Ch1. Describing Data with Graphs

Many sets of measurements are samples selected from larger populations. Other sets constitute the entire population, as in a national census. In this chapter, you will learn what a **variable** is, how to classify variables into several types, and how measurements or data are generated. You will then learn how to use **graphs to describe data sets**.

윤 연 옥

1.1 Variables and Data

- A **variable** is a characteristic that changes or varies over time and/or for different individuals or objects under consideration.

Examples: Hair color, white blood cell count, time to failure of a computer component.

- An **experimental unit** is the individual or object on which a variable is measured.
- A **measurement** results when a variable is actually measured on an experimental unit.
- A set of measurements, called **data**, can be either a **sample** or a **population.**

✓ *A population is the set of all measurements of interest to the investigator.*

✓ *A sample is a subset of measurements selected from the population of interest.*

1.1 Variables and Data

◆ Example1

Variable :

Hair color

Experimental unit

Person

Typical Measurements

Brown, black, blonde, etc.

◆ Example2

Variable

Time until a light bulb burns out

Experimental unit

Light bulb

Typical Measurements

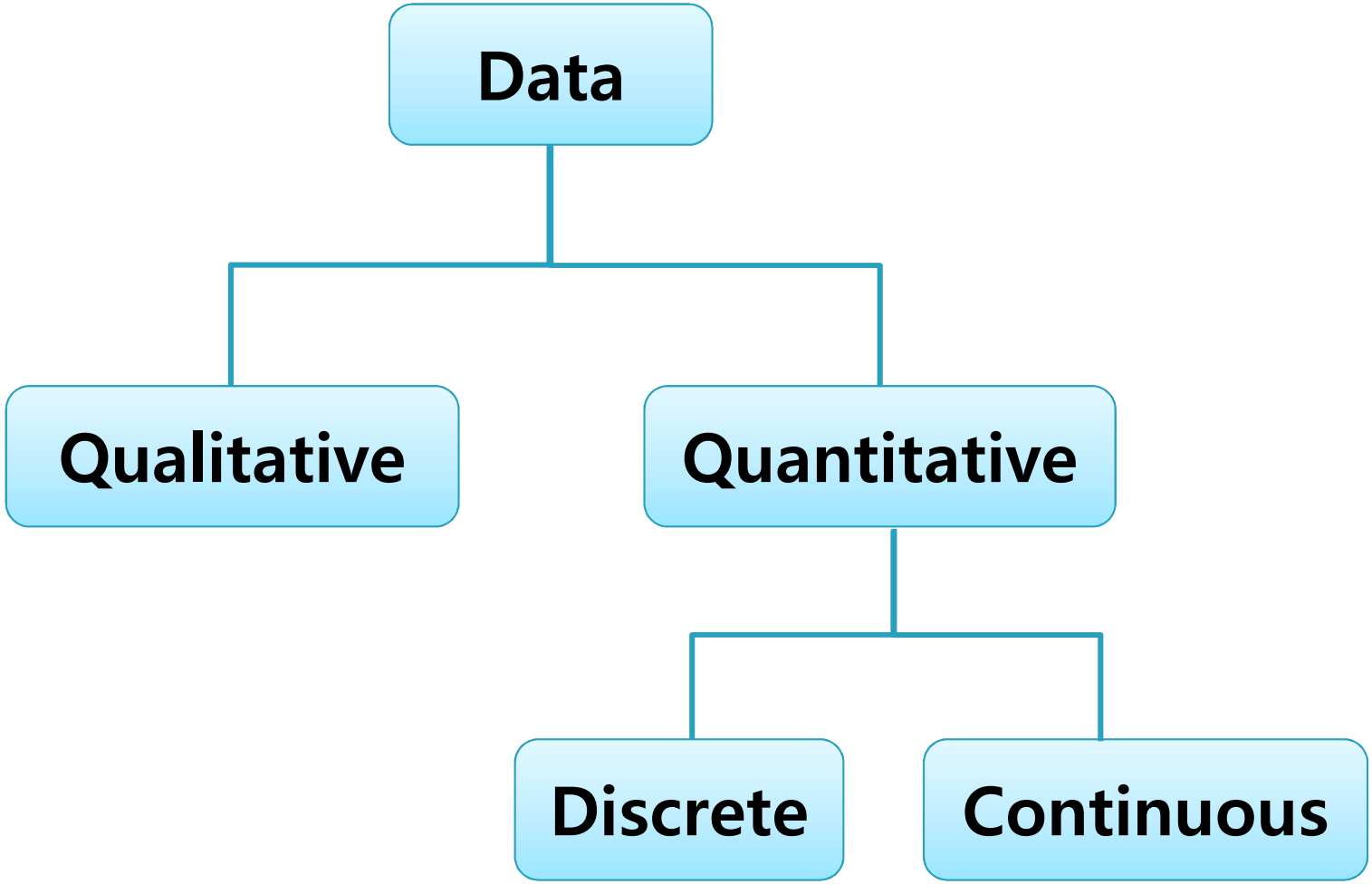
1500 hours, 1535.5 hours, etc.

