SESUG Paper 101-2024

SAS® and Analytical Research: A Primer for Qualitative Analysis

Smith, Kelly D., AEWAS Consulting

ABSTRACT

New SAS® users are often encouraged to use lexjansen.com as a resource when they hit a coding bump. As of May 31, 2024, the website contained more than 38,000 conference papers that can be searched by title, author, or keyword (thank you Lex!).

How can a new SAS user figure out the paper(s) most likely to be helpful? Perhaps they could find a "primer" or "introduction" paper. A search for "primer" yielded fewer than 100 papers; "introduction" returned nearly 800 papers.

Searching for a specific proc can return an overwhelming number of papers. For example, there are nearly 4,800 papers focused on "proc freq." What if a new SAS user isn't sure what procedure is the best choice? What if they are looking for information in a more general sense, perhaps qualitative analysis?

Papers focused on qualitative analysis are rare at lexjansen.com; a search for "qualitative," "qualitative analysis," and "text analysis" yielded fewer than 50 papers total. A search for the terms "categorical" and "categorical analysis" yielded fewer than 100 papers.

To help fill the gap, this session is intended for new SAS users. Information presented is intended to serve as a primer for qualitative analysis with SAS 9.4 or Enterprise Guide. Takeaways include an overview of the types of qualitative analysis that can be performed, useful procs, and a set of helpful hints on qualitative analysis with SAS.

INTRODUCTION

As a SAS coder who primarily has access to Base SAS and Enterprise Guide, I wanted to focus on introducing qualitative analysis in this context.

Qualitative analysis is the investigation of non-numerical data. Qualitative analysis can still yield numeric (quantitative) results but the underlying data is textual. In addition to items with preset answer choices, surveys often provide the opportunity for responders to provide text answers to open-ended questions.

The test data used for the analysis was collected from a survey of higher education professionals. The professionals were asked about their education and career paths. Some of the survey items included:

- Was this your first career choice? (yes / no)
- What level(s) of education have you completed? (multiple selections permitted)
- What was your original major? (open-ended)

DESCRIPTIVE STATISTICS

Descriptive statistics do not examine the underlying structure of the data; rather they present an overview that helps the researcher better understand the data in general. Frequency and percent analysis are two commonly used descriptive analyses. PROC FREQ offers a simple approach to analyzing the question “was this your first career choice?”:

```
PROC FREQ DATA=prac01;
TABLES Initial_Career;
RUN;
```

As Figure 1 shows, less than 20% of respondents started in this career. Adding the NOCUM option in the TABLES statement would have hidden the cumulative frequency and cumulative percent columns in the output.

Initial_Career	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	43	84.31	43	84.31
Yes	8	15.69	51	100.00

Figure 1. Output from analysis of first question.

KEYWORD SEARCHES

The final data for the second question was more complex than for the first question. The second question, “what level(s) of education have you completed?” offered a number of preset answers and an open-ended “Other” option. Some respondents only selected their highest level of educational attainment; others selected all their degree types; a few respondents used the “Other” option to include explanatory text. The preset answers created a set of known keywords that could be used to explore the data.

A useful analytical option for the second question is the FIND function, which utilizes the syntax FIND(string, ”substring”, modifier, start position). FIND searches the string for the first occurrence of the substring text and returns a value of 0 (zero) if the substring text is not present in the string; otherwise, the location number is returned (may be equal to 0). The following code demonstrates the use of the FIND function to analyze question two:

```
DATA prac02;
  SET prac01;
Grad=FIND(degrees,'Master','i');
    IF grad NE 0 THEN Master='Yes'; ELSE Master='No';
Grad2=FIND(degrees,'Doctoral','i');
    IF grad2 NE 0 THEN Doctorate='Yes'; ELSE Doctorate='No';
```

Since the FIND function can return a wide range of numbers, two reporting variables, Master and Doctorate, were created. In addition to discrete analysis of the variables Master and Doctorate, PROC FREQ can be used to produce a crosstabulation of the two reporting variables (Figure 2 displays the output):

```
PROC FREQ DATA=prac02;
TABLES Master Doctorate / NOCUM;
TABLES Master*Doctorate / NOROW NOCOL NOPERCENT; /* crosstabulation */
RUN;
```

Master	Frequency	Percent
No	14	27.45
Yes	37	72.55

Doctorate	Frequency	Percent
No	31	60.78
Yes	20	39.22

Frequency

Table of Master by Doctorate

Master	Doctorate		
	No	Yes	Total
No	2	12	14
Yes	29	8	37
Total	31	20	51

Figure 2. Output from analysis of second question.

The FIND function is also useful for analyzing the third question, “what was your original major?” In this analysis, I was interested specifically in education-oriented majors. Since the answer values were not preset, I removed all capitalization with the LOWCASE function prior to using FIND. I also included a filtering WHERE statement in the PROC FREQ analysis to eliminate any “No” responses (partial output, Figure 3):

```
Concentration=LOWCASE(concentration);
Major=FIND(concentration,'education','i');
    IF major NE 0 THEN Education='Yes'; ELSE Education='No';


PROC FREQ DATA=prac02 ORDER=FREQ;
WHERE Education EQ 'Yes';
TABLES Education / NOPERCENT NOCUM
TABLES Concentration / NOCUM NOPERCENT OUT=FreqCount;
RUN;
```

Education	Frequency
Yes	17

Figure 3. Partial output from analysis of third question.

