Chapter 2—Organizing and Summarizing Data

The Statistical Process (p. 4) involves four steps:

1. *Identify the research objective.*
2. *Collect information to answer the question posed in (1).*
3. *Organize and summarize the information.*
4. *Draw conclusions from the information.*

Chapter 2 is concerned with step 3, organizing raw data using tables, graphs, etc. (Chapter 3 will continue step 3 by summarizing the data using measures of central tendency and dispersion.)

Raw data—data that are not organized. The objective of chapter 2 is to learn techniques to organize data.

2.1 Organizing Qualitative Data

Remember: *Qualitative data* provide nonnumerical measures that categorize (or classify) individual observations.

Frequency distribution table—is developed using frequency and relative frequency (see example below).

Frequency—the number of observations in a given category.

Relative frequency (p. 56)—is the proportion or percent of observations within a category:

$$\text{Relative frequency} = \frac{\text{frequency}}{\text{sum of all frequencies}}$$

Birthplace of Presidents

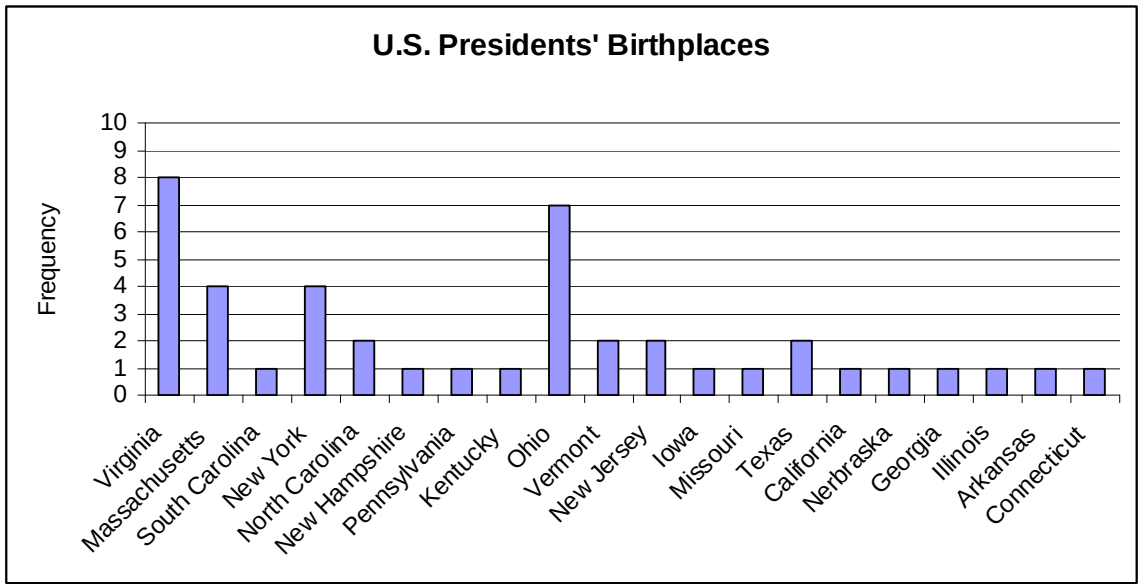
	Presidents	State of Birth
1	Washington	Virginia
2	J.Adams	Massachusetts
3	Jefferson	Virginia
4	Madison	Virginia
5	Monroe	Virginia
6	J.Q. Adams	Massachusetts
7	Jackson	South Carolina
8	Van Buren	New York
9	W. H. Harrison	Virginia
10	Tyler	Virginia
11	Polk	North Carolina
12	Taylor	Virginia
13	Fillmore	New York
14	Pierce	New Hampshire
15	Buchanan	Pennsylvania
16	Lincoln	Kentucky
17	A. Johnson	North Carolina
18	Grant	Ohio
19	Hayes	Ohio
20	Garfield	Ohio
21	Arthur	Vermont
22	Cleveland	New Jersey
23	B. Harrison	Ohio
24	Cleveland	New Jersey
25	McKinley	Ohio
26	T. Roosevelt	New York
27	Taft	Ohio
28	Wilson	Virginia
29	Harding	Ohio
30	Coolidge	Vermont
31	Hoover	Iowa
32	F. D. Roosevelt	New York
33	Truman	Missouri
34	Eisenhower	Texas
35	Kennedy	Massachusetts
36	L. B. Johnson	Texas
37	Nixon	California
38	Ford	Nebraska
39	Carter	Georgia
40	Reagan	Illinois
41	George H. Bush	Massachusetts
42	Clinton	Arkansas
43	George W. Bush	Connecticut

