3.2 Boxplots

In the previous section we explored the stem and leaf display developed by John Tukey. In this section of chapter 3, we will explore the boxplot another of Tukey's development for organizing and describing distributions of data. A **boxplot** is a graphical chart or diagram that illustrates both the central tendency and variability of a data set; the boxplot was earlier referred to as the **box-and-whisker plot**. The boxplot displays the following statistics or measures of a distribution: a. the *lower whisker* (often the minimum score), b. the *first quartile* (Q_1), c. the *median*, d. the *third quartile* (Q_3), d. the *upper whisker*, and e. extreme scores called **outliers (**indicated by an asterisk, *, on a boxplot**)**. Figure 3.2.1 shows the various components or statistics of a boxplot.

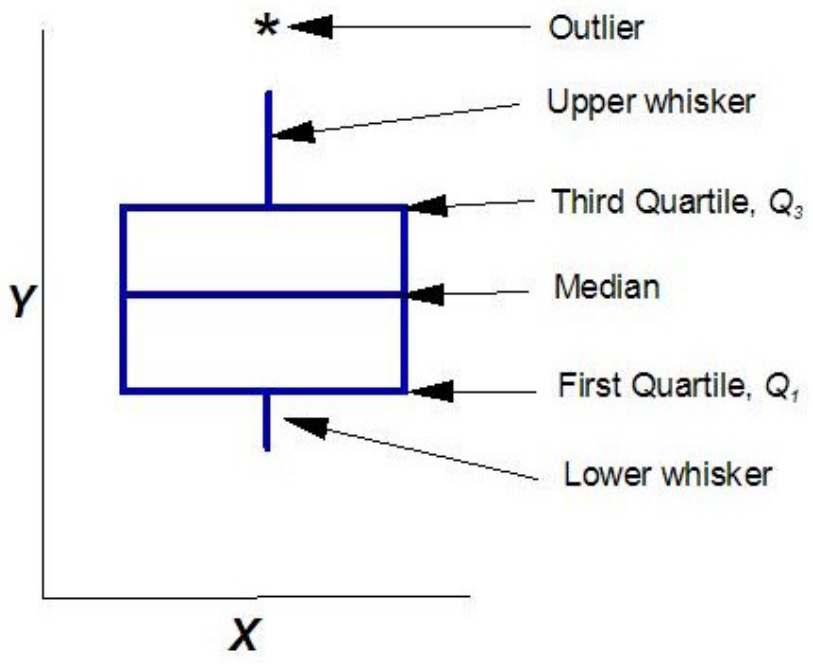

Figure 3.2.1 Description of a boxplot

A **boxplot or box-and-whisker plot** is a graphical representation of the variability or dispersion of a sample distribution.

An **outlier** is an extreme score that does not belong to the sample distribution or a score more extreme than the whiskers of a boxplot.

When the boxplot was first developed in the late 1970's some of the statistics of the diagram were complicated to compute, namely the quartiles (see Frequency Distribution chapter for examples of constructions of percentiles). Today, most applied statistics methods use more simplified computational procedures.

The boxplot consists of:

1. The **median**: For a distribution of *N* scores, the median is the center score of an ordered set of the data (series). The median location for a set of *N* scores is $(N + 1)/2$.
    a. When the median location is a whole number, as it will be when *N* is odd, the median is the score that occupies that location in an ordered arrangement of data.
    b. When the median location is a decimal number (when *N* is even), the median is the average of the two scores on either side of that location.
2. The **quartiles**: The quartiles are the 25 and 75 percentiles (Q_1 and Q_3) of the data distribution. Though there are many ways of finding these points, the **Tukey hinges** method will be used here (SPSS and other statistical programs use this approach). The locations of the first and second quartiles are called **hinges**. The quartiles (or hinges) are obtained from the **quartile location**, which is defined as:

$$\text{Quartile location} = \frac{\text{Median location} + 1}{2}$$

(Drop decimal or fraction from median location is necessary before computing the quartile location)

The quartile location is to a quartile what the median location is to the median (i.e., the medians of the lower and upper 50% of the distribution). It tells where in an ordered dataset (series), the quartile scores are to be found.

