

# Exploring Data: Interpreting Graphical Displays of Distributions of Univariate Data

A **frequency distribution** is a listing that pairs each value of a variable with its frequency. Frequency distributions in table form are useful but do not give the viewer a feel of what patterns might exist. Graphical representations of the data provide a better picture of the distribution. No one best choice exists when it comes to a graphical display. The most common types of graphical displays include dotplots and bar charts, stemplots, histograms, and cumulative frequency plots.

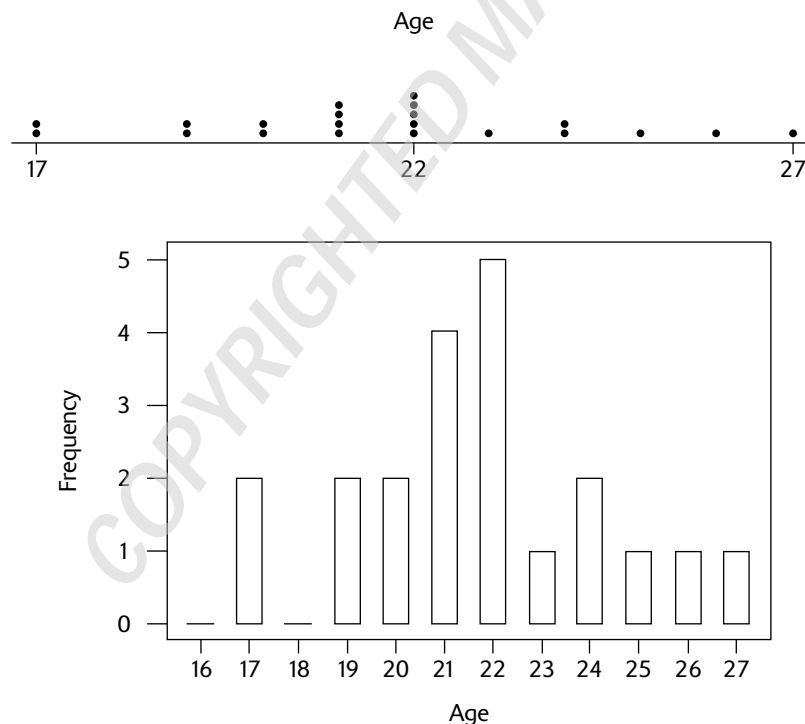
## Dotplots and Bar Charts

A **dotplot** represents each piece of data as a dot positioned along a scale or axis. The scale can be either horizontal or vertical. The horizontal position is more common. The frequency is represented by the other axis.

### EXAMPLE:

The following represent the ages of 21 club members. Describe the distribution in general terms.

22 24 19 17 20 27 24 23 26 17 19  
22 25 21 21 22 22 21 21 20 22

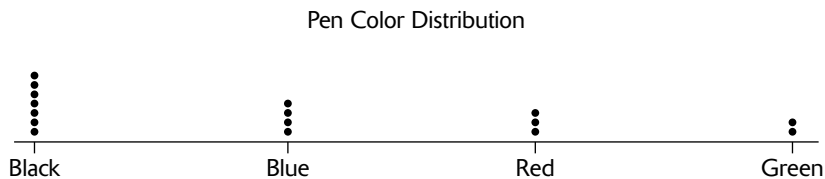


The data is mound-shaped and fairly symmetric. Both the dotplot and the **bar chart** convey the same information. The relative frequency of the data is shown by the height of the dots or the height of the bar.

Dotplots and bar charts are useful in depicting categorical or qualitative data. Each column (row) of dots in a dotplot or each column in a bar chart is used to represent a category of data.

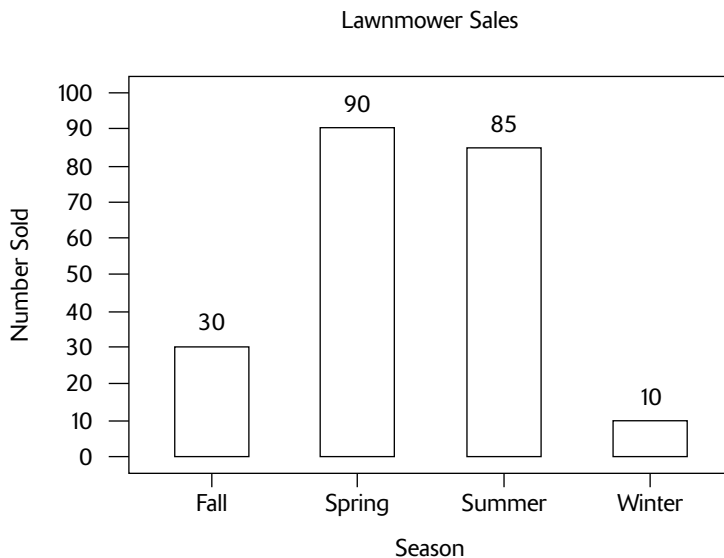
**EXAMPLE:**

Bryan found 16 pens in his desk: 7 black, 4 blue, 3 red, and 2 green. The following dotplot displays this data.



**EXAMPLE:**

Lawnmower sales vary depending on the season. A garden supply store sold 90 lawnmowers during the spring, 85 during the summer, 30 during the fall, and 10 during the winter. Display this information in a bar chart.



Listing the value at the top of each bar can be useful when the scale is spread out and the actual value is difficult to judge.

## Stemplots

---

**Stemplots**, sometimes known as **stem and leaf plots**, use digits to represent the data. Leading digits are used to form the stem, and trailing digits are used to form the leaves. What to use as the stem and what to use as the leaf should be determined by the data in question. The following example shows two different configurations.

**EXAMPLE:**

Construct a stemplot to display the following quiz scores:

58	62	62	63	65	65	65	68
69	72	72	75	76	78	79	81
84	84	85	92	94	95	98	

Grouping by 10:

**Quiz Scores**

```

5 | 8
6 | 22355589
7 | 225689
8 | 1445
9 | 2458
    
```

Grouping by 5:

**Quiz Scores**

```

5 |
5 | 8
6 | 223
6 | 55589
7 | 22
7 | 5689
8 | 144
8 | 5
9 | 24
9 | 58
    
```

Both of the plots are useful in showing the distributions of the data.

Placing two stemplots back-to-back can be useful in comparing two distributions.

**EXAMPLE:**

Compare the following two sets of test scores using back-to-back stemplots:

Rudy’s scores on 20 tests were 39, 40, 43, 44, 47, 49, 53, 55, 55, 56, 58, 59, 62, 62, 63, 65, 75, 78, 87, and 88.

Sonya’s scores on the same 20 tests were 46, 47, 47, 48, 54, 55, 56, 58, 62, 62, 63, 65, 65, 68, 68, 69, 75, 75, 76, and 87.

