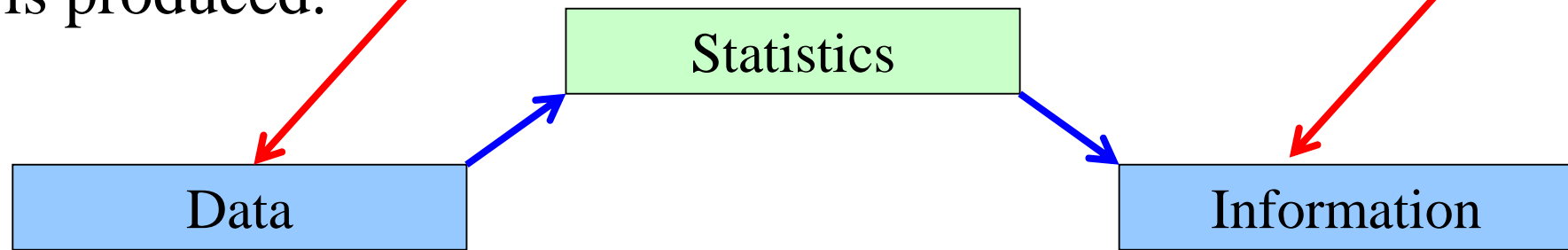


Chapter Two

Graphical and Tabular Descriptive Techniques

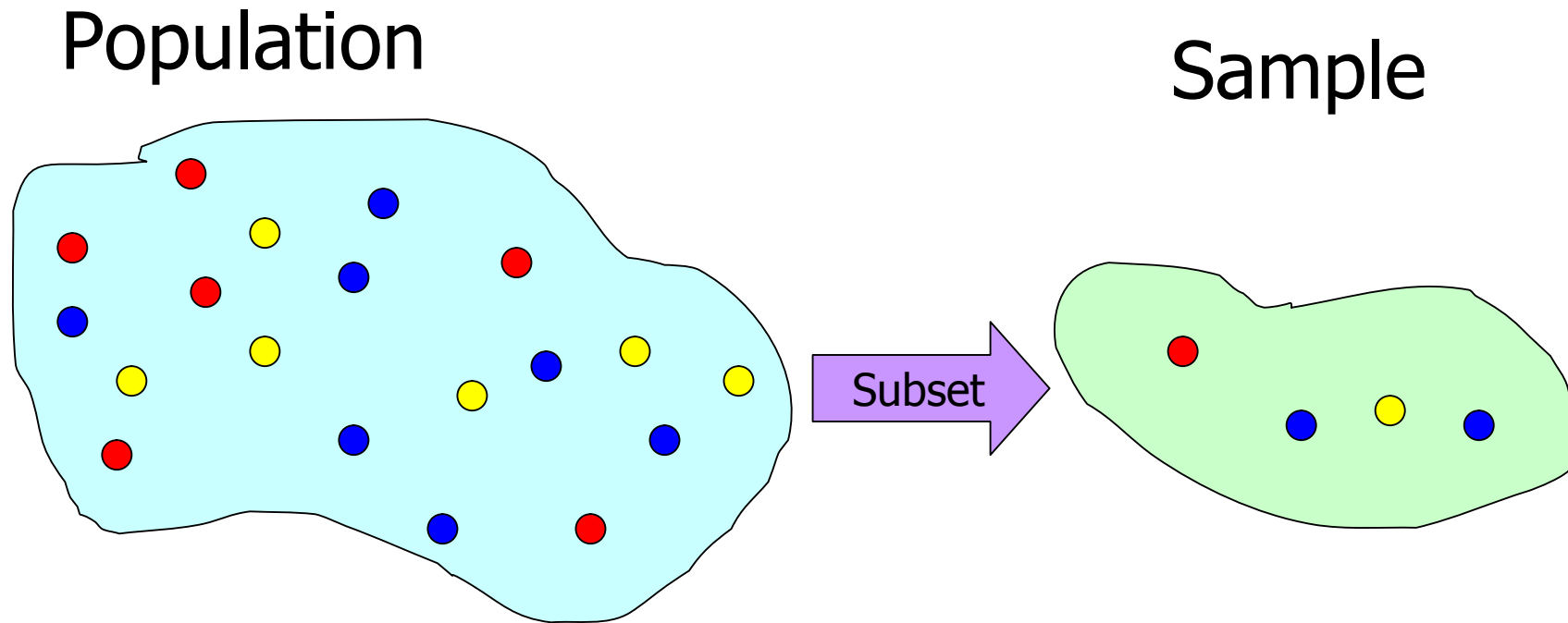
Introduction & Re-cap...

Descriptive statistics involves arranging, summarizing, and presenting a set of data in such a way that useful information is produced.



Its methods make use of graphical techniques and numerical descriptive measures (such as averages) to summarize and present the data.

Populations & Samples



The graphical & tabular methods presented here apply to both entire populations *and* samples drawn from populations.

Definitions...

A **variable** is some characteristic of a population or sample.

E.g. student grades.

Typically denoted with a capital letter: X, Y, Z...

The **values** of the variable are the range of possible values for a variable.

E.g. student marks (0..100)

Data are the *observed values* of a variable.

E.g. student marks: {67, 74, 71, 83, 93, 55, 48}

Types of Data & Information

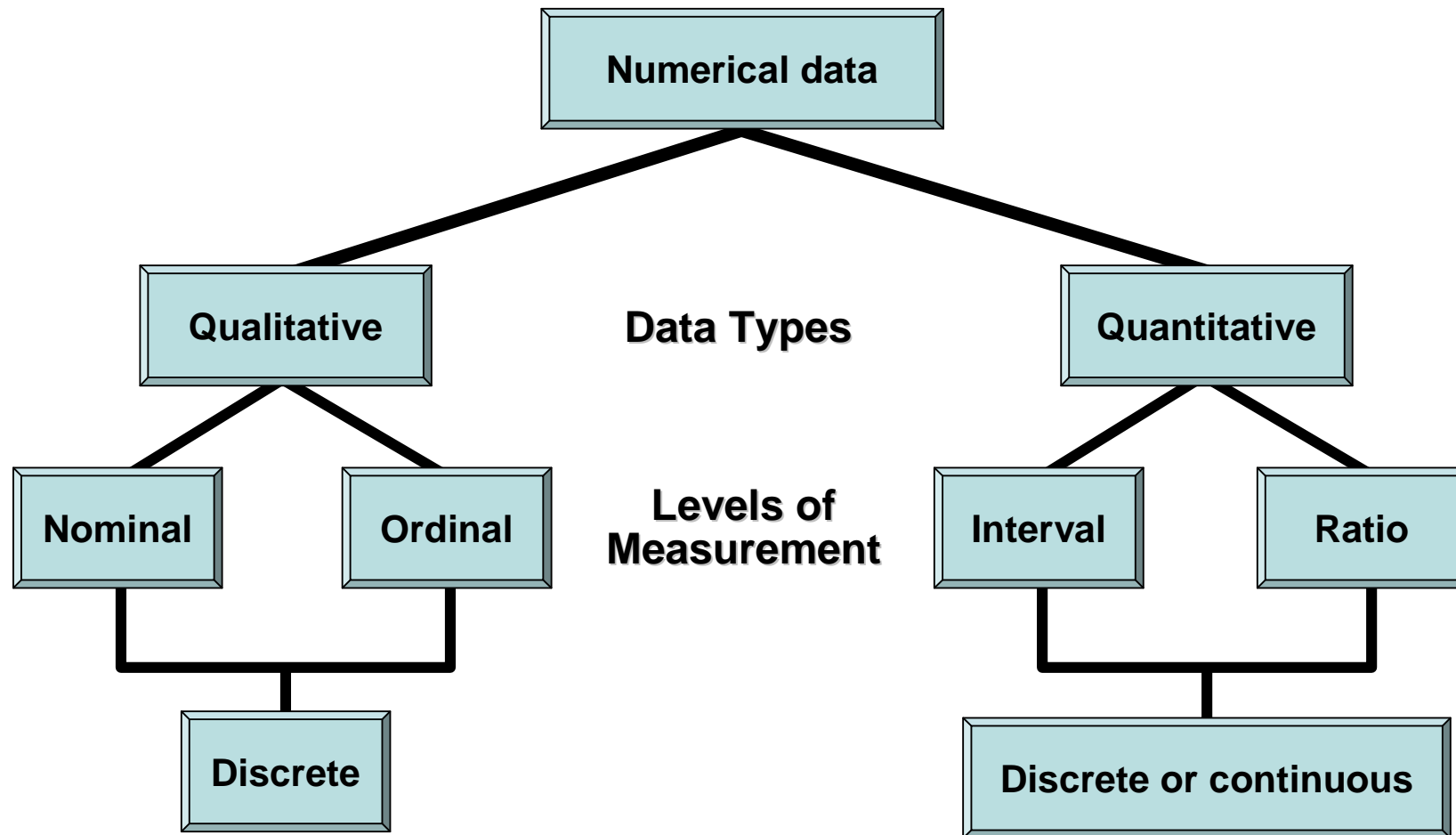
Data (at least for purposes of Statistics) fall into three main groups:

Interval Data

Nominal Data

Ordinal Data

Types of Data



Types of Data

EXAMPLES OF DISCRETE DATA

1. **Nominal:** Ownership status of resident dweller
(1 = own, 2 = rent)
2. **Ordinal:** Level of customer satisfaction
(1 = very dissatisfied, 2 = somewhat dissatisfied, 3 = somewhat satisfied, 4 = very satisfied)
3. **Interval:** Person's score on IQ test
4. **Ratio:** Number of defective lightbulbs in a carton

EXAMPLES OF CONTINUOUS DATA

1. **Interval:** Actual temperature, ° F
2. **Ratio:** Weight of packaged dog food

Interval Data...

Interval data

- Real numbers, i.e. heights, weights, prices, etc.
- Also referred to as **quantitative** or **numerical**.

Arithmetic operations can be performed on Interval Data, thus its meaningful to talk about $2 * \text{Height}$, or $\text{Price} + \$1$, and so on.

Nominal Data...

Nominal Data

- The values of **nominal** data are *categories*.

E.g. responses to questions about marital status, coded as:

Single = 1, Married = 2, Divorced = 3, Widowed = 4

These data are **categorical** in nature; arithmetic operations don't make any sense (e.g. does $\text{Widowed} \div 2 = \text{Married}?!)$

Nominal data are also called **qualitative** or **categorical**.

Ordinal Data...

- **Ordinal Data** appear to be categorical in nature, but their values have an *order*; a ranking to them:
 - E.g. College course rating system: poor = 1, fair = 2, good = 3, very good = 4, excellent = 5
- While its still not meaningful to do arithmetic on this data (e.g. does $2 * \text{fair} = \text{very good}?!)$, we can say things like:
 - **excellent** > **poor** or **fair** < **very good**
- That is, order is maintained **no** matter what numeric values are assigned to each category.

Notes

- It is not valid to compute and interpret differences between values of the ordinal variable.
- The critical difference between ordinal and interval data is that the intervals or differences between values of interval data are consistent and meaningful.
- Some of interval data contain the true zero, such as height and weight, but some of interval data don't contain the true zero, such as temperature and pH.

Calculations for Types of Data

As mentioned above,

- All calculations are permitted on **interval** data.
- Only calculations involving a ranking process are allowed for **ordinal** data.
- No calculations are allowed for **nominal** data, save counting the number of observations in each category.

