Paper 335-2013

Checking Out Your Dates with SAS®

Christopher J. Bost, MDRC, New York, NY

ABSTRACT

Checking the quality of date variables can be a challenge. PROC FREQ is impractical with a large number of dates. PROC MEANS calculates summary statistics but displays results as SAS® date values. PROC TABULATE, however, can calculate summary statistics and format the results as dates. This paper reviews these approaches plus the STACKODS option in SAS 9.3 that might make PROC MEANS the preferred method for checking out your dates.

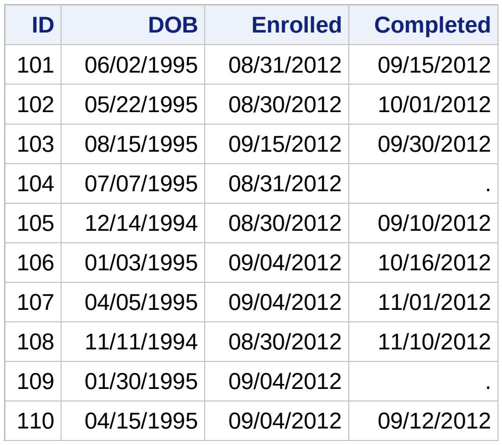
INTRODUCTION

Data quality checks might include running PROC FREQ on categorical variables and PROC MEANS on continuous variables. SAS date values, however, might be considered neither categorical nor continuous. PROC TABULATE is generally preferred. It can calculate statistics and format results as dates where appropriate.

The pros and cons of using PROC FREQ, PROC MEANS, and PROC TABULATE to check the quality of date variables are detailed below.

SAMPLE DATA

Data set PREP is used in this paper. It contains data on ten high school students who have enrolled in a college entrance exam preparation course.



ID	DOB	Enrolled	Completed
101	06/02/1995	08/31/2012	09/15/2012
102	05/22/1995	08/30/2012	10/01/2012
103	08/15/1995	09/15/2012	09/30/2012
104	07/07/1995	08/31/2012	.
105	12/14/1994	08/30/2012	09/10/2012
106	01/03/1995	09/04/2012	10/16/2012
107	04/05/1995	09/04/2012	11/01/2012
108	11/11/1994	08/30/2012	11/10/2012
109	01/30/1995	09/04/2012	.
110	04/15/1995	09/04/2012	09/12/2012

Output 1. Data set PREP

Data set PREP contains ten observations and four variables:

ID is a unique identifier for each student
DOB is the student's date of birth
ENROLLED is the date the student started the prep course
COMPLETED is the date the student finished the prep course

Note the variation in date values. DOB values vary widely, from 11/11/1994 through 08/15/1995. ENROLLED values vary little, from 08/30/2012 through 09/15/2012; most dates are clustered in late August/early September. COMPLETED values range from 09/10/2012 to 11/10/2012. Note that two values are missing (i.e., for students who have not completed the prep course).

We want to know the variable name, label, sample size, number of missing values, minimum value (earliest date), maximum value (latest date), median value (the date at or before which 50% of the values fall), and range of values (number of days between the minimum date and the maximum date). In some cases, we might also want to know the frequencies and percentages of values.