<i>Rudy</i>		<i>Sonya</i>
9	3	
97430	4	6778
986553	5	4568
5322	6	22355889
85	7	556
87	8	7

## Histograms

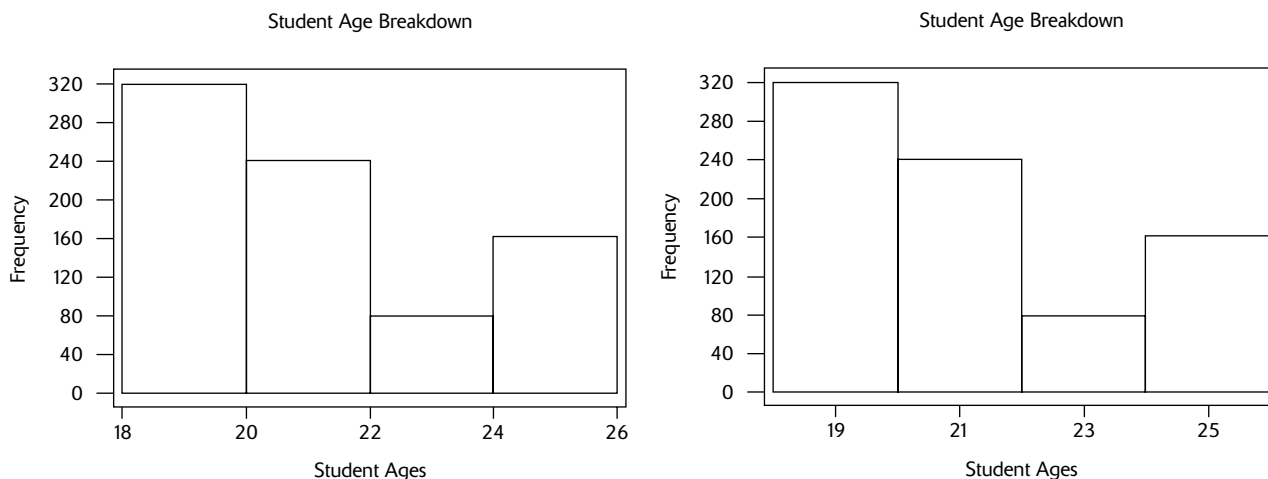
Bar charts represent categorical or qualitative data. Histograms are bar charts that represent a frequency distribution of a variable that is quantitative. The horizontal scale represents values of the variable and is labeled with class boundaries or class midpoints. The vertical scale represents the frequency (quantity) or relative frequency (percentage) of the values in each class. Bars in a histogram touch each other. Although histograms can be constructed from given data, the exam does not stress this technique. Questions on the exam use histograms to display data and ask test takers to interpret the histograms.

Important items concerning histograms:

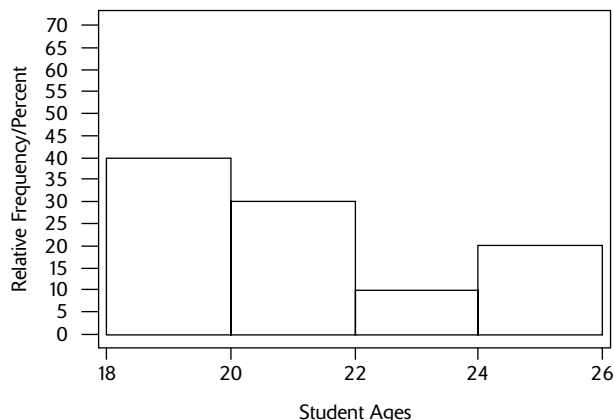
- Bar charts are used for categorical, qualitative data. Histograms are used for quantitative data.
- Classes, or bars, are of equal width and touch each other.
- The number of classes depends on the quantity of data and what you are trying to display.
- If a data element falls on a class boundary, it belongs to the class on the right. If the boundaries of a class are 20 and 30, and  $x$  is a data element, then  $20 \leq x < 30$ .
- All data elements must fit into a specific class. Do not extend the histogram far beyond data values.

**EXAMPLE:**

The following is a summary of results from a survey concerning the ages of 800 students at a private university: 320 were between 18 and 20 years old ( $18 \leq x < 20$ ); 240 were between 20 and 22 years old ( $20 \leq x < 22$ ); 80 were between 22 and 24 years old ( $22 \leq x < 24$ ); and 160 were between 24 and 26 years old ( $24 \leq x < 26$ ). The following frequency histograms represent this data, one using class boundaries (or cut points) and the other using class midpoints.









The data could also be represented by a relative frequency histogram. First calculate the relative frequency for each class:  $320/800 = .4$ ,  $240/800 = .3$ ,  $80/800 = .1$ ,  $160/800 = .2$ . Notice that the shapes of the two histograms are the same. The only difference is in the vertical scale.

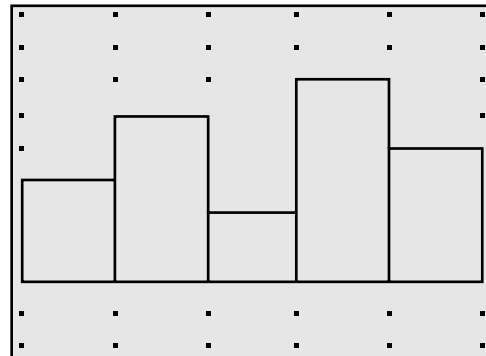


The TI-83/4 can display histograms. In the Window settings, Xscl refers to the class width. Xmin and Xmax are set to the left and right side boundaries of the histogram. Set Ymin and Ymax so that the entire graph is visible. The following screen shots of a TI-83/4 show the steps required to plot a histogram of the following data values: 7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 10, 10, 10, 10, 10, 10, 11, 11, 11, 11, 11, 12, 12, and 13.

L1	L2	L3	1
7	-----	-----	
7			
7			
8			
8			
8			
8			
L1(1)=7			

WINDOW  
 Xmin=7  
 Xmax=12  
 Xscl=1  
 Ymin=-2  
 Ymax=8  
 Yscl=1  
 Xres=1 ■







Plot1 Plot2 Plot3  
 On Off  
 Type:     
    
 Xlist: L1  
 Freq: 1 ■

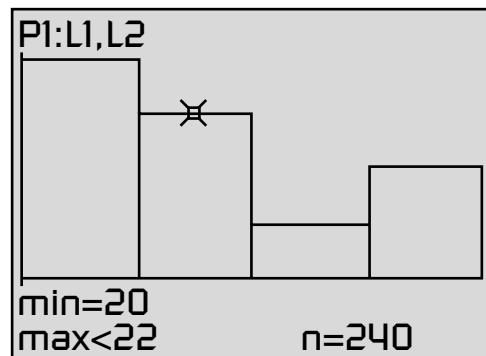


The TI-83/4 can also display a histogram given a frequency distribution table, such as the data from the preceding example. In this case, enter the midpoints of the classes into L1 and the frequency for each class into L2. In the Window settings, set Xmin to 18, Xmax to 26, Xscl to 2. Set Ymin and Ymax to show the entire graph. Some room can be left at the bottom and top to display TRACE information. Enter STATPLOT and set Xlist to L1 and Freq to L2. The following screenshots demonstrate this procedure.

L1	L2	L3	2
19	320	-----	
21	240		
23	80		
25	160		
-----	-----		
L2(5) =			

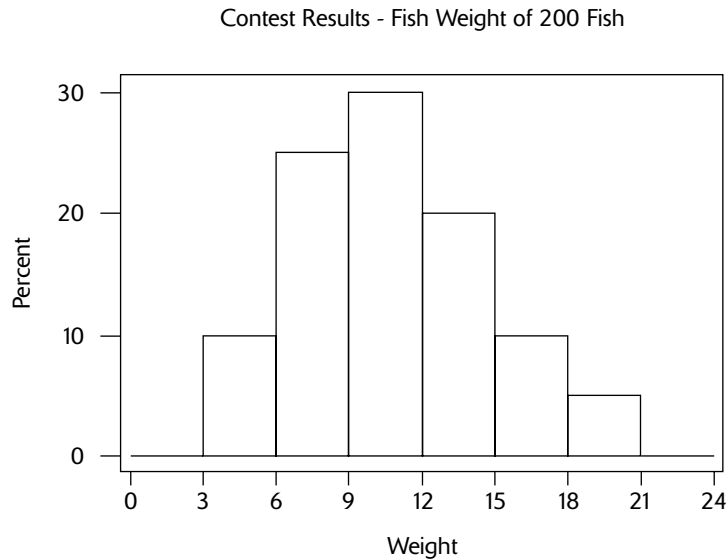
WINDOW  
 Xmin=18  
 Xmax=26  
 Xscl=2  
 Ymin=-100  
 Ymax=380  
 Yscl=1  
 Xres=1 ■

Plot1 Plot2 Plot3  
 On Off  
 Type:     
    
 Xlist: L1  
 Freq: L2 ■



**EXAMPLE:**

Based on the following histogram, what could be said about the weights of the 200 fish caught during the contest?



