Chapter 7

Data Visualization

What you will learn in this chapter

- ✓ univariate analysis for continuous and categorical variables
- ✓ how to produce and interpret bar charts and plots by groups
- ✓ how to interpret histograms, QQ Plots, box plots, and stem-leaf plots
- ✓ how to produce two way frequency tables
- ✓ how to create a scale and interpret Cronbach's Alpha
- ✓ how and why to reverse code a variable

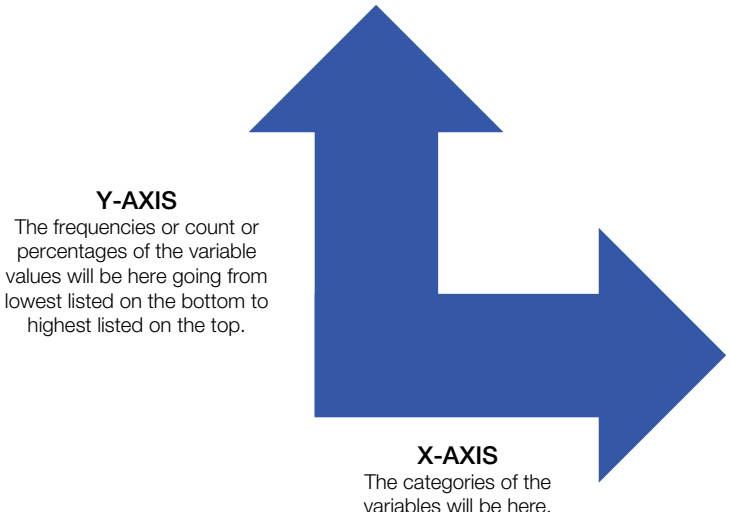
Univariate Analysis

When conducting research, one of the first steps that a researcher takes is to develop a basic understanding of a sample of a data to explain what the data indicate. Applied social scientists have to provide very basic answers such as, how many, or what percentage of the people who took the survey were of a particular race or were employed or were married. Answering these types of descriptions of the data using one variable at a time is **univariate analysis**. The methods of conducting univariate analysis vary for categorical and continuous variables. Researchers should examine both types of variables visually and produce frequencies in count or percentages. Univariate analyses include measures of central tendency, as appropriate, for categorical and continuous variables.

Categorical Variables

One of the best visual depictions of categorical or qualitative variables is the bar chart. Bar charts use rectangles, or bar shapes, to represent the values of a variable on a number line or a continuum. **Bar charts** have spaces between each rectangle or bar shape so as not to lead one

to believe that the values that they represent are continuous. The categories of the variables are along the X-axis and the frequencies are listed on the Y-axis.



View and Interpret Data

Bar Charts

Again, bar charts are used with less than interval level data. Bar charts graphically depict the frequency or count or percentage of values on a given variable. Bar charts help researchers determine the structure the values of the variable have in the data set. You create bar charts using PROC GCHART procedure. The DISCRETE option generates whole values when making categories, not midpoints. **Midpoints** are the middle value of a category and is the default GCHARTS selection when making categories. The following PROC GCHART of a Categorical Variable syntax generates a chart of the variable RANK entitled "Chart of Year in College". Notice the subcommand is VBAR, not VAR

```
PROC GCHART;
VBAR RANK / DISCRETE;
TITLE 'CHART OF YEAR IN COLLEGE';
RUN;
```

Program with Guided Interpretation of PROC GCHART

Program 7.1 PROC GCHART of a Categorical Variable creates a bar chart for the variable RANK. Building on the data set created at the end of Chapter 6 Get to know the Data, write the following program in the Program Editor and run the program. Remember that the variables in the INPUT statement must be in the exact order provided.