This lends itself to the following “hierarchy of data”...

Hierarchy of Data...

Interval

Values are real numbers.

All calculations are valid.

Data may be treated as ordinal or nominal.

Ordinal

Values must represent the ranked order of the data.

Calculations based on an ordering process are valid.

Data may be treated as nominal but not as interval.

Nominal

Values are the arbitrary numbers that represent categories.

Only calculations based on the frequencies of occurrence are valid.

Data may not be treated as ordinal or interval.

2.2 Graphical & Tabular Techniques for Nominal Data

- The only allowable calculation on nominal data is to
 - count the frequency of each value of the variable.
- We can summarize the data in a table that presents the categories and their counts called a *frequency distribution*.
- A *relative frequency distribution* lists the categories and the proportion with which each occurs.

Example 2.1 Light Beer Preference Survey

- In 2006 total light beer sales in the United States was approximately 3 million gallons.
- With this large market breweries often need to know more about
 - who is buying their product.
- The marketing manager of a major brewery wanted to analyze the light beer sales among college and university students who do drink light beer.
- A random sample of 285 graduating students was asked to report which of the following is their favorite light beer.

Things to check on Example 2.1

- Objective of the study?
- Population?
- Parameters?
- Sample data?
- Statistics?
- Data type?

Example 2.1

1. Budweiser Light
2. Busch Light
3. Coors Light
4. Michelob Light
5. Miller Lite
6. Natural Light
7. Other brand

- Data: The responses were recorded using the codes.
- Construct a frequency and relative frequency distribution for these data.
- Graphically summarize the data by producing a bar chart and a pie chart.

Sample data for Example 2.1

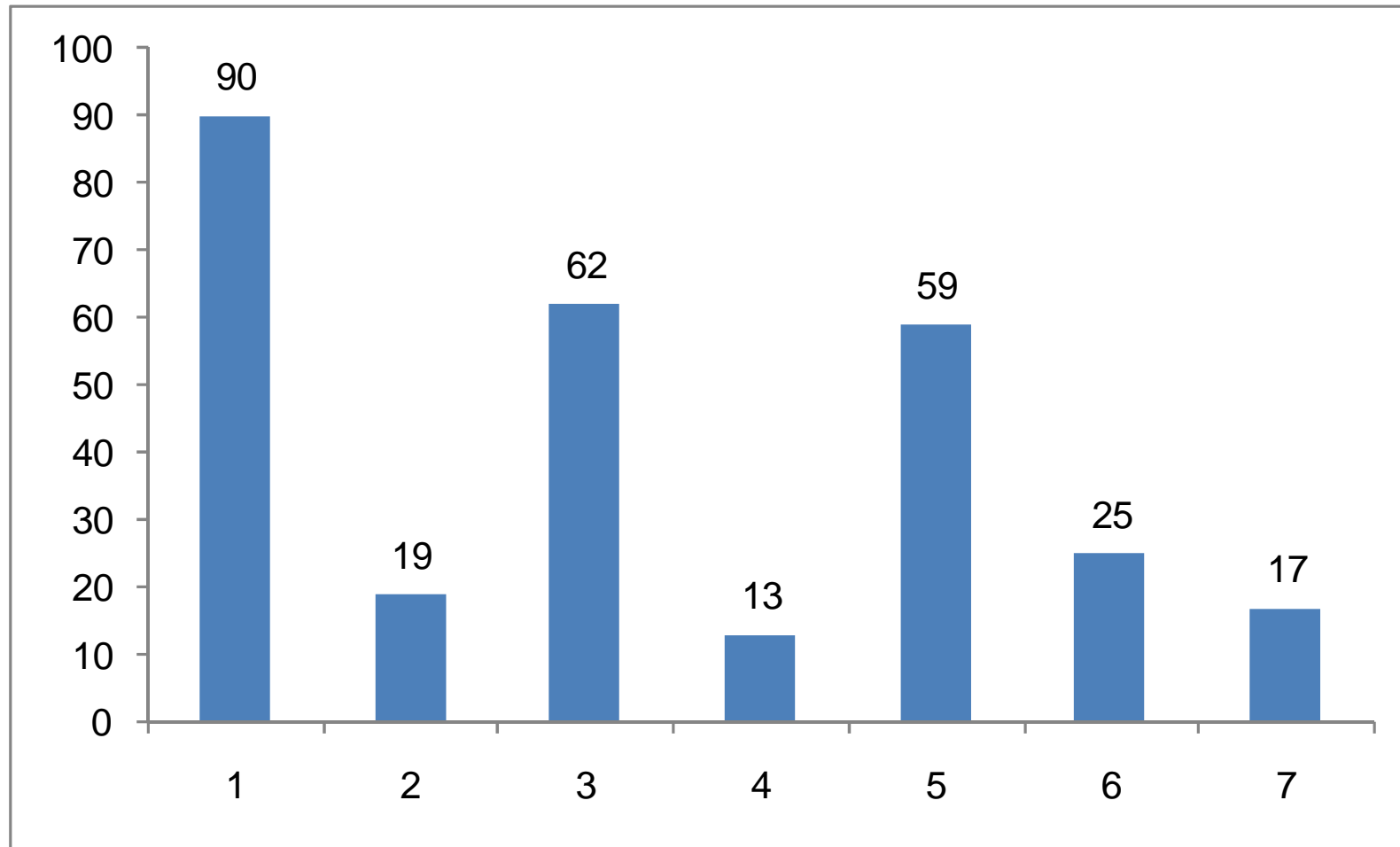
Xm02-01*

1	1	1	1	2	4	3	5	1	3	1	3	7	5	1
1	5	2	1	5	1	3	3	3	1	1	5	3	1	5
5	1	1	3	3	5	5	6	3	5	3	5	5	5	1
1	2	1	1	5	5	3	2	1	6	1	1	4	5	1
3	3	5	4	7	6	6	4	4	6	5	2	1	1	5
3	3	1	3	5	3	3	7	3	7	2	1	5	7	
3	6	2	6	3	6	6	6	5	6	1	1	6	3	
7	1	1	1	5	1	3	1	3	7	7	2	1	1	
2	5	3	1	1	3	1	1	7	5	3	2	1	1	
6	5	7	1	3	2	1	3	1	1	7	5	5	6	
1	4	6	1	3	1	1	5	5	5	5	1	5	5	
6	1	3	3	1	3	7	1	1	1	2	4	1	1	
3	3	7	5	5	1	1	3	5	1	5	4	5	3	
4	1	4	5	3	1	5	3	3	3	1	1	5	3	
5	6	4	3	5	6	4	6	5	5	5	5	3	1	
2	3	2	7	5	1	6	6	2	3	3	3	1	1	
5	1	4	6	3	5	1	1	2	1	5	6	1	1	
5	1	3	5	1	1	1	3	7	3	1	6	3	1	
2	2	5	1	3	5	5	2	3	1	1	3	6	1	
1	1	1	7	3	1	5	3	3	3	5	3	1	7	

Frequency and Relative Frequency Distributions

<u>Light Beer Brand</u>	<u>Frequency</u>	<u>Relative Frequency</u>
Budweiser Light	90	31.6%
Busch Light	19	6.7
Coors Light	62	21.8
Michelob Light	13	4.6
Miller Lite	59	20.7
Natural Light	25	8.8
<u>Other brands</u>	<u>17</u>	<u>6.0</u>
Total	285	100

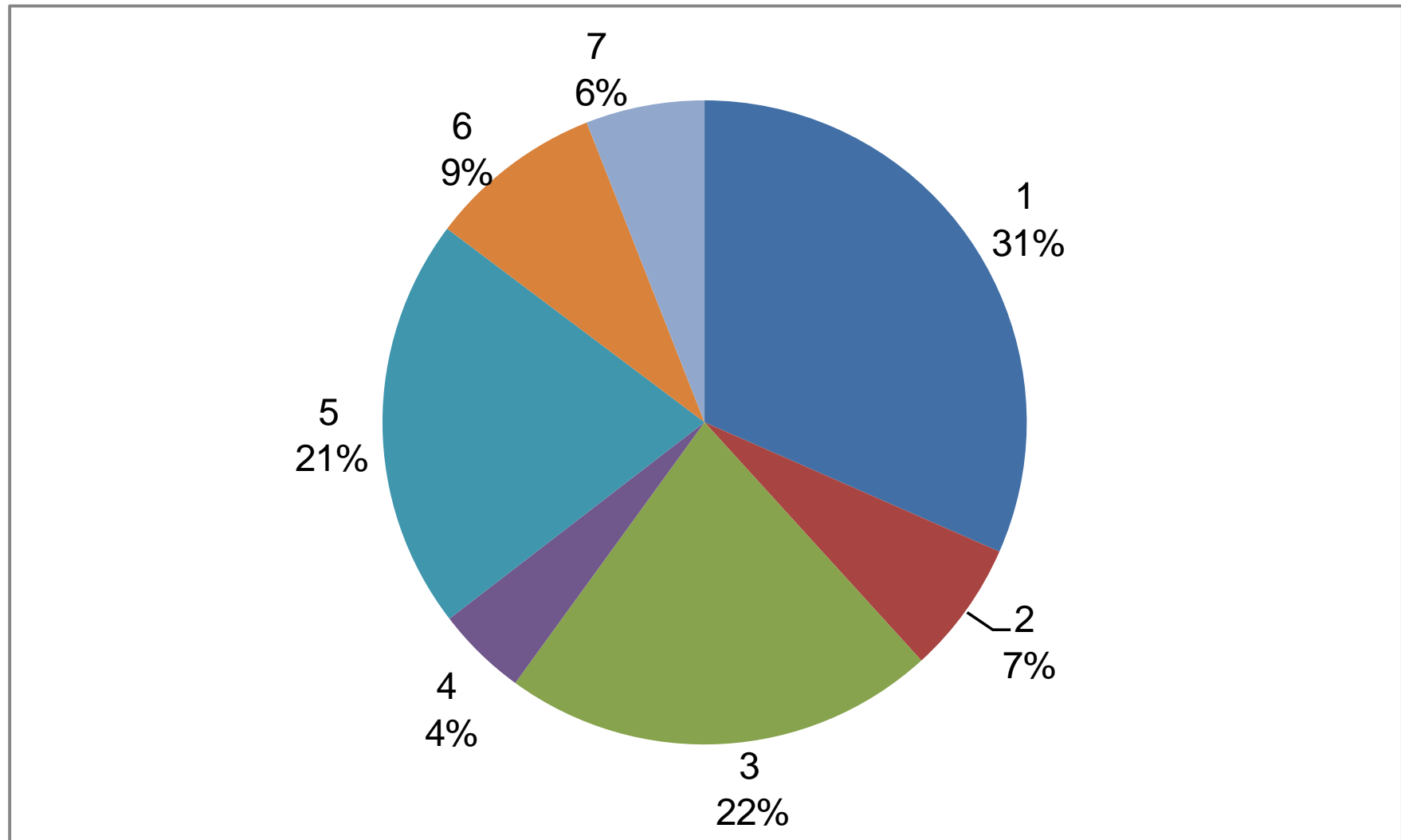
Nominal Data (Frequency)



Bar Charts are often used to display *frequencies*...

