11

Simple Boxplot

PURPOSE OF GRAPHS YOU ARE ABOUT TO BUILD

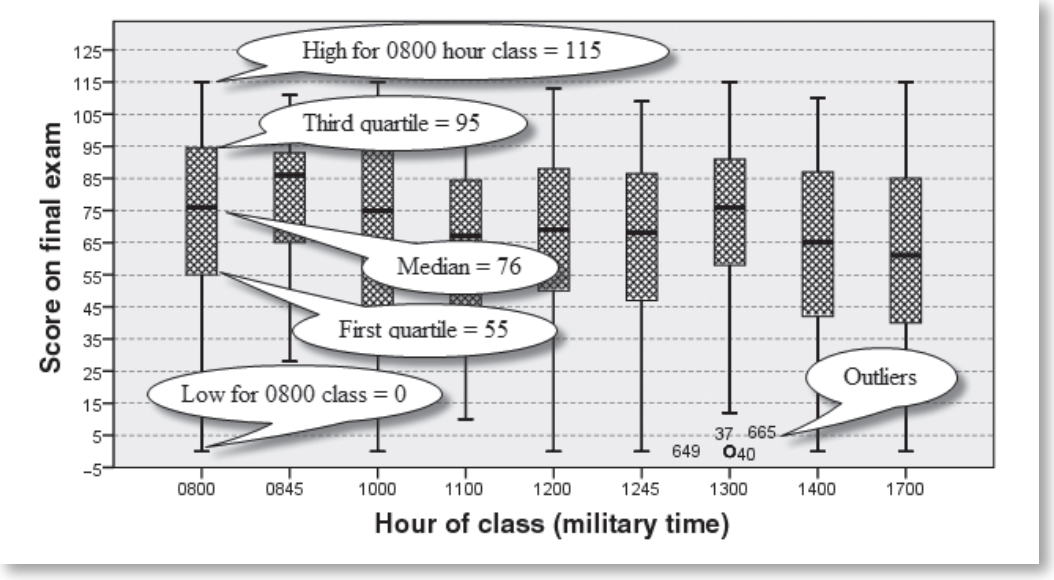
- To compare the variability of a continuous variable categorized by a discrete variable

★ 11.1 Introduction to the Simple Boxplot

First, we advise you not to be misled by what we consider to be a confusing name especially when this graph is compared with the 1-D boxplot described in the previous chapter. The simple boxplot discussed in this chapter is not as "simple" as the 1-D boxplot. The simple boxplot is not more difficult to build; however, it does convey considerably more information making it much more complex than the 1-D boxplot.

The simple boxplot displays the same five statistics as you did in the previous chapter (minimum, first quartile, median value, third quartile, and maximum value). Recall that these statistics were calculated for a single continuous variable. In that chapter, we used a single continuous variable, but the simple boxplot takes it a step beyond. The simple boxplot displays the same five statistics but separates the continuous variable into the several categories of a discrete variable. For an example of the simple boxplot, look at Figure 11.1 that uses data describing 1,050 students in terms of test performance (score on final exam) and hour of their class.

Figure 11.1 Simple Boxplot Displaying a Continuous (Scores) and Discrete (Time) Variable



The authors have chosen the default vertical orientation for this graph. You see a discrete variable with its nine categories (hour of class) displayed on the horizontal axis. The continuous variable is score on the final exam (vertical axis), which ranges from 0 to 125. The –5 value on the vertical axis represents a graph-editing procedure to eliminate the lower whiskers from landing directly on the horizontal axis. We can assure you that the professor did not issue minus points on the exams.

Figure 11.1 shows nine separate box and whisker plots, each giving the same statistics as the 1-D boxplot in the previous chapter. In this figure, you see our information bubbles giving the values for just one of the nine categories of this variable. For the 0800 hour (8 a.m.) class students scored between 0 and 115 points. The middle 50% scored between 55 and 95 points; thus, the interquartile range is 40 points (95 – 55). The median value for the entire grade distribution of the 0800 hour class is 76. We can say that the lowest scoring 25% earned between 0 and 55 points, while the highest scoring 25% earned between 95 and 115.

