

Check Your Data: Tools for Automating Data Assessment

Paul Stutzman

Paul is the Manager of Programming Infrastructure and a Senior Statistical Programmer at Axio Research in Seattle, WA. He has worked with SAS for more than 25 years – the last 13 of which have been in the pharmaceutical industry.

Topics Covered

- ▶ Introduction
- ▶ Example Tools for Assessing Data
- ▶ Components for Building and Automating Assessment Tools
- ▶ Using these Components to Building the Tools
- ▶ Concluding Remarks

Introduction

- ▶ Why are we doing this?
 - Validity and Accuracy
 - Consistency
 - Quality and Efficiency
- ▶ What do we mean by *automation*?

Tools for Assessing Data

Data Assessment Examples

- ▶ Related Variables Within Data Sets
- ▶ Related Variables Across Data Sets
- ▶ Related Variables Across Time
- ▶ All Data Sets, Variables, and Their Data Over Time
- ▶ Chronological Relationships

Assessing Related Variables

Data standards are our friends!

- ▶ VISIT / VISITNUM
- ▶ --TEST / --TESTCD / --STRESU
- ▶ QNAM / QLABEL
- ▶ PARAM / PARAMCD / PARAMN

Related Variables Within Data Sets

Tests, Test Codes, and Units (__STRESU): DOMAIN=LB				
Test Names	Test Codes	Units	Records	
Bilirubin	BILI	umol/L	5959	
C Reactive Protein	CRP		17	
		mg/L	973	
Calcium	CA	mmol/L	5947	
Creatinine	CREAT	umol/L	6065	
Hemoglobin	HGB	g/L	7791	
	HGZ	g/L	3	
Lactate Dehydrogenase	LDH	U/L	5591	
Leukocytes	WBC	10^9/L	7773	

Related Variables Across Data Sets



Compare VISIT/VISITNUM Pairs Across All Domains									
VISIT	EG	LB	MB	MS	PE	SV	TV	VS	XO
SCREENING	1	1	1	1	1	1	1	1	
DAY 1			2	2			2		
DAY 7		8	8	8	8	8	8	9	
DAY 13		14	14	14		14	14	14	
DAY 14			14.01	Missing					
EOS		18	18	18	18	18	18	18	18

Related Variables Across Time



Compare QLABEL/QNAM Pairs			
For the SUPPCM Domain			
	QLABEL	Old QNAM	New QNAM
New QLABEL	Conmed Treament for Adverse Event 3		AELINE3
QLABEL Not in New	Tconmed reament for Adverse Event 3	AELINE3	
	Concomitant Medication Route Others	ROUTEOTH	
QNAMs Match	Conmed Treament for Adverse Event 1	AELINE1	AELINE1
	Conmed Treament for Adverse Event 2	AELINE2	AELINE2

Data Sets, Variables, and Data Over Time



		Record Counts			
				Change	
Dataset	Status	Old	New	Count	Percent
AE		14095	14298	203	1.40%
AM		6913	6962	49	0.70%
CM		26015	24532	-1483	(5.7%)
DM		797	795	-2	(0.3%)
DS		10312	5542	-4770	(46.3%)
DV		4242	4717	475	11.20%
EX		11087	11090	3	0.00%
...					
PD		7650	7742	92	1.20%
PR	Not in Old		2093		
QS		371242	380121	8879	2.40%

Data Sets, Variables, and Data Over Time



		Old Transfer			New Transfer		
Dataset	Variable	Label	Type	Length	Label	Type	Length
AE	AEACN	Action Taken with Study Treatment	Char	200	Action Taken with Study Treatment	Char	40
	AEACNOTH	Other Action Taken	Char	200	Other Action Taken	Char	200
	AEBDSYCD	Body System or Organ Class Code	Num	8	Body System or Organ Class Code	Num	8
	AEBODSYS	Body System or Organ Class	Char	80	Body System or Organ Class	Char	80
	AECAT	Category for Adverse Event	Char	40	Category for Adverse Event	Char	40
	AECONTRT	Concomitant or Additional Trtmnt Given	Char	1	Concomitant or Additional Trtmnt Given	Char	2
	AEDECOD	Dictionary-Derived Term	Char	200	Dictionary-Derived Term	Char	200
	AEENDTC	End Date/Time of Adverse Event	Char	64	End Date/Time of Adverse Event	Char	64
	AEENDY	Study Day of End of Adverse Event	Num	8	Study Day of End of Adverse Event	Num	8
	AEENRF	End Relative to Reference Period	Char	20			
	AEENRTPT	End Relative to Reference Time Point	Char	40	End Relative to Reference Time Point	Char	40

