

## Section 2.1 - Introduction

- Graphs are commonly used to organize, summarize, and analyze collections of data. Using a graph to visually present a data set makes it easy to comprehend and to describe the data.

**Raw Data** is data before it has been arranged in a useful manner or analyzed using statistical techniques.

## Section 2.2 - Classifications of Data

A **variable** is a type of information, usually a property or characteristic of a person or thing that is measured or observed.

(a type of information or property about the thing being observed)

- Weight
- Height
- Color
- Sex
- GPA
- Temperature

**Value** of the variable is a specific measurement or observation for a variable.

- 163 pounds
- 5'11"
- Blue
- Male
- 3.5
- 32°

A **data set** is a collection of several data pertaining to one or more variables.

There are two types of variables: **Categorical** and **Numerical**:

A **categorical variable** represents categories, and is non-numerical in nature. These values are known as categorical data.

(Data which falls into categories)

Categorical:

- Non-numerical
- Color, nationality, vehicle type, met fan...

A **numerical variable** is a variable where the value is a number that results from a measurement process. The specific values of numerical variables are called numerical data.

(Data values which are the result of measurements)

Numerical data can be further classified as either continuous or discrete depending on the numerical values it can assume:

**Continuous data** are numerical measurements that can assume any value between two numbers.  
(Data which can take on any value between two numbers)

Continuous:

- Measurements that can take any value between any two numbers
- Usually rounded numbers
- Height, weight, distance, temperature, annual salary, blood pressure...

**Discrete data** are numerical values that can assume only a limited number of values.  
(Data which can take on only certain values)

Discrete:

- Represent precise counts at a point in time
- Full-time students at NCC
- The number of unemployed in a city
- Accidents in a state
- shoe size
- number of children in a family

*See page 35 for more examples of continuous & discrete*

### **Section 2.3 - Exploring Data Using the Stem-and-Leaf Display**

A **distribution** of a numerical variable represents the data values of the variable from the lowest to the highest value along with the number of times each data value occurs. The number of times each data value occurs is called its *frequency*.

(A presentation of data along with the number of times each data value occurs)

*See page 36 – Identifying Characteristics of a Distribution*

An **outlier** is an individual data value which lies far (above or below) from most or all of the other data values within a distribution.

A **stem-and-leaf display** is a visual exploratory data analysis technique that shows the shape of a distribution. The display uses the actual data values of the variable to present the shape of the distribution of data values.

(A form of data presentation which uses the data to visually display how the data are distributed)

**Procedure to Construct a Stem-and-Leaf Display:**

1. Identify the stem and leaf portion of your data values. Generally, the stem can have as many digits as needed to represent the beginning digit(s) of each data value. The leaf should only contain the last or terminating digit of the data values. (***SORT the data in your calculator***)
2. List each possible stem once, in a vertical column starting with the smallest stem on top and ending with the largest stem at the bottom. List even the stems with no corresponding leaves. Draw a vertical line to the right of the column of stems.
3. For each data value, record the leaf within the corresponding stem row and to the right of the vertical line.
4. Arrange the leaves within each stem row in increasing order from left to right. This provides a more informative stem-and-leaf display. (***NOT necessary if you SORT the data in your calculator first in step 1***)

At times, it may be of interest to compare two distributions using a stem-and-leaf display with a common stem. This type of display is called a **back-to-back stem-and-leaf display**.

**Example:** Cholesterol levels of 50 middle-aged men on a regular diet. (mg of Cholest/100ml of blood)

263 258 240 233 225 222 199 282 239 236 232 283 200  
 212 225 235 240 258 263 274 250 259 241 237 226 213  
 269 199 253 201 265 226 238 242 259 233 238 229 215  
 202 319 277 229 239 243 248 245 219 276 246

**Example 2.3:** Cholesterol levels of 50 middle-aged men on a low-fat diet. (mg of Cholest/100ml of blood)

161 172 193 190 205 214 217 205 168 176 188 195 209  
 219 189 209 220 211 200 195 184 175 198 209 211 228  
 199 188 172 165 177 197 202 211 198 184 171 200 218  
 220 210 189 231 176 180 193 230 200 181 190

- a) Construct a stem-and-leaf display.
- b) Describe the characteristics of the distribution.
- c) Determine the center of the distribution and then decide between which two values on a stem are the greatest concentration of data values located?

**Back-to-Back Stem-and-Leaf Display**

**Example 2.4:**

- a) Construct a back-to-back stem-and-leaf display to compare the distribution of cholesterol levels for middle-aged men who are on a regular diet to those who are on a low-fat diet.
- b) Use the back-to-back stem-and-leaf display to compare the two distributions.

## Section 2.4 - Frequency Distribution Tables

A **frequency distribution table** is a table in which a data set has been divided into distinct groups, called classes, along with the number of data values that fall into each class, called the frequency. Frequency distribution tables are useful in summarizing large data sets.

