Data Analysis Using SAS®

9. Comprehensive Descriptive Analysis and Normality Test

Contributors: C. Y. Joanne Peng
Print Pub. Date: 2009
Online Pub. Date:
Print ISBN: 9781412956741
Online ISBN: 9781452230146
DOI: 10.4135/9781452230146
Print pages: 163-207

10.4135/9781452230146.n9

[p. 163 ↓]

9. Comprehensive Descriptive Analysis and Normality Test

Objective

This chapter covers the most comprehensive descriptive analysis of data using the UNIVARIATE procedure. Included in this comprehensive analysis are typical indices of central tendencies, spread, skewness, kurtosis, percentiles, quartiles, missing data, valid data, and so on. Furthermore, the distribution of data is displayed in three different graphs: (1) the stem-and-leaf plot, (2) the box plot, and (3) the normal probability plot. The shape of the data distribution can also be tested against the theoretical normal distribution or examined for extreme values or outliers.

[p. 164 ↓]

9.1 Comprehensive Descriptive Analysis and the Test of Normality

A comprehensive descriptive data analysis is best carried out by the UNIVARIATE procedure in SAS. This procedure is capable of in-depth examination of quantitative data, variable by variable, including (a) three indices of central tendency (mean, median, mode); (b) six indices of variability (range, standard deviation, variance, coefficient of variation, interquartile range, standard error of the mean); (c) indices of skewness and kurtosis; (d) various percentiles and quartiles; (e) miscellaneous descriptive values such as the total, sample size, sum of weights, and so forth; (f)

identification of extreme values in a frequency distribution; (g) three visual displays of data (stem-and- leaf plot, box plot, and normal probability plot); and (h) three tests of central tendency (the *t* test, the sign test, and the signed rank test) plus four tests of normality (the Shapiro-Wilk test, the Kolmogorov-Smirnov test, the Cramérvon Mises test, and the Anderson-Darling test). This list of features reads like Who's Who in descriptive data analysis, doesn't it? Obviously, you recognize the similarity between the MEANS and the UNIVARIATE procedures. So you ask, “Why bother with the UNIVARIATE procedure?” The extra features listed in (f) through (h) are the reasons why you need this procedure at times when the MEANS procedure can't quite cover all bases, though the MEANS procedure is still handy for quickly summarizing data for multiple variables. These unique features about PROC UNIVARIATE are demonstrated in Examples 9.1 through 9.5.

Four statistical tests of normality are particularly useful for examining the degree to which a data distribution is approximately normal (Example 9.4). Because many parametric statistical procedures require that the underlying population be normally distributed, these tests can help you verify the normality assumption based on a sample.

In addition to the parametric *t* test, PROC UNIVARIATE performs the sign test based on positive or negative signs of data, deviating positively or negatively from the hypothesized population median. This test allows you to make inferences about the population median with 20 or fewer observations or when data are measured at the ordinal level or higher.

Another test available from PROC UNIVARIATE, but not from PROC MEANS, is the signed rank test. This test is also used to make inferences about a population median (and mean, if the distribution is symmetric). The signed rank test improves over the sign test by using both signs (+ or -) and ranks of data, after adjusting for the sample median. Thus, it is a more powerful test than the sign test when the underlying population is unimodal and symmetric. Both the sign test and the signed rank test are particularly useful for small data sets or with crude measurements, such as ordinal-level data (Example 9.5). In addition to making inferences about a population median, both tests are also applicable to testing the difference between two paired as well as two independent population medians.

[p. 165 ↓]

With PROC UNIVARIATE, you may also (a) stratify the sample into strata (or subgroups), analyze each stratum separately (Example 9.6), and (b) create an output data set for subsequent analysis (Example 9.7)—these features are available in PROC MEANS as well.

9.2 Examples

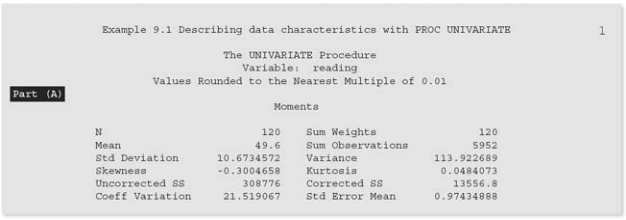
There are seven examples demonstrated in this chapter. All examples use either achieve.dat or mydata.dat for demonstration. Details about these data files are given in Appendix B.

Example 9.1 Describing Data Characteristics With PROC UNIVARIATE

This example shows how to compute simple descriptive statistics for the variable reading. The first statement is PROC UNIVARIATE, followed by an option ROUND= 0.01. This option rounds off variable values at the second decimal place prior to computing statistics. The second statement is VAR, which specifies reading to be analyzed.

```
/* The following bolded SAS statements establish the SAS data set 'achieve' */

DATA achieve;
     INFILE 'd:\data\achieve.dat';
     INPUT iv1 1 grade 2 iv2 3 sex $ 4 id $ 6-8 vocab 25-26 reading 27-28
           spelling 29-30 capital 31-32 punc 33-34 usage 35-36
           total1 37-38 maps 39-40 graphs 41-42 refer 43-44
           total2 45-46 concepts 47-48 problem 49-50 total3 51-52
           composite 53-54;
RUN;

TITLE 'Example 9.1 Describing data characteristics with PROC UNIVARIATE';

PROC UNIVARIATE DATA=achieve ROUND=0.01;
     VAR reading;
RUN;
```

Output 9.1 Describing Data Characteristics With PROC UNIVARIATE



Example 9.1 Describing data characteristics with PROC UNIVARIATE 1

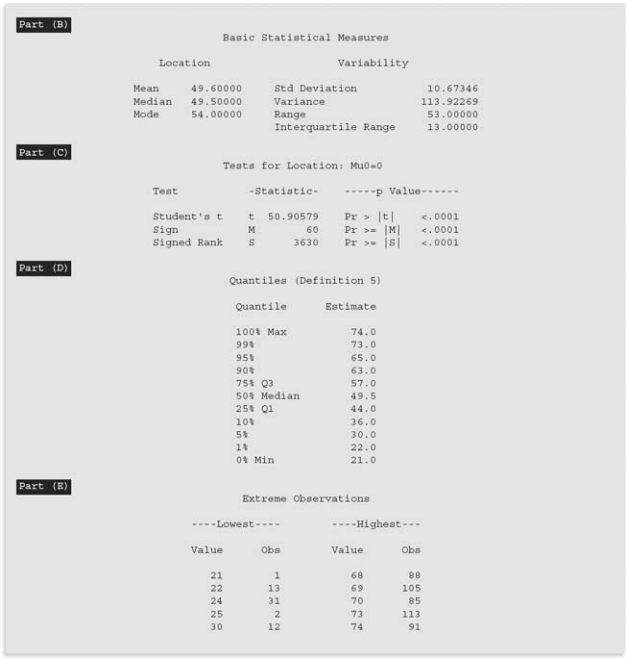
The UNIVARIATE Procedure
Variable: reading
Values Rounded to the Nearest Multiple of 0.01

Part (A)