Data Sets, Variables, and Data Over Time



		Counts			Percentages		
Variable	Values	Old	New	Difference	Old	New	Difference
DOMAIN	QS	371242	380121	8879	100.00%	100.00%	0.00%
EPOCH	FOLLOW-UP	37731	45953	8222	10.00%	12.00%	1.90%
	SCREENING	26117	28354	2237	7.00%	7.50%	0.40%
	TREATMENT	307394	305645	-1749	83.00%	80.00%	(2.4%)
	Missing		169			0.00%	
QSCAT	BPI SHORT FORM	105732	104697	-1035	28.00%	28.00%	(0.9%)
	FACT-P	265510	268639	3129	72.00%	71.00%	(0.8%)
	KPS Scale		6785			1.80%	
QSDTC	More than 20 discrete values						
QSDY	More than 20 discrete values						

Data Sets, Variables, and Data Over Time



Variable	N		Minimum		Maximum		Mean		STD	
	Old	New	Old	New	Old	New	Old	New	Old	New
QSDY	369651	378165	-742	-742	716	778	105.0908	108.6304	107.7306	112.5284
QSSEQ	371242	380121	1	1	1479	1507	359.2554	370.0487	254.9425	264.1602
QSSTRESN	353946	364065	0	0	155	155	7.86212	9.207232	19.8587	22.10026
VISITNUM	371122	379889	-1	-1	600120	600140	132786.2	138146.6	201805.8	206520.2

Chronological Relationships



Project X

Project X														
Main Object							Testing/QC							Issues
Exists?			Dates and Times				Exists?			Dates and Times				
Name	Object	Program	Log	Object	Program	Log	Name	Program	Log	Output	Program	Log	Output	
ADAE	Yes	Yes	Yes	28JAN17:09:54	02DEC16:09:16	28JAN17:09:54	TADAE	Yes	Yes	Yes	06DEC16:12:32	19JAN17:10:27	19JAN17:10:27	
ADEX	Yes	Yes	Yes	28JAN17:09:54	02DEC16:14:47	28JAN17:09:54	TADEX	Yes	Yes	Yes	09DEC16:15:27	19JAN17:10:27	19JAN17:10:27	
ADLB	Yes	Yes	Yes	28JAN17:09:55	06DEC16:11:52	28JAN17:09:55	TADLB	Yes	Yes	Yes	06DEC16:13:41	19JAN17:10:27	19JAN17:10:27	
ADMB	Yes	Yes	Yes	02FEB17:08:46	23JAN17:10:50	02FEB17:08:46	TADMB	Yes	Yes	Yes	18JAN17:10:54	19JAN17:10:27	19JAN17:10:27	Test program is older than program. Program is newer than the test program log. Program is newer than the test output.
ADMH	Yes	Yes	Yes	28JAN17:09:54	02DEC16:08:34	28JAN17:09:54	TADMH	Yes	Yes	Yes	05DEC16:16:30	19JAN17:10:27	19JAN17:10:27	

Components for Building and Automating

Components for Building and Automating



- ▶ Structural Information
- ▶ Information About Data
- ▶ Output and Formatting
- ▶ Automation

Structural Information

Information about data sets, variables, and the file system

- ▶ Dictionary Tables
- ▶ The CONTENTS Procedure
- ▶ Pipes

Dictionary Tables

Read-only tables that contain information about the SAS environment and SAS data sets – accessible by PROC SQL only

Columns Dictionary Table:

- LIBNAME
- MEMNAME (data set name)
- NAME (variable name)
- TYPE
- LENGTH
- LABEL

Dictionary Tables

Reading the Columns Dictionary table

```
proc sql noprint;  
  create table columns as  
  select *  
  from dictionary.columns  
  where libname ne "WORK";  
quit;
```


The CONTENTS Procedure



Data set and variable information

```
proc contents data=Old._all_ noprint out=OldData;  
run;
```

Pipes



A way to send commands to the operating system and retrieve the commands results via a FILENAME

- ▶ Establish the pipe

```
filename DirInfo pipe "dir C:\SASData /s" lrecl=32767;
```

- ▶ Get the results

```
data Dir;  
    length InRec $256;  
    infile DirInfo length=LRecl;  
    input InRec $varying256. LRecl;
```

Information About Data

- ▶ ODS OUTPUT and ODS TRACE
- ▶ MEANS and SUMMARY Procedures
- ▶ FREQ Procedure

ODS OUTPUT and ODS TRACE

- ▶ ODS OUTPUT provides information from SAS procedures in SAS datasets, beyond what is available from OUT=
- ▶ ODS TRACE shows the information available as ODS OUTPUT

```
ods trace on;  
proc freq data=DS1 nlevels;  
    tables Var1 / out=OutDS;  
run;  
ods trace off;
```

```
Output Added:  
-----  
Name:      NLevels  
Template:   Base.Freq.NLevels  
Path:      Freq.NLevels  
-----  
  
Output Added:  
-----  
Name:      OneWayFreqs  
Label:     One-Way Frequencies  
Template:   Base.Freq.OneWayFreqs  
Path:      Freq.Table1.OneWayFreqs  
-----
```

ODS OUTPUT and ODS TRACE

- ▶ ODS OUTPUT in action

```
ods output NLevels=Levels;  
proc freq data=DS1 nlevels;  
    tables Var1 / out=LBVis;  
run;
```

- ▶ ODS OUTPUT does not work with the NOPRINT option

```
ods exclude all;  
proc freq ...;  
    ...  
run;  
ods exclude none;
```

MEANS and SUMMARY

Simple descriptive statistics

```
proc means data=DataSet1 noprint;  
    var _NUMERIC_;  
    output out=Means;  
run;
```


FREQ: Basic Frequencies and Percentages



```
proc freq data=SDTM.LB noprint;  
    tables VISITNUM * VISIT / out=LBVis;  
run;
```

► Frequencies for several variables

```
proc freq data=DS1 noprint;  
    tables Var1 Var2 Var3 / out=AllFreqs;  
run;
```

FREQ: Basic Frequencies and Percentages



```
proc freq data=SDTM.LB noprint;  
    tables VISITNUM * VISIT / out=LBVis;  
run;
```

► Frequencies for several variables

```
ods exclude all;  
ods output OneWayFreqs=AllFreqs;  
proc freq data=DS1;  
    tables Var1 Var2 Var3;  
run;  
ods exclude none;
```

FREQ: Number of Discrete Values



```
ods exclude all;  
ods output NLevels=Levels;  
proc freq data=Analysis.ADSL nlevels;  
    tables SEX;  
run;  
ods exclude none;
```

	TableVar	TableVarLabel	NLevels	NMissLevels	NNonMissLevels
1	SEX	Sex	3	1	2

Output and Formatting

- ▶ Display results in a clear, concise manner
- ▶ Make discrepancies and possible issues obvious
 - Traffic Lighting
 - ODS Output Destinations

Traffic Lighting



The use of colors, bolding, etc. to highlight items of interest in order to make them stand out.