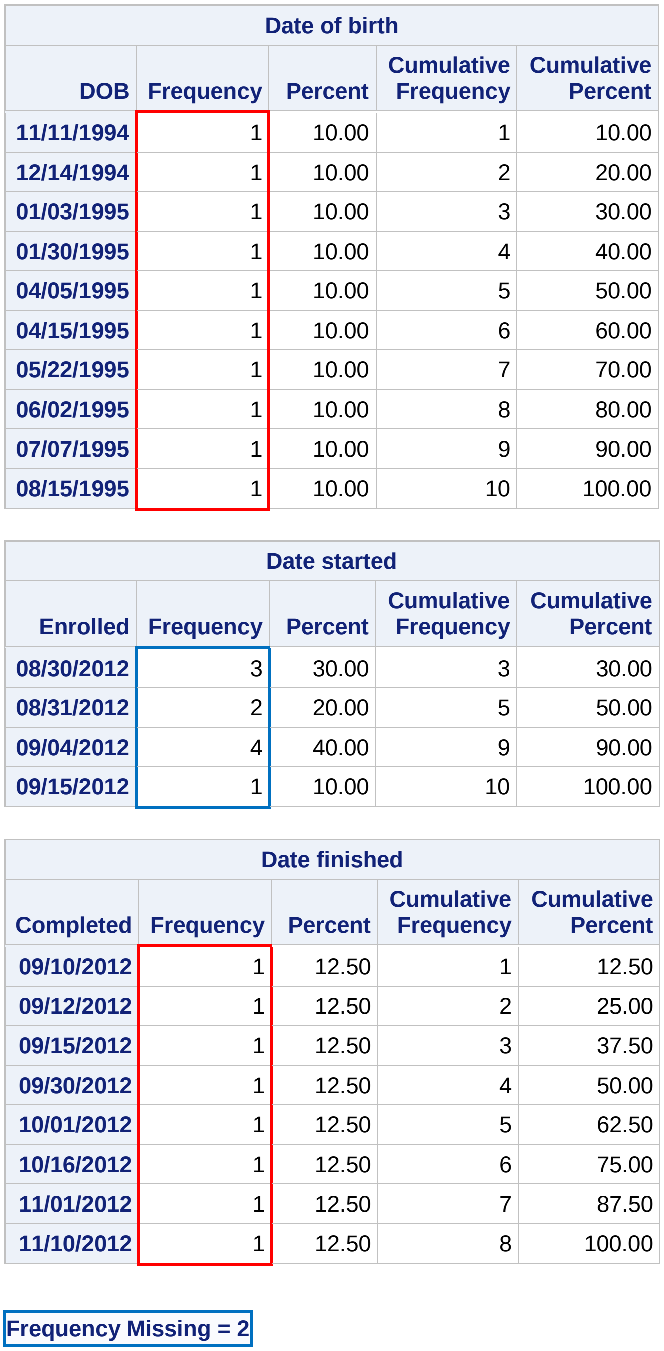
CHECKING DATES WITH PROC FREQ

PROC FREQ can be used to check (some) date variables. The syntax is:

```
proc freq data=prep;
tables DOB Enrolled Completed;
run;
```

The PROC FREQ statement starts the procedure.

The TABLES statement specifies one-way frequencies for DOB, ENROLLED, and COMPLETED.



Date of birth

DOB	Frequency	Percent	Cumulative Frequency	Cumulative Percent
11/11/1994	1	10.00	1	10.00
12/14/1994	1	10.00	2	20.00
01/03/1995	1	10.00	3	30.00
01/30/1995	1	10.00	4	40.00
04/05/1995	1	10.00	5	50.00
04/15/1995	1	10.00	6	60.00
05/22/1995	1	10.00	7	70.00
06/02/1995	1	10.00	8	80.00
07/07/1995	1	10.00	9	90.00
08/15/1995	1	10.00	10	100.00

- Each value of DOB has a Frequency of 1.

Date started

Enrolled	Frequency	Percent	Cumulative Frequency	Cumulative Percent
08/30/2012	3	30.00	3	30.00
08/31/2012	2	20.00	5	50.00
09/04/2012	4	40.00	9	90.00
09/15/2012	1	10.00	10	100.00

- Values of ENROLLED vary in Frequency.

Date finished

Completed	Frequency	Percent	Cumulative Frequency	Cumulative Percent
09/10/2012	1	12.50	1	12.50
09/12/2012	1	12.50	2	25.00
09/15/2012	1	12.50	3	37.50
09/30/2012	1	12.50	4	50.00
10/01/2012	1	12.50	5	62.50
10/16/2012	1	12.50	6	75.00
11/01/2012	1	12.50	7	87.50
11/10/2012	1	12.50	8	100.00

- Each value of COMPLETED has a Frequency of 1.

