Chapter 4

Box-and-Whisker Plots[1]

What is this chapter about? It's about taking data - possibly thousands of numbers - and finding a few measures (values) that help you make sense of the data and represent it effectively. You are probably already familiar with many of these tools, but may not have used them in the way that we describe here.

- Section 4.1 (page 104) of the chapter shows you how to reduce the data to a single number representing the central tendency of the data.
- Section 4.2 (page 111) of the chapter shows you how to reduce the data to several numbers and then represent these numbers in a graph.
- *As a result of this chapter, students will learn*
  - √ What a statistic is and what it is used for
  - √ What an average is and what the common ways of determining an average are
  - √ What quartiles are and what they tell you about data
  - √ What an outlier is
  - √ What a boxplot is, how to read the information in a boxplot, and how to interpret boxplots
  - √ How to compare data sets in order to answer real-world problems
- *As a result of this chapter, students will be able to*
  - √ Compute various summary statistics by hand, with Excel, and with add-ins like StatPro
  - √ Make a boxplot by hand or with StatPro
  - √ Incorporate graphs made in Excel into a Word document effectively to support your work
  - √ Refer to cells in Excel in order to use them in calculations
  - √ Explain what happens to various statistics if the data is increased by a constant amount or by a fixed percentage

4.1 What Does "Typical" Mean?

So far, we've got a lot of information: spreadsheets filled with data that we arranged into variables and observations. But what do we do with all this? Unless you're really special, you probably can't learn a lot from looking at a list of one thousand numbers. You probably know even less from looking at a thousand observations for each of four different variables. Sets of data in business and science are usually larger than this, so we need to think of something fast.

The key is to take it slowly. Rather than look at the entire set of data, we want to look at the data one variable at a time in order to find out what that one variable tells us about the situation about which we collected data. To make things even easier, we want to reduce the data down to one number that represents the "typical" data point for that variable. In general, a number used to represent an entire variable is called a statistic. If that statistic is meant to represent the typical data point, we call it an average.

Watch out, though, the word "average" doesn't really mean what you probably think it does. It has a much more general meaning than "add up the data and divide by the number of data points." That's only one method of computing an average. There are many others. In this chapter, we're interested in the three most common averages: the mean, the median, and the mode.

Another way to think of an average comes from the phrase central tendency. This refers to the middle of the data. You'll always have some data above the average and some below it. The average is a way of talking about the middle of the data. The three described here (mean, median, mode) are the most commonly used ways to compute the middle. Each has a different meaning and has different applications. All are correct ways to compute the middle; it's just that sometimes one is more appropriate than the others. When you go about computing an average you may need to check all three statistics (mean, median, and mode) of these in order to determine which of these will be the most appropriate measure of the typical data point.

If you've understood the ideas above, you might be amused by the statement below, which was issued by Joan Barb Briggs, the president of Generic University, in a moment of administrative desperation:

By the end of the next academic year, I want all of our instructors to have above average course evaluations.

4.1.1 Definitions and Formulas

Statistic Any number used to represent many observations of a single variable or that relates several variables together

Average A statistic that is intended to provide a measure of what a "typical" data point is for a single variable

Mean An average computed by adding all the observations of a variable together and then dividing by the number of observations. In symbols, the mean of the data x_i is $\bar{x} = \sum x_i/n$. This is more properly called the arithmetic mean. Excel uses AVERAGE for

the mean, which is the most commonly used average, and it is the most robust average (it will change the least under repeated sampling of the population)

Median An average computed by first ordering the observations from smallest to largest and then finding the number that splits the observations in half. Observe that this number may or may not be a data point, depending on whether there are an even or odd number of observations. 50% of the observations are less than or equal to the median and 50% are greater than or equal to the median. If there are an even number of points, the median is in between the two center numbers (see example 2 (page 105))

Mode An average computed by determining which observation(s) is repeated most often (or most frequently). The mode is not necessarily unique, nor is it guaranteed to even exist. This is really only useful for discrete numerical data with a few possible values or for categorical data

4.1.2 Worked Examples

Example 4.1. Computing Mean and Median with an Odd Number of Data Points For this example, we want to compute the mean, median and mode of a set of test scores:

55, 60, 67, 70, 78, 81, 84, 88, 90, 95, 99

The mode is the most frequently occurring observation. Since none of the test scores are repeated, there is no mode. We computed the mean of this data in example 1 (page 66) and found it to be about 78.82. Computing the median of the data requires us to put the data in order (this has been done already) and identify the data point in the middle of the ordered list. There are 11 points, so we want the 6th data point (that leaves five numbers less than that observation and five greater than that observation). This makes the median 81, which is slightly higher than the mean, indicating that many students did "above average" on the test. We call a distribution like this "skewed to the left", since the mean is smaller than (to the left of) the median.

55, 60, 67, 70, 78,		81, 84, 88, 90, 95, 99
Lowest five observations	⇑ Median	Highest five observations

Example 4.2. Computing Mean and Median with an Even Number of Data Points

Suppose that we have the same test scores as above, but a student who was absent finally comes to take the test. So now we have twelve test scores:

55, 60, 67, 70, 70, 78, 81, 84, 88, 90, 95, 99

We now have 70 repeated twice, making it the most frequently occurring test score, so the mode of this set of data is 70. To compute the median, we note that with 12 data points, we need to find a score between the 6th and 7th data points. This would be

$$\frac{78+81}{2} = 79.5.$$

55, 60, 67, 70, 70, 78,	81, 84, 88, 90, 95, 99
Lowest five observations	Middle two observations	Highest five observations

The mean can be computed using the same technique as above. A faster approach would be to realize that the first 11 scores (from example 1 (page 66)) have a mean of about 78.82. These will contribute a total of 11*78.82 = 867 to the sum of all the data . Then we add in the new data point, 70, for a total of 937, and divide by the total number of points, 12, to get the mean of the new data as approximately 78.08.

Example 4.3. Comparing Sales Performances

The data below shows the total monthly sales for each branch of Cool Toys for Tots in two different regions of the country, the north-east region and the north-central region. (See file "C04 Tots.xls".) Which of these two regions is performing better?

Sales NE	Sales NC
$95,643.20	$668,694.31
$80,000.00	$515,539.13
$543,779.27	$313,879.39
$499,883.07	**$345,156.13**
$173,461.46	$245,182.96
$581,738.16	$273,000.00
$189,368.10	$135,000.00
$485,344.87	$222,973.44
$122,256.49	$161,632.85
$370,026.87	$373,742.75
$140,251.25	$171,235.07
$314,737.79	$215,000.00
$134,896.35	$276,659.53
$438,995.30	$302,689.11
$211,211.90	$244,067.77
$818,405.93	$193,000.00
	$141,903.82
	$393,047.98
	$507,595.76

One way to answer this question is to compare the mean and median sales in each region. We find that the northeast region has mean sales of $325,000 and median sales of $262,974.85. The north-central region has mean sales of $300,000 and median sales of $273,000.

Based on this information, we might have a hard time deciding which region is performing better. Notice that the mean sales favor the north-east region, indicating higher sales across the region, but the median sales favor the north-central region. In fact, there are more stores in the north-central region and half of them had sales of greater than $273,000. This means that the top half of the stores in the north-central region are doing better in general than the top half of the stores in the north-east region.

Also notice that there is one store in the north-east region with sales of $818,405.93. This is much higher than the sales for the other stores in either region. This single high value is pulling the mean for the north-east region up, even though the stores in the north-central region are typically doing better.

This sensitivity to high or low scores is one of the drawbacks of the mean. This is why the Olympics (and many other sports bodies) drop the high and low scores for a competitor before computing the mean. In Chapter 3B, you'll learn what data points like this are called and gain a powerful graphic tool for determining which data points are likely to have too much influence on the mean.

4.1.3 Exploration 4A: Koduck Salary Increases

Koduck, a local company that makes pictures of water fowl, has 10 employees and needs to give raises to each of them. The company wants to know if it would be better financially (for the company) to give everyone a 3% raise or to add $1000 to each employee's yearly salary.