The histogram shows the following:

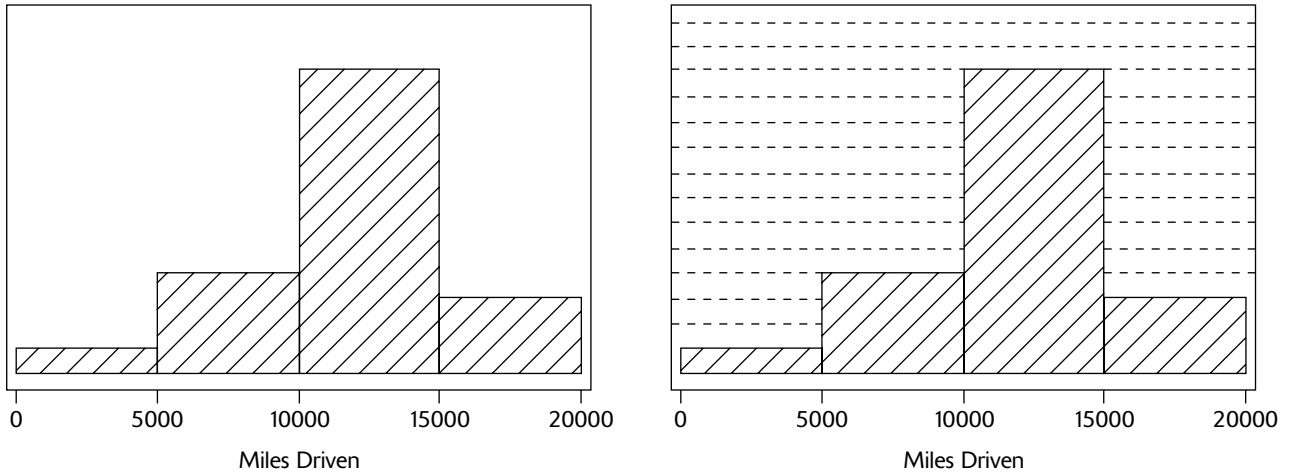
0-3	0	0%
3-6	20	10%
6-9	50	25%
9-12	60	30%
12-15	40	20%
15-18	20	10%
18-21	10	5%
21-24	0	0%

Observations could include the following: fish weights were between 3 and 21 pounds; 60% of the fish weighed between 9 and 18 pounds; 35% of the fish weighed more than 12 pounds; 110 fish weighed between 6 and 12 pounds.

**EXAMPLE:**

Since the heights of the columns of a histogram represent the frequency of each class, you can determine the relative frequency of classes even if the frequency is not known. Divide each column (class) into approximately equal sized rectangles. Count the total number of rectangles. The relative frequency can be approximated by dividing each column total by the grand total.

The following histograms show the number of miles driven annually by the employees of a small company:



To determine the relative frequency, divide as follows:  $1/20 = 5\%$ ,  $4/20 = 20\%$ ,  $12/20 = 60\%$  and  $3/20 = 15\%$ . Therefore, 75% of the employees drive more than 10000 miles annually, and 20% drive between 5000 and 10000 miles annually.

## Cumulative Frequency Charts

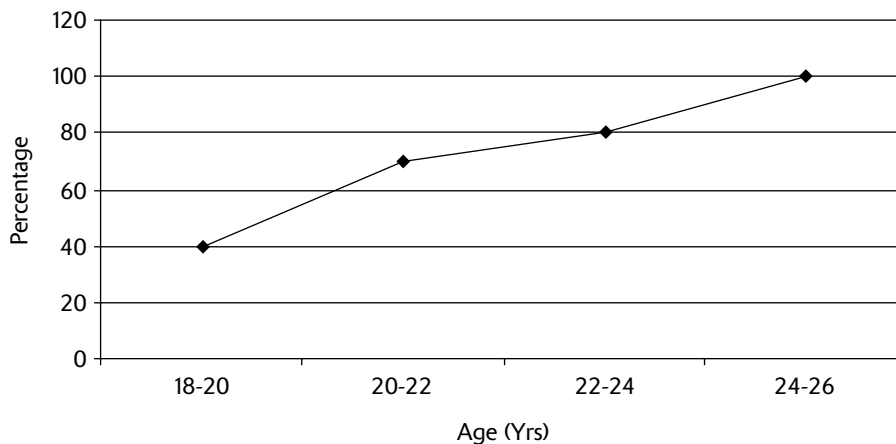
A **cumulative frequency chart**, often known as an *ogive*, can represent either total frequency or total relative frequency.

**EXAMPLE:**

The following chart shows the relative frequencies of ages of freshman at a private university. The cumulative frequency chart can be drawn using this data.

<b>Age Range</b>	<b>Percentage of Enrollment</b>	<b>Cumulative Percentage</b>
18–20	40	40
20–22	30	70
22–24	10	80
24–26	20	100

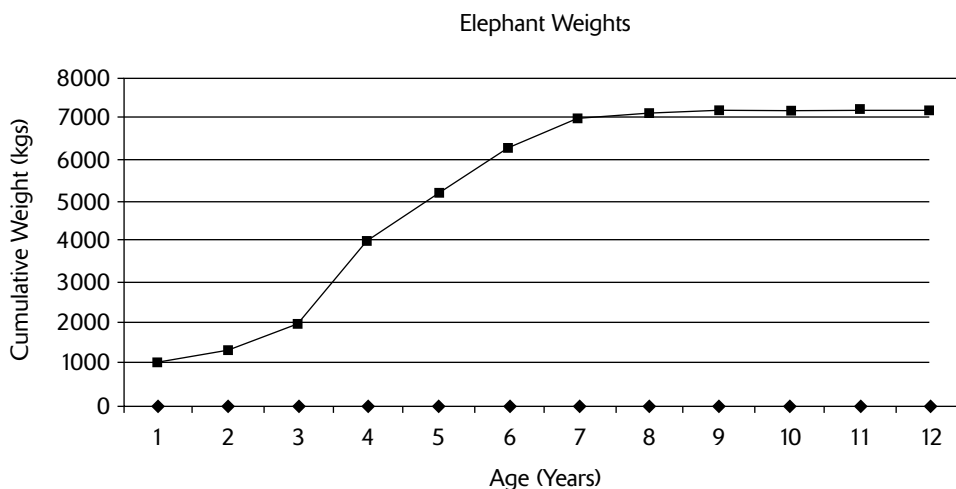
Cumulative Percentage of Enrollment



**EXAMPLE:**

Cumulative frequency charts can be used to show how fast values are changing over one interval compared to another interval. For example, this cumulative frequency chart shows that this elephant’s most rapid growth occurred between year 3 and 4, and most growth occurred before year 7.

Age (Yrs)	1	2	3	4	5	6	7	8	9	10	11	12
Weight Gain (kg)	1000	300	700	2000	1200	1100	700	100	60	40	20	10
Cumulative Weight (kg)	1000	1300	2000	4000	5200	6300	7000	7100	7160	7200	7220	7230



Cumulative line graphs can be difficult to read. The shape of a cumulative line graph can show whether the distribution is shifted to the left, shifted to the right, or is symmetric. See the discussion that follows on “Shape” and the example within that section for a comparison of shifted distributions and how they affect a cumulative line graph.