Moments

N	120	Sum Weights	120
Mean	49.6	Sum Observations	5952
Std Deviation	10.6734572	Variance	113.922689
Skewness	-0.3004658	Kurtosis	0.0484073
Uncorrected SS	308776	Corrected SS	13556.8
Coeff Variation	21.519067	Std Error Mean	0.97434888

[p. 166 ↓]



Part (B)

Basic Statistical Measures

Location		Variability	
Mean	49.60000	Std Deviation	10.67346
Median	49.50000	Variance	113.92269
Mode	54.00000	Range	53.00000
		Interquartile Range	13.00000

Part (C)

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	50.90579	Pr > |t|	<.0001
Sign	M	60	Pr >= |M|	<.0001
Signed Rank	S	3630	Pr >= |S|	<.0001

Part (D)

Quantiles (Definition 5)

Quantile	Estimate
100% Max	74.0
99%	73.0
95%	65.0
90%	63.0
75% Q3	57.0
50% Median	49.5
25% Q1	44.0
10%	36.0
5%	30.0
1%	22.0
0% Min	21.0

Part (E)

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
21	1	68	88
22	13	69	105
24	31	70	85
25	2	73	113
30	12	74	91

For purposes of explanation, the output is divided into five parts: **(A) Moments, (B) Basic Statistical Measures, (C) Tests for Location: Mu0=0, (D) Quantiles, and (E) Extreme Observations.**

Part (A) Moments: This seemingly nonstatistical term refers to four moments (or characteristics) of the reading distribution and other descriptive indices. The first moment is the mean, the second moment

is variance, the third moment is skewness, and the fourth moment is kurtosis. Table 9.1 provides a line-by-line interpretation of the results.

Example 9.2 Frequency Tabulation via PROC UNIVARIATE

This example illustrates how to compile a frequency distribution for reading. A frequency distribution is a tabulation of raw scores and the frequency of their occurrences. The tabulation sometimes can be more useful than the summary statistics presented in Output 9.1. To request a frequency distribution, simply type in the option **FREQ;** see the program below:

```
/* See Example 9.1 for the DATA step in creating the SAS data set 'achieve' */

TITLE 'Example 9.2 Frequency tabulation via PROC UNIVARIATE';

PROC UNIVARIATE DATA=achieve FREQ;
     VAR reading;
RUN;
```

Output 9.2 Frequency Tabulation via PROC UNIVARIATE

```
                Example 9.2 Frequency tabulation via PROC UNIVARIATE                    1

                            The UNIVARIATE Procedure
                               Variable:  reading

Part (A)
                                     Moments

            N                         120    Sum Weights                120
            Mean                     49.6    Sum Observations          5952
            Std Deviation       10.6734572    Variance             113.922689
            Skewness            -0.3004658    Kurtosis              0.0484073
            Uncorrected SS           308776    Corrected SS           13556.8
            Coeff Variation      21.519067    Std Error Mean       0.97434888

Part (B)
                            Basic Statistical Measures

                  Location                        Variability

               Mean     49.60000     Std Deviation           10.67346
               Median   49.50000     Variance               113.92269
               Mode     54.00000     Range                   53.00000
                                     Interquartile Range     13.00000
```

[p. 170 ↓]

Part (C)

Tests for Location: Mu0=0

Test	-Statistic-	-----p Value------
Student's t	t 50.90579	Pr > |t| <.0001
Sign	M 60	Pr >= |M| <.0001
Signed Rank	S 3630	Pr >= |S| <.0001

Part (D)

Quantiles (Definition 5)

Quantile	Estimate
100% Max	74.0
99%	73.0
95%	65.0
90%	63.0
75% Q3	57.0
50% Median	49.5
25% Q1	44.0
10%	36.0
5%	30.0
1%	22.0
0% Min	21.0

Part (E)

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
21	1	68	88
22	13	69	105
24	31	70	85
25	2	73	113
30	12	74	91

Example 9.2 Frequency tabulation via PROC UNIVARIATE 2

The UNIVARIATE Procedure
Variable: reading

Part (F)

Frequency Counts

Value	Count	Percents Cell	Percents Cum	Value	Count	Percents Cell	Percents Cum	Value	Count	Percents Cell	Percents Cum
21	1	0.8	0.8	45	5	4.2	34.2	58	1	0.8	78.3
22	1	0.8	1.7	46	5	4.2	38.3	59	4	3.3	81.7
24	1	0.8	2.5	47	3	2.5	40.8	60	5	4.2	85.8
25	1	0.8	3.3	48	8	6.7	47.5	61	1	0.8	86.7
30	3	2.5	5.8	49	3	2.5	50.0	62	3	2.5	89.2
33	1	0.8	6.7	50	3	2.5	52.5	63	4	3.3	92.5
34	1	0.8	7.5	51	3	2.5	55.0	64	2	1.7	94.2
36	4	3.3	10.8	52	3	2.5	57.5	65	2	1.7	95.8
37	5	4.2	15.0	53	1	0.8	58.3	68	1	0.8	96.7
39	1	0.8	15.8	54	10	8.3	66.7	69	1	0.8	97.5
40	5	4.2	20.0	55	3	2.5	69.2	70	1	0.8	98.3
42	3	2.5	22.5	56	4	3.3	72.5	73	1	0.8	99.2
43	2	1.7	24.2	57	6	5.0	77.5	74	1	0.8	100.0
44	7	5.8	30.0								

[p. 171 ↓]

The five parts on page 1 of the output are identical to those in Output 9.1. The new result is **Part (F**) under the heading **Frequency Counts** on page 2. Included in this table are four kinds of information. The first column, under the heading **Value**, presents the raw score value. The second column, under **Count**, gives the frequency of each score's occurrence. The last two columns display the percentages (frequency divided by N) and cumulative percentages, respectively. Thus, the last number in the last column is 100.0, or 100%.

Technically, the frequency distribution shown in **Part (F**) is an ungrouped frequency distribution because every raw score is displayed and counted. This presentation format is thorough because it includes every nuance of information in the data set. It can also be tedious, if your data are voluminous.

Example 9.3 Three Visual Displays of Data: Stem-and-Leaf, Box Plot, and Normal Probability Plot

Option **PLOT** in the PROC UNIVARIATE statement produces three visual displays of data: the stem-and-leaf plot, the box plot, and the normal probability plot. The stem-and-leaf plot and the box plot are simple, yet effective, techniques for exploring data. They were pioneered by John W. Tukey in the late 1970s. The normal probability plot provides a way to evaluate sample data against a normal curve. This plot is useful for checking if data meet the normal assumption, which is assumed by many inferential statistical procedures.

```
/* See Example 9.1 for the DATA step in creating the SAS data set 'achieve' */

TITLE 'Example 9.3 Three visual displays';

PROC UNIVARIATE DATA=achieve PLOT;
     VAR reading;
RUN;
```

Output 9.3 Three Visual Displays of Data: Stem-and-Leaf, Box Plot, and Normal Probability Plot

```
                          Example 9.3 Three visual displays                    1

                              The UNIVARIATE Procedure
                                 Variable:  reading
Part (A)
                                      Moments

           N                         120    Sum Weights               120
           Mean                     49.6    Sum Observations         5952
           Std Deviation       10.6734572    Variance            113.922689
           Skewness            -0.3004658    Kurtosis             0.0484073
           Uncorrected SS          308776    Corrected SS           13556.8
           Coeff Variation     21.519067     Std Error Mean      0.97434888
```