It is not valid to discuss the distribution of bar charts

Nominal Data (Relative Frequency)

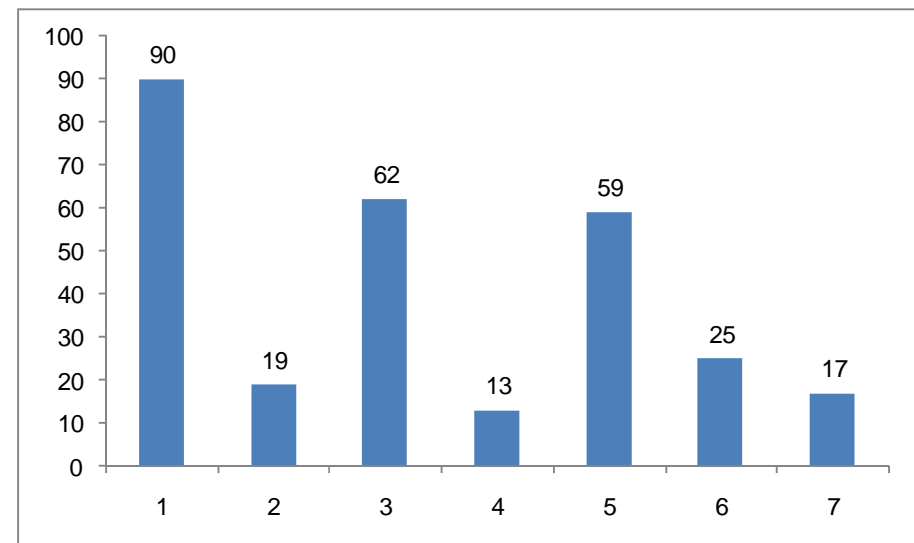
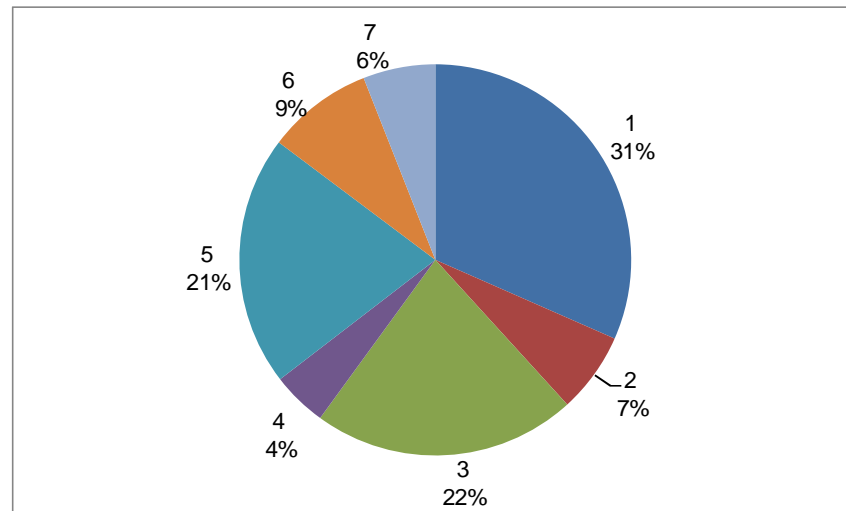


Pie Charts show *relative frequencies...*

Nominal Data

Light Beer Brand	Frequency	Relative Frequency
Budweiser Light	90	31.6%
Busch Light	19	6.7
Coors Light	62	21.8
Michelob Light	13	4.6
Miller Lite	59	20.7
Natural Light	25	8.8
Other brands	17	6.0

**It all the same *information*,
(based on the same *data*).
Just different *presentation*.**



Excel comment: COUNTIF (p20)

Example 2.2

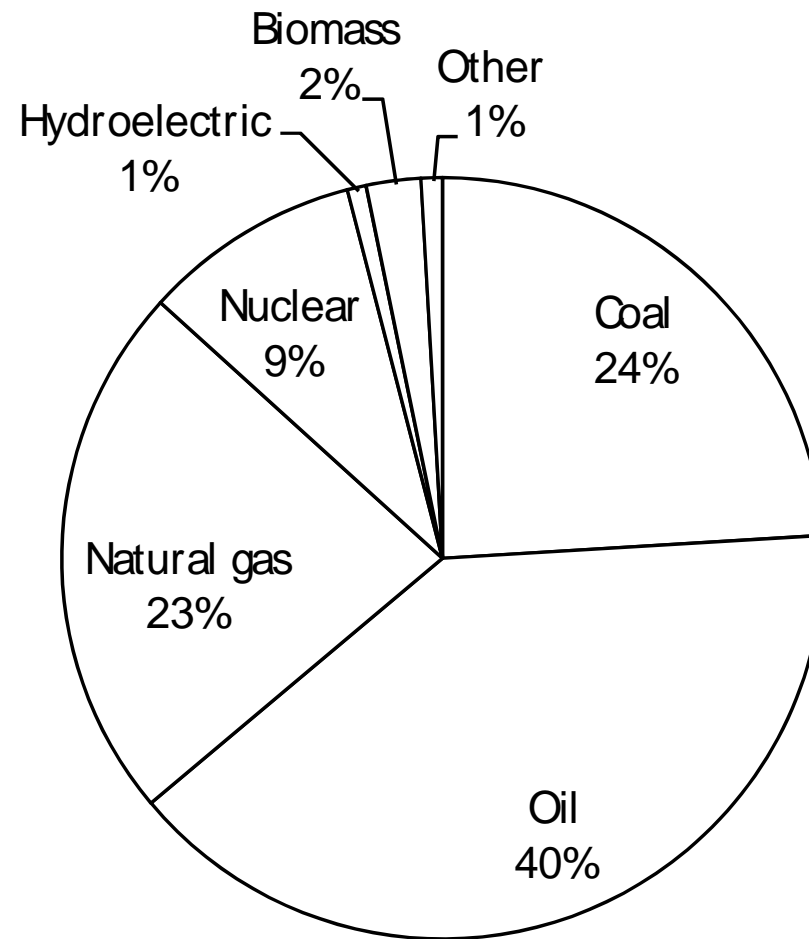
- Table 2.3 lists the total energy consumption of the United States from all sources in 2005.
- To make it easier to see the details the table measures the heat content in metric tons (1,000 kilograms) of oil equivalent.
 - For example, the United States burned an amount of coal and coal products equivalent to 545,259 metric tons of oil.
- Use an appropriate graphical technique to depict these figures.

Table 2.3

Xm02-02*

<u>Non-Renewable Energy Sources</u>	<u>Heat Content</u>
Coal & coal products	545,258
Oil	903,440
Natural Gas	517,881
Nuclear	209,890
<u>Renewable Energy Sources</u>	
Hydroelectric (水力發電)	18,251
Solid Biomass	52,473
Other (Liquid biomass, geothermal, solar, wind, and tide, wave, & Ocean)	20,533
Total	2,267,726

Example 2.2



Graphical Techniques for Interval Data

- There are several graphical methods that are used when the data are *interval* (i.e. numeric, non-categorical).
- The most important of these graphical methods is the *histogram*.
- The histogram is not only a powerful graphical technique used to *summarize* interval data, but it is also used to help *explain* probabilities.

Example 2.4

- Following deregulation of telephone service, several new companies were created to compete in the business of providing long-distance telephone service.
- In almost all cases these companies competed on price since the service each offered is similar.
- Pricing a service or product in the face of stiff competition is very difficult.
- Factors to be considered include
 - supply, demand, price elasticity, and the actions of competitors.

Example 2.4

- Long-distance packages may employ
 - per-minute charges, a flat monthly rate, or some combination of the two.
- Determining the appropriate rate structure is facilitated by acquiring information about the behaviors of customers and in particular the size of monthly long-distance bills.
- As part of a larger study, a long-distance company wanted to acquire information about the monthly bills of new subscribers in the first month after signing with the company.

Example 2.4

- The company's marketing manager conducted a survey of 200 new residential subscribers wherein the first month's bills were recorded.
- These data are stored in file [Xm02-04](#).
- The general manager planned to present his findings to senior executives.
 - What information can be extracted from these data?

Example 2.4

- In Example 2.1 we created a frequency distribution of the 5 categories.
- In this example we also create a frequency distribution by counting the number of observations that fall into a series of intervals, called classes.
- I'll explain later why I chose the classes I use below.

Example 2.4

- We have chosen eight classes defined in such a way that each observation falls into one and only one class.
- These classes are defined as follows:

Classes

Amounts that are less than or equal to 15

Amounts that are more than 15 but less than or equal to 30

Amounts that are more than 30 but less than or equal to 45

Amounts that are more than 45 but less than or equal to 60

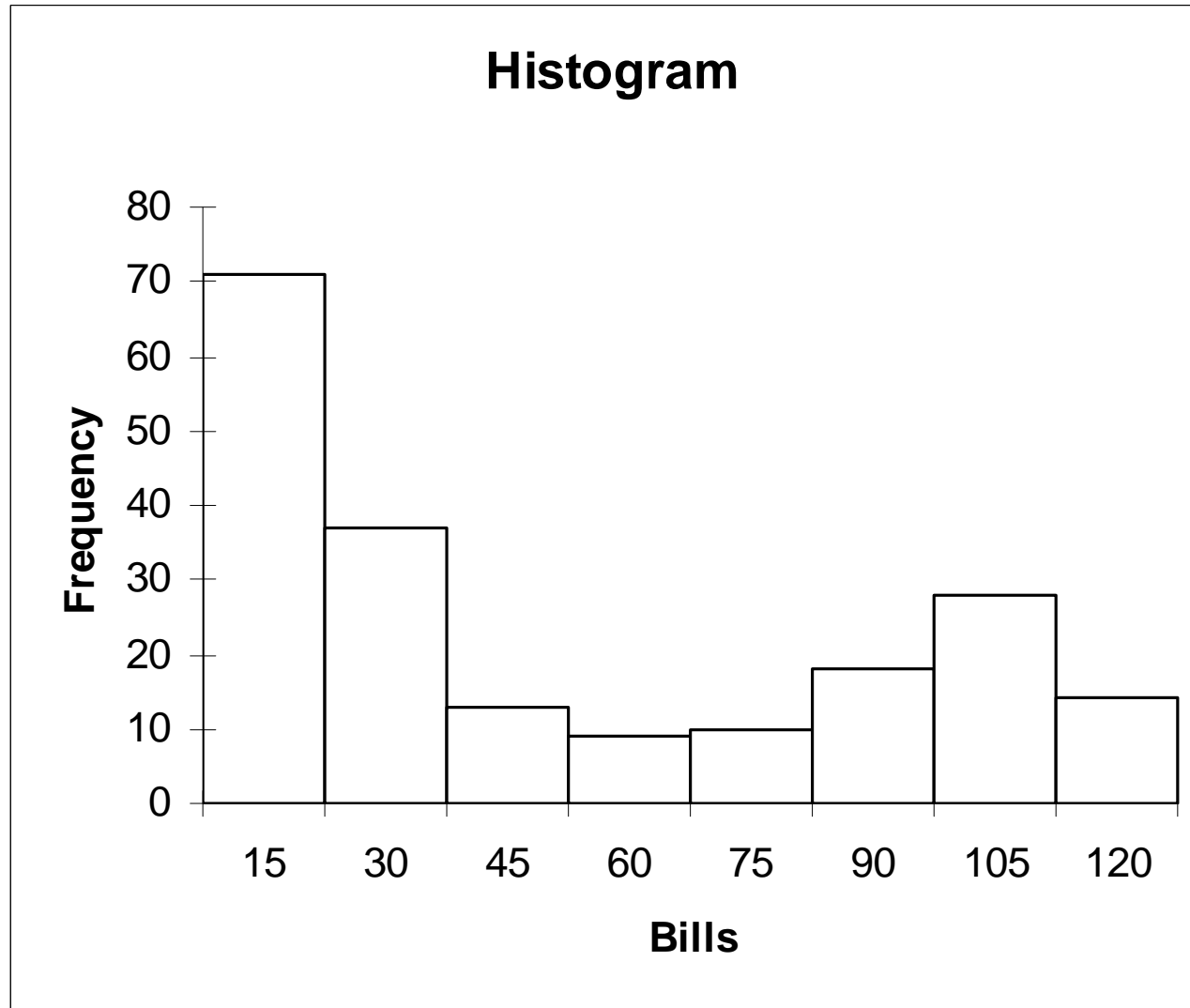
Amounts that are more than 60 but less than or equal to 75