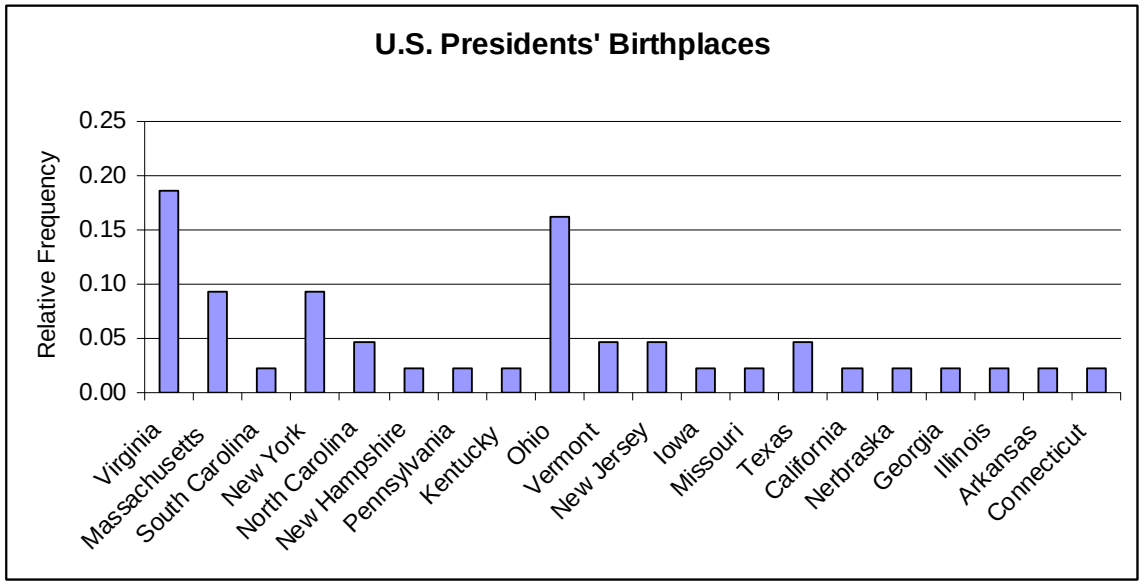
What does the data to the left reveal about the birthplace of Presidents?

Frequency Table

State	Tally	Frequency	Relative Frequency
Arkansas			
California			
Connecticut			
Georgia	|	1	0.023
Illinois	|	1	0.023
Iowa	|	1	0.023
Kentucky	|	1	0.023
Massachusetts	||||	4	0.093
Missouri	|	1	0.023
Nebraska	|	1	0.023
New Hampshire	|	1	0.023
New Jersey	||	2	0.047
New York	||||	4	0.093
North Carolina	||	2	0.047
Ohio	||||| ||	7	0.163
Pennsylvania	|	1	0.023
South Carolina	|	1	0.023
Texas	||	2	0.047
Vermont	||	2	0.047
Virginia	||||| |||	8	0.186
		43	1.000

Vertical bar graph—U.S. Presidents' Birthplaces



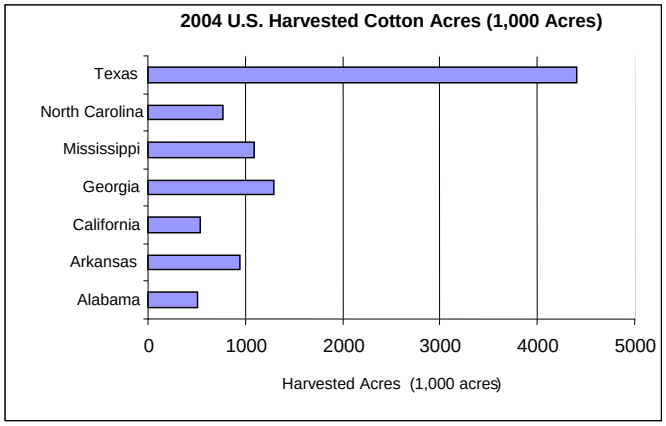


***Be sure to use Print/Preview to see that graph will print correctly. In most cases it will be necessary to change the page orientation to "Landscape" mode (more fits on a page).
To accomplish this: **File menu/Page Set-up/Page tab/Orientation (check Landscape**).

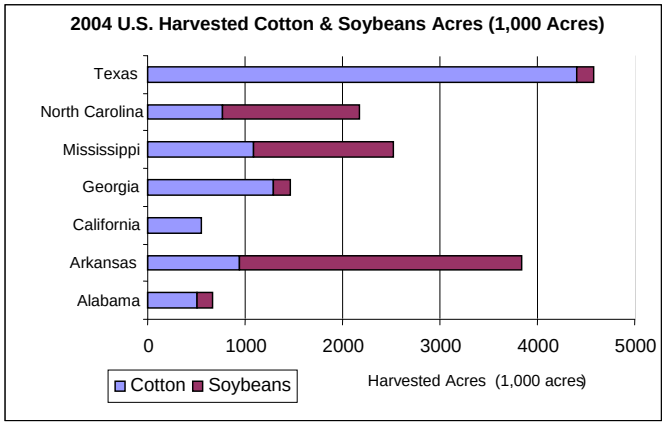
- **Bar graph**—is drawn from a frequency distribution table. There are several types of bar graphs:

  1) Vertical
     - Side-by-side (grouped)
     - Component-part
  2) Horizontal
     - Side-by-side (grouped)
     - Component-part

