Printing a Random Sample of a Large Data Set

Hany Aboutaleb, Pfizer Inc., NYC NY.

ABSTRACT

If you have a large data set and you wish to print 5%, 10%, or more of your data, this article can help you. It describes and illustrates how to print a percent of your data with proc print. The macro will pick the percent of your data that you specify and then help you to print it. The percent that is printed is a random sample. The database you are working with can be an entire library or just one file.

OVERVIEW

Proc print does produce satisfactory result with large dataset. The output is difficult to read and use unnecessary amount of paper. However if you use the OBS statement with proc print, you can specify the amount of observation to print. The remaining problem is to select a subset of the data using a key variable. This paper will describe how you can write a program to classify your data with a key variable, choose percent of your data to print, determine the size of the subgroup, select it randomly and print it. For example if the dataset have 100 entry and you choose to print 10% of it the key variable will determine that there are 10 elements to print. The RANUNI function will generate 10 uniform random number from 1-100. The records with those positions in the data will be identified and printed.

METHOD

1. Find a dataset or a library you want to print.
2. Identify your Key variable.
3. Choose how much you want to print (percent).
4. Print the simulated random sample or the random subset of your key variable.
5. Choose the header or the title of your output.

KEY STEPS

1. Reference the location of your library or a dataset.
2. The program will identify the member of your library and create macro variables ds1……dsn where n is the total number of members.
3. The program will identify the first element of your data to get the subgroup population.
4. The program will calculate from your percent the seed that is use with the RANUNI function. This seed number should be a positive integer so that you will be able to replicate the same random assignment.
5. The program will output the random sample of your subgroup and create a macro variable to be used in printing the subset data.

THE PROGRAM

PROGRAM NAME: PCTPRINT.SAS
PROJECT NAME: PFIZER Inc.
PROGRAMMER: Hany Aboutaleb
CREATION DATE: (11/21/97)
LANGUAGE: PC SAS
PURPOSE: THIS PROGRAM SELECTS PATIENTS FOR RANDOM PRINT, REQUIRES A DATASET OR LIBRARY TO BE PRINT.
MACRO FILE DIRECTORY:
INPUT DATA DIRECTORY:
OUTPUT FILE DIRECTORY:
OUTPUT AND PRINT SPECIFICATIONS: PAGESIZE=(60) LINESIZE=(170)
INITIALLY REVIEWED BY:
INITIAL REVIEW DATE:
REMARKS: THIS MACRO HELP CUT PAPER COSTS AND SAVE MORE TREES. IN ADDITION IT CAN BE USED FOR AUDITING PURPOSE.

Here you can change the name of the sub-directory to the new study.
LIBNAME FMTLIB 'C:\SASFMTS';
LIBNAME IN1 'C:\PROJECTS\SASDATA1';
LIBNAME IN7 'C:\PROJECTS\SASDATA';

Set the options for SAS

OPTIONS PAGENO=1 FORMS=PK150
LINESIZE=164 PAGESIZE=57 MACROGEN
MPRINT MAUTOSOURCE NOFMTERR
FMTSEARCH=(FMTLIB.HAN);

The Macro start with the macro call.

%MACRO PCTPRINT (LIBIN=, DATAIN=, VAR=, PERCENT=, OUT=, HEADER=);

The logic statement will determine whether you want to print a library or a member of the dataset. If you want to print a library, it will use PROC CONTENTS to get the members’ names from the datasets and create macro variables for the members.

%IF %LENGTH(&LIBIN)^=0 %THEN %DO;
%LET LIB=%UPCASE(&LIBIN);
PROC CONTENTS DATA=&LIB.._ALL_
OUT=CONTS(KEEP=MEMNAME)
NOPRINT;
PROC SORT DATA=CONTS NODUPKEY;
BY MEMNAME;
RUN;
DATA _NULL_; 
SET CONTS;
BY MEMNAME;
IF FIRST.MEMNAME THEN 
CALL SYMPUT
('DATAIN',TRIM(UPCASE(MEMNAME)));
RUN;
DATA _NULL_; 
SET CONTS END=LAST;
BY MEMNAME;
IF LAST.MEMNAME THEN 
CALL SYMPUT
('DATAIN',TRIM(UPCASE(MEMNAME)));
RUN;
%END;

The macro uses the logic statement to create macro variables for each member of the ID variable to be used for the subsiding of the data.

