PROC PARCAT: A PROCEDURE FOR TESTING AVERAGE PARTIAL ASSOCIATION IN THREE-WAY CONTINGENCY TABLES

Ronald G. Fleming, Mead Johnson and Company
J. Richard Landis, University of Michigan
Douglas P. Schnautz, Mead Johnson and Company

ABSTRACT

PROC PARCAT is a SAS procedure which provides tests of average partial association in three-way contingency tables within the framework of the multiple hypergeometric probability model. Primary attention is directed at the relationship between two of the variables, controlling for the effects of a covariable. This approach is essentially a multivariate extension of the Cochran-Mantel-Haenszel test to sets of $(s \times r)$ tables. Scores can be assigned to categories which are ordinally scaled. In particular, if ridit scores with midranks assigned for ties are utilized, this procedure is equivalent to a partial Kruskal-Wallis test when one variable is ordinally scaled, and is equivalent to a partial Spearman rank correlation test when both variables are ordinally scaled.

INTRODUCTION

The original computer program, PARCAT (Landis, et.al., 1979), was written in FORTRAN to implement this methodology as a stand-alone program. PROC PARCAT, a user written SAS PROC created by Ron Fleming and Doug Schnautz, enables SAS users to more easily use these techniques. Although the original program is relatively unchanged, minor alterations have been made in order to 1) permit the program to use data from a SAS dataset, 2) reduce the number of required input parameters, 3) establish an output SAS dataset containing the results of the analysis.

AREA OF APPLICATION

Certain research situations exist in which the basic data are measured in terms of discrete categories based on nominal or ordinal scales. Primary attention is directed at the relationship between two of the variables (e.g., treatment and response), controlling for the effects of a third variable, better known as the covariable (e.g., investigator or initial severity).

Let $i = 1, 2, \ldots, s$ index a set of distinct sub-populations and let $j = 1, 2, \ldots, r$ index the outcome categories associated with the dependent variable under study. In addition, let $h = 1, 2, \ldots, q$ index a set of $(s \times r)$ contingency tables which correspond to distinct levels of a relevant covariable or combination of several covariables. As a result, the data obtained from such studies can be summarized in a set of $q:(s \times r)$ contingency tables. In this formulation, the basic hypothesis can then be expressed in terms of 'no partial association' between the sub-populations and the response profiles, after adjusting for the levels of the covariable. For example, suppose either treatment A or B is assigned at random to subjects within each of a series of q clinics and the response to treatment is classified as either improved or not improved. In this situation, the resulting data can be displayed in a set of $q:(2 \times 2)$ tables for which the primary question is whether, on the average, across the q clinics, the percentages of subjects classified as improved are the same for the two treatments.

Cochran (1954) proposed a test statistic for this hypothesis with respect to a set of (2×2) tables from the point of view of asymptotic binomial model results (which require moderately large sample sizes, e.g., $n_h \geq 20$ for each table). Alternatively, Mantel and Haenszel (1959) noted that this same problem could be approached within the framework of a hypergeometric model which requires only the overall sample size $n = \Sigma_h n_h$ to be reasonably large for asymptotic methods to be applicable. In fact, their procedure is appropriate for matched case-control studies with only two subjects in each of the q tables. For other cases, the Mantel-Haenszel method is preferred because it only requires asymptotic considerations on an across-table basis rather than both across and within tables.

More recently, Landis, Heyman and Koch (1978) summarized and contrasted various alternative approaches for investigating the underlying concept of 'average partial association' in three-way contingency tables. In particular, they presented a unified notation and matrix formulation for the Generalized Cochran-Mantel-Haenszel (CMH) approach to the analysis of $q:(s \times r)$ contingency tables in terms of the corresponding multiple hypergeometric probability model. The purpose of this paper is to describe a new SAS procedure, PROC PARCAT, which implements the CMH analysis of three-way contingency tables.

PROCEDURE PARCAT

The frequency counts for each cell in each of the q tables must be input into PROC PARCAT (generally using an output dataset from PROC

FREQ). The original program has been altered to reduce the number of required input parameters. For example, the number of levels of the covariate, sub-population, and response variables are determined within the procedure. However, the user must comply with certain constraints:

1) The frequency counts must be sorted by covariate and sub-population (BY VARIABLES also require sorting, if they are being used).

```
PROC SORT; BY COVAR SUBPOP;
```

Example: Treatment A or B is assigned at random to two clinics and the response to treatment is either improved or not improved. The sorted data set must look like

Clinic (COVAR)	Treatment (SUBPOP)	Improved (VAR1) (Freq count)	Not Improved (VAR2) (Freq count)
1	A	5	2
1	B	4	3
2	A	6	1
2	B	3	4

where the overall sample size is 28.