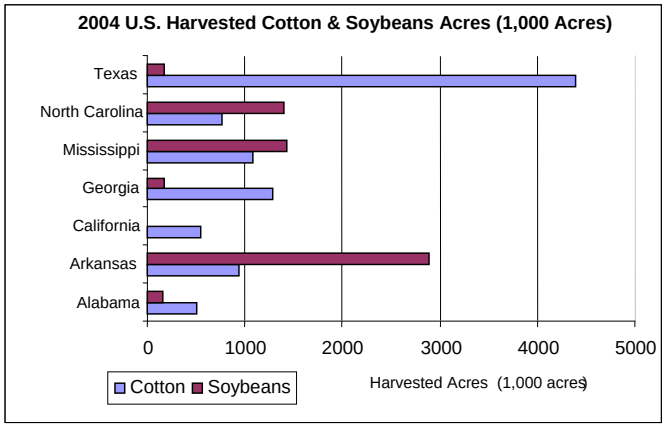
State	2004 Harvested Acres (1,000 acres) Cotton	Soybeans	Total
Alabama	510	160	670
Arkansas	945	2,890	3,835
California	545	0	545
Georgia	1,290	180	1,470
Mississippi	1,090	1,430	2,520
North Carolina	770	1,400	2,170
Texas	4,400	180	4,580



Horizontal Bar Graph



Component-Part Horizontal Bar Graph



Side-by-Side (Grouped) Horizontal Bar Graph

Side-by-Side Vertical Bar Graph—Educational Attainment (of Adults 25 years and older) in 1990 vs. 2003 (p. 59) (The data are in 1,000's)

Table 4

Educational Attainment	1990	2003
Less than 9th grade	16,502	12,276
9th to 12th grade, no diploma	22,842	16,323
High school diploma	47,643	59,292
Some college, no degree	29,780	31,762
Associate's degree	9,792	15,147
Bachelor's degree	20,833	33,213
Graduate or professional degree	11,478	17,169
Totals	158,870	185,182

Source: U.S. Census Bureau

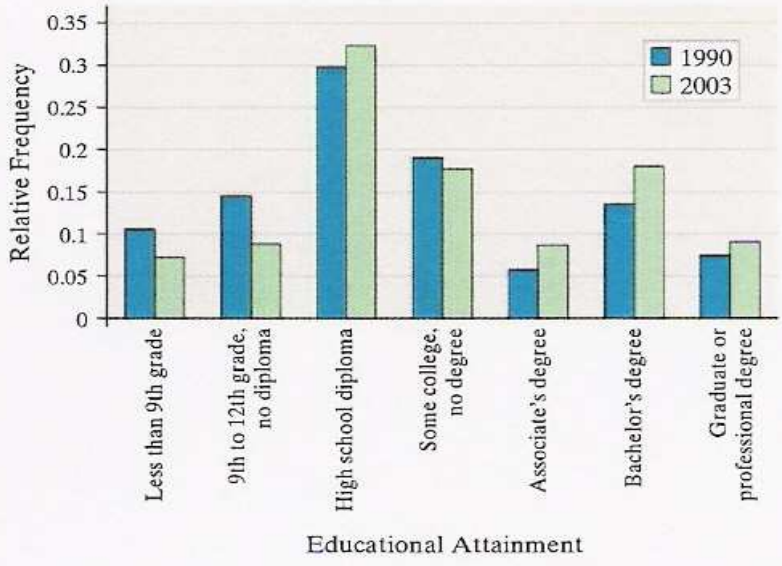
What can you learn about educational attainment from studying Table 4?

Table 5

Educational Attainment	1990	2003
Less than 9th grade	0.1039	0.0663
9th to 12th grade, no diploma	0.1438	0.0881
High school diploma	0.2999	0.3202
Some college, no degree	0.1874	0.1715
Associate's degree	0.0616	0.0818
Bachelor's degree	0.1311	0.1794
Graduate or professional degree	0.0722	0.0927

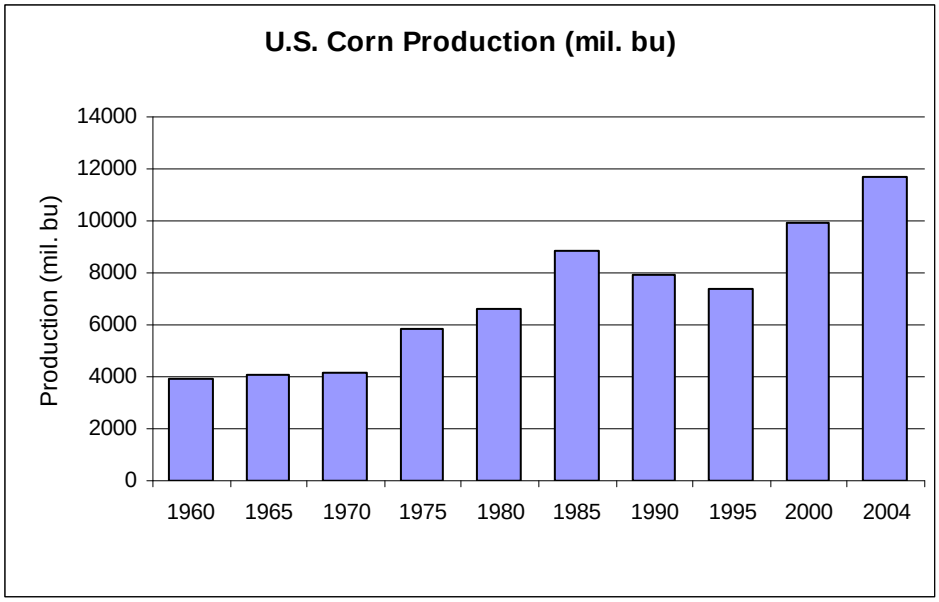
Solution: Table 5 shows the relative frequency for each category.