## Center and Spread

The center and spread of a distribution can be defined as follows:

*Center*—The point in the distribution where approximately half of the values (area) lie to the left and approximately half of the values (area) lie to the right.

*Spread*—Sometimes known as the range, the spread extends from the minimum value to the maximum value in a distribution.

**EXAMPLE:**

What is the center and spread of the following data?

- 58 62 62 63 65 65 65 68
- 69 72 72 75 76 78 79 81
- 84 84 85 92 94 95 98

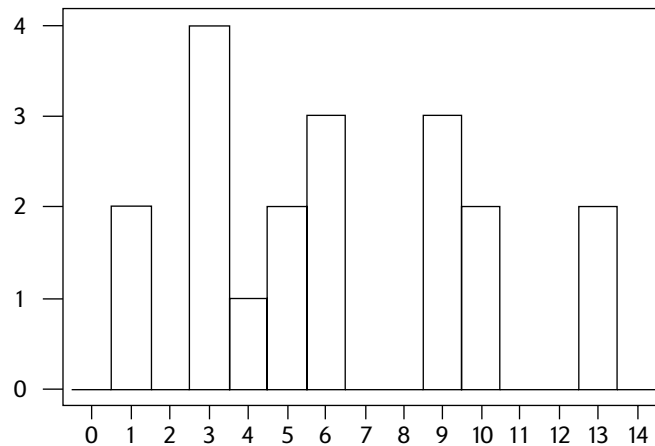
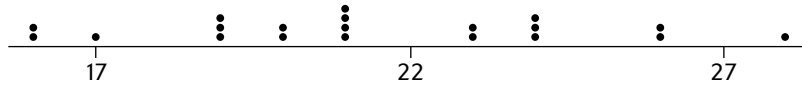
The center is 75, and the spread is from 58 to 98.

## Clusters and Gaps

Some distributions are concentrated about several values with space between these concentrations. These concentrations are called *clusters*, and the spaces between them are called *gaps*. Note that the gaps contain no members of the distribution.

### EXAMPLE:

These two distributions contain clusters and gaps:

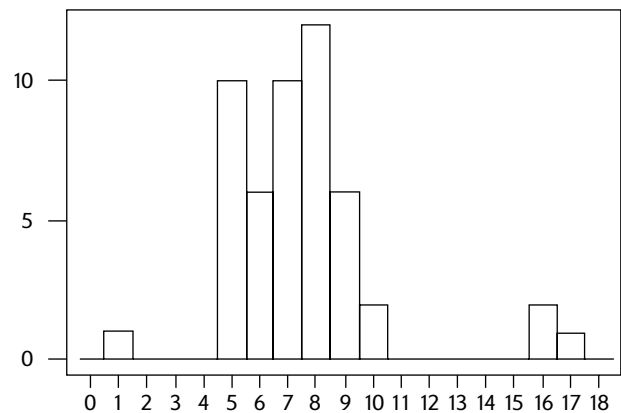
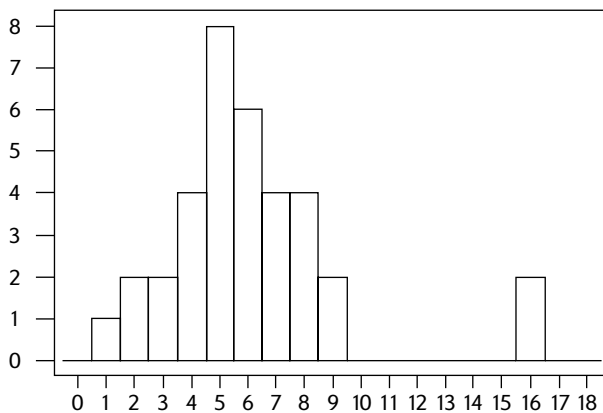


## Outliers and Other Unusual Features

An **outlier** is a data entry that is far removed from the other data entries. Outliers can have an adverse effect on some statistical measures and, therefore, must be treated with care. The decision to include or exclude an outlier should be considered carefully. (A more formal definition will be discussed in the next chapter.)

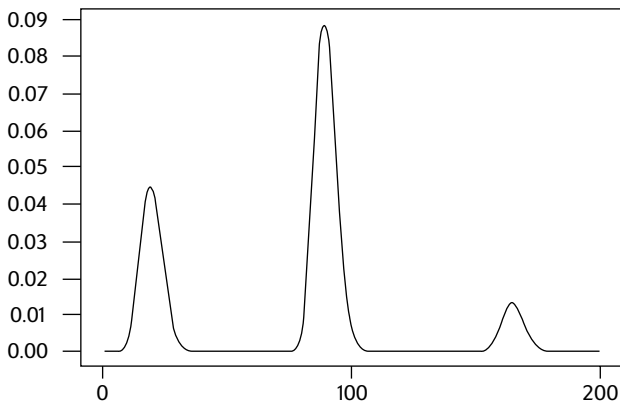
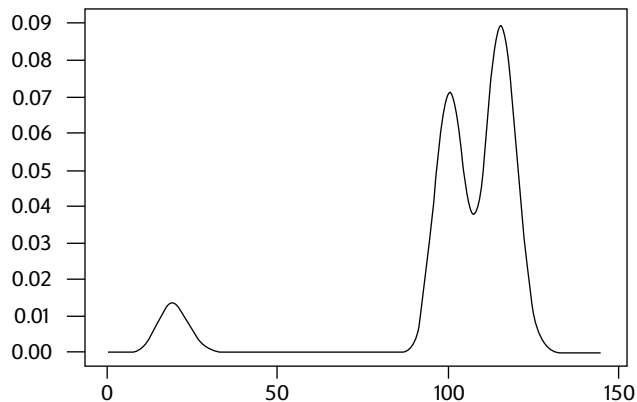
### EXAMPLE:

The following histograms contain outliers. Note that outliers may be on one side of the distribution or on both sides of the distribution.



**EXAMPLE:**

The following continuous distributions contain outliers:



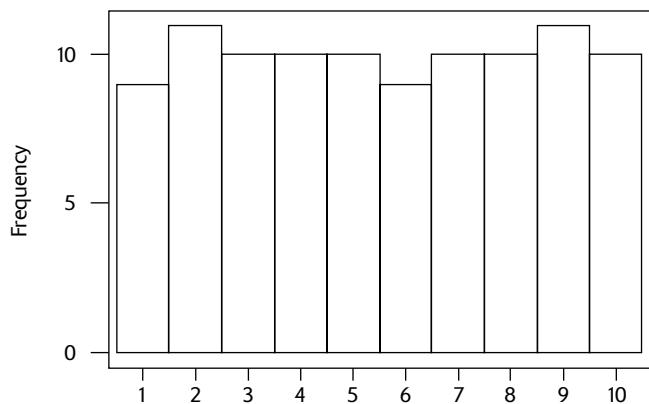
## Shape

One very important characteristic of a distribution is its shape. Distributions made up of a single mound can be classified as **symmetric**, **skewed left**, or **skewed right**. If the values of the distribution are evenly distributed, it is called a uniform distribution.

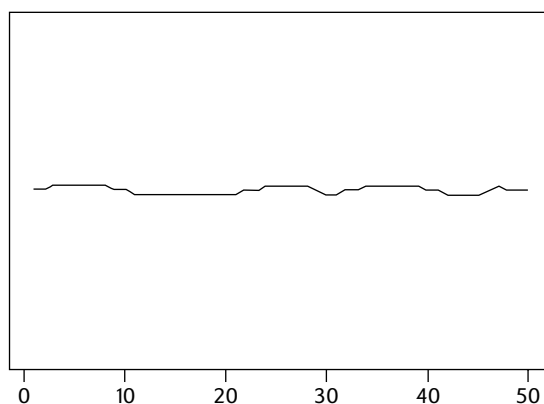
**EXAMPLE:**

Each of these distributions is made up of a single mound. Note that the *skew* is the *tail* of the distribution. The uniform distributions are uniform or nearly uniform. The symmetric distributions are symmetric or nearly symmetric.