3. The **interquartile range** (*IQR*): The interquartile range is the range between the quartiles ($Q_3 - Q_1$).
4. The **whiskers**: The whiskers are lines that extend above and below the quartiles of the plot. A whisker is the furthest point that is no more than 1.5 times the *IQR* (quartiles ± 1.5*IQR*) from the top and bottom of the box (quartiles). If the computed whiskers are more extreme than the actual minimum or maximum scores of the data distribution, we use these (minimum or maximum) instead of the calculated whiskers.
5. **Outliers**: Outliers are scores that are more extreme than the whiskers. An outlier could be an error in measurement or observation, data recording, or in data entry, or could represent a valid score that just happens to be extreme.

An **outlier** is an error in measurement or observation, data recording, or in data entry, or could represent a valid score that just happens to be extreme.

To compute the components of a boxplot, follow the construction guide below. If unable to construct a boxplot by hand using a graphical application, just state the scores for the various statistics of the diagram or plot.

Construction Guide: Creating a boxplot display (*N* is even)

Data: *X* ={*17.5, 10.8, 20.5, 14.2, 9.5, 11.5, 14.5, 17.5, 16.5, 15.2*}

Step 1. Order data: *X* = {*9.5, 10.8, 11.5, 14.2,* ***14.5, 15.2****, 16.5, 17.5, 17.5, 20.5*}

Step 2. Compute the various components of the boxplot:

(a) median: median location = (*N* +1)/2 = 11/2 = 5.5

Median = (14.5 + 15.2)/2 = **14.85**

(b) quartiles: quartile location = (median location + 1)/2 = (5 + 1)/2 = 3

The third scores from top and bottom of ordered data are:

Quartiles = **11.5** and **17.5**

IQR = (17.5 – 11.5) **= 6.0**

(c) upper whisker end: whisker length = quartiles ± 1.5(**IQR**)

Maximum of upper whisker = 17.5 + 1.5(6) = 26.5

Upper whisker end = **20.5**

(score closest to but *not exceeding* whisker length)

(d) lower whisker end

Maximum of lower whisker = 11.5 - 1.5(6) = 2.5

Lower whisker end = **9.5**

(score closest to but *not less* than whisker length)

The following are the SPSS outputs for the above dataset. Observe how the Tukey quartiles are the same as above and the boxplot shows the various statistics of the dataset. SPSS uses the 10 and 90 percentiles for the whiskers of the boxplot. Figure 3.2.2 shows the percentiles and boxplot output from the SPSS program for the data evaluated above.

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	X	9.5000	9.6300	11.3250	14.8500	17.5000	20.2000	.
Tukey's Hinges	X			11.5000	14.8500	17.5000		