Figure 4 Educational Attainment in 1990 versus 2003



Notes about bar graphs:

1) Bars do not touch each other (p. 57)
2) Distance between the bars is approximately one-half the bar width (see Fig 1, p. 57)
3) Use vertical bars for temporal (time series) data
4) When comparing two or more data sets, use relative frequencies (see Ex #5, p. 59)



2.2—Organizing Quantitative Data and online 'E-book' 2.3—Additional Displays of Quantitative Data.

Objectives: (1) Construct a frequency distribution table
(2) Draw graphs from the frequency distribution table
- Histogram (p. 72)
- Frequency polygon
- Cumulative frequency polygon

These two items are discussed in the online 'E-book' version found in MYMATHLAB. [Go to Chapter Contents, Chapter 2, Chapter Opener, 2.3 begins on page 97.]

Definitions of terms for a frequency distribution table:

1) **Classes**—intervals of numbers.
2) **Lower class limit**—smallest value within the class.
3) **Upper class limit**—largest value within the class.
4) **Class width**—difference between consecutive lower (or upper) class limits.
5) **Class midpoint**—the sum of consecutive lower class limits divided by 2.
6) **Frequency**—the number of observations in a given class.
7) **Cumulative frequency**—sum of the frequencies for all classes up to and including the given class.
8) **Relative frequency**—is the proportion or percent of observations within a given class.

$$Relative\ frequency = \frac{frequency}{sum\ of\ all\ frequencies}$$

9) **Cumulative relative frequency**—sum of the relative frequencies for all classes up to and including the given class.

Example of a frequency distribution table:

- Data are for 3-Year Rate of Return of Mutual Funds from Table 12, p. 74, text.
- Frequency distribution table is shown in Tables 18 and 19, Additional Displays of Quantitative Data, pp. 98-99 (online 'E-book')

The construction of a frequency distribution table is demonstrated on the following page.

3-Year Rate of Return of Mutual Funds (Table 12, p. 74)

27.4	16.7	10.8	24.1	35.9
12.7	28.5	22.2	18.4	17.4
22.6	29.6	11.6	45.9	16.6
32.1	47.7	10.9	18.4	23.3
18.2	32.0	25.5	23.7	38.1
23.7	14.7	12.8	31.1	21.9
18.4	21.3	27.0	19.6	15.8
14.7	37.0	19.2	18.5	29.1

What can you understand about the rate of return of mutual funds from studying the data to the left?

Frequency Distribution Table, 3-Year Rate of Return of Mutual Funds

Class (3-Yr Rate of Return, %)	Class Midpoint	Tally	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
10.0 - 14.9	12.5					
15.0 - 19.9	17.5	||||| ||||| |	11			
20.0 - 24.9	22.5	||||| |||	8			
25.0 - 29.9	27.5	||||| |	6			
30.0 - 34.9	32.5	|||	3			
35.0 - 39.9	37.5	|||	3			
40.0 - 44.9	42.5		0			
45.0 - 49.9	47.5	||	2			
			40		1.00	

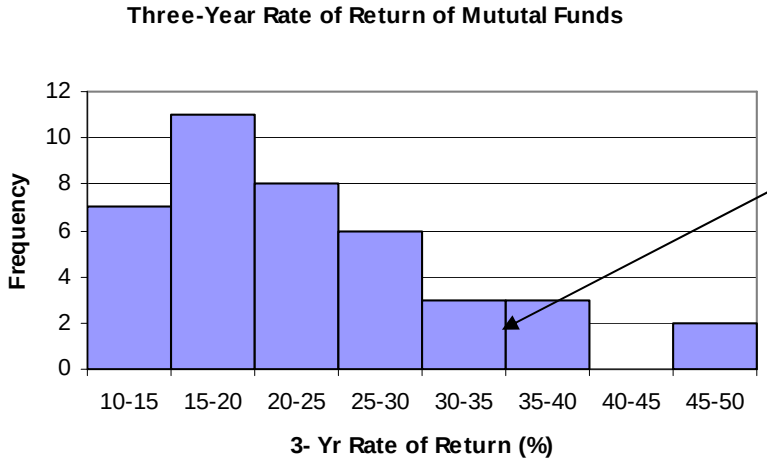
Notes about a frequency distribution table (p. 74-75):

1) The goal in constructing a frequency distribution table is <u>to reveal interesting features of the data</u>.
2) The choice of class width is arbitrary.
3) Typically the number of classes is between 5 and 20.
4) When the data set is small, use fewer classes; when the data set is large, use more classes.