[p. 172 ↓]

Part (B)

Basic Statistical Measures

Location		Variability	
Mean	49.60000	Std Deviation	10.67346
Median	49.50000	Variance	113.92269
Mode	54.00000	Range	53.00000
		Interquartile Range	13.00000

Part (C)

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------
Student's t	t 50.90579	Pr > |t|	<.0001
Sign	M 60	Pr >= |M|	<.0001
Signed Rank	S 3630	Pr >= |S|	<.0001

Part (D)

Quantiles (Definition 5)

Quantile	Estimate
100% Max	74.0
99%	73.0
95%	65.0
90%	63.0
75% Q3	57.0
50% Median	49.5
25% Q1	44.0
10%	36.0
5%	30.0
1%	22.0
0% Min	21.0

Part (E)

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
21	1	68	88
22	13	69	105
24	31	70	85
25	2	73	113
30	12	74	91

Example 9.3 Three visual displays 2

The UNIVARIATE Procedure
Variable: reading

Part (F)

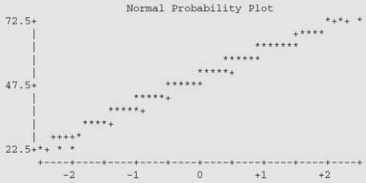
Stem	Leaf	#	Boxplot
7	034	3	|
6	5589	4	|
6	000001222333344	15	|
5	555666777777899999	18	+-----+
5	000111222344444444444	20	*--+--*
4	55555666667778888888999	24	| |
4	000002223344444444	17	+-----+
3	6666777779	10	|
3	00034	5	|
2	5	1	|
2	124	3	0

----+----+----+----+----+----
Multiply Stem.Leaf by 10**+1

[p. 173 ↓]

Part (G)

Normal Probability Plot



Parts (A) to **(E**) on page 1 should look familiar to you by now since they are identical to their counterparts in Output 9.1. **Parts (F**) and **(G**) on page 2 are new; they result from the option **PLOT.** The first plot is the stem-and-leaf plot. As the name implies, it consists of stems and several leaves. The stems are "tens" digits; thus, 7 = 70, 6 = 60, and so on. The leaves are the "units" digits. Hence, the first line shows three reading scores: 70, 73, and 74. The second line displays four more scores: 65, 65, 68, and 69. The bottom line shows the three lowest scores: 21, 22, and 24. The frequency of scores in each interval is indicated by the integers, in this case 3, 4,…, 3, under the "#" sign.

You may ask, "How about central tendency scores, such as mean or median? Where can I find those?" The answers are found in the plot next to the stem-and-leaf plot, namely, the **Boxplot.** Inside the square box in the **Boxplot**, the symbol "+" refers to the mean. Thus, we know that the mean is in the interval of 50 to 54. The precise value for the mean is **49.6**, reported under **Moments.** The dashed line (—) drawn inside the box corresponds to the median. It too lies in the interval from 50 to 54. The precise value for the median is **49.5**, found under **Quantiles.**

The top and bottom borders of the square box in Boxplot are defined by the 75th and 25th percentiles, respectively. The 75th percentile is located in the score interval of 55 to 59; the 25th percentile is located in the score interval of 40 to 44. Unfortunately, you cannot be more precise about these percentiles in a Boxplot. The precise information is available from **Part (D**) under **Q3** and **Q1.** The vertical lines extended from the borders out in both directions (north and south) are *whiskers.* Whiskers convey range information, because they are drawn as far as the largest or smallest data, but no more than 1.5 times the value of interquartile (= Q3 - Q1). If extreme scores are present, they are indicated by **"0"**, if they exceed 1.5 times the interquartile, or by "#" for extreme scores beyond three interquartiles.

Now let's turn to the Normal Probability Plot—**Part (G).** The plot shows a match between the sample distribution and the standard normal curve. The horizontal axis is the *z* score with a mean of 0 and a standard deviation of 1. The vertical axis is based on the sample data. Two symbols are used to draw the normal probability plot: pluses (**+**) and asterisks $(*)$. The straight **[p. 174 ↓]** line formed by the pluses (+) represents a theoretical reference line based on the normal curve. The raw data values are drawn as asterisks $(*)$ for which the vertical coordinate is the data value, and the horizontal is

the normalized *z* score. Thus, if data can be fitted to a normal curve, all asterisks are in the same position as the pluses. For a perfectly normal distribution, only the pluses (**+**) are shown. According to the current normal probability plot, data appear not to deviate noticeably from the normal curve.

Example 9.4 Test of Normality

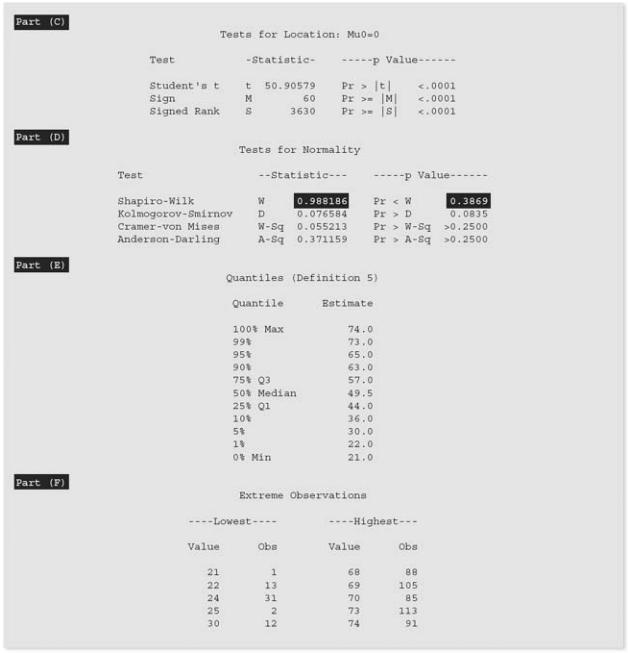
In addition to the normal probability plot, you may carry out a test of normality in the UNIVARIATE procedure. The test helps us to determine if the underlying population of scores is normally distributed. If the normality assumption holds, the sample data should preserve the normal distribution to a certain degree. For this reason, it is a good idea to carry out a test of normality before conducting a parametric procedure. The test is invoked by the option **NORMAL.** The following program tests both reading and punc score distributions to determine if they are normally distributed.

```
/* See Example 9.1 for the DATA step in creating the SAS data set 'achieve' */
TITLE 'Example 9.4 Test of normality';

PROC UNIVARIATE DATA=achieve NORMAL;
     VAR reading punc;
RUN;
```

Output 9.4 Test of Normality

```
                          Example 9.4 Test of normality                          1

                             The UNIVARIATE Procedure
                                Variable:  reading
Part (A)
                                    Moments

N                         120    Sum Weights               120
Mean                     49.6    Sum Observations         5952
Std Deviation       10.6734572    Variance            113.922689
Skewness            -0.3004658    Kurtosis             0.0484073
Uncorrected SS         308776    Corrected SS          13556.8
Coeff Variation     21.519067    Std Error Mean      0.97434888
Part (B)
                          Basic Statistical Measures

          Location                        Variability

     Mean     49.60000     Std Deviation          10.67346
     Median   49.50000     Variance              113.92269
     Mode     54.00000     Range                  53.00000
                           Interquartile Range    13.00000
```