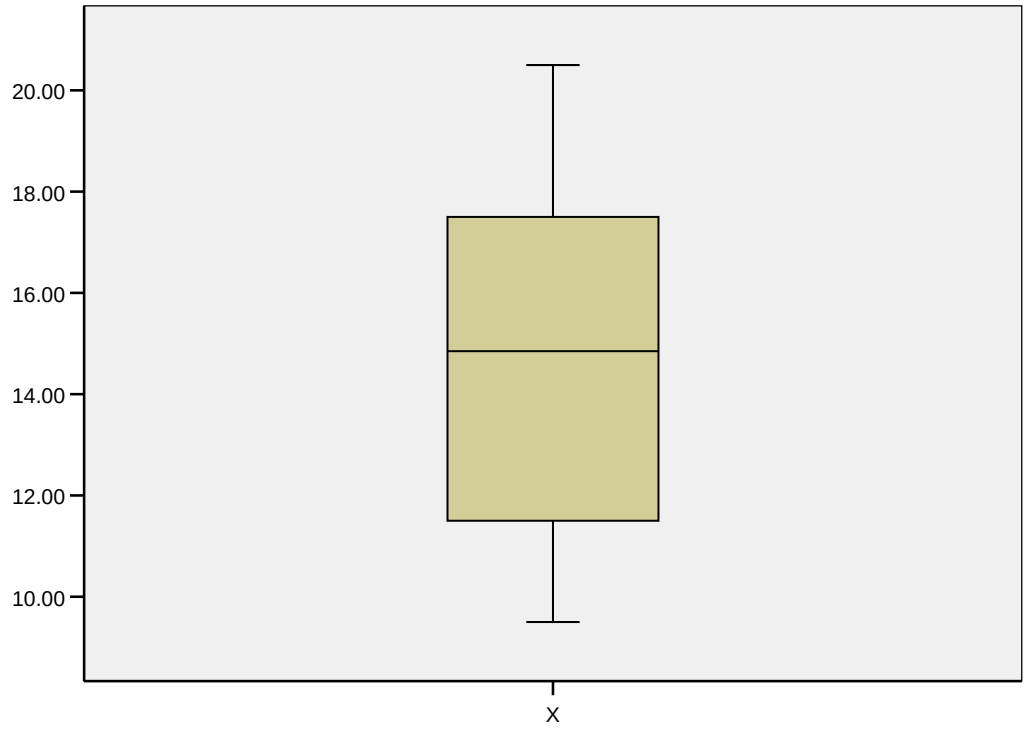

Figure 3.2.2 SPSS Percentile and Boxplot Output (odd *N)*.

Figure 3.2.3 shows the SPSS procedure for creating a boxplot (**Analyze** -> **Explore** -> Select variable, *Statistics* select *Descriptive* and *Percentile*, for *Plots* select *Boxplots*).

