

Taking the PROC FREQ a Step Beyond

Priya Suresh and Elizabeth Heath

RTI International, RTP, NC

ABSTRACT

Some surveys administer the same set of questions, at particular time intervals, in order to study changes over time. In longitudinal studies, the questions are administered to the same person, while in multiple-year cross-sectional studies, the questions are administered to different cohorts or different people within the same demographic group. Studies that we have supported have followed respondent's drug use, academic success, and social well-being, over time, and collected the same data year after year. Changes over time are identified as variable values that significantly differ by more than a specific percentage threshold. One way to identify these changes involves the use of the SAS® frequency procedure PROC FREQ. We will review this simple but elegant approach to highlighting changes.

INTRODUCTION

We collect the same data variables year after year and derive key analytical variables from these data variables. We use Proc Freq to review and verify that analytical variables have been derived correctly. We also use Proc Freq to quickly identify variations in values and the deviation from one year to the next. If a new value occurs or if the deviation is significant, it could mean that a scenario has occurred that has not happened before. In that case, we do further investigation to confirm that the deviation is valid.

BRUTE-FORCE BROWSING IS INSUFFICIENT AND INEFFICIENT

Browsing the Proc Freq output for a year's worth of data is not sufficient because you cannot guarantee that you will catch all of the changes. Of course, you can compare the frequency distributions side by side but that is very inefficient, not to mention mind-numbing! Ideally, you want to highlight only the frequency values that are significantly different between the datasets.

You can imagine the magnitude of the review task, given that each annual dataset holds hundreds of variables and the frequency output can be hundreds of pages. Since anomalies occur in data, it is imperative that we review variable frequencies for a specific year dataset and compare them against the prior years' data.

Reviewing and re-approving the frequencies that existed in a prior year is literally a waste of time. So, the obvious next step is to highlight the differences in an easy to spot manner. How does one go about highlighting the anomalies? In other words, how do we take the Proc Freq a step beyond?

THE BIG PICTURE

The technique developed is as follows:

- a) Generate the Proc Freq output as a dataset for a specific variable from the current dataset
- b) Generate the Proc Freq output as a dataset for the same variable from the prior dataset and rename the count and percentage variables.
- c) Merge the two Proc Freq output datasets by the variable name.
- d) In the merge step:
 1. Create a customized variable called "chg", for change, that will highlight significant deviations.
 2. Assign values of variable 'chg' when the difference between the percentages is more than a threshold value.
- e) Use Proc Print to display the merged output.
- f) Review the highlighted frequencies.

OVERVIEW WITH A FICTITIOUS EXAMPLE

We will first review a summary of fictitious data to explain the code for our approach. Our fictitious data about the number of times a respondent traveled more than 500 miles roundtrip in 2004 and 2008 is summarized in variable traveltm in Table 1. Variable traveltm has four categories or values that include: no travel or 0 times, one to five times, six to ten times, and more than 10 times.

Table 1. *Summary of fictitious travel data*

Variable traveltm - number times traveled more than 500 miles roundtrip	Percentages		Percent change	Flag above 5 % and below -5 % change
	2004	2008		
0	50	70	+20	+
1-5	20	10	-10	-
6-10	25	15	-10	-
> 10	5	5	0	
TOTAL	100	100		

The fictitious data show that the percent change decreased in 2008 versus 2004 for the data categories of one to five times and six to ten times and increased in 2008 versus 2004 for respondents who reported that they did not travel more than 500 miles roundtrip. Table 1 also contains flags to mark when the percent change with respect to 2008 is greater than a specific percent cutoff. A '+' flag denotes that the change is greater than 5 percent, while a '-' flag denotes that the change is less than minus five percent.

IDENTIFYING DATA VARIATIONS

There are two general steps in identifying data variations. First, the analyst obtains data counts and percentages for the same variable in two different years, and secondly, the analyst identifies and flags differences between the percentages for the two years. We will now review how SAS ® can do these two tasks.

