

An Algorithm for Creating Multiple Response Variables

Paul Pope, Texas Agricultural Extension Service, Texas A&M University System

ABSTRACT

This paper presents a parsimonious algorithm for creating multiple response variables from multiple dichotomous variables. The algorithm is based on concatenation and character insertion via the SUBSTR (substring) function.

INTRODUCTION

Multiple response questions, common in evaluation and survey work, permit more than one response. The following questions from a clip art usage survey serve as examples:

Q5. WITH WHAT TYPE OF DOCUMENTS ARE YOU USING CLIP ART (check all that apply):

- ☐ A. Newsletters
- ☐ B. Calendars
- ☐ C. Announcements
- ☐ D. Evaluation forms (surveys, pre/post tests, etc.)
- ☐ E. Letters (letterhead, logos)
- ☐ F. Awards
- ☐ G. Programs
- ☐ H. Handbooks, fact sheets
- ☐ I. Other _____

Q9. WHAT PICTURES DO YOU USE NOW FROM THE FOLLOWING SUBJECT GROUPS (check all that apply):

- ☐ A. Farm/ranch animals
- ☐ B. Wildlife
- ☐ C. Crops/vegetables/fruits
- ☐ D. Family/people scenes
- ☐ E. Holiday symbols/scenes
- ☐ F. Extension logos
- ☐ G. 4-H
- ☐ H. Other _____

For questions such as these, ease of data entry dictates that a separate, dichotomous variable be created for each question item (with each variable coded as 1 if the corresponding item is selected or 0 if the item is not selected). For example, Q5 would require nine dichotomous variables (Q5A, Q5B, Q5C, Q5D, Q5E, Q5F, Q5G, Q5H, and Q5I) with the variables coded as 0 1 1 1 0 0 0 1 0 for a respondent who checked B C D and H. Frequency distributions on the dichotomous variables would reveal the frequency that each item was selected. *However, there remains the challenge of creating single, multiple response variable (Q5) from the multiple dichotomous variables (Q5A-Q5I).*

OTHER METHODS

Because SAS currently does not contain a dedicated procedure for handling multiple response data, a number of methods have been developed to create multiple response variables from multiple dichotomous variables.

The SAS Institute, Inc. (1987), in its Applications Guide, offered a method based on a formula that performed a binary to decimal transformation. Keene (1989) presented PROC TABULATE as a solution. More recently, Pope and Fannin (1992) offered a facilitated implementation of SAS' method and Pope (1993) presented two other methods, one based on concatenation, the other based on validated data entry.

All of these methods have merit. However, in my view, the algorithm presented in this paper is a superior method in terms of efficiency and ease of use.

THE ALGORITHM

The program below demonstrates the algorithm for Q5 and Q9. The algorithm concatenates 0s and 1s into a single character variable and substitutes the appropriate item label for 1s in the character variable. 0s are replaced with spaces. Then the variable is compressed.

The shaded portion of code requires modification for your particular application. Simply list the new (multiple response) variables in the LENGTH and ARRAY COMBINE statements. Then create an ARRAY for each new variable.

