LESSON: DESCRIBING DATA *GRAPHICALLY*

This lesson includes an overview of the subject, instructor notes, and example exercises using Minitab.

Describing Data *Graphically*

Lesson Overview

Statistics is the discipline concerned with the **optimal acquisition** (where garbage in equals garbage out) and **analysis of data** in order to **model** a population or process.

We can begin to analyze a data set by describing it both numerically and graphically. This lesson considers important **graphical summaries** of data, including dotplots, histograms, and stem-and-leaf plots. In this lesson, we are only considering quantitative (numeric) data, not qualitative (categorical) data. For the data sets of interest, we will select only one variable of interest; that is, we will be working with univariate data, not bivariate or multivariate data. Note that boxplots are discussed in a separate lesson.

Prerequisites

This lesson requires 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:

- Construct a dotplot for a data set
- Construct a histogram for a data set
- Construct a stem-and-leaf plot for a data set
- Consider the shape, center, spread, and skewness of a data set

Time Required

It will take the instructor 20 minutes in class to introduce the graphical summaries. 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 40 minutes, and they can be used as homework or quiz problems.

Materials Required

- Minitab 17 or Minitab Express
- Minitab worksheet of sample data, entitled ***DescribingDataGraphically_Lesson.mtw***

Assessment

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

Possible Extensions

This lesson provides good introductory examples for students new to statistics. The instructor may want to do the ***Describing Data Numerically*** lesson first so that students know how quantitative measures of center and spread for a data set are calculated.

References

Minitab 17 Online Support: What is a stem-and-leaf plot?
http://support.minitab.com/minitab/17/topic-library/basic-statistics-and-graphs/graphs/graphs-of-distributions/stem-and-leaf-plots/stem-and-leaf-plot/

Instructor Notes with Examples

Dotplots

Often the graphical method selected for a data set is determined by the type of data that you have (e.g. quantitative versus qualitative), the size of the data set (e.g. n small or large), and the main characteristic(s) of the data that you would like to show (e.g. spread, skewness).

One of the most basic graphs for a single data set is the **dotplot**.

Definition: Dotplots or dot diagrams represent each observation by a dot on a single numerical axis.

The dotplot:

- Is used for smaller data sets, such as *n* < 50.
- Displays the basic shape, center, and spread of data.
- Can highlight points that are unusual observations or **outliers**.
- Is easy to construct and read.

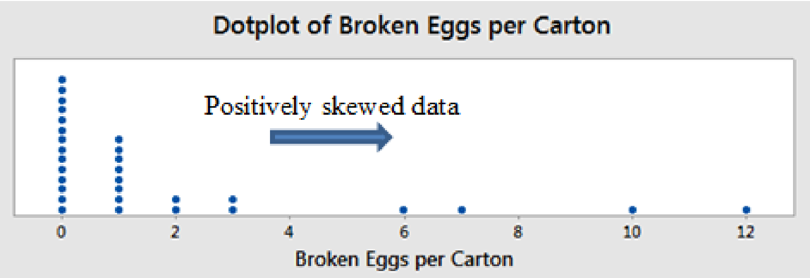
Minitab instructions for creating dotplots will be discussed later in this lesson.

Example 1

You take a trip to your local grocery store to pick up a carton of a dozen eggs. You open a carton, and there is one cracked egg inside. This gets you thinking – how many of the cartons on the grocery store shelf contain cracked eggs? Further, how many of the eggs in each carton are cracked? You select 30 cartons of a dozen eggs and record the number of cracked eggs in each carton.

Below is the raw data, sorted from smallest to largest, for the sample of size *n* = 30 cartons. Below the data is its dotplot. This data is in the column “Broken Eggs per Carton” in the Minitab worksheet ***DescribingDataGraphically_Lesson.mtw***.

0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	1	1	1	1	1	1	2	2	3	3	6	7	10	12



As seen in the dotplot, the data is **skewed**, where **skewness** is the extent to which the data is not symmetrical. We say the data is **positively skewed** or **right skewed** because the “tail” of the graph pulls to the right.