OBTAIN DATA COUNTS AND PERCENTAGES

We use the SAS ® frequency procedure PROC FREQ to output the counts and percentages of variable traveltm to a sas dataset called temp04 for the year 2004 data, with the code shown in lines A1a and A1b. Figure 1 shows the counts and percentages for the 2004 data.

```
proc freq data = data04;                                (A1a)
tables traveltm / out=temp04 noprint;run;                (A1b)
```

Figure 1. *Dataset temp04 counts and percentages*

Obs	traveltm	COUNT	PERCENT
1	0	50000	50
2	1-5	20000	20
3	6-10	25000	25
4	> 10	5000	5

We use similar code for the year 2008 data, but rename the variables percent and count to pct08 and count08, respectively, in order to distinguish the 2008 data, using the code shown in lines B1a and B1b. Figure 2 shows the counts and percentages for the 2008 data.

```
proc freq data = data08;                                (B1a)
tables traveltm / out=temp08
(rename=(percent=pct08 count=count08)) noprint; run;    (B1b)
```

Figure 2. *Dataset temp08 counts (count08) and percentages (pct08)*

Obs	traveltm	count08	pct08
1	0	70000	70
2	1-5	10000	10
3	6-10	15000	15
4	> 10	5000	5

DETERMINE AND FLAG PERCENT DIFFERENCES

Percent difference cutoffs and flags can be defined in several ways. For our discussion data, we want to identify where response values differ by more than five percent or less than negative five percent, with respect to the 2008 data. When the percent difference is more than five percent, we want to flag the counts with '+', and when the percent difference is less than negative five percent, we want to flag the counts with '-'. Now we will review how to do this using SAS ® statements.

We merge the datasets temp04 and temp08 by variable traveltm in lines C1a through C1c. In lines C1d through C1f, we check the percent difference for each value or category of variable traveltm and flag values above five percent and below negative five percent. We summarize variations in variable traveltm by printing variable traveltm, counts and percentages for 2004 and 2008, variable diffpct which is the difference in percentages for 2004 and 2008, and variable chg, which is the change flag, using the code in lines D1a and D1b. The print output is shown in Figure 3.

```
data comb;                                (C1a)
merge temp04 temp08;                      (C1b)
by traveltm;                              (C1c)
diffpct = pct08 - percent;                 (C1d)
if diffpct < -5 then chg = '-';           (C1e)
else if diffpct > 5 then chg = '+'; run;   (C1f)

proc print data = comb;                    (D1a)
var traveltm count count08 percent pct08 diffpct chg; run; (D1b)
```

Figure 3. Summary of variations in variable traveltm for 2004 and 2008 data.

Obs	traveltm	COUNT	count08	PERCENT	pct08	diffpct	chg
1	0	50000	70000	50	70	20	+
2	1-5	20000	10000	20	10	-10	-
3	6-10	25000	15000	25	15	-10	-
4	> 10	5000	5000	5	5	0	

AN ACTUAL EXAMPLE

Using publicly available data from the General Social Survey (<http://sda.berkeley.edu/archive.htm>), the variables Race, Ethnic(ity) and Educ(ation) have been compared between year 2002 and 2000, in Figures 4, 5, and 6, respectively. In this example, 'diffperc' is the difference in 2002 and 2000 percentages; 'star' flags percentage differences for a threshold of +/- 0.5% as well as when values are missing in one of the datasets.

Figure 4. Comparison of 2002 and 2000 race data

General Social Surveys, 1972-2006							1
GSS VARIABLES (out.GSS_2002 vs out.GSS_2000)							
19:32 Thursday, June 18, 2009							
Obs	RACE	COUNT	count2	PERCENT	cent2	diffperc	star
1	WHITE	2188	2213	79.1320	78.5588	0.57326	*
2	BLACK	410	429	14.8282	15.2290	-0.40076	
3	OTHER	167	175	6.0398	6.2123	-0.17250	

Figure 5. Comparison of 2002 and 2000 ethnicity data

General Social Surveys, 1972-2006
GSS VARIABLES (out.GSS_2002 vs out.GSS_2000)
19:32 Thursday, June 18, 2009

