

MODULE 5

Identifying and Addressing Outliers

DATA FILES FOR THIS MODULE

module5_jobsatis.sav
module5_jobsatis_final.sav

KEY LEARNING OBJECTIVES

The student will learn to

- use summary descriptive statistics (e.g., skewness and kurtosis) to help determine the shape of a continuous variable's distribution
- create and interpret stem-and-leaf plots and boxplots to help determine the shape of a distribution and identify outliers
- create, interpret, and compare a set of boxplots for a continuous variable by groups of a categorical variable
- conduct and compare t -tests on data with outliers and data without outliers to determine whether the outliers have an impact on results.

A. Description of Researcher's Study

This module is a continuation of Dr. Mendoza's collaboration with a company that is conducting experimental research to improve the level of job satisfaction among its employees. In Module 4, he examined the distributions for descriptive variables in his current data file as well as scores on an assessment of knowledge and aptitude. Using frequency tables, crosstabs, and histograms to display information about the distributions, he made modifications to condense the number of groups for the nominal and ordinal variables and to change the measurement scale of one variable from ratio (continuous) to ordinal (categorical).

Here, Dr. Mendoza will examine the final two continuous variables in the data file, which contain scores on two instruments designed to measure motivation and job satisfaction prior to the beginning of the 3-month study. This module demonstrates how he used boxplots to look at the shape of the distributions, identify potential outliers, and decide how outliers will be handled when analyzing the data. Upon completion of the study, the instruments will be administered again to measure potential change in the constructs for employees who received the job enrichment program and to compare results to those who did not receive the program.

△ B. A Look at the Data

The data file used in this module (*module5_jobsatis.sav*) contains all nine original variables in Dr. Mendoza's initial file plus the newly modified variables he created in Module 4. There are 265 cases (participants) and a total of 12 variables. As before, each case represents one employee who agreed to be in the study. Employees are in one of seven geographic locations of the company. A total of 127 participants are receiving the job enrichment program, and the remaining 138 participants are in the control group. Table 5.1 lists each variable and its description.

Table 5.1 Variables in the Job Satisfaction Data File

Variable Name	Description
id	numerical identifier for each participant 1 to 265
location	geographic location of the company in which the participant is employed, coded 1 to 7
program	0 = no (participant is not receiving the program, he or she is in the control group); 1 = yes (participant is receiving the program)
size	total number of employees in each geographic location of the company across all its sectors. This variable is a constant value for participants within the same location.
yrs_exper	total number of years of experience within the company for each participant. Codes for the categories are 2 = 0 to 2 years of experience, 4 = 3 to 4 years of experience, 6 = 5 to 6 years of experience, 8 = 7 to 8 years of experience, 10 = 9 to 10 years of experience, and so on until the final code of 32 = 31 years of experience or more.

Variable Name	Description
reprimand	total number of reprimands received by each employee during his or her total years of experience with the company
aptitude	score on an aptitude instrument prior to implementing the study
motivation	score on a motivation survey prior to implementing the study
job_satis	score on a job satisfaction survey prior to implementing the study
location_r	revised variable with only five codes (1, 3, 6, 7, and 9). Due to the small number of participants in locations 2, 4, and 5, these locations were combined and assigned a new code of 9.
yrs_exper_r	revised variable with only six codes (2, 4, 6, 8, 10, and 30). The first original five codes remained unchanged. Due to small sample sizes in the remaining codes, they were combined into a new code of 30 that represents participants with 11 or more years of experience.
reprimand_r	revised variable is now an ordinal variable with three levels: 0 reprimands, 1 reprimand, and 2 or more reprimands. Original values of 2, 3, and beyond were collapsed into one level due to small sample sizes.

C. Planning and Decision Making ▲

To examine prescores on the two constructs of motivation and job satisfaction, Dr. Mendoza decided to create boxplots in SPSS. There are several beneficial features of this type of graphic display. First, it allows you to view aspects of the distribution in a way that histograms do not. The length of the “box” spans the middle 50% of the values, that is, from the 1st quartile (25th percentile) to the 3rd quartile (75th percentile), and the median appears as a solid line in the box. In a distribution with no outliers, the length of the two “whiskers” represent the bottom 25% of values and the top 25% of values. When a distribution is approximately normal, the median will be in the center of the “box” and the two “whiskers” will be equal in length. The extent to which this does not occur indicates potential positive or negative skewness or kurtosis.

A second beneficial feature of the boxplot over the histogram is that it can identify potential outliers. Outliers are values at the lower or

**1 Dr. Mendoza says: “There is not a hard and fast rule for identifying outliers in a distribution. SPSS uses one particular method, but others exist. For example, standardized values can be used with a general guideline that absolute z values larger than 3 are considered to be outliers. However, for large samples, some statisticians use a cutoff z value of 4 or greater, and for small sample sizes, a cutoff of 2.5” (Stevens, 2009).

upper end that lie apart from the distribution. These values are identified on the box plot as cases below or above the end of each “whisker.” More specifically, SPSS identifies outliers as cases that fall more than 1.5 box lengths from the lower or upper hinge of the box. The box length is sometimes called the “hsread” and is defined as the distance from one hinge of the box to the other hinge. It is also called the interquartile range. SPSS further distinguishes “extreme” outliers by identifying values more than 3 box lengths from either hinge. ****Dr. Mendoza (1)**

Statistical inferential tests can be quite sensitive to outliers, often because the calculations rely on squared deviations from the mean. One or two values that are far from the mean can alter the results considerably. Therefore, if outliers are identified, Dr. Mendoza must decide how to handle them. First, he will go back to the data collection instrumentation to determine whether the outlier was due to a data entry error or an instrumentation error. The former can be corrected, and the latter probably should be deleted. However, if the outlier was not due to one of these reasons and was an actual value obtained from the participant, then he has a few options. The most undesirable option is to delete the case from further analysis. This is not the best solution because the value is a legitimate case in the data file, and with large samples, it can be expected that a few outliers may occur and probably will not greatly impact results. Another option is to conduct data analysis with and without the outlier(s) and compare the two outcomes. If results are the same, then the outlier(s) did not have a great influence in the distribution of the variable. If results are not the same, both outcomes can be reported. A third option is to transform the variable and hopefully reduce the influence of the outlier(s). Finally, outliers could be recoded into the lowest (or highest) value that is not determined to be an outlier by SPSS (or any other method that is used).