2) The above dataset is readily obtained as an output SAS dataset from PROC FREQ. The SPARSE option should be used to insure that all cells are represented in this output dataset. Also, any frequency counts of "missing" should be set equal to zero (0) before processing the dataset with PROC PARCAT.

3) Choose appropriate scores for the response variable and sub-populations (if applicable).

4) If the number of sub-populations times the number of responses is greater than 100, then the calculation of the multivariate statistics is not performed. The maximum number of sub-population or response variables is 50.

PROCEDURE SYNTAX

The PROC PARCAT statement
PROC PARCAT options;
The options below may appear in the PROC PARCAT statement.

TYPECOL= ____

The TYPECOL= option allows the user to specify the scoring method for the response profiles (column scores). A brief description of the scores is given later. The types of scores available to the user are:

UNIFORM = Uniform (default)
CRIDIT = Combined ridit-type
MRANK = Marginal rank
MRIDIT = Marginal ridit
USER = User specified

If neither the TYPECOL option nor CSCORE statement is specified, UNIFORM scores are assumed. If TYPECOL= USER, the CSCORE statement must also be used.

TYPEROW= ____

The TYPEROW= option allows the user to specify the scoring method for the sub-populations (row scores). The types of scores available are displayed in the previous TYPECOL= option. If neither the TYPEROW option nor RSCORE statement is specified, UNIFORM scores are assumed. If TYPEROW= USER, the RSCORE statement must also be used.

MULT, MEAN, CORR

These three options allow the user to specify the types of test statistics. If none of these options are specified, all 3 test statistics listed below will be computed.

MULT = Multivariate Test

For a situation in which both dimensions of the table represent data measured on nominal scales.

MEAN = Mean Score Test

For situations where the response categories are ordinally scaled with progressively larger intensities, but sub-populations are scored on a nominal scale.

CORR = Correlation Test (Mean Score Test is also performed)

For situations where the response categories and the sub-population categories are ordinally scaled with progressively larger intensities.

DATA= ____

Use DATA= to give the name of the SAS data set to be used by PARCAT. If it is omitted, the most recently created data set will be used.

SUMMARY

Use SUMMARY to request output of summary statistics and CMH statistics only. Without the SUMMARY option, the individual tables and their associated statistics will be output in addition to the summary.

NOPRINT

Use to suppress all printed output.

Statements used with PARCAT

1. BY variable names; (optional)

2. CLASS covariate name sub-population name;
(required)
The CLASS statement indicates the names of the covariate and sub-population variables, in that order. The data must be sorted by "BY VARIABLES" (if any), covariate and sub-population.

Example: PROC SORT; BY ITEM BASE TRT;
PROC PARCAT; CLASS BASE TRT; BY ITEM;
Where BASE is the covariable, TRT (treatment) is the sub-population, and ITEM is the BY VARIABLE.

3. VAR variable names; (required)
The VAR statement will define the response variables. These variables contain the frequency counts for each combination of levels of the CLASS variables.

4. RSCORE variable name; (optional)
5. CSCORE variable names; (optional)
RSCORE and CSCORE are statements that define user-specified scores for the sub-population (rows) and response (columns) variables respectively. The order of variables in the CSCORE statement corresponds directly to that of the variables in the VAR statement as illustrated below.

EXAMPLE: Treatment 1 or 2 or 3 is assigned at random to 70 subjects having baseline severity (BASE) of either 1 or 2 and the response categories are none, mild, and severe. The scoring may look like RSC1 - RSC3 for the response variables and SUBPOPSC for the sub-population variable in the following data set.

```
PROC SORT; BY ITEM BASE TRT;
PROC PARCAT; CLASS BASE TRT; BY ITEM;
     VAR    NONE  MILD SEVERE;
     CSCORE RSC1  RSC2 RSC3  ;
     RSCORE SUBPOPSC; * CORRESPONDS TO TRT;
```