PROGRAM 7.1 PROC GCHART of a Categorical Variable

```
DATA ONE; ***THIS LINE CREATES THE TEMPORARY DATASET NAMED
ONE;
INFILE DATALINES N=400; *REQUIRED TO MAKE ADJUSTMENTS TO
NUMBER OF CHARACTERS THAT SAS WILL READ ON A LINE;
INPUT VERSION $ CASEID $    AGE   DOB : MMDDYY8. GENDER $
     RANK $   CREDIT_HRS COLLEGE $ TOPS  PARENTFI    JOB
     SCHOLARSHIP GRANT   OTHER1 OTHER_FIN  ADVOCATE  RADIO TV
     INTERNET REVEILLE   PARENTS   FRIENDS  NONEWS
     OTHER2   OTHER_NEWS INET_ACCESS  SKIP_CLASS     EAT_OUT
     HRS_WORK BIRTH_ST $ BIRTH_CO $   IM    TAX       NO_DATES
     LIFE     MED_CABINET    RECYCLE_GIFT CALL_IN_SICK;
LABEL
   AGE = 'YEARS OF AGE'
   CREDIT_HRS = 'TOTAL CREDIT HOURS'
HRS_WORK = 'HOURS WORKED PER WEEK'
EAT_OUT = 'TIMES WENT OUT TO EAT'
RANK = 'YEAR IN COLLEGE'
;
*BELOW WE RECODE ORIGINAL VALUES TO SETTING INVALID DATA
VALUES TO MISSING;
IF CREDIT_HRS=-99 THEN CREDIT_HRS=.;
FORMAT DOB MMDDYY8.;
DATALINES;

[Copy data from Program 5.2 and paste it in this area-then
delete this note]

B  56   21   12/17/83    MALE  3     65    2     1     1
   1    0    0     0     -97   0     0     1     0     1
   1    1    0     0     -97   4     3     4     25    LA
   USA  0    1     1     5     1     0     0
B  57   19   10/23/85    FEMALE      3     65    2     1
   1    1    1     1     1     1     0     1     0     1
   1    1    0     0     0     -97   4     2     6     20
   LA   USA  1     0     20    5     1     0     1
B  58   21   11/12/83    FEMALE      4     106   7     0
   1    1    1     1     1     1     0     0     1     1
   1    1    1     0     0     -97   4     2     2     15
   LA   USA  1     1     3     4     1     0     1
B  59   19   11/08/85    FEMALE      2     33    5     1
   0    1    0     0     0     -97   0     1     1     1
   1    1    1     0     1     1     4     2     3     30
   CO   USA  1     0     3     3.5   0     0     1
B  60   19   11/19/85    MALE  2     30    5     0     1
   0    1    0     1     2     0     0     1     1     1
   0    1    0     0     -97   3     2     2     15    LA
   USA  0    0     3     3     1     0     0
```

```
B  61   20   03/15/84    MALE   3      63     2      0      1
   0    0    0      0     -97    0      1      1      1      0
   0    1    0      0     -97    3      2      0      0      WI
   USA  1    1      6     5      0      1      1
B  62   19   07/08/85    FEMALE        2      33     2      0
   0    1    1      0     0      -97    0      0      1      0
   1    1    1      0     0      -97    4      4      3      0
   AR   USA  1      1     1      5      0      0      0
B  63   21   12/23/83    MALE   4      96     2      1      1
   0    0    0      1     1      1      1      1      0      1
   0    1    0      0     -97    3      4      15     20     LA
   USA  0    1      6     5      1      0      1
B  64   20   10/19/84    FEMALE        2      57     5      1
   0    0    0      0     1      1      0      0      0      1
   0    0    0      0     0      -97    4      2      5      35
   LA   USA  1      0     2      4      0      0      1
B  65   18   11/13/86    FEMALE        1      29     5      1
   1    0    0      1     1      1      0      1      1      0
   1    1    1      0     0      -97    4      2      2      0
   LA   USA  1      1     1      5      0      1      1
B  66   21   04/06/83    FEMALE        3      75     2      1
   1    0    0      0     0      -97    0      0      1      0
   0    0    0      0     0      -97    3      3      2      10
   LA   USA  1      0     1      3      0      -99    0
B  67   20   06/08/84    FEMALE        3      68     5      1
   0    0    0      0     1      2      0      0      1      1
   1    1    1      0     0      -97    4      3      2      20
   LA   USA  1      1     2      5      0      0      1
B  68   21   01/07/83    MALE   3      63     2      0      1
   0    0    0      1     2      1      0      1      1      1
   0    0    0      0     -97    4      2      10     0      LA
   USA  0    0      0     3      1      0      1
B  69   20   04/16/84    FEMALE        3      61     5      0
   0    1    0      0     1      1      0      0      0      1
   1    0    1      0     0      -97    4      3      1      20
   LA   USA  1      1     10     5      1      1      0
B  70   21   12/07/83    MALE   4      60     2      1      1
   0    0    0      0     -97    1      0      1      1      1
   0    0    0      0     -97    4      3      5      20     LA
   USA  0    0      5     5      1      1      1
B  71   20   11/30/84    FEMALE        3      62     2      1
   1    0    0      0     1      1      0      1      1      1
   1    1    1      0     0      -97    4      5      5      0
   TX   USA  1      1     1      4      0      0      1
B  72   20   02/07/84    MALE   3      63     2      1      1
   0    0    1      1     1      0      0      1      0      0
   0    0    0      0     -97    4      2      3      15     LA
   USA  1    1      4     4      0      0      1
B  73   22   05/17/82    MALE   3      67     5      0      0
   1    0    0      0     1      0      0      1      1      0
```

```
   0    0    0     0     -97   4     3     3     22    LA
   USA  1    0     1     3     1     0     0
B  74   19   03/15/85    FEMALE      2     36    2     1
   1    0    0     0     0     -97   0     0     1     1
   0    0    0     0     0     -97   4     2     7     0
   LA   USA  0     1     0     4     0     0     1
B  75   20   03/08/84    FEMALE      3     57    2     1
   1    0    0     0     0     -97   0     0     1     0
   1    0    0     0     0     -97   3     3     3     0
   TX   USA  0     1     -99   4     0     0     1
B  76   19   11/02/85    FEMALE      2     31    -99   1
   0    0    0     0     0     -97   0     1     1     1
   1    1    1     0     0     -97   4     3     2     0
   LA   USA  1     0     3     4     0     1     1
B  77   19   04/20/85    MALE  2     37    2     0     1
   0    0    0     0     -97   0     0     1     1     1
   0    0    0     0     -97   3     3     5     0     LA
   USA  0    1     5     5     0     0     1
B  78   19   09/13/85    FEMALE      2     33    2     1
   0    0    1     1     0     -97   0     0     1     1
   1    1    1     0     0     -97   4     2     3     18
   -98  VNM  1     0     1     4     0     0     1
B  79   19   02/07/85    FEMALE      2     30    2     1
   0    0    0     0     0     -97   0     0     0     0
   1    0    0     0     0     -97   4     2     2     5
   LA   USA  0     0     3     3     0     1     1
B  80   21   05/01/83    MALE  3     59    5     1     1
   1    0    0     0     -97   0     0     1     0     1
   1    1    0     0     -97   4     3     3     4     LA
   USA  1    1     2     3     0     0     1
B  81   19   09/08/85    MALE  2     39    2     1     0
   0    0    0     0     -97   1     0     1     1     1
   1    1    0     0     -97   4     3     11    0     LA
   USA  1    0     0     4     0     0     1
A  82   21   04/14/83    MALE  4     94    5     0     1
   1    0    0     1     1     0     0     1     1     0
   0    0    0     0     -97   4     1     5     0     FL
   USA  0    0     2     3     0     1     0
A  83   20   12/11/84    MALE  3     60    2     0     0
   0    1    0     0     -97   1     0     1     1     1
   0    0    0     0     -97   4     3     2     25    TX
   USA  0    0     5     5     0     1     0
A  84   19   10/14/85    FEMALE      2     33    2     1
   1    0    0     0     0     -97   0     1     1     1
   0    1    1     0     0     -97   2     3     7     0
   LA   USA  1     1     0     4     0     0     0
A  85   21   12/08/83    MALE  4     75    2     1     0
   0    0    0     0     -97   0     0     1     0     0
   0    0    0     0     -97   3     2     5     30    LA
   USA  0    0     1     4     0     0     1
```

```
A  86   20   10/12/84    MALE   3      70     2      0      1
   0    0    0      0     -97    0      0      1      0      0
   0    0    0      0     -97    4      4      8      0      LA
   USA  1    0      5     4      1      0      0
A  87   21   12/15/83    MALE   3      57     2      1      1
   1    1    0      0     -97    0      1      1      1      1
   1    1    0      0     -97    4      3      4      20     LA
   USA  0    1      20    6      1      0      0
A  88   19   11/07/85    FEMALE        3      63     2      0
   1    0    1      0     0      -97    0      0      1      0
   1    0    0      0     0      -97    4      3      4      0
   TX   USA  0      1     4      5      1      0      1
A  89   20   11/12/84    MALE   3      61     2      1      1
   0    0    0      0     -97    0      0      1      0      0
   0    0    0      0     -97    4      3      3      0      LA
   USA  0    0      1     4      1      0      1
A  90   21   06/27/83    FEMALE        4      85     2      0
   0    0    1      0     0      -97    0      0      1      1
   1    0    0      0     0      -97    4      3      4      0
   LA   USA  1      1     3      5      1      0      0
A  91   19   08/19/85    MALE   2      37     2      1      0
   1    1    0      1     1      0      1      1      0      1
   0    1    0      0     -97    4      3      3      0      LA
   USA  0    1      4     3      0      0      1
A  92   20   02/10/84    FEMALE        3      69     5      1
   0    1    0      0     0      -97    0      1      1      1
   1    1    1      0     0      -97    4      3      1      30
   LA   USA  1      0     4      4      1      1      1
A  93   19   11/28/85    FEMALE        4      98     2      0
   0    1    0      0     1      4      0      1      1      0
   1    1    1      0     0      -97    3      2      2      30
   MO   USA  0      0     1      4      0      0      0
A  94   19   08/09/85    FEMALE        2      34     5      1
   1    0    0      1     1      1      0      0      1      1
   1    1    1      0     0      -97    4      3      0      0
   LA   USA  1      1     5      5      1      1      0
A  95   19   10/14/85    MALE   2      -99    2      1      0
   0    1    0      1     -97    1      0      1      1      0
   1    1    0      0     -97    4      4      5      4      LA
   USA  0    1      4     3      1      0      1
A  96   20   03/13/84    FEMALE        3      68     2      1
   1    0    0      1     0      -97    1      0      1      0
   1    1    1      0     0      -97    4      3      4      15
   MS   USA  0      0     4      5      0      0      1
A  97   19   04/09/85    FEMALE        2      39     2      0
   0    0    1      0     1      -97    0      0      0      0
   0    0    0      1     0      -97    4      3      4      0
   TX   USA  1      0     4      5      1      1      0
A  98   19   07/19/85    MALE   3      80     2      1      0
   0    0    0      0     -97    0      1      0      0      0
```

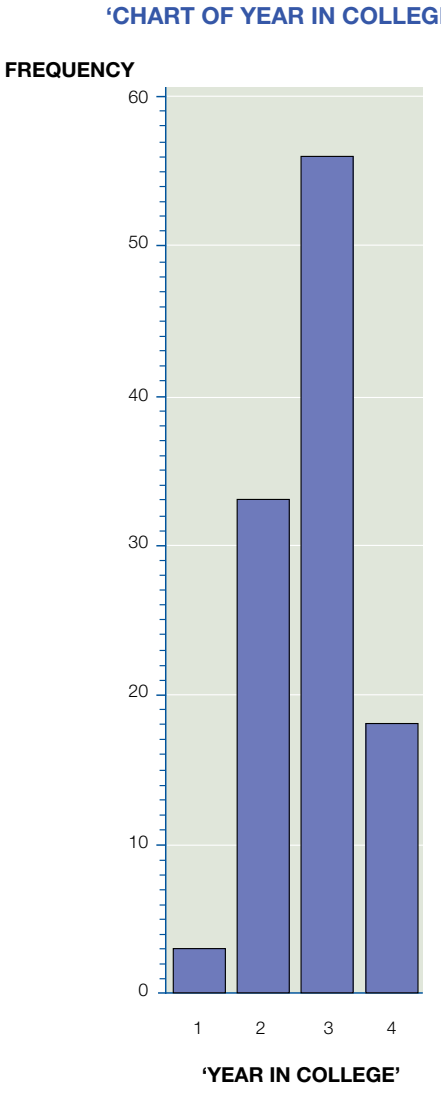
```
   0    0    1     0     -97   4     4     5     15    LA
   USA  1    0     4     4     1     0     0
A  99   20   04/23/84    FEMALE      3     63    2     0
   1    0    0     0     0     -97   0     1     1     1
   1    0    0     0     0     -97   4     2     2     0
   LA   USA  1     0     1     3.5   0     0     0
A  100  20   07/01/84    FEMALE      3     60    2
   1    0    0     0     0     1     -97   0     1     1
   0    0    0     1     0     0     -97   4     2     3
   20   LA   USA   0     1     4     5     0     0     1
A  101  20   06/17/84    FEMALE      3     60    2
   1    0    0     1     1     1     -97   0     1     1
   1    1    1     1     0     0     -97   4     2     2
   20   LA   USA   1     -99   1     4     1     0     1
A  102  21   08/17/83    FEMALE      3     -99   2
   1    1    1     1     1     1     -97   0     1     1
   0    1    1     1     0     0     -97   3     2     3
   12   LA   USA   -1    0     2     5     0     0     1
A  103  20   08/23/84    FEMALE      3     60    7
   0    1    0     0     0     0     -97   0     1     1
   0    1    0     1     0     0     -97   4     3     6
   0    LA   USA   1     1     10    5     1     0     1
A  104  19   04/17/85    FEMALE      2     41    2
   1    1    0     0     0     1     5     0     0     1
   0    1    1     1     0     0     -97   4     2     2
   0    LA   USA   1     1     4     4     0     1     1
A  105  19   02/09/85    FEMALE      2     44    2
   0    1    0     0     0     1     -97   0     1     1
   1    1    1     1     0     0     -97   4     2     3
   0    LA   USA   1     0     2     4     1     1     1
A  106  19   06/16/85    MALE  2     24    2     1
   0    0    0     1     1     -97   0     1     1     1
   1    0    1     0     0     -97   4     1     2     20
   LA   USA  1     1     8     3     0     0     0
A  107  21   04/08/83    MALE  3     82    5     0
   0    0    0     1     0     -97   1     0     1     0
   1    1    1     0     0     -97   2     3     5     20
   MS   USA  0     1     1     4     0     1     0
A  108  21   04/22/83    MALE  3     72    5     1
   1    0    0     0     1     -97   0     1     0     0
   1    0    1     0     0     -97   4     1     8     0
   LA   USA  1     1     10    5     0     0     0
A  109  29   10/15/75    MALE  3     69    2     1
   0    1    0     1     1     -97   0     0     1     1
   1    1    1     0     0     -97   4     1     3     30
   TX   USA  0     0     4     4     0     0     1
A  110  21   10/18/83    FEMALE      4     90    2
   0    0    0     0     1     0     1     0     1     1
   0    0    1     0     0     0     -97   4     2     5
   20   MS   USA   0     1     1     5     1     1     1
```

```
;
PROC FREQ;
TABLES RANK;
RUN;
PROC GCHART;
VBAR RANK / DISCRETE;
TITLE 'CHART OF YEAR IN COLLEGE';
RUN;
*** PROGRAM ENDS HERE ***;
```

