Minitab

LESSON: FIVE-NUMBER SUMMARY AND BOXPLOTS
This lesson includes an overview of the subject, instructor notes, and example exercises using Minitab.

Five-Number Summary and Boxplots

Lesson Overview

The **five-number summary** of a data set provides descriptive statistics for the center, spread, and range of the data. Specifically, the five-number summary of a data set includes the following:

- minimum
- first quartile (Q_1)
- median or second quartile (Q_2)
- third quartile (Q_3)
- maximum

A **boxplot** is a graphical display of the five-number summary displaying the shape, center, spread, and extreme points of a data set. The first and third quartiles frame this rectangular plot.

In this lesson, the five-number summary statistics will be defined and examples will be provided. The construction of the boxplot using the five-number summary will be shown.

Prerequisites

This lesson requires basic mathematical operations and knowledge of basic graphing techniques. In Minitab, graphs will be constructed on single and multiple columns of data.

Learning Targets

This lesson teaches students how to:

- Generate the five-number summary statistics
- Construct a boxplot by hand and in Minitab
- Construct a boxplot using the five-number summary and identify any skewness or outliers
- Construct side-by-side boxplots for comparing data sets

Time Required

It will take the instructor 30 minutes in class to introduce the five-number summary, a single boxplot, and multiple comparison boxplots. We recommend starting the activity sheet in class so that students can ask the instructor questions while working on it. The exercises on the activity sheet will take an additional 30 minutes, and they can be used as homework or quiz problems.

Materials Required

- Minitab 19 or Minitab Express
- Minitab worksheet of sample data, entitled ***Boxplot_Lesson.mtw***
- Internet access (optional example)

Assessment

The activity sheet contains exercises for students to assess their understanding of the learning targets for this lesson.

Possible Extensions

This lesson provides a good introduction to new summary statistics for a data set and the boxplot. The instructor may want to do the lessons ***Describing Data Numerically*** and ***Describing Data Graphically*** before this lesson so that students see other mainstream descriptive statistics and more basic plots first.

References

The Minitab Blog: Tooltips, Assistant Menu, and Help: The 5 Coolest Things You Didn't Know You Could Copy From Minitab: http://blog.minitab.com/

The Impossible Quiz website: http://www.notdoppler.com/theimpossiblequiz.php

Instructor Notes with Examples

Five-Number Summary

Definition: The **five-number summary** of a data set provides descriptive statistics for the center, spread, and range of the data. Specifically, the five-number summary includes:

- minimum
- first quartile (Q_1)
- median or second quartile (Q_2)
- third quartile (Q_3)
- maximum

Definition: The **first quartile (Q_1)** of a data set is the ordered data value such that 25% of the values in the data set are less than or equal to it.

Definition: The **second quartile (Q_2)**, or **sample median**, of a data set is the middle ordered data value. 50% of the values in a data set are less than or equal to the second quartile.

Definition: The **third quartile (Q_3)** of a data set is the ordered data value such that 75% of the values in the data set are less than or equal to it. Alternatively, 25% of the values in the data set are greater than or equal to the third quartile.

Definition: Subtracting the first quartile from the third quartile, $Q_3 - Q_1$, is the **interquartile range**, or **IQR.** It is another measure of spread, besides the standard deviation and range. The IQR is the spread of the middle 50% of the data set.

Example 1

The following data set is a sample of retirement ages. The data set is already ordered from the minimum to the maximum value.

59 60 64 67 68 68 70 71 72 73 73

Using Minitab's formulas for calculating quartiles, which will be discussed after this example, the five-number summary of this data set is:

Minimum = 59 Q_1 = 64 Median Q_2 = 68 Q_3 = 72 Maximum = 73

For the moment, don't worry about how the quartiles are being calculated. The point of this example is to show that approximately:

- 25% of the n = 11 data values are less than or equal to Q_1 = 64
- 50% of the n = 11 data values are less than or equal to Q_2 = 68
- 75% of the n = 11 data values are less than or equal to Q_3 = 72

The interquartile range for this data set is $Q_3 - Q_1 = 72 - 64 = 8$. This means that the spread of the middle 50% of the data is 8. The range of the entire data set is Maximum – Minimum = 73 – 59 = 14.

