EXST 700x

Lab #3: Frequencies and Chart Graphics

Tip: Read previous tips.

Objectives

1. Use a LABEL statement to clarify information about a variable.
2. Use a LENGTH statement.
3. Use a FREQ procedure (frequency) to produce a frequency table.
4. Use a CHART procedure to
    a) produce a horizontal bar chart with frequencies.
    b) produce a vertical bar chart (histogram).
    c) produce other descriptive charts (e.g. BLOCK and PIE charts).

More details on basic SAS Statements and Procedures

The "**results viewer**" window is the default HTML window for SAS output. However, output accumulates in this window with each successive run of the program with the most recent output on the bottom. The following statements will result in clearing the results window each time these statements are run.

```
ods html close; ods html;
```

Obviously, this program should go near the top; I suggest right after the

```
dm 'log;clear;output;clear';
```

code, assuming you include that. That code will clear the log and the list each time the program is executed from the top. For this week's assignment I suggest you also include,

```
ods graphics on;
```

as this may produce some prettier pictures. I would still include the statement to get output

```
ODS listing;
```

it would probably be wise to put all four of these as the first four lines in your program.

Assignment 3 example

The **example** program for today has is much more complicated (16 variables) than the **assignment** dataset (only 5 variables), but I needed a large dataset to find a range of values that would produce graphics similar to those in the assignment. The dataset is the "Healthy Breakfast" dataset from DASL (The Data and Story Library; http://lib.stat.cmu.edu/DASL/).

Since the example dataset is more complex, I had to use some statements that you will not have to use. For example, by default the length of a character variable is equal to the length of the first value that is read for that variable. If you have a variable GENDER and the first value is "male" then when the value "female" occurs it will read "fema". If the first occurrence is "female", then there would be no problem reading "male". By default, the

length of a variable should not exceed 8 characters. However, longer variables can be read if the length is specified in a "LENGTH" statement. In my example dataset some of the cereal names (variable NAME) are very large. I specified the following statement after the DATA statement to get 24 characters, and some names exceeded that length (the maximum length is 32 characters).

```
LENGTH name $ 24;
```

By the way, SAS variable names must begin with a letter or an underscore. They can include numbers, but not blanks or special characters.

Since I had so many variables, I also included a large LABEL statement. Notice that everything, from the word LABEL to the semicolon 14 lines later, is a single statement. The variable specifying one of the 8 manufacturers was represented with a single character. I included a definition of these codes as a COMMENT prior to the DATA step. You will need a LABEL statement, but not as large as mine.

New procedures

PROC FREQ: The FREQ procedure is used to produce frequency tables with percentages. Eventually we will also use this procedure to do Chi Square Tests of Independence and Chi Square Tests of Goodness of Fit.

Other statements can be used together with the PROC FREQ statement such as:

1) TABLE statement: – controls the tables to be printed. When a single variable is listed in the table statement the procedure will output a table with the frequency, percent (relative frequency), cumulative frequency and cumulative percent (relative cumulative frequency). The variable in the TABLE statement can be either quantitative or qualitative. The following statements will produce this table for the cereal calorie variable.

```
proc freq data=Cereal;
   Title3 'Frequencies of calories';
   table calories;
run;
```

When more than one variable is listed in a PROC FREQ TABLE statement the procedure produces a two-way table, or series of two-way tables for the variables. The following will produce a two-way table of the shelf where the cereal is presented in the supermarket (shelf: 1=bottom, 2=middle and 3=top) and the list of the 8 cereal manufacturers (mfr). In addition to the frequency the table will have the overall table percent, row percent and column percent for each cell of the table and the marginal frequencies (row and column) with percentages.

```
proc freq data=Cereal;
   Title3 'Frequencies of shelf & manufacturer categories';
   table shelf * mfr;
run;
```

2) BY statement: – can be used as with most other procedures.

3) And of course the ever popular TITLE statements can be included.