[p. 175 ↓]



Part (C)

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	50.90579	Pr > |t|	<.0001
Sign	M	60	Pr >= |M|	<.0001
Signed Rank	S	3630	Pr >= |S|	<.0001

Part (D)

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.988186	Pr < W	0.3869
Kolmogorov-Smirnov	D	0.076584	Pr > D	0.0835
Cramer-von Mises	W-Sq	0.055213	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.371159	Pr > A-Sq	>0.2500

Part (E)

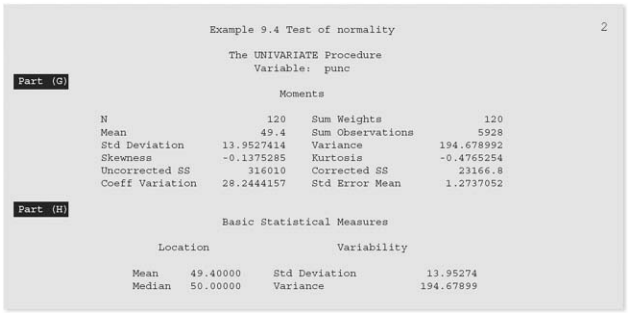
Quantiles (Definition 5)

Quantile	Estimate
100% Max	74.0
99%	73.0
95%	65.0
90%	63.0
75% Q3	57.0
50% Median	49.5
25% Q1	44.0
10%	36.0
5%	30.0
1%	22.0
0% Min	21.0

Part (F)

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
21	1	68	88
22	13	69	105
24	31	70	85
25	2	73	113
30	12	74	91



Example 9.4 Test of normality 2

The UNIVARIATE Procedure
Variable: punc

Part (G)

Moments

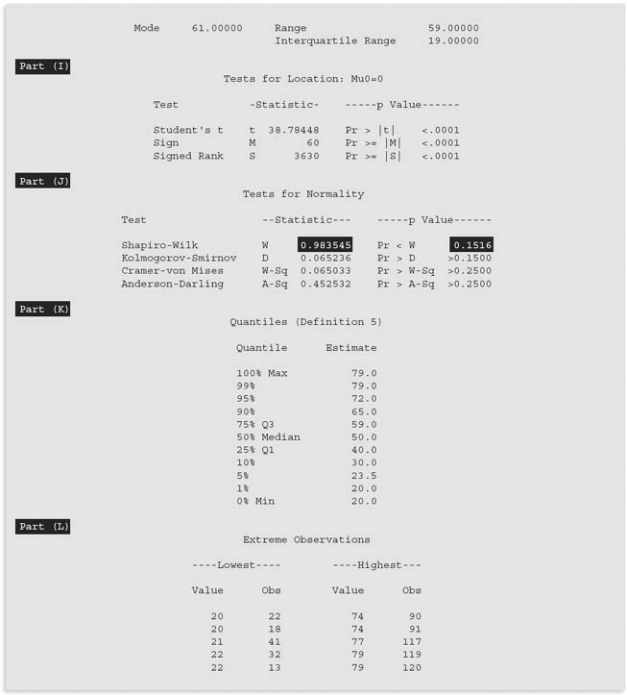
N	120	Sum Weights	120
Mean	49.4	Sum Observations	5928
Std Deviation	13.9527414	Variance	194.678992
Skewness	-0.1375285	Kurtosis	-0.4765254
Uncorrected SS	316010	Corrected SS	23166.8
Coeff Variation	28.2444157	Std Error Mean	1.2737052

Part (H)

Basic Statistical Measures

Location		Variability	
Mean	49.40000	Std Deviation	13.95274
Median	50.00000	Variance	194.67899

[p. 176 ↓]



```
Mode     61.00000     Range                    59.00000
                      Interquartile Range      19.00000

Part (I)

                Tests for Location: Mu0=0

    Test           -Statistic-    -----p Value------

    Student's t    t  38.78448    Pr > |t|    <.0001
    Sign           M        60    Pr >= |M|   <.0001
    Signed Rank    S      3630    Pr >= |S|   <.0001

Part (J)

                   Tests for Normality

Test                  --Statistic---    -----p Value------

Shapiro-Wilk          W      0.983545    Pr < W      0.1516
Kolmogorov-Smirnov    D      0.065236    Pr > D      >0.1500
Cramer-von Mises      W-Sq   0.065033    Pr > W-Sq   >0.2500
Anderson-Darling      A-Sq   0.452532    Pr > A-Sq   >0.2500

Part (K)

                Quantiles (Definition 5)

                 Quantile      Estimate

                 100% Max          79.0
                 99%               79.0
                 95%               72.0
                 90%               65.0
                 75% Q3            59.0
                 50% Median        50.0
                 25% Q1            40.0
                 10%               30.0
                 5%                23.5
                 1%                20.0
                 0% Min            20.0

Part (L)

                   Extreme Observations

              ----Lowest----        ----Highest---

              Value      Obs        Value      Obs

                 20       22           74       90
                 20       18           74       91
                 21       41           77      117
                 22       32           79      119
                 22       13           79      120
```

For purpose of explanation, Output 9.4 is divided into 12 parts: Parts (A) to (F) pertain to variable reading, whereas Parts (G) to (L) pertain to punc. In Part (D), the Tests for Normality section contains four tests for normality. The highlighted value (W = 0.988186) and its significance level (Pr < W = 0.3869) are the Shapiro-Wilk test of normality for reading. The large p level indicates that the null hypothesis of normally distributed reading scores cannot be rejected at α = 0.05. Is this good news or bad news? The answer depends on your reason for conducting the test. Recall that the **[p. 177 ↓]** null hypothesis states that sample data conform to a normal curve. Thus, if you are interested in verifying a normal assumption for the reading distribution, you should be delighted in such a large *p* level as it leads to the retention of the null hypothesis.

Similarly, the normality test of the punc distribution, in **Part (J**), yields a *W* statistic of 0.983545, significant at 0.1516. Again, you can conclude that the null hypothesis cannot be rejected at α = 0.05 and that the punc distribution is approximately normal.