1.1 Variables and Data

◆ How many variables have you measured?

- **Univariate data**: One variable is measured on a single experimental unit.
- **Bivariate data**: Two variables are measured on a single experimental unit.
- **Multivariate data**: More than two variables are measured on a single experimental unit.

1.2 Types of Variables



1.2 Types of Variables

◆ **Qualitative variables** measure a quality or characteristic on each experimental unit.

✓ *Qualitative variables produce data that can be separated into categories. Hence they are called* ***categorical variables***

Examples:

- Hair color (black, brown, blonde…)
- Make of car (Dodge, Honda, Ford…)
- Gender (male, female)
- Political affiliation(Republican, Democrat, Independent)

1.2 Types of Variables

◆ **Quantitative variables** measure a numerical quantity or amount on each experimental unit.

✓ **Discrete** if it can assume only a finite or countable number of values.

✓ **Continuous** if it can assume the infinitely many values corresponding to the points on a line interval.

Tip

- Discrete refers to the discrete gaps between the possible values

  ex) number of family members, number of new car sales, number of defective tires

- Continuous means a third value can always be found between two values.

  ex)height, weight, time, distance, etc

1.2 Types of Variables

Examples 1.2

Identify each of the following variables as qualitative or quantitative:

1. The most frequent use of your microwave oven (reheating, defrosting, warming, other) : *qualitative*
2. The number of consumers who refuse to answer a telephone survey : *quantitative discrete*
3. The door chosen by a mouse in a maze experiment (A, B, or C): *qualitative*
4. The winning time for a horse running in the Kentucky Derby : *quantitative continuous*
5. The number of children in a fifth-grade class who are reading at or above grade level : *quantitative discrete*

1.3 Graphs for Categorical Data(Qualitative Variables)

◆ Use a data distribution to describe:

✓ **What values** of the variable have been measured

✓ **How often** each value has occurred

◆ "How often" can be measured 3 ways:

- **Frequency** (*number of measurements in each category*)
- **Relative frequency = Frequency/n**
- **Percent = 100 x Relative frequency**

- Once the measurements have been categorized and summarized in a statistical table, you can use either a **pie chart** or a **bar chart** to display the distribution of the data.
- **Pareto chart** : a bar chart in which the bars are ordered from largest to smallest.

Examples 1.3

In a survey concerning public education, 400 school administrators were asked to rate the quality of education in the United States. Their responses are summarized in Table 1.1. Construct a pie chart and a bar chart for this set of data.