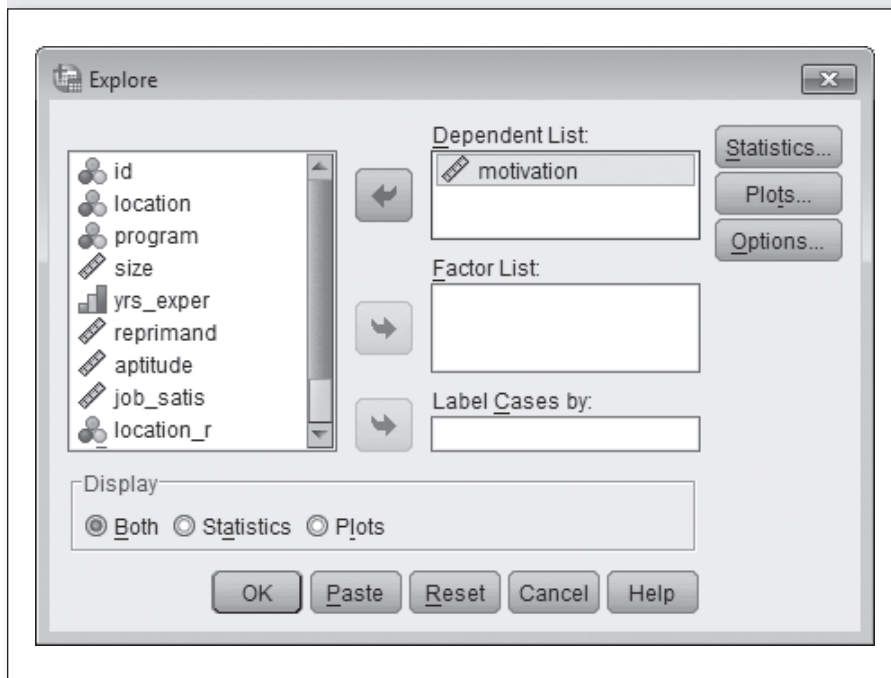
Dr. Mendoza will obtain the boxplots for *motivation* and *job_satis* using the **Explore procedure**. Although there are a few other ways to get boxplots in SPSS, he chose this procedure because it produces descriptive statistics such as skewness and kurtosis as well as a stem-and-leaf plot, which is another type of visual display of data.

D. Using SPSS to Address Issues and Prepare Data ▲

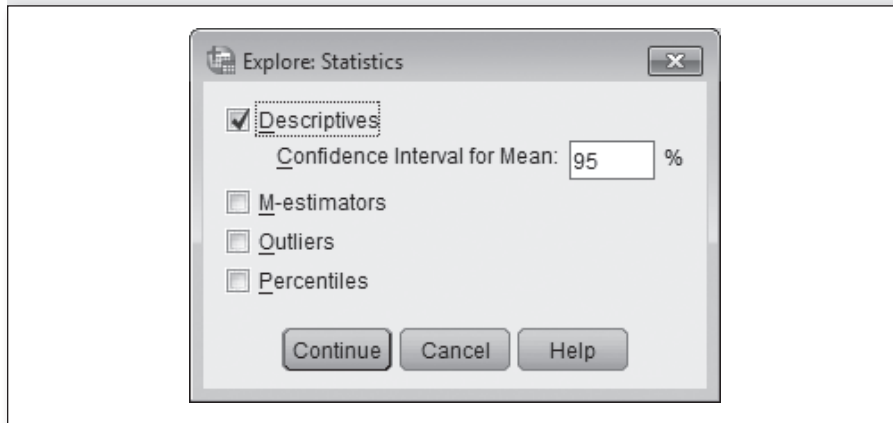
Motivation

A visual scroll through the data file is sometimes the first indication a researcher has that potential outliers may exist. For *motivation*, Dr. Mendoza noticed that a few low scores seemed to stand apart from the rest of the distribution. To help him determine whether these low values are actually outliers, he obtained a boxplot under the **Analyze** menu. Selecting **Descriptive Statistics** and **Explore** produced the dialog box shown in Figure 5.1. He placed *motivation* in the **Dependent List** box.