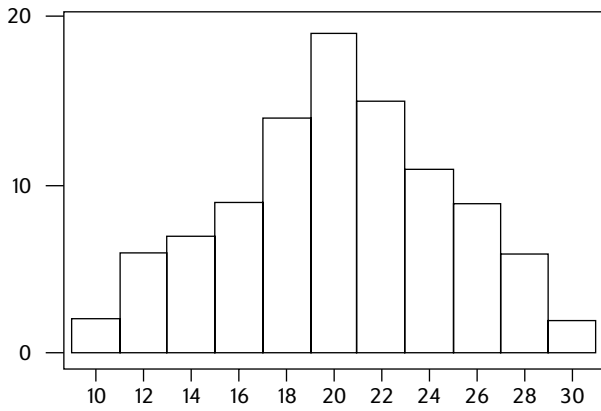
Uniform Distribution



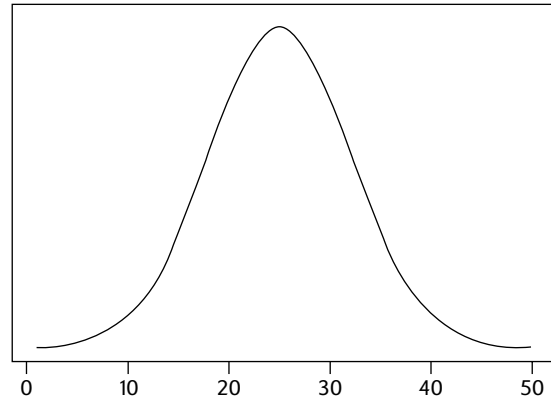
Uniform Distribution



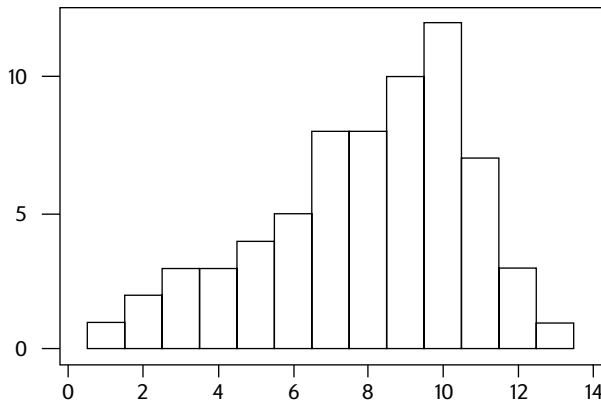
Symmetric Distribution



Symmetric Distribution



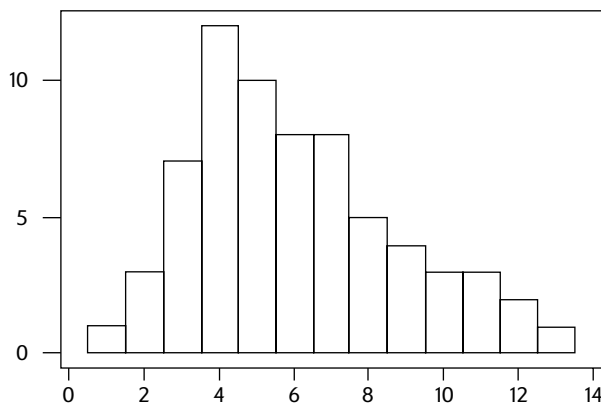
Skewed Left Distribution



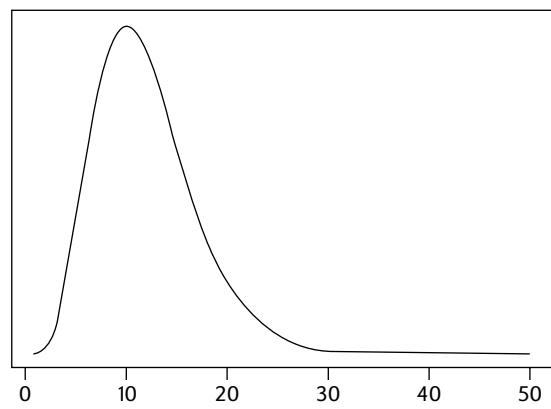
Skewed Left Distribution



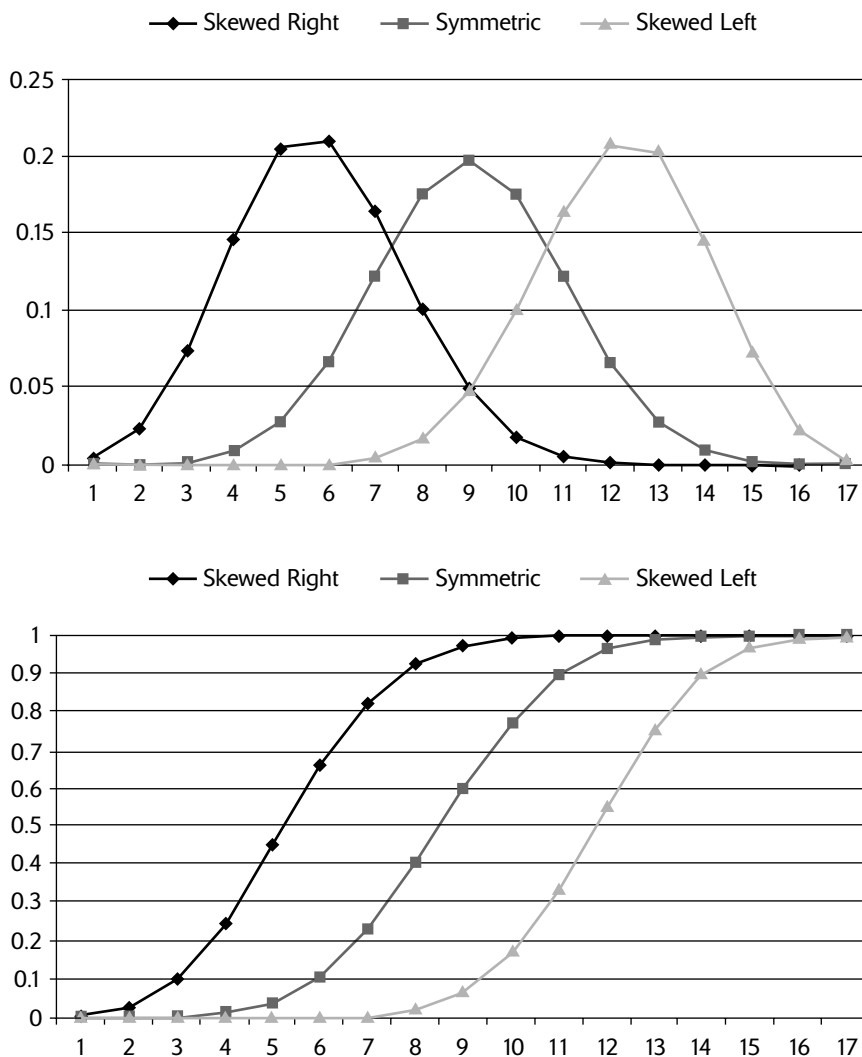
Skewed Right Distribution



Skewed Right Distribution



The relationship between a relative frequency graph and its cumulative relative frequency graph is an important one to note. The following two sets of graphs demonstrate this relationship. The first diagram is a set of three relative frequency graphs (skewed right, symmetric, and skewed left) and the second diagram shows the corresponding cumulative frequency graphs. Notice the shape of the cumulative graphs.