PROC REPORT:

```
Col2 = cat("~{style [fontweight=bold foreground=red]", Col2, "}");

compute PctDiff;
    if abs(PctDiff) >= 5 or nmiss(O_Count, P_Count) = 1 then
        call define(_row_, "style", "style=[font_weight=bold
                                foreground=red]");
endcomp;
```

ODS Output Destinations

- ▶ ODS RTF, PDF, and HTML
- ▶ ODS EXCELXP Tagset (EXCEL as XML)
- ▶ ODS EXCEL Tagset (True EXCEL XLSX)
Requires SAS 9.4 (TS1M3)

Automation

- ▶ Setting Macro Variables
- ▶ CALL EXECUTE
- ▶ Keywords for Sets of Variables

Setting Macro Variables

- ▶ CALL SYMPUT (set macro variables from within a DATA step)

```
call symput("Domains", strip(DomNames));
```

- ▶ PROC SQL's INTO Keyword

```
proc sql noprint;  
    select distinct memname into: DataSets separated by " "  
    from dictionary.columns  
    where libname eq "SDTM";  
quit;
```

CALL EXECUTE

Call SAS macros and execute DATA steps and SAS procedures from within DATA steps

```
%QCheck(domain=SUPPDM) ;
```

```
data _null_;  
    set SUPPDoms;  
    call execute(' %QCheck(domain=' || strip(memname) || '); ');  
run;
```

Keywords for Sets of Variables

Select all variables or all variables of a type:

CHARACTER, _NUMERIC_, and _ALL_

```
array Char{*} _CHARACTER_;
```

```
proc means data=DataSet1 noprint;
```

```
var _NUMERIC_;
```

```
...
```

```
proc contents data=SDTM._ALL_ ...;
```

Building the Tools

Related Variables Within Data Sets

- ▶ Find all data sets that contain the --TEST, --TESTCD, and --STRESU: COLUMNS Dictionary table or PROC CONTENTS (DATA=*libname._ALL_*)
- ▶ Open an ODS output destination (RTF, PDF, or EXCELXP Tagset)
- ▶ Process each data set individually:
Macro or DATA _NULL_ and CALL EXECUTE
 - Find all distinct paired values of --TEST/--TESTCD/--STRESU:
PROC FREQ
 - Find and highlight inconsistencies (ODS formatting information)
 - Present the results:
PROC REPORT with ODS

All Data Sets and Variables over Time



- ▶ Get information for all data set (including record counts) and all variables (including variable attributes) for both new and old data transfers

PROC CONTENTS (DATA=libname._ALL_)

- ▶ Open an ODS EXCELXP Tagset destination
- ▶ Find and highlight discrepancies
 - Data sets or variables not in both transfers
 - Data sets with fewer records in new transfer
 - Variables with differing attributes
- ▶ Present the results in individual sheets in the workbook:

PROC REPORT

All Data Sets and Variables over Time



- ▶ Compare categorical variable's values for each data set in both Old and New:
Macro or DATA _NULL_ and CALL EXECUTE for each data set individually
 - Find number of unique values in each variables (both Old and New)
ODS OUTPUT NLEVELS information from PROC FREQ
 - Create a list of all variables whose NLEVELS are below a threshold, and find and compare values for those variables
 - – Find all values for those variables
 - ODS OUTPUT OneWayFreqs information from PROC FREQ
 - – Merge Old and New; Find and highlight discrepancies
 - Present the results in individual sheets in the workbook

Chronological Relationships

Define a set of content and temporal rules:

- ▶ Programs must have corresponding output and logs.
- ▶ A program's timestamp must precede the timestamps for its output and its log.
- ▶ A QC program's timestamp must come after the timestamp of the program it validates.

Chronological Relationships

- ▶ Get OS file system info for programs, QC programs, output, and logs
Pipes submit DIR commands
- ▶ Parse data from the FILENAMES that established the pipes, getting time stamps for all object
- ▶ Merge to get put related objects together (program, output, log, QC program, QC output, QC log), highlighting inconsistencies (ODS formatting information)
- ▶ Open an ODS output destination (RTF, PDF, or EXCELXP Tagset)
- ▶ Present the results:
PROC REPORT

Concluding Remarks

- ▶ These are just a few examples of what can be done.
- ▶ I'd like to know what tools and techniques you use, and what ideas you have for tools you'd like to implement.
- ▶ Check out the paper for more detailed information and references.

Paul Stutzman
Axio Research LLC
pauls@axioresearch.com