CONTENT OVERVIEW

The topic of this unit is the **five-number summary** and its associated graph, the **box-and-whisker plot** or **boxplot**. The five-number summary of a set of data consists of the minimum, **first quartile**, median, **third quartile** and maximum. You already know how to calculate the minimum, median, and maximum. In this overview, we will provide algorithms for calculating the quartiles. It should be noted, however, that there are several different algorithms for calculating the quartiles. So, check with your textbook or software to see how it calculates quartiles.

First, we discuss a rationale for the five-number summary for describing a data set. The five-number summary provides information on both the center of a distribution and its spread. The median is a useful measure of the *center* of a set of observations. The median is the midpoint, the point with half of the data at or below it and half above. However, the median alone is not an adequate description of a set of data. For example, it is not enough to know that the median number of candies in bags of candy is 60 pieces. It is quite a different story if (1) some bags have as few as 40 and others have as many as 75 compared to (2) some bags have as few as 55 and others have as many as 65. To quantify these two situations, we'll need information about the *spread* or *variability* of the data.

Because the median is the "halfway" point in a data set, one way to show spread is by giving the two quartiles along with the median. The first quartile is the one-quarter point in the data: one-fourth of the data values are at or below the first quartile and three-quarters above. The third quartile is the three-quarters point, with three-quarters of the data at or below it. The two quartiles capture the middle half of the data between them. So, the distances from the median out to the quartiles and between the quartiles show how spread out the data are, or at least how spread out the middle 50% of the data are. The distances between the first and third quartiles, $Q_3 - Q_1$, is called the **interquartile range** or **IQR**.

To calculate the quartiles, first locate the median in an ordered data list. The median divides the ordered data into a lower half and an upper half.

- If there is an odd number of data values, the median is the middle data value in the ordered list. Omit this value when forming the lower half and upper half of the ordered data.
- If there is an even number of observations, the median is between the middle two data values. So, the ordered data can be divided into a lower half and upper half about the median.

The first quartile, Q_1, is the median of the lower half of the ordered data and the third quartile, Q_3, is the median of the upper half of the ordered data.

So far, we have discussed using the median to describe the center of a distribution and the interquartile range to describe the spread of the middle half of the data. We can add information about how far the data are spread by giving the distances from the median out to the minimum and maximum data values and between the minimum and maximum. The distance between the minimum and maximum, maximum – minimum, is called the **range**.

Now, we work through an example. Grades from an exam are displayed in the stemplot in Figure 5.5. We use the stemplot to order the test scores from smallest to largest.

```
4 | 01
5 |
6 | 89
7 | 2699
8 | 0256788
9 | 01236
```

Figure 5.5. Stemplot of test scores.

Since n = 20, the median is at the (20 + 1)/2 or 10.5 position, midway between 82 and 85; hence, the median = 83.5. The median divides the data into a lower half and upper half:

Lower half:	40	41	68	69	72	76	79	79	80	82
Upper half:	85	86	87	88	88	90	91	92	93	96

The median of the lower half is at the (10 + 1)/2 or 5.5 position, midway between 72 and 76; so, $Q_1 = 74$. The median of the upper half is midway between 88 and 90; so, $Q_3 = 89$.

Here's our five-number summary of the exam grades:

minimum = 40, Q_1 = 74, median = 83.5, Q_3 = 89, maximum = 96

We can use the median, 83.5, as a measure of center for the test scores. The spread of the middle 50% of the test scores is given by the interquartile range, IQR = 89 – 74 = 15. The spread as measured from the smallest test score to the largest is given by the range = 96 – 40 = 56. Notice that the overall spread of the test scores is more than three times the spread of the middle 50% of the test scores.

In its basic form, a **boxplot** (or **box-and-whisker plot**) is a graphical display of the five-number summary. It can be drawn either vertically or horizontally depending on your preference. Once you have the five-number summary, it takes only three steps to draw a basic boxplot as outlined below.

Constructing a Basic Boxplot

The instructions below are for horizontal boxplots but easily can be adapted for vertical boxplots.

Step 1: Draw a number line. Add a scale that begins at or below the minimum and ends at or above the maximum.

Step 2: Directly above the number line, draw a rectangular box that extends from Q_1 to Q_3. Divide the box with a vertical line at the median.

Step 3: Draw two whiskers: one from the middle left side of the box to the minimum and the other from the middle right side of the box to the maximum.

Figure 5.6 shows the result of applying these steps to create a basic boxplot from the five-number summary for the test scores.