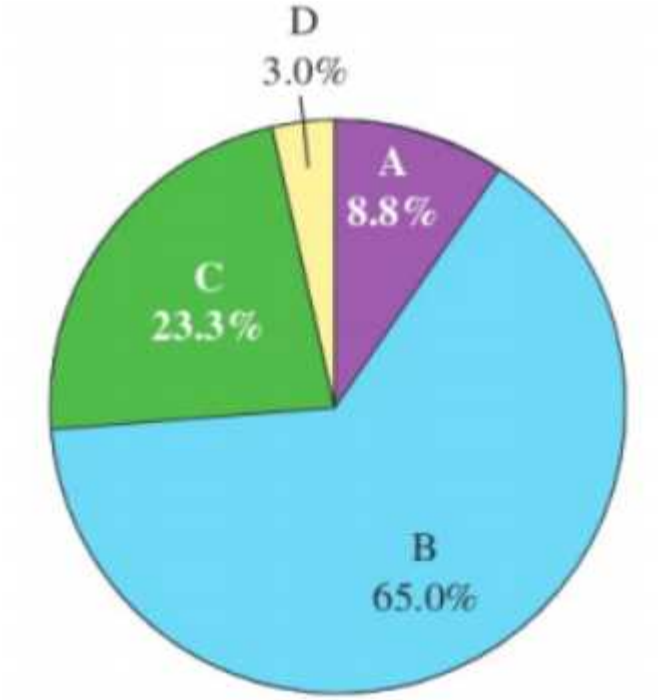
Table 1.1

Rating	Frequency
A	35
B	260
C	93
D	12
Total	400

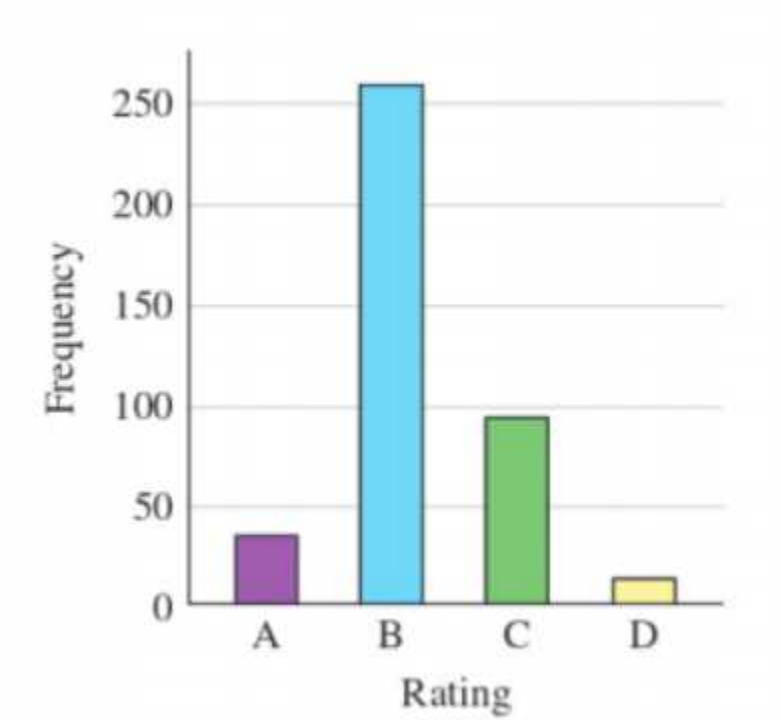
✓ Calculation for the Pie Chart

Rating	Frequency	Relative Frequency	Percent	Angle
A	35	35/400 = .09	9%	.09 × 360 = 32.4°
B	260	260/400 = .65	65%	234.0°
C	93	93/400 = .23	23%	82.8°
D	12	12/400 = .03	3%	10.8°
Total	400	1.00	100%	360°

Pie chart



Bar chart



Examples 1.4

A snack size bag of peanut M&M'S candies contains 21 candies with the six different colors.

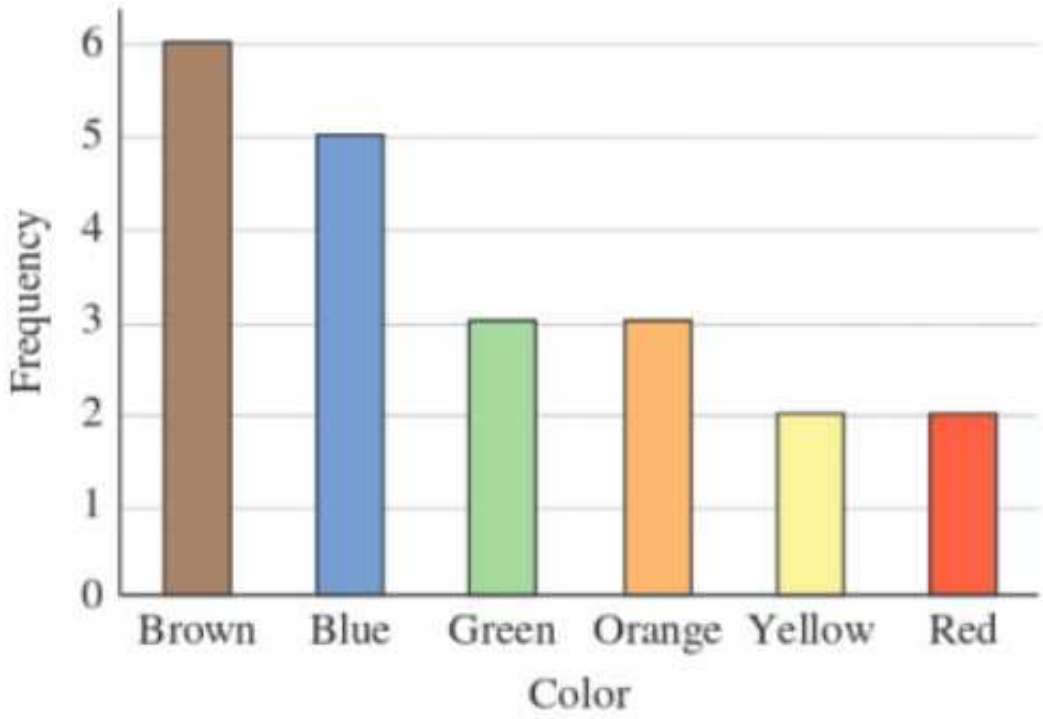
< Raw data: Color of 21 Candies>

Brown	Green	Brown	Blue
Red	Red	Green	Brown
Yellow	Orange	Green	Blue
Brown	Blue	Blue	Brown
Orange	Blue	Brown	Orange
Yellow			

< Statistical Table: M&M Data>

Category	Tally	Frequency	Relative Frequency	Percent
Brown	𝍸 I	6	6/21	28%
Green	III	3	3/21	14
Orange	III	3	3/21	14
Yellow	II	2	2/21	10
Red	II	2	2/21	10
Blue	𝍸	5	5/21	24
Total		21	1	100%

