

What's *really* in your data?

A quick and easy utility to analyze any SAS dataset

SESUG Author, Michele Laughlin, Catalyst Technologies, Snellville, GA
SESUG Co-Author, Thom Clay, Catalyst Technologies, Snellville, GA
SESUG Co-Author, Jamila Ervin, Catalyst Technologies, Snellville, GA

ABSTRACT

Many companies produce Quality Assurance reports daily in order to provide the highest quality of service and keep information up to date. The reports may consist of general demographic information, statistical data, cost analysis, date validation, etc. In order to produce all of these reports, programmers may generate code to create each of these analyses. The effort in creating separate utilities can be time consuming, expensive and an inefficient use of company resources. When Catalyst Technologies' management realized the number of hours expended on creating code for these Quality Assurance reports, they came up with an answer. This paper discusses that solution, an easy to use utility that combines a variety of data to create numerous reports usable by programmers and non-programmers alike.

TECHNICAL OVERVIEW

The utility uses a PROC CONTENTS statement to read in all of the variables on the input dataset. It will then evaluate each individual variable in the dataset to determine which report to produce based on the variable type. If a numeric variable is discovered, then a PROC UNIVARIATE will be performed. If a character variable is found, then a PROC FREQUENCY will be performed. The utility also looks for a date variable and, if found, will perform a PROC FREQUENCY using the year of the date. There are additional reports that can be produced from this utility. Examples include age, date of birth frequency and reporting of values that fall outside of a specified range.

```
%macro FreqUniv(indsn);
  %inc "G:\Macros\DateIt.sas";
  %inc "G:\Macros\DateRpt.sas";
  %inc "G:\Macros\FreqIt.sas";
  %inc "G:\Macros\HighVol.sas";
  %inc "G:\Macros\UnivIt.sas";

  Proc contents data=&indsn
                out=contents (keep=name type length format formatl formatd
                             label nob) noprint;
run;

filename sasparm "E:\Customer_Regions\Sample_Customer\ sastest.txt";

data contents;
  set contents(rename=(type=vtype));
  file sasparm;
  if type = 1 and format in ('DATE' 'DDMMYY' 'JULIAN' 'MMDDYY' 'MMYY'
                           'MONYY' 'WORDDATE' 'YYMMDD' 'YYMON')
     or
     informat in ('DATE' 'DDMMYY' 'JULIAN' 'MMDDYY' 'MMYY'
                  'MONYY' 'WORDDATE' 'YYMMDD' 'YYMON')
  then do;
    dtflag = 'Y';
```

```

        put @1 '%dateit(' name ');';
    end;
    else if type = 1 then put @1 '%univit(' name ');';
    else if type = 2 and (length > 3 or in_hv) then
        put @1 '%highvol(' name ');';
    else
        put @1 '%freqit(' name ');';
    run;
    %inc sasparm;
%mend FreqUniv;

%FreqUniv(custds.sample);

```

After the PROC CONTENTS is performed, the utility decides what to do with the data it has been given. If a numeric variable is found, the PROC UNIVARIATE will be run by calling the macro 'univit'.

```

        *** Macro to run univariates based on &var ***;
%macro univit(var);
    title2 "Univariate Statistics of &var from &indsn";
    proc univariate data=&indsn;
        var &var;
        output out=temp_unv
            NOBS    = NOBS
            N       = N
            MEAN    = MEAN
            SUM     = SUM
            STD     = STD
            MAX     = MAX
            MIN     = MIN
            MEDIAN  = MEDIAN
            VAR     = VAR
            MODE    = MODE
            P10     = P10
            P90     = P90
            nmiss   = nmiss;
    run;

    %if &freqout ~= %then %do;
        data temp_unv;
            length custid $4 datatype $7 varname $20 value $50 count percent cumpct
            rundeate 8;
            set temp_unv;
            varname = "&var";
            rundeate = datetime();
            value    = 'Number of Missing values';
            count    = nmiss;
            percent  = round(nmiss/nobs,.0001);
            output;
            value    = 'Number of Nonmissing values';
            count    = n;
            percent  = round(n/nobs,.0001);
            output;
            value    = 'The Sum';
            count    = sum;
            percent  = .;
            output;
            value    = 'The Largest Value';
            count    = max;
        %end;
    %end;
%macro univit;

