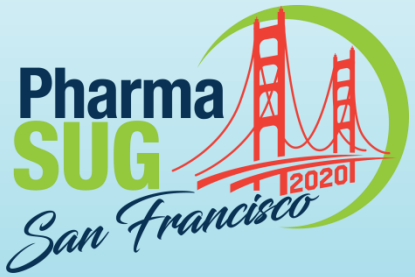
Library Data Sets Summary
Macro %DATA_SPECS

Jeffrey Meyers, Mayo Clinic:

Statistical Programmer Analyst III within Mayo Clinic's Cancer Center Statistics team.

Research focus on gastrointestinal and respiratory cancers as well as meta-analysis.

10 years of SAS programming experience focusing on macros, graphics, SQL and reports.



Library Data Sets Summary Macro %DATA_SPECS

Jeffrey Meyers, Mayo Clinic

%DATA_SPECS

- Creates a data dictionary style document of all datasets in a library
  - Outputs to an Excel file (requires SAS 9.4+)
- 7 parameters (2 required, 5 optional)
- Creates overview worksheet and one additional summary worksheet per dataset
- Contains error checking, documentation, and cleans up after itself
- Contains navigation from first work sheet to dataset specific worksheets

Macro Parameters

- Required
  - LIBN: designates the library (libname not path)
  - OUT: full file path and file name of output Excel document
- Optional
  - INDEX: designate one or more variables as unique identifiers (example: Patient ID)
  - CAT_THRESHOLD: cut-point for numeric variables that determines continuous vs. categorical
  - WHERE: subset dictionary table to exclude certain datasets or variables
  - FORMAT: determines if summary worksheets are listed LONG, WIDE, or CONDENSED
  - ORDER: determines if variables are ordered alphabetically or the order they appear in the dataset

Macro Call Example

```
%data_specs (
    LIBN=sashelp,
    OUT=data_specs_example.xlsx,
    INDEX=id,
    WHERE=memname ^in('LEUTEST' 'LEUTRAIN') and
              substr(memname,1,1)^='V');
```

Macro Error Checking

- The following items are checked by the macro:
  - If the library exists and has been assigned
  - If the *OUT* and *LIBN* parameters are missing in the macro call or set to null
  - If the *CAT_THRESHOLD* parameter is not a number greater than 0
  - If the *FORMAT* parameter is not set to a value of LONG, WIDE, or CONDENSED
  - If the *ORDER* parameter is not set to a value of VARNUM or ALPHA
  - If the current session's SAS version is not at least 9.4 or greater

Macro Error Checking

```
69          %data_specs(LIBN=sashelp2, OUT=data_specs_example.xlsx,
70              INDEX=id, WHERE=memname ^in('LEUTEST' 'LEUTRAIN')
71                                      and substr(memname,1,1)^='V');
ERROR: Library sashelp2 is not assigned
ERROR: Please enter a valid libname
DATA_SPECS has finished processing, runtime:  0:00
```

```
68
69          %data_specs(LIBN=sashelp, OUT=data_specs_example.xlsx, format=wider,
70              INDEX=id, WHERE=memname ^in('LEUTEST' 'LEUTRAIN')
71                                      and substr(memname,1,1)^='V');
ERROR: (Global: FORMAT): wider is not a valid value
ERROR: (Global: FORMAT): Possible values are condensed|wide|long
ERROR: 1 pre-run errors listed
ERROR: Macro DATA_SPECS will cease
DATA_SPECS has finished processing, runtime:  0:00
```

Overview Worksheet

- Table 1:
  - Lists each dataset alphabetically
  - Number of observations
  - Number of unique index values (Number of patients)
  - Number of variables
- Table 2:
  - Displays any variables that exist in multiple datasets
  - Lists all datasets variable exists within
  - Lists all possible labels for that variable

Overview Worksheet (Table 1)