Graphs drawn from a frequency distribution table:

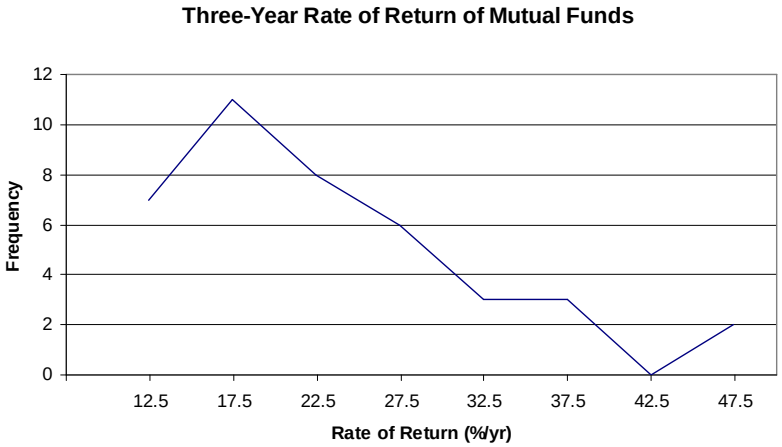
1) Histogram
2) Frequency polygon—Fig. 17, p. 98, Additional Displays of Quantitative Data, online 'E-book'.
3) Cumulative frequency polygon (or ogive)— Fig. 18, p. 99, Additional Displays of Quantitative Data, online 'E-book'.

Note: Graphs can be drawn with either <u>frequency</u> or <u>relative</u> frequency on the vertical (Y) axis.

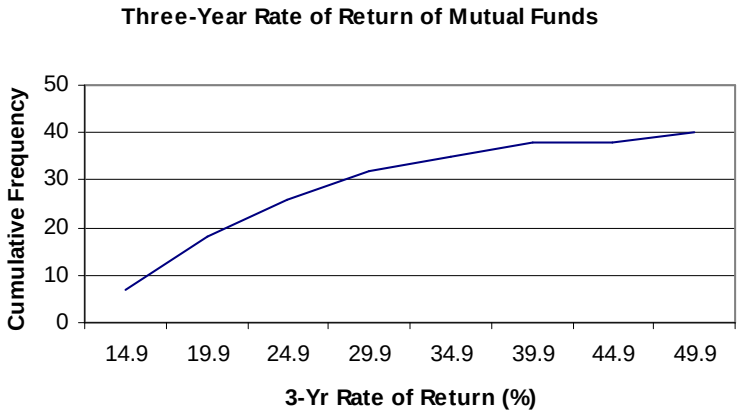


Histogram

An Excel bar graph has space between the bars as the default. To remove the space, right click on a bar and select **Format Data Series** and this will show the **Format Data Series** window. Click **Options** and set **Gap width**=0.



Frequency Polygon

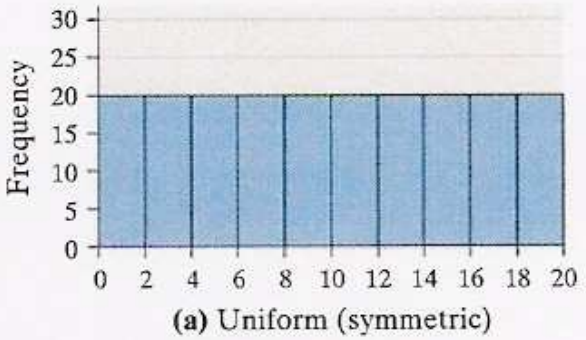


Cumulative Frequency Polygon (or Ogive)

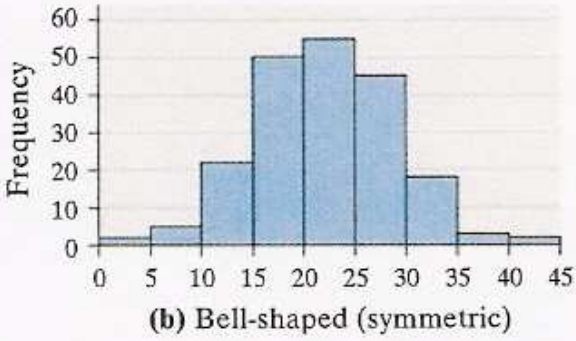
Distribution Shapes (p. 80):

A distribution comes from a *frequency distribution table*. The word distribution refers to the distribution of observations into classes. Some examples of distribution shapes are shown below.