To learn more about these professionals, I sorted the output in descending frequency (ORDER=FREQ) and created an output data set (OUT=FreqCount). I then used PROC PRINT to create a list of the education-related programs pursued by the respondents (Figure 4):

```
PROC PRINT DATA=FreqCount NOOBS;
VAR Concentration;
RUN;
```

concentration
higher education
higher education administration
educational leadership
adult and community college education
adult education
education
education leadership
education policy
educational leadership and administration
educational research and policy analysis
educational research, measurement, and evaluation
higher education, administration

Figure 4. Program concentration of seventeen survey respondents.

The LIKE condition provides another method for exploring the data; this condition is particularly useful with open-ended questions where respondents may either misspell or truncate terms (e.g., “educ” rather than “education”). The LIKE condition can be placed in a WHERE statement to create a subset of the data based on matches with the specified pattern. The LIKE condition is case-sensitive, so consider the use of the UPCASE or LOWCASE function to avoid lost matches. The LIKE condition may be used in a DATA step or with PROC SQL:

```
DATA prac02;
    SET prac01;
WHERE LOWCASE(Concentration) LIKE “education’;

          -or-

PROC SQL;
SELECT Concentration
FROM prac01
WHERE LOWCASE(Concentration) LIKE “education”;
QUIT;
```

The underscore (_) and percent (%) wildcards can be used to expand the number of pattern matches. An underscore matches any single character (two underscores would match any two characters in a row) while the percent matches any number of characters. Underscore and percent wildcards may be placed at the beginning, middle, or end of the pattern. If the percent sign is placed at the start of the patter, enclose the pattern with single quotes to avoid “WARNING: Apparent invocation of macro” in the log. Table 1 demonstrates the use of the wildcard characters with the LIKE condition to search the word list: analysis, analytics, reanalyze, predictive analytics, psychoanalyst, analyte.

Matching Pattern	Will Match These Terms
LIKE “analysis”	analysis
LIKE “anal%”	analysis, analytics, analyte
LIKE “anal___” (3_)	analyte
LIKE ‘%anal%’	analysis, analytics, reanalyze, predictive analytics, psychoanalyst, analyte

Table 1. Examples of Pattern Matching with Wildcards in LIKE

CHI SQUARE

Chi square analysis is one way to examine potential relationships among categorical variables. Due to the small number of observations (*N*=51), the test data was not well-suited for chi square analysis. Chi square analysis is an option in PROC FREQ; the following code would perform a chi square analysis to determine if there is a relationship between initial career (yes / no) and education major (yes / no):

```
PROC FREQ DATA=prac02;
TABLES Initial_Career * Education / NOCUM;
RUN;
```

For a deeper exploration of chi-square with sample analyses, refer to Waller (2015) and Smith (2021).

TEXT ANALYSIS

Descriptive qualitative analysis and hypothesis testing are available on any SAS platform; however, text analysis is generally more difficult without Enterprise Miner or Visual Text Analytics. Text analysis can be used to uncover sentiment (negative, neutral. positive) and identify keywords or themes.

IDENTIFY KEYWORDS

A word cloud is frequently used to identify keywords (words that appear most frequently in text). Parker (2019) described a method using SAS in combination with Python to create a word cloud. Although I was not able to create a true word cloud using Base SAS alone, I was able to create an approximation. An eight step process was used to identify and visualize the primary keywords in the respondents’ answers to the third question (original majors).

1. Use LOWCASE to eliminate all capital letters in the text.
2. Use TRANWRD function to remove punctuation and “stop words” from the text. (“stop words” are common words such as a / an / and / the)
3. Use the STRIP function to remove leading blanks in the final text.
4. Use SCAN function to segment the text (each word in a new variable).
5. Use PROC TRANSPOSE to flip the data set and create a new variables of individual words.
6. Use PROC FREQ to analyze data, sort in descending order, and produce an output data set of frequency (count) and percentage information.
7. Use CEIL function to round percentage information to the nearest integer greater than the percentage.
8. Use PROC REPORT to visualize the results; output font size based on frequency count.

Steps one to four can be completed one DATA step; I chose to use two DATA steps for clarity. Note that when using the TRANWRD function, I included blank spaces before and after the word pattern so as not to remove letters within words (e.g., removing “and” from “standard”). The syntax for TRANWORD Is (source, target, replacement). The required syntax for SCAN is (string, count); an optional argument can be used to identify non-default delimiters. The general code for steps one to four follows:

```
DATA prac03;
  SET prac02;
WHERE Concentration NE " ";
LENGTH Major $100.
  Major=LOWCASE(Concentration);   /* removes all capitalization */
  Major=TRANWRD(Major, ", ", " "); /* replaces commas with blank space */
  Major=TRANWRD(Major, " the ", " "); /* replaces "the" with blank space */
  Major=STRIP(Major); /* removes leading blanks */
RUN;

DATA prac04;
  SET prac03;
  P1 = SCAN(Major,1); /* first word extracted to new variable P1 */
  P2 = SCAN(Major,2); /* second word extracted to new variable P2 */
  P3 = SCAN(Major,3);
  P4 = SCAN(Major,4);
  P5 = SCAN(Major,5);
  P6 = SCAN(Major,6);
RUN;
```