Important notes:

- Quartiles are not necessarily observations in the data set as shown in this example. Quartiles are calculated values, and it is often necessary to interpolate between two ordered data values to determine a quartile.
- Because quartiles are not affected by extreme observations, the median and interquartile range are better measures of center and spread, respectively, than the mean and standard deviation for highly skewed data.

Quartile Calculations

Surprisingly, statisticians do not use one standard formula for computing quartiles; there are a variety of methods for a sample data set of size n. Since we will be using Minitab to calculate descriptive statistics and to make boxplots, we will use Minitab's definitions for computing quartiles. For the computations that follow, assume the sample size is n.

Calculating Q_1: Compute the value $(n+1)/4$. The first quartile is the ordered data value in position $(n+1)/4$. Interpolation between two ordered data values may be necessary to compute the first quartile.

For example, if $n = 30$, then $(n+1)/4=7.75$ and Q_1 is between the 7th and 8th ordered data values. Interpolation is required since Q_1 is closer to the 8th data value than the 7th.

Calculating Q_2: The second quartile is the middle ordered data value if the sample size n is odd and the average of the middle two ordered data values if the sample size n is even.

Calculating Q_3: Compute the value $(3(n+1))/4$. The third quartile is the ordered data value in position $(3(n+1))/4$. Interpolation between two ordered data values may be necessary to compute the third quartile.

For example, if $n = 30$, then $(3(n+1))/4=23.25$ and Q_3 is between the 23rd and 24th ordered data values. Interpolation is required since Q_3 is closer to the 23rd data value than the 24th.

Example 2

For the sample of retirement ages given above, n = 11, $\frac{n+1}{4} = 3$ and $\frac{3(n+1)}{4} = 9$. Thus, Q_1 is in the 3rd ordered data position and is the value 64. Q_3 is in the 9th ordered data position and is the value 72. No interpolation is required to determine Q_1 and Q_3.

Example 3

Below is a sample of n = 26 trials of rolling a fair 6-sided die until a 2 is obtained.

13	11	2	10	1	2	4	7	4	15	1	6	9
2	3	11	14	3	1	3	7	7	10	29	2	2

In order to determine the five-number summary of this data set, we first need to order the data from the minimum to the maximum.

1	1	1	2	2	2	2	2	3	3	3	4	4
6	7	7	7	9	10	10	11	11	13	14	15	29

Important observations:

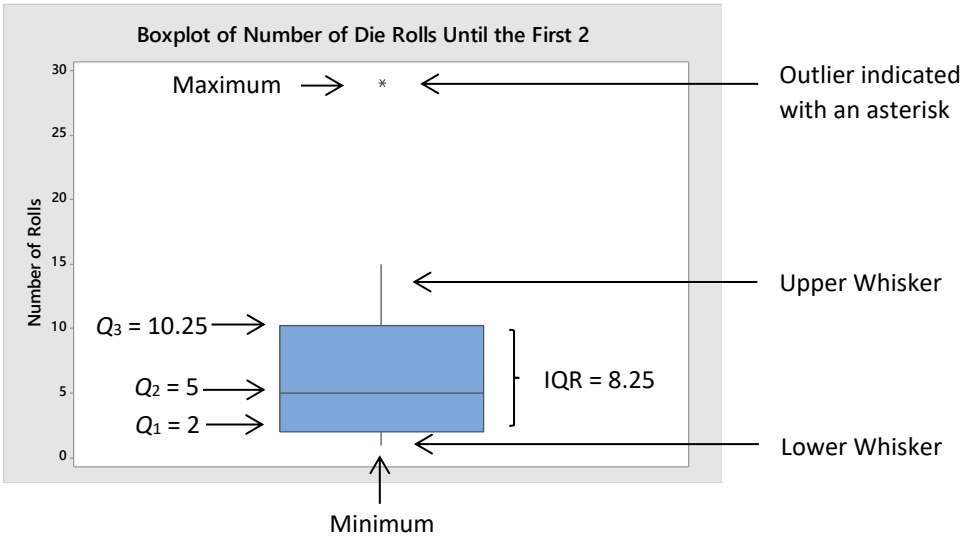
- Since n is even, the median is the average of the 13th and 14th ordered data points; it is 5. Approximately 50% of the data is less than or equal to the value 5.
- The position of the first quartile is between the 6th and 7th ordered data points since $\frac{n+1}{4} = 6.75$. Since the 6th and 7th data points are both 2, then Q_1 = 2. Approximately 25% of the data values are less than or equal to 2.
- The position of the third quartile is between and 20th and 21st ordered data points since $\frac{3(n+1)}{4} = 20.25$. The value 20.25 indicates that we calculate Q_3 as the value that is 0.25 of the way between the 20th and 21st ordered data points; Q_3 is 10.25. Approximately 75% of the data values are less than or equal to 10.25.
- The interquartile range is IQR = 10.25 – 2 = 8.25. Approximately 50% of the data is spread between 2 and 10.25.
- The minimum data value is 1 and the maximum is 29. The range of the data set is 28.