Amounts that are more than 75 but less than or equal to 90

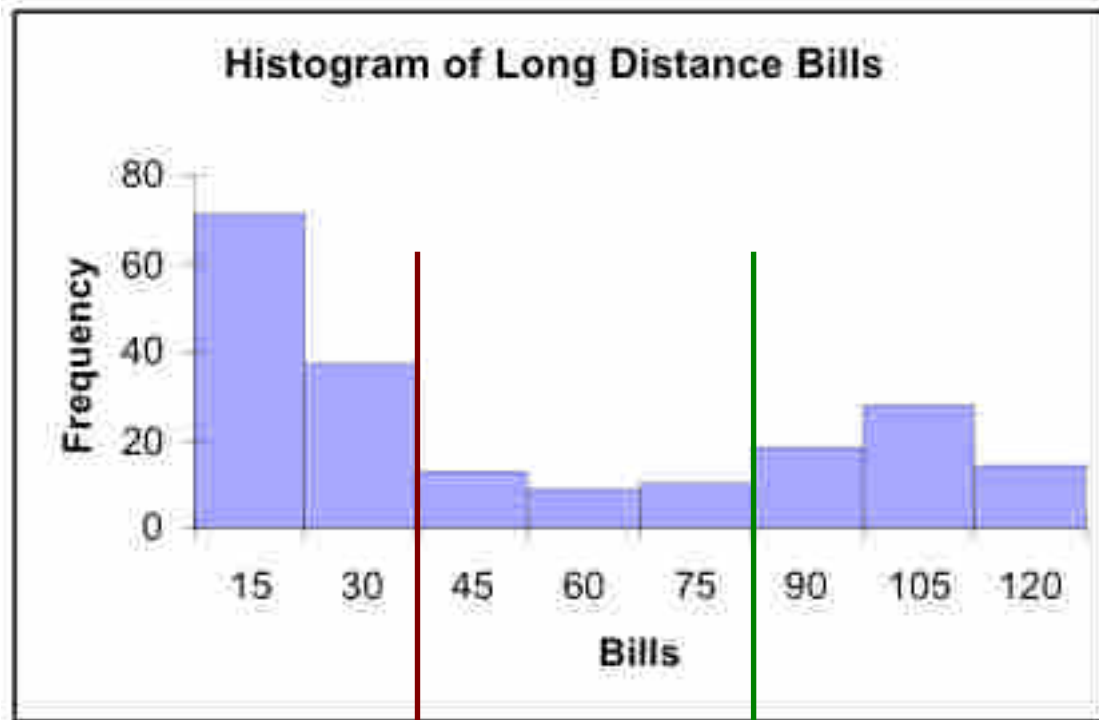
Amounts that are more than 90 but less than or equal to 105

Amounts that are more than 105 but less than or equal to 120

Example 2.4



Interpret...



about half ($71+37=108$)
of the bills are "small",
i.e. less than \$30

$(18+28+14=60) \div 200 = 30\%$
i.e. nearly a third of the phone bills
are \$90 or more.

There are only a few telephone
bills in the middle range.

Building a Histogram...

- 1) Collect the Data
- 2) Create a frequency distribution for the data...

How?

- a) Determine the number of *classes* to use...

How?

Refer to table 2.6:

With 200 observations,
we should have
between 7 & 10
classes...

Table 2.6 Approximate Number of Classes in Frequency Distributions

<u>Number of Observations</u>	<u>Number of Classes</u>
Less than 50	5 - 7
50 - 200	7 - 9
200 - 500	9 - 10
500 - 1,000	10 - 11
1,000 - 5,000	11 - 13
5,000 - 50,000	13 - 17
<u>More than 50,000</u>	<u>17 - 20</u>

Alternative, we could use Sturges' formula:
Number of class intervals = $1 + 3.3 \log(n)$

Building a Histogram...

- 1) Collect the Data
- 2) Create a frequency distribution for the data...

How?

- a) Determine the number of *classes* to use. [8]
- b) Determine how large to make each class...

How?

Look at the *range* of the data, that is,

Range = Largest Observation – Smallest Observation

Range = \$119.63 – \$0 = \$119.63

Then each class width becomes:

Range ÷ (# classes) = 119.63 ÷ 8 ≈ 15

Building a Histogram...

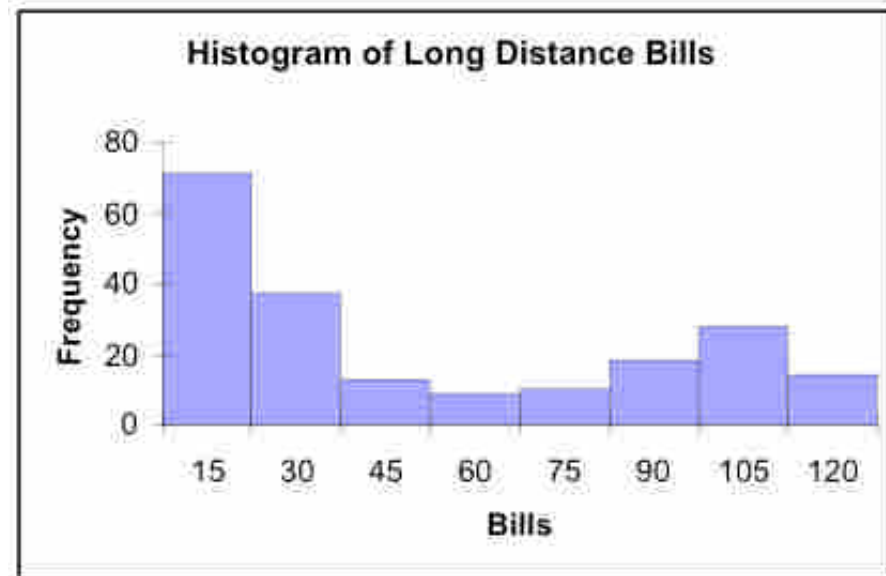
Table 2.5 Frequency Distribution of the Long-Distance Bills in Example 2.4

<u>Class Limits</u>	<u>Frequency</u>
0 to 15	71
15 to 30	37
30 to 45	13
45 to 60	9
60 to 75	10
75 to 90	18
90 to 105	28
<u>105 to 120</u>	<u>14</u>
Total	200

Building a Histogram...

Table 2.5 Frequency Distribution of the Long-Distance Bills in Example 2.4

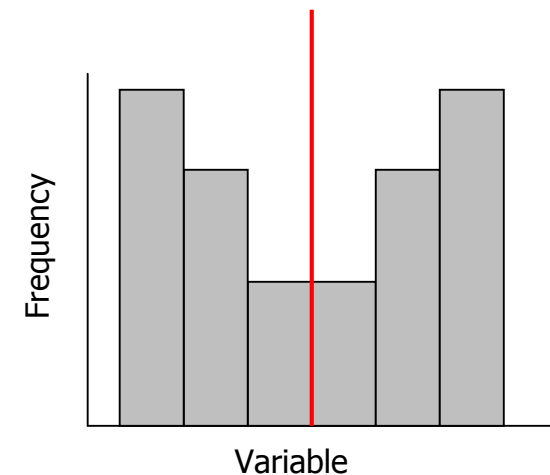
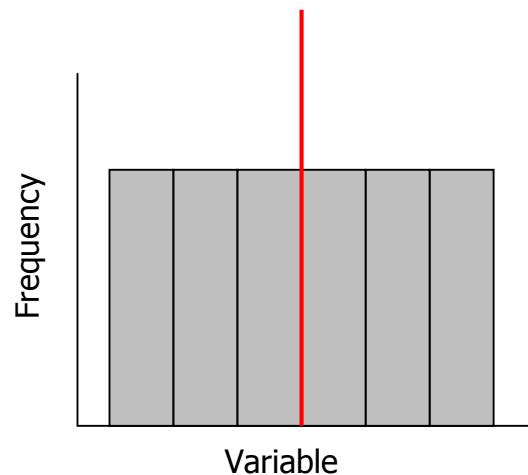
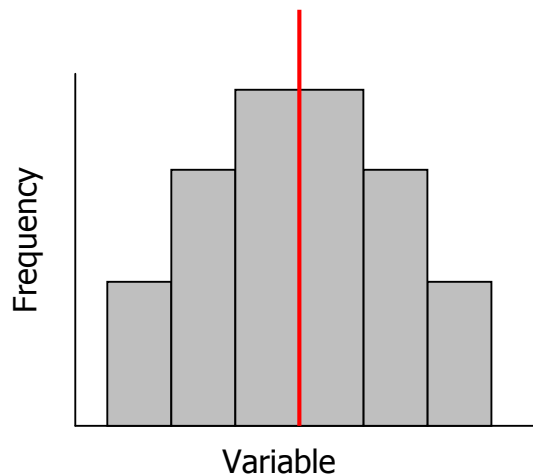
<u>Class Limits</u>	<u>Frequency</u>
0 to 15	71
15 to 30	37
30 to 45	13
45 to 60	9
60 to 75	10
75 to 90	18
90 to 105	28
<u>105 to 120</u>	<u>14</u>
Total	200



Shapes of Histograms...

Symmetry

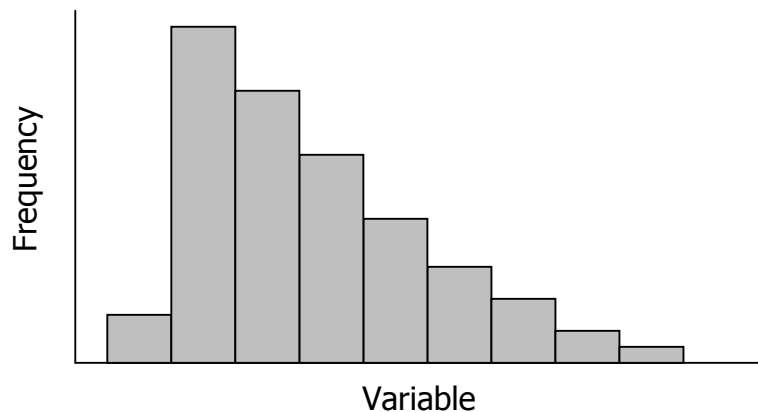
A histogram is said to be *symmetric* if, when we draw a **vertical line** down the center of the histogram, the two sides are identical in shape and size:



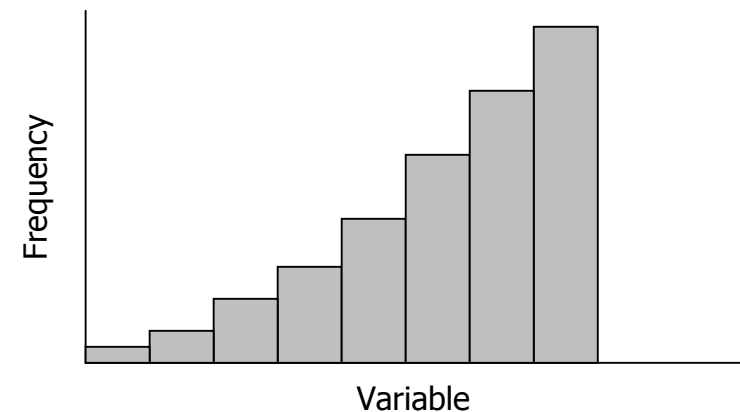
Shapes of Histograms...