Histograms

One of the most popular graphs for a single set of data is the **histogram**.

Definition: A **histogram** is a graphical way to display the frequency of data points within a particular data set.

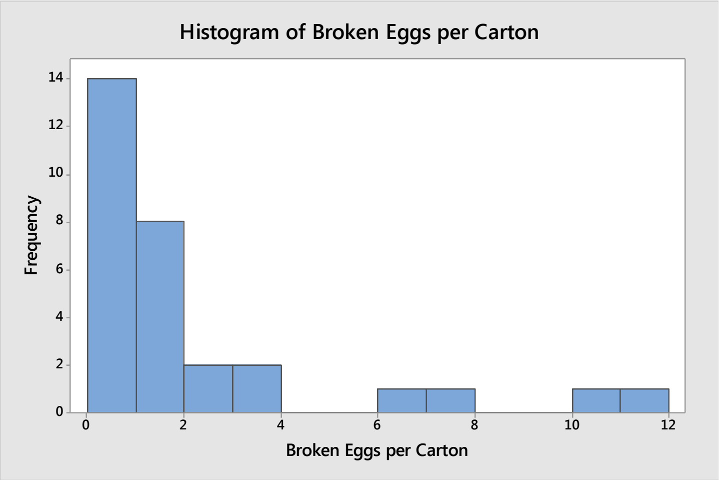
The histogram:

- Is typically used for larger data sets, such as $n > 50$. It is not a good graph if you only have a few data points.
- Displays the basic shape, center, and spread of data.
- Condenses large data sets into manageable and readable graphs.
- Is easy to construct and read once the histogram binning method is determined and clearly defined.

Pointers for constructing histograms by hand:

- It's best to keep the bins the same width; otherwise, the histogram can be hard to read.
- Be careful not to use too many or too few bins. In Minitab, you are able to adjust the binning structure, which is useful in displaying the features of a data set that you want your reader to notice.
- ALWAYS CAREFULLY LABEL ALL PARTS OF A GRAPH – including axes, titles, and units! Assume your reader knows nothing about your data and is gathering information about it from your graph.

Below is a histogram of the number of "Broken Eggs per Carton." The bins contain their left endpoints. That is, if x is a data value, then the first bin from 0 to 1 contains x's such that $0 \le x < 1$. In this histogram, all of the 0's are in the first bin, all of the 1's are in the second bin, etc.



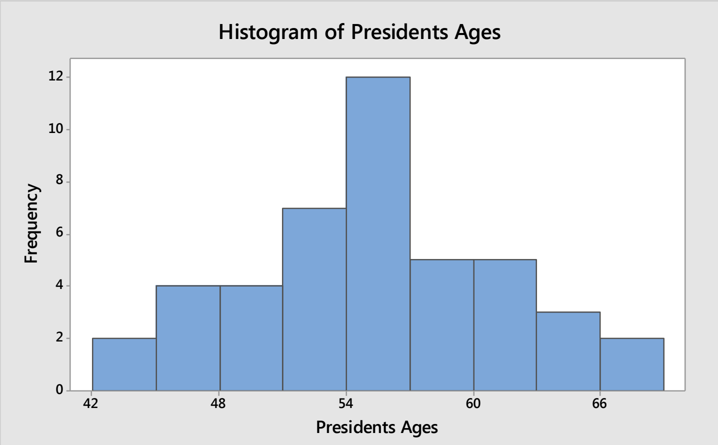
Minitab instructions for creating histograms will be discussed later in this lesson.

Example 2

Below are the ages at which the U.S. presidents began their first (non-consecutive) terms, increasing in order from George Washington to Barack Obama.