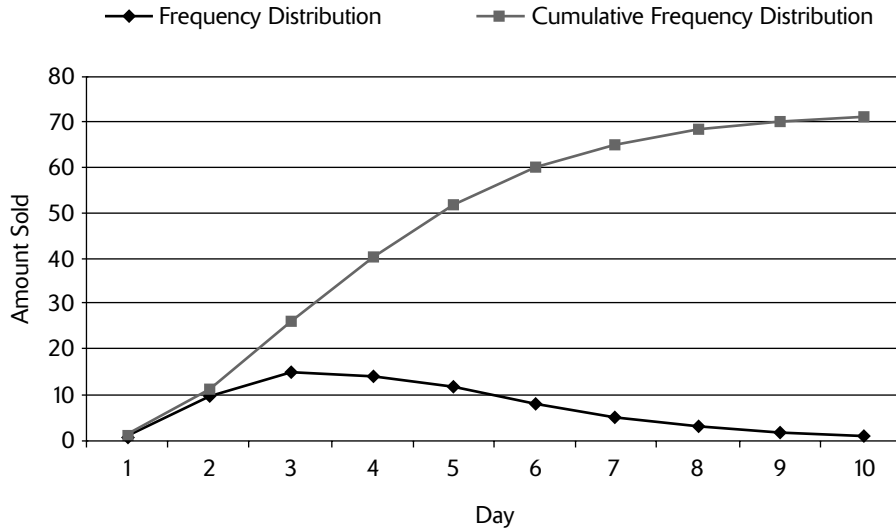


**EXAMPLE:**

The following represents the amount of product sold each day for ten days for a salesman. Graph the frequency distribution and the cumulative frequency distribution on the same graph.

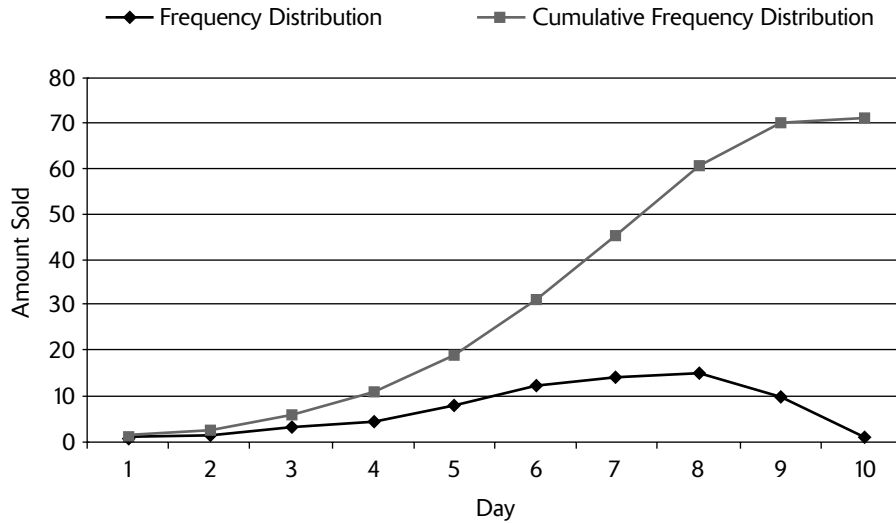
Day	1	2	3	4	5	6	7	8	9	10
Amount Sold	1	10	15	14	12	8	5	3	2	1

This frequency distribution shows a skew to the right. Notice that the cumulative graph rises more quickly during the first few days as compared to the last few days.



If the amounts sold were reversed, the distribution would be skewed left and the cumulative graph rises more rapidly during the last few days.

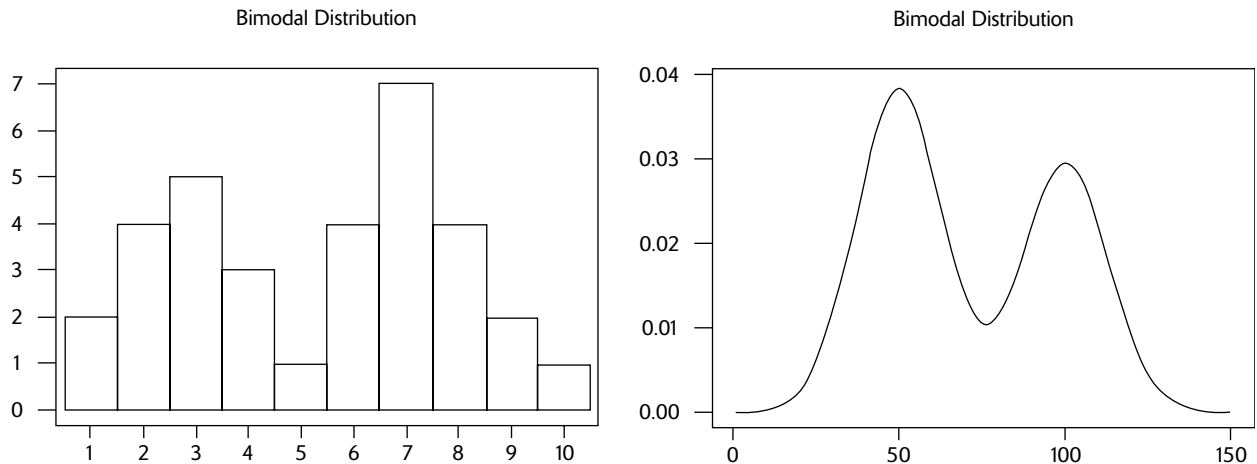
Day	1	2	3	4	5	6	7	8	9	10
Amount Sold	1	2	3	5	8	12	14	15	10	1



Distributions made up of a single mound of data are said to have one **mode**. A mode is a high point in the distribution. Each of the distributions in the previous example (demonstrating shape) have a single mode. Some distributions have more than one mode. If the distribution had two main mounds, then it is bimodal. Distributions with more than two mounds are considered multimodal.

**EXAMPLE:**

These distributions contain more than one mound of connected data. They have more than one mode.