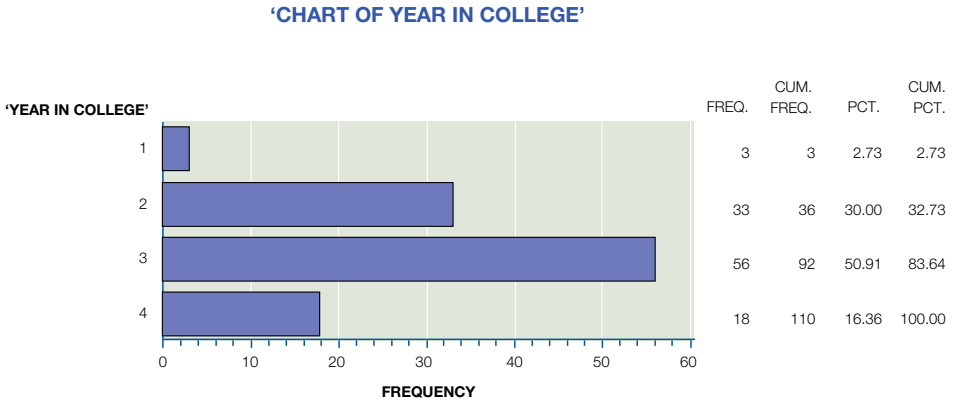

Output 7.1 Chart of Year in College was generated by Program 7.1 PROC GCHART of a Categorical Variable. From this chart you can determine that most respondents are juniors in college. You can also conclude that the categories of year in college are not equal or approximately equal; there are fewer freshmen than any other group. Researchers should speculate how the patterns revealed here may influence the study. To consider how the distribution may influence the study, you first consult prior research in the field to determine if the topic of study varies by year in school. Another, but less sophisticated approach is to brainstorm about the possible influences the distribution has on the question. Explanations based on prior literature and on theory application are always considered better than merely brainstorming and suggesting untested possibilities. Always try to utilize prior research or applications of theory to explain your position.

In Program 7.1 PROC GCHART of a Categorical Variable you could have changed the orientation of the output to horizontal instead of vertical by changing the subcommand VBAR to HBAR. Re-run Program 7.1 PROC GCHART of a Categorical Variable making this minor adjustment and examine the output. You can see the output generated by using the HBAR subcommand is somewhat different from the output generated using the VBAR subcommand. Output 7.2 PROC GCHART of Year in College using HBAR Subcommand includes frequencies, cumulative frequencies, percentages, and cumulative percentages.

OUTPUT 7.1 Output 7.1 PROC GCHART of a Categorical Variable



OUTPUT 7.2 PROC GCHART of Year in College using HBAR Subcommand



By Subgroups

To see patterns across two categorical variables at the same time use the GROUP option on the PROC CHART and PROC GCHART procedures. Add Partial Program 7.3 PROC GCHART using GROUP Option to Program 7.2 to generate Output 7.3 PROC GCHART using GROUP Option.

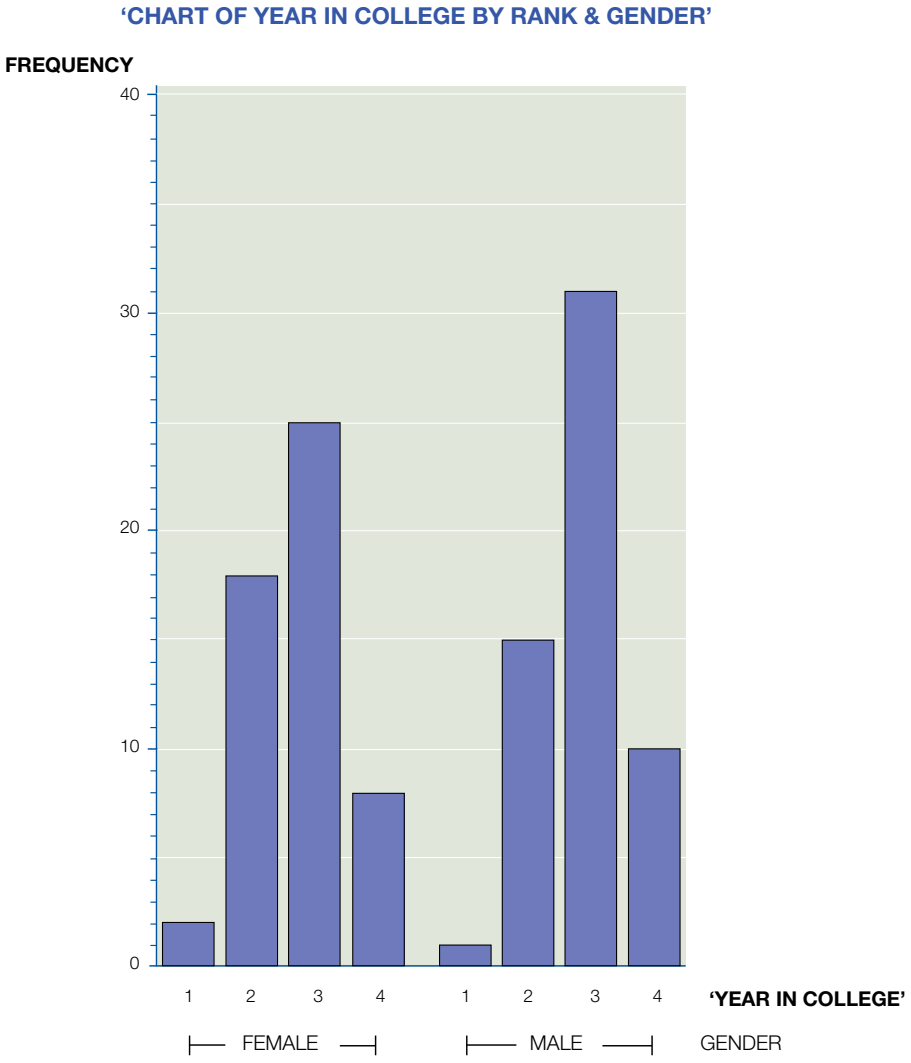
PROGRAM 7.2 PROC GCHART using GROUP Option

```
PROC GCHART;
VBAR RANK / DISCRETE GROUP=GENDER;
TITLE 'CHART OF YEAR IN COLLEGE BY RANK & GENDER';
RUN;
```

There are numerous options available for creating charts and graphs in SAS software. You need to take some time to explore SAS software options available for creating charts and graphs. Using SAS statements and options you can specify the number of bars on your charts, the color, size (width or height as appropriate), the patterns, and much more. Take time to explore this area of SAS on your own.

From Output 7.3 PROC GCHART using GROUP Option below you may be inclined to think that rank of student in the data follows a similar pattern across males and females.

OUTPUT 7.3 PROC GCHART using GROUP Option



Two-Way Frequency Tables

There are at least two procedures in SAS software to generate frequency tables. PROC TABLES procedure and PROC UNIVARIATE with FREQ option. The syntax for the **PROC TABLES** procedure is

```
PROC TABLES;
TABLES {INSERT VARIABLE NAME HERE};
RUN;
```

The PROC UNIVARIATE syntax to create frequencies is

```
PROC UNIVARIATE FREQ;
VAR {INSERT VARIABLES HERE};
RUN;
```

You may list as many variables after the TABLES and VAR subcommands as you need. The output will be one variable frequency distribution. If you want to have the frequency of one variable broken down by another variable, you use syntax similar to the following syntax while substituting the variables you are interested in for the variables in Program 7.4 Table of Categorical Variable by Categorical Variable. Add Partial Program 7.4 to Program 7.3 PROC GCHART using GROUP Option. Run Program 7.4 Table of Categorical Variable by Categorical Variable and examine the output.