Skewness

A skewed histogram is one with a long tail extending to either the right or the left:



Positively Skewed



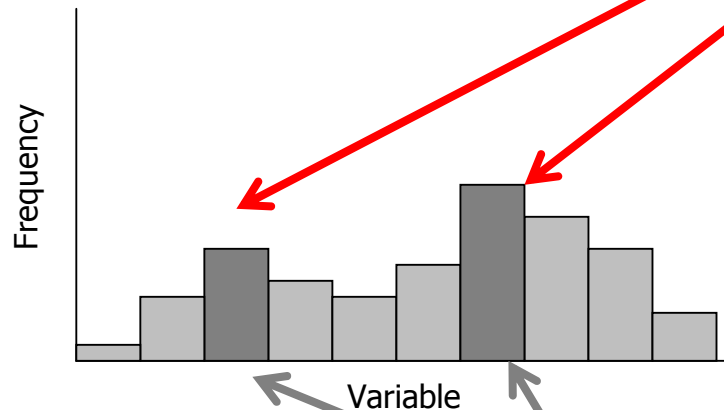
Negatively Skewed

Shapes of Histograms...

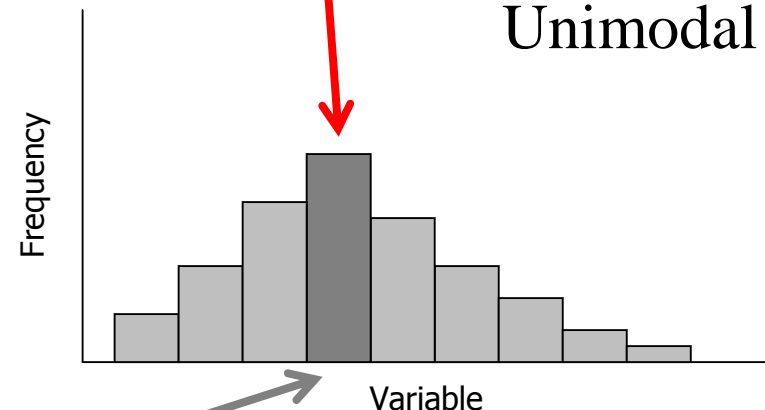
Modality

A *unimodal* histogram is one with a single peak, while a *bimodal* histogram is one with two peaks:

Bimodal



Unimodal



A *modal class* is the class with the largest number of observations

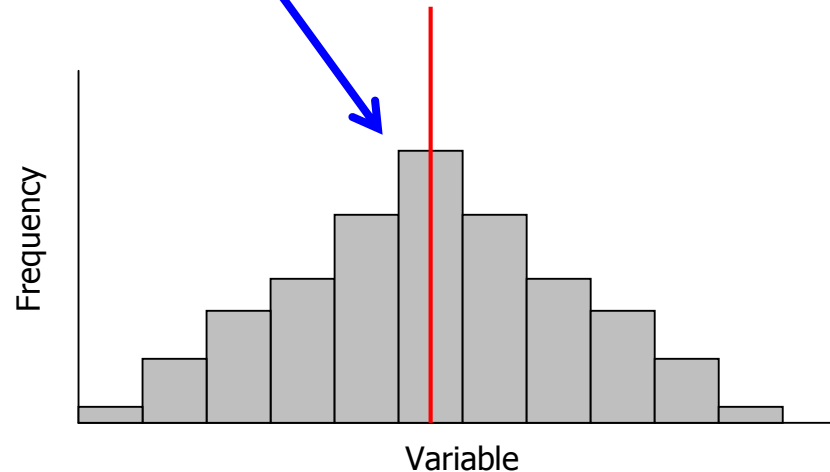
Shapes of Histograms...

Bell Shape

A special type of *symmetric unimodal* histogram is one that is bell shaped:

Many statistical techniques require that the population be bell shaped.

Drawing the histogram helps verify the shape of the population in question.



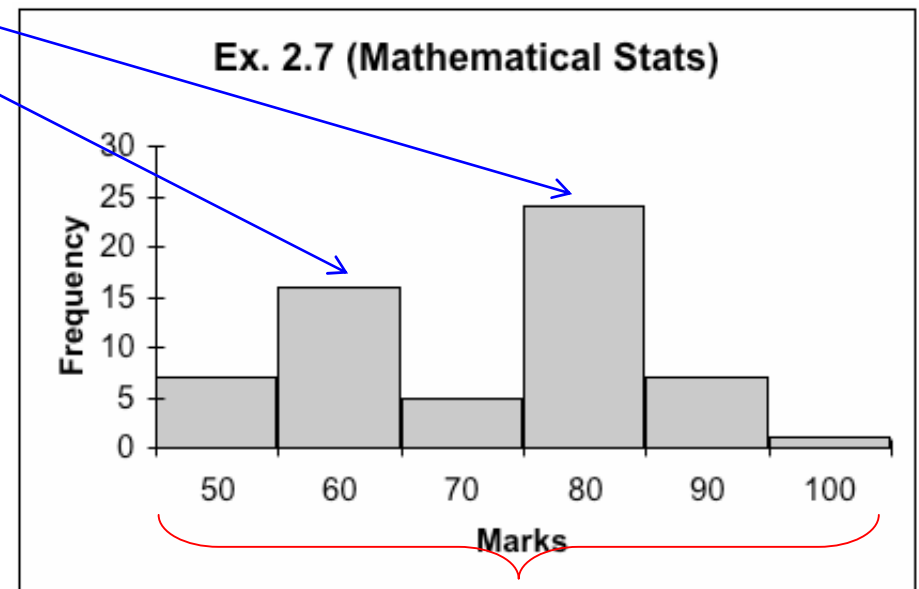
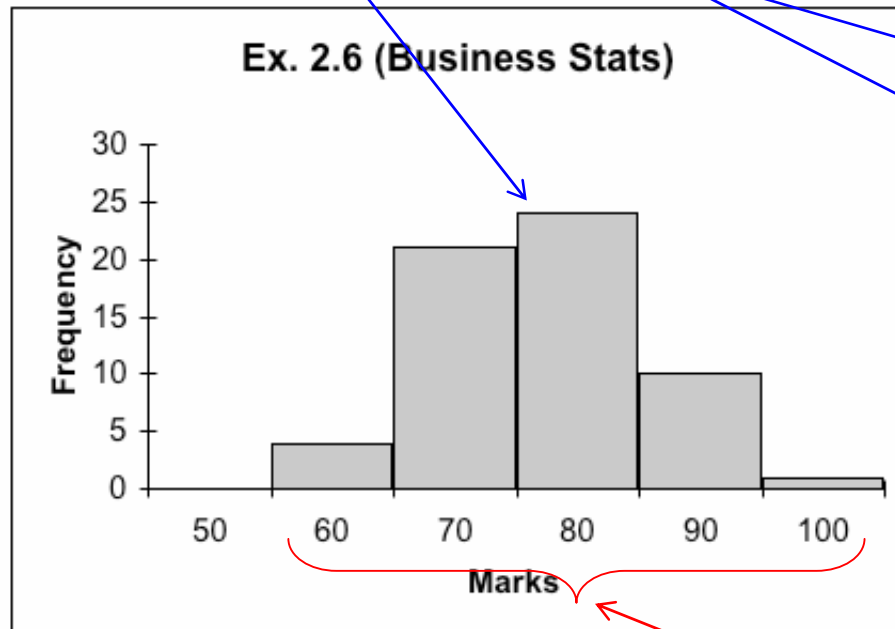
Bell Shaped

Histogram Comparison...

Compare & contrast the following histograms based on data from [Ex. 2.6](#) & [Ex. 2.7](#):

The two courses, Business Statistics and Mathematical Statistics have very different histograms...

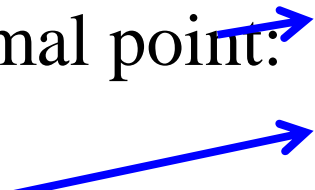
unimodal vs. bimodal



spread of the marks (narrower | wider)

Stem & Leaf Display...

- Retains information about individual observations that would normally be lost in the creation of a histogram.
- Split each observation into two parts, a *stem* and a *leaf*:
- e.g. Observation value: **42.19**
- There are several ways to split it up...
- We could split it at the decimal point:
- Or split it at the “tens” position (while rounding to the nearest integer in the “ones” position)