Obs	ETHNIC	COUNT	count2	PERCENT	cent2	diffperc	star
1	.	630	572	.	.	.	
2	AFRICA	228	232	10.6792	10.3341	0.34508	
3	AUSTRIA	10	13	0.4684	0.5791	-0.11068	
4	FRENCH CANADA	32	37	1.4988	1.6481	-0.14928	
5	OTHER CANADA	16	12	0.7494	0.5345	0.21489	
6	CHINA	14	8	0.6557	0.3563	0.29939	
7	CZECHOSLOVAKIA	19	25	0.8899	1.1136	-0.22366	
8	DENMARK	14	17	0.6557	0.7572	-0.10150	
9	ENGLAND & WALES	231	271	10.8197	12.0713	-1.25160	*
10	FINLAND	7	7	0.3279	0.3118	0.01606	
11	FRANCE	48	53	2.2482	2.3608	-0.11256	
12	GERMANY	330	383	15.4567	17.0601	-1.60346	*
13	GREECE	11	16	0.5152	0.7127	-0.19747	
14	HUNGARY	12	6	0.5621	0.2673	0.29480	
15	IRELAND	266	258	12.4590	11.4922	0.96681	*
16	ITALY	125	115	5.8548	5.1225	0.73231	*
17	JAPAN	7	7	0.3279	0.3118	0.01606	
18	MEXICO	84	90	3.9344	4.0089	-0.07448	
19	NETHERLANDS	27	32	1.2646	1.4254	-0.16075	
20	NORWAY	40	49	1.8735	2.1826	-0.30909	
21	PHILIPPINES	18	25	0.8431	1.1136	-0.27049	
22	POLAND	67	63	3.1382	2.8062	0.33194	
23	PUERTO RICO	31	37	1.4520	1.6481	-0.19612	
24	RUSSIA	39	45	1.8267	2.0045	-0.17776	
25	SCOTLAND	82	80	3.8407	3.5635	0.27728	
26	SPAIN	18	17	0.8431	0.7572	0.08585	
27	SWEDEN	35	33	1.6393	1.4699	0.16941	
28	SWITZERLAND	4	8	0.1874	0.3563	-0.16899	
29	WEST INDIES	4	6	0.1874	0.2673	-0.07991	
30	OTHER	6	29	0.2810	1.2918	-1.01073	*
31	AMERICAN INDIAN	118	109	5.5269	4.8552	0.67170	*
32	INDIA	11	15	0.5152	0.6682	-0.15293	
33	PORTUGAL	6	8	0.2810	0.3563	-0.07532	
34	LITHUANIA	5	6	0.2342	0.2673	-0.03307	
35	YUGOSLAVIA	3	9	0.1405	0.4009	-0.26038	
36	RUMANIA	6	.	0.2810	.	.	*
37	BELGIUM	6	2	0.2810	0.0891	0.19194	
38	ARABIC	8	1	0.3747	0.0445	0.33016	
39	OTHER SPANISH	34	29	1.5925	1.2918	0.30075	
40	NON-SPAN WINDIES	16	13	0.7494	0.5791	0.17035	
41	OTHER ASIAN	15	15	0.7026	0.6682	0.03442	
42	OTHER EUROPEAN	17	8	0.7963	0.3563	0.43991	
43	AMERICAN ONLY	65	56	3.0445	2.4944	0.55006	*

Figure 6. *Comparison of 2002 and 2000 education data*

3

General Social Surveys, 1972-2006
GSS VARIABLES (out.GSS_2002 vs out.GSS_2000)
19:32 Thursday, June 18, 2009

Obs	EDUC	COUNT	count2	PERCENT	cent2	diffperc	star
1	.	12	9	.	.	.	
2	0	5	1	0.1816	0.0356	0.14601	
3	1	2	.	0.0726	.	.	*
4	2	15	3	0.5449	0.1068	0.43802	
5	3	2	6	0.0726	0.2137	-0.14103	
6	4	6	6	0.2179	0.2137	0.00427	
7	5	8	9	0.2906	0.3205	-0.02992	
8	6	25	22	0.9081	0.7835	0.12462	
9	7	14	20	0.5085	0.7123	-0.20371	
10	8	66	95	2.3974	3.3832	-0.98581	*
11	9	60	69	2.1794	2.4573	-0.27782	
12	10	88	106	3.1965	3.7749	-0.57842	*
13	11	137	153	4.9764	5.4487	-0.47233	
14	12	818	823	29.7130	29.3091	0.40392	
15	13	265	290	9.6259	10.3276	-0.70177	*
16	14	374	374	13.5852	13.3191	0.26609	
17	15	157	127	5.7029	4.5228	1.18008	*
18	16	377	393	13.6942	13.9957	-0.30157	
19	17	93	90	3.3781	3.2051	0.17300	
20	18	128	114	4.6495	4.0598	0.58964	*
21	19	43	39	1.5619	1.3889	0.17304	
22	20	70	68	2.5427	2.4217	0.12103	

CONCLUSION

We showed how to quickly identify variations in data from multiple rounds of data collection. This simple yet powerful and reliable technique can be easily extended using SAS ® macros to automatically flag variations in hundreds of variables.

REFERENCES

SAS® Version 9 On-line Documentation

General Social Survey 1972-2006 data from UC Berkeley's SDA archive; <http://sda.berkeley.edu/archive.htm>

CONTACT INFORMATION

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

Priya Suresh

Phone: (919) 541-7428

E-mail: psoh@rti.org

Elizabeth Heath

Phone: (919) 485-2786

E-mail: eah@rti.org

The authors can be reached at the mailing address:

RTI International

PO Box 12194

RTP NC 27709-2194.

TRADEMARK NOTICE

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.