For CPU and disk space efficiency, reduce the length of all variables to equal the number of items for the multiple response question with the largest number of items and eliminate unneeded IF-THEN-ELSE statements within the COMBINE array. For example, the lengths of Q5 and Q9 can be reduced to 9 and all but the first nine IF-THEN-ELSE statements within the COMBINE array can be eliminated since there are no items beyond I (for Q5).

```
DATA MRD;
INPUT (Q5A Q5B Q5C Q5D Q5E Q5F Q5G Q5H Q5I Q9A Q9B Q9C Q9D Q9E Q9F Q9G Q9H) (2.);
LENGTH Q5 Q9 $16.;
ARRAY C1 Q5A-Q5I; DO OVER C1; Q5=COMPRESS(Q5||C1); END;
ARRAY C2 Q9A-Q9H; DO OVER C2; Q9=COMPRESS(Q9||C2); END;
ARRAY COMBINE Q5 Q9;
DO OVER COMBINE;
  IF SUBSTR(COMBINE,1,1)='1' THEN SUBSTR(COMBINE,1,1)='A'; ELSE SUBSTR(COMBINE,1,1)=' ';
  IF SUBSTR(COMBINE,2,1)='1' THEN SUBSTR(COMBINE,2,1)='B'; ELSE SUBSTR(COMBINE,2,1)=' ';
  IF SUBSTR(COMBINE,3,1)='1' THEN SUBSTR(COMBINE,3,1)='C'; ELSE SUBSTR(COMBINE,3,1)=' ';
  IF SUBSTR(COMBINE,4,1)='1' THEN SUBSTR(COMBINE,4,1)='D'; ELSE SUBSTR(COMBINE,4,1)=' ';
  IF SUBSTR(COMBINE,5,1)='1' THEN SUBSTR(COMBINE,5,1)='E'; ELSE SUBSTR(COMBINE,5,1)=' ';
  IF SUBSTR(COMBINE,6,1)='1' THEN SUBSTR(COMBINE,6,1)='F'; ELSE SUBSTR(COMBINE,6,1)=' ';
  IF SUBSTR(COMBINE,7,1)='1' THEN SUBSTR(COMBINE,7,1)='G'; ELSE SUBSTR(COMBINE,7,1)=' ';
  IF SUBSTR(COMBINE,8,1)='1' THEN SUBSTR(COMBINE,8,1)='H'; ELSE SUBSTR(COMBINE,8,1)=' ';
  IF SUBSTR(COMBINE,9,1)='1' THEN SUBSTR(COMBINE,9,1)='I'; ELSE SUBSTR(COMBINE,9,1)=' ';
  IF SUBSTR(COMBINE,10,1)='1' THEN SUBSTR(COMBINE,10,1)='J'; ELSE SUBSTR(COMBINE,10,1)=' ';
  IF SUBSTR(COMBINE,11,1)='1' THEN SUBSTR(COMBINE,11,1)='K'; ELSE SUBSTR(COMBINE,11,1)=' ';
  IF SUBSTR(COMBINE,12,1)='1' THEN SUBSTR(COMBINE,12,1)='L'; ELSE SUBSTR(COMBINE,12,1)=' ';
  IF SUBSTR(COMBINE,13,1)='1' THEN SUBSTR(COMBINE,13,1)='M'; ELSE SUBSTR(COMBINE,13,1)=' ';
  IF SUBSTR(COMBINE,14,1)='1' THEN SUBSTR(COMBINE,14,1)='N'; ELSE SUBSTR(COMBINE,14,1)=' ';
  IF SUBSTR(COMBINE,15,1)='1' THEN SUBSTR(COMBINE,15,1)='O'; ELSE SUBSTR(COMBINE,15,1)=' ';
  IF SUBSTR(COMBINE,16,1)='1' THEN SUBSTR(COMBINE,16,1)='P'; ELSE SUBSTR(COMBINE,16,1)=' ';
  COMBINE=COMPRESS(COMBINE);
END;
CARDS;
1 1 1 0 0 0 1 1 0 1 1 0 1 0 0 1 0
1 0 1 0 1 0 1 1 0 1 0 0 1 0 0 0 0
  (continue data)
;
PROC FREQ; TABLES Q5 Q9; RUN;
```

The frequency distributions for the multiple response variables (Q5 and Q9) are shown below.

Q5	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ABC	1	1.6	1	1.6
ABCEGH	1	1.6	2	3.1
ABCFGH	1	1.6	3	4.7
ABCGH	24	37.5	27	42.2
ABCGHI	2	3.1	29	45.3
ACE	1	1.6	30	46.9
ACEFGH	1	1.6	31	48.4
ACEGH	19	29.7	50	78.1
ACEGHI	2	3.1	52	81.2
ACEH	1	1.6	53	82.8
BCEGH	1	1.6	54	84.4
BCGH	2	3.1	56	87.5
CEGH	5	7.8	61	95.3
CGHI	1	1.6	62	96.9
EGHI	1	1.6	63	98.4
GHI	1	1.6	64	100.0

Q9	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	1	1.6	1	1.6
ABDFG	1	1.6	2	3.1
ABDG	24	37.5	26	40.6
ABDGH	1	1.6	27	42.2
ACD	1	1.6	28	43.7
AD	22	34.4	50	78.1
ADEF	1	1.6	51	79.7
ADF	2	3.1	53	82.8
ADH	2	3.1	55	85.9
BCD	1	1.6	56	87.5
BDEG	1	1.6	57	89.1
BDFG	2	3.1	59	92.2
BDG	2	3.1	61	95.3
D	1	1.6	62	96.9
DEF	1	1.6	63	98.4
DG	1	1.6	64	100.0

Finally, its important to note that for the algorithm to work correctly as presented, it is essential that all of the dichotomous variables have a code of 0 or 1. That is, there should be no missing values (.).

REFERENCES

Keene, T. (1989), "Analyzing Multiple Response Data Using the TABULATE PROCEDURE." Proceeding of the Fourteenth SAS Users Group International Conference, 14, Cary, NC: SAS Institute Inc.

Pope, P. and Fannin, D. (1992), "Analyzing Multiple Response Data: Facilitating An Existing Method and Presenting An Alternative." Proceedings of the Seventeenth SAS Users Group International Conference, 17, Cary, NC: SAS Institute Inc.

Pope, P. (1993), "Analyzing Multiple Response Survey Data: An Overview of Strategies and Methods." Proceedings of the Eighteenth SAS Users Group International Conference, 18, Cary, NC: SAS Institute Inc.

SAS Institute, Inc. (1987), *SAS Applications Guide, 1987 Edition*. Cary, NC: SAS Institute, Inc.

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries.
® indicates USA registration.

The author encourages you to share your experiences, evaluations, or comments concerning the concepts and applications presented in this paper:

Paul Pope
Texas Agricultural Extension Service
112A Special Services Bldg.
College Station, TX 77843-2468
Phone: (409) 845-3929
Internet: p-pope@tamu.edu