

Using SAS® Macros to Extract P-values from PROC FREQ

Rachel Straney, University of Central Florida

ABSTRACT

This paper shows how to leverage the SAS ® Macro Facility with PROC FREQ to obtain multiple chi-square test statistics and their associated p-values into one data set to achieve a quick solution to the common variable selection problem. The purpose of this paper is to provide a simplified macro function that can be used to identify important factors in a study. Although the use of PROC FREQ in this macro limits its use to categorical data, references to other SAS papers will be summarized for the readers to get a better understanding of how this concept can be expanded upon.

INTRODUCTION

Analysts and statisticians using data sets with large volumes of variables commonly run into the challenge of having too many factors to focus on and the scope can be overwhelming. This is particularly the case in situations where the goal is to identify which factors are related to a particular dependent variable. This paper shows how to leverage the SAS ® Macro Facility with PROC FREQ to obtain multiple chi-square test statistics and their associated p-values into one data set to achieve a quick solution to the common variable selection problem. There have been many past SAS papers written which have done similar tasks, however most are fairly specific to their field or their purpose is to simplify reporting of information. The purpose of this paper is to provide a simplified macro function, %COMPARE_DIST, which can be used to identify important factors in a study. Although the use of PROC FREQ in this macro limits its use to categorical data, references to other SAS papers will be summarized for the readers to get a better understanding of how this concept can be expanded upon.

LIMITATIONS

This SAS macro was written to easily evaluate multiple Pearson's Chi-square Tests of Independence against a particular target variable of interest, and so only categorical variables can be considered. Since the Chi-square test is sensitive to sample size, this macro should be used in situations where the number of observations in the data set is relatively small. Most importantly, this macro should be used to supplement the data exploration process and not solely relied upon to identify relationships between variables.

DATA USED IN MACRO EXAMPLE

The Graduating Student Survey is administered to graduating undergraduates at a university every year and asks students to rate their experiences while earning their degree. Since data is collected on a recurring basis, a common question is:

Are there differences over time in graduate perceptions regarding the services and academic support received while attending the university?

The initial data source consists of survey responses from graduates earning a degree during 2011-2012 to 2015-2016. There are 88 variables in the data set that individually correspond to a survey item. As mentioned previously, the Chi-square test is sensitive to sample size and a large number of observations can lead to meaningless results. To appropriately use the %COMPARE_DIST macro, we will conduct analysis for a subset of students who earned a degree in Civil Engineering (n = 521). The name of the SAS data set used in the macro example is GSS_PROGRAM.

To identify any changes in response to a particular survey item, we can perform a Pearson's Chi-square Test of Independence using that survey item and ACAD_YEAR_AWARDED, a variable which indicates the academic year the student earned their degree. To quickly summarize results for multiple Chi-square tests, we can use the %COMPARE_DIST SAS macro.

SAS OUTPUT FROM THIS MACRO

There are two types of output provided by the %COMPARE_DIST macro.

1. The macro will always print the final data set, CHISQ_ALL, which houses all Chi-square test results (PROC PRINT). Variables printed from the CHISQ_ALL data set include:

Variable	Description
N	Number of observations used to perform Chi-square test
CHI_TEST_STAT	Test statistic from Chi-square test
CHI_PVALUE	P-value from Chi-square test
VARNAME	Variable used in the Chi-square test (against the target variable)
CHI_DEG_FREEDOM	Degrees of Freedom for the test
CHI_WARNING	SAS warnings due to low expected frequency counts, if applicable (more than 20% of table cells have expected frequencies less than 5)
CHI_RESULT	Description of Chi-square test result

2. The macro will selectively print contingency tables for variables that have a significant relationship with the target variable (PROC TABULATE). These “SIGNIFICANT RESULT” tables summarize the joint distribution of the variables by displaying column percentages for the column variable.

THE %COMPARE_DIST MACRO

The following section describes in detail how the %COMPARE_DIST macro works. Any output generated by SAS is provided in the associated steps.

