Paper 162

Creating Quartiles from Continuous Responses: Making Income Data Manageable

Michelle M. Dahnke, DrPH, Florida A&M University; Tyra Dark, PhD, Florida A&M University

ABSTRACT

It is customary to collect data at the most granular level; however sometimes that requires consolidating responses in categories before using them in advanced analysis. For example, in the Collaborative Psychiatric Epidemiology Surveys the household income variable is a continuous variable with individual responses ranging from 0 to $200,000. Working with the data may require categorization so that the data is more manageable—like in quartiles. This paper illustrates the creation of household income quartiles from free response data, an important fundamental skill in working with large data sets.

INTRODUCTION

Household income is a common demographic variable to include when conducting analysis on data from survey participants. Sometimes household income data is collected categorically, which can make it easy to work with, but it can also be collected and organized in a continuous fashion, which can result in a wide range of individual responses. This is the case with the household income variable (V8683) from the Collaborative Psychiatric Epidemiology Surveys, which includes more than 100 discrete responses from more than 20,000 survey participants.

EVALUATING THE RESPONSES AND DETERMINING THE VALUES FOR QUARTLES

Creating categories for data, like quartiles, is straightforward in SAS®. The first step in creating quartiles is to evaluate the distribution of responses for the household income variable. This can be accomplished by running a PROC FREQ on the variable. In this instance, the household income variable is called “V8683”. An example of this PROC FREQ code is below.

```
PROC FREQ;
TABLES V08683;
RUN;
```

As previously mentioned, there are more than 100 discrete responses so the full output is not included, but rather an excerpt from the beginning of the output is shown in Figure 1.

The SAS System

The FREQ Procedure

Household Income : Topcode

V08683	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	327	1.99	327	1.99
200	2	0.01	329	2.00
250	3	0.02	332	2.02
294	1	0.01	333	2.03
400	2	0.01	335	2.04
444	1	0.01	336	2.05
477	1	0.01	337	2.05
500	90	0.55	427	2.60
531	1	0.01	428	2.61
545	1	0.01	429	2.61

Figure 1 Excerpt of the Beginning of a PROC FREQ Output

The output includes the response, frequency, percent, cumulative frequency and cumulative percent. The cumulative percent column identifies where the cut points for the quartiles should be—as close to 25, 50 and 75 cumulative percent as possible. Figures 2, 3 and 4 show an excerpt from the output around these cut points.

16980	1	0.01	4012	24.43
16999	10	0.06	4022	24.49
17000	87	0.53	4109	25.02
17036	1	0.01	4110	25.03

Figure 2 Excerpt from PROC FREQ Output around 25 Cumulative Percent

37000	72	0.44	8210	49.99
37044	1	0.01	8211	50.00
37099	1	0.01	8212	50.00
37188	1	0.01	8213	50.01
37232	1	0.01	8214	50.02

Figure 3 Excerpt from PROC FREQ Output around 50 Cumulative Percent

66998	1	0.01	12302	74.91
66999	2	0.01	12304	74.92
67000	33	0.20	12337	75.12
67084	2	0.01	12339	75.13

Figure 4 Excerpt from PROC FREQ Output around 75 Cumulative Percent

The end of the output for PROC FREQ shows the cumulative percent is 100.00. The end of the output table also provides the frequency of missing responses for the variable as demonstrated in Figure 5.

199998	2	0.01	15995	97.39
199999	1	0.01	15996	97.40
(200000.00000) TOP CODED AT $200,000	427	2.60	16423	100.00
Frequency Missing = 3590				

Figure 5 Excerpt from the End of a PROC FREQ Output

It is simple to identify the values for the household income quartiles by observing these figures. The quartiles are (1) $0-$17,000, (2) $17001-$37,099, (3) $37,100-$67,000 and (4) $67,001-$200,000.

CREATING THE QUARTILES

Now that the quartile cut points are identified, use an IF-THEN statement to have SAS® create the quartiles. An example of IF-THEN code is below.

```
IF (V08683 <= 17000) and (V08683 >=0) THEN Income = '1';
IF (V08683 >= 17001) and (V08683 <= 37099) THEN Income = '2';
IF (V08683 >= 37100) and (V08683 <= 67000) THEN Income = '3';
IF (V08683 >= 67001) THEN Income = '4';
RUN;
```

This IF-THEN statement also creates a meaningful variable name changing what was “V08683” to “Income”. Finally, run another PROC FREQ with the new variable name to see the newly created quartiles. The updated PROC FREQ code is below and the output showing the new quartiles is in Figure 6.

```
PROC FREQ;
TABLES Income;
RUN;
```

The SAS System

The FREQ Procedure

Income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	4109	25.02	4109	25.02
2	4103	24.98	8212	50.00
3	4125	25.12	12337	75.12
4	4086	24.88	16423	100.00
Frequency Missing = 3590				

Figure 6 Newly Created Quartiles

In this instance, there was not exactly the same number of responses per quartile because the cumulative percent was not exact for 25.0 or 75.0, but it is very close as the frequency column shows. One way to check the data and verify all of the identified responses are included in the newly created quartiles is by comparing the frequency of missing responses. In both the original presentation of the data and in the table with the newly created quartiles, the frequency missing is 3590.

CONCLUSION

Quartiles can be a practical way to organize data before using it for analysis. This paper was written to highlight the ease with which free response data can be categorized into quartiles, even if there are thousands of responses.

REFERENCES

Collaborative Psychiatric Epidemiology Surveys. *Welcome to CPES.* July 1, 2015. Available at http://www.icpsr.umich.edu/icpsrweb/CPES

ACKNOWLEDGMENTS

The first author would like to thank the second author, Dr. Tyra Dark, and the other members of her dissertation committee being that this SAS® skill was learned while completing analysis for her dissertation. The remaining committee members are Dr. Gebre Kiros (FAMU), Dr. Saleh Rahman (FAMU), Dr. Hong Xiao (UF) and Dr. Roger Boothroyd (USF).

CONTACT INFORMATION

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

Name: Michelle M. Dahnke, DrPH
Enterprise: Florida A&M University Institute of Public Health
Address: 1515 S. Martin Luther King, Jr. Blvd, SRC207
City, State ZIP: Tallahassee, Florida 32307
Phone: 954.494.7471
E-mail: mdahnke@gmail.com