Frequency Missing = 2

- COMPLETED has 2 missing values.

Output 2. Checking dates with PROC FREQ

DOB values each have a Frequency of 1. This table is not very useful.

COMPLETED values each have a Frequency of 1. Again, this is not useful. The two missing values are noted.

ENROLLED values, however, include four dates. The sample size (Cumulative Frequency) is 10. The minimum value (08/30/2012) and the maximum value (09/15/2012) are easily determined. Frequency and Percent columns are useful.

Pros PROC FREQ output includes the variable name, label, sample size, number of missing values, minimum value, and maximum value. Counts and percentages are also calculated.

Cons Output could be prohibitively long with a large number of date values. PROC FREQ does not calculate the median value or the range of values.

Recommendation Use PROC FREQ to check date variables with a limited number of values.

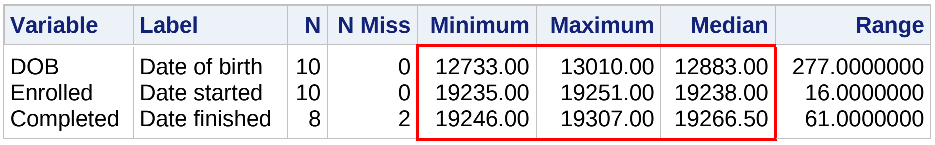
CHECKING DATES WITH PROC MEANS

PROC MEANS (seems like it) can be used to check date variables. The syntax is:

```
proc means data=prep n nmiss min max median range;
var DOB Enrolled Completed;
run;
```

The PROC MEANS statement starts the procedure. The type and order of summary statistics is specified (i.e., N NMISS MIN MAX MEDIAN RANGE).

The VAR statement specifies the analysis variables DOB, ENROLLED, and COMPLETED.



Variable	Label	N	N Miss	Minimum	Maximum	Median	Range
DOB	Date of birth	10	0	12733.00	13010.00	12883.00	277.0000000
Enrolled	Date started	10	0	19235.00	19251.00	19238.00	16.0000000
Completed	Date finished	8	2	19246.00	19307.00	19266.50	61.0000000

▪ Results are SAS date values.

Output 3. Checking dates with PROC MEANS

Pros PROC MEANS output includes the variable name, label, sample size, number of missing values, and range of values.

Cons The minimum value, maximum value, and median value are displayed as SAS date values (i.e., number of days from January 1, 1960). Results cannot be formatted as dates.

Recommendation Do not use PROC MEANS output to check date variables.

CHECKING DATES WITH PROC TABULATE

PROC TABULATE can be used to check date variables. The syntax is:

```
proc tabulate data=prep;
var DOB Enrolled Completed;
table DOB Enrolled Completed,
      n nmiss (min max median)*f=mmddyy10. range;
run;
```

The PROC TABULATE statement starts the procedure.

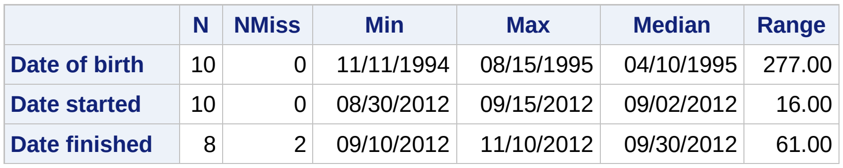
The VAR statement specifies the analysis variables DOB, ENROLLED, and COMPLETED.

The TABLE statement defines the table. Row dimensions are specified before the comma and column dimensions are specified after the comma.

DOB, ENROLLED, and COMPLETED will be in the rows of the table.

N, NMISS, MIN, MAX, MEDIAN, and RANGE will be in the columns of the table.

MIN, MAX, and MEDIAN are in parentheses followed by *F=MMDDYY10. This applies the specified date format to all three columns.