The yearly salaries of each employee are $24,300; $25,000; $45,000; $40,000; $36,700; $70,000; $19,000; $44,000; $15,000; $43,000.

1. Write down which method (3% raise or flat $1000 increase) you think would be better.

2. For whom is your method better, the management, all the employees, or only certain employees?

3. Why did you select this option?

4. What would help you to make a more informed decision?

5. Now, try this in Excel. Enter the salary data in one column, and then create formulas to have Excel compute the salaries after each of the two methods for the raise. Then, compute the mean and median of each data set using the Excel formulas for mean and median (see the "How to Guide" for this chaper).

6. Compare the mean and median before and after each raise. What happened?

7. Explain why you think this happened.

As it turns out, there is a mathematical explanation for why each change happened the way it did. Using algebra, we can calculate what will happen to the mean and median of any set of data after a fixed amount is added to each data value or after a fixed percentage increase.

4.1.4 How To Guide

For all of the information below, assume that the spreadsheet shown in figure 3.1 (page 75) is being used. It contains data on sample salaries (from Exploration 4.1.3 (page 108)).

Computing Medians in Excel (Method #1)

To compute the median of the data in cells A2:A11, we enter the formula

```
= MEDIAN(A2:A11)
```

into any cell on the spreadsheet. Remember, though, that if you later move or copy the cell, the cell references will be changed since we used relative cell references. Also remember that if you change any of the data in cells A2:A11 the median will be re-calculated instantly. If, however, you add data outside this range, you will need to change the formula.

Computing Medians in Excel (Method #2)

Since the median is also the second quartile (see section 4.2 (page 111)), we could use the formula below to compute the median.

```
= QUARTILE(A2:A11,2)
```

Notice that this function uses two inputs, a range of data (A2:A11) and a number indicating which quartile is desired, 1=first quartile, 3=third quartile.

Computing the Mode in Excel

The mode is computed with the formula

```
=MODE(A2:A11)
```

You may get the result #N/A if there is no mode. If there is more than one mode, Excel just guesses and gives one of them. The fact that there may be more than one mode, or no mode at all, is why this statistic is rarely used except for categorical data.

4.2 Thinking inside the box

Very often, we find that the measures of central tendency - mean, median and mode - are not enough to describe the data we are exploring. These numbers give us some idea of what a typical data point looks like, but they cannot answer questions like:

- How much of the data is less than the average? How much is more than the average?
- What is the largest value in the data? What is the smallest value?
- Where is "most" of the data? Is it close to the average?
- Which measure of central tendency best describes this data?

To answer these questions, we will need to have more tools available. This means that we need more information. If you think about it, we start with a collection of data. This might include thousands of observations of each variable. No human mind can process that much data in order to draw conclusions to make decisions. Therefore, we tried the easiest thing possible: reduce all the data down to a single statistic that represents the central tendency of the data. Now we can see some of the limitations of this approach. Any time we reduce thousands of pieces of data to a single number we have lost information about the data. Consider the following statement:

The mean number of children in a U.S. family is 2.2.

Certainly, this does not mean that every family in the U.S. is made up of 2.2 children. In fact, even to claim that the typical family has 2.2 children is a little strange since the number of children in a family is a discrete numerical quantity. Based on this statement only, which of the following statements most closely seems to describe family structures in the U.S.?

- Most families in the U.S. have two children. A few families have zero, or one child. A few more families have more than two children.
- There are more families with two or fewer children than there are families with more than two children.
- The number of families with two or fewer children is the same as the number of families with three or more children.

In fact, without more information, only the third statement can be ruled out. This one is based on the definition and computations used to compute the mean. (See if you can figure out why the third statement is definitely false.) We cannot decide which of the two remaining statements is more accurate without additional information. One common set of statistics used to get more information about a set of data are called quartiles. The idea behind quartiles is to take the data, put it in order from smallest to largest, and then break it into four quarters, each with the same number of data points in it. We then keep track of the data points at the places where the data is broken up, and we call these statistics quartiles. This gives us some idea of how the data is distributed. Graphically, we can represent the quartiles and other information about the spread of the data in a boxplot, which is a type of graph that contains about seven pieces of information to describe the data.

4.2.1 Definitions and Formulas

Minimum The smallest observation of a variable

Maximum The largest observation of a variable

Range The difference between the largest and smallest observations: Range = Maximum - Minimum

Quartiles These divide the data into four equal-sized groups of observations, based on an ordered list of data from smallest to largest

First quartile (Q1) The first quartile is the numerical value that exactly 25% of the observations are less than or equal to.

Third quartile (Q3) The third quartile is the numerical value that exactly 75% of the observations are less than or equal to.

Interquartile Range (IQR) The distance between the first and third quartiles: IQR = Q3 - Q1. Exactly 50% of the data falls inside the IQR.

Outliers These are data points that are not large enough or small enough to "fit in" with the other data. A **mild outlier** is an observation that is more than 1.5 IQR above Q3 or more than 1.5 IQR below Q1. An **extreme outlier** is an observation that is more than 3 IQR above Q3 or more than 3 IQR below Q1

Boxplot This is a graph of all the basic summary measures of a single variable. It combines all of the above information, the mean, and the median. It is sometimes called a box-and-whisker plot. A sample plot is shown below.

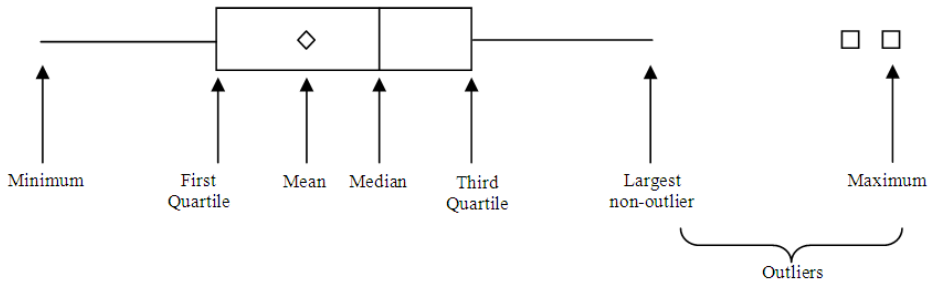

Figure 4.1: Sample boxplot (without scale) showing the major features

4.2.2 Worked Examples

Example 4.4. Making a boxplot when there are an odd number of data points
Consider the list of test scores below:

55, 60, 67, 70, 78, 81, 84, 88, 90, 95, 99

We already determined that the mean of this data is 78.82 and the median is 81. We now divide the list into four equal parts to determine the quartiles. Start by dividing the data into two equal parts, as with finding the median. Then divide each of these into two equal parts. For this data, each quartile should include three data points, since there are 11 total. Notice that the middle data point, the median, is in both the upper half and the lower half of the data when we divide it up.

Lowest 50%					Median	Upper 50%				
55	60	67	70	78	81	84	88	90	95	99
Lowest 25%			Lowest 25%			Lowest 25%			Lowest 25%	

We now have almost everything that we need to make the boxplot. We just need to check whether there are any outliers in this data. An outlier is more than 1.5 IQR from Q1 or Q3. The interquartile range (IQR) for this data is IQR = Q3 - Q1 = 89 - 68.5 = 20.5. Thus, outliers must be more than 1.5*20.5 = 30.75 from the quartiles. Outliers on the low end would be less than (Q1 - 30.75) = (68.5 - 30.75) = 37.75. Outliers on the high end would be greater than (Q3 + 30.75) = (89 + 30.75) = 119.75. Since there are no data points outside this range, there are no outliers in this data.