```

```

output;
value    = 'The Smallest Value';
count    = min;
output;
value    = 'The Mean';
count    = mean;
output;
value    = 'The Median';
count    = median;
output;
value    = 'The Most Frequent value';
count    = mode;
output;
value    = 'The Variance';
count    = var;
output;
value    = 'The Standard Diviation';
count    = std;
output;
value    = 'The 90th percentile';
count    = p90;
output;
value    = 'The 10th percentile';
count    = p10;
output;
drop nobs n mean sum std max min median mode var p10 p90 var nmiss;
retain cumpct .;
format rundate datetime. count comma12. percent cumpct percent9.2;
label custid   = 'Customer ID'
      datatype = 'Data Type'
      varname   = 'Variable Name'
      value     = 'Value'
      count     = 'Count'
      percent   = 'Percent'
      cumpct    = 'Cumulative Percent'
      rundate   = 'Run Date'
      ;
run;

proc append base=freqout.&freqpfx._freqtbl data=temp_unv;
run;

%end;
%mend univit;

```

This produces an output report containing useful statistics such as standard deviations, means, moments, sum weights, base statistical measures, quantiles, maximum, mean and minimums.

OUTPUT REPORT

Sample Customer
Univariate Statistics of DAYSSPLY from custds.sample

The UNIVARIATE Procedure
Variable: DAYSSPLY (Days Supply)

Moments

N	32000	Sum Weights	32000
Mean	27.7229063	Sum Observations	887133
Std Deviation	23.1790809	Variance	537.26979
Skewness	2.43620053	Kurtosis	22.0769411
Uncorrected SS	41786001	Corrected SS	17192096
Coeff Variation	83.6098519	Std Error Mean	0.129575

Basic Statistical Measures

Location		Variability	
Mean	27.72291	Std Deviation	23.17908
Median	30.00000	Variance	537.2697
Mode	30.00000	Range	610.00000
	Interquartile Range		20.00000

Quantiles (Definition 5)

Quantile	Estimate
100% Max	520
99%	90
95%	90
90%	34
75% Q3	30
50% Median	30
25% Q1	10
10%	4
5%	0
1%	0
0% Min	-90

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
-90	19358	336	7406
-90	15038	365	11454
-90	9103	365	11455
-90	9102	520	13029
-90	2451	520	13031

If a character variable is found, the utility will determine the length of the field. If the length is less than three then a PROC FREQUENCY will be performed on all of the observations by calling the macro 'freqit'.

```
%macro freqit(var);
  title2 "Frequency distribution of &var from &indsn";
  proc freq data=&indsn;
    table &var / missing out=temp_frq;
  run;

  %if &freqout ~= %then %do;
    data temp_frq;
      length custid $4 datatype $7 varname $20 value $50 count percent cumpct
      runde 8;
      set temp_frq;
      varname = "&var";
      value   = &var;
      percent = round(percent/100,.0001);
    end;
  %end;
```

```

        cumcnt + count;
        cumpct = round(cumcnt/&subset1,.0001);

        rundate = datetime();
        drop &var cumcnt;
        retain cumcnt 0;
        format rundate datetime. count comma12. percent cumpct percent9.2;
        label varname = 'Variable Name'
              value    = 'Value'
              count    = 'Count'
              percent  = 'Percent'
              cumpct   = 'Cumulative Percent'
              rundate  = 'Run Date'
              ;
    run;

    proc append base=freqout.&freqpfx._freqtbl data=temp_frq;
    run;
%end;
%mend freqit;