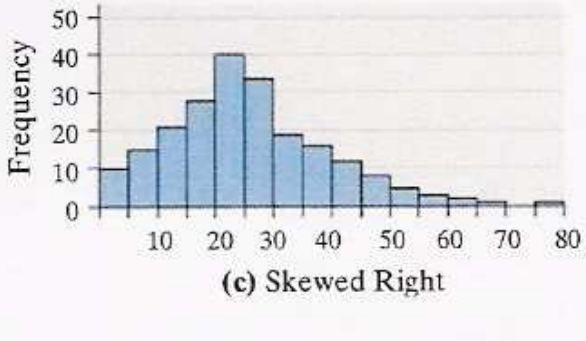
Figure 15



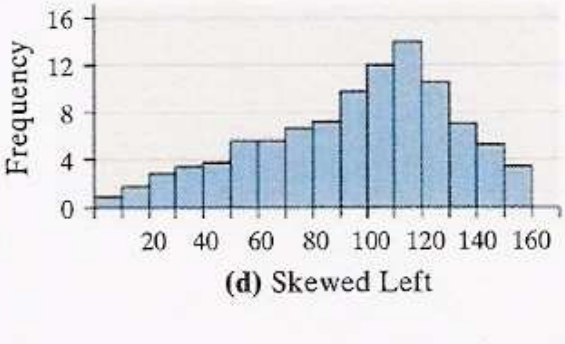
(a) Uniform (symmetric)



(b) Bell-shaped (symmetric)



(c) Skewed Right



(d) Skewed Left

EXAMPLE—Construction a Stem-and-Leaf Diagram

The employment ratio is the number of employed to population ratio. It is found by dividing the number of employed individuals in a population by the size of the population. The following data represent the employment ratio by state in the United States for 1999. Construct a stem-and-leaf diagram.

State	Employment Ratio (%)	State	Employment Ratio (%)	State	Employment Ratio (%)
Alabama	60.3	Kentucky	61.5	North Dakota	68.1
Alaska	68.8	Louisiana	59.4	Ohio	64.0
Arizona	63.3	Maine	65.1	Oklahoma	62.9
Arkansas	60.0	Maryland	67.3	Oregon	64.3
California	62.8	Mass.	66.5	Pennsylvania	61.6
Colorado	71.4	Michigan	66.0	Rhode Island	64.4
Connecticut	65.4	Minnesota	73.0	South Carolina	62.7
Delaware	64.7	Mississippi	58.0	South Dakota	71.1
Dist. Of Col.	63.4	Missouri	66.4	Tennessee	63.6
Florida	60.1	Montana	65.6	Texas	65.6
Georgia	66.8	Nebraska	71.0	Utah	69.6
Hawaii	63.2	Nevada	66.0	Vermont	69.9
Idaho	66.1	New Hamp.	70.3	Virginia	65.6
Illinois	66.7	New Jersey	64.1	Washington	66.9
Indiana	66.2	New Mexico	58.5	West Virginia	52.7
Iowa	70.1	New York	59.7	Wisconsin	70.1
Kansas	70.0	North Carolina	65.1	Wyoming	67.8

We let the **stem represent the integer portion** of the number and the **leaf will be the decimal portion**. For example, the stem of Alabama will be 60 and the leaf will be 3.

52 7
53
54
55
56
57
58 0 5
59 4 7
60 3 0 1
61 5 6
62 8 9 7
63 3 4 2 6
64 7 1 0 3 4
65 4 1 6 1 6 6
66 8 1 7 2 5 0 4 0 9
67 3 8
68 8 1
69 6 9
70 1 0 3 1
71 4 0 1
72
73 0

Advantage of Stem-and-Leaf Diagrams over Histograms
The raw data can be retrieved from the stem-and-leaf plot. By comparison, once a frequency distribution or histogram of continuous data is created, the raw data is lost (unless reported with the frequency distribution).

The data below are for **cottonseed oil prices** at Valley Points (cents/lb) for the crop years 1980/81 through 2004/05.

Year beginning October 1	Price, Avg. Valley Points 1/ (cents /lb)	Rounded to integer (cents/lb)
1980/81	25.86	26
1981/82	20.10	20
1982/83	21.80	22
1983/84	32.80	33
1984/85	29.20	29
1985/86	16.91	17
1986/87	17.67	18
1987/88	21.67	22
1988/89	19.71	20
1989/90	23.30	23
1990/91	22.30	22
1991/92	20.10	20
1992/93	30.07	30
1993/94	30.30	30
1994/95	29.23	29
1995/96	26.53	27
1996/97	25.58	26
1997/98	28.84	29
1998/99	27.32	27
1999/00	21.56	22
2000/01	15.98	16
2001/02	17.98	18
2002/03	37.75	38
2003/04	32.00	32
2004/05	24.00	24

Stem-and-Leaf Graph--Price of Cottonseed Oil at Valley Points (cents/lb)

1	10-19	7	8	6	8										
2	20-29	6	0	2	9	2	0	3	2	0	9	7	2	4	
3	30-39	3	0	0	8	2									

Stem-and-Leaf Graph--Price of Cottonseed Oil at Valley Points (cents/lb)

1	10-14									
1	15-19	7	8	6	8					
2	20-24	0	2	2	0	3	2	0	2	4
2	25-29	6	9	9	7	6	9	7		
3	30-34	3	0	0	2					
3	35-39	8								

Split stems, see p. 79 of text.

1/ PBSY, basis Greenwood, MS, beginning 1992.

Sources: The Bureau of the Census and Agricultural Marketing Service, USDA.