	A	B	C	D
1	Summary of Datasets within Library - NEWSTUD			
2	**Data Set Name**	**Observations**	**Unique Index Values (DCNTR_ID)**	**Number of Variables**
3	CASECRFS	3197	104	48
4	CASECRSE	104	104	105
5	CASEMATS	3427	104	43
6	CASENFVA	3385	0	17
7	CASE_FOLLOWUP	2850	104	18
8	CASE_NFA	0		19
9	CHKLIST	104	104	105
10	CHKLIST_HISTORY	112	101	105
11	COMMENTS	1701	104	17
12	CRSE	104	104	52
13	CRTOX	3987	104	24
14	CYCLE	373	101	84
15	CYTOX	8807	104	24
16	DATACHANGES	1042	65	32
17	DUMPPARMS	1	0	11
18	DUMPSTATUS	669	0	7
19	DZ_TYPE	477	104	58
20	EDITCHKS	869	104	106
21	END_AT	104	104	23
22	EVENT	273	101	115

Overview Worksheet (Table 2)

	A	B	C
42		**Variables that Exist Across Multiple Datasets**	
43	**Variable Name**	**Datasets Containing Variable**	**Variable Label(s)**
44	ACCRUAL	PROTREF, TOX45	Current#Accrual
45	ADJRSN1	CYCLE, PROTDATA	Reason#Dose#Adjusted#or#Held
46	ADJRSN2	CYCLE, PROTDATA	Reason#Dose#Adjusted#or#Held
47	ADJSPEC1	CYCLE, PROTDATA	Specified#Reason#Dose#Adjusted
48	ADJSPEC2	CYCLE, PROTDATA	Specified#Reason#Dose#Adjusted
49	ADJSPEC3	CYCLE, PROTDATA	Specified#Reason#Dose#Adjusted
50	ADRFORM	NONAER, NONAER_HISTORY, TOX45	ADR#Form
51	ADR_TYPE	NONAER, NONAER_HISTORY	Type#of#Expedited#Reporting
52	AGE	CASECRSE, CHKLIST, CHKLIST_HISTORY, CRSE, TOX45	Age
53	AGENT1	CYCLE, PROTDATA	Agent1
54	AGENT2	CYCLE, PROTDATA	Agent2
55	AGENT3	CYCLE, PROTDATA	Agent3
56	ALT	CHKLIST, CHKLIST_HISTORY	ALT
57	ALT_UNL	CHKLIST, CHKLIST_HISTORY	ALT#Upper#Limit
58	ANC	CHKLIST, CHKLIST_HISTORY	ANC#Value
59	ANCILLRY	CASECRSE, PROTREF	Ancillary#Study
60	ARM	CASECRFS, CASECRSE, CASEMATS, CASENFVA, CASE_NFA, CHKLIST, CHKLIST_HISTORY, COMMENTS, CRSE, CRTOX, CYCLE, CYTOX, DATACHANGES, DZ_TYPE, END_AT, EVENT, FISHYDAT, HIDDEN, LABTRACK, LASA, NONAER, NONAER_HISTORY, ONSTUDY, PATHFORM, PAT_PROB, PROTDATA, QOL_COMP, TOX45, TOXICITY	Arm
	CASE	CASECRFS, CASECRSE, CASEMATS, CASENFVA, CASE_FOLLOWUP, CASE_NFA, CHKLIST, CHKLIST_HISTORY, COMMENTS, CRSE, CRTOX, CYCLE, CYTOX, DATACHANGES, DZ_TYPE, EDITCHKS, END_AT, EVENT, FISHYDAT, HIDDEN, LABTRACK, LASA, MATERIAL, NONAER, NONAER_HISTORY, ONSTUDY, PATHFORM, PAT_PROB, PROTDATA, QOL_COMP, SITEHIST, TOX45, TOXICITY	Case Case#Number

Summary Worksheet

- One per dataset
- Displays name, label, format, and a short distribution for each variable
  - Continuous variables
    - N observations, N missing, Median, and range
    - Applies the format to the median/range (works well for dates)
  - Categorical variables
    - Frequency of each value + missing values
    - If formatted value is not equal to unformatted value then both the formatted and unformatted values are shown