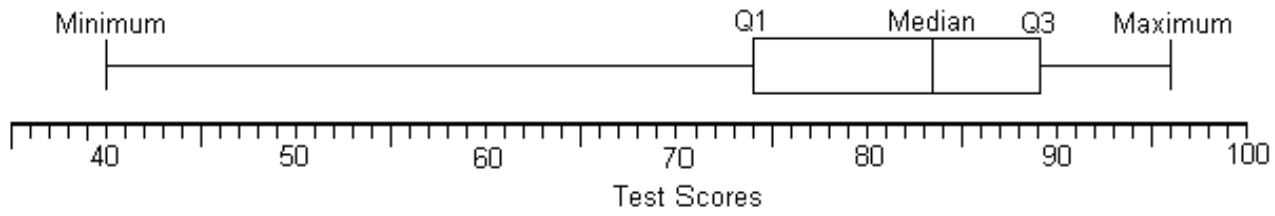

Figure 5.6. Basic boxplot of exam grades.

$Q_1 = 74$, median = 83.5, $Q_3 = 89$, maximum = 96

Each part of the boxplot – the left whisker, the box from Q_1 to the median, the box from the median to Q_3, and the right whisker – represents the spread of one quarter of the data. So, for example, because the box from Q_1 to the median is longer than the box from the median to Q_3, we know that the second quarter of the test scores are more spread out than the third quarter of the test scores.

Notice also the long left whisker that extends from $Q_1 = 74$ all the way down to the minimum test score of 40. We don't know if that long whisker is the result of a single low grade, an outlier, or if the pattern of the lower quarter of the test scores spreads out over the interval from 40 to 74. A modified boxplot, which separates out the outliers and adjusts the lengths of the whiskers so that they are unaffected by outliers, will help us sort out this issue. Here are the steps needed to convert a basic boxplot into a modified boxplot (the generally preferred plot).

Constructing a Modified Boxplot

Step 1: After making a basic boxplot, remove the whiskers.

Step 2: Compute the IQR = $Q_3 - Q_1$; compute a step = 1.5 × IQR.

Step 3: Calculate the inner fences (one step on either side of the box ends):

$$Q_1 - \text{ 1 step} \text{ and } Q_3 - \text{ 1 step}.$$

Calculate the outer fences (two steps on either side of the box ends):

$$Q_1 - 2 \text{ steps} \text{ and } Q_3 + 2 \text{ steps}.$$

Step 4: Identify the mild outliers. Use an asterisk (*) to plot any data values that lie between the two fences. Identify the extreme outliers. Use another symbol, such as an open circle, to plot any data values that are more extreme than the outer fences.

Step 5: Attach a whisker from the left end of the box to the smallest data value that is not an outlier. Then attach a whisker from the right end of the box to the largest data value that is not an outlier.

Next, we convert the basic boxplot from Figure 5.6 into the modified boxplot shown in Figure 5.7. We begin by removing the whiskers from the basic boxplot. Then we calculate the inner and outer fences as follows:

IQR = 89 – 74 = 15
Step = 1.5 × IQR = 1.5(15) = 22.5
Inner fences: Q_1 - 1 step; Q_3 + 1 step: 74 - 22.5 = 51.5; 89 + 22.5 = 111.5
Outer fences: Q_1 - 2 steps; Q_3 + 2 steps: 74 - 2(22.5) = 29; 89 + 2(22.5) = 134

Two test scores, 40 and 41, fall between the lower inner fence and lower outer fence and hence are classified as mild outliers. Mark each of their locations with an asterisk. (There are no extreme outliers.) Attach the left end of the box at Q_1 to 68, the smallest test score that is not an outlier. Redraw the original right whisker (since all test scores were smaller than the upper fences).

The completed modified boxplot appears in Figure 5.7.

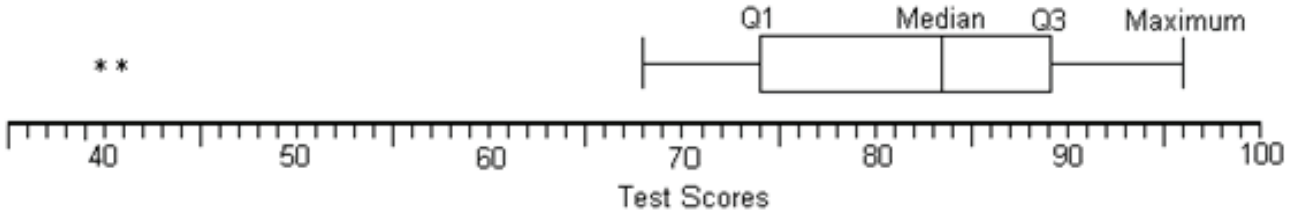

Figure 5.7. Modified boxplot of test scores.

Notice that in the modified boxplot, the length of the lower whisker is about the same as the upper whisker, which indicates that with outliers removed the lower quarter of the test scores had about the same spread as the upper quarter of the test scores. The long left whisker in the basic boxplot was due to two students whose grades were outliers.