*STEP 1

Define a macro variable &TARGET using the name of a variable of interest (target) you wish to run multiple Chi-square tests against other variables in your data set;

```
%LET TARGET = ACAD_YEAR_AWARDED;
```

*STEP 2

Create an empty data set, CHISQ_ALL to hold all results from your Chi-square tests;

```
DATA CHISQ_ALL;
```

```
  LENGTH N CHI_TEST_STAT CHI_PVALUE 8 VARNAME $15;
```

```
  STOP;
```

```
run;
```

*STEP 3

Define the COMPARE_DIST macro;

```
%MACRO COMPARE_DIST (VAR) ;
```

*STEP 4

Define a macro variable, &PRINT_TABLE, which is later used to print contingency tables for any variables found to have a significant relationship to your target variable;

```
%LET PRINT_TABLE = 0;
```

***STEP 5**

Run PROC FREQ to conduct a Chi-square test using your target variable and the secondary variable passed as an argument in the macro call;

```
PROC FREQ DATA = GSS_PROGRAM NOPRINT;
```

```
  *The WARN=OUTPUT option will save an indicator variable that flags
  when more than 20% of the table cells have expected frequencies
  less than 5 during the test;
```

```
  TABLE &VAR*&TARGET / CHISQ WARN=OUTPUT;
```

```
  *Save results from the Chi-square test to a temporary data set,
  using the name of your secondary passed variable;
```

```
  OUTPUT OUT = CHI_&VAR N PCHI;
```

```
RUN;
```

***STEP 6**

Create a few new variables in your temporary data set for ease of interpretation;

```
DATA CHI_&VAR;
```

```
SET CHI_&VAR (RENAME=(_PCHI=CHI_TEST_STAT P_PCHI=CHI_PVALUE
DF_PCHI=CHI_DEG_FREEDOM));
```

```
  LENGTH VARNAME $15. CHI_WARNING $50. CHI_RESULT $60.;
```

```
  *Insert variable name that was tested against the target;
```

```
  VARNAME = SYMGET("VAR");
```

```
  *Insert variable describing whether there was a SAS warning due to
  more than 20% of table cells having expected frequencies less
  than 5;
```

```
  IF WARN_PCHI = 1 THEN CHI_WARNING = "Pearson Chi-square may not be a
  valid test.";
```

```
  *Insert variable describing the result of the Chi-square test;
```

```
  IF (WARN_PCHI = 0 AND CHI_PVALUE LT .05) THEN DO;
```

```
    CHI_RESULT = "Evidence to suggest these variables are not independent.";
```

```
  *If Chi-square test is valid and significant, change PRINT_TABLE to 1;
```

```
  CALL SYMPUT('PRINT_TABLE' ,1);
```

```
END;
```

```
DROP WARN_PCHI;
```

```
RUN;
```

***STEP 7**

Append Chi-square result to the initially created data set CHISQ_ALL;

```
DATA CHISQ_ALL;
```

```
  SET CHISQ_ALL CHI_&VAR;
```

```
RUN;
```

***STEP 8**

If Chi-square test is valid and significant (PRINT_TABLE = 1), print contingency table with column percentages;

```
%IF &PRINT_TABLE = 1 %THEN %DO;
```

```
PROC TABULATE DATA = GSS_PROGRAM;
```

```
  CLASS &VAR &TARGET;
```

```
  TABLE &VAR=" " ALL, &TARGET* (N="Count" COLPCTN="Col %") ALL* (N="Count"
  COLPCTN="Col %") / BOX=&VAR;
```

```
  TITLE "SIGNIFICANT RESULT: &&VAR BY &&TARGET";
```

```
RUN;
```

```
%END;
```

NOTE: Output from Step 8 is shown on next page

Output 1. Output from PROC TABULATE in the %COMPARE_DIST macro

SIGNIFICANT RESULT: CATALOG BY ACAD_YEAR_AWARDED

CATALOG	Academic Year Awarded										All	
	2011-12		2012-13		2013-14		2014-15		2015-16			
	Count	Col %	Count	Col %	Count	Col %	Count	Col %	Count	Col %	Count	Col %
1	2	1.69	4	3.57	.	.	.	.	.	.	6	1.26
2	8	6.78	8	7.14	9	10.23	1	1.35	9	10.47	35	7.32
3	50	42.37	39	34.82	30	34.09	19	25.68	23	26.74	161	33.68
4	37	31.36	49	43.75	33	37.50	26	35.14	24	27.91	169	35.36
5	21	17.80	12	10.71	16	18.18	28	37.84	30	34.88	107	22.38
All	118	100.00	112	100.00	88	100.00	74	100.00	86	100.00	478	100.00