57	61	57	57	58	57	61	54	68	51	49	64	50
48	65	52	56	46	54	49	51	47	55	55	54	42
51	56	55	51	54	51	60	62	43	55	56	61	52
69	64	46	54	47								

Note: Clearly define your bin "boundaries" so your reader knows which bins contain the borderline values. Label your bins clearly when there is a question involved; don't leave your reader to decipher your binning scheme! For example, the first bin contains ages *x* such that 42 ≤ *x* < 45.



Bins contain their left endpoints.

As seen in the histogram, the data is **symmetric** about its center. Since the data is symmetric, the mean and median ages will be close to the same value.

Stem-and-Leaf Plots

The last graph in this lesson is the **stem-and-leaf plot**. This graph is often new to students taking statistics for the first time.

Definition: A **stem-and-leaf plot** is similar to a histogram, but is turned on its side. Instead of displaying bins, a stem-and-leaf plot displays digits from the actual data values to denote the frequency of each value.

The stem-and-leaf plot:

- Is used for medium-sized data sets, such as n = 50.
- Displays the basic shape, center, and spread of data.
- Shows the actual data values in the graph.
- Is more complicated to construct and explain to first time users of the graph, since the “stem” and “leaf” splits may be difficult to understand initially. Also, too few or too many stems may render the plot non-informative.

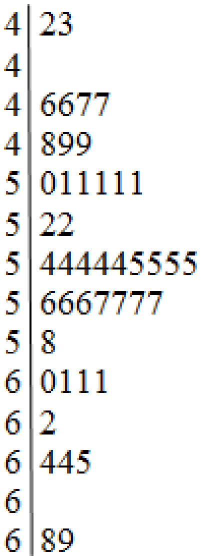
How to construct a stem-and-leaf plot by hand:

1. Select one or more digits for the stem values. The trailing digits become the leaves.
    - The “leaf” is usually the last digit of the number, and the other digits to the left of the “leaf” form the “stem.”
    - The number 125 could be split as: stem 12, leaf 5. The number 8124 could be split as: stem 812, leaf 4.
2. List possible stem values in a vertical column.
3. Record the leaf for every observation beside its corresponding stem value.
4. Indicate the units for stems and leaves in the display!

Example 3

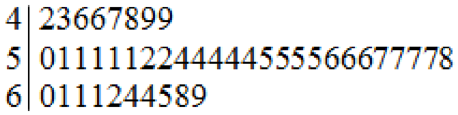
Using the “Presidents Ages” data set from **Example 2**, construct a stem-and-leaf plot.

(a) Choose stem units as tens and the leaf units as ones. Increment the stem rows by twos.



4	23
4	
4	6677
4	899
5	011111
5	22
5	444445555
5	6667777
5	8
6	0111
6	2
6	445
6	
6	89

(b) Choose stem units as tens and the leaf units as ones. Increment the stem rows by tens.



```
4 | 23667899
5 | 0111112244444555566677778
6 | 0111244589
```

Both stem-and-leaf plots show that the data is symmetric about its center.

However, the stem-and-leaf plot in part **(a)** displays the shape and spread of the data better than the plot in part **(b)**.

The advantage of the stem-and-leaf plot over the histogram is that we can see each data value that is represented within a given row (e.g. bin). Also, computing the median, mean, and mode are possible—in fact, fairly easy—given how the data are displayed in the stem-and-leaf plot.

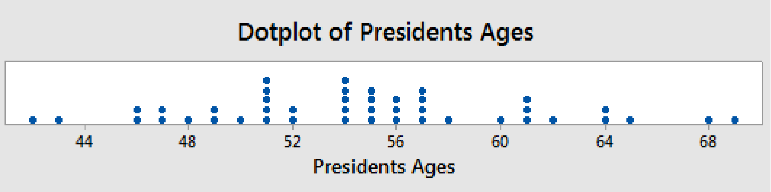
Minitab Calculations

The graphs described in the first three examples can be constructed in Minitab. We'll use the "Presidents Ages" data in the Minitab worksheet ***DescribingDataGraphically_Lesson.mtw*** to make a dotplot, histogram, and stem-and-leaf plot. The data are in column C2.