## Review Questions and Answers

---

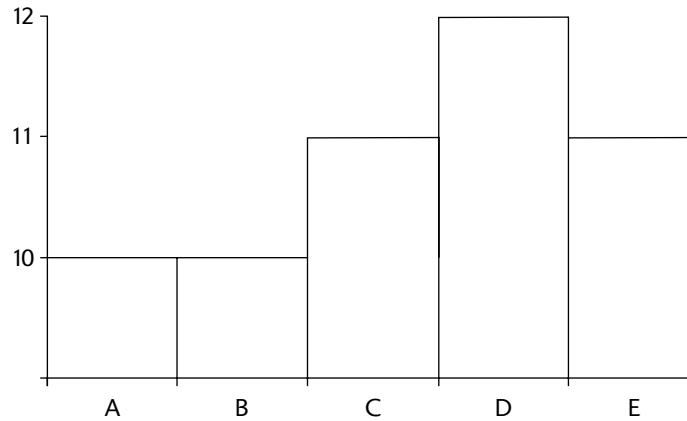
### Multiple Choice Questions

**Directions:** Solve each of the following problems. Decide which is the best of the choices given.

---

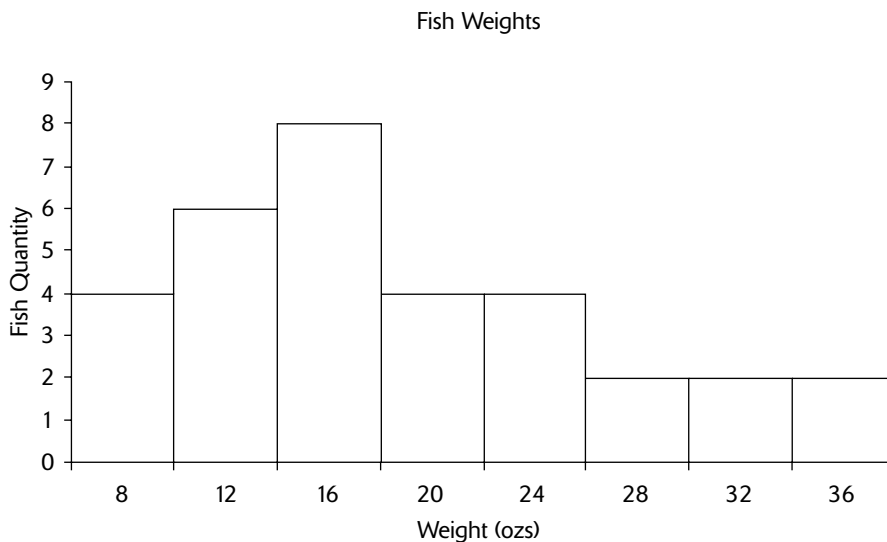
1. Which of the following is/are suitable to determine whether a distribution is skewed?
  - I. Stemplot
  - II. Histogram
  - III. Cumulative frequency charts
  - A. III only
  - B. I and II
  - C. I and III
  - D. II and III
  - E. I, II, and III
  
2. Which of the following statements is false?
  - A. Continuous distributions may have gaps or outliers.
  - B. A skewed right distribution and a skewed left distribution cannot be compared using a back-to-back stemplot.
  - C. Histograms are never continuous.
  - D. In histograms, the area of the bars may be used to compute relative frequency.
  - E. A frequency histogram can be constructed from a cumulative frequency chart.

3. Which statement is best represented by this histogram?



- A. The histogram is skewed right.
  - B. Division D sold more than A and B combined.
  - C. Division D made  $\frac{1}{3}$  of all sales.
  - D. Division C sold more than division B.
  - E. Division C sold twice as much as division B.
4. If a relative frequency distribution is symmetric, the sum of the relative frequencies is:
- A. 0.5
  - B. 0.68
  - C. 1.00
  - D. 100
  - E. Cannot be determined from the information given
5. Which of the following are true statements about histograms?
- I. Midpoints are just as useful in determining class width as boundaries.
  - II. They are useful in displaying cumulative frequencies.
  - III. They are used to display categorical data.
- A. I only
  - B. II only
  - C. I and II
  - D. I and III
  - E. II and III

6. This histogram represents the weights of 16 fish caught in Lake Thomas on April 1. Which of the following statements is true based on this histogram? The  $x$ -values represent the midpoints of the class.



- A. More than half of the fish caught weighed less than 16 ounces.
  - B. Half the fish caught weighed between 16 and 24 ounces.
  - C. The relative frequency for fish weighing between 22 and 30 ounces is 0.1875.
  - D. The distribution is skewed left.
  - E. More than half of the fish caught weighed at least 16 ounces.
7. In a histogram, integer data is grouped into five classes. The classes contain the following age ranges:

Class	I	II	III	IV	V
Age Range	12–17	18–23	24–29	30–35	36–41

- What is the numeric width of each class?
- A. 5.5
  - B. 6.0
  - C. 6.5
  - D. 7.0
  - E. 7.5
8. A 4-class histogram is used to display information about the height (to the closest inch) of 30 club members. The minimum height is 57 inches, and the maximum is 80 inches. If the right boundary of the fourth class is 82 and the width of each class is an integer value, what would be the largest possible value of the boundary between the first and second classes?
- A. 60 inches
  - B. 61 inches
  - C. 62 inches
  - D. 63 inches
  - E. 64 inches

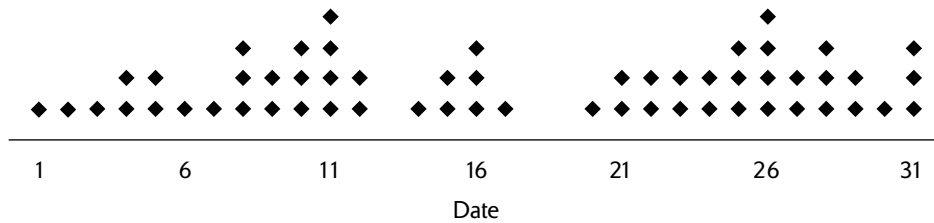
9. Which of the following is not true about histograms?

- A. The bars must touch each other.
- B. The width of each class must be the same.
- C. A histogram for a given set of data may be symmetric or skewed depending on the selection of class width and boundaries.
- D. Histograms may have gaps and clusters.
- E. Histograms may represent categorical data.

10. Which of the following is true?

- A. Histograms convey more information than stemplots.
- B. Dotplots are not used for categorical data.
- C. Relative frequency histograms convey less information than frequency histograms.
- D. The largest value of a set of data can be determined by observing a histogram of the data.
- E. If a stemplot for a set of data shows a gap, then a histogram for the same data will also show a gap.

11. This dotplot represents the number of people applying for jobs at a company during March.



If the first is a Monday, what is the approximate relative frequency of people applying for a job on a weekend (Saturday and Sunday)?

- A. 0.17
- B. 0.19
- C. 0.21
- D. 0.23
- E. 0.25

12. Which of the following is true about this back-to-back stemplot?

<i>Team A</i>		<i>Team B</i>
994	16	68
8776442110	17	47889
99888863	18	346679
94431	19	113367799
986554	20	01136688
97755	21	144566

- I. The team with the highest score was the team with the lowest score.
- II. The distribution for Team A is skewed right, and the distribution for Team B is skewed left.
- III. Team A had a higher average score than Team B.

- A. I only
- B. I and II
- C. I and III
- D. II and III
- E. I, II, and III

13. Which of the following statements is false?
- A. A symmetric distribution can have outliers.
  - B. A skewed left distribution cannot have outliers on the right.
  - C. A skewed right distribution may have outliers on the right.
  - D. A symmetric distribution may have more outliers than a skewed distribution.
  - E. Stemplots are suitable for showing outliers.

## Multiple Choice Answers

1. **E.** Stemplots and histograms directly show whether a distribution is skewed. Using a cumulative frequency chart, the frequency chart can be constructed, and a determination can be made if the distribution is skewed. Also, left and right skewed distributions have distinctive cumulative frequency charts.
2. **B.** Back-to-back stemplots are very appropriate to use to compare the shapes of two distributions. Response **C** is true because histograms are used for discrete data—not continuous data. Response **D** is true if the scales of the histogram are accurate. Response **E** is true because frequency histograms and cumulative frequency histograms can be determined from each other. Relative frequency can be determined from frequency, but frequency cannot be determined from relative frequency.
3. **D.** The histogram is skewed left. Because the scale on the left does not start at zero, the actual areas that the bars show are misleading.
4. **C.** The symmetry of a distribution does not have anything to do with the fact that the sum of all the relative frequencies must be equal to 1.
5. **C.** The class width of a histogram can be determined by calculating the difference between the midpoints of adjacent classes or by calculating the difference between the left and right boundaries of a class. Histograms can represent either frequency or cumulative frequency. Histograms are not used for categorical data.
6. **C.** Choices **A**, **B**, and **E** are each false because 16 is a midpoint, and the distribution of fish weights within the class is unknown. Choice **D** is false because the distribution is skewed right. The relative frequency for the fish in the classes between 22 and 30 can be determined by adding to find the total frequency for the classes. That total is 32. Then find the total frequency for the classes between 22 and 30. This total is 6. Six divided by 32 is 0.1875.
7. **B.** The width of a class is determined by finding the difference between midpoints of adjacent classes or by subtracting the left and right boundary values of a class. The second class, for example, has left and right boundaries of 17.5 and 23.5, respectively. Therefore, the class width is 6.
8. **C.** The statement of the problem implies that the values of the boundaries are integers. A chart demonstrates that the largest possible value for the required boundary of 61 occurs when the class width is 7. If the class width is 6, then the left boundary of the histogram is 58. This is not possible because the smallest data value is 57.