PROC SORT DATA=&DATAIN
OUT=&DATAIN;
BY &VAR ;
RUN;

DATA DATAIN;
SET DATAIN;
BY &VAR;
IF FIRST.&VAR;
RUN;

DATA _NULL_; SET DATAIN END=LAST;
IF LAST THEN CALL
SYMPUT('CNT',LEFT(_N_));
RUN;

The macro will use the percent to calculate the seed which will be used in the RANUNI function to generate a random number.

DATA _NULL_; 
SEED=INT((&PERCENT*&CNT)/100);
CALL SYMPUT('SEED',SEED);
RUN;

DATA TEMP(DROP=SAMPLE) 
 TEMP1(KEEP=SAMPLE);
SET DATAIN(KEEP=&VAR) END=MAX;
RND=RANUNI(&SEED);

A positive integer is used to retain the ability to replicate the same random assignment and round up to the nearest integer.

By using the sort, and the set statements the macro will output the random sample of your subgroup with the help of the ID variable.

IF MAX=1 THEN DO;
SAMPLE=&SEED;
OUTPUT TEMP1;
END;
OUTPUT TEMP;
RUN;
PROC SORT DATA=TEMP;
BY RND;
RUN;
DATA &OUT(DROP=RND SAMPLE);
SET TEMP;
IF _N_=1 THEN SET TEMP1;
IF _N_<=SAMPLE;
RUN;

Create macro variables Vs1 ... VsN, which hold the variable names.

PROC SORT DATA=&OUT;
  BY &VAR;
RUN;
DATA HANY;SET &OUT;
  BY &VAR;
  IF LAST.&VAR;
    GROUP=&VAR;
  KEEP GROUP &VAR;
PROC SORT;BY &VAR;
RUN;
DATA HANY;
  SET HANY;
  BY &VAR;
  RETAIN GROUPNM 0;
  IF FIRST.&VAR THEN GROUPNM+1;
DATA _NULL_; SET HANY END=LAST;
  CALL SYMPUT('VS'||LEFT(_N_),TRIM(UPCASE(GROUP)));
  IF LAST THEN CALL SYMPUT('TOT',LEFT(_N_));
RUN;

Create the macro variables to hold the subgroup values.

%LET SUBGROUP=;
%LET COMA=%NRSTR(,);
%DO I=1 %TO &TOTAL;
  %IF %LENGTH(&SUBGROUP)^=0 %THEN %DO;
    %LET SUBGROUP=&SUBGROUP&COMA%STR("&&VS&I");
  %END;
%END;

Now the macro will print your library or whatever size random sample of the dataset you specify with your percent.

%IF %LENGTH(&LIBIN)=0 %THEN %DO;
  PROC PRINT DATA=&DATAIN;
  WHERE &VAR IN &SUBGROUP;
  TITLE1 &HEADER;
  TITLE2 'LIST OF RANDOMLY SELECTED PATIENTS ("&PERCENT"%)';
  TITLE3 "&DATAIN";
  RUN;
%END;

%ELSE %DO;
%DO I=1 %TO &TOTAL;
  %IF %LENGTH(&HEADER)> 0 %THEN %DO;
    TITLE1 "&HEADER";
    TITLE2 "&&DS&I";
  %ELSE %DO;
    TITLE "&LIB..&&DS&I";
  %END;
  PROC CONTENTS DATA=&LIB..&&DS&I;
  PROC PRINT DATA=&LIB..&&DS&I;
  WHERE &VAR IN &SUBGROUP;
  RUN;
%END;
%MEND PCTPRINT;

This a sample call for the macro.
CONCLUSION:

A great deal of flexibility can be gained by using this macro to enhance SAS proc print. It will also help cut paper costs. In addition it can be used for Auditing purpose by printing a random sample of the data.

TRADEMARKS:

SAS® is a registered trademark of the SAS Institute, Inc., in the USA and in other countries.
Windows® is a registered trademark of the Microsoft Corporation.

ACKNOWLEDGMENT:

The author would like to thank Prof. Judy Coomes of William Paterson University for her comments in his preparation of this paper.

REFERENCES:


CONTACT:

Hany Aboutaleb
Pfizer Inc.
235 East 42nd street
New York, NY 10017-5755
Mail Stop:150-37-30
Office: (212) 733-4718
Fax: (212) 309-4346
Internet:abouth@pfizer.com