Technology Insert[i]

How to execute concepts using Microsoft® Excel

Birthplace of Presidents

	Presidents	State of Birth
A	B	C
3	Washington	Virginia
4	J.Adams	Massachusetts
5	Jefferson	Virginia
6	Madison	Virginia
7	Monroe	Virginia
8	J.Q. Adams	Massachusetts
9	Jackson	South Carolina
10	Van Buren	New York
11	W. H. Harrison	Virginia
12	Tyler	Virginia
13	Polk	North Carolina
14	Taylor	Virginia
15	Fillmore	New York
16	Pierce	New Hampshire
17	Buchanan	Pennsylvania
18	Lincoln	Kentucky
19	A. Johnson	North Carolina
20	Grant	Ohio
21	Hayes	Ohio
22	Garfield	Ohio
23	Arthur	Vermont
24	Cleveland	New Jersey
25	B. Harrison	Ohio
41	Carter	Georgia
42	Reagan	Illinois
43	George H. Bush	Massachusetts
44	Clinton	Arkansas
45	George W. Bush	Connecticut

Frequency Table:

	State	Frequency	Relative Frequency
D	E	F	G
	Virginia	8	0.1860
	Massachusetts	4	0.0930
	South Carolina	1	0.0233
	New York	4	0.0930
	North Carolina	2	0.0465
	New Hampshire	1	0.0233
	Pennsylvania	1	0.0233
	Kentucky	1	0.0233
	Ohio	7	0.1628
	Vermont	2	0.0465
	New Jersey	2	0.0465
	Iowa	1	0.0233
	Missouri	1	0.0233
	Texas	2	0.0465
	California	1	0.0233
	Nerbraska	1	0.0233
	Georgia	1	0.0233
	Illinois	1	0.0233
	Arkansas	1	0.0233
	Connecticut	1	0.0233
		43	

Steps:

1. Enter the data in the tabular format given on the left.

2. To create a Frequency Table as shown above:

a. Filter the unique states:
From menubar--select "**Data/Filter/Advance Filter"**
and this opens the "**Advance Filter**" window as shown on the left.

b. Select "**Copy to another location**" radio button.

c. Select the range of "**State of Birth**" as "**List range"**

d. "**Copy to**" will be a cell where you want to create the Frequency Table.

e. Select option "**Unique records only**."

f. Press **OK** and that will filter the unique records and paste them in cells "**E3:E22**"

3. To find the frequencies in cells "F3:F22":

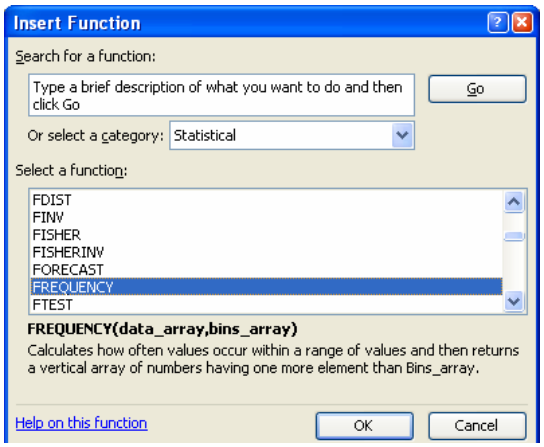
a. Enter the following formula in cell "**F3**"
=COUNTIF(range,criteria)
where range includes cells containing the States of Birth; and criteria is a particular State of Birth.
=COUNTIF(C$3:C$45,E3)

b. Copy this formula (from "**F3**") into cells "**F4:F22**"

Filling the Frequency Column by Counting Frequencies.

3-Year Rate of Return of Mutual Funds (Table 12, p. 74)

	C	D	E	F	G
5	27.4	16.7	10.8	24.1	35.9
6	12.7	28.5	22.2	18.4	17.4
7	22.6	29.6	11.6	45.9	16.6
8	32.1	47.7	10.9	18.4	23.3
9	18.2	32.0	25.5	23.7	38.1
10	23.7	14.7	12.8	31.1	21.9
11	18.4	21.3	27.0	19.6	15.8
12	14.7	37.0	19.2	18.5	29.1

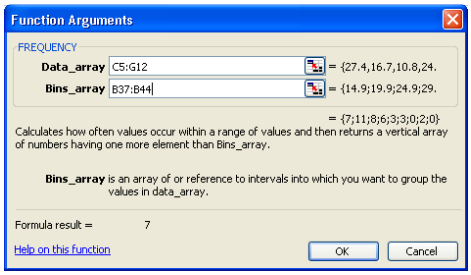


Two ways to count frequencies into the frequency column in a frequency distribution table:

1. Using the Statistical function FREQUENCY:

1. Create headings for your frequency table by typing "Bins" in **B36** and "Frequency" in **C36**.
2. Type the **Bins_array** in **B37:B44** (the Bins_array includes the upper class limits).
3. Highlight the "**Frequency Column,**" cells **"C37:C44."**
4. From the **Excel Menubar** select **Insert** ---> **Function** and this will show the **Insert Function** window.
5. From the **Insert Function** window ---> **Or select a category**, select **Statistical**.
6. From the **Select a function** list, select **FREQUENCY** and press **OK and the Function Arguments** window will appear.
7. From the **Function Arguments** window**,**
   a) Press the button on the right of the **Data_array** textbox and highlight cells **C5:G12.**
   b) Press the button on the right of the **Bins_array** textbox and highlight the **Bins** column cells **B37:B44**. (Note: The **Bins_array** in Excel is called "Upper Class Limits" in a frequency distribution table.)
8. Press **Ctrl+Shift+Enter** together and the frequency values will appear in cells "**C37:C44."**

The example below illustrates the process of counting frequencies using the Statistical function FREQUENCY.