Pareto Chart



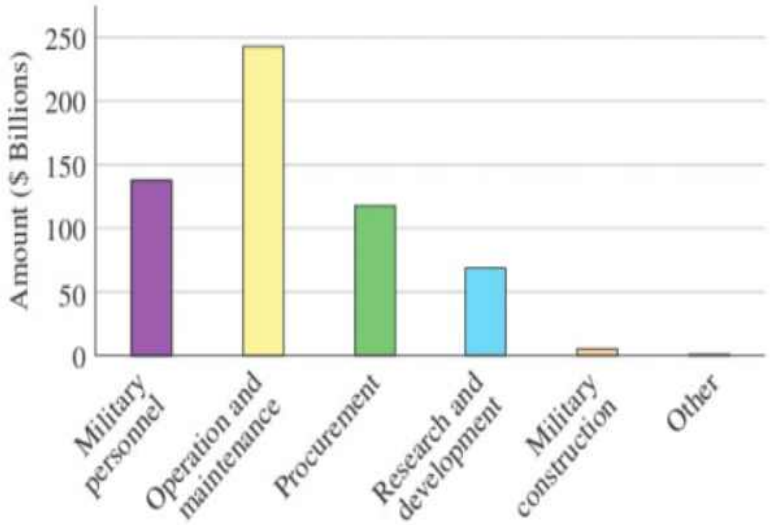
1.4 Graphs for Quantitative Data

- A single quantitative variable measured for different population segments or for different categories of classification can be graphed using a **pie** or **bar chart**.

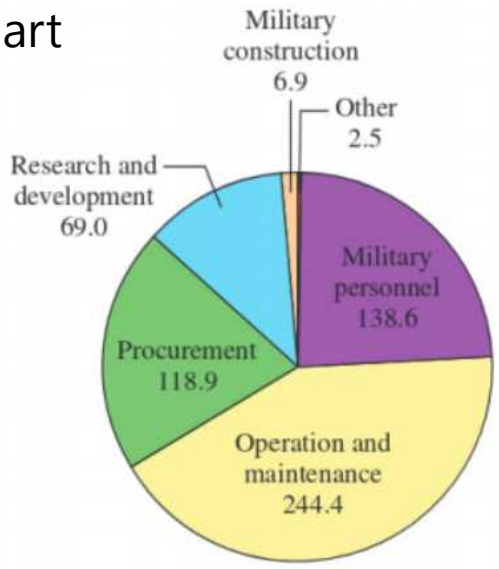
Example 1.5
Expenses by categories
U.S. Department of Defense

Category	Amount ($ billions)
Military personnel	138.6
Operation and maintenance	244.4
Procurement	118.9
Research and development	69.0
Military construction	6.9
Other	2.5
Total	**580.3**

Bar chart



Pie chart



- A single quantitative variable measured over time is called a **time series.** It can be graphed using a **line** or **bar chart.**

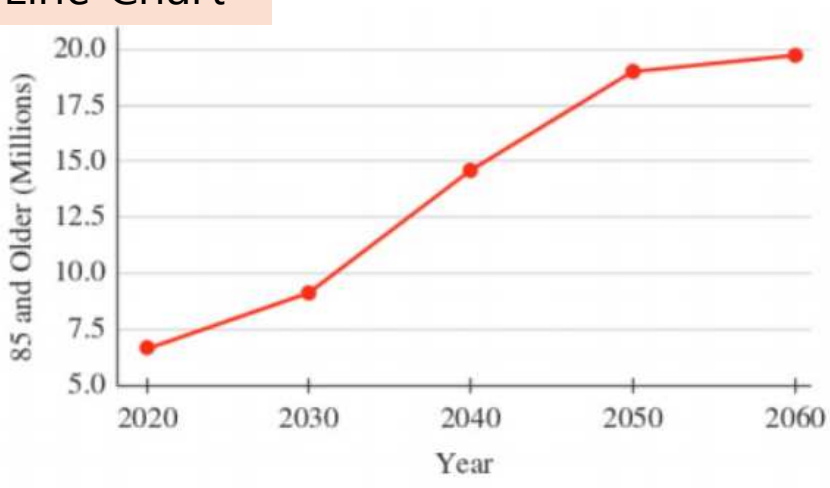
Example1.6

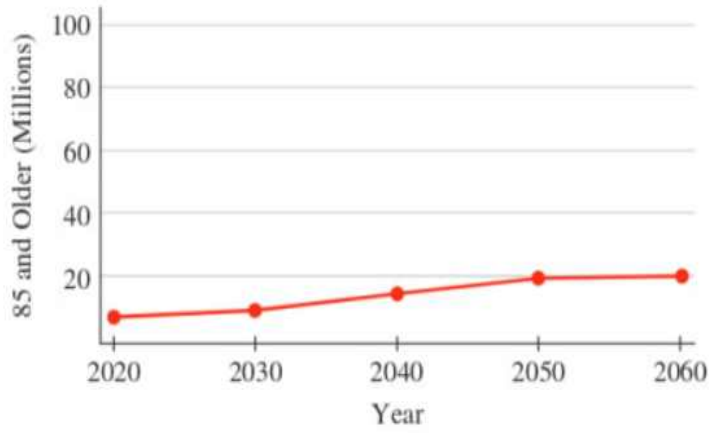
U.S. Population Growth Projections (*United States Bureau of the Census*)