PROC CHART: PROC CHART is one of a number of graphical procedures often used for data exploration and examination. This procedure can be used to produce a number of different styles of graphic depending on the statements that are included. The variable to be processed is named in the statement. Some of these statements are

HBAR – a horizontal bar chart that will also include information on frequency, percent (relative frequency), cumulative frequency and cumulative percent (relative cumulative frequency)

```
proc chart data=Cereal;
   hbar calories;
run;
```

VBAR – a vertical bar chart often called a histogram

```
proc chart data=Cereal;
   vbar sugars / midpoints=0 to 14 by 2;
run;
```

BLOCK – produces a 3D plot with two variables (sugars and shelf) on a surface and blocks who's height represent a third "response" variable. The default for the response is frequency of occurrence in each combination of the first two variables. The response variable can also be percents, sums or means.

```
proc chart data=Cereal;
   block sugars / discrete group=shelf midpoints=0 to 15 by 5;
run;
```

PIE, STAR, DONUT – yields pie chart and similar charts

```
proc chart data=Cereal;
   pie mfr;
run;
```

PROC CHART OPTIONS: A number of options are available to modify the appearance of charts. We will not discuss size and resolution options here, but some other important options are listed below. The options below are placed on the chart type statement following a slash (i.e. /).

MIDPOINTS = *midpoint_list* – as discussed in class, when a quantitative variable is turned into a frequency table it is often necessary to divide the numbers into size categories. The same is true of bar charts. By default SAS will determine groupings, or midpoints for groupings. However, you can set your own midpoints with the MIDPOINT option; see examples below.

```
vbar sugars / midpoints= 0 to 14 by 2;  * by range and interval;
vbar sugars / midpoints= 3 6 9 12 15;   * by specific values;
vbar sugars / midpoints= 2 4 8 16;      * by unequal spacing;
```

DISCRETE – indicates that the quantitative values are to be treated as discrete categories and not as a quantity so midpoints will not be calculated.

GROUP = *variable_name* – This option produces a groups of bars with the HBAR and VBAR statements. It specifies the second axis in the BLOCK statement.

Assignment 3

The dataset (Table 1.1 from Freund, Wilson & Mohr): The data is responses of 50 respondents on their level of happiness (Likert scale: 1=Not too happy, 2=Pretty happy, 3=Very happy). Additional information includes the age and sex of the respondent and the average number of hours of TV watched daily. The complete dataset is available on the Lab web page.

respondent	age	sex	happy	tvhours
1	41	1	2	0
2	25	2	1	0
3	43	1	2	4
4	38	1	2	2
5	53	2	3	2
6	43	2	2	5
7	56	2	2	2
8	53	1	2	2
9	31	2	1	0
10	69	1	3	3
. . .				
45	74	2	2	3
46	37	2	3	0
47	48	1	2	3
48	42	2	2	6
49	77	2	2	2
50	75	1	3	0

Suppose you are going to examine this data for happiness. Answer any questions posed using SAS. Turn in your program log and, for each question, please turn in the relevant SAS output. Try to organize your responses for clarity. (1 point)

1) You will want to include in your program the “usual statements” with option, comments and titles similar to those in ASSIGNMENT 01 and 02. (2 points)

Include appropriate title statements. (1 point)

Create a data step to enter the data set above. The data step will include an input statement and a data statement. To these statements add the following LABEL statement: (1 point)

```
label happy = 'Happyness index: 3=happiest'
      sex   = '1 = Male';
```

2. Print data for all the respondents. (1 point)

3. Do a frequency procedure creating a table for sex * tvhours. (1 point)

4. Prepare a horizontal bar chart for the variable tvhours. (1 point)

5. Create a vertical bar chart of the variable age. (1 point)

6. Redo the vertical bar chart of the variable age grouped from 20 to 80 by 10. (1 point)

7. Create a BLOCK chart of the variable HAPPY and the group SEX. Be sure to specify the option DISCRETE so the procedure does not try to calculate midpoints for the variable HAPPY. (1 point)

8. Do the pie chart for the variable age, also specifying discrete. (1 point)