PROGRAM 7.3 Table of Categorical Variable by Categorical Variable

```
PROC FREQ;
TABLES SKIP_CLASS*GENDER;
title 'Table of Categorical Variable by Categorical
Variable';
RUN;
```

OUTPUT 7.4 Table of Categorical Variable by Categorical Variable

The FREQ Procedure

Frequency Percent Row Pct Col Pct	TABLE OF SKIP_CLASS BY GENDER			
		GENDER		
	SKIP_CLASS	FEMALE	MALE	Total
	1	5	12	17
		4.55	10.91	15.45
		29.41	70.59	
		9.43	21.05	
	2	29	21	50
		26.36	19.09	45.45
		58.00	42.00	
		54.72	36.84	
	3	17	17	34
		15.45	15.45	30.91
		50.00	50.00	
		32.08	29.82	
	4	1	6	7
		0.91	5.45	6.36
		14.29	85.71	
		1.89	10.53	
	5	1	1	2
		0.91	0.91	1.82
		50.00	50.00	
		1.89	1.75	
	Total	53	57	110
		48.18	51.82	100.00

To understand the Output 7.4, examine the top left corner of the output. Here you find a list of words: frequency, percent, row percent, and column percent. These items represent the order and the number in each box. So, for the first box with statistics we see the numbers listed as follows

5

4.55

29.41

9.43

From this data we determine that five females did not skip class and 4.55% of all respondents were females who did not skip class. Among

respondents who did not skip class 29.41% were female. Among females, 9.43% did not skip class.

At the bottom of each column for female and male you find two numbers. The first number is the total number of respondents (i.e., frequency) represented by the column. So in the column for females we can conclude that there are 53 females in the data set with valid data values. The number below 53 is 48.18 and represents the percentage that women comprise of the total sample. To the far right you see the number 110. This number represents the total number of cases in the data set. Practice interpreting the results.

When you construct a two-way frequency table, you need to consider what information is important to report and what patterns develop in the table.

Continuous Variables

One of the best visual depictions of continuous variables is the histogram. **Histograms**, like bar charts, use rectangles or bar shapes to represent groups of values of a variable on a number line or a continuum. Unlike the bar chart, the rectangles on the histogram are connected with no spaces between them in order to represent the true continuous nature of the variable values. As with bar charts, the values of the variable are reported along the X-axis and the frequencies of the variable are along the Y-axis. The tops of the boxes/rectangles represent the highest value on the variable's values. Because continuous variables usually have many unique values, it is often necessary to create groupings of the data rather than having every value on the variable represented by a bar. Generally, keeping the number of bars to less than thirty allows the researcher to recognize patterns without minimizing too much of the variation in the variable's values. The groups of values on the variable are then represented on the histogram rather than individual values of the variable. Thirty bars or fewer should allow the researcher to determine what patterns exist in the values, such as bi-modal, skewness, and normal distribution. You should experiment with the groupings to get used to how these types of changes affect the visual display of the data. Next we review an example of grouping a continuous variable.

Consider a variable representing respondents' annual income that is measured in dollars and cents. If you want to view this variable on a histogram every unique income value would create a bar. There would be too many bars to interpret the histogram. One way to view the distribution is to create groupings of income values say, every $10,000 would create a new group category. While we do not review exactly how to accomplish this change in the income variable here, we cover it later. For now, just conceptually consider that you would create a new income variable with income categories from the original income variable measured in dollars and cents. If you used the income categories on your plot, you may be better able to see the pattern on the variable.

Histograms and QQ Plots

Visually examining data is necessary. Researchers commonly use histograms to examine the **distribution**, **spread**, and **outliers** of variables and use **QQ Plots (Quantile-Quantile Plot)** to determine if the distribution of the variable values is normal (although you can specify other distribution patterns other than normal, such as Weibull).

Histograms

Histograms differ from bar charts in that they are used with continuous variables where the 'bars' are touching to represent the quantitative nature of the variable. Histograms are appropriate to use when you can have valid values between whole numbers. For example, if we asked respondents in a survey to report their annual household income, we can easily imagine that one respondent said $12,234.56, a second reported $45,980.20, and another reported $102,300.23. In this example there are values that are not whole numbers: the .56, the .20, and the .23 (i.e., the change). We can also determine that any values in the "change" may be any place between .01, and .99. Because all of the values between whole numbers are valid, when we create histograms we must make sure that all "bars" are touching to symbolically represent the values are on a true continuum and that values between whole numbers are valid. Whereas in the Bar Chart, the bars or rectangles were distinct and not touching the other bars. The spaces represent that values between whole numbers do not exist or are invalid. The Y-Axis (Vertical Axis) represents the scale of the variable, usually a frequency or percentage of the cases with the value on the variable. The X-Axis (Horizontal Axis) represents the number of times the values occurred in that range or interval represented by the bar. The bars on the histogram indicate two characteristics about the variable: The wider the bar, the bigger the length of the group or time represented by the bar; the higher the bar, the more observations, frequencies, or cases it represents. Ordinarily, a visual display can aid in determining whether or not data distributions are normal. One can use a stem-and-leaf display or a box-and-whisker plot, for example. Another common display is a **normal probability plot**, specifically the quantile-quantile (QQ) plot.

QQ Plot

You use the QQ Plot to compare the distribution of variable values with a theoretical population distribution. It is the plotting of the quantiles of the sample with the quantiles of the theoretical population where you determine the shape of the distribution. Usually researchers are interested in knowing if the distribution of the variable in their data set mirrors a population with a normal distribution. You examine the QQ Plot to determine the distribution of the variable values. More advanced researchers may adjust the type of theoretical distribution to compare with the actual data. There are other plots such as the PP Plot that can be used to graphically analyze the sample distribution for normality. The QQ Plot is preferred because it is able to better detect variation from normality in the tails than other approaches (Gnanadesikan, 1997).

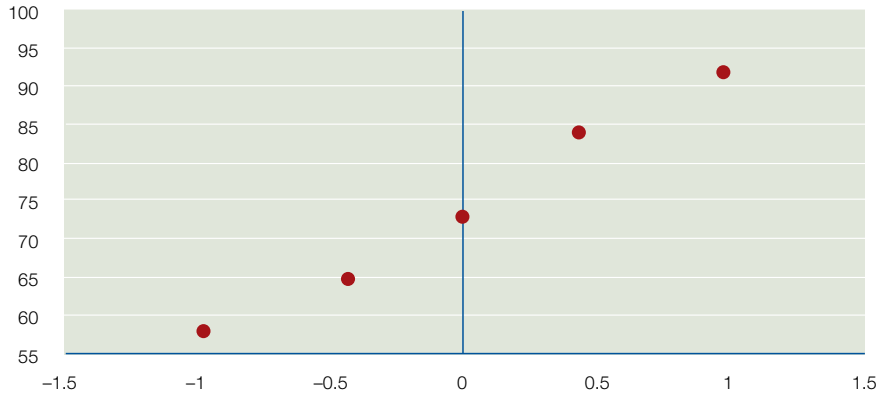
Let's consider how the QQ plot is constructed. Simulation studies for repeated random samples selected from normal distributions have shown that the ith score of n values in an ordered array has a standard **Z-score** value such that i/(n+1) proportion of the area under the normal curve is below that Z-score. A Z-score is also referred to as a standard score and as such it allows for comparison across distributions. A Z-score is tells you how far the value is from the mean. A Z-score of zero would indicated that score is equal to the mean, while a Z-score of 1 tells you the value is one standard deviation above the mean, and a Z-score of -1 tells you the value is one standard deviation below the mean. Positive Z-scores represent values greater than the mean and negative Z-scores represent values lower than the mean. So for example, you would expect a random sample of 5 observations to have z-scores as shown in Table 7.1 Random Sample with Expected Z Scores.

TABLE 7.1 Random Sample with Expected Z- Scores

X	Observation Number	Cumulative Area Below Z	Z-score
58	1	0.1667	–0.97
65	2	0.3333	–0.43
73	3	0.5000	0.00
84	4	0.6667	0.43
92	5	0.8333	0.97

Notice that the observations are put in order, so that i = 1, 2, 3, 4, and 5. The cumulative area for the first observation with i=1 is calculated using i/(n+1) = 1/(5+1) = 0.1667, the second observation has cumulative area = 2/(5+1) = 0.3333, etc. The Z-score for each value of X is found using the cumulative Z-table (See Appendix C Z-table). The QQ plot is a graphical display of the order pairs (X,Z) for each of the observations and is as follows:

FIGURE 7.1 Q-Q Plot for Random Sample of Five Observations



If the points follow along a 45-degree straight line (i.e., Θ must equal 45°), then the data are considered approximately normal. The interpretation of the QQ plots require some judgment, so the closer the points fall around the 45-degree line, the stronger the evidence that the data are normally distributed. To pull together several of these techniques, use the PROC UNIVARIATE with QQ PLOT AND HISTOGRAM with NORMAL option commands. The NORMAL option places a normal curve over the actual data on the histogram enhancing the visual comparison of the data to a normal curve.