Remember that the statistics given in the preceding paragraph are for only one class time. By carefully reading the graph, you have these same statistics for all nine class times. The major purpose of this graph is to permit a convenient way to visually compare *all* nine class times on these five statistics

at the same time. The analyst can look for differences and then conduct further investigations to establish plausible explanations for differences. In an experimental research approach, you might specify expected differences before the analysis. Regardless of your research approach, under certain conditions, various statistical tests could be conducted to determine if differences are significant or due to chance fluctuations in the data.[1]

★ 11.2 DATABASE AND QUESTIONS

For your simple boxplot-building experience, you use the **1991 U.S. General Social Survey.sav** consisting of 1,517 cases measured on 43 variables. The discrete variable is labeled *Race of respondent* and named *race*. This variable has three categories of White, Black, and Other. The continuous variable for this exercise is labeled *Age of respondent* and named *age*.

11.2.1 Questions for the Simple Boxplot

1. What are the median ages, ranked from the lowest to the highest, for the three race categories?
2. What are the race category, age, and gender of the one outlier in the graph? (Notice that you must first examine the graph and then check the database to answer this question.)
3. What are the ranks, from the lowest to the highest, for the three race categories in terms of overall ranges of ages? (Give the values for these three ranges.)
4. What are the interquartile ranges for the "White," "Black," and "Other" race categories? (Exclude the outlier from your calculations.)
5. Just by looking at the graph, what can you say about the distributions?

In the next section, you will build the simple boxplot graph that will answer questions about this variable.

---

[1]There are different statistical tests that might be used if one wished to determine if the statistics varied significantly between meeting times for these classes. For instance one may want to determine if the medians differed significantly by using the independent-samples Kruskal–Wallis test. Perhaps, you may wish to check for significant differences in variability for the nine distributions by using a chi-square test of equality.

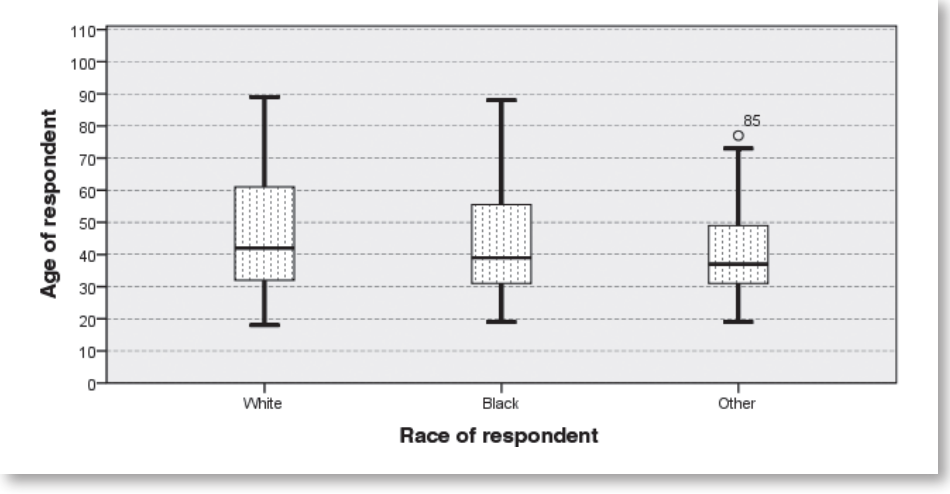
11.3 USING SPSS TO BUILD THE SIMPLE BOXPLOT ★

In this section, you will build the basic simple boxplot and then use the *Chart Editor* to make one similar in appearance to the graph shown in Figure 11.1.

- Open **1991 U.S. General Social Survey.sav** (found in the SPSS Sample files).
- Click **Graphs**, then click **Chart Builder** to open *Chart Builder* window.
- Click **Boxplot**, click and drag the **Simple Boxplot** (the first icon) to the *Chart preview* panel.
- Click and drag **Race of Respondent** to the *X-Axis* box.
- Click and drag **Age of Respondent** to the *Y-Axis* box.
- Click **OK** (the basic graph now appears in the *Output Viewer*).
- Double click the **graph** to open the *Chart Editor*.
- Click **any number** on the *y*-axis.
- In the *Properties* window, click the **Scale tab**, then in the *Range* panel, change *Major Increment* to **10**.
- Click **Apply**.
- Click the **Show Grid Lines icon** (the fifth icon from the right).
- In the *Properties* window, click the **Lines tab**, and in the *Lines* panel, click the **black arrow** beneath *Weight* and click **0.25**, then click the **black arrow** beneath *Style* then click the **first dotted line**.
- Click **Apply**.
- Click on **any whisker of a boxplot** (a faint line appears around all whiskers).
- If *Properties* window is not open, click the **Properties Window icon**.
- In the *Properties* window, click the **Lines tab**, and in the *Lines* panel, click the **black arrow** beneath *Weight*, then click **2**.
- Click **Apply**.
- Click **any boxplot box** (a faint frame appears around all boxes).
- In the *Properties* window, make sure that the *Fill & Border* tab is highlighted.
- In the *Color* panel, click the **white rectangular box**.
- In the *Color* panel, click the **black arrow** beneath *Pattern*, and click the **first pattern** in the **second** row.
- Click **Apply**.
- Click the **X** in the upper right-hand corner of the *Chart Editor* (graph is moved to the *Output Viewer*).