How to create a dotplot in Minitab:

Minitab 17

1 Choose **Graph > Dotplot**.
2 Choose **One Y - Simple**, then click **OK**.
3 In **Graph variables**, enter *'Presidents Ages.'*
4 Click **OK**.



Minitab Express

1 Open the dotplot dialog box.
- Mac: **Graphs > Dotplot > Simple**
- PC: **GRAPHS > Dotplot > Simple**

2 In **Variables**, enter *'Presidents Ages.''*
3 Click **OK**.

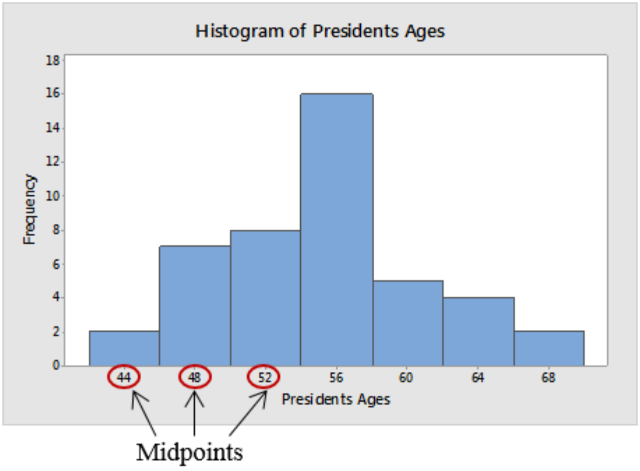
How to create a histogram in Minitab:

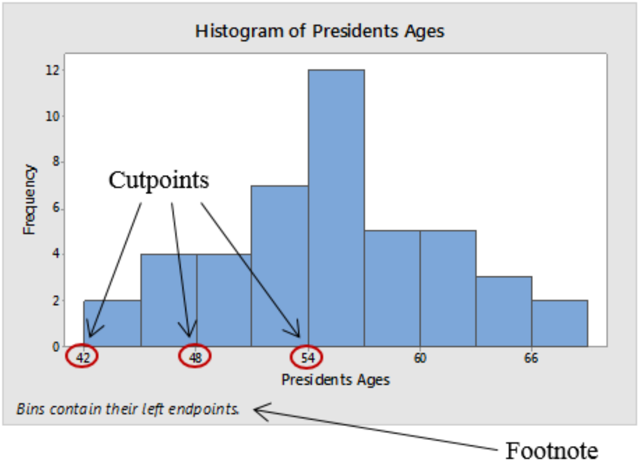
Minitab 17

1. Choose **Graph > Histogram**.
2. Choose **Simple**, then click **OK**.
3. Under **Graph variables**, enter *'Presidents Ages.'*
4. Click **OK**.

Minitab Express

1. Open the histogram dialog box.
   - Mac: **Graphs > Histogram > Simple**
   - PC: **GRAPHS > Histogram > Simple**
2. In **Variables**, enter *'Presidents Ages.'*
3. Click **OK**.





Minitab produces the histogram shown above on the left. By default, Minitab's bins are defined by their center values, or **midpoints**. It's hard to read the histogram with midpoints because you can't easily tell where each bin starts and ends. The histogram shown above on the right defines bins by their boundary values, or **cutpoints**.

How to display cutpoints on a histogram in Minitab:

Minitab 17

1. Double-click the histogram bins.
2. Click the **Binning** tab.
3. Under Interval Type, choose **Cutpoint**.
4. Click **OK**.

Minitab Express

1. Click the graph to select it.
2. Click the plus sign to open the graph elements menu.
3. Click the arrow next to **Binning**.
4. Choose **Cutpoint**. Select 8 for **Number of Bins**.
5. Click **OK**.

When using cutpoints instead of midpoints, Minitab constructs histograms such that the bins include their left cutpoints. It's important to state this in the graph to alleviate confusion as to whether bins contain their right or left cutpoints. We can display this fact as a **Footnote** at the bottom of the histogram graph.