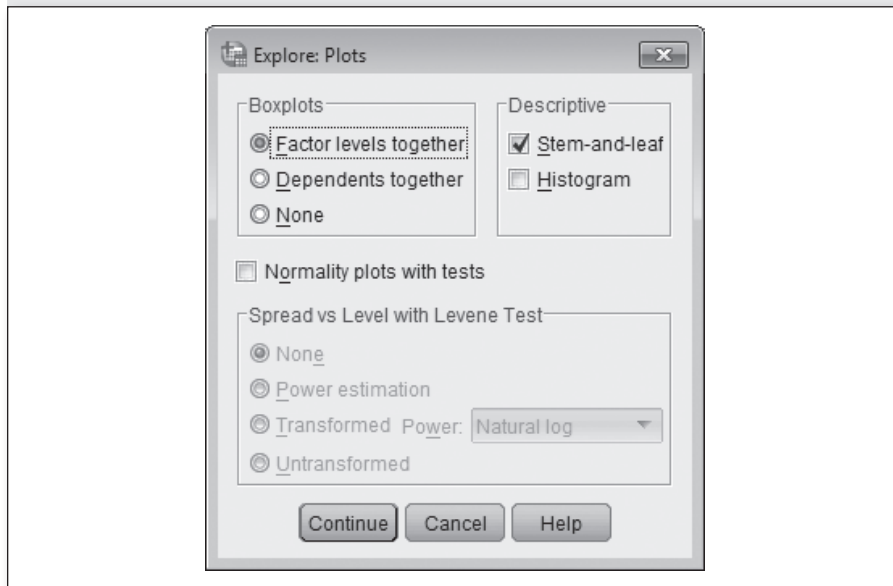
Figure 5.1 Analyze → Descriptive Statistics → Explore



Dr. Mendoza kept the default settings under the **Statistics** and **Plots** buttons as is. The **Options** button offers different ways to treat missing values, but he has none in his study, so he does not need the options. As shown in Figure 5.2, the default for **Statistics** is **Descriptives**. This will produce a variety of descriptive summary statistics for *motivation*, including the skewness and kurtosis values.

Figure 5.2 Statistics in Explore

For **Plots**, the default is to produce a boxplot and a stem-and-leaf plot, as shown in Figure 5.3.

Figure 5.3 Plots in Explore

After he clicked **OK** in the **Explore** dialog box, Dr. Mendoza obtained output that includes a table of values, a stem-and-leaf plot, and a boxplot. Table 5.2 displays the summary descriptive statistics. He noticed that the mean, median, and trimmed mean are nearly identical. This is one indication

that the distribution is not skewed in one direction or another. To examine skewness and kurtosis, Dr. Mendoza used the standard errors provided for each value in order to obtain standardized values for each statistic. Dividing skewness by the standard error ($-.602$ divided by $.150$) yields a standardized value of -4.01 , which does indicate a somewhat negatively skewed distribution. In a similar fashion, he divided kurtosis by its standard error (1.891 divided by $.298$) to obtain a standardized value of 6.35 , which indicates a peaked, or slender and narrow, distribution. ****SPSS Tip 1** Dr. Mendoza kept these values in mind as he looked at the next section of output from the **Explore procedure**.

SPSS Tip 1: A distribution with a considerably high positive kurtosis value is called leptokurtic, meaning that it is slender and narrow. A distribution with a considerably high negative kurtosis value is called platykurtic, meaning flat or broad. Low absolute values close to 0 for kurtosis are said to be mesokurtic or intermediate.

Table 5.2 First Portion of SPSS Explore Output: Summary Statistics for Motivation

Descriptives			
		Statistic	Std. Error
Motivation	Mean	20.02	.221
	95% Confidence Interval for Lower Bound	19.58	
	Mean Upper Bound	20.45	
	5% Trimmed Mean	20.11	
	Median	20.00	
	Variance	12.932	
	Std. Deviation	3.596	
	Minimum	4	
	Maximum	30	
	Range	26	
	Interquartile Range	4	
	Skewness	-.602	.150
	Kurtosis	1.891	.298

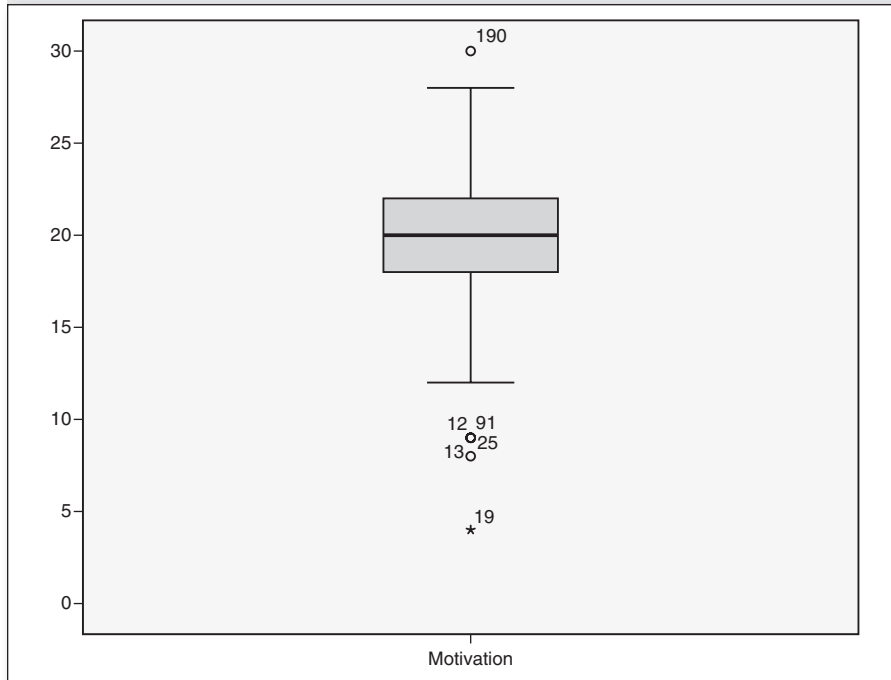
Figure 5.4 displays the stem-and-leaf plot. The stems represent the two-digit data values for *motivation*. Each leaf represents a case with that particular data value. The frequency column represents the total number of cases

Finally, the boxplot in Figure 5.5 is produced. Dr. Mendoza saw the five outliers at the lower end of the motivation scale and the one outlier at the upper end. The values next to each represent the case numbers. ****SPSS Tip 3** Five of the six values are denoted by a circle. Recall earlier (from Section C in this module) that SPSS makes a distinction between outliers that are more than 1.5 box lengths from one hinge of the box (using a circle) and outliers that are more than 3 box lengths from a hinge (using an asterisk).