Year	2020	2030	2040	2050	2060
85 and over (millions)	6.7	9.1	14.6	19.0	19.7

Source: The World Almanac and Book of Facts 2017, p. 618

Line Chart





◆ Dotplots

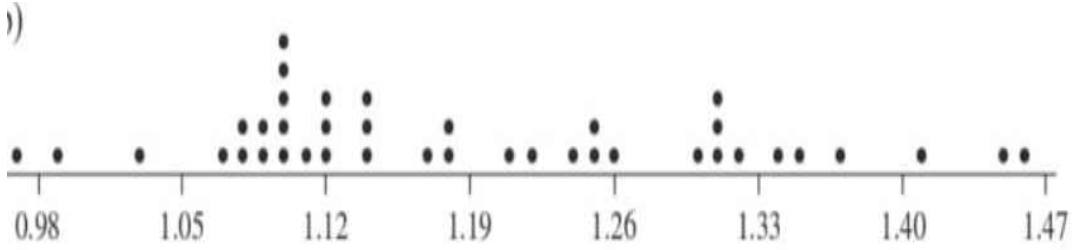
- The simplest graph for quantitative data
- Plots the measurements as **points** on a horizontal axis, stacking the points that duplicate existing points.

Example :

For Small Set 2, 6, 9, 3, 7, 6

Small Set

For Large Set



Large Set

Hard to interpret

◆ Stem and Leaf Plots

- A simple graph for quantitative data
- Uses the actual numerical values of each data point
- How to Construct a Stem and Leaf Plot

1. Divide each measurement into two parts: the **stem** and the **leaf**.
2. List the stems in a column, with a **vertical line** to their right.
3. For each measurement, record the leaf portion in the **same row** as its matching stem.
4. **Order** the leaves from lowest to highest in each stem.

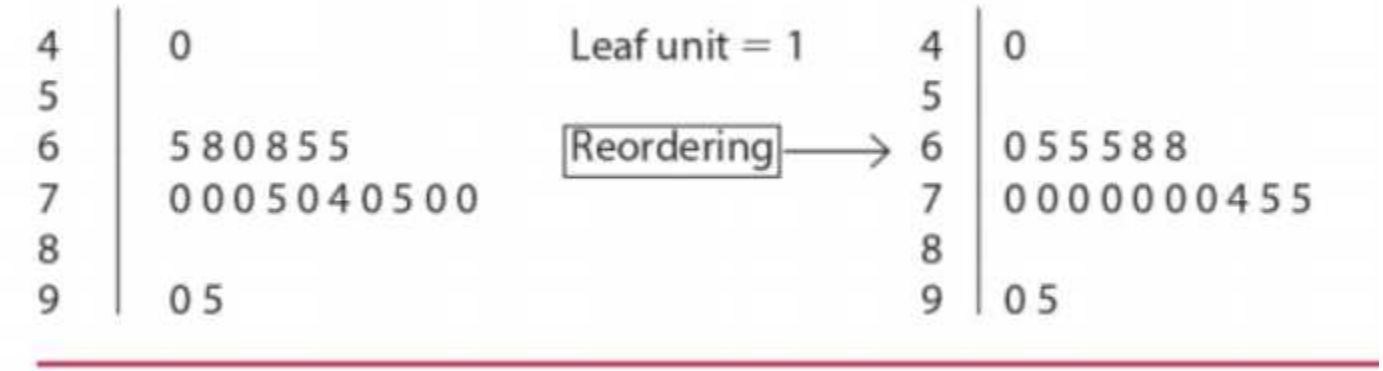
Example

The prices ($) of 19 brands of walking shoes:

Data

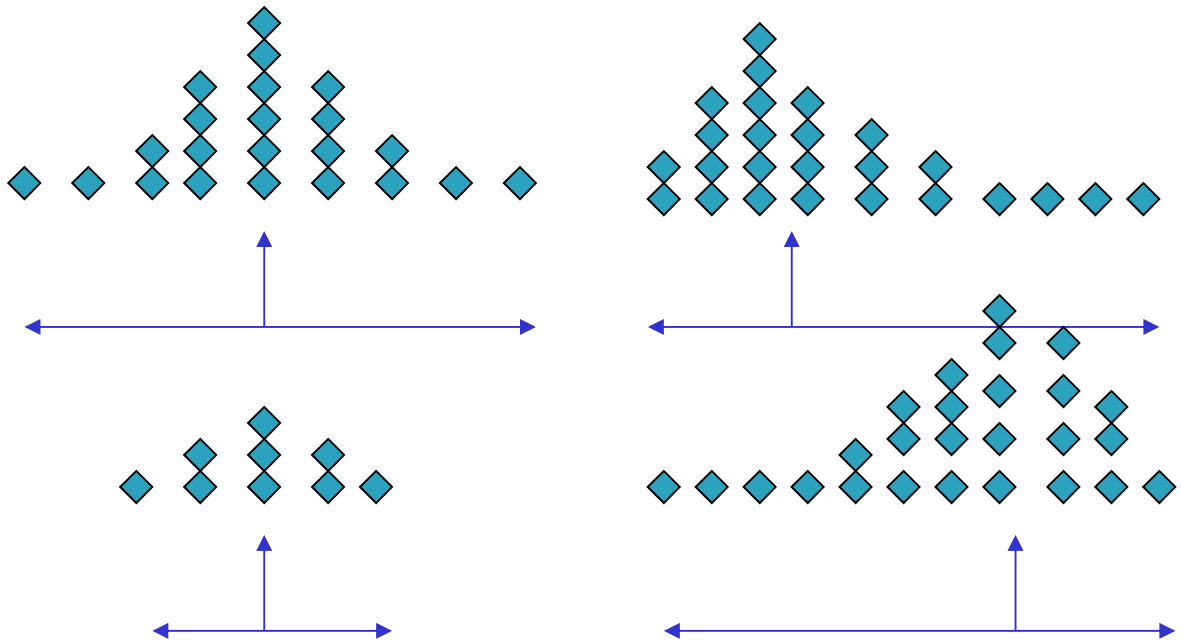
90	70	70	70	75	70
65	68	60	74	70	95
75	70	68	65	40	65
70					