6. OUTPUT OUT= data set name; (optional)

The OUTPUT statement asks PARCAT to create a new SAS data set. All CMH statistics, overall sample size, and any BY VARIABLES will be in the new data set. The OUT= option gives the name of the new data set. If it is omitted, SAS names the new data set using the DATAn convention (see Chapter 7, the SAS Users Guide). The output data set will contain these variables:

SSIZE - Overall Sample Size
QCMH - Multivariate statistic with a Chi-Square distribution
DFMN - QCMH degrees of freedom
PMH - Chi-Square p-value for QCMH
QCMMS - Mean Score statistic with a Chi-Square distribution
DFMS - QCMMS degrees of freedom
PMS - Chi-Square p-value for QCMMS
QCMMA - Correlation statistic with a Chi-Square distribution
DFMA - QCMMA degrees of freedom
PMA - Chi-Square p-value for QCMMA

Additionally, any "BY VARIABLES" are included in this SAS dataset.

OUTPUT

The resulting summary output for item 1 from the example given above follows. The available output data set statistics are underlined in the summary with their respective variable names listed directly below them. Individual tables and their associated statistics are also available.

ITEM	BASE	TRT (Treatment)	SUBPOPSC	NONE (freq)	RSC1	MILD (freq)	RSC2	SEVERE (freq)	RSC3
1	1	1	1	2	1	3	3	8	7
1	1	2	2	7	1	5	3	1	7
1	1	3	4	2	1	3	3	3	7
1	2	1	1	3	1	2	3	7	7
1	2	2	2	5	1	5	3	2	7
1	2	3	4	5	1	5	3	2	7

```
PROC FREQ; BY ITEM;
  TABLES BASE * TRT * RESP/SPARSE NOPRINT OUT=FREQDATA;
DATA COMPLETE; SET FREQDATA; BY ITEM BASE TRT;
  RETAIN NONE MILD SEVERE;
  IF FIRST.TRT THEN DO; NONE=.; MILD=.; SEVERE=.; END;
  IF RESP='NONE' THEN NONE=COUNT;  IF RESP='MILD' THEN MILD=COUNT;
  IF RESP='SEVE' THEN SEVERE=COUNT;
  RSC1=1; RSC2=2; RSC3=3;
  IF TRT=1 THEN SUBPOPSC=1;  IF TRT=2 THEN SUBPOPSC=2;
  IF TRT=3 THEN SUBPOPSC=4;
  IF LAST.TRT THEN OUTPUT;
PROC SORT; BY ITEM BASE TRT;
PROC PARCAT TYPECOL=USER TYPEROW=USER; BY ITEM;
  CLASS BASE TRT; VAR NONE MILD SEVERE;
  RSCORE SUBPOPSC; CSCORE RSC1 RSC2 RSC3;
```