The lowest value came from case/id number 19. It had a value of 4. Four additional values were also identified as outliers: case/id numbers 25, 13, 91, and 12.

SPSS Tip 3: Prior to running the Explore procedure, Dr. Mendoza's file had been sorted by id in ascending order, which is a sequential match with the case numbers in SPSS from 1 to 265. If the file was sorted in a different way, then the SPSS case number shown on the boxplot would not coincide with the id variable in the data file. This is not a problem, but when going back to the data file to examine outlying cases, care must be taken to ensure that you are looking at the correct case and not mistaking the value in the id variable for the actual SPSS case number.

Figure 5.5 Third Portion of SPSS Explore Output: Boxplot for Motivation



Their values for *motivation* were 8, 9, 9, and 9, respectively. Although it was not initially apparent to Dr. Mendoza when he visually scrolled through his data file, the highest value of 30 for case number 190 was also identified as an outlier.

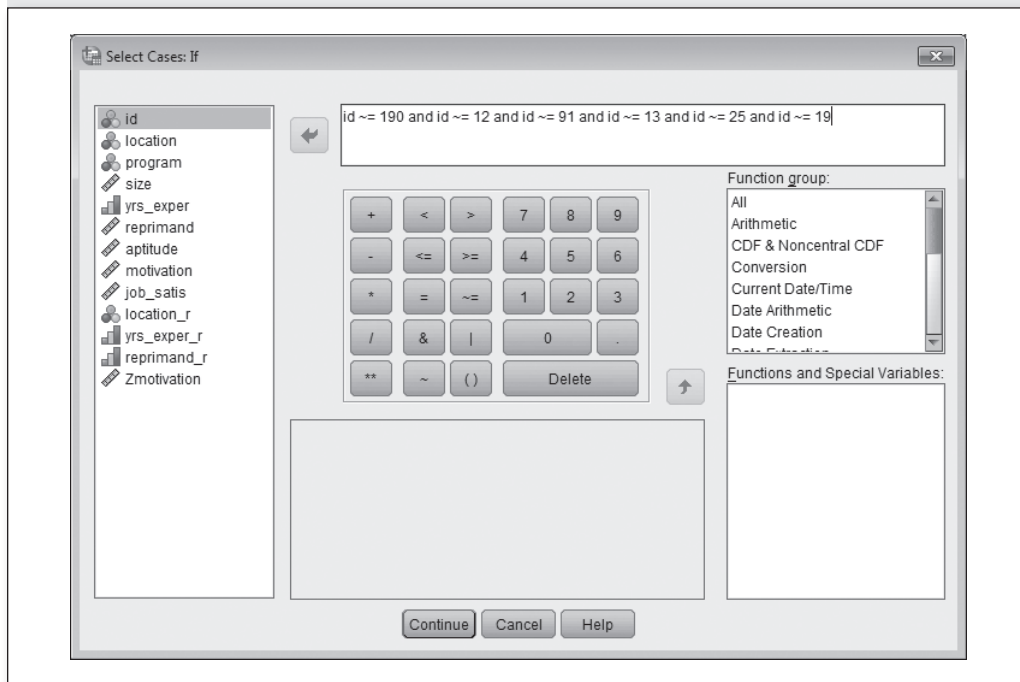
To decide how to handle the outliers, Dr. Mendoza's first step was to go back to the data collection materials and ensure that there were no data entry or instrument errors. There were not, so this indicates that the values are legitimate. He would rather not delete the cases because the sample is somewhat large, and it is to be expected that a few values will appear in the far lower and upper ends of the distribution. Although these cases were identified as outliers according to SPSS specifications, he wanted to check their standardized values. Under the **Analyze menu**, he selected **Descriptive Statistics** and **Descriptives**, put *motivation* in the **Variable(s)** box, and clicked the **Save standardized values as variables** box. This created a new variable, *Zmotivation*, that contains the corresponding *z* values for all values in *motivation*. Using a cutoff value of 4, only one of the cases was identified as an outlier. Case number 19 had a score of 4 for *motivation* and a *z* value of -4.45 for *Zmotivation*. (For a more detailed description of this SPSS procedure, refer to Module 2 in Section 1.)

Dr. Mendoza was still interested in conducting an analysis with and without the outliers that SPSS identified. If results are the same (that is, produce similar statistical decisions), then the outliers are not influential in that particular analysis. If results differ, then he will report both sets of results. Even though he does not have the full set of data because the study is not yet complete, he decided to conduct an independent *t*-test to determine if the prescores on *motivation* are statistically similar for the two groups of *program*. His hypothesis is that there is no difference between the average motivation prescores for participants in the program versus participants not in the program.

In order to conduct the *t*-test without outliers, Dr. Mendoza selected only those cases that are nonoutliers. To do so, he used the **Select Cases**

procedure under the **Data menu**. (Refer to Module 2 in Section 1 for a detailed description and example of the use of the Select Cases procedure.) After he clicked the **If condition is satisfied** button, he created the logical expression in the dialog box shown in Figure 5.6 to tell SPSS that he did not want cases with id values equal to 190, 12, 91, 13, 25, and 19 to be included in the analysis. After he clicked **OK**, a variable called *filter_\$* (the default name) was

SPSS Tip 4: If he didn't create this variable, Dr. Mendoza would need to recreate the logical expression and rerun Select Cases each time he wanted to conduct analysis with and without outliers. This is time-consuming and can lead to errors if any part of the expression is retyped incorrectly.