Program with Guided Interpretation of Proc Univariate with Histogram and Normal Distribution Overlay

Program 7.5 PROC UNIVARIATE with QQ PLOT and HISTOGRAM with NORMAL option is used to generate a histogram with a normal distribution overlay and a QQ Plot of EAT_OUT. Read the program and try to understand what takes place in it. Run the program by adding it to the bottom of Program 7.4 Table of Categorical Variable by Categorical Variable. Delete the procedures that were contained in Program 7.4 Table of Categorical Variable by Categorical Variable, or comment them out by placing /* prior to the first procedure and adding */ after the last unwanted procedure. Be careful to not comment out the "RUN", as it is required at the end of the program to run it. Then be sure to save your file as Program 7.5 PROC UNIVARIATE with QQ PLOT and HISTOGRAM with NORMAL Option.

PROGRAM 7.4 PROC UNIVARIATE with QQ PLOT and HISTOGRAM with NORMAL Option

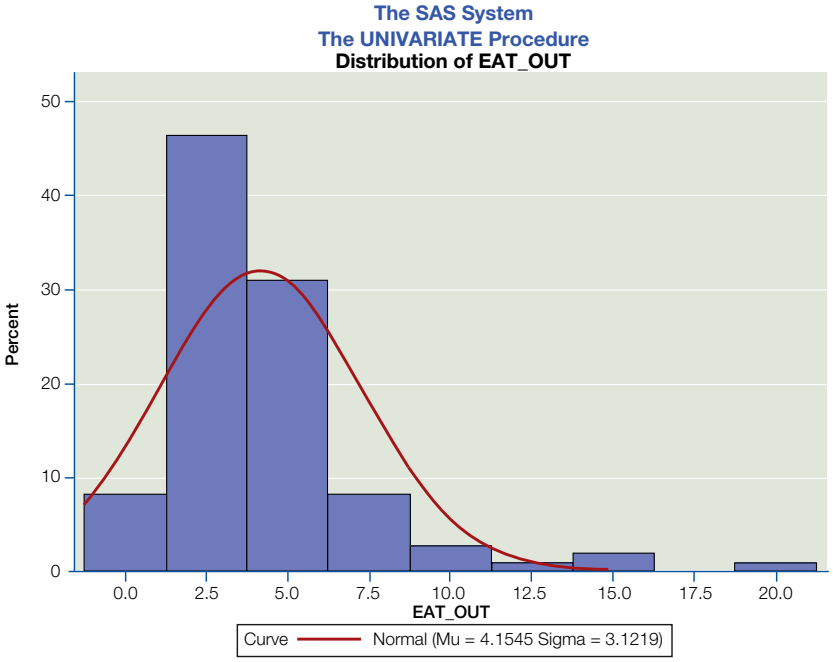
```
PROC UNIVARIATE;
VAR EAT_OUT;
QQPLOT /NORMAL (MU=EST SIGMA=EST); *REQUESTS THE QQ PLOT
USING NORMAL DISTRIBUTION AND SETTING THE MEAN (MU) TO
THE SAMPLE MEAN AND THE STANDARD DEVIATION (SIGMA) TO
THE SAMPLE DISTRIBUTION AND CREATES THE LINE SHOWING THE
NORMAL POPULATION VALUES;
HISTOGRAM /NORMAL; *NORMAL REQUESTS THAT THE LINE SHOWING
THE NORMAL CURVE BE OVERLAYED ON THE HISTOGRAM;
RUN;
```

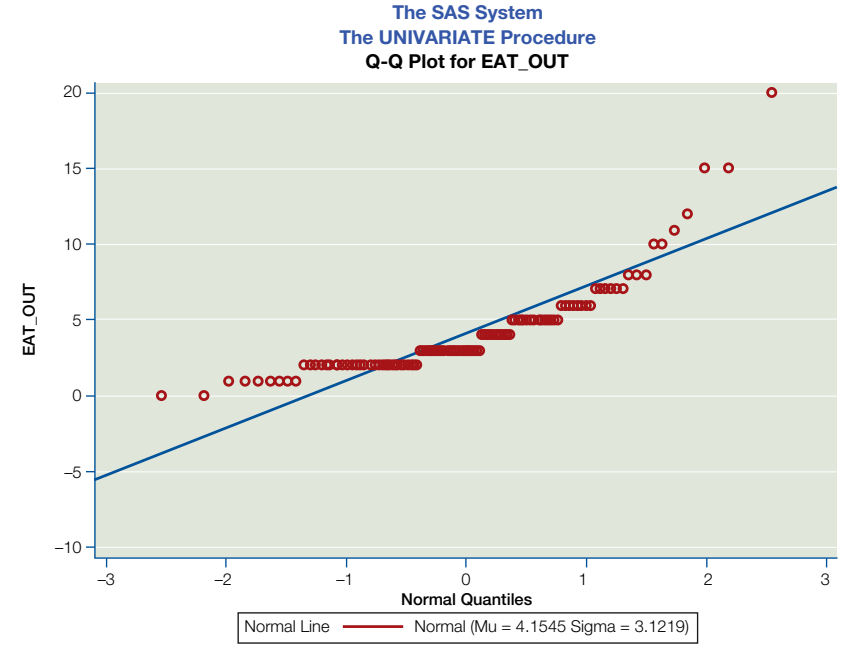

The histogram in Output 7.5. Histogram with NORMAL Curve Overlay and QQ Plot demonstrates some concerns about issues of normality. You should notice that the left tail of the normal distribution overlay is not on the histogram, signaling a first clue that the sample did not come from a population with a normal distribution. Another visible sign of a non-normal distribution is that the bars on the right of the histogram extend beyond the normal distribution tail. Finally, the normal distributions are symmetrical around the mean, which is not the case. One bar representing one case on the far right in the histogram is an outlier. Outliers are values on a variable that are dissimilar to the general values on the variable. Often there is a gap between outliers and the next closest valid value on the variable. In this histogram, the suspected outlier is one case with the value of "20", and there is a gap between this case's value and the next closest case that has a value of "15". Notice also how the histograms' bars are connected except for the one on the far right of the histogram with a value of "20". Taken together, these characteristics of the histogram suggest that the variable is not normally distributed. Remember that many statistical procedures assume that the variable has a normal distribution. When the variable does not have a normal distribution, you must determine the appropriate statistical test to use, or consider making transformations to the variable in order to obtain a normal distribution on the variable. Additional statistical tests reported by the PROC UNIVARIATE procedure also demonstrate that the variable EAT_OUT does not have a normal distribution. We cover these statistical tests in a later chapter.

Additional results in Output 7.5 provide the QQ Plot of EAT_OUT. Again, there is graphical evidence in the QQ Plot that the data were not drawn from a population with a normal distribution. The data points in a population with a normal distribution would fall along the 45 degree angle and go from the lower left corner to the upper right corner. But the results here show that almost none of the data points fall on the 45 degree angle and that there is heavy clustering in the lower left side of the plot reflecting a nonsymmetrical pattern in the data values. The

suspected outlier case is in the extreme upper right corner. You can see how it is farther from the 45 degree line than any other case and how it appears very much separated from all other cases. Again, the QQ Plot visually depicts the non-normal distribution of EAT_OUT. Next we review another tool to visually examine the distribution of a continuous variable.

OUTPUT 7.5 Histogram with NORMAL Curve Overlay and QQ Plot





Stem-and-Leaf Box Plot

Researchers use two additional techniques to examine the distribution of a variable in order to know if the distribution is approximately normal. The first is the **stem-and-leaf plot**. The stem-and-leaf plot depicts all values of a variable and how many times these values occur. You can determine frequency using the stem-and-leaf plot. You can create boxplots using **PROC BOXPLOT** procedure. SAS software has many options when using this procedure, but we use PROC UNIVARIATE in this section.