Summary Worksheet (Long View)

	A	B
1	**Dataset Name: CRSE**	
2	**Specification**	**Value**
3	Variable	AGE
4	Label	Age
5	Format	Numeric with format F3.
6	Values	N (N Missing): 104 (0) Median: 61 Range: 37 - 81
7	Variable	ARM
8	Label	Arm
9	Format	Character string of length 1 and format $CHAR1.
10	Values	A: 49 (47.1%) B: 55 (52.9%)
11	Variable	BIRTH_DT
12	Label	Date#of#Birth
13	Format	Numeric with format MMDDYY10.
14	Values	N (N Missing): 104 (0) Median: 07/30/1948 Range: 05/28/1928 - 02/06/1973
15	Variable	CASE
16	Label	Case#Number
17	Format	Numeric with format F6.
18	Values	N (N Missing): 104 (0) Median: 255359 Range: 228838 - 263950
19	Variable	DATE_ON
20	Label	Date#on
21	Format	Numeric with format MMDDYY10.
22	Values	N (N Missing): 104 (0) Median: 04/27/2010 Range: 06/23/2009 - 08/11/2010

Summary Worksheet (Wide View)

	A	B	C	D
1				
2	Variable	AGE	ARM	BIRTH_DT
3	Label	Age	Arm	Date#of#Birth
4	Format	Numeric with format F3.	Character string of length 1 and format $CHAR1.	Numeric with format MMDDYY10.
5	Values	N (N Missing): 104 (0) Median: 61 Range: 37 - 81	A: 49 (47.1%) B: 55 (52.9%)	N (N Missing): 104 (0) Median: 07/30/1948 Range: 05/28/1928 - 02/06/1973

Summary Worksheet (Condensed View)

1	Dataset Name: BASEBALL			
2	Variable	Label	Format	Values
15	CrRuns	Career Runs	Numeric with format BEST12.	N (N Missing): 322 (0) Median: 266 Range: 18 - 2165
16	CrRbi	Career RBIs	Numeric with format BEST12.	N (N Missing): 322 (0) Median: 250 Range: 9 - 1659
17	CrBB	Career Walks	Numeric with format BEST12.	N (N Missing): 322 (0) Median: 178.5 Range: 8 - 1566
18	League	League at the End of 1986	Character string of length 8 and format $8.	American: 175 (54.3%) National: 147 (45.7%)
19	Division	Division at the End of 1986	Character string of length 8 and format $8.	East: 157 (48.8%) West: 165 (51.2%)

Summary Worksheet (Continuous Variable)

3	Variable	AGE
4	Label	Age
5	Format	Numeric with format F3.
6	Values	N (N Missing): 104 (0) Median: 61 Range: 37 - 81
11	Variable	BIRTH_DT
12	Label	Date#of#Birth
13	Format	Numeric with format MMDDYY10.
14	Values	N (N Missing): 104 (0) Median: 07/30/1948 Range: 05/28/1928 - 02/06/1973

Summary Worksheet (Categorical Variable)

8	Label	Arm
9	Format	Character string of length 1 and format $CHAR1.
10	Values	A: 49 (47.1%) B: 55 (52.9%)
23	Variable	DCNTR_ID
24	Label	Data#Center#ID
25	Format	Character string of length 25 and format $CHAR25.
26	Values	N (N Missing): 104 (0)
47	Variable	DZ_G
48	Label	Measurable#Disease
49	Format	Numeric with format YESNO7.
50	Values	1 (Yes): 100 (96.2%) 2 (No): 4 (3.8%)

Conclusion

- The DATA_SPECS macro is a powerful tool for summarizing new data quickly or for creating a quick data dictionary
- An example of the utility ODS EXCEL has for creating multi-sheet reports
- The macro is available for download on the SAS Communities page

Name: Jeffrey Meyers
Organization: Mayo Clinic
E-mail: meyers.jeffrey@mayo.edu / jpmeyers.spa@gmail.com