- Click the **graph** (a frame appears around the graph), and then click and grab the **lower right corner** of the frame (marked by a small black square), hover the mouse pointer until you see a *double-headed arrow*, and move it diagonally up and to the left to reach the approximate size of the graph in Figure 11.2.

Figure 11.2 Simple Boxplot for a Discrete (Race) and Continuous (Age) Variable



★ 11.4 INTERPRETATION OF THE SIMPLE BOXPLOT

This section repeats those questions presented earlier (Section 11.2.1) but this time with the answers provided by the information presented in the graph just built.

11.4.1 Questions and Answers for the Simple Boxplot

The information to answer the following questions may be found in Figure 11.2.

1. What are the median ages, ranked from lowest to highest, for the three race categories?

The youngest median age of 36 years is for the "Other" race category, next is for "Black" at 39, and finally, the oldest median age is for the "White" category at 42.

2. What are the race category, age, and gender of the one outlier in the graph? (Notice that you must first examine the graph and then check the database to answer this question.)

The one outlier shown on the graph is for the "Other" race category and is labeled as Case Number 85. If the database is not open do so now, go to Case Number 85 and determine that this respondent is a female and aged 77.

3. What are the ranks, from the lowest to highest, for the three race categories in terms of overall ranges of ages? (Give the values for these three ranges.)

The lowest range for age is found in the "Other" category at 55 (73 – 18) and the next lowest is for the "Black" category at 69 (88 – 19). Those with the largest range is the white category at 71 (89 – 18).

4. What are the interquartile ranges for the "White," "Black," and "Other" race categories? (Exclude the outlier from your calculations.)

The "White" category interquartile age range is 61 (third quartile) – 32 (second quartile) = 29 years. The "Black" category has an interquartile range of 55 (third quartile) – 31 (second quartile) = 24 years. Finally, the "Other" category has an interquartile range of 49 (third quartile) – 31 (second quartile) = 18 years.

5. Just by looking at the graph, what can you say about the distributions?

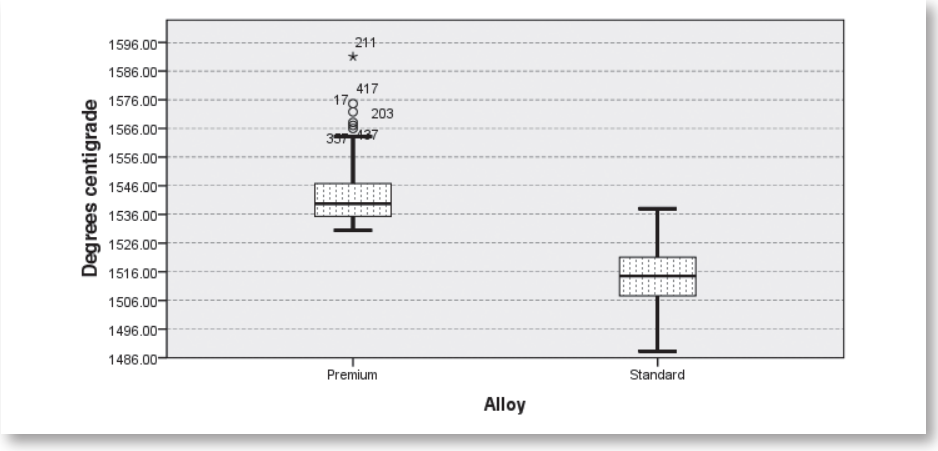
The distributions appear to be roughly the same shape with all having a slight to moderate degree of positive skew.

11.5 REVIEW EXERCISES ★

1. Open **ceramics.sav** in the SPSS Sample files. Select a discrete variable named *batch* and labeled *Alloy*. This variable has two categories, 1 = *premium* and 2 = *standard*. The continuous variable is named *temp* and labeled *Degrees centigrade*. Build the simple boxplot in Figure 11.3 and answer the following questions. (Hint: you will have to make some changes in the *Chart Editor* > *Properties* window > *Scale* tab > *Range* panel > *Minimum* = 1,486, *Maximum* = 1,592, *Major Increment* = 10, and *Origin* = 0.)