Figure 5.6 Data → Select Cases

added to the file. It contains values of 0 and 1. Dr. Mendoza renamed this variable *motivation_nonoutliers* in case he decides to use it again in future analysis. This new variable as well as *Zmotivation* are included in the file *module5_jobsatis_final.sav*. **SPSS Tip 4

Dr. Mendoza was curious about the nature of the distribution without the outliers ($n = 259$), so prior to running the t -test, he decided to run the **Explore procedure** again. He found that it was no longer negatively skewed or peaked. Standardized values for skewness and kurtosis were .013 and $-.391$, respectively. Also, the box plot in Figure 5.7 indicates no outliers, and the distribution looks approximately normal.

Now Dr. Mendoza performed the t -test by choosing **Compare Means** and **Independent-Samples T Test** under the **Analyze menu**. The dependent variable (**Test Variable**, as SPSS calls it) is *motivation*, and the independent variable (**Grouping Variable**, as SPSS calls it) is *program* (see Figure 5.8). **SPSS Tip 5

SPSS Tip 5: Do not make the mistake of using *motivation_nonoutliers* as the dependent variable in this t -test. If you do, no results will be produced because the values for this variable are constant (all 1). The *motivation_nonoutliers* variable is only used in this context to select cases that are not outliers.

Figure 5.7 SPSS Boxplot for the Distribution of Motivation After Outliers Were Removed

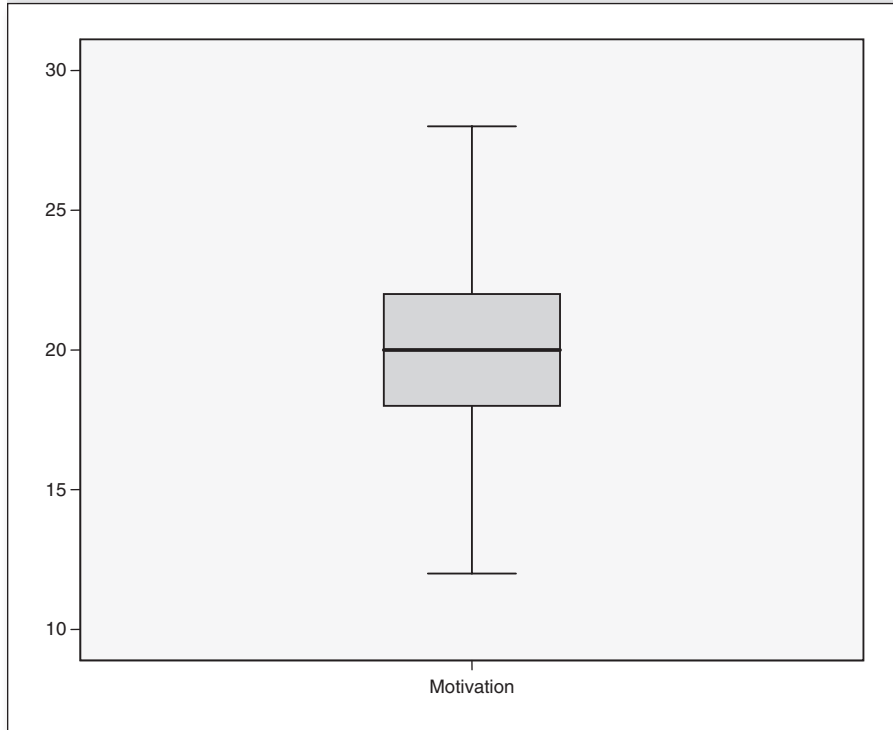
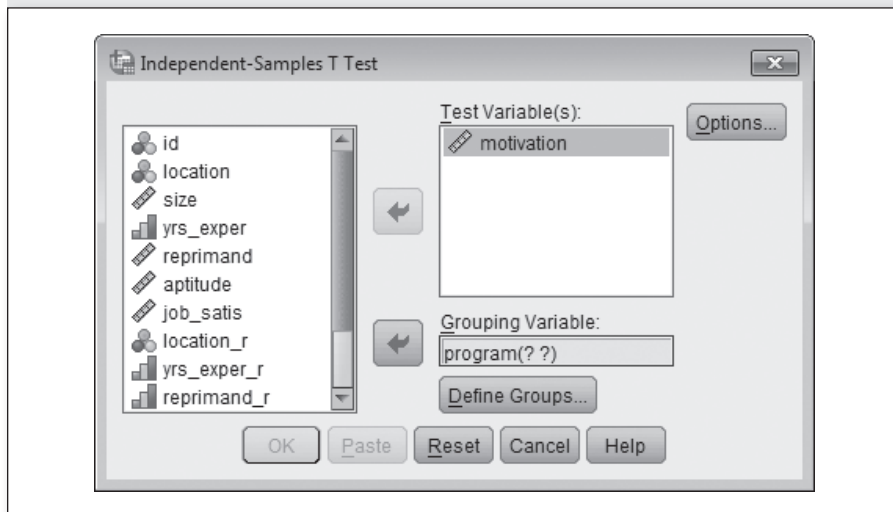
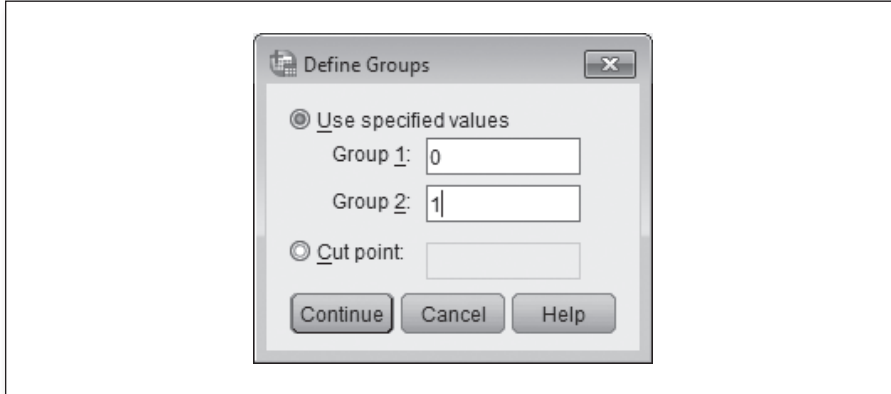


