AP Statistics
Summer Assignment
Chapter 1:
Exploring Data

This pie chart shows how much pie I ate while making this chart.
BIZARRO.COM
Dist. by King Features
©DAN PIRARO 12-1-04

Name ____

Introduction

statistics (p.4) –

~ any set of data contains information about some group of *individuals*, and this information is organized in *variables*

individuals (p.4) -

variables (p.4) -

- **categorical variable (p.5) -**

- **quantitative variable (p.5)-**

with a new set of data, we always ask:

- **who (p.6)** –

- **what (p.6)** –

- **why (p.6)** –

distribution of a variable (p.6) -

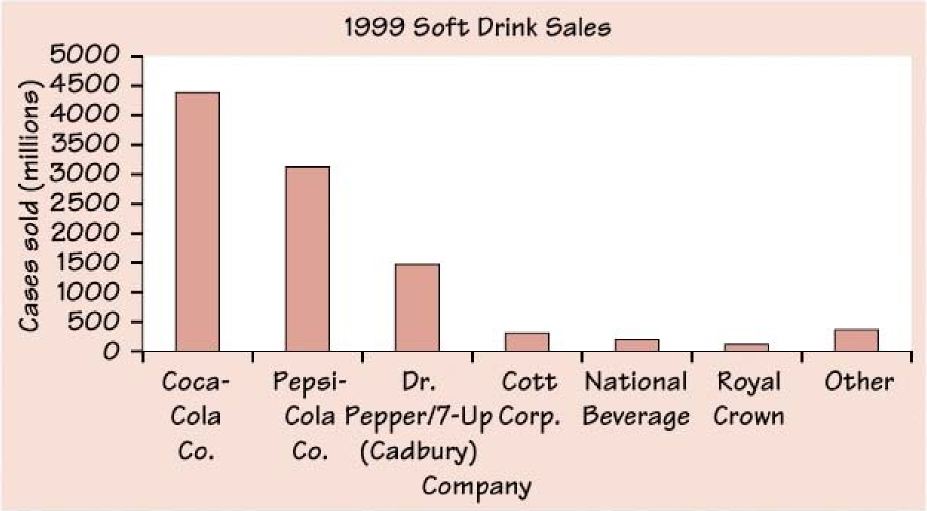
exploratory data analysis (**EDA**) (p.6) -

1.1 - displaying distributions with graphs

displaying categorical variables we use: bar graphs and pie charts

(p.8) distribution of a categorical variable lists the categories and gives either the __________ or the__________ of individuals who fall in each category.

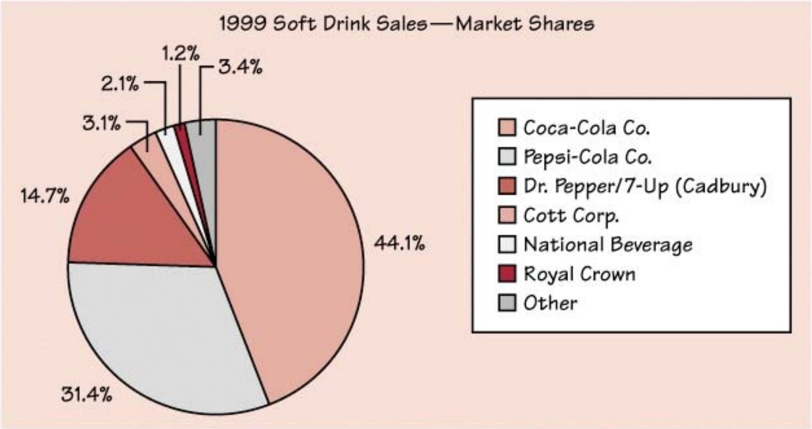
bar graph – *quickly compares; heights of bars show the counts*



(a)

• **remember**: *label axis, title graph, scale axis, leave space between bars*

pie chart - *helps to see what part of the whole each group forms*



(b)

• **remember**: must include all categories that make up a whole (= 100%)

advantages: helps an audience grasp the distribution quickly

disadvantages: takes time and space

1.6. In 1997 there were 92,353 deaths from accidents in the United States. Among these were 42,340 deaths from motor vehicle accidents, 11,858 from falls, 10,163 from poisoning, 4,051 from drowning, and 3,601 from fires.

a. Find the percent of accidental deaths from each of these causes, rounded to the nearest percent. What percent of accidental deaths were due to other causes?