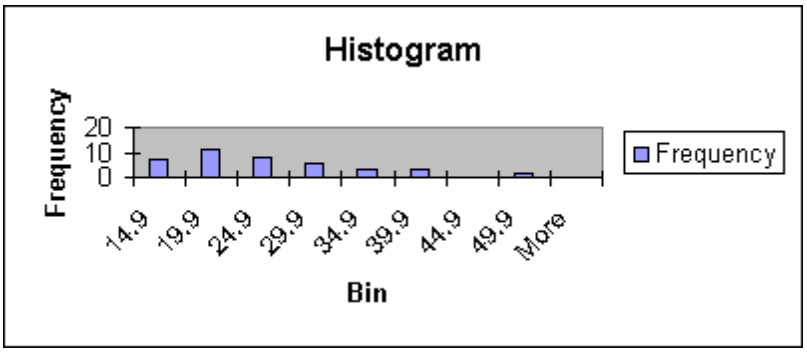


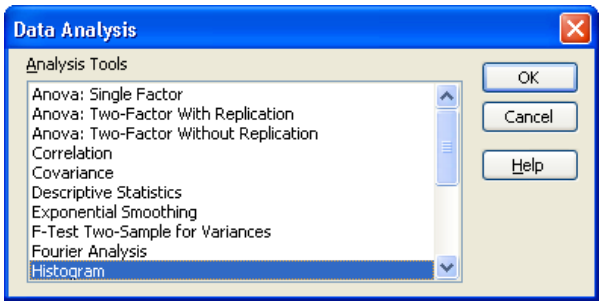
	B	C	D
36	Bins	Frequency	Excel Formula
37	14.9	7	{=FREQUENCY(C5:G12,B37:B44)}
38	19.9	8	{=FREQUENCY(C5:G12,B37:B44)}
39	24.9	6	{=FREQUENCY(C5:G12,B37:B44)}
40	29.9	5	{=FREQUENCY(C5:G12,B37:B44)}
41	34.9	3	{=FREQUENCY(C5:G12,B37:B44)}
42	39.9	1	{=FREQUENCY(C5:G12,B37:B44)}
43	44.9	0	{=FREQUENCY(C5:G12,B37:B44)}
44	49.9	2	{=FREQUENCY(C5:G12,B37:B44)}

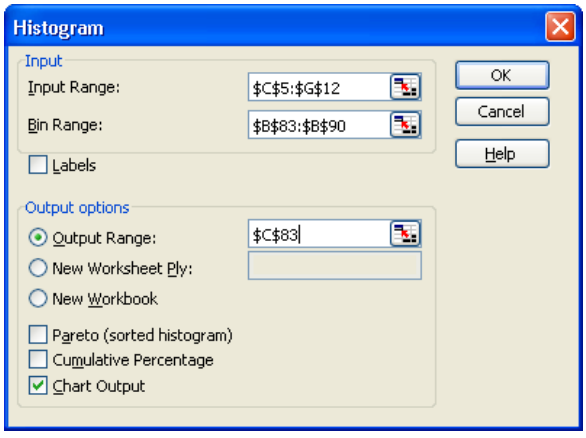
2. Using the Histogram Analysis Tool:

1. Create a heading for the **Bins_array** by typing “Bins” in **B82**.
2. From the **Excel Menubar** select **Tools** ---> **Data Analysis** and the **Data Analysis** window will appear.
3. From the **Analysis Tools** list select **Histogram** and the **Histogram** window will appear.
4. Press the button on the right of the **Input Range** textbox and highlight cells **C5:G12** (on previous page).
5. Press the button on the right of the **Bin Range** textbox and highlight cells **B83:B90** (below left).
6. Under **Output options**,
   a) Click the **Output Range** radio button. Press the button on the right of the **Output Range** textbox and highlight cell **C82**.
   b) Check **Chart Output** if you would like a histogram drawn.
7. Click **OK** and a frequency distribution and histogram will appear (see below). Right-click on the histogram chart area to make changes in the chart.

	B	C	D
82	**Bins**	***Bin***	***Frequency***
83	14.9	14.9	7
84	19.9	19.9	11
85	24.9	24.9	8
86	29.9	29.9	6
87	34.9	34.9	3
88	39.9	39.9	3
89	44.9	44.9	0
90	49.9	49.9	2
		More	0







[i] The concepts presented in the Technology Insert may be formally demonstrated as time permits; but this supplement is not intended to be a focus of this course.