Constructing Boxplots By Hand

A useful graphical representation of a data set including its five-number summary is a **boxplot**. Besides indicating the shape, center, and spread of a data set, boxplots can be used to identify

extreme values, called **outliers**. Examples of boxplots and constructing them are on the following pages.

Below is a boxplot of this data constructed in Minitab.



Important components of the boxplot as seen above:

- **Minimum** (1): It is located at the end of the lower whisker.
- **Maximum** (29): It is denoted by an asterisk.
- **First Quartile, Q_1** (2): It is the bottom of the rectangular box.
- **Third Quartile, Q_3** (10.25): It is the top of the rectangular box.
- **Second Quartile, Q_2 or median** (5): It is the line between and parallel to the top and bottom of the rectangular box.
- **Lower Whisker** (1): It is the line that extends from the bottom of the box (Q_1) to the smallest observation within the lower limit, or lower fence. This limit will be more clearly explained later.
- **Upper Whisker** (15): It is the line that extends from the top of the box (Q_3) to the largest observation within the upper limit, or upper fence. This limit will be more clearly explained later.
- **IQR** (8.25): It is the length of the rectangular box from top (Q_3) to bottom (Q_1).
- **Outlier** (29): It is an actual data value that is beyond the upper or lower whisker; it is a large or small data value beyond a given upper or lower limit.
  - On a boxplot, outliers are identified by an asterisk (*).
  - Outliers in a data set have an impact on some descriptive statistics, such as the mean, range, and standard deviation of that data set.

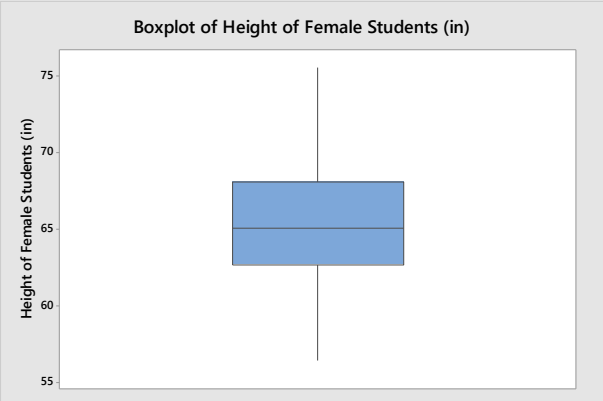
When to construct a boxplot for a data set:

- A boxplot is typically used when the sample size is at least $n = 20$.
- If the sample size is too small, the boxplot's quartiles and outliers may be meaningless.
- If the sample size is less than 20, consider using an individual value plot, dotplot, or stem-and-leaf plot as described in the ***Describing Data Graphically*** lesson.

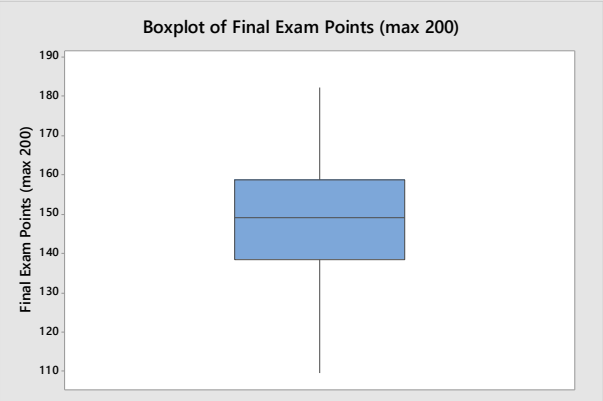
Boxplots are particularly useful in showing the shape of a data set.