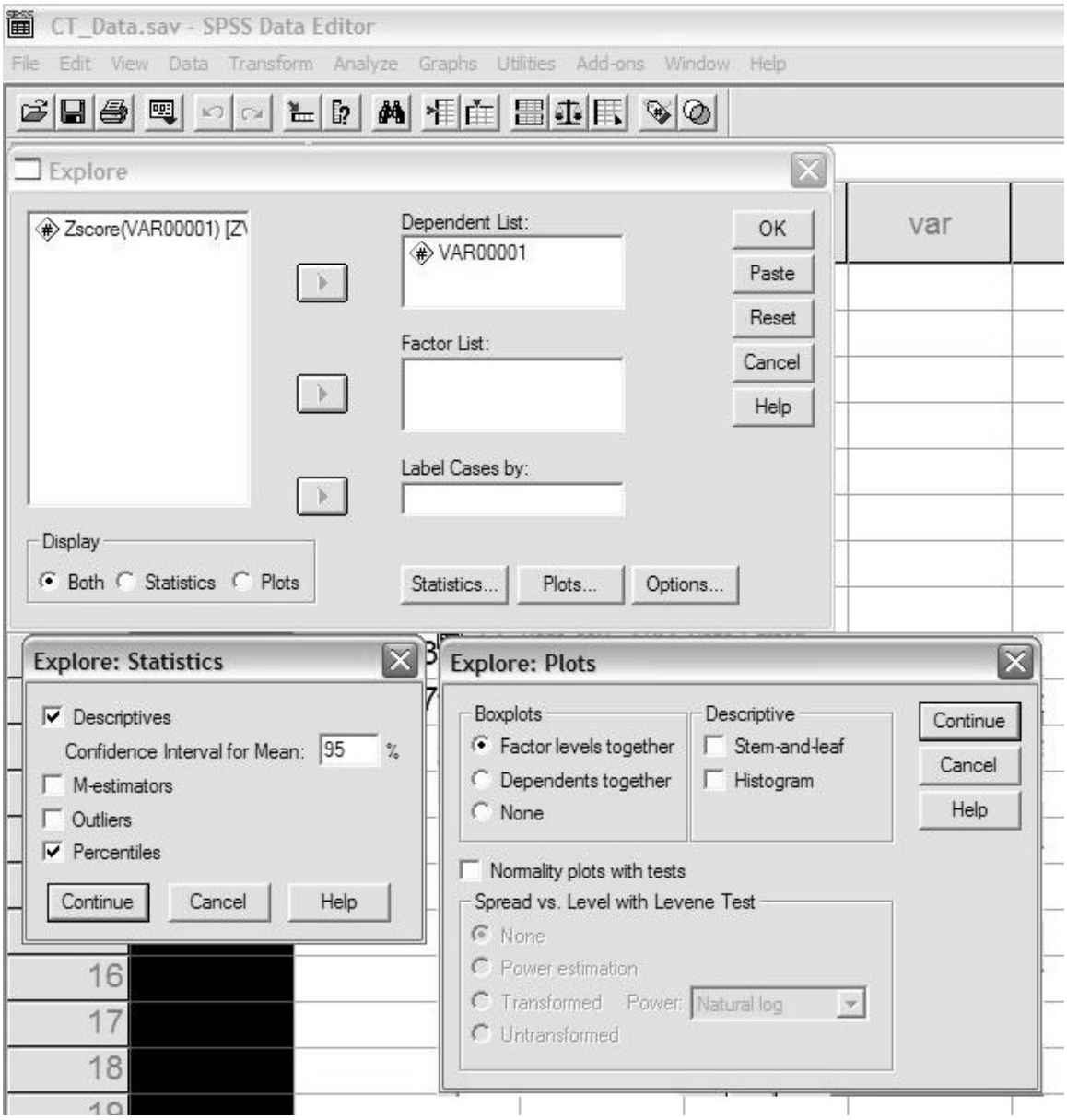

Figure 3.2. 3 SPSS procedure for Boxplot.

Figure 3.2.4 shows the SPSS outputs for the pass9th variable of the ODS.csv dataset. The outliers are indicated with a small circle (**o**). Not all scores identified by SPSS as extreme values are outliers, only scores that are more extreme than the whiskers. There are two scores identified on the boxplot as outliers: 100 and 28 (these are more extreme than the whiskers).

Descriptives

			Statistic	Std. Error
pass9th	Mean		65.8617	1.40371
	95% Confidence Interval for Mean	Lower Bound	63.0742	
		Upper Bound	68.6492	
	5% Trimmed Mean		66.1052	
	Median		67.0000	
	Variance		185.217	
	Std. Deviation		13.60945	
	Minimum		28	
	Maximum		100	
	Range		72.00	
	Interquartile Range		17.50	
	Skewness		-.258	.249
	Kurtosis		.447	.493

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	pass9th	39.2500	49.5000	56.7500	67.0000	74.2500	82.5000	85.2500
Tukey's Hinges	pass9th			57.0000	67.0000	74.0000		

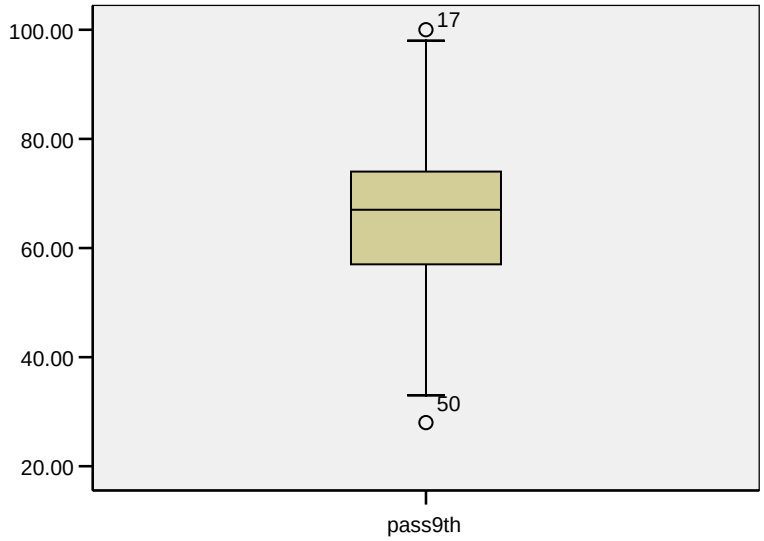

Figure 3.2. 4 SPSS Output: Boxplot and Statistics Outputs from ODE.cvs dataset

Below is another example for constructing a boxplot for a dataset with *N* odd scores.