How to add a footnote to a graph in Minitab:

Minitab 17

1. Right-click anywhere inside the graph.
2. Choose **Add > Footnote**.
3. Enter footnote text, such as "Bins contain their left endpoints."
4. Click **OK**.

Minitab Express

1. Click the graph to select it.
2. Click the plus sign to open the graph elements menu.
3. Select **Footnote**.
4. Select the graph, then click the footnote to edit it.
5. Enter footnote text, such as "Bins contain their left endpoints."
6. Click **OK.**

How to create a stem-and-leaf plot in Minitab:

Minitab 17

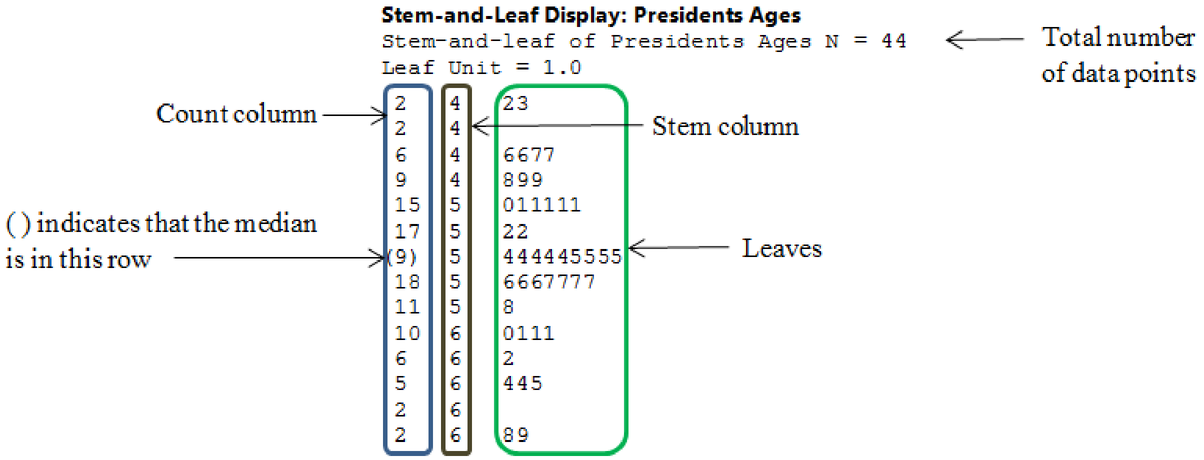
1. Choose **Graph > Stem-and-Leaf**.
2. In **Graph variables**, enter *'Presidents Ages.'*
3. Click **OK**.

Minitab Express

1. Open the stem-and-leaf-plot dialog box.
    - Mac: **Graphs** > **Stem-and-Leaf Plot**
    - PC: **GRAPHS** > **Stem-and-Leaf Plot**

2 In **Variable**, enter *'Presidents Ages.'*

3 Click **OK**.



```
Stem-and-Leaf Display: Presidents Ages
Stem-and-leaf of Presidents Ages  N = 44
Leaf Unit = 1.0
 2   4  23
 2   4
 6   4  6677
 9   4  899
15   5  011111
17   5  22
(9)  5  444445555
18   5  6667777
11   5  8
10   6  0111
 6   6  2
 5   6  445
 2   6
 2   6  89
```

Total number of data points
Count column
Stem column
( ) indicates that the median is in this row
Leaves

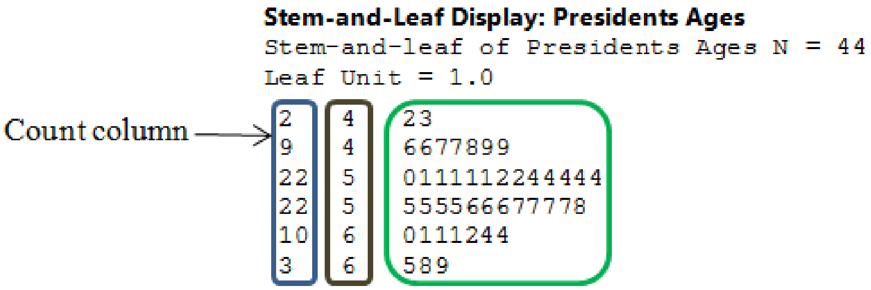
Note that the "count column" appears as the first column in the plot. Here is what the count column is doing:

- Each row of the stem-and-leaf plot displays the count, stem, and leaf.
  - The 1st line of the plot above has the count as 2, the stem as 4, and the leaves as 2 and 3.
  - The 2nd line has the count as 2, the stem as 4, and no leaves.
  - The 3rd line has the count as 6, the stem as 4, and leaves 6, 6, 7, 7.
- The row that contains the median has parentheses around the count. The count for this row represents the number of data values (or leaves) in the row.
  - The 7th row contains the median, and the count is denoted by (9). This means there are 9 data values in this row.
- The count for a row before the median represents the total count for that row and the rows before it.
  - The 1st line's count is 2, meaning there are two data values less than or equal to 43.
  - The 2nd line's count is 2, meaning there are two data values less than or equal to 45.
  - The 3rd line's count is 6, meaning there are six data values less than or equal to 47.
- The count for a row after the median represents the total count for that row and the rows after it.
  - The 8th line's count is 18, meaning there are eighteen data values greater than or equal to 56.
  - The 9th line's count is 11, meaning there are eleven data values greater than or equal to 58.
  - The 10th line's count is 10, meaning there are ten data values greater than or equal to 60.

Here are some notes for interpreting the stem-and-leaf plot for “Presidents Ages”:

- The first stem is 4 with two leaves: 2 and 3. This means that one president was age 42 and one president was age 43 at the beginning of their first terms.
- Eleven presidents began their first terms at age 58 or older.
- The median for the presidents’ ages is 54.5, the average of the 22nd and 23rd data points.
- Minitab chose the increment value of 2. This indicates the difference in value between stems is 2 years. In other words, the stem-and-leaf bins have a width of 2 years. The first line (bin) represents presidents whose ages are 42 and 43; the second line represents presidents whose ages are 44 and 45, etc.
- Using a different increment value, such as 5 or 10, will provide a different view of the stem-and-leaf plot. Too few or too many stems may result in the plot revealing less information.

Before ending this lesson, let’s look at another stem-and-leaf plot of “Presidents Ages” using an increment of 5 years, instead of 2 years as shown above:



Stem-and-Leaf Display: Presidents Ages

```
Stem-and-leaf of Presidents Ages N = 44
Leaf Unit = 1.0
2    4  23
9    4  6677899
22   5  0111112244444
22   5  555566677778
10   6  0111244
3    6  589
```

Here are some notes for interpreting this stem-and-leaf plot for “Presidents Ages”:

- The 1st stem is 4 with two leaves: 2 and 3. This means that one president was age 42 and one president was age 43, at the beginning of their first terms.
- The 2nd stem is 4 with seven leaves: 6, 6, 7, 7, 8, 9, 9. This means that the ages of seven presidents at the beginning of their first terms were ages 46, 46, 47, 47, 48, 49, and 49.
- Twenty-two presidents began their first terms at age 54 or younger.
- The median for the presidents’ ages is 54.5. It is the average value between the age 54 in row 3 and 55 in row 4. Since the median is not contained in a row, there is no indication of the median using parentheses in the plot.
- Since the increment value is 5, the difference in value between stems is 5 years. In other words, the stem-and-leaf bins have a width of 5 years. The first line (bin) represents presidents whose ages were between and including 40 to 44; the second line represents presidents whose ages were between and including 45 and 49, etc.
- You can choose to increment by another value, such as 10, if you want a different view of the stem-and-leaf plot. For an increment of 10, there are too few stems for the plot to be as informative as it currently is with increments of 2 or 5.