- A symmetric data set has its median roughly in the middle of the boxplot.
- A lopsided boxplot indicates that the data set is skewed and non-symmetric.
  - If the median is closer to Q_1 than Q_3, the data is said to be positively or right skewed. Also, the upper tail (whisker and outliers) of the boxplot is typically longer than the lower tail when the data is right skewed.
  - If the median is closer to Q_3 than Q_1, the data is said to be negatively or left skewed. Also, the lower tail (whisker and outliers) of the boxplot is typically longer than the upper tail when the data is left skewed.

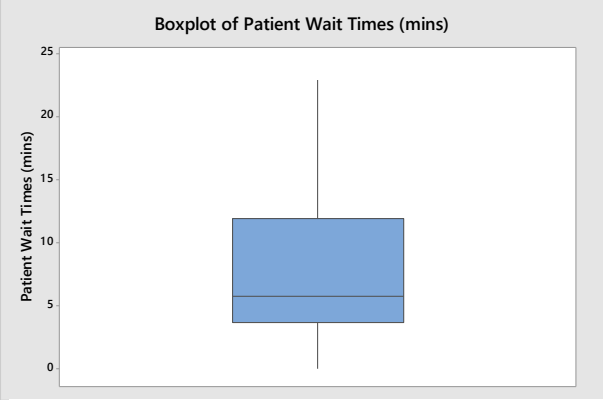
Below are boxplots showing symmetric and skewed data sets.



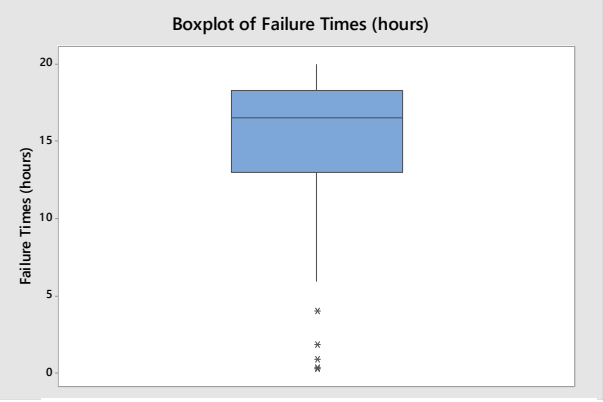
Female heights displaying symmetry



Exam grades displaying symmetry



Patient wait times displaying positive skewness



Failure times displaying negative skewness

How to construct a boxplot by hand:

1. $\boldsymbol{Q_1}$ and $\boldsymbol{Q_3}$: Parallel to the axis of interest, **construct a box** with its **bottom** as the first quartile (Q_1) and its **top** as the third quartile (Q_3).
2. $\boldsymbol{Q_2}$: Draw a **line through the box** at the second quartile or median (Q_2).
3. **Whiskers**: Draw a line from $\boldsymbol{Q_1}$ to the smallest value that falls within the **lower limit**, and draw a line from $\boldsymbol{Q_3}$ to the largest value that falls within the **upper limit**.
    a. The lower limit, or lower **fence**, is equal to $Q_1 - 1.5 * IQR$.
    b. The upper limit, or upper **fence**, is equal to $Q_3 + 1.5 * IQR$.
    c. In some boxplots, though not Minitab's, the lower and upper fences are drawn on the plot using dotted lines.
4. **Outliers**: If any data values lie beyond the upper or lower whiskers, then draw an **asterisk** (*) for each outlier.

Example 4

The following data are a random sample of $n = 50$ times (in minutes) that you read a statistics book for enjoyment at night before falling asleep. The data values have already been sorted from minimum to maximum.