Stem and Leaf



Too few stems and a large number of leaves within each stem
→ stretch the stems by dividing each one into several lines
ex) two lines: 0-4, 5-9

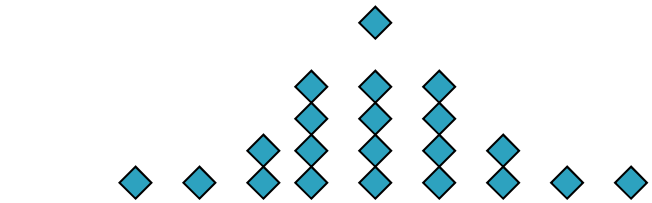
◆ Interpreting Graphs: Location and Spread



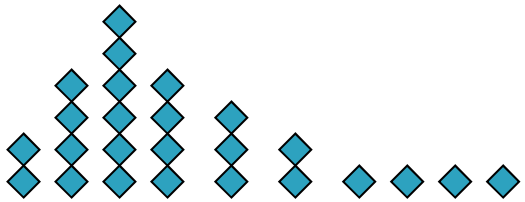
✓ Where is the data centered on the horizontal axis?

✓ And how does it spread out from the center?

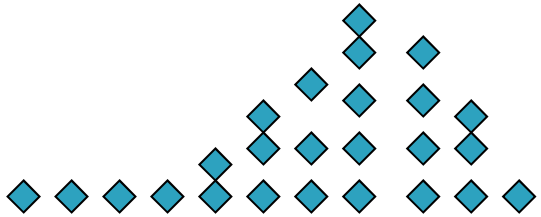
◆ Interpreting Graphs: Shapes



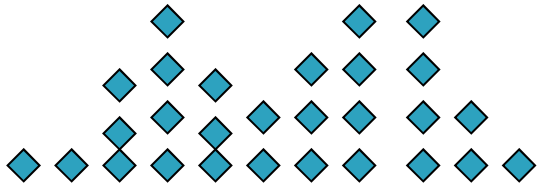
Mound shaped and symmetric (mirror images)