To make the boxplot, we simply draw an axis scaled from 55 to 99 (for ease of reading, let's go from 50 to 100 in steps of 5). We then draw the box part of the graph, extending from Q1 to Q3. We put a vertical line in the box at the median. We add a star or a diamond for the mean, and then we extend the "whiskers" of the box from the edges out to the minimum and maximum, since there are no outliers. The final result is shown in figure 4.2 (page 113)

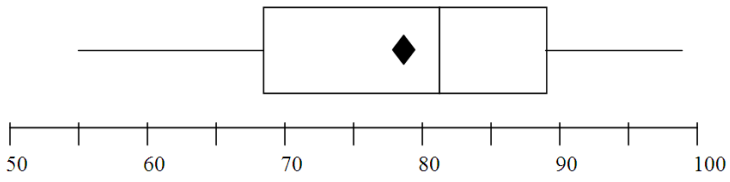

Figure 4.2: Boxplot of test scores

Example 4.5. Reading an Interpreting a Boxplot
Consider the boxplot shown in figure 4.3 (page 114). It represents the distribution of test scores in a class of 120 students. What can we learn about the class performance from this graph?

To analyze the graph, we will consider a series of questions:

1. What is the minimum test score? What is the maximum? What is the range? From the data, the lowest score is a 50 (which is an outlier) and the highest score is a 95. The range is 95 - 50 = 45.

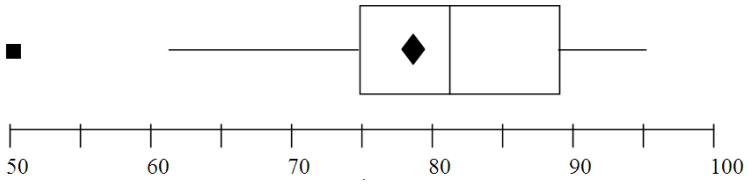

Figure 4.3: Boxplot of test scores for 120 students. What does it say about class performance?

2. What are the quartiles of the data? The first quartile is 75. The median (second quartile) is about 82 and the third quartile is about 89. The IQR is 89 - 75 = 14. This number is about one-third of the range, indicating a fairly tight spread of data (lots of similar test scores).

3. How many students scored between Q1 and Q3? We know that 50% of the data is always between Q1 and Q3. This means that 50% of the observations (in this case student test scores) fall between 75 and 89. Since 50% of 120 (the total number of observations) is 0.50*120 = 60, we know that 60 students scored between 75 and 89 on the test. However, this is a little misleading. It is possible that there are multiple students with the same tests score. If these duplications happen to be at the quartiles, then a few more students would be in the 75 to 89 range.

4. Assuming that a score of 90 is sufficient to earn an "A", how many students got an "A" on the test? This is harder to answer. Notice that the third quartile is 89. This means that 25% of the class (0.25*120 = 30 students) got an 89 or higher. So we only really know that at most 30 students earned an "A". It is possible that most of the scores in the third quartile are right at 89 and only a few of them are between 89 and 95, which would lead to a smaller number of A's on the test.

5. Did most students do well or poorly on this test? We see that the median is slightly higher than the mean. This shows that more than 50% of the class earned a score above the mean. We do not know exactly what percentage scored above the mean, only that it is between 50% and 75% (since the mean is between Q1 and the median). This indicates that the data is negatively skewed, so that more of it is piled up above the mean than below it. Overall, then, it seems that the class did a little better than average on this test.

6. What other questions could we ask about the data?

Example 4.6. Side-by-Side Boxplots
Consider the sales data given above in example 3 (page 106). (Data file "C04 Tots.xls".) Let's use boxplots to compare the two sales regions and select the region that has the better performance. If you enter the data above into Excel and use StatPro to create side-by-side boxplots (see "How to guide" for details) you should get the graph in figure 4.4 (page 115).

As you can see, the boxplot shows that the sales in the North East region are spread over a much greater range of sales figures than the sales in the North Central region. In addition,

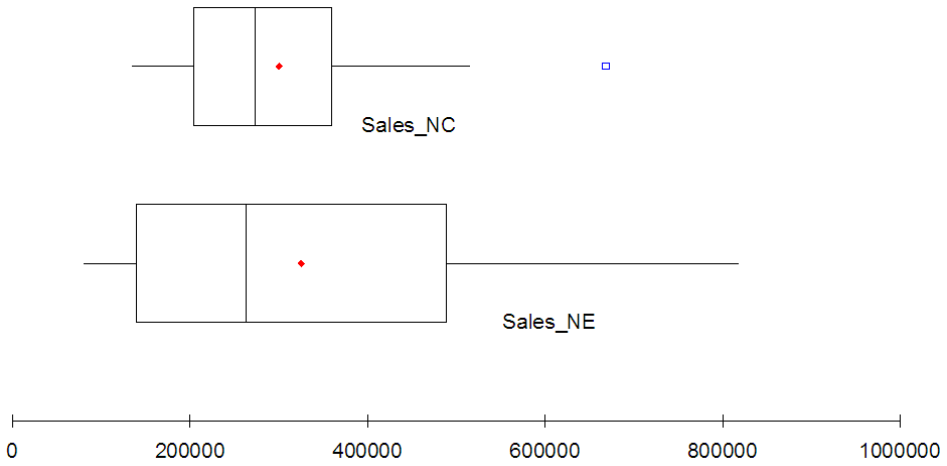

Figure 4.4: Side-by-side boxplots of sales from two regions of the Toys-For-Tots company.

the highest performing store in the NC region is an outlier and is not at all representative of the region's performance. However, the lowest 25% of the stores in the NE region are performing worse than all of the stores in the NC region (the minimum for the NC region is about equal to the first quartile for the NE region). By the same token, the upper 25% of stores in the NE region seem to be doing better than all the stores in the NC region (the third quartile of NE region sales is about equal to the maximum for the NC region, if we ignore the outlier). The middle of each region seems to be about the same, with the medians of the two regions almost equal. The mean sales of the NE region (indicated by the small dot) are higher than the mean sales in the MC region, but not by a very significant amount.

Given just these graphs, it might be difficult to determine which region is performing better overall. In general though, it seems that the NE region has more stores performing well than the NC region. Also, the highest performing store is in the NE region. Overall, it looks like the NE region has better sales, but we must remember that the NE region has fewer stores, so each quartile refers to fewer data points. The real question is what is causing the NE region to do better. Is it better management? Less overhead? Wealthier clients? Better marketing? Better service? Some other factor?

Example 4.7. Skewness of data
Generic University offers three sections of its math course for business majors. At the end of the semester, all sections take the same final exam. The boxplots in figure 4.5 (page 116) show the results of the final separately for each section. The graphs are oriented vertically, rather than horizontally, just for variety. As you can see from the graphs, the minimum scores are the same in each section, as are the maximum scores and the means. Which section did best on the final exam?

In order to decide which section did best on the final exam, we need to picture how the data itself looks, based on the boxplots. For example, the test scores in section 1 seem to be unevenly spread throughout the range of the data (low score: 40, high score: 99). We can tell this because the median of the data is very close to the upper end of the spread. Half

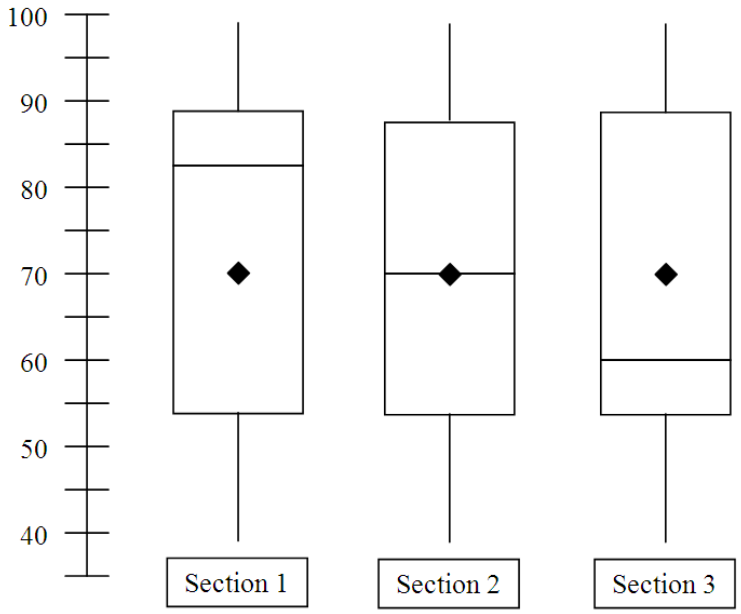

Figure 4.5: Side-by-side boxplots of final exam scores from three sections of a course.