Figure 5.8 Analyze → Compare Means → Independent-Samples T Test



Codes for *program* must be defined. He clicked **Define Groups**, and it opened the box shown in Figure 5.9. He entered the two codes (0, 1) he used for *program*.

Figure 5.9 Define Groups in *T*-Test



The first table of output produced by the *t*-test is shown in Table 5.3. As a double-check that the *t*-test was conducted only on the nonoutliers, Dr. Mendoza made sure that the total *N* for both groups sums to 259, and it does. He noticed that the group means are quite similar to each other, approximately 20, which likely indicates no significant difference between the two means.

Table 5.3 SPSS Output for Descriptive Statistics of the *t*-Test on Motivation Without Outliers

Group Statistics					
	Program	<i>N</i>	Mean	Std. Deviation	Std. Error Mean
Motivation	0 no	135	20.26	3.357	.289
	1 yes	124	20.16	2.892	.260

The second set of output from the *t*-test is not reprinted here. The *t* statistic and *p* value indicate that the difference between the average motivation score for participants who are not receiving the job enrichment program compared to participants who are receiving the program was not statistically significant [$t(257) = .251, p = .802$].

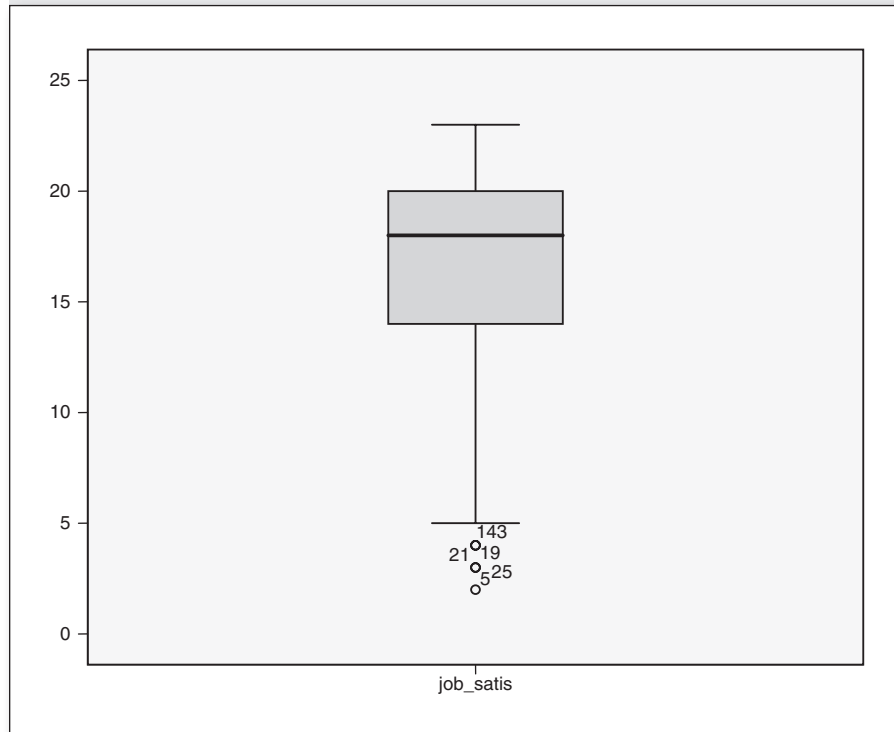
Dr. Mendoza's final step for *motivation* was to conduct another *t*-test using the full sample with outliers. Before doing so, however, he went to the **Data menu**, chose **Select Cases**, and clicked the box next to **All cases** to tell SPSS to run the analysis on all 265 cases. Then he conducted the *t*-test, and the statistical results led to the same conclusion. There is no significant difference between motivation scores for employees who are not receiving the program and those who are [$t(263) = -.173, p = .862$]. The means for the two groups (19.98 and 20.06, respectively) were nearly identical to the means produced without the outliers. The overall conclusion is that the outliers are not affecting the outcomes of the *t*-test.

Job Satisfaction

The final variable Dr. Mendoza examined in this module was *job_satis*. Results after running the **Explore procedure** (in the same way he did for *motivation*) are in Table 5.4 and Figures 5.10 and 5.11. The descriptive statistics indicate a negatively skewed distribution, but no kurtosis. Using the values in Table 5.4 for each statistic and its standard error, he calculated the standardized values to be -6.707 for skewness and 2.218 for kurtosis.