If, in the future, this test is statistically significant at, say, $\alpha = 0.05$ and the null hypothesis of normality is rejected, you may wish to examine the three plots introduced in Example 9.3 to understand how and where the distribution deviates from the normal curve. For Output 9.4, we focus on the Shapiro-Wilk test because the sample size is less than 2,001. For samples larger than 2,001, we recommend the Kolmogorov-Smirnov test. Details about these four tests of normality are found in **Section 9.4: Tips.**

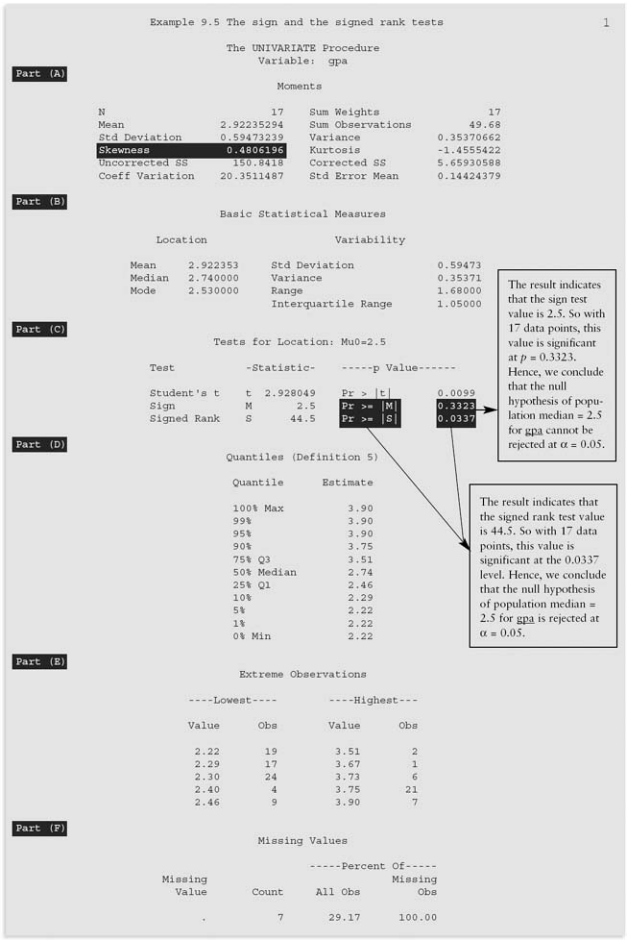
Example 9.5 The Sign Test and the Signed Rank Test for Small Samples

Sometimes you may wish to draw inferences about an underlying population average based on a small sample that may not be normally distributed. How do you draw conclusions based on small samples? In this example, we demonstrate solutions for this challenge. Let's suppose that we are interested in finding out if students' median grade point average (GPA) in the population is 2.5 on a 4-point scale. Seventeen students provided their GPA information in the raw data file mydata.dat. The null hypothesis of a 2.5 population median is specified by the option MU0= 2.5 in the PROC UNIVARIATE statement. If the null hypothesis is rejected, we know that these 17 students' gpa median is either significantly *higher* or *lower* than 2.5. Failing to reject the null hypothesis indicates that data might come from a population of students whose median GPA is 2.5.

```
/* The following bolded SAS statements establish the SAS data set 'mydata' */

DATA mydata;
     INFILE 'd:\data\mydata.dat';
     INPUT id $ sex $ age gpa critical polpref $ satv;
RUN;

TITLE 'Example 9.5 The sign and the signed rank tests';

PROC UNIVARIATE DATA=mydata MU0=2.5;
     VAR gpa;
RUN;
```

[p. 178 ↓]

Output 9.5 The Sign Test and the Signed Rank Test for Small Samples



[p. 179 ↓]

The sign test of gpa leads to the conclusion that this particular sample of students represented a population with a 2.5 median GPA because the null hypothesis of the population median = 2.5 cannot be rejected. The sign test carried out by the UNIVARIATE procedure is always a two-tailed test. If you are interested in a one-tailed test, simply half the output probability, printed next to (Pr >= IMI). For this example, a one-tailed probability would be 0.3323/2, or 0.16615.

According to the signed rank test, however, the same null hypothesis about the population median can be rejected at a significance level of 0.0337. This conclusion is inconsistent with that reached by the sign test. How can the inconsistency be resolved? Before you accept either test result, you should realize that there is an assumption associated with the signed rank test. It assumes that the underlying population is symmetric around its median. Failing to meet this assumption weakens the validity of the test. The symmetry of the gpa distribution can be inferred from the plot of its distribution and its skewness, which equals 0.4806196. This value suggests that the sample distribution is skewed to the right with a heavier density of low scores than high scores. It is therefore likely that the assumption of symmetry, required by the signed rank test, is violated. Thus, it is better to rely on the sign test result, rather than the signed rank result. In other words, the median gpa of this group of students is probably around 2.5.

A population median is the value above and below which 50% of all cases lie. In a skewed distribution, median and mean are two different values; this is the case with the gpa distribution. They are, however, identical in a symmetric distribution. Therefore, the test of the population median is the same as the test for the population mean for symmetric distributions.

Example 9.6 Comprehensive Descriptive Analysis of Subgroups

Examples 9.1 to 9.5 demonstrate the nuts and bolts of PROC UNIVARIATE for an entire data set. In this example, you will learn to replicate these helpful features for subsets of a sample. Specifically, the program below requests separate descriptions of men's and women's performance on satv (SATVerbal score) from the file mydata.dat. As this program illustrates, data need to be first sorted into two groups: “F” and “M”, or females and males. Afterward, separate analyses for each subgroup are specified by the BY sex; statement.

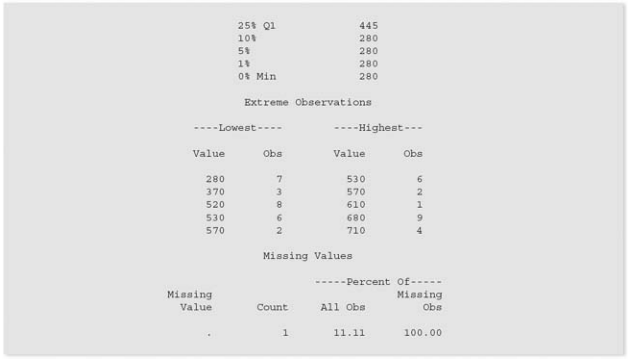
[p. 180 ↓]

```
/* See Example 9.5 for the DATA step in creating the SAS data set 'mydata' */

PROC SORT DATA=mydata;
     BY sex;
RUN;

TITLE 'Example 9.6 Comprehensive descriptive analysis of subgroups';

PROC UNIVARIATE DATA=mydata;
     VAR satv;
     BY sex;
RUN;
```

Output 9.6 Comprehensive Descriptive Analysis of Subgroups

```
Example 9.6 Comprehensive descriptive analysis of subgroups                1

---------------------------------------- sex=F ----------------------------------------

                          The UNIVARIATE Procedure
                              Variable:  satv

                                Moments

N                           8    Sum Weights                 8
Mean                    533.75    Sum Observations         4270
Std Deviation      146.866071    Variance            21569.6429
Skewness           -0.6997707    Kurtosis            -0.2403955
Uncorrected SS        2430100    Corrected SS          150987.5
Coeff Variation    27.5158916    Std Error Mean      51.9249974

                     Basic Statistical Measures

          Location                     Variability

     Mean     533.7500     Std Deviation          146.86607
     Median   550.0000     Variance                 21570
     Mode        .         Range                  430.00000
                           Interquartile Range    200.00000

                     Tests for Location: Mu0=0

     Test           -Statistic-     -----p Value------

     Student's t    t  10.27925     Pr > |t|    <.0001
     Sign           M          4     Pr >= |M|   0.0078
     Signed Rank    S         18     Pr >= |S|   0.0078


                     Quantiles (Definition 5)

                      Quantile      Estimate

                      100% Max           710
                      99%                710
                      95%                710
                      90%                710
                      75% Q3             645
                      50% Median         550
```