	N	NMiss	Min	Max	Median	Range
Date of birth	10	0	11/11/1994	08/15/1995	04/10/1995	277.00
Date started	10	0	08/30/2012	09/15/2012	09/02/2012	16.00
Date finished	8	2	09/10/2012	11/10/2012	09/30/2012	61.00

Output 4. Checking dates with PROC TABULATE

Pros PROC TABULATE output includes the variable label, sample size, number of missing values, minimum value, maximum value, median value, and range of values. The minimum, maximum, and median values are formatted as dates.

Cons PROC TABULATE output includes the variable name or label (if present) but not both. Syntax is, arguably, less intuitive than other procedures.

Recommendation Use PROC TABULATE to check date variables with a large number of values.

CHECKING DATES WITH PROC MEANS REVISITED

SAS output is managed by the Output Delivery System (ODS). Results can be saved in different formats (e.g., HTML, RTF, or PDF) as well as to SAS data sets. PROC MEANS has one ODS output object named SUMMARY. It can be saved to a SAS data set, formatted, and printed.

SAS 9.3 supports the new STACKODS option. It is specified on the PROC MEANS statement (and is an alias for STACKODSOUTPUT). The STACKODS option produces an ODS output data set that is similar to PROC MEANS printed output. The syntax to create a “stacked” data set is:

```
ods output summary=stacked;
proc means data=prep n nmiss min max median range stackods;
var DOB Enrolled Completed;
run;
```

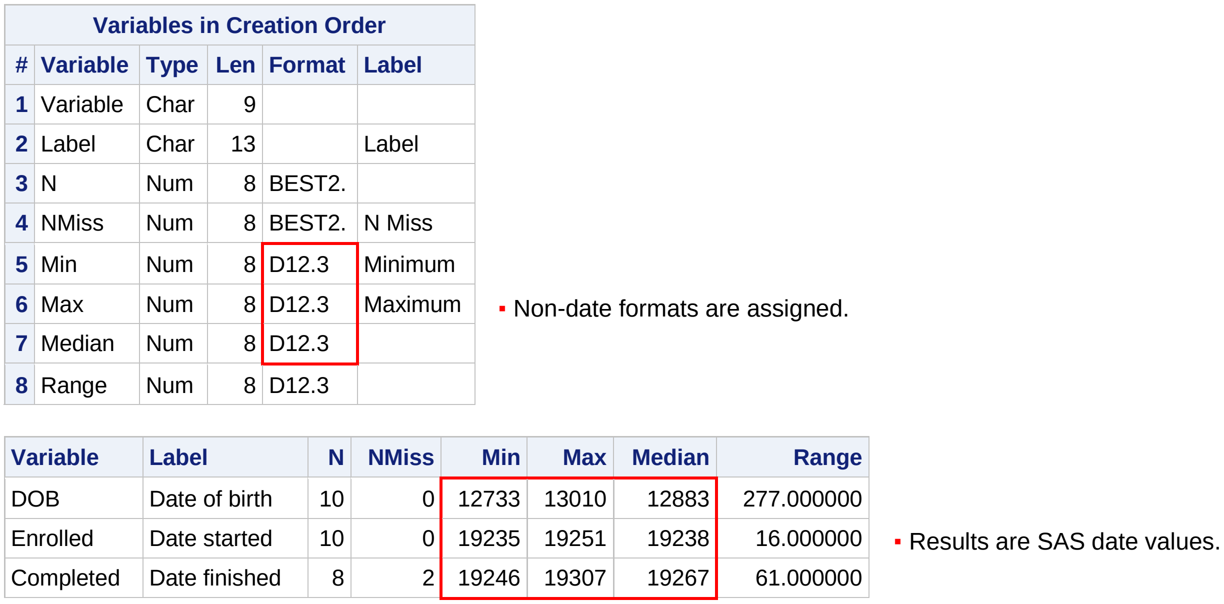
The ODS OUTPUT statement stores information from SUMMARY in a SAS data set named STACKED.