SIGNIFICANT RESULT: LISTEN BY ACAD_YEAR_AWARDED

LISTEN	Academic Year Awarded										All	
	2011-12		2012-13		2013-14		2014-15		2015-16			
	Count	Col %	Count	Col %	Count	Col %	Count	Col %	Count	Col %	Count	Col %
1	3	2.38	3	2.46	2	2.06	2	2.67	1	1.12	11	2.16
2	24	19.05	12	9.84	3	3.09	2	2.67	3	3.37	44	8.64
3	37	29.37	47	38.52	36	37.11	21	28.00	31	34.83	172	33.79
4	45	35.71	43	35.25	41	42.27	31	41.33	34	38.20	194	38.11
5	17	13.49	17	13.93	15	15.46	19	25.33	20	22.47	88	17.29
All	126	100.00	122	100.00	97	100.00	75	100.00	89	100.00	509	100.00

SIGNIFICANT RESULT: SUPPORT BY ACAD_YEAR_AWARDED

SUPPORT	Academic Year Awarded										All	
	2011-12		2012-13		2013-14		2014-15		2015-16			
	Count	Col %	Count	Col %	Count	Col %	Count	Col %	Count	Col %	Count	Col %
1	1	0.78	3	2.44	1	1.02	2	2.74	1	1.10	8	1.56
2	12	9.38	9	7.32	5	5.10	6	8.22	6	6.59	38	7.41
3	42	32.81	48	39.02	25	25.51	18	24.66	16	17.58	149	29.04
4	51	39.84	45	36.59	38	38.78	25	34.25	31	34.07	190	37.04
5	22	17.19	18	14.63	29	29.59	22	30.14	37	40.66	128	24.95
All	128	100.00	123	100.00	98	100.00	73	100.00	91	100.00	513	100.00

* STEP 9

Delete temporary data set with Chi-square result for current variable that was passed to macro;

```
PROC DATASETS LIBRARY=WORK NOLIST;
```

```
  DELETE CHI_&VAR;
```

```
RUN;
```

```
%MEND;
```

*STEP 10

Call %COMPARE_DIST macro for all variables interested in running Chi-square test against the target variable;

```
%COMPARE_DIST(OVERALL);
```

```
%COMPARE_DIST(RECOMM);
```

```
%COMPARE_DIST(SUPPORT);
```

```
%COMPARE_DIST(CATALOG);
```

```
%COMPARE_DIST(LEARN);
```

```
...
```

NOTE: Although data set GSS_PROGRAM has 88 variables, only 10 are shown to save space in this paper

```
%COMPARE_DIST(SPEAK);
%COMPARE_DIST(LISTEN);
%COMPARE_DIST(PROFPRAC);
%COMPARE_DIST(OMBUDS);
%COMPARE_DIST(FRIENDS);
```

```
*STEP 11
Sort data set with all Chi-square results;
PROC SORT DATA = CHISQ_ALL;
  BY CHI_WARNING CHI_PVALUE;
RUN;
```

```
*STEP 12
Print all Chi-square results;
PROC PRINT DATA = CHISQ_ALL NOOBS;
  TITLE "PEARSON CHI-SQUARE RESULTS : SORTED BY CHI_WARNING
        AND CHI_PVALUE";
RUN;
```

NOTE: Output from Step 12 is shown below

Output 2. Output from PROC PRINT in the %COMPARE_DIST macro

PEARSON CHI-SQUARE RESULTS : SORTED BY CHI_WARNING AND CHI_PVALUE

N	CHI_TEST_STAT	CHI_PVALUE	VARNAME	CHI_DEG_FREEDOM	CHI_WARNING	CHI_RESULT
478	45.9883	0.0001	CATALOG	16		Evidence to suggest these variables are not independent.
509	35.96	0.00293	LISTEN	16		Evidence to suggest these variables are not independent.
513	33.1508	0.00706	SUPPORT	16		Evidence to suggest these variables are not independent.
500	23.5093	0.10078	OMBUDS	16		
505	22.3013	0.13373	LEARN	16		
507	20.2811	0.20782	SPEAK	16		
508	19.7652	0.23098	FRIENDS	16		
506	41.5835	0.00046	PROFPRAC	16	Pearson Chi-square may not be a valid test.	
517	41.0531	0.00055	RECOMM	16	Pearson Chi-square may not be a valid test.	
516	31.0412	0.01329	OVERALL	16	Pearson Chi-square may not be a valid test.	