At this point, the data set is structured Response Concentration Major P1 P2 P3 P4 P5 P6. To identify the most frequent word choices, P1 – P6 needed to be stacked vertically and identified with the same variable name (e.g., "Word"). The data is then ready for analysis with PROC FREQ:

```
PROC TRANSPOSE DATA=prac04 OUT=Prac05  /* output data = prac05 */
  (RENAME=(col1=Word)); /* new variable renamed "Word" */;
  VAR P1 - P6; /* variables being transposed (stacked) */
RUN;


PROC FREQ DATA=prac05 ORDER=FREQ;
  WHERE Word NE " "; /* filters out empty variables */
  TABLES Word / NOCUM OUT=Counts; /* output data set of frequencies */
RUN;
```

The data set Counts has three variables: Word Count Percent. The CEIL function was used to round the percentage data for later use (output data set displayed in Figure 5):

```
DATA prac06;
  SET Counts;
WHERE Count GE 4; /* restricts to most frequent word choices */
Size = CEIL(Percent); /* rounding */
RUN;
```

VIEWTABLE: Work.Prac06

	Word	COUNT	PERCENT	Size
1	education	12	9.4488188976	10
2	administration	9	7.0866141732	8
3	higher	8	6.2992125984	7
4	leadership	6	4.7244094488	5
5	business	5	3.937007874	4
6	educational	5	3.937007874	4
7	psychology	5	3.937007874	4
8	science	4	3.1496062992	4

Figure 5. Most frequent keywords describing original major.

In a traditional word cloud, the most frequent keyword is displayed in the largest sized font; color or font style may also be used as distinguishing characteristics. To create a similar effect, the style characteristics of the values of Word were related to the Size variable:

```
PROC REPORT DATA=prac06;
COLUMN Size Word;  /* order of columns in table */
DEFINE Size / CENTER DISPLAY WIDTH=20; /* characteristics, Size column */
COMPUTE Word; /* characteristics depend on Size variable */
IF Size EQ 10 THEN
    CALL DEFINE(_col_,"style","style=[background=yellow fontsize=25
    fontstyle=italic fontweight=bold");
IF Size EQ 9 THEN
    CALL DEFINE(_col_,"style","style=[background=lime fontsize=22
    fontweight=bold");
IF Size EQ 8 THEN
    CALL DEFINE(_col_,"style","style=[background=cyan fontsize=18
    fontstyle=italic fontweight=bold");
IF Size EQ 7 THEN
    CALL DEFINE(_col_,"style","style=[background=vpab fontsize=15
    fontweight=bold");
IF Size EQ 6 THEN
    CALL DEFINE(_col_,"style","style=[background=vlipb fontsize=12
    fontstyle=italic fontweight=bold");
IF Size EQ 5 THEN
    CALL DEFINE(_col_,"style","style=[background=bio fontsize=10
    fontweight=bold");
IF Size LT 5 THEN
    CALL DEFINE(_col_,"style","style=[background=ligrpr fontsize=8
    fontstyle=italic fontweight=bold");
RUN;
```

In Figure 5, the first record has Word = education and Size = 10. In the report, “education” will be in bold, italic print, with the largest font size against a yellow background. The last record has Word = science and Size = 4. In the report, “science” will be in bold, italic print, but with the smallest font size against a mauve background. Figure 6 displays the final output.

Size	Word
10	education
8	administration
7	higher
5	leadership
4	business
4	educational
4	psychology
4	science

Figure 6. Base SAS "word cloud" of keywords related to respondents' original major.

CONCLUSION

A primer is an introduction to a subject. In this paper, several techniques of qualitative analysis were described and applied to test data with Base SAS: descriptive statistics, chi square, keyword identification and keyword visualization (aka "word cloud"). Additional techniques are included as recommended readings.

RECOMMENDED READING

- Hampton, J. (SAS Global Forum, 2012). *Dickens vs. Hemingway: Text analysis and readability statistics in Base SAS®*. Available at lexjansen.com
- Parker, C. (2019, October 16). *Creating a word cloud using Python and SAS® software*. https://blogs.sas.com/content/sgf/2019/10/16/word-cloud-python/
- Smith, K. D. (SESUG, 2021). *Moving beyond frequency and percentage to chi square, t-tests, and correlation analysis*. Available at lexjansen.com
- Suraweera, W., Weerasooriya, J., & Fernando, N. (SAS Global Forum, 2018). *A simple approach to text analysis using SAS functions*. Available at lexjansen.com
- Waller, J. (WUSS, 2015). *Chi-square and t-tests using SAS®: Performance and interpretation*. Available at lexjansen.com

CONTACT INFORMATION

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

Kelly D. Smith, Ed.D.
AEWAS Consulting
kds.aewas@gmail.com
www.aewasconsulting.com