%FREQ1VAR: Frequency of one variable with format: a macro to standardize proc FREQ output data sets

Ronald Fehd, Centers for Disease Control and Prevention, Atlanta GA

ABSTRACT

SAS(r) software provides the ability to associate formats, or look up tables, with variables in a data set. Encapsulating the format labels in a summary data set provides a complete and independent set of information about a variable to other procedures.

This macro was written to standardize summary data sets and bring them into conformance with data sets produced by Fehd's (1997) CHECKALL and SHOWCOMB macros, which provide information about multiple response data. The standardized data sets, or objects, have the same structure and are used for quick and concise reporting of summary information of large survey questionnaire data sets. Their identical structure enables easy access by other methods and software.

INTRODUCTION

Proc FREQ provides a summary of a variable: when a format is associated with the variable the procedure prints the value using its associated format. While a paper report is often sufficient, the increasing demand for graphics reports and presentations creates an expectation for an object that contains a complete set of information about a data set and the variable(s) being reported.

This macro processes a proc FREQ output data set and adds format labels and attributes to the data set. An attribute is a single item of information and includes the name of the data set against which the proc was run, the number of identifiers and number of observations of the data set, the number of valid responses in the variable and the response rate: the ratio of valid response to number of observations. These attributes are placed in variable labels and the first observation of the summary data set.

Setup: sets of values, autoexec, proc FORMAT

Observations in a data set contain one of three sets of values. When a variable is defined, its value is initialized to blank for character variables, or missing for numeric variables. As data entry or data manipulation occurs, variables receive valid values. In a small minority of cases, variables obtain invalid values.

The standard labels for blank/missing and invalid are set as global macro variables in the autoexec for the session. This allows the labels to be assigned with proc FORMAT in a separate program from the summarization program. The autoexec contains these statements:

%LET BLANK = 'BLANK';
%LET INVALID = 'INVALID';

During a data review process, it is necessary to report all three sets of data values, but for summary purposes, blank/missing and invalid values are expected to be excluded. This is accomplished in this macro by using standardized labels in proc FORMAT value statements. Valid values have their description. Character variables have space and dot labeled as "&BLANK.". Numeric variables have missing and in some cases, zero, labeled as "&BLANK.". Values to be excluded are grouped with the phrase: other="&INVALID.".

A program which contains formats would have these statements:

proc FORMAT;
  value Num 1 = 'one'  2 = 'two'
    . = 'BLANK.' other = 'INVALID.';
  value $Chr '1' = 'one' '2' = 'two'
    ' ' = 'BLANK.' other = 'INVALID.';

Refer to the test data at the end of the macro for further examples.

The process of standardization

This macro consists of several steps: initialization, using proc FREQ to produce an output data set, making attribute data sets, reading the data to prepare the other attributes and finally writing the summary data set to the library.

In the initialization step, all local macro variables are declared, the output data-set name is initialized, if not provided as a parameter, and options for macro debugging are turned on or off.

proc FREQ is the first major step in the macro. Data are excluded with a where statement which compares the formatted value of the variable to the macro variables BLANK and INVALID. Parameters are provided for an additional where clause with which to subset the data and to sort the output data set by descending Count.

The %NOS macro provides the number of observations of the data set. The %NOS macro is based on the %OBSNVAR macro (SAS Macro Language Reference, First Edition). If the output data set contains no observations, then the macro stops processing and returns a completion code of zero.

The second step consists of preparing two data sets which contain the attributes Number-of-Ids and Number-of-Observations. These data sets are merged into the final data set.

In the third step, the format labels are added to the data set. The width of the three major variables, Label, Count and Percent, are calculated. The number of responses is accumulated from Count.

The label of the analysis variable is copied to a macro variable in the fourth step; other macro variables are created containing the various widths and the number of responses. When the data step is complete then the percentage of response is calculated.

During the final data step, the attributes are placed in the respective variable labels as the summary data set is written to the library.

Summary data set and attributes: definitions

The FREQ1VAR macro provides several attributes in addition to the three variables of the proc FREQ data set.

N-IDS: Number of Identifiers: When reporting summary information, either the number of observations or the number of identifiers is typically presented. Any data set either is unique on its identifiers or has multiple occurrences of its identifiers, therefore the value of N-IDS is less than or equal to the number of observations. This item is stored in the first observation of a character variable.

N_OBS: Number of Observations with valid values: data may be excluded from the summary; therefore the number of observations...
with valid values is saved. This item is stored in two places: as the first observation of this character variable, and in the label of that variable. Three other items are stored in the label as well: the name and total number of observations of the data set, and the response rate (Response/Total). Here is an example of the label: N=5 data:TESTDATA Obs:15 Resp:33%

LABEL: the variable’s value, from its format; a character variable whose length is determined by the widest label. The label of this variable is the label of the analysis variable.

COUNT: number of observations with this value, a numeric integer.

PERCENT: of observations with this value, a numeric real; note that the denominator is the sum of COUNT.