8	9	9	9	9	10	10	11	11	12	12	12	12	12
12	13	13	13	13	14	14	14	14	15	15	15	15	15
15	16	16	16	17	17	17	17	17	17	17	18	18	19
19	19	19	20	20	20	22	23						

The five-number summary for the data is:

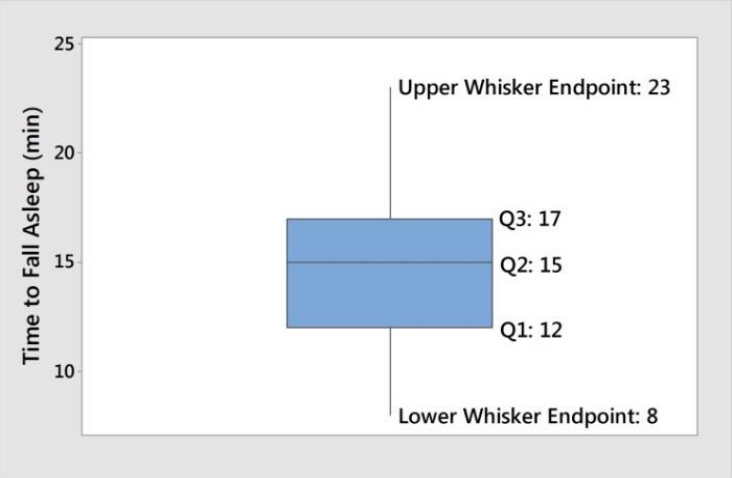
- Minimum = 8 minutes
- Maximum = 23 minutes
- Second quartile or median (Q_2) = 15 minutes (in ordered position 25.5)
- First quartile (Q_1) = 12 minutes (in ordered position 12.75)
- Third quartile (Q_3) = 17 minutes (in ordered position 38.25)

We need to calculate the lower and upper fences to determine the endpoints of each whisker and if the data set has any outliers.

- Lower fence = 12 – 1.5 * 5 = 4.5 minutes
- Upper fence = 17 + 1.5 * 5 = 24.5 minutes

Since there are no data points beyond the fences, this data set has no outliers. The upper whisker extends to 23 (the maximum) and the lower whisker extends to 8 (the minimum). The data is only slightly skewed.

Below is a basic boxplot of the "Time to Fall Asleep (min)" data.



Example 5

A random sample of n = 26 pesticide contamination levels in parts per million (ppm) from Pennsylvania lakes was collected. The data values have been sorted from minimum to maximum.

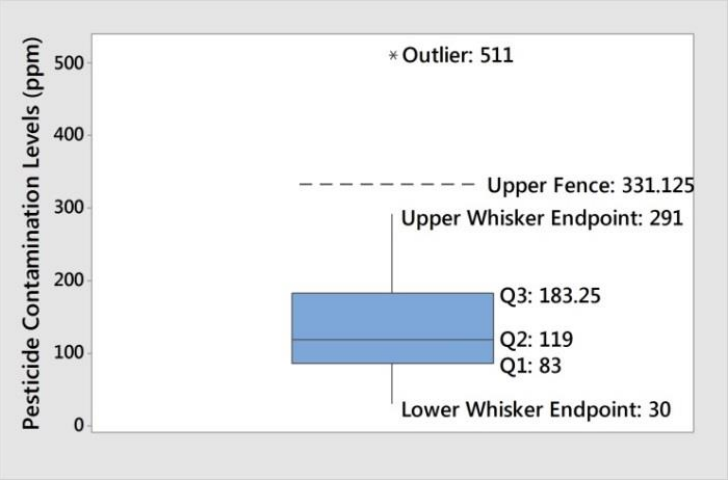
30	30	60	63	70	79	87	90	101	102	115	118	119
119	120	125	140	145	172	182	183	191	222	244	291	511

The five-number summary, along with lower and upper fences, for the data is:

- Minimum = 30 ppm
- Second quartile or median (Q_2) = 119 ppm (in ordered position 13.5)
- First quartile (Q_1) = 83 ppm (in ordered position 6.75)
- Third quartile (Q_3) = 182.25 ppm (in ordered position 20.25)
- Lower fence = 83 – 1.5 * 99.25 = -65.875 (but realistically for ppm, the lower fence should be set to 0)
- Upper fence = 182.25 + 1.5 * 99.25 = 331.125 ppm

The data value 511 is beyond the upper fence and is denoted as an outlier with an asterisk. The data is slightly positively skewed since the median is closer to Q_1 than Q_3 and the upper tail is longer than the lower tail.