Questions: (a) What are the median heat values for both batches? (b) What are the minimum and maximum heats for the premium batch (exclude the outliers)? (c) What are the minimum and maximum temperatures for the standard batch? (d) Which of the two batch types show the more normal distribution?

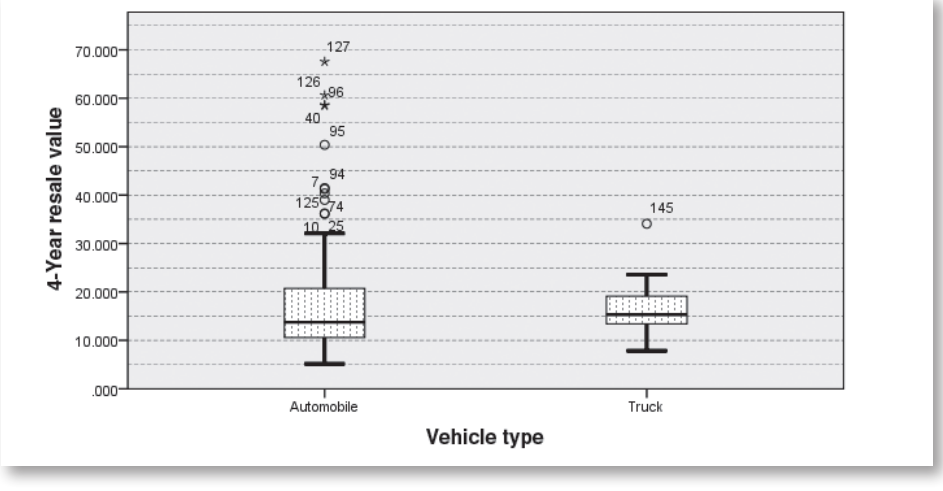
Figure 11.3 Review Exercise: Simple Boxplot for Degrees Centigrade and Alloy



2. Open **car_sales.sav** and select the discrete variable named *type* and labeled *Vehicle type*. The continuous variable is named *resale* and labeled *4-year resale value*. Build the simple boxplot as in Figure 11.4 and answer the following questions.

Questions: (a) Which distribution approximates the normal distribution? (b) Which of the two distributions has the highest median resale value and what is this value? (c) What are the highest resale values for trucks and automobiles when outliers and extremes are excluded? (d) What is the interquartile range for automobiles? (e) What are the minimum and maximum values for the resale value of the middle 50% of the trucks?

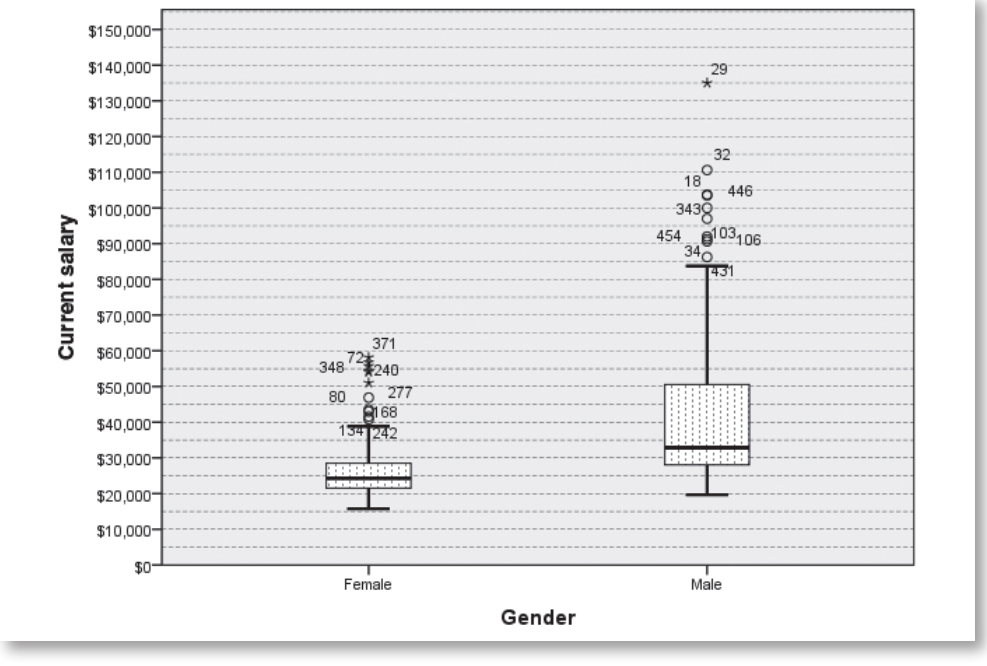
Figure 11.4 Review Exercises: Simple Boxplot for Resale Value and Vehicle Type



3. Open **Employee data.sav**, and select *gender* as the discrete variable. Select the continuous variable named *Salary* and labeled *Current salary*. Build the simple boxplot in Figure 11.5 and answer the following questions.