<---Class 1--->	<---Class 2--->	<---Class 3--->	<---Class 4--->
58	64	70	76
54	61	68	75

9. **E.** This is an important distinction between histograms and bar charts. Histograms are not used for categorical data.
10. **C.** A relative frequency distribution can be constructed from a frequency distribution. A frequency distribution cannot be constructed from a relative frequency distribution. Therefore, a frequency distribution conveys more information than does a relative frequency distribution.
11. **B.** If the first is on a Monday, then the weekends fall on days 6, 7, 13, 14, 20, 21, 27, and 28. The total number of dots above weekends is 11. The total number of dots is 57. Dividing, you get approximately 0.19.

12. **B.** Choice III is not accurate. Team A had a lower average score than Team B. Choices I and II are accurate.
13. **B.** Outliers usually lie in the direction of the skew, but not always. Distributions can be drawn easily with an outlier on the opposite side from the skew. The other choices are accurate.

## Free-Response Questions

**Directions:** Show all work. Indicate clearly the methods you use. You will be graded on method as well as accuracy.

1. The following are test scores from 40 students in a tenth-grade math class. Construct a histogram with 6 classes and use it to describe the shape of the distribution.

97	95	88	74	69	86	70	79
99	90	89	97	59	67	78	96
88	66	76	92	98	77	62	84
76	85	48	69	63	98	49	91
80	58	99	85	63	60	97	96

2. Compare and contrast these two distributions using a back-to-back stemplot.

Distribution A:

105	65	27	33	94	79	48	45
53	85	66	108	59	70	22	71
54	47	23	39	93	89	73	56
101	77	73	66	96	22	82	34
99	76	77	34	86	68	54	108
109	88	64	47	98	67	96	84
35	88	59	68	97	28		

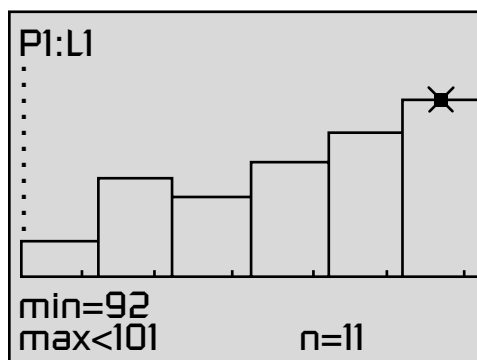
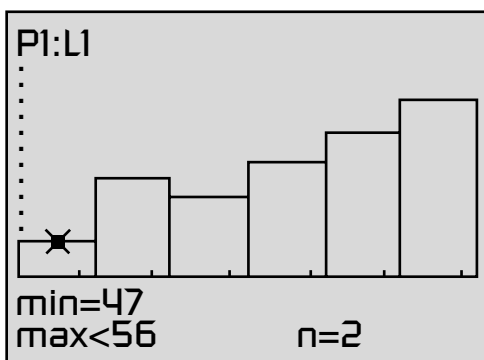
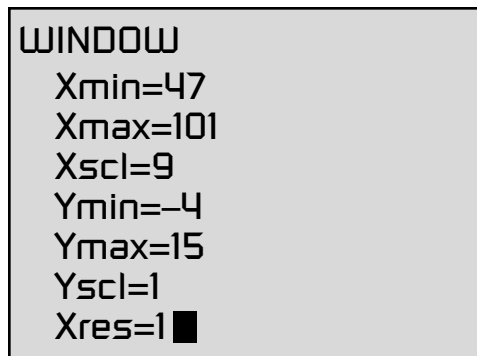
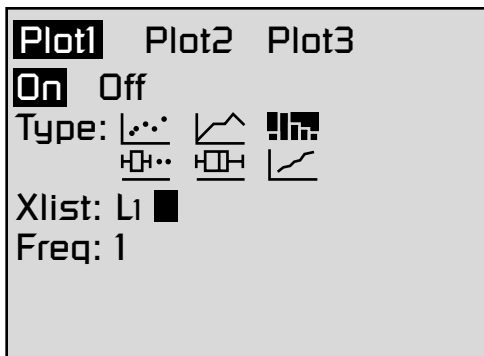
Distribution B:

43	59	102	86	48	27	74	66
31	99	112	95	61	20	57	50
22	63	21	31	79	57	26	56
33	66	32	45	104	89	43	82
63	77	44	57	83	47	21	34
37	96	33	36	96	76	83	65
106	37	104	84	48	79		

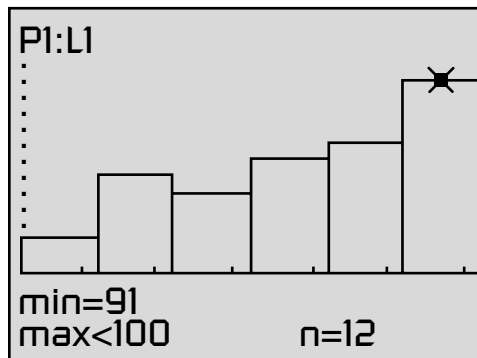
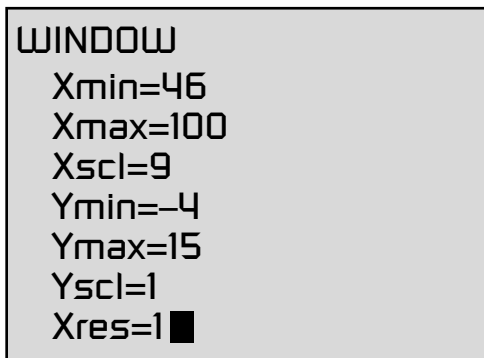
## Free-Response Answers

1. The minimum value in the distribution is 48, and the maximum value is 99. Because both of these numbers plus all those in between must be contained in the distribution, you can subtract to find the range of values of the data. The difference between 99 and 48 is 51. If you keep the class width as an integer, you could make the class width 9. The six classes then would be 54 units wide. The actual choice for the beginning and end points of the histogram will determine its shape. You must start lower than the minimum data value of 48 and extend higher than the maximum data value of 99. Enter the 40 numbers into a list, such as L1, in the TI-83/4.

If the left edge of the histogram begins at 47 and there's a class width of 9, then the class boundaries would be 47, 56, 65, 74, 83, 92, and 101. The following screenshots display the output of the TI-83/84:



If the left edge of the histogram begins at 46 and there's a class width of 9, then the class boundaries would be 46, 55, 64, 73, 82, 91, and 100. The following screenshots display the output of the TI-83/84:



Although the frequency of the last two classes changes by one each as the boundaries are shifted, the shape of the distribution remains approximately the same. This distribution is definitely skewed left. One half of the entire distribution is located in the last two classes.

2. Create a back-to-back stemplot.

<b>Distribution A</b>		<b>Distribution B</b>
87322	2	011267
95443	3	112334677
8775	4	3345788
996443	5	067779
8876654	6	133566
97763310	7	46799
9886542	8	233469
9876643	9	5669
98851	10	2446
	11	2

Distribution A is skewed to the left (the low numbers) and distribution B is skewed to the right (the high numbers). Distribution B has the smallest value (20) of both distributions; it also has the largest value (112). The average value of distribution A appears to be greater than that of distribution B. Neither distribution has any outliers, and neither distribution has gaps or clusters. Both distributions contain the same number of data values (54).

