure is then used to generate the report.

```
%MACRO CONDES (dataset,pdf=Y,html=N)
```

dataset -	Name of SAS data set including the libname if other than WORK.
pdf -	If set to Y then a PDF report is generated as well as the text report.
html -	If set to Y then an HTML reported is generated as well as the text report

REPORT CREATED

The appendix gives an example of the text report generated of the SAS data set. There is one row per variable and after every five variables a line is skipped. This makes for easy visual reading of the report. The sequence number is the order of the variable on the data set. This is useful for reference and discussion between users of the data set. A small column indicates whether the variable is numeric (N) or character (C), and the length of the variable. This is useful for quickly identifying which variables are characters and their corresponding lengths, which can sometimes be surprisingly bigger than expected. Any associated format is also displayed.

Note that date variables are formatted with a date format for the minimum, maximum, and mean values. The standard deviation is displayed in its unformatted numeric value which is the variation in days. Whatever date format is associated with the variable the same format is used to display the statistics. Character variables such as clinical site (values A-D) do not have a mean or SD, but do have a minimum, maximum, and number of non-missing values displayed. The text under the description is taken from the label assigned to the variable. With these summaries one can learn much about the data set: when subjects were enrolled, the distribution of blood pressure, and the percentage of subjects smoking cigarettes. The user can also get a good feel for the amount of missing data across the variables. In the example report we see that there are 902 subjects on the data set, that there is no missing data collected on Form 10 (baseline form); there are 3 subjects that did not complete the smoking question on From 30, and that there are probably 98 smokers in the study (based on the question “amount smoked”, answered only by current smokers). The form 200 data has more missing data as this was a follow-up visit.

PROGRAM OUTLINE

To use the macro it is not required to understand the programming syntax. However, a basic understanding may be useful for modifying the program, if desired, and for learning methods for other programming. The program is approximately 200 lines of code. The SAS data set for which a report is generated is read once for the MEANS procedure and once to calculate statistics for character variables (a KEEP statement is used to keep only character variables to improve efficiency). In most cases the macro will run very quickly.

A basic description of the program follows these steps:

- The OUTPUT statement is used from the CONTENTS procedure to have access to the variable name, format, label, type, and length.
- The OUTPUT statement from the MEANS procedure is used to generate a data set containing the statistics for numeric variables.
- A data step is used to generate a data set with the statistics for character variables.
- The two statistics data sets are merged with the contents data set, sorted by variable name. After being merged the data set is then sorted by variable number, i.e. the order on the data set.
- The REPORT procedure is used to generate the report. For PDF output a modified journal style is used.

CONCLUSION

This macro has proven to be a useful tool for statisticians here at the Division of Biostatistics. Whenever a data set is created a corresponding documentation report is generated. The report generated from the macro has become the icon of the data set, the place where the user starts. When we send a data set to a user outside the Division we send the descriptive file along with the data set. From our experience users like and appreciate having the documentation report file. It is certainly possible to add additional information for each variable to the report (such as frequency counts for categorical variables, modes for character variables, and other features). However, we think the simplicity and readability of the current report is a good balance of information and utility.

The macro code is available at www.biostat.umn.edu/~greg-g.

CONTACT INFORMATION

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

Greg Grandits
Division of Biostatistics, University of Minnesota
2221 University Avenue
Suite 200
Minneapolis, MN 55414
612-626-9033
grand001@umn.edu
www.biostat.umn.edu/~greg-g

APPENDIX (Report Generated)

Seq	Name	T	Format	Variable Label	N	Mean	Std Dev	Minimun	Maximum
1	patient_id	C6		Subject ID	902			A00001	D02136
2	clinic	C1	$CLINIC.	Clinical Center	902			A	D
3	drug_group	N8	GROUP.	Rand: Drug Group	902	3.788248	1.787413	1	6
4	bottle_number	C6		Rand: Bottle Number Assigned	902			A0001	D0189
5	rand_date	N8	DATE9.	Rand: Randomization Date	902	20AUG87	144.3987	28OCT86	31MAR88
6	death_date	N8	DDMMYY10.	Death Date	10	08/01/90	517.2403	16/04/88	21/04/92
7	f010_date	N8	MMDDYY10.	Form 10: Date	902	06/03/87	134.4276	09/02/86	12/18/87
8	f010_sex	N8		Form 10: Sex (1=M,2=F)	902	1.382483	0.486263	1	2
9	f010_race	N8		Form 10: Ethnicity	902	1.238359	0.493861	1	4
10	f010_dob	N8	MMDDYY10.	Form 10: Birth Date	902	02/21/32	2352.264	02/14/17	11/13/42
11	f010_sbp	N8		Form 10: Systolic BP	902	137.9113	15.57357	99	223
12	f010_dbp	N8		Form 10: Diastolic BP	902	88.10865	7.29567	59	99
13	f010_bpmeds	N8		Form 10: On BP Meds	902	1.391353	0.488324	1	2
14	f010_source_meds	C4		Form 10: Where Meds Obtained	902			0080	7000
15	f030_eversmoke	N8		Form 30: Ever Smoke Cigarettes	899	1.532814	0.4992	1	2
16	f030_nowsmoke	N8		Form 30: Curretly Smoke Cigarettes	424	1.768868	0.422055	1	2
17	f030_quansmoke	N8		Form 30: Amount Smoke Cigarettes	98	17.14286	12.58292	1	60
18	f030_cigars	N8		Form 30: Currently Smoke Cigars	898	1.949889	0.218296	1	2
19	f030_ursod	N8		Form 30: Urinary Sodium Excretion	883	53.67452	27.94486	3.8	197.5
20	f030_urpot	N8		Form 30: Urinary Potassium Excretion	883	15.41121	7.531775	3.1	81.7
21	f200_date	N8	MMDDYY10.	Form 200: Date	897	02/23/90	146.1651	04/25/89	11/24/90
22	f200_wt	N8		Form 200: Weight (lb)	793	179.5032	30.55028	111	286
23	f200_pulse	N8		Form 200: Pulse	792	34.85354	5.04631	21	55
24	f200_dbp	N8		Form 200: Diastolic BP	793	78.40227	8.210356	47	107
25	f200_sbp	N8		Form 200: Systolic BP	793	126.1059	15.78146	87	195
26	f200_fever	N8		Form 200: Experience Fever?	805	1.069565	0.334769	1	4
27	f200_sweat	N8		Form 200: Experience Sweating?	805	1.036025	0.228436	1	4
28	f200_drows	N8		Form 200: Experience Drowsiness?	805	1.085714	0.317573	1	3
29	f200_tired	N8		Form 200: Experience Tiredness?	805	1.173913	0.419743	1	3
30	f200_weak	N8		Form 200: Experience Weakness?	805	1.034783	0.196443	1	3