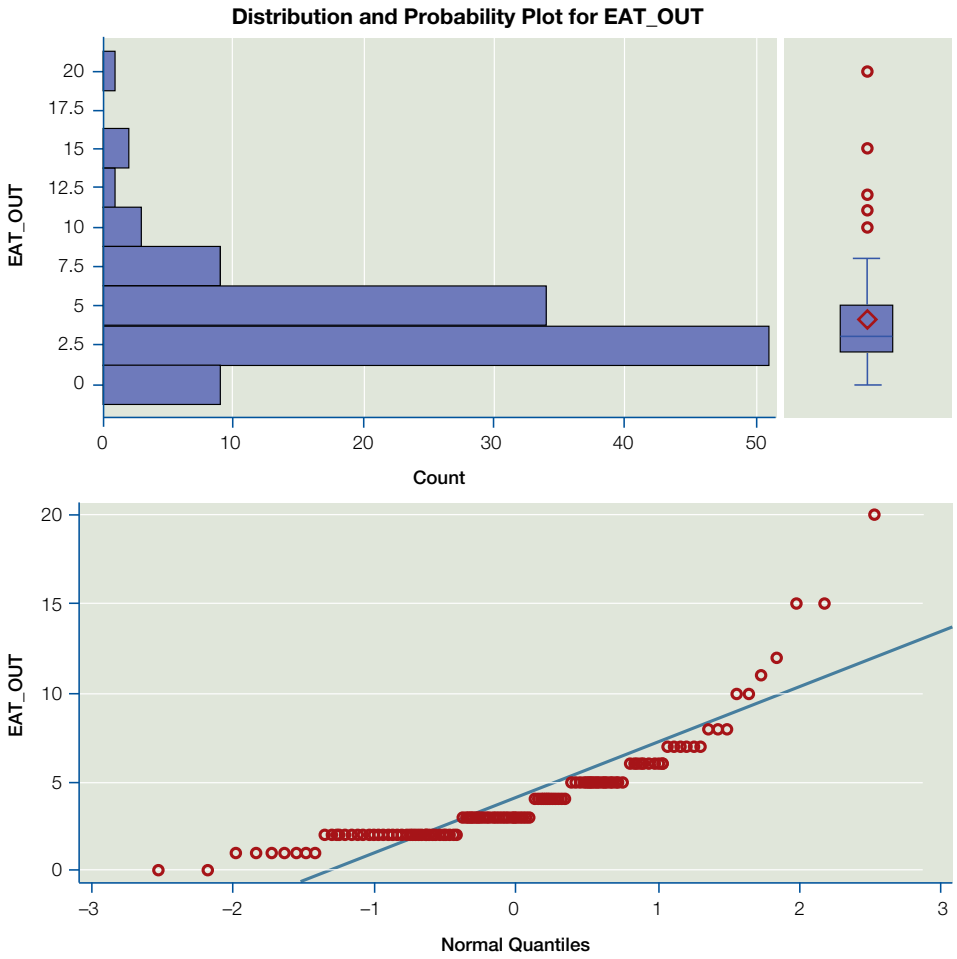
Guided Program with Interpretation of PROC UNIVARIATE with Option PLOT and NORMAL

Notice in Program 7.6 Stem Leaf and Box Plot of EAT_OUT using PROC UNIVARIATE with PLOT and NORMAL Options that PLOT and NORMAL are options for the PROC UNIVARIATE procedure, and are listed after PROC UNIVARIATE with only a space in between. Then, you proceed to the VAR subcommand to specify the variables of interest. The stem-and-leaf plot and the box plot are produced side-by-side in SAS output using PROC UNIVARIATE. First, we will review the BOX PLOT portion of Output 7.6 Stem Leaf and Box Plot of EAT_OUT using PROC UNIVARIATE with PLOT and NORMAL Options and will later review the Stem and Leaf results. So, first add Program 7.6 Stem Leaf and Box Plot of EAT_OUT using PROC UNIVARIATE with PLOT and NORMAL Options to Program 7.5 and either delete or comment-out the other procedures.

Program 7.5 Stem Leaf and Box Plot of EAT_OUT using PROC UNIVARIATE with PLOT and NORMAL Options

```
PROC UNIVARIATE PLOT NORMAL; /*THE PLOT OPTION WILL GEN-
ERATE THE STEM-AND-LEAF AND BOX PLOT*/
VAR EAT_OUT;
RUN;
```

Output 7.6 Stem Leaf and Box Plot of EAT_OUT using PROC UNIVARIATE with PLOT and NORMAL Options



The Output 7.6 Stem Leaf and Box Plot of EAT_OUT using PROC UNIVARIATE with PLOT and NORMAL Options provides the example of a stem-and-leaf and box plot of the variable, EAT_OUT. The box plot is interpreted as follows. The stem represents the variable value and the leaf represents the number of cases that held the specific value. SAS software totals these on the Count axis. You can tell that only one person reported eating out 20 times per week. Two people reported eating out 15 times per week, and so on. Again, as with our two previous graphical representations of the variable, you can see that the value "20" seems very different from the other values, as a large space exists between it and any other value. Thus, we would again conclude that it is an outlier. The shape of the distribution again suggests non-normal distribution. The variable EAT_OUT is **positively skewed**: more

values are on the right side of the distribution than are on the left side of the distribution. Now, we will review the box plot interpretation.

Boxplot

The **boxplot** (also known as the **BOX AND WHISKERS PLOT**) is useful for depicting distribution, spread, and outlier (Tukey, 1977). Examine the boxplot in Output 7.6 Stem Leaf and Box Plot of EAT_OUT using PROC UNIVARIATE with PLOT and NORMAL Options. You have to learn some key terminology to be able to discuss and describe the box plot results. The box is visually displayed as a box on the box plot. It represents the middle 50% of the variable's values. Another way to describe the box plot is that the box represents the range of values from the 25th percentile to the 75th percentile. The area represented by the box represents the interquartile range (IQR). The **IQR** is

IQR=75th percentile-25th percentile

Thus, 75% of the data values fall below the 75th percentile while 25% of cases fall below the 25th percentile. Again, think of the IQR as the values between the 25th percentile and the 75th percentile that comprise the middle half of all data values for the variable. The value represented by the 25th percentile is the median of the lower half of the data values and the 75th percentile value is the median for the upper half of the data values. A single line crossing the image represents the median value for the variable.

Another key term used to describe the box plot is **whiskers**. The whiskers are the lines extending away from the box. The ends of the whiskers represent the highest and lowest values on the variable that are not considered outliers. Also, a longer whisker on one end, compared to the other also suggests non-normal distribution. Our output again reveals that one whisker is longer than the other. When the box is located in the middle of the whiskers, that is, when the whiskers are approximately equal in length, you may be inclined to conclude the distribution is approximately normal. But, other factors also have to be considered.

Outlier is often another term used when describing distributions on boxplots. Outliers are represented by single symbols that do not touch the whiskers. In Output 7.6 Stem Leaf and Box Plot of EAT_OUT using PROC UNIVARIATE with PLOT and NORMAL Options the outliers are represented by the "o" symbol. In addition, it highlights the extreme outlier values "20" and "15". There are different types of outliers. Extreme outliers are those outliers farthest from the whiskers. Outliers are not as far from the whiskers as extreme outliers.

Yet another symbol on the boxplot represents the mean value. The mean is represented by a single symbol. In Output 7.6 Stem Leaf and Box Plot of EAT_OUT using PROC UNIVARIATE with PLOT and NORMAL Options the mean is represented by a "◊". It is near the middle of the box. When the mean and the median are close on the box plot, the

data may be normally distributed. When the mean and the mean are far from one another, the data are not normally distributed.

Thus, we have several indications in this example that EAT_OUT is not normally distributed:

- Box is not in the middle of values.
- The median and the mean are not close together.
- The median is outside of the box.
- The whiskers are not approximately equal in length.
- There are outliers.

Taken together, these characteristics represent a non-normal distribution.

Bar charts, histograms, and scatter plots are often in reports written for lay audiences. However, it would be very unlikely to see a box plot included in such a report. Rather, we suggest that you include information about the IQR and explain that a box plot was used to identify outliers, etc.

Plots

Within SAS software there are a number of ways to visually examine variables. One additional SAS procedure is **PROC PLOT**. This procedure can provide a variety of visual depictions of a variable. PROC PLOT is commonly used to examine two variables with interval or higher level of measurement, simultaneously. You will examine the number of times a respondent went out to eat and the number of hours the respondent worked per week by inserting the following syntax into your previous program. Note here that you generate two plots using two SAS procedures: PROC PLOT and **PROC GPLOT**.

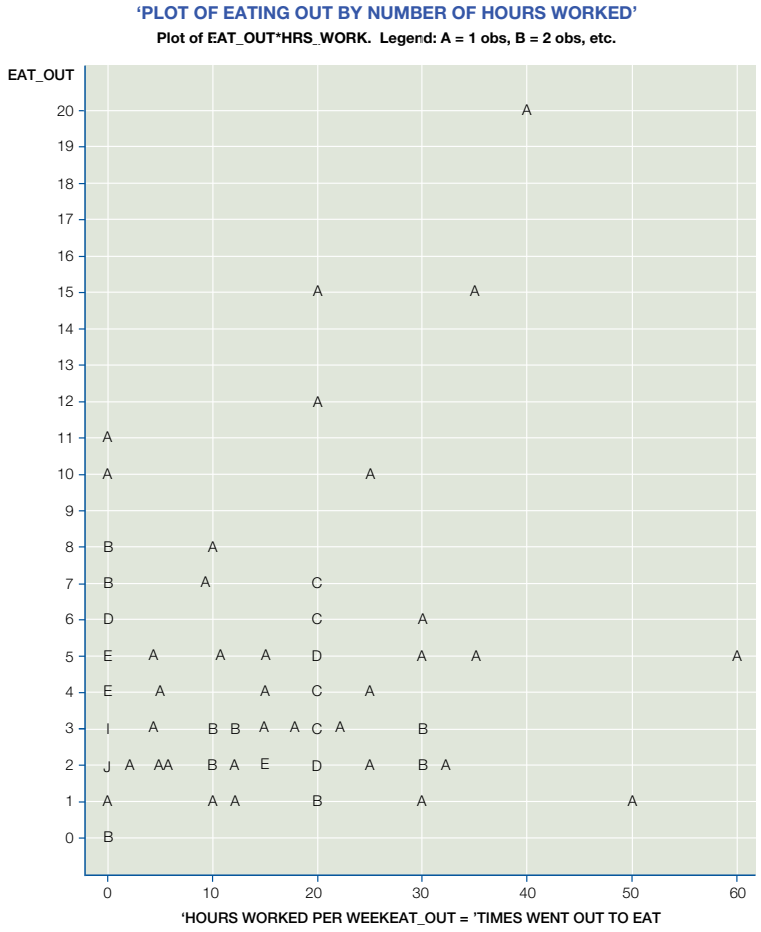
Review Program 7.7 PROC PLOT and PROC GPLOT and insert it into Program 7.6 that you previously created. Again, either delete the procedures that you just completed, or comment them out.