Here is a basic boxplot of the "Contamination Levels" data:



Constructing Boxplots in Minitab

Now we'll construct a boxplot in Minitab for a single set of data. As is the case with the boxplot constructed by hand, a boxplot in Minitab consists of:

- A box, whiskers, and outliers.
- A line drawn in the box to indicate the second quartile or median (Q_2).
- A line drawn at the bottom of the box and parallel to the median line to indicate the first quartile (Q_1).
- A line drawn at the top of the box and parallel to the median line to indicate the third quartile (Q_3).
- Whiskers, which are the lines that extend from the box to the smallest and largest data values within the fences. The lower fence is 1.5 * IQR from Q_1, and the upper fence is 1.5 * IQR from Q_3.
- Outliers, which are data values beyond the fences.

Example 6

Use the same random sample of $n = 26$ pesticide contamination levels in parts per million (ppm) collected from Pennsylvania lakes in **Example 5**.

To compute the five-number summary for "Contamination Levels" column in the Minitab worksheet ***Boxplot_Lesson.mtw***:

Minitab 19 (Mac and PC)

1 Choose **Stat > Basic Statistics > Display Descriptive Statistics**.
2 In **Variables**, enter *'Contamination Levels'*.
3 Click **Statistics,** and then select **Minimum**, **Maximum**, **First quartile**, **Median**, **Third quartile**, **Interquartile range**, and **N total**.
4 Click **OK** in each dialog box.

Minitab Express

1 Open the descriptive statistics dialog box.
   - Mac: **Statistics > Summary Statistics > Descriptive Statistics**
   - PC: **STATISTICS > Describe > Descriptive Statistics**
2 In **Variable**, enter *'Contamination Levels'*.
3 Click the **Statistics** tab, and then select **Minimum**, **Maximum**, **First quartile**, **Median**, **Third quartile**, **Interquartile range**, and **N total**.
4 Click **OK**.

The Minitab output is:

Statistics

Variable	Total Count	Minimum	Q1	Median	Q3	Maximum	IQR
Contamination Levels	26	30.0	85.0	119.0	182.3	511.0	97.3

Notice that Minitab rounds Q_3 and the IQR, compared to the manual calculations on page 10.

To create a boxplot in Minitab:

Minitab 19

1 Choose **Graph > Boxplot**.
2 Choose **One Y - Simple**, then click **OK**.

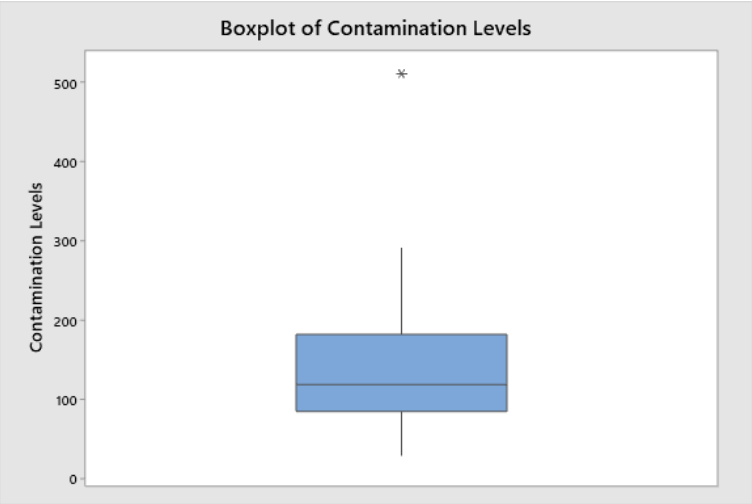
3 Under **Graph variables**, enter '*Contamination Levels*'.
4 Click **OK**.

Minitab 19 Mac

1 Choose **Graph > Boxplot**.
2 Choose **One Y Variable - Simple**.
3 In **Y-variable**, enter '*Contamination Levels*'.
4 Click **OK**.