[p. 181 ↓]



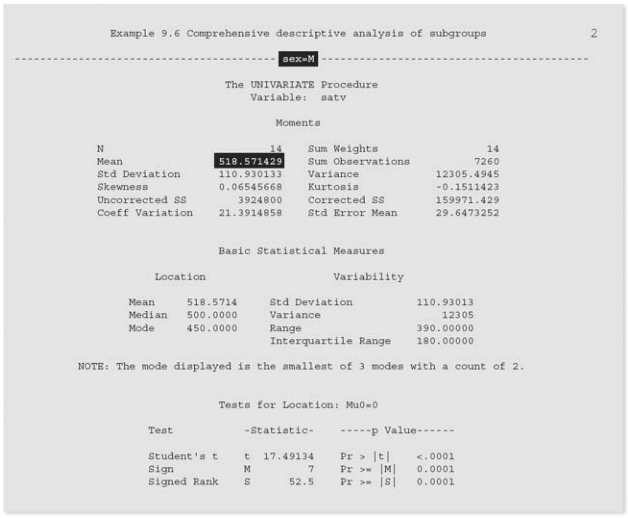
25% Q1	445
10%	280
5%	280
1%	280
0% Min	280

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
280	7	530	6
370	3	570	2
520	8	610	1
530	6	680	9
570	2	710	4

Missing Values

Missing Value	Count	-----Percent Of----- All Obs	Missing Obs
.	1	11.11	100.00



Example 9.6 Comprehensive descriptive analysis of subgroups 2

sex=M

The UNIVARIATE Procedure
Variable: satv

Moments

N	14	Sum Weights	14
Mean	518.571429	Sum Observations	7260
Std Deviation	110.930133	Variance	12305.4945
Skewness	0.06545668	Kurtosis	-0.1511423
Uncorrected SS	3924800	Corrected SS	159971.429
Coeff Variation	21.3914858	Std Error Mean	29.6473252

Basic Statistical Measures

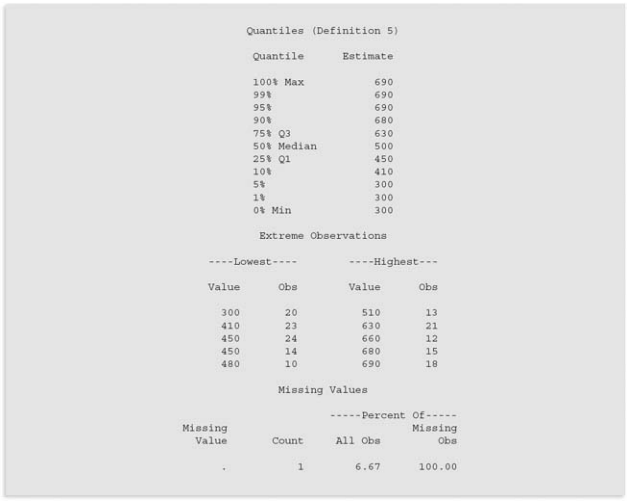
Location		Variability	
Mean	518.5714	Std Deviation	110.93013
Median	500.0000	Variance	12305
Mode	450.0000	Range	390.00000
		Interquartile Range	180.00000

NOTE: The mode displayed is the smallest of 3 modes with a count of 2.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	17.49134	Pr > |t|	<.0001
Sign	M	7	Pr >= |M|	0.0001
Signed Rank	S	52.5	Pr >= |S|	0.0001

[p. 182 ↓]



Quantiles (Definition 5)

Quantile	Estimate
100% Max	690
99%	690
95%	690
90%	680
75% Q3	630
50% Median	500
25% Q1	450
10%	410
5%	300
1%	300
0% Min	300

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
300	20	510	13
410	23	630	21
450	24	660	12
450	14	680	15
480	10	690	18

Missing Values

Missing Value	Count	-----Percent Of----- All Obs	Missing Obs
.	1	6.67	100.00

This output shows that women and men are similar in average satv scores. For women, the average is 533.75, based on 8 women. For men, the average is 518.571429, based on 14 men. As for variance (or standard deviation), the women's group seemed to have greater variability than the men's group. The women's verbal scores are more skewed than the men's. Both curves are about equally flat. These indices are within the ballpark of national SAT-Verbal scores. Other descriptive statistics, such as the median, the mode, and the interquartile range, also confirm that the sample data are similar to the national norm.

It is important to point out that the **Mode** value (missing expressed as a period, ".") is misleading for the women's group because all raw scores for women appear only once in the data. None of these scores qualifies to be a mode. Therefore, PROC UNIVARIATE cannot compute the mode for women and reports the mode to be a missing score. When the sample distribution is multimodal, namely, having more than one mode, SAS reports only the lowest modal value. To compute all modal values, you must specify the MODES option in the PROC UNIVARIATE statement.

[p. 183 ↓]

Example 9.7 Creating an Output Data Set in PROC UNIVARIATE

This example demonstrates how to save analysis results from PROC UNIVARIATE into a SAS data set for subsequent use. Specifically, the **OUTPUT** statement is illustrated for its versatile utility and syntax. The leading word OUTPUT is followed by an output data set name **(OUT=myout**) and several statistical indices, to be included in this output data set. Thus, **MEAN=meanage meansatv** names the mean of age and satv, respectively, and includes them in myout. Likewise, **STD=stdage stdsatv** names the standard deviation of age and satv, respectively, and saves them into myout as well. The keyword **PCTLPTS=20 80** requests the 20th and 80th percentiles for both age and satv. The next and final keyword **PCTLPRE=age satv** defines a prefix name for the percentiles requested. The use of this keyword will become clearer when you read the explanation of Output 9.7. Once the output data set, myout, is established, the PRINT procedure displays its content.

```
/* See Example 9.5 for the DATA step in creating the SAS data set 'mydata' */

TITLE 'Example 9.7 Creating an output data set in PROC UNIVARIATE';

PROC UNIVARIATE DATA=mydata NOPRINT;
     VAR age satv;
     OUTPUT OUT=myout MEAN=meanage meansatv STD=stdage stdsatv PCTLPTS= 20 80 PCTLPRE= age satv;
RUN;

PROC PRINT DATA=myout;
RUN;
```

Output 9.7 Creating an Output Data Set in PROC UNIVARIATE

```
                Example 9.7 Creating an output data set in PROC UNIVARIATE              1

Obs    meanage    meansatv     stdage    stdsatv    age20    age80    satv20                1

 1     24.1176     524.091    2.86972    121.916      22       27       450       660
```

The information displayed in Output 9.7 is precisely what was asked for in the program. Thus, meanage (24.1176) is the mean of the age variable and meansatv (524.091) is the mean of satv; stdage and stdsatv are standard deviations of age and satv, respectively. The next pair of values (age20 and age80) corresponds to the 20th and 80th percentiles of the age variable. Similarly, satv20 and satv80 represent the 20th and 80th percentiles of the satv variable, respectively. These percentiles reveal that neither distribution is symmetric around its mean. The satv distribution is skewed more to the right than the age distribution because satv80 (= 660) is further away than **[p. 184 ↓]** satv20 (= 450) from their mean (= 524.091), compared with age80 (= 27) and age20 (= 22) relative to their mean (= 24.1176).