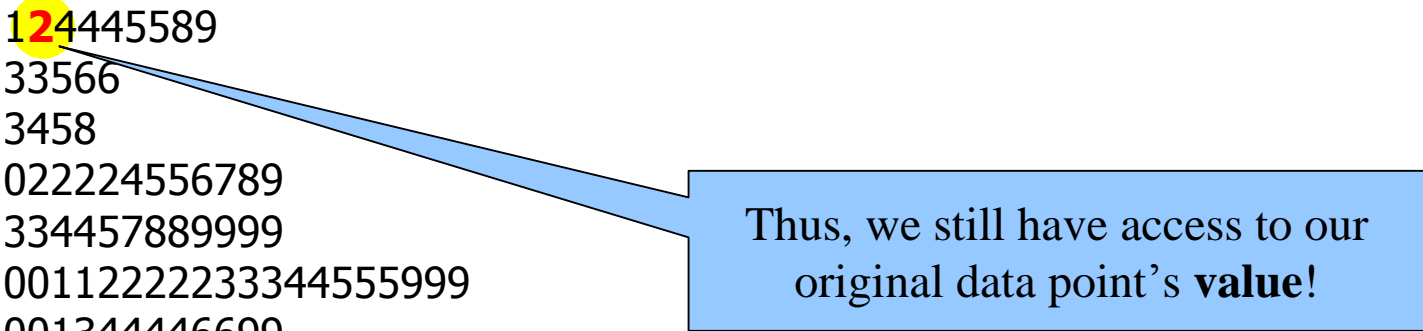


Stem	Leaf
42	19
4	2

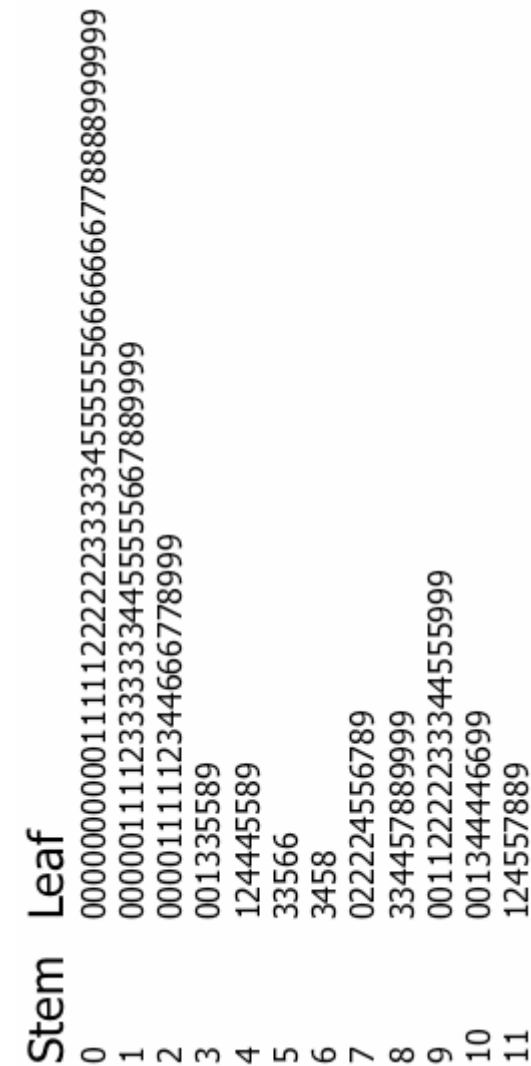
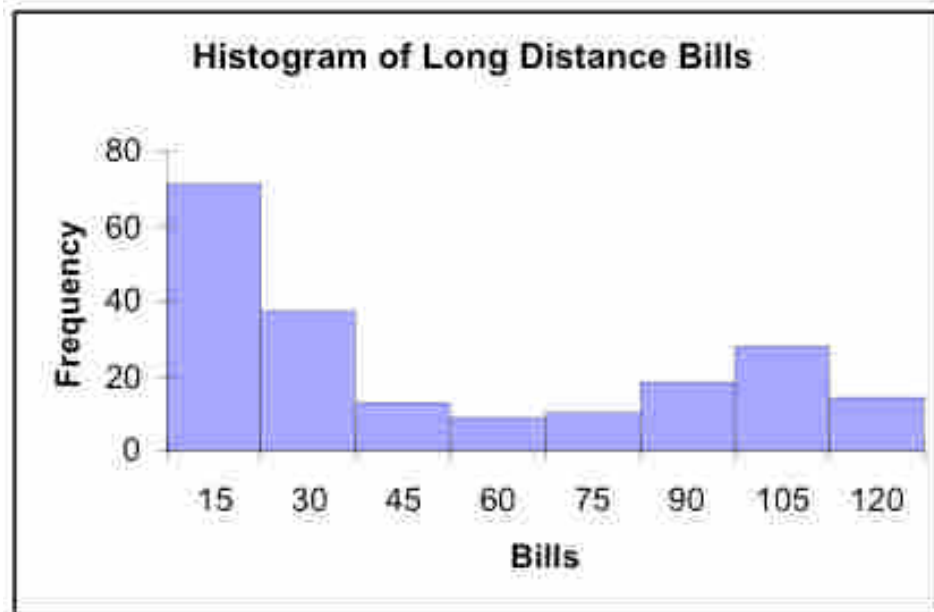
Stem & Leaf Display...

- Continue this process for all the observations. Then, use the “stems” for the classes and each leaf becomes part of the histogram (based on Example 2.4 data) as follows...

Stem	Leaf
0	0000000000111112222223333345555556666666778888999999
1	000001111233333334455555667889999
2	0000111112344666778999
3	001335589
4	124445589
5	33566
6	3458
7	022224556789
8	334457889999
9	00112222233344555999
10	001344446699
11	124557889



Thus, we still have access to our original data point's **value**!



Ogive...

- (pronounced “Oh-jive”) is a graph of a *cumulative frequency distribution*.
- We create an ogive in three steps...
- First, from the frequency distribution created [earlier](#), calculate *relative frequencies*:
- $$\text{Relative Frequency} = \frac{\text{\# of observations in a class}}{\text{Total \# of observations}}$$

Relative Frequencies...

- For example, we had 71 observations in our first class (telephone bills from \$0.00 to \$15.00). Thus, the relative frequency for this class is $71 \div 200$ (the total # of phone bills) = 0.355 (or 35.5%)

Table 2.7 Relative Frequency Distribution for Example 2.4

<u>Class Limits</u>	<u>Relative Frequency</u>
0 to 15	$71/200 = .355$
15 to 30	$37/200 = .185$
30 to 45	$13/200 = .065$
45 to 60	$9/200 = .045$
60 to 75	$10/200 = .050$
75 to 90	$18/200 = .090$
90 to 105	$28/200 = .140$
<u>105 to 120</u>	<u>$14/200 = .070$</u>
Total	$200/200 = 1.0$

Ogive...

- Is a graph of a *cumulative frequency distribution*.
- We create an ogive in three steps...
- 1) Calculate relative frequencies. ✓
- 2) Calculate *cumulative relative frequencies* by adding the current class' relative frequency to the previous class' cumulative relative frequency.
- (For the first class, its cumulative relative frequency is just its relative frequency)

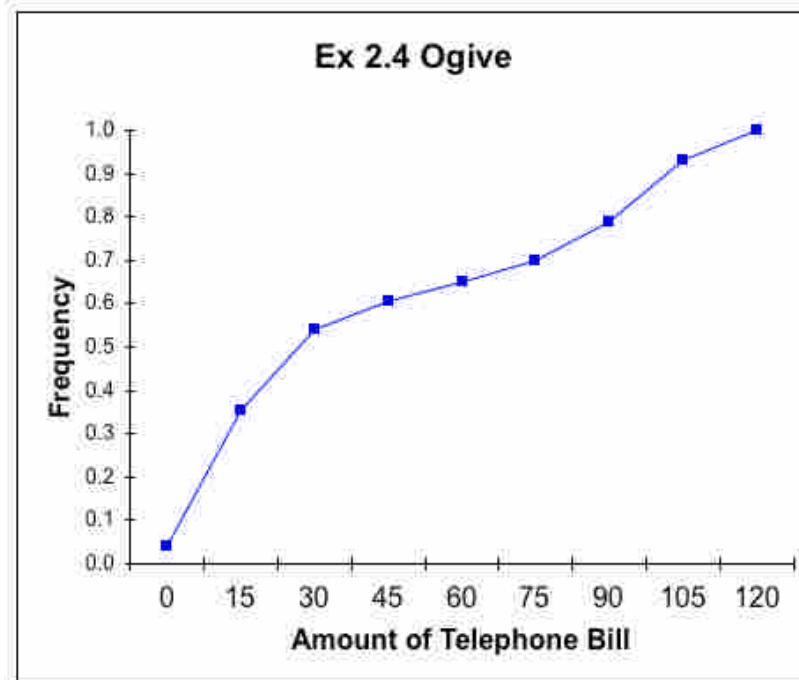
Cumulative Relative Frequencies...

Table 2.8 Cumulative Relative Frequency Distribution for Example 2.4

Class Limits	Relative Frequency	Cumulative Relative Frequency	
0 to 15	$71/200 = .355$	$71/200 = .355$	first class...
15 to 30	$37/200 = .185$	$108/200 = .540$	next class: $.355 + .185 = .540$
30 to 45	$13/200 = .065$	$121/200 = .605$	
45 to 60	$9/200 = .045$	$130/200 = .650$:
60 to 75	$10/200 = .05$	$140/200 = .700$:
75 to 90	$18/200 = .09$	$158/200 = .790$	
90 to 105	$28/200 = .14$	$186/200 = .930$	
105 to 120	$14/200 = .07$	$200/200 = 1.00$	last class: $.930 + .070 = 1.00$

Ogive...

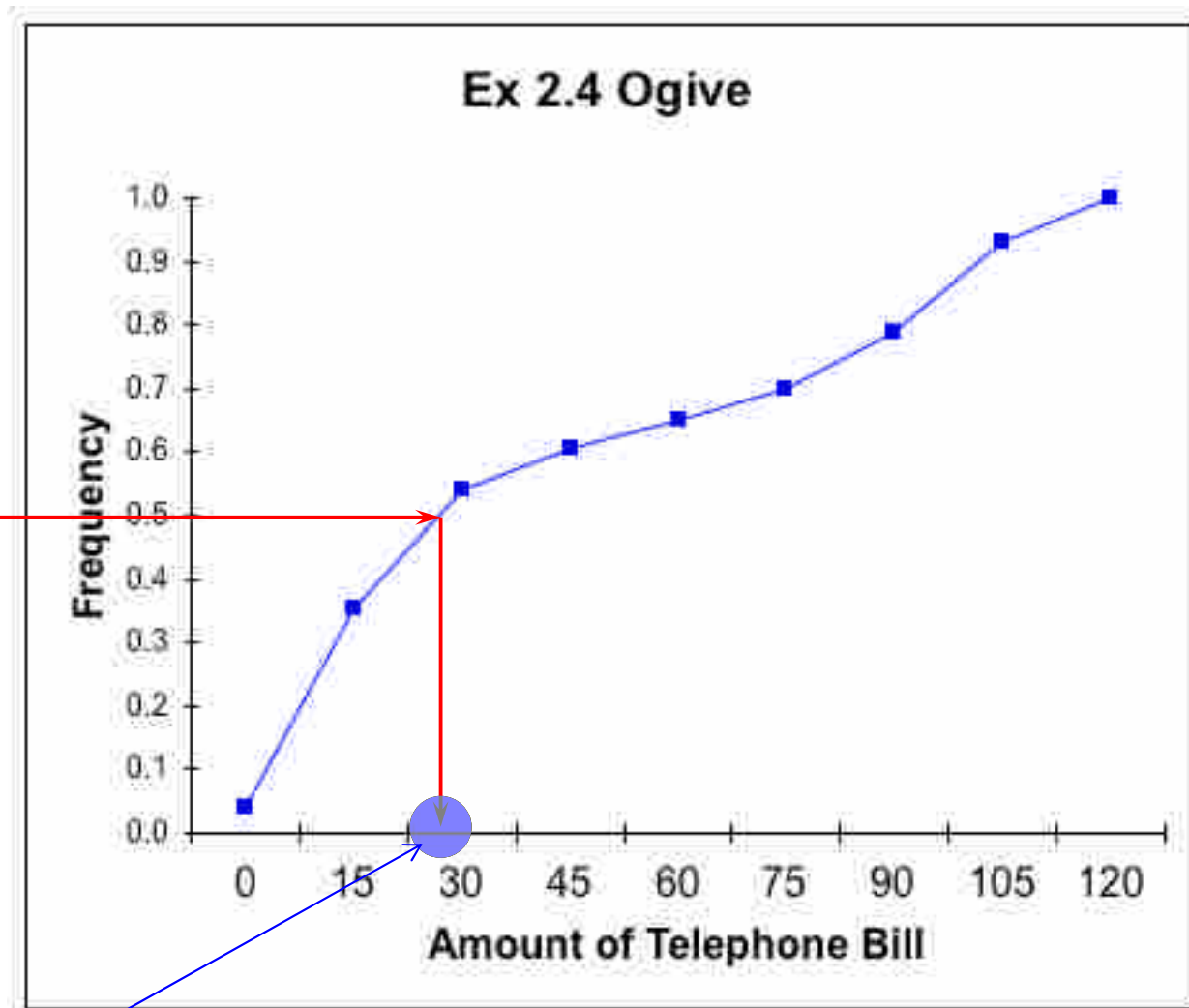
- Is a graph of a *cumulative frequency distribution*.
- 1) Calculate relative frequencies. ✓
- 2) Calculate cumulative relative frequencies. ✓
- 3) Graph the cumulative relative frequencies...



Ogive...

The ogive can be used to answer questions like:

What telephone bill value is at the 50th percentile?