Skewed right: a few unusually large measurements

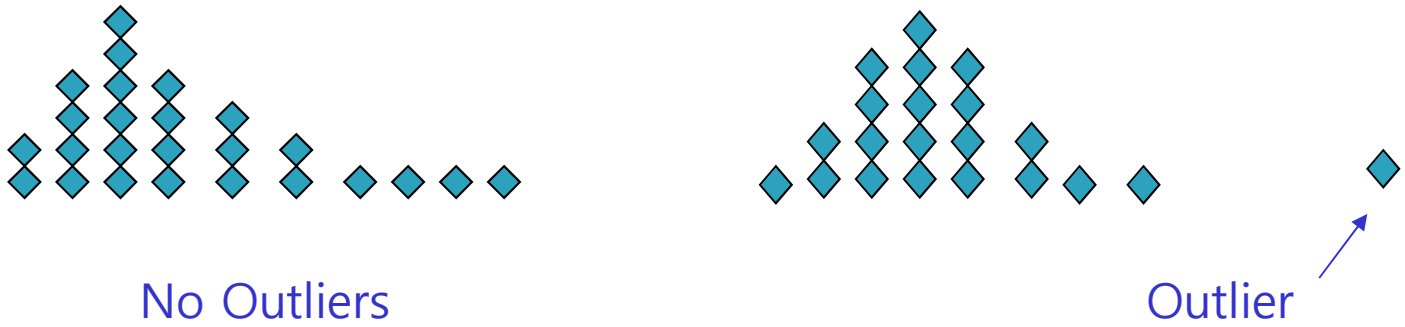


Skewed left: a few unusually small measurements



Bimodal: two local peaks

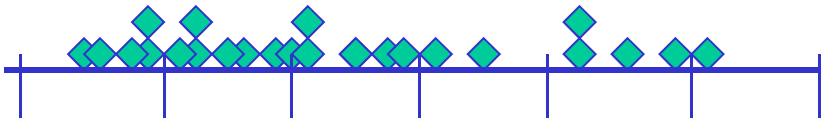
◆ Interpreting Graphs: Outliers



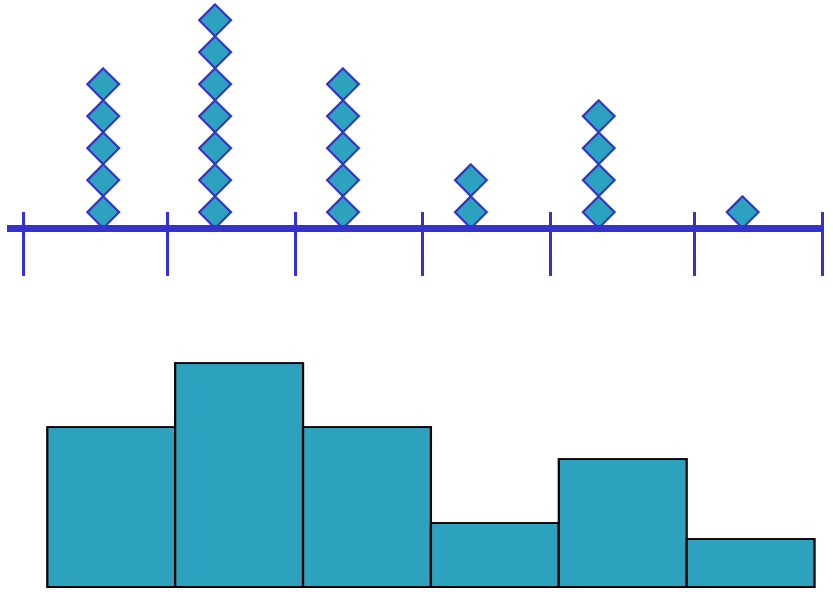
✓ Are there any strange or unusual measurements that stand out in the data set?

1.5 Relative Frequency Histogram

A **relative frequency histogram** for a quantitative data set is a bar graph in which the height of the bar shows "how often" (measured as a proportion or relative frequency) measurements fall in a particular class or subinterval.



Create intervals



Stack and draw bars

1.5 Relative Frequency Histogram

◆ How to Construct a Relative Frequency Histogram

1. Divide the range of the data into **5-12 subintervals** of equal length.
2. Calculate the **approximate width** of the subinterval as Range/number of subintervals.
3. Round the approximate width up to a convenient value.
4. Use method of **left inclusion** including the left endpoint, but not the right boundary point in the class.

   ✓*each measurement falls into one and only one class*

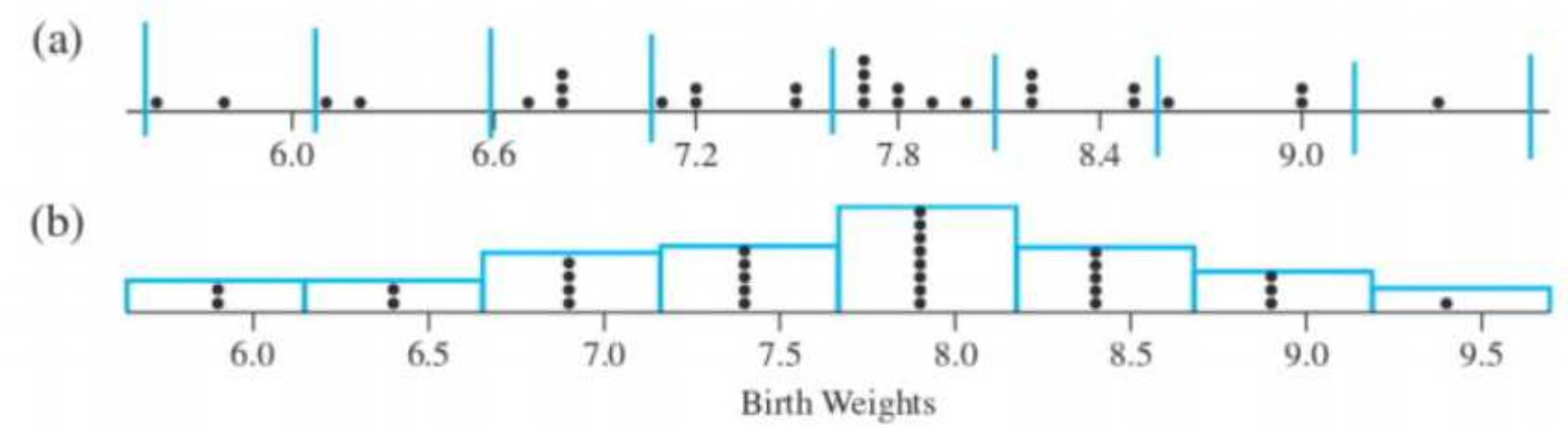
5. Create a **statistical table** including the subintervals, their frequencies and relative frequencies.
6. Draw the **relative frequency histogram** plotting the subintervals on the horizontal axis and the relative frequencies on the vertical axis.