```

If the length is greater than three, then the macro will limit the number of observations output by performing a PROC FREQUENCY on the first 100 observations in the dataset. If the utility finds that there are more than three observations, the macro 'highvol' is invoked.

```

*** Macro to run high volumn frequency for &var ***;
%macro highvol(var);
    proc freq data=&indsn noprint order=freq;
        table &var / missing out=temp_hv;
    data temp_hv;
        set temp_hv(obs=100);
    proc sort data=temp_hv;
        by &var;
    data temp_hv;
        set temp_hv;
        percent = round(percent/100,.0001);
        cumcnt + count;
        cumpct = cumcnt;
    run;
    title2 "Top 100 from frequency distribution of &var from &indsn";
    proc print data=temp_hv split='*' noobs;
        var &var count cumcnt percent cumpct;
        format percent cumpct percent9.2;
        label count    = 'Frequency'
              cumcnt    = 'Cumulative*Frequency'
              percent    = 'Percent*of Total'
              cumpct     = 'Cumulative*Percent';
    run;
%mend highvol;

```

The output report below shows a frequency distribution of the sample variable. The number of observations output can be changed to whatever you wish. The default used in this example is 2.

OUTPUT REPORT

Sample Customer
Top 10 from frequency distribution of SAMPVAR

SAMPVAR	Frequency	Cumulative Frequency	Percent of Total	Cumulative Percent
	31940	31940	99.81%	99.81%
ANTIBIO06	60	32000	0.19%	100.00%

If a date variable is found, a PROC FREQUENCY is performed on the variable using a year format. A macro is invoked called 'dateit'. The current macro is for year reports only, but it can also be set up to report the month and year.

```

*** Macro to run univariates based on %varlst ***;
%macro dateit(var);
  title2 "Frequency distribution of &var. by year of &indsn.";
  proc freq data=&indsn.;
    table &var / missing out=temp_dt;
    format &var year4.;
  run;
  /*%if &freqout ~= %then %do;*/
    data temp_dt;
      length datatype $7 varname $20 value $50 count percent cumpct rundate 8;
      set temp_dt;
      varname = "&var.";
      value = put(&var.,year4.);
      percent = round(percent/100,.0001);
      rundate = datetime();
      cumcnt + count;
      *cumpct = round(cumcnt/&subset1,.0001);
      *drop &var. cumcnt;
      retain cumcnt 0;
      format rundate datetime. count comma12. percent cumpct percent9.2;
      label custid = 'Customer ID'
             datatype = 'Data Type'
             varname = 'Variable Name'
             value = 'Value'
             count = 'Count'
             percent = 'Percent'
             cumpct = 'Cumulative Percent'
             rundate = 'Run Date'
             ;
    run;
  /*
  proc append base=freqout.&freqpfx._freqtbl data=temp_dt;
  run;
  */
  /*%end;*/
%mend dateit;

```

The report below shows the FREQ procedure for the variable DOS summarized by year.

OUTPUT REPORT

Sample Customer
Frequency distribution of DOS by year from custds.Sample

The FREQ Procedure

Date of Service

DOS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ff				
2005	27	0.08	27	0.08
2006	31973	99.92	32000	100.00

CONCLUSION

The FREQUINIV utility is a general purpose data analysis tool for use with any SAS data set. The program dynamically determines the report set based on individual variable types and reduces valuable resource time by eliminating redundant programming for multiple data sources.

CONTACT INFORMATION

Thom Clay
2386 Clower Street
Suite C201
Snellville, GA 30078
770-982-8022 ext. 164
tclay@catalysttech.com
www.catalysttech.com

Jamila Ervin
2386 Clower Street
Suite C201
Snellville, GA 30078
770-982-8022 ext. 126
jervin@catalysttech.com
www.catalysttech.com

Michele Laughlin
2386 Clower Street
Suite C201
Snellville, GA 30078
770-982-8022 ext. 133
milaughlin@catalysttech.com
www.catalysttech.com

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.

FREQUINIV FLOWCHART