Other descriptive indices, besides mean, standard deviation, and percentiles, can be named and saved into an output data set as well. For details, read the next section on general syntax of the OUTPUT statement.

9.3 How to Write the PROC UNIVARIATE Codes

The following PROC UNIVARIATE statements were illustrated in one or more of the seven examples in **Section 9.2.** Their syntax is summarized as follows:

PROC	UNIVARIATE	DATA= sas_dataset_name <options>;
	VAR	variable(s);
	BY	classification_variable(s);
	WEIGHT	weighting_variable;
	FREQ	frequency_variable;
	OUTPUT	<OUT= sas_dataset_name>
		<keywords for outputted statistic>
		<PCTLPTS= percentiles PCTLPRE = prefix for percentiles
		PCTLNAMES = suffix for percentiles>;

The **PROC UNIVARIATE** statement initiates the procedure and specifies a data set to be analyzed. Five useful options may be specified:

ROUND= a round-off unit (such as 0.01)	This option saves memory space while processing the SAS program because it truncates, or rounds off, variable values according to the round-off unit (say, the second decimal place) prior to computing statistics.
VARDEF=DF	This option specifies that degrees of freedom is to be used in calculating the sample variance and standard deviation. Other choices besides degrees of freedom are possible: sample size (VARDEF=N), sum of weights (VARDEF=WGT), or sum of weights minus one (VARDEF=WDF). The default is degrees of freedom (DF=N - 1).
PLOT (or PLOTS)	This option produces three visual displays of data: the stem-and-leaf plot, the box plot, and the normal probability plot. For illustration, refer back to Example 9.3.
NORMAL (or NORMALTEST)	This option requests a statistical test of the normal population hypothesis. For illustration, refer back to Example 9.4.
MU0= a list of numerical values (such as 0 2.5 4)	This option specifies the hypothetical population parameter for the t test, the sign test, and the signed rank test.
FREQ	This option requests a frequency table to be compiled; in the frequency table, variable values, frequencies, percentages, and cumulative percentages are tabulated.

MODES	This option requests a list of all modes when the sample distribution is multimodal.
NOPRINT	This option suppresses the display of the output.

The second statement, **VAR**, is used to list one or more variables in the data set for which analyses are sought. The third statement, **BY**, is used to perform subgroup analyses. It divides the data set into subgroups according to diverse values of the BY variable. Within each subgroup, the same advanced descriptive analyses and the test of normality are conducted. If more than one BY variable is listed, all possible combinations of the BY variables' values are used in dividing up the data set. Be sure to presort the data set in the ascending order of all BY variables, if the BY statement is included in the UNIVARIATE procedure. Presorting a data set can be accomplished with the SORT procedure.

The fourth statement, **WEIGHT**, is used to apply differential weights to data values. All weights must be nonnegative. Zero or positive fractional numbers are allowed; negative or missing values are substituted by a zero. When a zero weight is applied to a variable, the variable's summary statistics, such as mean, variance, standard deviation, and so on, will be missing in the output; yet the minimum and the maximum will continue to be the smallest and the largest actual data values, respectively.

The fifth statement, **FREQ**, is needed for grouped data. By grouped data, we mean data that have been grouped together by identical data values or into intervals, represented by a midpoint and frequency of occurrences in that interval. To describe a grouped data set, two pieces of information are needed: (1) the data value or the midpoint of an interval and (2) its frequency count. In SAS terminology, the frequency count is the value of the FREQ variable.

The last statement, **OUTPUT**, is used to create an output data set so that results from PROC UNIVARIATE can be saved for subsequent analysis by another procedure. Immediately after the *OUT=sas_dataset_name*, one or **[p. 186 ↓]** more statistics may

be requested by their keywords. The list below presents keywords you may specify for inclusion in an output data set:

N=	no. of valid data points
SKEWNESS=	the skewness index
KURTOSIS=	the kurtosis index
MEAN=	the arithmetic average
MEDIAN=	the median or the 50th percentile or Q2
MODE=	the modal value
STD=	the standard deviation
VAR=	the variance
STDMEAN=	the standard error of the mean
MAX=	the maximum
MIN=	the minimum
RANGE=	the range
Q1=	the first quartile
Q3=	the third quartile
P1=	the 1st percentile
P5=	the 5th percentile
P10=	the 10th percentile
P90=	the 90th percentile
P95=	the 95th percentile
P99=	the 99th percentile
T=	the t test value
PROBT=	the significance level of the above t test

MSIGN=	the sign test value
PROBM=	the significance level of the above sign test
SIGNRANK=	the signed rank test value
PROBS=	the significance level of the above signed rank test
NORMALTEST=	the statistic for the normality test. If the sample size is less than or equal to 2,000, this statistic is the Shapiro- Wilk test statistic. Otherwise, PROC UNIVARIATE calculates the Kolmogorov-Smirnov statistic.
PROBN=	the significance level of the above normal test

How do you name the statistic? It is easy; simply create a name that jointly represents the statistic and the variable. For instance, combine mean with age or mean with satv and list both after the keyword, such as **MEAN=meanage meansatv.**

As for percentiles, they are requested via the option **PCTLPTS=.** This option is followed by the percentage point(s) corresponding to the specific percentile. If more than one variable's percentiles are requested, you should use the option **PCTLPRE=** to define a prefix name for two or more variables' percentiles. For **[p. 187 ↓]** an illustration of these two options, refer back to Example 9.7. Sometimes there may be a need to request an unusual percentile, such as 58.9. This percentile is acceptable to PROC UNIVARIATE. It will be included in the output under the name of, say, **age58_9** or **satv58_9**, for age and satv, respectively.

Once an output data set is created, it is recommended that its content be displayed by PROC PRINT, preferably right after the data set is created. A responsible data analyst should examine the content of each data set at least once to ensure that no mistake is made in naming variables or choosing the kind of statistic for the output data set.

9.4 Tips

Missing data means that you do not have any information about an observation on a particular variable. Missing data should be treated differently than 0. A score of 0 means that the information is available, and the score happens to be exactly 0, whereas missing data means you have no information.

The count of missing observations on each variable is automatically printed under headings such as **Missing Value, Count, Percent of All Obs**, and **Percent of Missing Obs.** If an observation has missing information on one or more variables, the observation is excluded from the calculation of descriptive statistics and also from any statistical test performed on these variables. The observation itself is not removed from the SAS data set though; it is simply discounted from analyses.

As Example 9.4 demonstrates, the test of normality is carried out by four statistical tests in the UNIVARIATE procedure: the Shapiro-Wilk test, the Kolmogorov-Smirnov test, the Anderson-Darling test, and the Cramér-von Mises test. Each test has its own special features and sensitivity. The default is the Shapiro-Wilk test, if the sample size is 2,000 or less. The test statistic is denoted by *W* and ranges from 0 to 1. Small values of *W* lead to the rejection of the null hypothesis of normality. The sampling distribution of *W* is highly skewed to the left. Hence, values above 0.90 may still lead to the rejection of the null hypothesis.