Example

data: Birth Weights of 30 Full-Term Newborn Babies

7.2	7.8	6.8	6.2	8.2	8.0	8.2	**5.6**	8.6	7.1
8.2	7.7	7.5	7.2	7.7	5.8	6.8	6.8	8.5	7.5
6.1	7.9	**9.4**	9.0	7.8	8.5	9.0	7.7	6.7	7.7

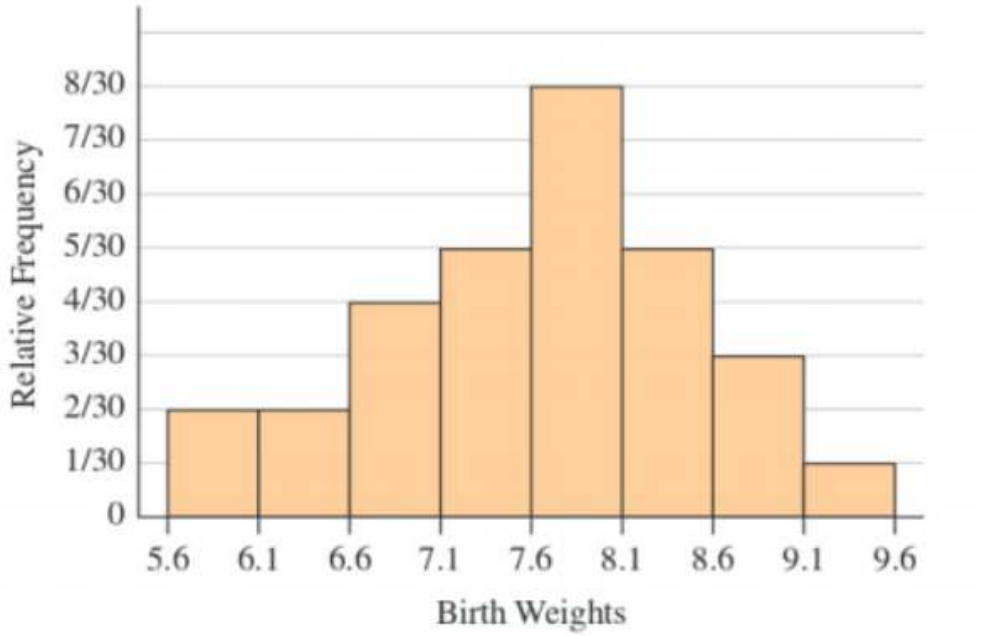
1. We choose 8 intervals
2. Minimum class width = (9.4−5.6)/8 = 3.8/8 = 0.475
3. Convenient class width = 0.5
4. Use 8 classes of length 0.5, starting at 5.6.



Relative Frequency Table

Class	Class Boundaries	Tally	Class Frequency	Class Relative Frequency
1	5.6 to <6.1	II	2	2/30
2	6.1 to <6.6	II	2	2/30
3	6.6 to <7.1	IIII	4	4/30
4	7.1 to <7.6	𝍸	5	5/30
5	7.6 to <8.1	𝍸 III	8	8/30
6	8.1 to <8.6	𝍸	5	5/30
7	8.6 to <9.1	III	3	3/30
8	9.1 to <9.6	I	1	1/30

Relative Frequency Histogram



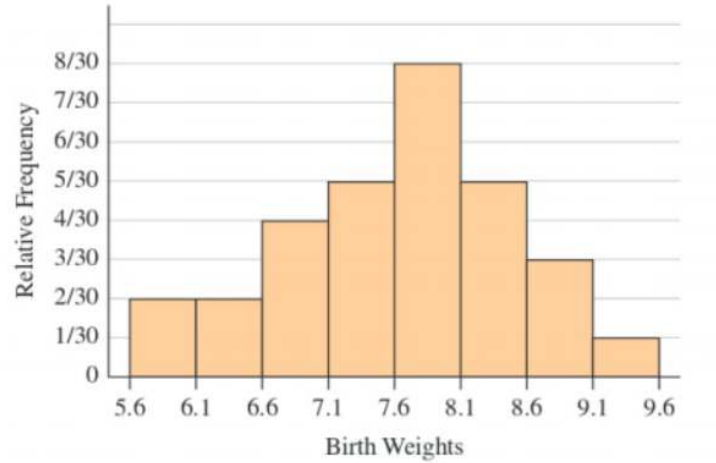
Describing the Distribution

Shape? Symmetric

Outliers? No

What proportion of the newborns have birth weights of 7.6 or higher?

(8+5+3+1)/30 =17/30=0.57



Key Concepts

I. How Data Are Generated

1. Experimental units, variables, measurements
2. Samples and populations
3. Univariate, bivariate, and multivariate data

II. Types of Variables

1. Qualitative or categorical
2. Quantitative
    a. Discrete
    b. Continuous

III. Graphs for Univariate Data Distributions

1. Qualitative or categorical data
    a. Pie charts,
    b. Bar charts
2. Quantitative data
    a. Pie and bar charts
    b. Line charts
    c. Dotplots
    d. Stem and leaf plots
    e. Relative frequency histograms
3. Describing data distributions
    a. Shapes—symmetric, skewed left, skewed right, unimodal, bimodal
    b. Proportion of measurements in certain intervals
    c. Outliers