Construction Guide: Creating a boxplot display (*N* is odd)

Data: $X = \{1, 3, 3, 5, 8, 8, 9, \mathbf{12}, 16, 17, 17, 18, 20, 21, 30\}$

Step 1. Data already an order dataset or a series

Step 2. Compute the various components of the boxplot:

(e) median: median location = $(N+1)/2 = 16/2 = 8$

Median = **12**

(f) quartiles: quartile location = (median location + 1)/2 = (8 + 1)/2 = 4.5

The third scores from top and bottom of ordered data are:

Quartiles = **6.5** and **17.5** (Tukey's hinges)

$Q_1 = (5 + 8)/2 = 6.5$ and $Q_3 = (17 + 18)/2 = 17.5$

IQR = (17.5 – 6.5) **= 11**

(g) upper whisker end: whisker length = quartiles ± 1.5(**IQR**)

Maximum of upper whisker = 17.5 + 1.5(11) = 34

Upper whisker end = **30**

(score closest to but *not exceeding* whisker length)

(h) lower whisker end

Maximum of lower whisker = 6.5 - 1.5(11) = -10

Lower whisker end = **1**

(score closest to but *not less* than whisker length)

Figure 3.2.5 shows the SPSS outputs for the boxplot for the dataset above. Observed how the ***IQR = 13*** statistics is based on the weighted percentile computation, however, ***IQR = 11*** is based on the quartiles computed from Tukey's algorithm (mathematical procedure). The Tukey's *IQR* is preferred for this course.

Descriptives

			Statistic	Std. Error
X	Mean		12.2667	2.09186
	95% Confidence Interval for Mean	Lower Bound	7.7801	
		Upper Bound	16.7533	
	5% Trimmed Mean		11.9074	
	Median		12.0000	
	Variance		65.638	
	Std. Deviation		8.10173	
	Minimum		1.00	
	Maximum		30.00	
	Range		29.00	
	Interquartile Range		13.00	
	Skewness		.536	.580
	Kurtosis		-.084	1.121

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	X	1.0000	2.2000	5.0000	12.0000	18.0000	24.6000	.
Tukey's Hinges	X			6.5000	12.0000	17.5000		

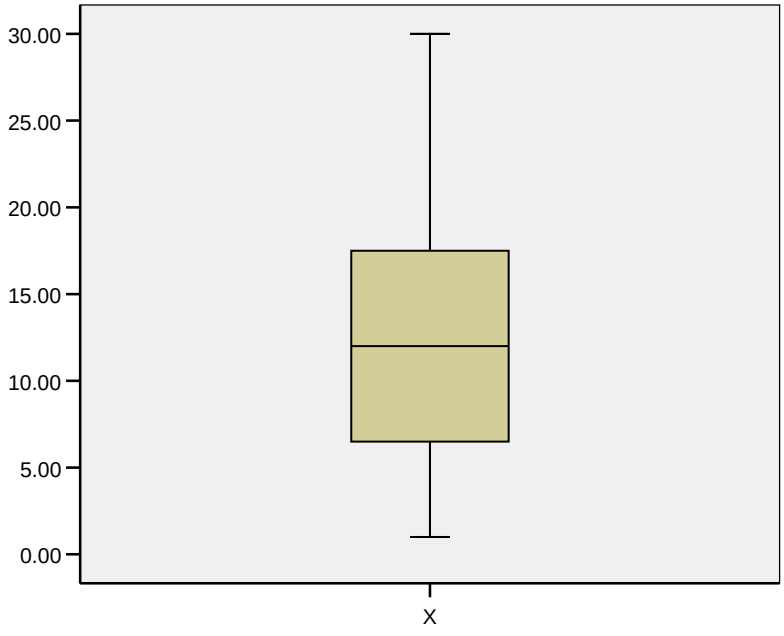

Figure 3.2.5 SPSS Percentile and Boxplot Output (even *N*).