The PROC MEANS statement starts the procedure. The type and order of summary statistics is specified.

The STACKODS option structures the output data set like PROC MEANS output.

The VAR statement specifies the analysis variables DOB, ENROLLED, and COMPLETED.

The stacked data set is different from the standard PROC MEANS output data set and merits inspection:



Variables in Creation Order

#	Variable	Type	Len	Format	Label
1	Variable	Char	9		
2	Label	Char	13		Label
3	N	Num	8	BEST2.	
4	NMiss	Num	8	BEST2.	N Miss
5	Min	Num	8	D12.3	Minimum
6	Max	Num	8	D12.3	Maximum
7	Median	Num	8	D12.3	
8	Range	Num	8	D12.3	

▪ Non-date formats are assigned.

Variable	Label	N	NMiss	Min	Max	Median	Range
DOB	Date of birth	10	0	12733	13010	12883	277.000000
Enrolled	Date started	10	0	19235	19251	19238	16.000000
Completed	Date finished	8	2	19246	19307	19267	61.000000

▪ Results are SAS date values.

Output 5. PROC CONTENTS and PROC PRINT of STACKODS output data set

The minimum value, maximum value, and median value are displayed as SAS date values. (Note that the values of RANGE are also formatted with D12.3, which is excessive given our data.) Because results are saved in a SAS data set, however, variables can be formatted as needed. For example:

```
proc print data=stacked noobs;
format min max median mmddyy10. range 3.0;
run;
```

PROC PRINT prints the values in data set STACKED. The NOOBS option suppresses observation numbers.

The FORMAT statement prints MIN, MAX, and MEDIAN as *mm/dd/yyyy* and RANGE as *nnn:*

Variable	Label	N	NMiss	Min	Max	Median	Range
DOB	Date of birth	10	0	11/11/1994	08/15/1995	04/10/1995	277
Enrolled	Date started	10	0	08/30/2012	09/15/2012	09/02/2012	16
Completed	Date finished	8	2	09/10/2012	11/10/2012	09/30/2012	61

Output 6. Checking dates with PROC MEANS revisited

Pros PROC PRINT output includes the variable name, label (if present), sample size, number of missing values, minimum value, maximum value, median value, and range of values. The minimum, maximum, and median values are formatted as dates.

Cons This technique requires learning the new STACKODS option. Two steps are needed.

Recommendation Use the PROC MEANS stacked output data set to check date variables.

Note that the SAS 9.3 Windowing environment closes the LISTING destination and opens the HTML destination by default. (Settings can be adjusted with Tools > Options > Preferences… > Results tab.) To avoid printing the PROC MEANS output with unformatted dates, use ODS HTML CLOSE; before the PROC MEANS step and ODS HTML; after the PROC MEANS step (i.e., turn output off and on programmatically). Do not use the NOPRINT option on the PROC MEANS statement; if there is nothing to print, ODS OUTPUT has nothing to store in a SAS data set. Enterprise Guide users can control output with Tools > Options… > Results > Results General. When only Text output is checked under Result Formats, use ODS LISTING CLOSE; before the PROC MEANS step and ODS LISTING; after the PROC MEANS step (i.e., turn output off and on programmatically).

CONCLUSION

PROC FREQ, PROC MEANS, and PROC TABULATE can be used to check the quality of date variables. PROC FREQ counts the frequency of values and is useful with a moderate number of dates. PROC MEANS calculates summary statistics but cannot format results as dates. PROC TABULATE calculates summary statistics and can format results as dates. PROC MEANS with the STACKODS option and an ODS output data set also calculates summary statistics and results can be formatted as dates.

REFERENCES

SAS Institute Inc. 2011. *Base SAS® 9.3 Procedures Guide.* Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Christopher J. Bost
MDRC
16 East 34th Street
New York, NY 10016
(212) 340-8613
christopher.bost@mdrc.org
chrisbost@gmail.com