of the students in section 1 scored above 83 on the exam. Even so, the overall mean of this section's test scores was only 70 because the lower 50% of the class has scores from 83 on down to 40. This unevenness is referred to as skewness. When the mean is smaller than the median, we say the data is negatively skewed because the quantity (Mean - Median) would be less than zero. This is in stark contrast with section 3, where half of the students' scores are bunched together at the low end of the spread, from 40 to 60, and the top half of the class has scores ranging from 60 up to 99. In this case, the mean is larger than the median, so the data is positively skewed. What about section 2? The data for this section doesn't seem to be skewed at all; the mean and median are identical. This tells us that half of the students in section 2 scored above 70 and half of the students in section 2 scored below 70. Given all of this, it seems reasonable to conclude that section 1 had the best showing on the exam; more than half of the students in section 1 had exam scores above the median and mean scores of both sections 2 and 3.

4.2.3 Exploration 4B: Relationships Among Data, Statistics, and Boxplots

For this exploration, we'll go back to the sales data for Cool Toys for Tots in two sales regions. The data is shown above in example 3 (page 106) and can be found in "C04 Tots.xls". To set the data up for the exploration, do the following:

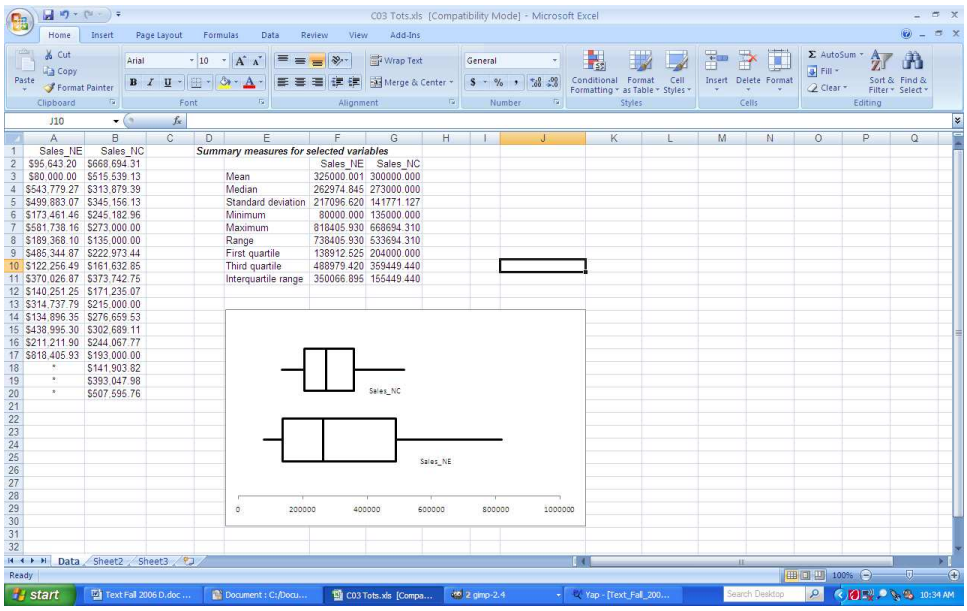

Figure 4.6: Side-by-side boxplots of Toys-For-Tots sales data.

1. Use StatPro to compute the mean, median, minimum, maximum, range, first quartile, third quartile and interquartile range for both regions. Place the statistics to the right of the data (not on a new worksheet, as usual). See the "How to Guide" for this section in order learn how to use StatPro to calculate the statistics.

2. Create side-by-side boxplots for the two regions. For instructions, see the "How to Guide" for this section.

3. Move the boxplot so that it is on the same worksheet as the data. To do this, right-click on the boxplot and select "Location..." from the context menu. Click the option for "As object in sheet1" and hit "OK". You should now have a worksheet with the data in cells A1:B20, the summary statistics in cells D1:G10, and a boxplot that you can move around (see image). Be sure to save this worksheet now so that you can return to the original version of it as often as needed by opening the file. Now, we want to explore what happens to the statistics and the boxplots as the sales information changes. You'll want to keep notes on what you observe happen (if anything) as you change the data in cells A1:B20. Explore the following questions.

4. What happens to the results (statistics and boxplots) as you change the sales figures? Be sure to keep notes on what kinds of changes you made. It may help to organize

these notes into a table with three columns labeled "Change I Made", "Change in Statistics", "Change in Boxplot".

5. What happens to the results (statistics and boxplots) if you decrease the number of stores in the data? What happens if you increase the number of stores? Can you explain this behavior?

6. What happens if many of the stores in the NE region have sales above $500,000? What if many of the stores in the NC region have sales below $225,000?

7. What changes need to be made so that the NE and NC regions perform about the same? What changes will make the NC region perform better? (Careful! There are easy answers to these questions, but go deep and find more realistic ways to get the results.)

4.2.4 How To Guide

For all of the information below, assume that the spreadsheet shown above in the "How To Guide" for section 4.1 (page 104) is being used. It contains data on sample sales figures from Toys-For-Tots (from Exploration 4.2.3 (page 117)) in cells A2:B20. For a picture of this spreadsheet, see figure 4.6 (page 117).

Summary statistics in StatPro

Statistical add-ins like StatPro often have a more convenient way to compute summary statistics than to enter formulas for the mean, median, and so forth, separately. Often, they include a routine that will compute all of the possible statistics, or a selection of them, at one time. Each routine in StatPro is structured the same way. The six steps below will take you through any of StatPro's useful routines. The only differences among the routines occur in step 5.

1. Select the region of the worksheet that contains the data.

   When using any StatPro routine, you must first select the data to which the routine will be applied. To do this, simply click on any cell inside the region of the worksheet that contains the data. For example, if the data is in the region A1:B20 (including the variables names), clicking on any cell or collection of cells in that region will work. Be careful, though! StatPro assumes that the region is rectangular, and it will include any rows or columns that have any data that touches the region you are interested in. This is why it is always a good idea to make sure there is at least one blank row and one blank column between the data region and any other information on the worksheet.

2. Select the StatPro routine to apply to the data.

   In this case, we want to use the statistics routines. Select the Add-Ins ribbon. Click on the "StatPro" menu, then select "Summary Statistics..." and choose "One variable summary statistics...".

3. Verify that the data region is correct.

   Check to be sure the highlighted section of the worksheet includes your data (and only your data), then click "OK". StatPro will then create a list of variables that are included in your data so that each variable can be referred to by name rather than by row and column references. If there is an error in the data region that has been selected by StatPro you can either hit "Cancel" and start the procedure over, being certain to select the data region correctly this time at step 1, or you can type the correct region into the dialog box in the blank next to the prompt "Data range:" (see figure 4.7.)

4. Select the variables to which the routine will apply.

   From the list that StatPro generates, choose the variable or variables you wish to get statistics for, and hit "OK" (see figure 4.8.) To select multiple variables that are listed next to each other, either click and drag with the mouse, or select the first variable in the list that you want and hold down the SHIFT key while selecting the last variable

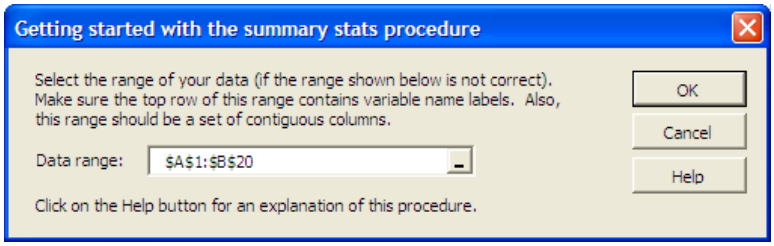

Figure 4.7: Dialog box in StatPro to verify that the correct data region is selected.

that you want. All the variables in between these two will be highlighted. If the variables that you want are not listed together, simply hold down the CONTROL (CTRL) key while selecting each variable with the mouse.