“around \$35”

(Refer also to Fig. 2.13 in your textbook)

Describing Time Series Data

Observations measured at the same point in time are called *cross-sectional* data.

Observations measured at successive points in time are called *time-series* data.

Time-series data graphed on a *line chart*, which plots the value of the variable on the vertical axis against the time periods on the horizontal axis.

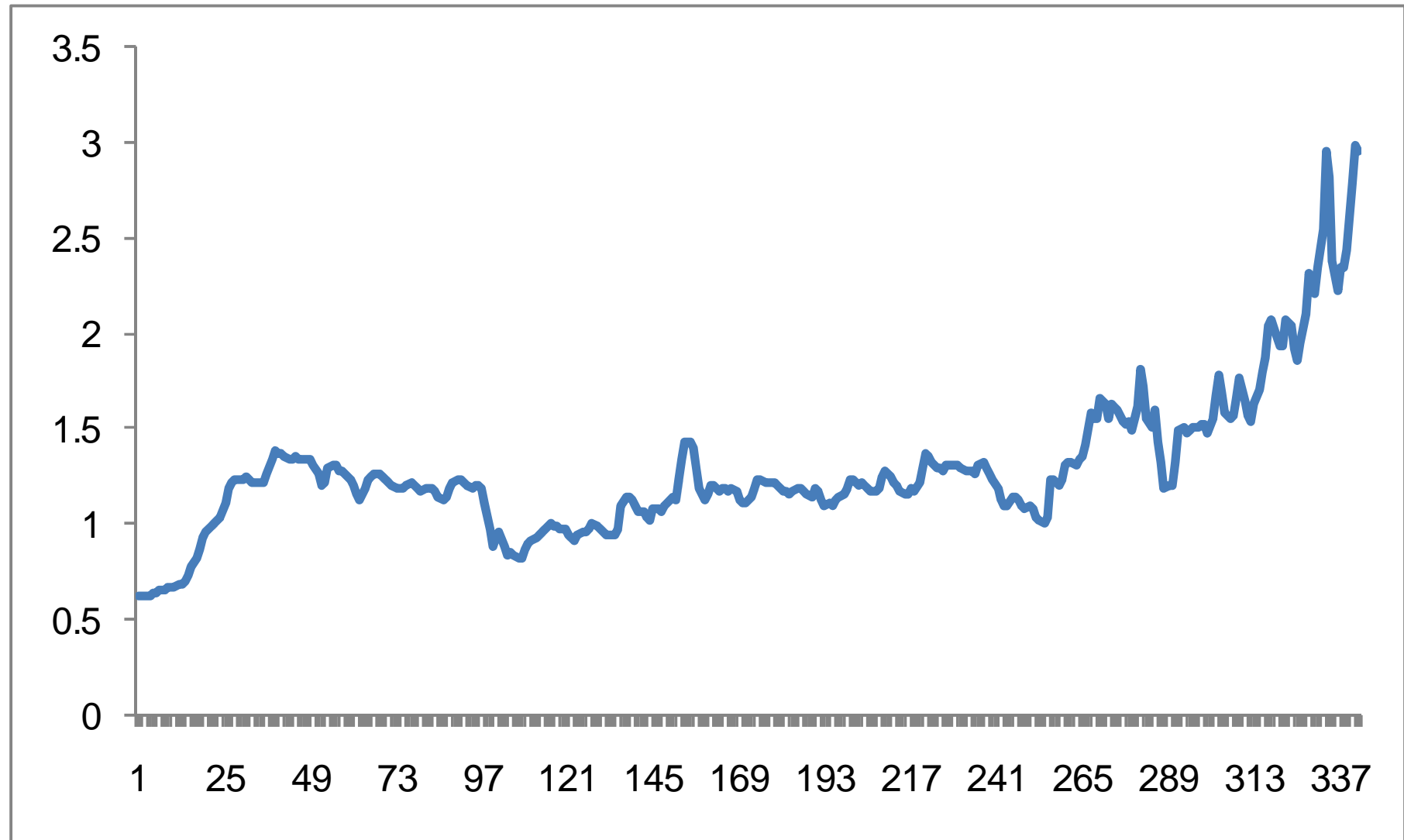
Example 2.8

We recorded the monthly average retail price of gasoline since 1978.

[Xm02-08](#)

Draw a line chart to describe these data and briefly describe the results.

Example 2.8



Example 2.9 Price of Gasoline in 1982-84 Constant Dollars

Xm02-09

Remove the effect of inflation in Example 2.8 to determine whether gasoline prices are higher than they have been in the past after removing the effect of inflation.

Example 2.9



Relationship between Two Nominal Variables...

So far we've looked at tabular and graphical techniques for one variable (either nominal or interval data).

A *cross-classification table* (or cross-tabulation table) is used to describe the relationship between **two** nominal variables.

A cross-classification table lists the *frequency* of *each combination* of the values of the two variables...

Example 2.10

In a major North American city there are four competing newspapers: the Post, Globe and Mail, Sun, and Star.

To help design advertising campaigns, the advertising managers of the newspapers need to know which segments of the newspaper market are reading their papers.

A survey was conducted to analyze the relationship between newspapers read and occupation.

Example 2.10

A sample of newspaper readers was asked to report which newspaper they read: Globe and Mail (1) Post (2), Star (3), Sun (4), and to indicate whether they were blue-collar worker (1), white-collar worker (2), or professional (3).

The responses are stored in file [Xm02-10](#).

Example 2.10

By counting the number of times each of the 12 combinations occurs, we produced the Table 2.9.

Newspaper	Occupation			Total
	Blue Collar	White Collar	Professional	
G&M	27	29	33	89
Post	18	43	51	112
Star	38	21	22	81
Sun	37	15	20	72
Total	120	108	126	354

Example 2.10

If occupation and newspaper are related, then there will be differences in the newspapers read among the occupations. An easy way to see this is to convert the frequencies in each column to relative frequencies in each column. That is, compute the column totals and divide each frequency by its column total.

Newspaper	Occupation		
	Blue Collar	White Collar	Professional
G&M	$27/120 = .23$	$29/108 = .27$	$33/126 = .26$
Post	$18/120 = .15$	$43/108 = .40$	$51/126 = .40$
Star	$38/120 = .32$	$21/108 = .19$	$22/126 = .17$
Sun	$37/120 = .31$	$15/108 = .14$	$20/126 = .16$

Example 2.10

Interpretation: The relative frequencies in the columns 2 & 3 are similar, but there are large differences between columns 1 and 2 and between columns 1 and 3.

Table 2.10 Column Relative Frequencies for Example 2.8

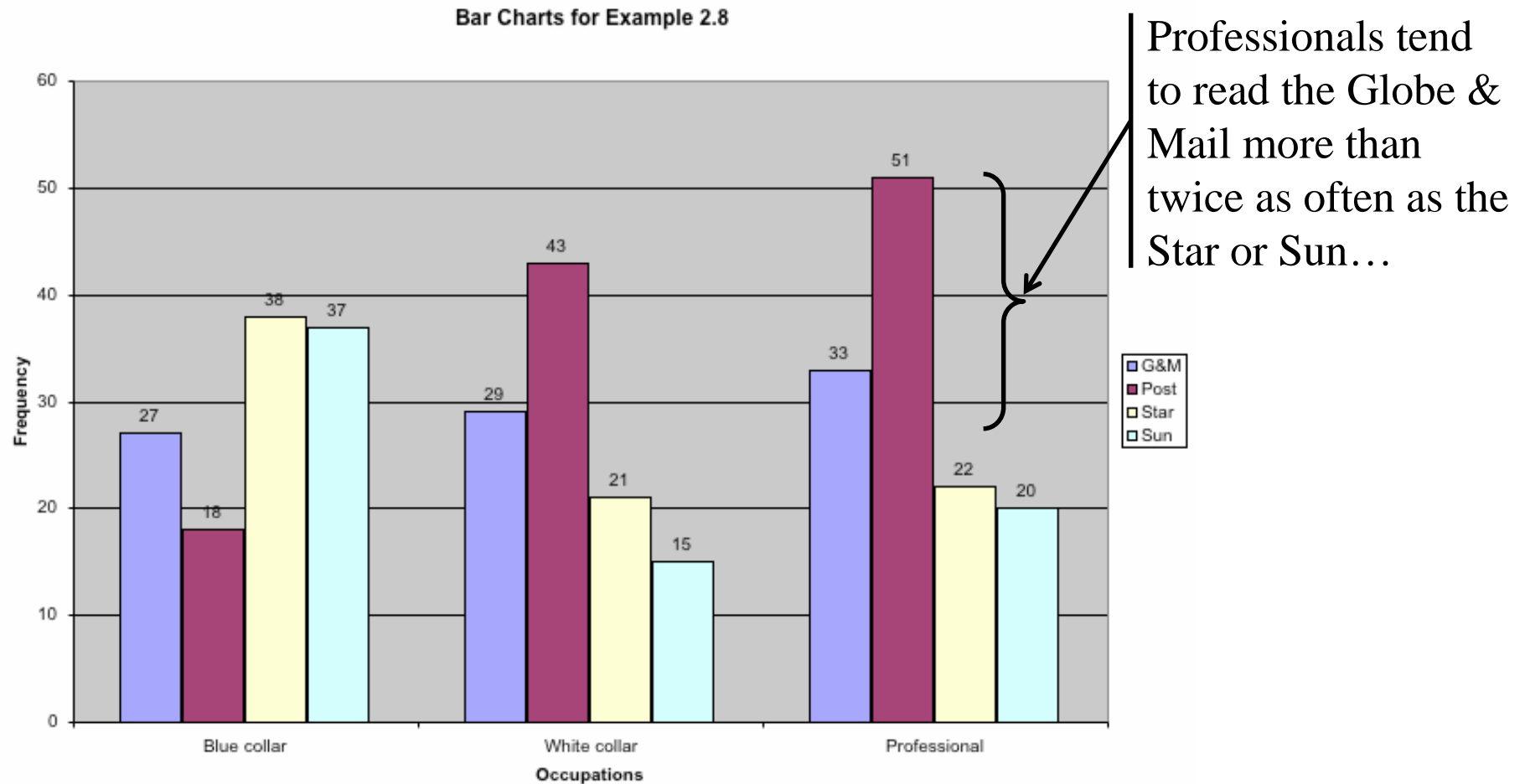
Newspaper	Occupation			
	Blue Collar	White Collar	Professional	
G&M	$27/120 = .23$	$29/108 = .27$	$33/126 = .26$	similar
Post	$18/120 = .15$	$43/108 = .40$	$51/126 = .40$	
Star	$38/120 = .32$	$21/108 = .19$	$22/126 = .17$	
Sun	$37/120 = .31$	$15/108 = .14$	$20/126 = .16$	

dissimilar

This tells us that blue collar workers tend to read different newspapers from both white collar workers and professionals and that white collar and professionals are quite similar in their newspaper choice.

Graphing the Relationship Between Two Nominal Variables...

Use the data from the cross-classification table to create bar charts...



Graphing the Relationship Between Two **Interval** Variables...

Moving from nominal data to interval data, we are frequently interested in how two interval variables are related.

To explore this relationship, we employ a *scatter diagram*, which plots two variables against one another.

The *independent* variable is labeled X and is usually placed on the horizontal axis, while the other, *dependent* variable, Y, is mapped to the vertical axis.