Table 5.4 First Portion of SPSS Explore Output: Summary Statistics for Job Satisfaction

Descriptives			Statistic	Std. Error
job_satis	Mean		16.59	.266
	95% Confidence Interval for	Lower Bound	16.07	
	Mean	Upper Bound	17.11	
	5% Trimmed Mean		16.90	
	Median		18.00	
	Variance		18.713	
	Std. Deviation		4.326	
	Minimum		2	
	Maximum		23	
	Range		21	
	Interquartile Range		6	
	Skewness		-1.006	.150
	Kurtosis		.661	.298

Figure 5.11 Third Portion of SPSS Explore Output: Boxplot for Job Satisfaction

line to the end of the bottom whisker). The plot also shows the case numbers of the five outliers. As before, Dr. Mendoza had the file sorted by *id* in ascending order so that the case numbers coincide with the *id* values for participants. He checked cases 143, 25, 21, 19, and 5 in his data file and found they had values for *job_satis* of 4, 4, 3, 3, and 2, respectively.

He also noticed that four of the five outliers were from Location 1, the other was from Location 6. This result made Dr. Mendoza curious about the shape of each individual *job_satis* distribution across *location*. Therefore, he decided to run the **Explore procedure** again in order to obtain a boxplot for each location. To do so, he placed the revised location variable, *location_r*, in the **Factor List** box in the **Explore** dialog box shown in Figure 5.12.

The five boxplots appear in Figure 5.13. When looking at these distributions for *job_satis* by *location_r*, Dr. Mendoza noticed additional

Figure 5.12 Analyze → Descriptive Statistics → Explore

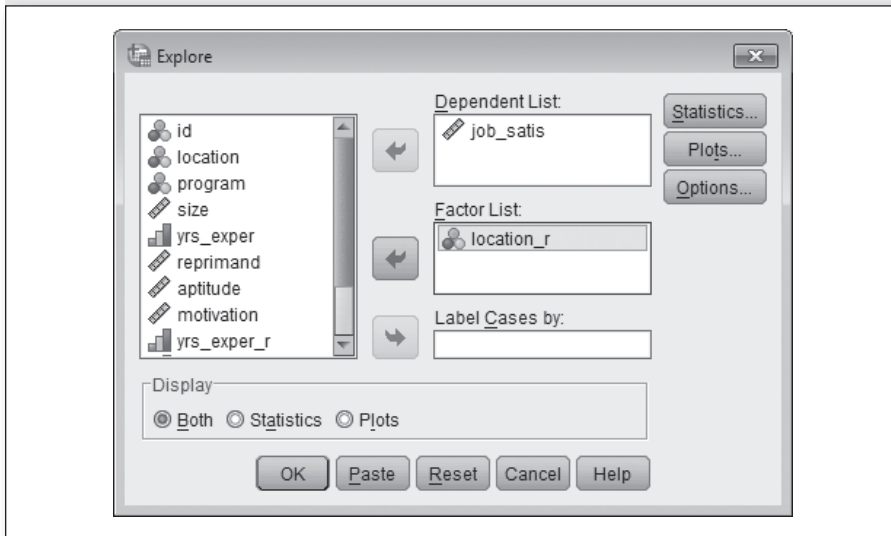
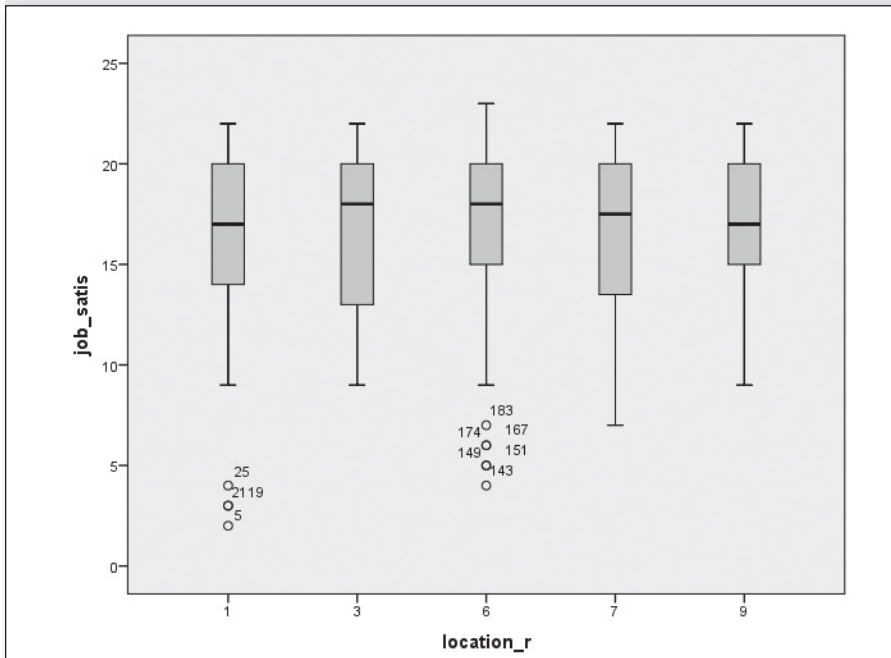


Figure 5.13 SPSS Boxplots Displaying Distributions for Job Satisfaction by Location



outliers. Within Location 6, there are six outliers (cases 183, 174, 167, 151, 149, and 143), and only one of them was previously identified as an outlier on the entire set of data. These outliers had values of 7, 6, 6, 5, 5, and 4, respectively. Three locations had no outliers. Dr. Mendoza has identified an important aspect of the prescores for *job_satis*. When post-data at the end of the study become available, he will need to ensure that analysis is conducted with and without the outliers.