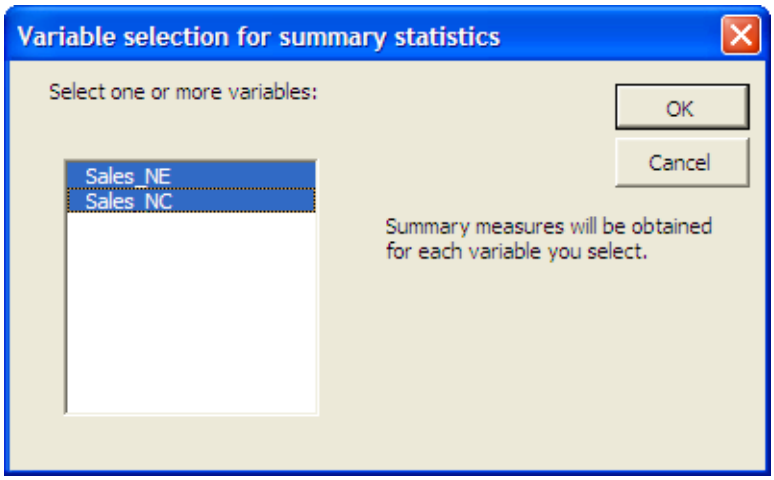

Figure 4.8: Dialog box in StatPro to select variables for the summary statistics procedure.

5. Fill in the details of the routine.

   Each StatPro routine has different information at this point, depending upon what the routine is designed to do with the data. For the one variable statistics you should see a dialog box like figure 4.9. Check off the statistics that you want computed (some are already checked by default) and hit "OK".

6. Select the placement for the output of the routine.

   Next, you will be asked where the statistics should be placed. You have three options, shown in figure 4.10. We recommend that you always place the computations, graphs, and statistics on separate worksheets in the workbook. Each worksheet should have a descriptive name, something like "Statistics," so that you know which worksheet contains the original data, and which contains the statistics on that data. Sometimes we will ask you to place the results of a routine on the same worksheet (to the right

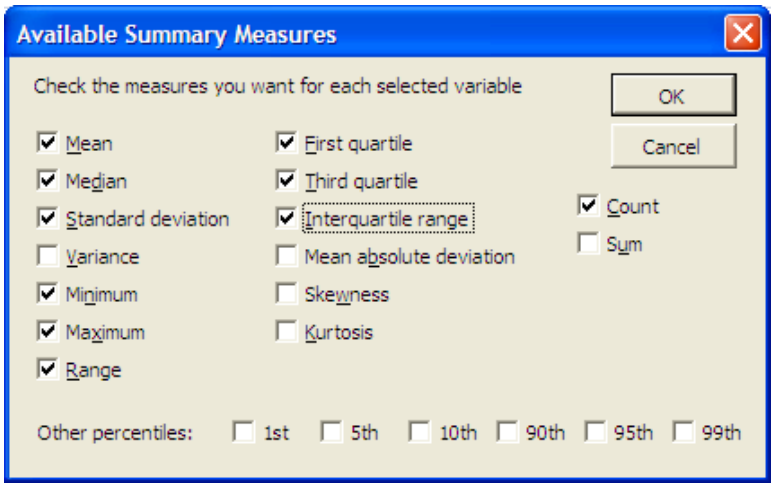

Figure 4.9: Dialog box in StatPro to select which summary statistics are computed.

of the data) so that you can see everything at once. We'll let you know when we want you to do this; otherwise, put everything on separate worksheets.

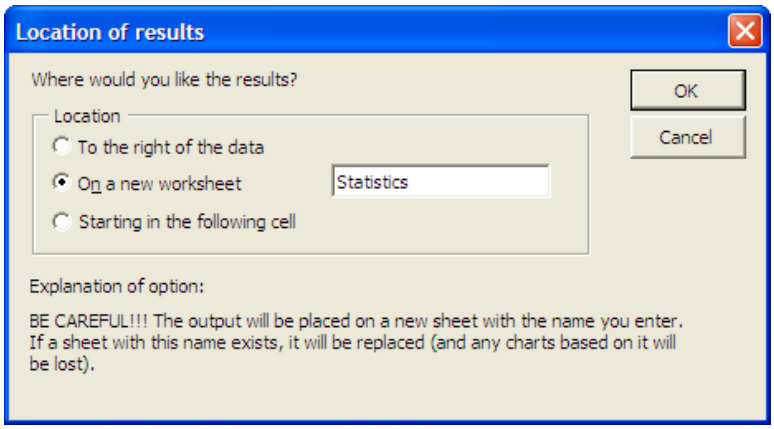

Figure 4.10: Dialog box in StatPro to select where to place the summary statistics.

Making Boxplots with StatPro

Microsoft Excel has the built-in ability to make many useful graphs; the boxplot, however, is not one of these. Fortunately, StatPro does include a powerful boxplot routine that has options for making one of three types of boxplots: a single boxplot, a side-by-side boxplot for several unstacked variables, or a side-by-side boxplot from a single variable that is stacked based on a second variable.

1. Select the region of the worksheet that contains the data. This step is the same as it is described above under "Summary Statistics in StatPro".

2. Select the StatPro routine to apply to the data. To make a boxplot, we want to select "StatPro", then "Charts" and then "BoxPlot(s)..."

3. Verify that the data region is correct. This is the same as above under "Summary Statistics in StatPro".

4. Select the variables to which the routine will apply. For boxplots, this step actually comes later, as part of step 5 because there are two types of boxplots that you can make. Each requires slightly different information.

5. Fill in the details of the routine. The next screen you see will ask you whether you want to make a single boxplot or side-by-side boxplots. Single boxplots are for displaying a boxplot of a single variable, while side-by-side boxplots are excellent for comparing two or more variables that are in the same units with about the same range.

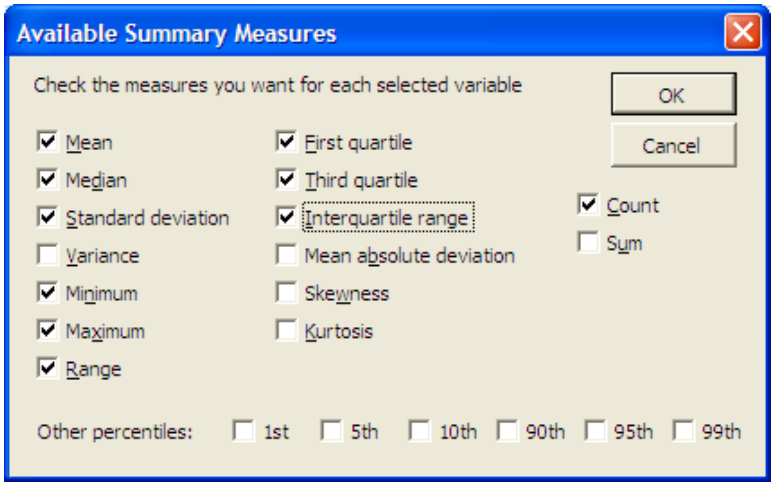

Figure 4.11: StatPro dialog box for selecting the statistics you want to compute.

6. (Option 1.) Making a single boxplot. After you select "single boxplot" from the list and hit "OK" you will be asked to select the variable (as in the usual step 4). After you hit "OK", StatPro will automatically create a new worksheet called "BOX - (your variable name)" and take you to it. Unfortunately, in Excel 2007, the boxplot procedure has a slight error. When it completes the graph, it selects the wrong line type for the box plot, producing something like the graph shown in figure 4.12.