The Kolmogorov-Smirnov test *D* is calculated on the basis of the largest vertical differences between the sample distribution (also called the empirical distribution) and the normal distribution. If the sample size is larger than 2,000, the UNIVARIATE procedure defaults to this test statistic.

Both the Anderson-Darling test and the Cramér-von Mises test are computed on the basis of squared difference between the sample distribution and the normal distribution. The Anderson-Darling test weighs the squared differences by the reciprocal of the product of the cumulative probability multiplied by (1 - the cumulative probability), over the entire range of the score scale. The Cramér-von Mises test does not apply any weight to the squared differences.

[p. 188 ↓]

When conducting the test of normality, you need to be aware of the power of such a test, and power is directly related to the sample size. If the sample size is large, in hundreds or thousands, the statistical test is so powerful that a minor departure from normality can lead to the rejection of the null hypothesis. Conversely, if the sample size is small, in the 10s or below 10, the test may not be powerful enough to detect serious departure from normality. In both cases, it is recommended that you rely on additional information such as the skewness, kurtosis, normal probability plot, frequency distribution, and so on, to make an informed judgment about the violation of the normality assumption. For small samples, you may wish to raise the level of significance (i.e., α) in order to compensate for the reduced power associated with small sample tests.

The Output Delivery System (ODS) in SAS allows you to (a) select part(s) of the output to be displayed, (b) export part(s) of the output into data set(s), and (c) save the output in formats other than the standard SAS output. To use the ODS, you need to know ODS table names corresponding to various portions of the output. Table 9.2 presents selected ODS table names for the UNIVARIATE procedure and their descriptions.

Table 9.2 Selected ODS Table Names and Descriptions for the UNIVARIATE Procedure

ODS Table Name	Description	Option in the PROC UNIVARIATE Statement
Moments	Sample moments	(default)
BasicMeasures	Basic statistical measures of location and variability	(default)
TestsForLocation	Tests for location	(default)
Quantiles	Quantiles	(default)
ExtremeObs	Extreme observations	(default)
TestsForNormality	Tests for normality	NORMAL

Frequencies	Frequency table	FREQ
Plots	The stem-and-leaf plot, the box plot, and the normal probability plot	PLOT
Modes	Modes	MODES

[p. 189 ↓]

These ODS table names and related details are tracked in the Log window if you ask for them with the ODS TRACE ON statement in the SAS program. You may turn off this tracking feature with the ODS TRACE OFF statement, also in the SAS program.

```
ODS TRACE ON;
PROC UNIVARIATE DATA=achieve;
     VAR reading;
RUN;
ODS TRACE OFF;
RUN;
```

After executing the program, the following will appear in the Log window listing all ODS table names for the PROC UNIVARIATE output:

```
Output Added:
-------------
Name:       Moments
Label:      Moments
Template:   base.univariate.Moments
Path:       Univariate.reading.Moments
-------------

Output Added:
-------------
Name:       BasicMeasures
Label:      Basic Measures of Location and Variability
Template:   base.univariate.Measures
Path:       Univariate.reading.BasicMeasures
-------------

Output Added:
-------------
Name:       TestsForLocation
Label:      Tests For Location
Template:   base.univariate.Location
Path:       Univariate.reading.TestsForLocation
-------------

Output Added:
-------------
Name:       Quantiles
Label:      Quantiles
Template:   base.univariate.Quantiles
Path:       Univariate.reading.Quantiles
-------------

Output Added:
-------------
Name:       ExtremeObs
Label:      Extreme Observations
Template:   base.univariate.ExtObs
Path:       Univariate.reading.ExtremeObs
-------------
```

[p. 190 ↓]

Based on the list of ODS table names, you may select certain results to be displayed in the Output window. For example, the following program selects **Part (A**) of Example 9.1 to be included in the output:

```
ODS SELECT Univariate.reading.Moments;
PROC UNIVARIATE DATA=achieve ROUND=0.01;
     VAR reading;
RUN;
```

Likewise, you may select certain result(s) to be exported as a SAS data set. For example, the following program exports **Part (A**) of Example 9.1 to the SAS data set descriptive:

```
ODS OUTPUT Moments = descriptive;
PROC UNIVARIATE DATA=achieve ROUND=0.01;
     VAR reading;
RUN;
```

Furthermore, you may select certain results to be stored in file formats other than the SAS standard output. For example, the following program saves the output of Example 9.1 in HTML format in its default style:

```
ODS HTML BODY = 'd:\result\Example9_1Body.html'
         CONTENTS = 'd:\result\Example9_1TOC.html'
         PAGE = 'd:\result\Example9_1Page.html'
         FRAME = 'd:\result\Example9_1Frame.html';
PROC UNIVARIATE DATA=achieve ROUND=0.01;
     VAR reading;
RUN;
ODS HTML CLOSE;
RUN;
```

For additional information about the ODS feature, consult with *SAS Output Delivery System: User's Guide* (SAS Institute Inc., 2006c) and *Base SAS 9.1.3 Procedures Guide* (SAS Institute Inc., 2006a) or the online documentation at http://www.sas.com.

9.5 Summary

This chapter covers the most comprehensive descriptive analysis procedure in SAS, that is, PROC UNIVARIATE. The UNIVARIATE procedure may be applied for descriptive as well as inferential analyses of the data. For descriptive analyses, PROC UNIVARIATE computes indices for central **[p. 191 ↓]** tendency, variability, skewness, kurtosis, extreme scores, quartiles, and selected percentiles.

For inferential analyses, PROC UNIVARIATE performs the one-sample *t* test, tests of normality, the sign test, and the signed rank test. Each is conducted as a two-tailed test. In addition to the test of normality, PROC UNIVARIATE also draws the normal probability plot, which can serve as a visual device for making inferences about the approximation of the sample distribution to normal.

Two other graphical representations of data include the stem-and-leaf plot and the box plot. Both can be examined to determine if outliers exist, or the shape implies a particular trend or pattern in data. Most of these handy features are not available in PROC MEANS, which is simpler to program and its output less complicated. If you need to learn more about the UNIVARIATE procedure, consult with *Base SAS 9.1.3 Procedures Guide* (SAS Institute Inc., 2006a) or the online documentation available

from http://www.sas.com. Now let's see how much material you have mastered from this chapter.

9.6 Exercises

VARIABLES: name	sex	age	room	s.e.s	income (in $1000)
ANDY	M	28	122	H	9.6
SOPHIA	F	42	412	M	7.4
LOUIS	M	34	213	L	2.3
LANDERS	M	40	216	M	5.8
TED	M	68	101	H	8.8
DICKENS	M	37	135	M	6.3
RUTH	F	39	430	L	1.3
CHARLIE	M	54	222	M	4.7
MICHAEL	M	25	118	M	6.0
WOLF	M	51	104	H	8.4
ANNE	F	58	404	H	10.2
REBECCA	F	43	423	M	7.1
DAVID	M	47	117	M	5.2
RICHARD	M	28	240	H	9.2
SAM	M	36	231	M	6.4
TINA	F	31	302	M	4.3
PETER	M	65	108	L	3.3
SHEILA	F	24	336	M	5.7
TIM	M	27	115	H	8.2
LINDA	F	20	425	M	5.5

9.7 Answers to Exercises

10.4135/9781452230146.n9