△ E. Reflection and Additional Decision Making

Dr. Mendoza made some interesting discoveries about the prescores for *motivation* and *job_satis*. He will go through the same steps described here in this module when the job enrichment program ends and all participants complete the motivation and job satisfaction instruments again. Recall that additional data are being collected during and after the study in the form of interviews with participants and supervisors, assessment of job productivity, and an evaluation of job skills and knowledge. Based on the information he obtained in the preliminary steps described in this module, especially for *job_satis*, he believes that results across locations are likely to be quite interesting.

In this module, he examined univariate outliers, which are outliers on one variable. If some of his research questions require a multiple regression analysis to determine which factors impact job satisfaction, he will need to examine multivariate outliers (defined as outliers on two or more variables). Procedures are available to identify cases that have unusual combinations of values on multiple variables as well as cases that are influential data points. (See Section 3, Module 7, for another research scenario that demonstrates how to examine these types of cases.)

△ F. Writing It Up

Dr. Mendoza's written summary of the outlier portion of his investigation on the distribution of variables is provided below. He will also include the boxplots in Figure 5.13 in his report to the company because they show a great deal of information about the five *job_satis* distributions by location.

"Skewness and kurtosis values as well as boxplots were obtained to examine the distributions for two variables that examined motivation

and job satisfaction prior to beginning the job enrichment program. Motivation contained five outliers at the lower end of the scale with values of 9 or less and one outlier at the upper end with a value of 30. Based on the standardized values for skewness (-4.01) and kurtosis (6.35) the distribution was somewhat negatively skewed and peaked. After the cases identified as outliers were removed from the sample, the distribution for motivation was examined again. No outliers were identified and the distribution appeared to be approximately normal which was supported by low skewness and kurtosis standardized values ($.013$ and $-.391$, respectively).

“To determine whether the outliers influenced results for a *t*-test on motivation, this inferential statistical test was conducted twice, with outliers (i.e., the entire sample) and without outliers. Results were nearly identical. Both analyses concluded there was no significant difference between average motivation prescores for participants who experience the job enrichment program and participants who do not. Both *p*-values were above a significance level of .05 ($p = .802$ without outliers and $p = .862$ with outliers).

“For job satisfaction, the boxplot showed a skewed distribution in the negative direction with a standardized skewness value of -6.707 . Scores were clustered at the upper end of the scale, and there were five outliers less than or equal to 4. All but one of these outliers came from location 1. Therefore, another analysis of the job satisfaction distribution was conducted by location. An individual boxplot was produced for each location and two locations (1 and 6) showed a cluster of outliers at the bottom of the scale. This indicates that prior to the beginning of the study, there was a very low level of job satisfaction for a small subset of employees at these two locations. As additional data become available upon conclusion of the study, analyses will determine whether these outlying cases, as well as other outliers that may occur on post scores of motivation and job satisfaction, will have an impact on the study’s outcomes.”

Reflective Questions ▲

-
- How can skewness and kurtosis values be used to describe the shape of a continuous variable’s distribution?
 - What is the process for creating stem-and-leaf plots and boxplots? What information do these visual displays provide about a variable’s distribution?

- Why is it important to identify outliers?
- After identifying outliers, how are they handled in future data analysis?

△ Extensions

- Dr. Mendoza created a variable to identify outliers and nonoutliers for motivation. Create a variable to identify outliers and nonoutliers for job satisfaction.
- Discuss standardized z values. Do you think Dr. Mendoza should be concerned about outliers for job satisfaction? Find the z scores for this variable. How many are above 3, above 4? Can high standardized values occur in a normal distribution? Look through research and statistics books to find out what researchers and statisticians say about outliers and how to handle them.
- Dr. Mendoza conducted t -tests on motivation to compare results with and without outliers. Conduct t -tests to determine if means for job satisfaction prescores differ for employees who are participating in the program versus employees who are not. Compare the results from the two t -tests (one with outliers and one without outliers). Do the outliers have an impact on the results?
- Examine descriptive statistics (e.g., mean, median, mode, skewness, kurtosis) for continuous variables in other research studies and discuss the information they provide about the distribution. For instance, what can differences in the values for mean, median, and mode tell you about the shape of a distribution?

△ Additional Resources for Section 2: The Nature and Distribution of Variables

- Abu-Bader, S. H. (2010). *Advanced and multivariate statistical methods for social science research with a complete SPSS guide*. Chicago: Lyceum. (For Module 5, see Section 2 on outliers.)
- Glass, G. V. & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Needham Heights, MA: Allyn & Bacon. (For Module 4, see Chapters 2 and 3 on variables, their scales and measurements, and visual display of data.)
- Gravetter, F. J., & Wallnau, L. B. (2013). *Statistics for the behavioral sciences* (9th ed.). Wadsworth. (See Chapter 1 for more information on scales of measurement.)
- Holcomb, Z. C. (2009). *SPSS basics: Techniques for a first course in statistics* (2nd ed.). Glendale, CA: Pyrczak.
- Huizingh, E. (2008). *Applied statistics with SPSS*. Thousand Oaks, CA: Sage. (For Module 3, see Section 9 on computation and classification of variables, which contains details about each SPSS function that computes new variables, what it does, and the information necessary to supply.)
- IBM. (2011). *SPSS Statistics 20 core systems user's guide*. Author. (A pdf version can be obtained from <http://www.ibm.com>)
- Kinnear, P. R., & Gray, C. D. (2009). *SPSS 16 made simple*. Sussex, UK: Psychology Press.
- Kulas, J. T. (2009). *SPSS essentials: Managing and analyzing social sciences data*. San Francisco: Jossey-Bass.
- Mertler, C. A., & Vannatta, R. A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation* (3rd ed.). Los Angeles: Pyrczak. (For Module 5, see Chapter 3, which gives information on identifying and handling outliers.)