   Fortunately, there is an easy way to fix this. First click on your newly created chart to activate the design ribbon for chart tools (figure 4.13.) On the left-hand side is the option to change the chart type. The only change you need to make is to select a different type of XY (scatter) chart. Instead of the one with the smooth lines connecting the points, boxplots should use the straight lines (no data points shown) option. See figure 4.14 for the chart type dialog box showing the correct option highlighted. One other change you may want to make for boxplots is to remove the legend (placed along the right-hand side of the graph) because it is totally meaningless.

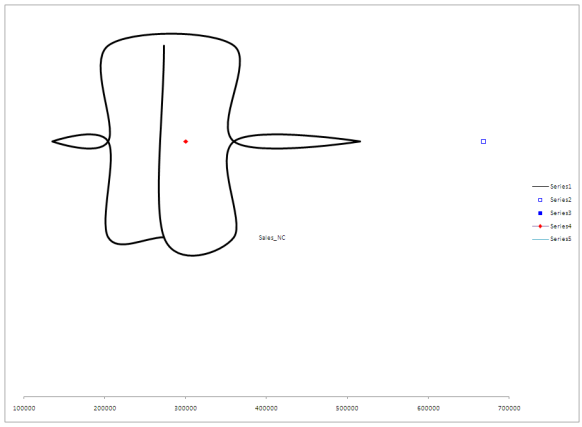

Figure 4.12: The default (incorrect) boxplot made by StatPro in Excel 2007.

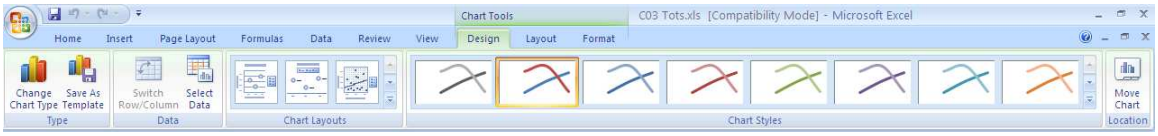

Figure 4.13: The Excel 2007 design ribbon for chart tools.

7. (Option 2.) Making side-by-side boxplots. If you select "side-by-side" boxplot, StatPro will ask you whether the data is stacked or unstacked. If you want to make several boxplots to compare different variables, select "unstacked", then choose your variables, and hit OK. The option to make boxplots of stacked data is only applicable if you have data in which one variable is a "Code" for another variable. For example, if you have one variable that lists house prices in a neighborhood, and another variable that describes the house type as "Ranch", "Cape Cod", or "Colonial", the stacked option will allow you to create three boxplots for house price: one for each type of house.

8. Select the placement for the output of the routine. This is handled automatically by the "Chart" routines in StatPro. All charts are placed on new worksheets.

Pasting graphs and charts from Excel into a Word document to make a report

Microsoft Windows-based computers have the powerful ability to share information between programs. This is especially useful when creating technical reports that involve data analysis. You can use the powerful data analysis routines of Excel and StatPro, and then share that information with Microsoft Word by copying and pasting the charts and computations. There are several ways to do this, however, and each is useful for certain types of information. When copying a chart from Excel into Word, the important thing to remember is that the chart is full-screen/full-page in Excel, but will only be about one-third as big in your report in Word. This means that all the text on your chart - the labels for the axes and the title of the chart, for example - will shrink to almost unreadable size when copying from Excel to

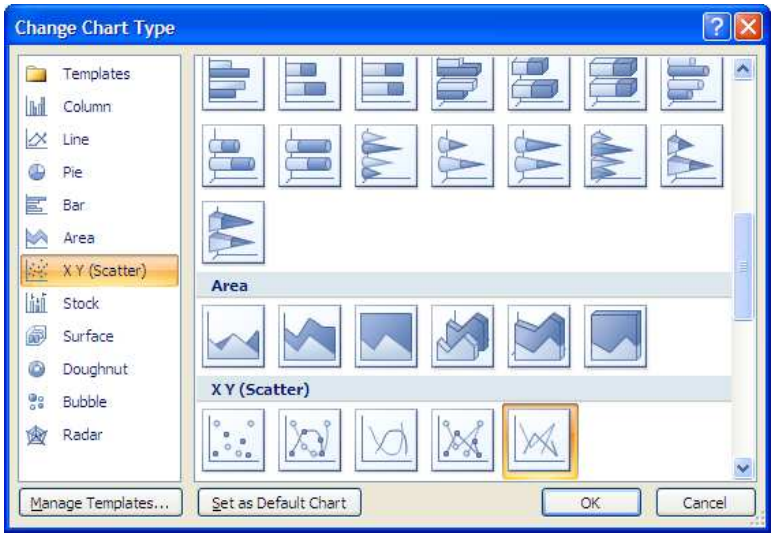

Figure 4.14: Selecting the correct chart type for boxplots in Excel 2007.

Word. This can be avoided by simply preparing the chart before you copy it.

1. Prepare the chart for copying.

   The first thing to do is to make the font size of the chart larger. To do this easily, right-click on the chart area and select "Font" from the context menu that appears (see figure 4.15.) Change the "size" setting to something larger. We recommend either 16- or 18-point font. That seems to carry over into Word very nicely. When you hit "OK" all the text on the chart should change to the size you selected.

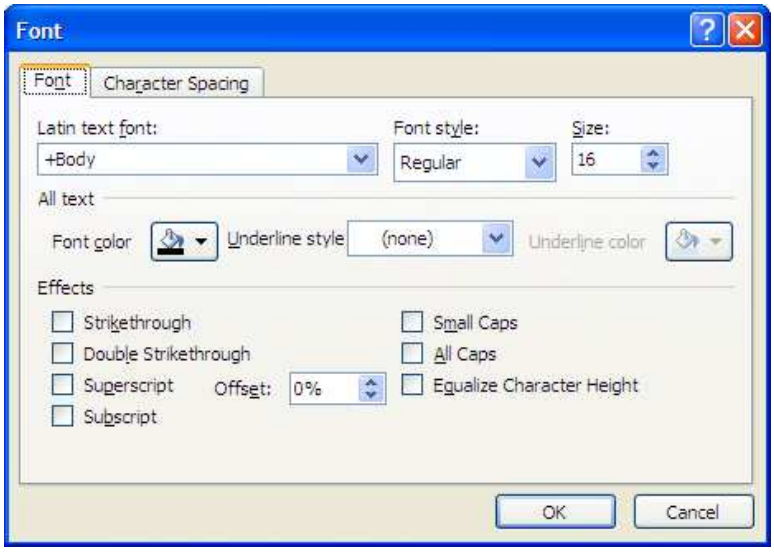

Figure 4.15: Modifying the chart for pasting into Word.

The other change you may want to make is to be sure the background color is turned off. Sometimes this doesn't come through well in Word, especially if you plan to print the report. To change the background, simply right-click on the chart and select "Format plot area..." from the context menu. Then, select "no fill" for the Fill setting.