Minitab Express

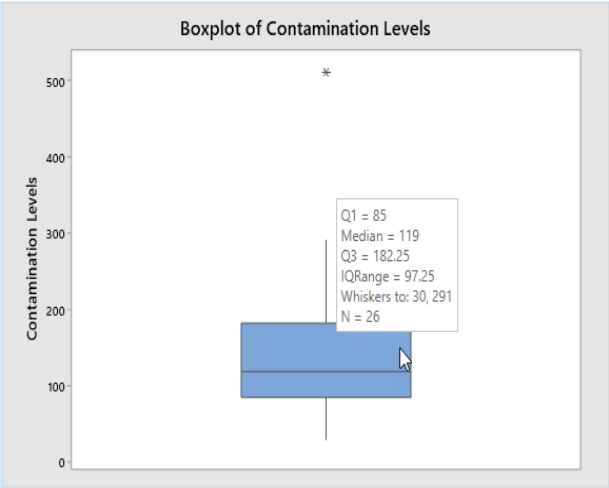
1 Open the boxplot of a single y variable dialog box.
   - Mac: **Graphs > Boxplot > Single Y Variable > Simple**
   - PC: **GRAPHS > Boxplot > Single Y Variable > Simple**
2 In **Y variable**, enter '*Contamination Levels*'.
3 Click **OK**.

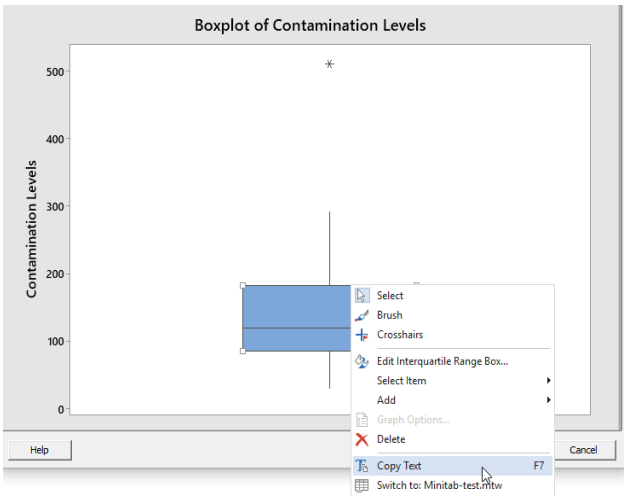


To obtain the outlier's value, hover the cursor over the outlier and Minitab will display its value. To obtain additional information about the boxplot, such as quartile values, hover over the boxplot itself. Minitab will display a tooltip containing the boxplot's quartiles, IQR, whisker endpoints, and sample size.

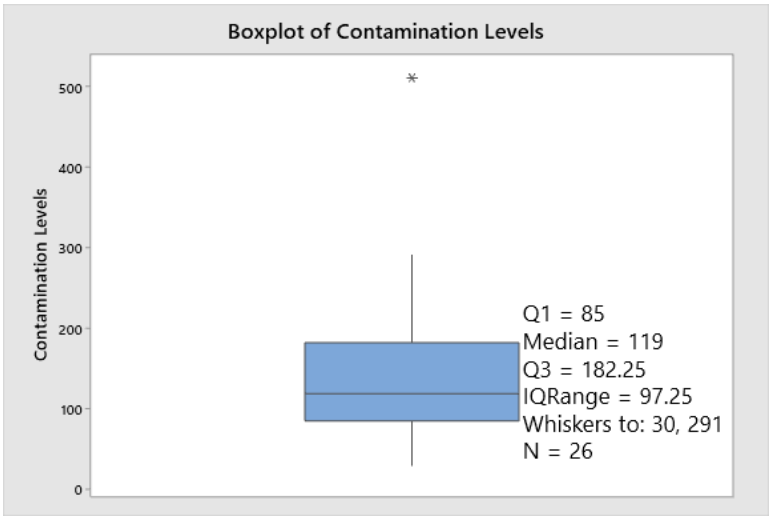
In Minitab 19, you can copy the tooltip information onto the plot using these steps:

1 Right-click on the graph and select Edit Graph.
2 In the Pop-out window, hover over the boxplot to display the tooltip (see the left boxplot below).
3 Right-click on the boxplot and select **Copy Text** (see the right boxplot below).





4 Press **Ctrl+V** to paste the tooltip information onto the boxplot.
5 (Optional) Click and drag the textbox to move it within the plot to the desired location.