### Procedure for Constructing a Frequency Distribution Table for Numerical Data:

1. **Number of Classes** (between 5 and 15):

- Given

2. **Class Width:**

- Find using  $\frac{\text{largest data value} - \text{smallest data value}}{\text{number of classes}}$  then go up to the **next whole number**.

3. **Class Limits (or just Classes):**

- First class limit starts at the smallest data value
- First class ends at the smallest data value *plus* one less than the class width.
- Class limits for each class after, add the class width to the lower and upper limit of the previous class.

4. **Class Frequency:**

- Find by counting the number of data values that fall within each class.
- *Class frequency - vertical axis on the Histogram*

5. **Class Mark:**

- Find using  $\frac{\text{Upper Class Limit} + \text{Lower Class Limit}}{2}$

6. **Class Boundary (only if continuous data):**

- Find by subtracting 0.5 to each lower class limit and adding 0.5 to each upper class limit.
- *Class Boundary - the horizontal axis on the Histogram*

7. **Relative Frequency** is the proportion (in decimal form) of data values within each class:

- First add up the *Class Frequency* column for the total number of data values.
- Next for each class divide the *Class Frequency* by the total number of data values or use

$$\text{Relative Frequency of a Class} = \frac{\text{class frequency}}{\text{total number of data values}}$$

8. **Relative Percentage** is the percentage of data values within each class:

- Find by multiplying each *Relative Frequency* by 100 which changes it to a percentage or use

$$\text{Relative Percentage Frequency} = (\text{relative frequency}) \cdot (100)$$

### Section 2.5 - Graphs

- A graph is a descriptive tool used to visualize the characteristics and the relationships of the data quickly and attractively. A well-constructed graph will reveal information that may not be apparent from a quick examination of a frequency distribution table.

A **histogram** is a vertical bar graph that represents a **continuous** variable. To depict the continuous nature of the data, the rectangles are connected to each other without any gaps or breaks between two adjacent rectangles. The width of each rectangle corresponds to the width of each class of the frequency distribution and the height of each rectangle corresponds to the frequency of the class. The vertical sides of each rectangle of a histogram correspond to the class boundaries of each class, and so, there are no breaks or gaps between the rectangles.

#### Procedure to Construct a Histogram (only for *Continuous* data)

1. Organize the data into a frequency distribution table.
2. Label the horizontal axis with the name of the variable. Scale the horizontal axis starting from the lowest class boundary to the highest class boundary, while marking off all the other class boundaries in increments equal to the class width.
3. Label the vertical axis as the frequency. Determine the largest frequency value and scale the vertical axis using an appropriate increment to represent the frequencies.
4. Construct a rectangle for each class where the class boundaries correspond to the vertical sides of the rectangle, and the height of the rectangle corresponds to the frequency. Each rectangle should have the same width, and this width is the class width.

#### Histogram with Graphing Calculator:

- Enter data into *List 1*
- If frequencies are given enter into *List 2*
- Set *Window* as follows:  
 $X_{\min}$  = the smallest data value  
 $X_{\max}$  = the largest data value + the class width.  
 $X_{\text{scl}}$  = the class width  
 $Y_{\min}$  = 0  
 $Y_{\max}$  = your estimate of the frequency of the largest class  
 Keep both  $Y_{\text{scl}}$  and  $X_{\text{res}}$  = 1
- 2<sup>nd</sup> STAT PLOT choose 1
- $x$  List: *List 1*
- Frequency: *List 2* **or** the number "1" (if frequencies were not given)
- Use the TRACE to get class limits and frequencies
- Change only  $X_{\min}$  to smallest data value minus 0.5 to get class boundaries
- Use the TRACE to get class boundaries and frequencies

### Section 2.7 - Identifying Shapes and Interpreting Graphs

- Symmetric bell-shaped distribution
- Skewed to the right distribution
- Skewed to the left distribution
- Uniform distribution
- U-shaped distribution
- Reverse J-shaped distribution
- Bimodal distribution

Class Worksheet - Chapter 2

**Frequency Distribution Table Example** (NOT in textbook)

A survey of 110 people asked the number of cups of coffee they drink in a week. The table below represents the data collected.

# of Cups	Frequency
0	2
1	4
2	10
3	6
4	3
5	7
6	4
7	5
8	4
9	2
10	3
11	6
12	0
13	1
14	3
15	7
16	12
17	9
18	7
19	8
20	7

a. Construct a Frequency Distribution Table with 7 classes. Include within the table: Classes, Class Frequency, Class Boundaries, Class Marks, Relative Frequency, and Relative Percentages.

Classes or Class Limits	Class Frequency	Class Boundaries	Class Mark	Relative Frequency	Relative Percentage

b. Construct a Histogram.