Program 7.6 PROC PLOT and PROC GPLOT

```
PROC PLOT;
PLOT EAT_OUT*HRS_WORK;
TITLE 'PLOT OF EATING OUT BY NUMBER OF HOURS WORKED';
RUN;

PROC GPLOT;
PLOT EAT_OUT*HRS_WORK;
TITLE 'GPLOT OF EATING OUT BY HOURS WORKED';
RUN;
```

REVIEW OUTPUT 7.7 PROC PLOT and PROC GPLOT



The primary difference between the two charts generated from Program 7.7 PROC PLOT AND PROC GPLOT is that the PROC GPLOT includes every data value on the chart, whereas the output from PROC PLOT results in grouping nearby data into a single symbol, for example, 2 observations in the same area are represented by "B".

Researchers must plot the relationship between two interval level variables to examine the relationship that appears. The patterns observed in the scatter plot also indicate direction of relationship between two variables, strength of the relationship between the two variables, and expose bivariate outliers. When you see clustering at one extreme and the other, you might conclude that there are two types of patterns in the data. You have to consider the variables and speculate about why this pattern is present. Other common relationships between variables are **positive** and **negative**.

One of the key reasons to visually examine the relationship between two continuous variables is to determine if the relationship is linear. A **linear relationship** is when two variables' values change unit for unit in a systematic pattern. A perfectly **positive linear relationship** occurs when for every unit increase on a variable you observe a unit increase on the second variable. Relationships between variables that do not approximate a linear relationship require either data analysis using nonparametric statistics, or for the variables to be recoded or transposed to a different scale, as was discussed earlier in the book.

In sum, tighter clustering along the 45 degree angle represents strong positive relationship. A pattern that tends to follow the 45 degree angle but where the dots are not tight represents a weaker positive relationship. Generally, the more "scatter" among the dots the lower the strength of the relationship between variables.

Program with Guided Interpretation of PROC PLOT-SCATTER PLOT by Statement

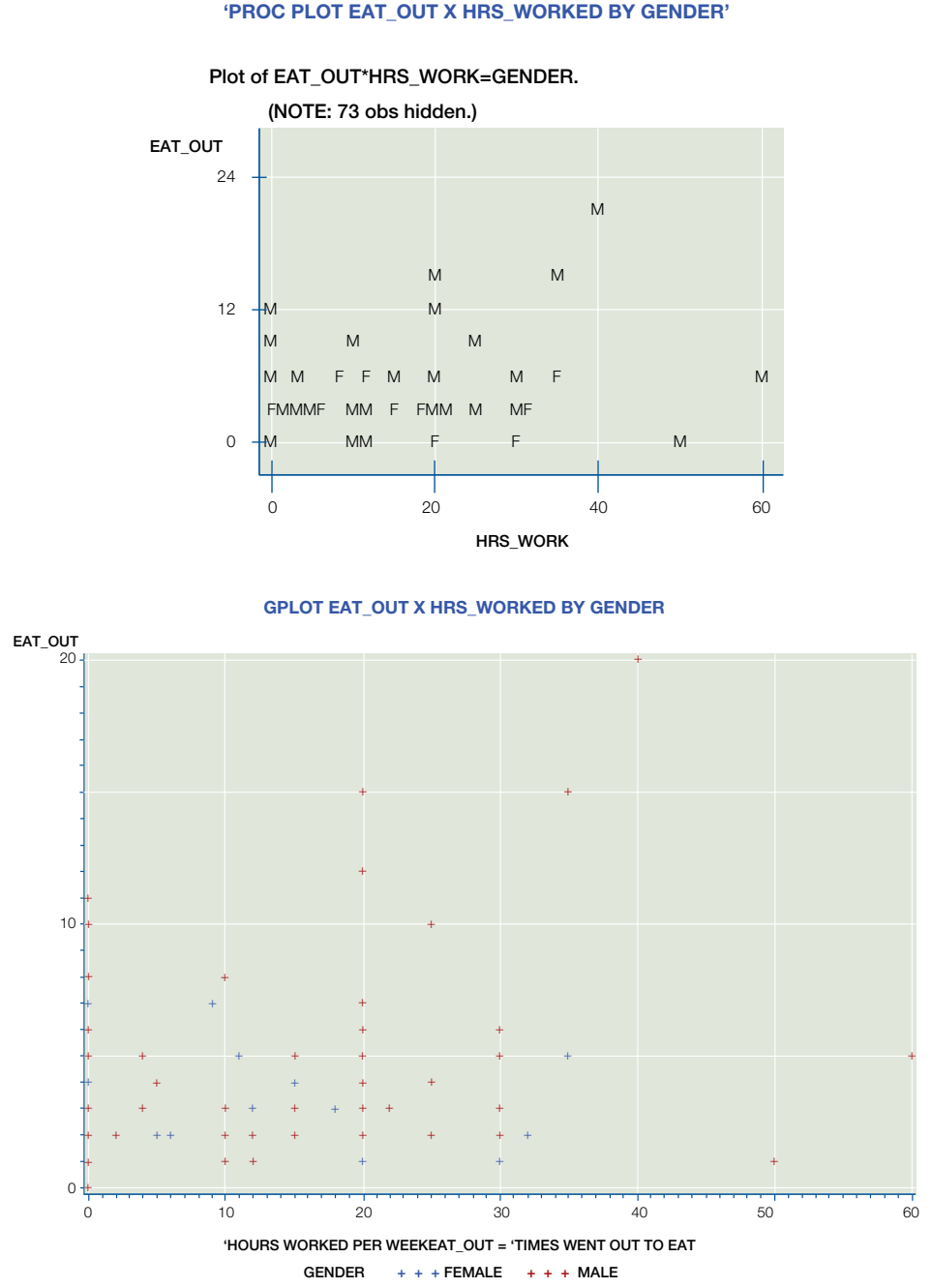
Below is an example of how to generate a **scatter plot** of EAT_OUT and HRS_WORK by GENDER. Add Program 7.8 Plot of Two Continuous Variables to the Program 7.7 PROC PLOT and PROC GPLOT and delete or comment out any procedures no longer needed. Save your program as Program 7.8 Plot of Two Continuous Variables.

Program 7.7 Plot of Two Continuous Variables

```
PROC PLOT HPERCENT=50 VPERCENT=33 NOMISS; *TO MAKE THE
CHART FIT TO ONE SCREEN, I RESET THE HORIZONTAL AND VER-
TICAL SIZES USING HPERCENT AND VPERCENT;
PLOT EAT_OUT*HRS_WORK=GENDER;
TITLE 'PROC PLOT EAT_OUT X HRS_WORKED BY GENDER';
RUN;
PROC GPLOT;
PLOT EAT_OUT*HRS_WORK=GENDER;
```

```
TITLE 'GPLOT EAT_OUT X HRS_WORKED BY GENDER';
RUN;
```


Output 7.8 Plot of Two Continuous Variables



The Output 7.8 Plot of Two Continuous Variables reflects that 73 observations are hidden. SAS does not show all values on the chart using PROC PLOT when the values are very close in the same physical location on the chart. Also, notice that the variable name EAT_OUT is used in place of the variable label because the variable label exceeds the default length in SAS software for use in the plot. In Program 7.8 Plot of Two Continuous Variables we also changed the variable labels as follows:

HRS_WORK = 'HOURS WORKED PER WEEK'

EAT_OUT = 'TIMES WENT OUT TO EAT'

The output from the PROC GPLOT procedure has no hidden values reported on it because using the SAS Graphics component of the software displays all values on the variable.