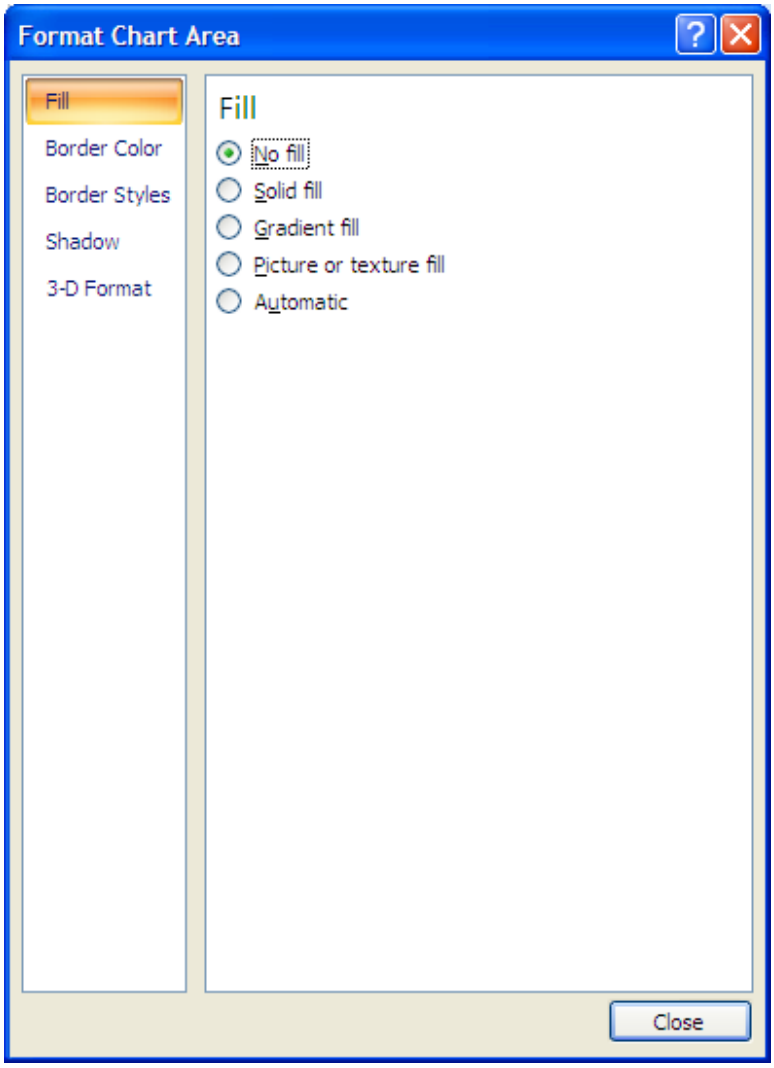

Figure 4.16: Changing the background fill of a chart before pasting into Word.

2. Copy the chart.

   As you probably know, there are many ways to copy and paste in Windows-based programs. We'll show you all of them here. Each begins with the same step: select the object to be copied. For a chart, you should click in the white region so that the entire chart is selected. This region is the same region where you double-clicked to change the font and is described under "step 1" above.

   (a) Using the ribbon. Now that the chart is selected, go to the Home menu and select "Copy".

(b) Using the keyboard. After selecting the chart, hit CTRL + C. This is the keyboard shortcut for the "Edit/ Copy" menu option above. You'll also notice that this keyboard shortcut appears on the edit menu next to the copy command.

(c) Using the mouse. Right-click on the chart and select "Copy".

Before we complete the process, it's important to understand a little about how Windows works with copying and pasting. Anytime you copy an object (a word, a chart, a paragraph, a file) it is placed in a temporary storage space called the Clipboard. Starting with Microsoft Office 2000, this clipboard can hold several objects at once. In Microsoft Office 2007, you can access the entire clipboard by clicking on the small arrow next to the word "clipboard" along the left-hand side of the Home ribbon. Once you have the entire clipboard displayed, you can select any item to paste.

3. Paste the chart into your report.

Now here is where it gets a little confusing. The easiest way to paste a chart is to place the cursor where you want the chart and either use the "Edit/ Paste" menu command, the keyboard shortcut CTRL + V, or right-click with the mouse and select "paste" from the context menu that appears. If you are pastign a chart from Excel, this will automatically select the format in which the object is pasted to be a Microsoft graphic object.

You can, however, choose other formats for pasting objects. By clicking on the small down arrow below "Paste" on the Home ribbon, you can use the "Paste Special.." option, which allows you to select many different formats. Usually, this is not necessary. In older versions of Word, the paste special feature could help by selecting a format that used less memory, making it easier to transfer and store files. The newer version is smarted about its default selection process, and modern memory storage and data transfer rates make saving a few kilobytes of space unnecessary.

4.3 Homework

4.3.1 Mechanics and Techniques Problems

4.1. Download the data file "C04 Salaries.xls". This data represents salaries for employees at a small company.

1. Add in two new columns of computed data: The first column should contain the salaries of each employee after a flat $1000 raise. The second column should contain the salaries after a 5% raise.
2. What are median and quartiles of these three different salaries? (Be sure to copy and paste these statistics from Excel).
3. What happened to the median and quartiles after the $1000 increase? Why?
4. What happened to the median and quartiles after the 5% increase? Why?
5. Describe (in words) how you think the boxplots would look different from the original boxplot for both (i) the fixed salary increase and (ii) the percent salary increase.

4.2. The boxplots below (figure 4.17) provide information about the people who tend to purchase your company's products. These data are reported as boxplots, one for the ages of the customers, one for the incomes, and one for the typical monthly credit card debt the customers carry. Use these boxplots to describe your typical customer. Make explicit reference to the quantities you can read from the boxplot directly and use these to describe your company's typical customer.

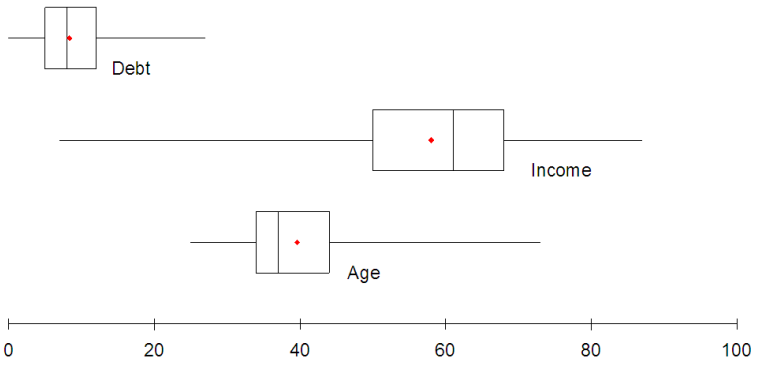

Figure 4.17: Boxplot for problem 2 showing income, age, and credit card debt distributions.

4.3. Consider the data shown in C04 MachineParts.xls. This data shows the diameter of 1,000 rods manufactured on your company's assembly line. The rods must be within 0.03 inches of being 0.50 inches in diameter to fit in the structure for which they are made.

1. Create a boxplot of this data. Determine how many data points are extreme outliers and how many data points are mild outliers.

2. Sort the original data and locate all the extreme outliers. Make a new column containing all the data except these outliers. Make a boxplot of the data without the extreme outliers.

3. Are there any outliers in the reduced data from part b? If so, eliminate them and redraw the boxplot. Continue doing this until there are no outliers in the data. (Hint: This should take two more rounds of eliminating mild and extreme outliers.)

4. Compare your final boxplot (with no outliers) to the original boxplot from part a. What can you learn about the data?

5. When reporting these data, should you include the outliers? Why or why not?

4.3.2 Application and Reasoning Problems

4.4. Place the data X1, X2, X3, X4, X5, X6, X7, X8, X9 along the horizontal axis of the boxplot in figure 4.18 so that it would produce a boxplot similar to the one shown. Assume that $X_1 < X_2 < X_3 < X_4 < X_5 < X_6 < X_7 < X_8 < X_9$.

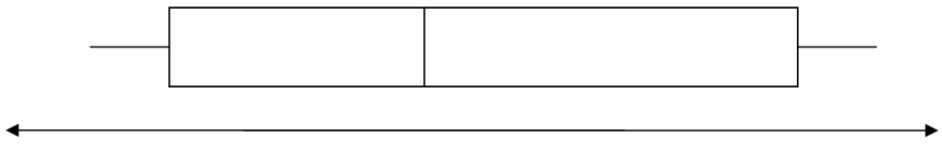

Figure 4.18: Boxplot for problem 4.

4.5. Place the data X1, X2, X3, X4, X5, X6, X7, X8 along the horizontal axis of the boxplot in figure 4.19 so that it would produce a boxplot similar to the one shown. Assume that $X_1 < X_2 < X_3 < X_4 < X_5 < X_6 < X_7 < X_8$.