Example 7

Boxplots are useful graphs for comparing data sets.

In Whoville, home to the merry Whos, the snowiest month of the year is December. The annual snowfall in Whoville, in inches, for each December from 1972 to 2013 is given below.

0.2	3.7	1.2	13.7	1.5	0.2	1.7	0.6	0.1	8.9	1.9	5.5	0.5	3.1
3.1	8.9	8.0	12.7	4.1	0.3	2.6	1.5	8.0	4.6	0.7	0.7	6.6	4.9
0.1	4.4	3.2	11.0	7.9	0.0	1.3	2.4	0.1	2.8	4.9	3.5	6.1	0.1

In Grinchville, home to the grouchy Grinch and his dog Max, the snowiest month of the year is also December. The annual snowfall in Grinchville, in inches, for each December from 1972 to 2013 is given below.

9.8	10.5	5.9	10.1	12.7	8.3	10.1	11.1	12.2	8.6	14.1	9.3	7.7	14.5
12.1	9.1	7.7	10.1	11.7	9.5	8.7	10.2	9.3	12.8	11.7	8.6	12.3	8.8
7.3	10.0	11.2	11.8	8.9	8.5	10.6	9.5	12.6	9.5	7.0	7.3	11.1	8.5

To create comparison boxplots in Minitab:

Minitab 19

1 Choose **Graph > Boxplot.**
2 Choose **Multiple Y's** - **Simple**, then click **OK**.
3 Under **Graph variables**, enter '*Snowfall in Whoville*' and '*Snowfall in Grinchville*'.
4 Click **OK**.

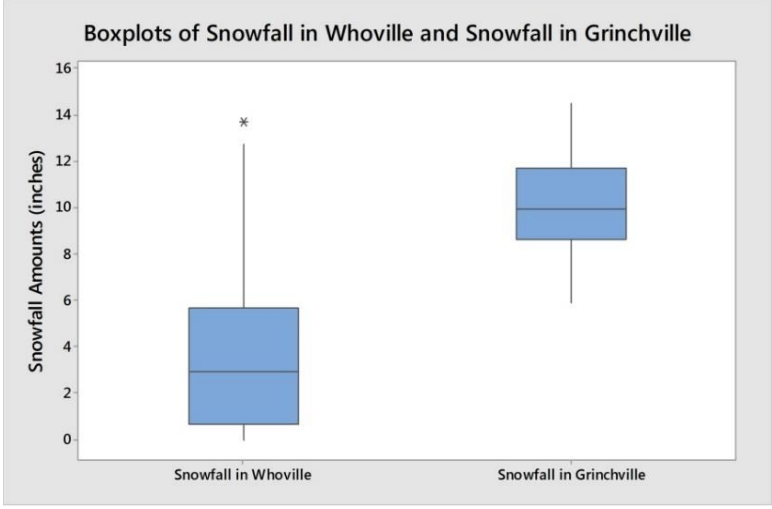
Minitab 19 (Mac)

1 Choose **Graph > Boxplot.**
2 Choose **Multiple Y Variables** – **With Categorical Variables**.
3 In **Y-variables**, enter '*Snowfall in Whoville*' and '*Snowfall in Grinchville*'.
4 Click **OK**.

Minitab Express

1 Open the boxplot of multiple y variables dialog box.
   - Mac: **Graphs > Boxplot > Multiple Y Variables > Simple**
   - PC: **GRAPHS > Boxplot > Multiple Y Variables > Simple**
2 In **Y variables**, enter '*Snowfall in Whoville*' and '*Snowfall in Grinchville*'.
3 Click **OK**.

Minitab produces the comparison boxplot below. We have retitled the plot and the snowfall axis. Editing graph text is discussed in the ***Describing Data Graphically*** lesson.



Notice the following about the comparison boxplots:

- The median snowfall for these 42 years is greater in Grinchville than in Whoville.
- Snowfall in Whoville demonstrates greater variability than in Grinchville, with an interquartile range of 4.975 inches.
- The annual snowfall amounts in Whoville are positively skewed, with an outlier snowfall amount of 13.7 inches in 1981.
- The annual snowfall amounts in Grinchville are close to being symmetric about the median.