Questions: (a) What is the value of the most extreme value in the male salary distribution? (b) What is the most extreme salary for the females? (c) If you exclude the outliers and extremes, what are the highest and lowest salaries for the males? (d) Excluding the outliers and extremes, what are the high and low salaries for females? (e) What are the median salaries for males and females?

Figure 11.5 Review Exercises: Simple Boxplot for Current Salary and Gender

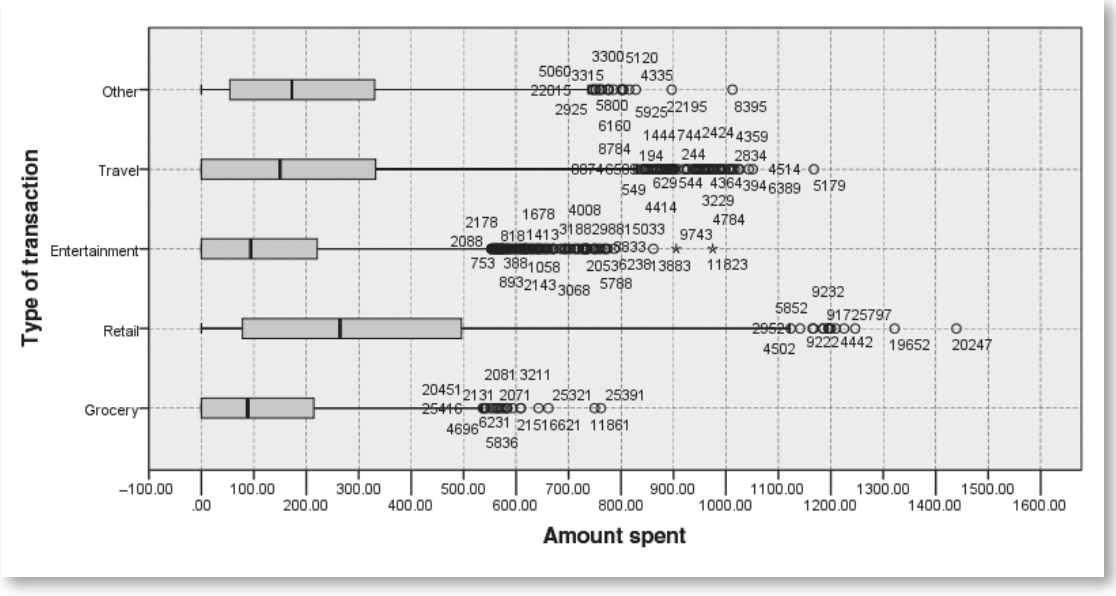


4. Open **credit_card.sav** from the SPSS Sample file, and select *Type of transaction* as the discrete variable. Select the continuous variable named *spent* and labeled *Amount spent*. Build the simple boxplot in Figure 11.6 and answer the following questions.

Questions: (a) What type of transaction has the largest interquartile range, and what is that value? (b) Which type of transaction recorded the highest expenditure as measured by the median, and what was that amount? (c) Which type of transaction recorded the highest expenditure, and what

was that amount? (d) Rank from high to low the types of transactions and their maximum spent. (e) Which transaction type has the highest minimum expenditure, and what is that amount?

Figure 11.6 Review Exercise: Simple Boxplot for Amount Spent and Transaction Type



5. Open **credit_card.sav** from the SPSS Sample file, and select *gender* as the discrete variable. Select the continuous variable named *spent* and labeled *Amount spent*. Build the simple boxplot in Figure 11.7 and answer the following questions.

Questions: (a) Which gender has the largest interquartile range, and what are both these values? (b) Which gender recorded the highest expenditure as measured by the median, and what was that amount? (c) Which gender recorded the highest expenditure, and what was that amount? (d) Excluding the outliers and extreme values, which gender had the maximum expenditure and what is that amount? (e) Which gender has the highest minimum expenditure, and what is that amount?

Figure 11.7 Review Exercise: Simple Boxplot for Amount Spent and Gender