Manitoba Centre for Health Policy

Interpretation of VIMO table

Say Pham Hong
3/27/2015

	Dataset Label: dummy dataset			**Records: 10000**			**Legend (Potential Data Quality Problems) :**						
	Dataset Name: dummy			**Period: yyyy**			**None or Minimal < 5%**	**Moderate 5-30%**	**Significant > 30%**	**Unknown or N/A**			
	= No variance or 100% missing value												
	= Min, Max values based on valid range						**Legend for comment column**						
							Blank = no format have been specified, variables have not been tested for invalid codes						
							✓ = Variables have been tested against the associated format but no invalid values found						

Type	Variable Name	Variable Label	Valid	Invalid	Missing	Outlier	Min	Max	Mean	Median	STD	Comment
ID	VAR1	variable1	100.00		.00							
	VAR2	variable2	.00		100.00							
	VAR3	variable3	99.85		.15							
Num	VAR4	variable4	95.07		4.64	.29	0.77	10.00	8.65	9.18	1.50	
	VAR5	variable5	82.14		.00	17.86	0.00	99.00	4.73	.00	20.80	
	VAR6	variable6	90.80		4.74	4.46	-2.00	92.00	6.26	3.76	8.30	
	VAR7	variable7	.00		100.00	.00						
	VAR8	variable8	59.94		40.06	.00	1.00	99.00	28.92	1.99	42.25	
	VAR9	variable9	93.27		.00	6.73	0.00	26.00	1.24	.00	5.45	
	VAR10	variable10	95.21		4.70	.09	0.00	10.00	8.14	8.85	2.01	
	VAR11	variable11	100.00		.00	.00	0.00	0.00	.00	.00	.00	
	VAR12	variable12	81.91		.00	18.09	0.00	99.00	4.79	.00	20.93	
	VAR13	variable13	87.04		.00	12.96	0.00	7.00	.56	.00	1.73	
	VAR14	variable14	89.18		.00	10.82	0.00	110.00	5.97	.00	22.93	
	VAR21	variable21	95.34		.00	4.66	1.00	6.00	1.12	1.00	.57	
							Top 10 Observed Values					
Codenum	VAR15	variable15	100.00		.00		0, 1, 99					✓
	VAR16	variable16	100.00		.00		0, 1					✓
	VAR17	variable17	98.98		1.02		7, 2, 12, 9, 5					✓
	VAR18	variable18	100.00		.00		1, 0, 99					✓
	VAR19	variable19	100.00		.00		0, 1, 99					✓
	VAR20	variable20	99.24	.31	.45		0, 1, -1					-1 (31 Invalid Obs. in total)
	VAR22	variable22	100.00		.00		1, 0					✓
Char	VAR23	variable23	.00		100.00							
	VAR24	variable24	92.46		7.54		21, 11, 07, 19, 14, 06, 09, 10, 02, 04					
	VAR25	variable25	100.00		.00		15, 138, 75, 137, 88, 84, 146, 24, 78, 148					
	VAR26	variable26	99.92		.08		2, 1					✓
Date	VAR27	variable27	99.76		.24		1955-05-15	2055-10-25				
	VAR29	variable29	47.14		52.86		2001-09-09	2008-12-11				
	VAR30	variable30	.13	43.52	56.35		2000-09-15	2006-03-31				4352 invalid obs. out of [2000-01-01, 2006-04-01] range
	VAR31	variable31	12.30		87.70		2006-01-16	2008-01-21				
Datetime	VAR28	variable28	87.98	12.02	.00		02JAN2001:04:20:32	01APR2006:21:27:55				1202 invalid obs. out of [01JAN2001:23:59:59, 01APR2006:23:59:59] range
Time	VAR32	variable32	100.00		.00		0:00:02	23:59:49				

How to read and interpret VIMO table

The VIMO table contains 13 columns, namely Type, Variable name, Variable label, Valid, Invalid, Missing, Outlier (VIMO), 5 summary statistics (Min, Max, Mean, Median and Standard Deviation) and Comment.

Explanation of the different columns:

The first column **Type** can be divided up to 7 categories depending on the type of variables a dataset contains and these categories are:

- **ID** – These are the ID variables in the dataset
- **Num** – Numeric variables in the dataset
- **Codenum** – This category usually contains character variables that are coded as numeric in the dataset and have formats associated with the variables.
- **Char** – All character variables in the dataset will be put in this category
- **Date** – This category contains SAS date variables
- **Datetime** – This category contains SAS datetime variables
- **Time** – This category contains SAS time variables

The second and third columns of the VIMO table contain the name of the variables and their associated labels. Columns 4 through 7 (Valid, Invalid, Missing, Outlier) are the four main columns of interest.

- **Valid** – Percentage of valid values in the variable
- **Invalid** – Percentage of invalid values in the variable
- **Missing** – Percentage of missing values in the variable
- **Outlier** – Percentage of outlier in the variable (Apply to the **Num** category only)

Note that the Legend(Potential Data Quality Problems) at the middle top of the table are associated with these four columns and it can be interpreted as follow:

1. If the combination of invalid, missing and outlier is less 5%, then data quality problems are considered to be none or minimal (or equivalently, valid column is greater 95%)
2. 5-30% then data quality problems are considered to be moderate (or equivalently, valid column is between 70% – 95%)
3. > 30 % then data quality problems are considered to be significant (or equivalently, valid column is less 70%)

The mean, median and standard deviation columns of the 5 summary statistics are calculated for the numeric variables only (**Num** category), minimum and maximum are calculated for the numeric variables as well as SAS date, date time and time variables. For **Codenum** and **Char** categories, top 10 values of the variables are outputted under the **Top 10 Observed Values** column.

The last column comment might contains the following:

- Blank – variables have not been tested (no formats have been specified for the variables)
- ✓ – variables have been tested against the associated formats and no invalid values found
- All or partial list of invalid codes with the total number of invalid observations in the dataset.