Scale Creation and Cronbach's Alpha

In the previous section you learned how to examine variable characteristics when only one variable was involved. Another type of variable can be created from several variables. This type of variable is a **scale variable**. Let's imagine that we want to study delinquency. First we need to determine how we want to define delinquency. We could ask just one question of respondents, such as, "Have you ever stolen an item from a store?" While a few people might believe this single question taps the delinquency concept, most would likely agree that we need to ask more than one question. In the field of criminology a relatively stable set of questions are used to develop a criminality scale. Most criminologists use the delinquency scale in their research on delinquency. Giordano, Cernkovich, and Holland (2003) used several scales in their research on desistence from crime. They created an adult criminality scale using a modified version of the Elliott et al. (1985) delinquency scale. They also created a friends' criminal involvement scale using 7 questions from their survey where respondents provided information about their friends' criminal involvement. The answer options were 1 (none), 2 (some), 3 (most), or 4 (all). The friends' criminal involvement scale was created by summing the responses to all seven items (See Giordano, Cernkovich, and Holland, 2003 for specific scale questions). The range on the newly created variable was 7–28. In the friends' criminal involvement scale all items had values ranging from 1–4 (Giodano, Cernkovich, and Holland, 2003). Below is a generic example of how one would create a scale item where all questions have the same number and types of answer options.

$$\text{SCALE=SUM(ITEM}_1\text{+ITEM}_2\text{+ITEM}_3\ldots\text{ITEM}_N\text{);}$$

Another programming option to create a scale from several variables that are listed in adjacent form in the data set (that are listed side-by-side) is:

Positive (direct) relationship

As the value on one variable goes up the value on the other variable goes up. We'd expect a positive relationship between income and cost of home; as a family income increases, the cost of the home increases.

Negative (inverse) relationship

As the value on one variable goes up, the value on the other variable goes down. We'd expect a negative relationship between Body Mass Index (a scale used to determine how overweight people are) and health and well-being; the more overweight people are, the lower their health and well-being.

SCALE=SUM (OF $ITEM_1$ - - $ITEM_N$);

The double hyphen directs SAS to include all items that are adjacent and have an ending numeric value. If you accidentally omit one of the hyphens, SAS software will interpret the syntax as a request for a mathematical subtraction sign and will calculate the difference between $ITEM_1$ and $ITEM_N$. If you forget the word OF, then SAS software will not create the scale in the correct manner.

Sometimes researchers have a series of questions that represent an underlying concept, but all of the answer options for each item are not on the same scale. That is, sometimes a question may have 5 answer options, and other questions may have 7 answer options or some other combination. When this is the case, you cannot create the scale by summing the items as we did above. First, the items have to be converted to a common scale.

Once such a scale is created, a researcher has to report **Cronbach's Alpha**, the **coefficient of reliability** (Cronbach, 1951). Cronbach's Alpha is an assessment of the internal consistency of the scale. Cronbach's Alpha ranges from 0–1. The closer the value is to 1, the higher the internal consistency of the scale. That is, the closer each item is related to the underlying construct of interest in the population. When Cronbach's Alpha is closer to zero, the group of items may not reflect the true scale. Commonly used acceptable level of Cronbach's Alpha is 0.7 (Nunnally, 1977). There is debate about the appropriateness of using scales with lower Cronbach's Alpha levels. The researcher has to decide if lower Cronbach's Alpha levels are justified by arguing that the items included in the scale actually represent the underlying concept, even if their inter-item correlations are not that high. It is a conceptual and theoretical argument rather than a statistical one. The syntax for calculating Cronbach's Alpha is

```
PROC CORR ALPHA NOMISS; *ALPHA GENERATES CRONBACH'S ALPHA
AND NOMISS DIRECTS SAS NOT TO INCLUDE CASES WITH MISSING
DATA;
VAR {LIST ALL VARIABLES HERE SEPARATED BY SPACE};
RUN;
```

Output would appear as

Cronbach Coefficient Alpha	
Variables	**Alpha**
Raw	0.172848
Standardized	0.432459

If you ever obtain a negative Cronbach's Alpha, it indicates a coding problem. All items in the scale must have values in which the increasing value represents the construct. If your scale items were based on how often your friends

OWN FRIENDS:DESTROYED PROPERTY

OWN FRIENDS:USED MARIJUANA

OWN FRIENDS:HIT SOMEONE

OWN FRIENDS:BROKEN INTO VEHICLE

OWN FRIENDS:SOLD HARD DRUGS

OWN FRIENDS:STOLEN WORTH > $50

OWN FRIENDS:USED PRESCRIP DRUGS

and the answer options were never (1), some (2), most (3), and all (4) where all answer options are on the same scale, you report Cronbach's Alpha for RAW variables. All of the items and the answer options are written in such a way that a higher value represents more friend criminality. If, however, some of the answer options were as noted above and others were reversed such as all (1), most (2), some (3), and none (4), then we have to **reverse code** these items so that a higher score reflects greater friend criminality.

Let us imagine then, that the last item "OWN FRIENDS: USED PRESCRIP DRUGS" had the following answer options on the survey, all (1), most (2), some (3), and none (4). A respondent who selected "never" has a value of "4" in the data set. Given that the name of the scale is "friend criminality", you want the higher value to indicate more criminal behavior by friends. In this instance a higher value of 4 indicates that this person's friends never used prescription drugs. You have to create a new variable that restructures the answers so that a 4 represents that all friends used drugs. In doing this we take all values of 1 for this variable and assign the value of 4, all values of 2 and assign the value of 3 and all values of 3 and assign the value of 2, and all values of 4 and assign the values of 1. Even reading this you can see that a simple task can be tricky to do manually. But, SAS software allows for simple syntax to handle this reverse coding. Assign a variable name of FPREDRUG for this variable. The syntax is

```
FSCRIPD=5-FPREDRUG;
```

and would be placed in the program before the data lines. Notice that to reverse code the variable values we create a new variable, FSCRIPD, by taking one more than the number of actual answer options (which were 1–4) and subtracting the original variable from it (5-FPREDRUG). Another approach to reverse coding several variables where the formula for the recode is the same, is to use the following **array** and do **loop**:

Annotated Activity 7.1

When would it be appropriate to use histograms?

_____

_____

_____

_____

```
array revers (i) item1 item2 item3 item4 item5;
do over revers;
revers=6-revers;
end;
scale=sum(of item1--item5);
```

The first line in the above partial program begins the array process in SAS and then names the array "revers" and then lists the elements of the array, items 1–5. If you were reverse coding a series of dummy variables, you would subtract the items from "1". Note that these items and the variables do not exist in our data set that we have been using. These are mere examples.

Handling Missing Data

One approach to retaining cases with some missing data on scale (index) questions is to replace the missing values with the mean as suggested by Afifi and Elashoff (1966). Missing values should be replaced prior to creating the summated scale variable. Using the newly created variables where the mean was imputed for missing, you create a new variable by summing the items.

Key Terms

array
bar charts
box and whiskers plot
boxplot
distribution
coefficient of reliability
Cronbach's Alpha
do loop
IQR
mean
median
midpoint
negative (inverse) relationship
normal probability plot
outliers
positive relationship
positively skewed
PROC BOXPLOT
PROC GPLOT
PROC PLOT
PROC TABLES
qq plots
quantile-quantile plot
reverse code
scale variable
scatter plot
spread
stem-and-leaf plot
whiskers
Z-score

Portfolio Assignment

Obtain three articles or applied reports that report variable characteristics/analysis in the text of the article or applied report. Write a summary of how each author explains and interprets variable characteristics.

Make a chart that explains when each item covered in this chapter may be used.

Review

How do you reverse code a variable? Why is it necessary?

When do you use a histogram versus a bar chart?

Questions

1. Univariate analysis
   a. varies for categorical and continuous variables.
   b. includes visually examining the data.
   c. is the final step a researcher takes to develop understanding of the data.
   d. both a and b

2. Categorical variables
   a. require use of histograms
   b. require use of bar charts
   c. are synonymous with qualitative variables
   d. both b and c

3. Continuous variables
   a. require use of histograms
   b. require use of bar charts
   c. are synonymous with qualitative variables
   d. both b and c

4. Bar charts
   a. are used with interval level data.
   b. graphically depict the frequency, count or percentage of values on a variable.
   c. have rectangles that touch to represent the continuous nature of the values.
   d. both a and b

5. Histograms
   a. are used with interval level data.
   b. graphically depict the frequency, count or percentage of values on a variable.
   c. have rectangles that touch to represent the continuous nature of the values.
   d. both a and b

6. You can turn the bar charts to a horizontal chart by using
   a. VBAR subcommand
   b. HBAR subcommand
   c. BARV subcommand
   d. BARH subcommand

7. You can have the bar charts as a vertical chary by using
   a. VBAR subcommand
   b. HBAR subcommand
   c. BARV subcommand
   d. BARH subcommand

8. To generate a plot of two interval level variables you use
   a. PROC PLOT
   b. PROC GPLOT
   c. PROC GCHART using GROUP option
   d. both a and b

9. To generate a plot of two categorical variables you use
   a. PROC PLOT
   b. PROC GPLOT
   c. PROC GCHART using GROUP option
   d. both a and b

10. Boxplots
    a. display the distribution of a variable.
    b. display the spread of a variable.
    c. both a and b
    d. none of the above