TABLE NO. 1

BASE=1

TRT	SCORES	NONE 1.00	MILD 3.00	SEVERE 7.00	TOTAL
1	1.00	2	3	8	13
2	2.00	7	5	1	13
3	4.00	2	3	3	8
	TOTAL	11	11	12	34

I. MULTIVARIATE TEST

Q(1) = 8.80 WITH 4 D.F. P = 0.0664

II. MEAN SCORE TEST

VECTOR OF MEAN SCORES

SUB-POPN: 1 2 3

F(1): 5.1538 2.2308 4.0000

QMS(1) = 8.57 WITH 2 D.F. P = 0.0138

III. CORRELATION TEST

CORRELATION COEFFICIENT : -0.15

QMA(1) = 0.70 WITH 1 D.F. P = 0.4037

TABLE NO. 2

BASE=2

TRT	SCORES	NONE 1.00	MILD 3.00	SEVERE 7.00	TOTAL
1	1.00	3	2	7	12
2	2.00	5	5	2	12
3	4.00	5	5	2	12
	TOTAL	13	12	11	36

I. MULTIVARIATE TEST

Q(2) = 6.48 WITH 4 D.F. P = 0.1663

II. MEAN SCORE TEST

VECTOR OF MEAN SCORES

SUB-POPN: 1 2 3

F(2): 4.8333 2.8333 2.8333

QMS(2) = 5.11 WITH 2 D.F. P = 0.0775

III. CORRELATION TEST

CORRELATION COEFFICIENT : -0.29

QMA(2) = 2.92 WITH 1 D.F. P = 0.0874

***** PARCAT *****

GENERALIZED COCHRAN-MANTEL-HAENSZEL TEST STATISTICS
FOR AVERAGE PARTIAL ASSOCIATION IN THREE-WAY CONTINGENCY TABLES

COLUMN SCORES : USER SPECIFIED
1.00 3.00 7.00

ROW SCORES : USER SPECIFIED
1.00 2.00 4.00

SUMMARY ACROSS TABLES

****

A. SUMMARY OF INDIVIDUAL TABLE STATISTICS

BASE	SAMPLE SIZE	MULTIVARIATE Q	D.F.	P	MEAN SCORE QMS	D.F.	P	CORRELATION QMA	D.F.	P
1	34	8.80	4	0.0664	8.57	2	0.0138	0.70	1	0.4037
2	36	6.48	4	0.1663	5.11	2	0.0775	2.92	1	0.0874
TOTAL	70	15.27	8	0.0541	13.68	4	0.0084	3.62	2	0.1637

B. GENERALIZED COCHRAN-MANTEL-HAENSZEL STATISTICS

	SAMPLE SIZE	MULTIVARIATE Q(CMH)	D.F.	P	MEAN SCORE Q(CMMS)	D.F.	P	CORRELATION Q(CMMA)	D.F.	P
	70	13.67	4	0.0084	12.41	2	0.0020	3.33	1	0.0680

****

VARIABLE: NAMES	SSIZE	QCMH	DFMH	PMH	QCMMS	DFMS	PMS	QCMMA	DFMA	PMA

SPECIFICATION OF SCORES

The choice of a particular set of scores depends on a variety of substantive and statistical issues which will not be elaborated further here. For such considerations, the reader is referred to Yates (1948), Williams (1952), Mantel (1963), Bross (1958), Bhapkar (1968), and Koch et. al. (1977).

Scores available to the procedure include:

User specified (see RSCORE or CSCORE statements) - for situations in which the levels of an ordinal variable may represent well-defined intervals of an underlying quantitative variable.

Uniform (UNIFORM - default) - for situations in which the response variable is ordinally scaled with progressively larger intensities.

Marginal rank (MRANK) - for situations in which ordinally scaled variables are approached from the point of view of various non-parametric rank procedures.

Marginal ridit-type (MRIDIT) - an alternative set of scores similar to MRANK, yet relative to the total sample size in the corresponding table.

Combined ridit-type (CRIDIT) - differentiates from MRIDIT and MRANK in that they utilize the marginal distributions of each table.

AVAILABILITY

This SAS procedure was written in double precision FORTRAN IV for an IBM 360/370/3033 series computer. Some minor alterations may be required in using compilers on other machines. The SAS subroutine library must be available to your system.

Although PROC PARCAT is only a slight modification of the original FORTRAN program, it has not been as extensively tested as its predecessor. The resulting statistics of the procedure should not differ from those from the original program. Inquiries about obtaining PROC PARCAT should be directed to the first author, c/o Mead Johnson and Company, Clinical Information and Statistics, 2404 Pennsylvania Avenue, Evansville, Indiana 47721.

REFERENCES

1. Bhapkar, V. P. (1968), On the analysis of contingency tables with a quantitative response. Biometrics 24: 329-338.

2. Bross, I. D. J. (1958), How to use 'ridit' analysis. Biometrics 14: 18-38.

3. Cochran, W. G. (1954), Some methods for strengthening the common χ^2 test. Biometrics 10: 417-451.

4. Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., Jr., and Lehnen, R. G. (1977), A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics 33: 133-158.

5. Landis, J. R., Heyman, E. R., and Koch, G. G. (1978), Average partial association in three way contingency tables: a review and discussion of alternative tests. Int. Stat. Review: 237-254.

6. Landis, J. R., Cooper, M. M., Kennedy, T., and Koch, G. G. (1979), A computer program for testing average partial association in three-way contingency tables. (PARCAT), Computer Programs in Biomedicine 9: 223-226.

7. Mantel, N., and Haenszel, W. (1959), Statistical aspects of the analysis of data from retrospective studies of disease. J. Nat. Cancer Inst. 22: 719-748.

8. Mantel, N. (1963), Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. J. Am. Stat. Assoc. 58: 690-700.

9. Williams, E. J. (1952), Uses of scores for the analysis of association in contingency tables. Biometrika 39: 274-289.

10. Yates, F. (1948), The analysis of contingency tables with groupings based on quantitative characters. Biometrika 35: 176-181.