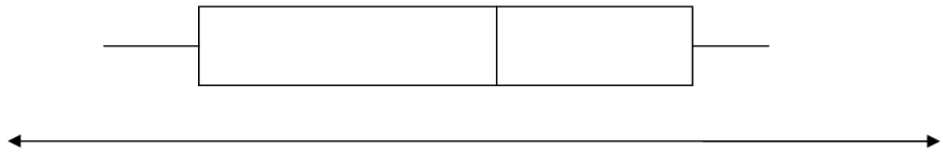

Figure 4.19: Boxplot for problem 5.

4.6. In figure 4.20, boxplot A represents the set Data 1 and Boxplot B represents the set Data 2. Given that the min and max of both data sets are the same answer the following questions:

1. In general, why is the length of box B longer than the length of box A?
2. Why is the median of B off to the right compared to A that is more central?
3. Even though the max of A and B are equal, why does the max of A appear as an isolated dot to the right of the whisker where as the max of B appears as the endpoint of the whisker? (Do not use calculations; give an "eyeball" explanation.)

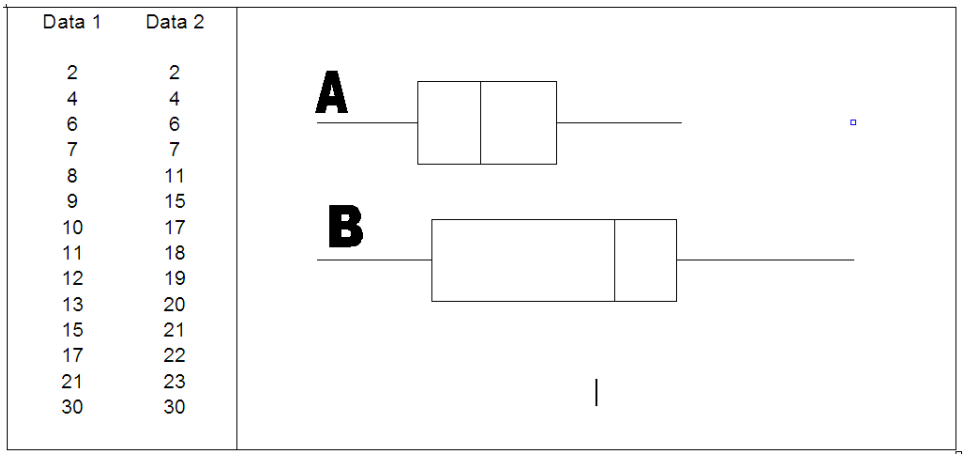

Figure 4.20: Boxplot for problem 6.

4.7. Consider the data on baseball players in 1992 that is in data file "C04 Baseball.xls". We are interested in whether the salaries of players eligible to be free agents in 1992 were significantly different from the salaries of players who were not eligible to be free agents. Use Excel to make side-by-side boxplots of the stacked salary data, using the variable "Free Agent Eligibility" as the code variable. Use your boxplot to explain whether being eligible to be a free agent has an effect on your salary. In your explanation, make use of all the statistics that the boxplots show you.

4.8. Consider the data and work you did in Mechanics and Techniques, problem 1.

1. Will these results occur for any fixed (flat) salary increase? What about a salary decrease? Explain your reasoning.
2. Will these results occur for any percentage increase in salary? What about a percentage decrease in salary? Explain your reasoning.

4.9. Download the data file "C04 Sales.XLS". The data shows sales figures from sixteen stores in our chain. We have plans to open new stores in the following cities: Honolulu, HI; Little Rock, AR; El Paso, TX; Tucson, AZ; and Hartford, CT. Generate sales figures at the five new stores that will result in the mean sales of the company exceeding the median sales of the company. To answer this problem, add five new data points to make the mean of the new data larger than the median of the new data. To demonstrate to me that your new data satisfies this criterion, copy the five data points you added, show me the box plot and summary statistics for your new set of data, and explain your thinking process for selecting these five points. In order to explain your thinking process, be sure to provide a box plot and the summary statistics for the original data in order to compare.

Organize your report as shown in figure 4.21. (NOTE: The image below is just a sample. Your graphs will probably not look like the ones below.) In addition, explain what else changed when you added the five new data points.

Keep in mind that it is unrealistic for all five of the new stores to turn in exceptional sales figures. It is more likely that the five new stores will exhibit a wide range of sales figures, with most of them falling in the inter-quartile range.

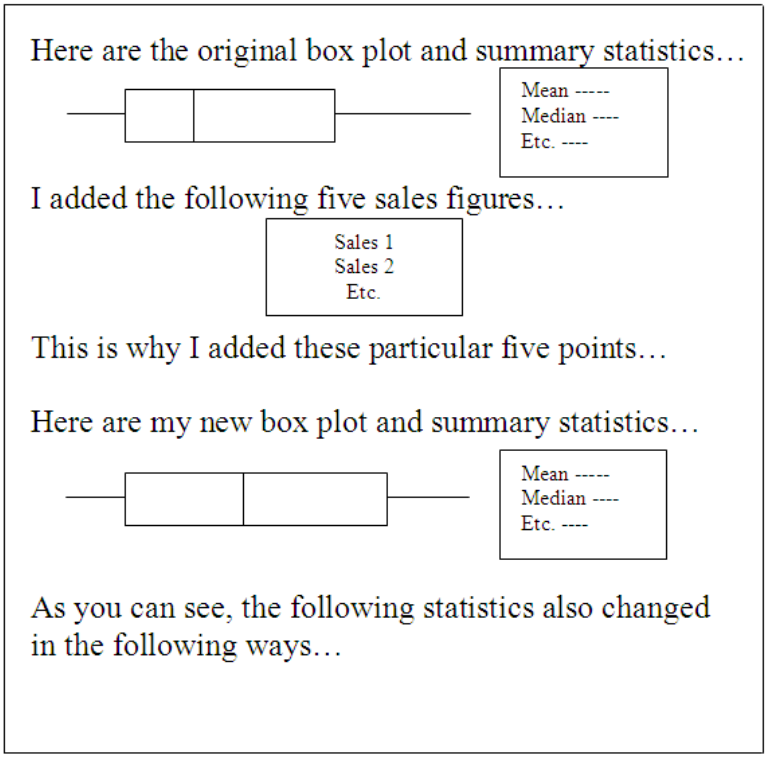

Figure 4.21: Sample report for problem 9.

4.3.3 Memo Problem

To: Job Placement Staff
From: Project Management Director
Date: May 18, 2008
Re: Placement of Managerial Clients

Since our company does management consulting, we have two middle-management clients who have come to us looking for management positions. Each of the clients is qualified to work at the four large companies in the local region. I need you to analyze the four companies in the attached data file and make a recommendation to each client as to which company each would be better suited to. The data file contains a list of the salaries at each of the four companies. There are about the same number of managers in each company with roughly the same ratios of middle- to upper managers in each.

Each of our clients has just moved out of the lower 25% management rank in his or her previous position. They are, however, quite different. Manager A is a confident go-getter who enjoys leaving the competition behind. Manager B, on the other hand, prefers to run with the pack. He wants to do well, of course, but stability and security are important.

To get started, you might consider generating comprehensive summary statistics and side-by-side box plots for these four companies. Based on what you learn from the box plots make a recommendation of a company for each client. Be sure to provide as much evidence as possible.

Attachment: Excel data file "C04 Companies.xls"

Follow-up Memo Problem

To: Job Placement Staff
From: Project Management Director
Date: May 20, 2008
Re: Follow-up on Placement of Managerial Clients

Our clients like the recommendations that we made, but they would each like an additional option. You are to make two recommendations for each client; that is, for A, you are to recommend two companies and for B you are to recommend two companies. If you haven't already, make sure you consider and discuss almost every part of the boxplot in making your recommendations. Moreover, be sure you point out the comparative advantages and disadvantages of your selected companies for each client according to that client's profile. I expect your work to be thoughtful and comprehensive.

Attachment: Excel data file "C04 Companies.xls"