VALUECHR, or VALUENUM: the analysis variable, renamed according to its type. This variable is provided in the data set as a subset: additional where clause used to subset the data set check that the correct format was applied.

TITLE: to be used, for graphics or other presentation, if not the other countries.

ACKNOWLEDGMENTS

This routine was developed over a period of ten years while I crunched the numbers of survey data collected by the Model Performance Evaluation Program (MPEP) of the Division of Laboratory Systems, Public Health Practice Program Office of the Centers for Disease Control and Prevention, Atlanta, Georgia. John Hancock, chief, Information Services Activity, encouraged me to write up these routines. I wish to thank Sharon Blumer, David Cross, Thomas Heam, and William Schalla of the MPEP group for their perseverance while I developed this routine.

CONCLUSION

Proc FREQ provides an output data set that contains three variables: the analysis variable, and Count and Percent. This macro renames the analysis variable, adds a character variable with the format label, calculates and adds attributes to the data set to create a package of information in a standardized object that can be used by other methods and applications for graphics and presentations.

REFERENCES


SAS is a registered trademark of SAS Institute, Inc. In the USA and other countries. © indicates SAS USA registration.

Author: Ronald Fehd e-mail: RjF2@cdc.gov Centers for Disease Control MS-G25 4770 Buford Hwy NE Atlanta GA 30341-3724 voice: 770/488-8102 to: SAScontribute@SAServ.uga.edu for %FREQVAR subject: cnb0038: download for %CHECKALL subject: cnb0032: download for %SHOWCOMB subject: cnb0033: download

ACKNOWLEDGMENTS

Posters

Proc FREQ provides an output data set that contains three variables: the analysis variable, and Count and Percent. This macro renames the analysis variable, adds a character variable with the format label, calculates and adds attributes to the data set to create a package of information in a standardized object that can be used by other methods and applications for graphics and presentations.

REFERENCES


SAS is a registered trademark of SAS Institute, Inc. In the USA and other countries. © indicates SAS USA registration.

Author: Ronald Fehd e-mail: RjF2@cdc.gov Centers for Disease Control MS-G25 4770 Buford Hwy NE Atlanta GA 30341-3724 voice: 770/488-8102 to: SAScontribute@SAServ.uga.edu for %FREQVAR subject: cnb0038: download for %CHECKALL subject: cnb0032: download for %SHOWCOMB subject: cnb0033: download

ACKNOWLEDGMENTS

Posters

Proc FREQ provides an output data set that contains three variables: the analysis variable, and Count and Percent. This macro renames the analysis variable, adds a character variable with the format label, calculates and adds attributes to the data set to create a package of information in a standardized object that can be used by other methods and applications for graphics and presentations.

REFERENCES


SAS is a registered trademark of SAS Institute, Inc. In the USA and other countries. © indicates SAS USA registration.

Author: Ronald Fehd e-mail: RjF2@cdc.gov Centers for Disease Control MS-G25 4770 Buford Hwy NE Atlanta GA 30341-3724 voice: 770/488-8102 to: SAScontribute@SAServ.uga.edu for %FREQVAR subject: cnb0038: download for %CHECKALL subject: cnb0032: download for %SHOWCOMB subject: cnb0033: download

ACKNOWLEDGMENTS

Posters

Proc FREQ provides an output data set that contains three variables: the analysis variable, and Count and Percent. This macro renames the analysis variable, adds a character variable with the format label, calculates and adds attributes to the data set to create a package of information in a standardized object that can be used by other methods and applications for graphics and presentations.

REFERENCES


SAS is a registered trademark of SAS Institute, Inc. In the USA and other countries. © indicates SAS USA registration.

Author: Ronald Fehd e-mail: RjF2@cdc.gov Centers for Disease Control MS-G25 4770 Buford Hwy NE Atlanta GA 30341-3724 voice: 770/488-8102 to: SAScontribute@SAServ.uga.edu for %FREQVAR subject: cnb0038: download for %CHECKALL subject: cnb0032: download for %SHOWCOMB subject: cnb0033: download

ACKNOWLEDGMENTS

Posters

Proc FREQ provides an output data set that contains three variables: the analysis variable, and Count and Percent. This macro renames the analysis variable, adds a character variable with the format label, calculates and adds attributes to the data set to create a package of information in a standardized object that can be used by other methods and applications for graphics and presentations.

REFERENCES


SAS is a registered trademark of SAS Institute, Inc. In the USA and other countries. © indicates SAS USA registration.

Author: Ronald Fehd e-mail: RjF2@cdc.gov Centers for Disease Control MS-G25 4770 Buford Hwy NE Atlanta GA 30341-3724 voice: 770/488-8102 to: SAScontribute@SAServ.uga.edu for %FREQVAR subject: cnb0038: download for %CHECKALL subject: cnb0032: download for %SHOWCOMB subject: cnb0033: download

ACKNOWLEDGMENTS