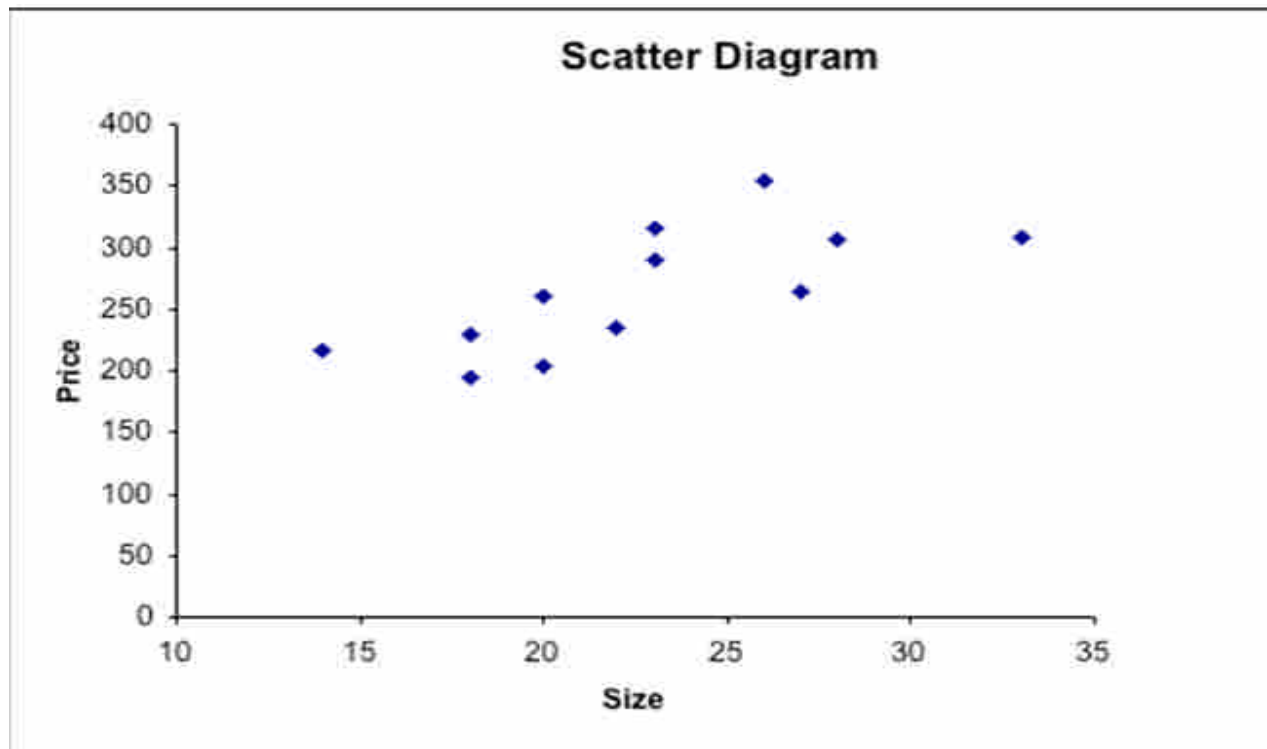
Example 2.12

A real estate agent wanted to know to what extent the selling price of a home is related to its size. To acquire this information he took a sample of 12 homes that had recently sold, recording the price in thousands of dollars and the size in hundreds of square feet. These data are listed in the accompanying table. Use a graphical technique to describe the relationship between size and price. [Xm02-12](#)

Size	23	18	26	20	22	14	33	28	23	20	27	18
Price	315	229	355	261	234	216	308	306	289	204	265	195

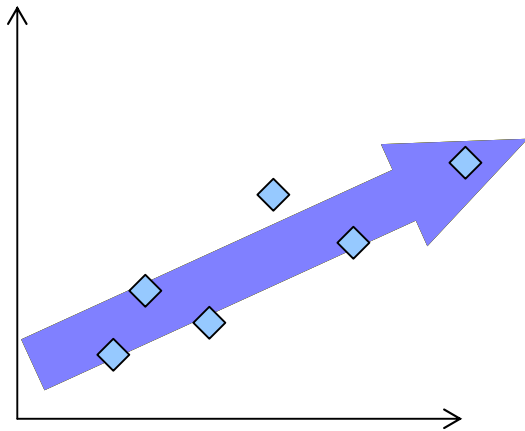
Example 2.12

It appears that in fact there is a relationship, that is, the greater the house size the greater the selling price...

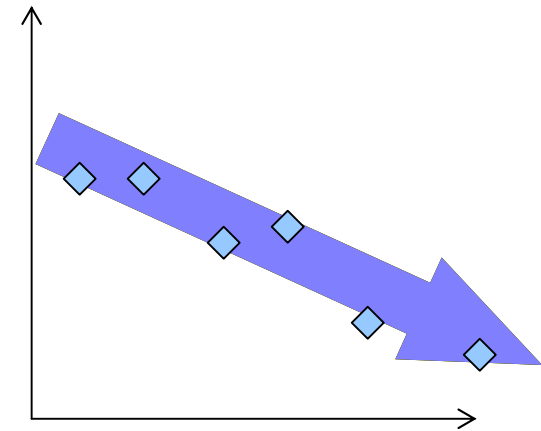


Patterns of Scatter Diagrams...

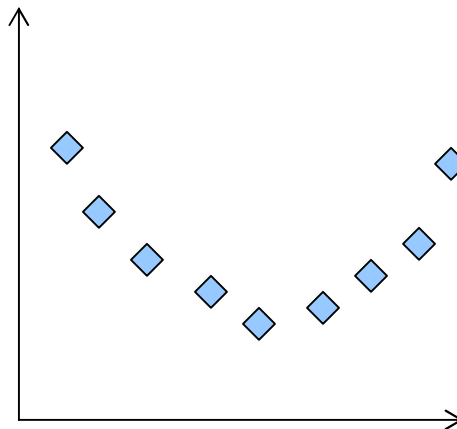
Linearity and Direction are two concepts we are interested in



Positive Linear Relationship



Negative Linear Relationship



Weak or Non-Linear Relationship

Chapter-Opening Example

WERE OIL COMPANIES GOUGING CUSTOMERS 1999-2006: SOLUTION

In December 1998 the average retail price of gasoline was \$0.913 per gallon and the price of oil (West Texas intermediate crude) was \$11.28 per barrel. Over the next 8 years the price of both substantially increased. Many drivers complained that the oil companies were guilty of price gouging. That is, they believed that when the price of oil increased the price of gas also increased, but when the price of oil decreased, the decrease in the price of gasoline seemed to lag behind.

Chapter-Opening Example

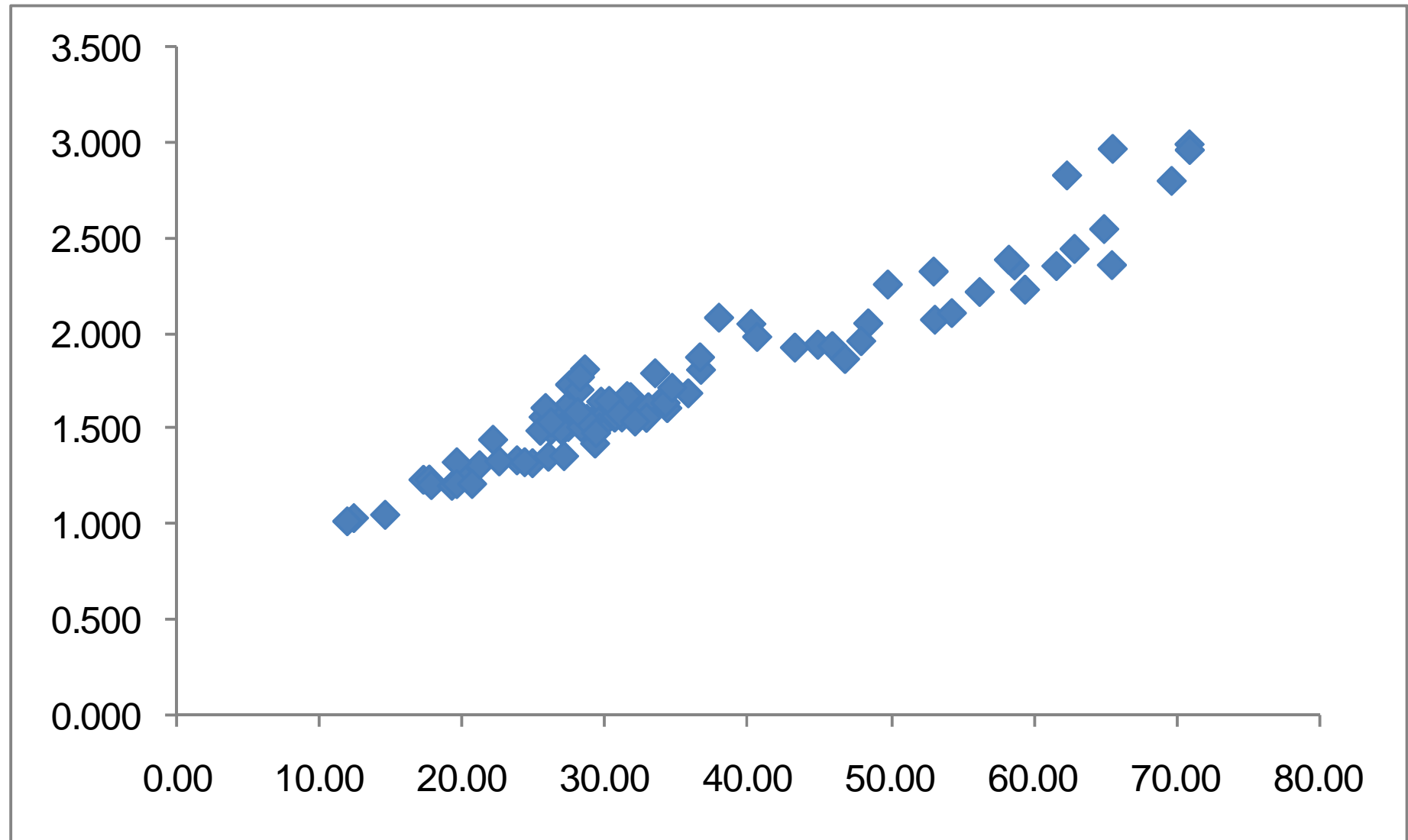
WERE OIL COMPANIES GOUGING CUSTOMERS 1999-2006: SOLUTION

To determine whether this perception is accurate we determined the monthly figures for both commodities.

[Xm02-00](#)

Graphically depict these data and describe the findings.

Chapter-Opening Example



Summary I...

Factors That Identify When to Use Frequency and Relative Frequency Tables, Bar and Pie Charts

1. Objective: Describe a single set of data.
2. Data type: Nominal

Factors That Identify When to Use a Histogram, Ogive, or Stem-and-Leaf Display

1. Objective: Describe a single set of data.
2. Data type: Interval

Factors that Identify When to Use a Cross-classification Table

1. Objective: Describe the relationship between two variables.
2. Data type: Nominal

Factors that Identify When to Use a Scatter Diagram

1. Objective: Describe the relationship between two variables.
2. Data type: Interval

Summary II...

	Interval Data	Nominal Data
Single Set of Data	Histogram	Frequency and Relative Frequency Tables, Bar and Pie Charts
Relationship Between Two Variables	Scatter Diagram	Cross-classification Table, Bar Charts

Exercises

- 2, 4, 8, 9, 10, 13, 16, 20, 35, 38, 42, 85, 88