RESOURCES FOR EXPANDING ON THIS MACRO

This section shares a number of other SAS papers that provide examples of how to expand on the %COMPARE_DIST macro. SAS programs using similar logic have been written to incorporate numeric type data in addition to character type. Other programs use these concepts to format and tabulate results for specific reporting needs. All papers can be referenced at the end of this paper.

A QUICK AND DIRTY DESCRIPTIVE DATA SUMMARY

NORA H. RUEL AND REBECCA A. NELSON

The macro in this paper can combine summary statistic results for both character and numeric data and display all the results in one table. The character type variables are summarized using output from PROC FREQ whereas the numeric variable types use output from PROC UNIVARIATE and PROC NPAR1WAY. The final table that displays all summary statistics is achieved using PROC REPORT.

GENERATING CUSTOMIZED ANALYTICAL REPORTS FROM SAS PROCEDURE OUTPUT

BRINDA BHASKAR AND KENNAN MURRAY

This paper provides two macros that are particularly useful when statistics must be summarized for a large number of variables. One macro can be used on character data and the other on numeric data types. The statistical tests summarized in the output include Chi-square tests from PROC FREQ and t-tests from PROC TTEST.

P-VALUE GENERATION SIMPLIFIED WITH A SINGLE SAS® MACRO

PETE ANDERSON AND CHRIS HORD

This paper explains how to write a SAS macro which incorporates p-values from a number of statistical tests (Chi-square, Cochran-Mantel-Haenszel, Fisher's exact, Kruskal-Wallis and a rank ANOVA). The macro merges p-values with other descriptive statistics from the original data set and displays it in an easy to read table. The macro can be easily altered to include additional statistical tests.

IS YOUR FAILED MACRO DUE TO MISJUDGED "TIMING"?

ARTHUR LI

Although this paper focuses on the SAS macro facility and how it interacts with DATA step execution, some of the examples are written to achieve the summary of multiple summary statistics into one data set for reporting. It also serves as a good resource to understand how to effectively use the SAS macro language.

CONCLUSION

The %COMPARE_DIST macro can be used to quickly summarize results for multiple Chi-square tests. The macro stores multiple chi-square test statistics and their associated p-values into one data set and summarizes the results using PROC PRINT. It also provides contingency tables for any variables found to have a significant relationship with a target variable of interest using PROC TABULATE. The concepts behind this macro can be generalized to include other statistical tests or other data types.

REFERENCES

Anderson, Pete and Hord, Chris. 2003. "P-Value Generation Simplified with a Single SAS® Macro." *Proceedings of the SAS Users Group International 28 Conference*. Seattle, Washington. Available at <http://www2.sas.com/proceedings/sugi28/209-28.pdf>

Bhaskar, Brinda and Murray, Kennan. 2004. "Generating Customized Analytical Reports from SAS Procedure Output." *Proceedings of the Northeast SAS Users Group 17 Conference*. Baltimore, Maryland. Available at <http://www.lexjansen.com/nesug/nesug04/ap/ap15.pdf>

Li, Arthur. 2015. "Is Your Failed Macro Due To Misjudged "Timing"?" *Proceedings of the PharmaSUG 2015 Conference*. Orlando, Florida. Available at <http://www.pharmasug.org/proceedings/2015/BB/PharmaSUG-2015-BB07.pdf>

Ruel, Nora H. and Nelson, Rebecca A. "A Quick and Dirty Descriptive Data Summary." 2013. *Proceedings of the Western Users of SAS Software Conference*. Las Vegas, Nevada. Available at http://www.lexjansen.com/wuss/2013/35_Paper.pdf

CONTACT INFORMATION

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

Rachel Straney
University of Central Florida
12424 Research Parkway, Suite 225
Orlando, FL 32826
407-882-0280
rstraney@ucf.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.