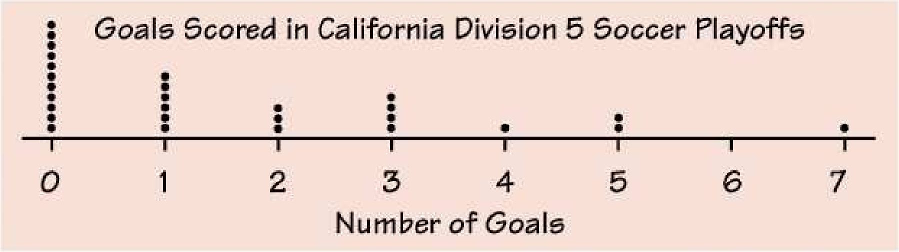
b. Make a well-labeled bar graph of the distribution of causes of accidental deaths. Be sure to include an “other causes” bar.

c. Would it also be correct to use a pie chart to display these data? If so, construct the pie chart. If not, explain why not.

* Note * The AP exam generally does not give you graph paper to construct graphs, get used to making them on blank paper like this one.

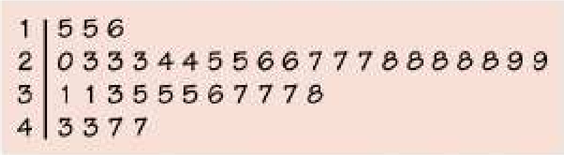
displaying quantitative variables we use: dotplots, stemplots, histograms

dotplots - *one of simplest types of graphs to display quantitative data*



• **remember**: *label axis, title graph, scale axis*

stem (and leaf) plot - *graphs quantitative data*



1 | 5 5 6
2 | 0 3 3 3 4 4 5 5 6 6 7 7 7 8 8 8 8 8 9 9
3 | 1 1 3 5 5 5 6 7 7 7 8
4 | 3 3 7 7

CAFFEINE CONTENT (MG) PER 8-OUNCE
SERVING OF VARIOUS SOFT DRINKS

Key:
3|5 means the soft drink contains 35 mg of caffeine per 8-ounce serving

• how to construct a stemplot:

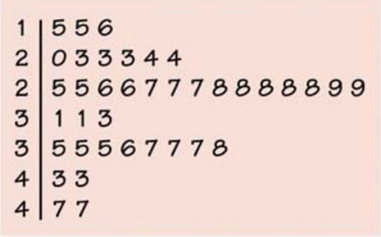
1. *separate each observation into a stem consisting of all but the rightmost digit (leaf)*

2. *write stems vertically in increasing order and draw a vertical line to their right; write each leaf to right of its stem*

3. *write stems again and rearrange leaves in increasing order out from stem*

4. *title graph and add key describing what stems and leaves represent*

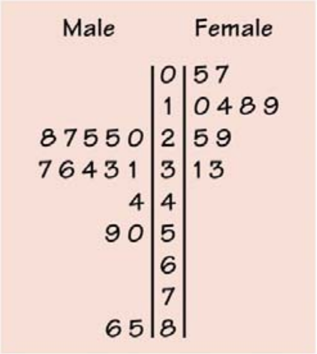
• **split-stem and leaf plot** - *when splitting stems, be sure each stem is assigned an equal number of possible leaf digits*



1 | 5 5 6
2 | 0 3 3 3 4 4
2 | 5 5 6 6 7 7 7 8 8 8 8 8 9 9
3 | 1 1 3
3 | 5 5 5 6 7 7 7 8
4 | 3 3
4 | 7 7

(b)

- **back-to-back stemplots** - *good to compare two sets of data*



Male		Female
	0	57
	1	0489
87550	2	59
76431	3	13
4	4	
90	5	
	6	
	7	
65	8	

- **remember**: 5 stems is a good minimum

advantages: easy to construct; displays actual data values

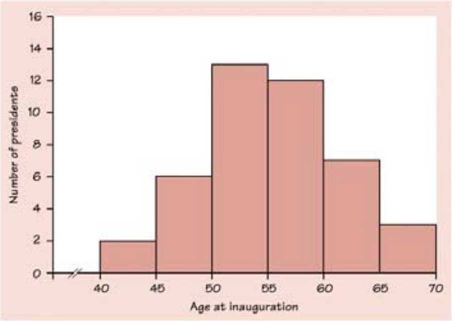
disadvantages: does not work well with large data sets

1.10 DRP TEST SCORES There are many ways to measure the reading ability of children. One frequently used test is the Degree of Reading Power (DRP). In a research of third-grade students, the DRP was administered to 44 students. Their scores were:

40	26	39	14	42	18	25	43	46	27	19
47	19	25	35	34	15	44	40	38	31	46
52	25	35	35	33	29	34	41	49	28	52
47	35	48	22	33	41	51	27	14	54	45

Display these data graphically. Write a paragraph describing the distribution of DRP scores.

histograms - *most common graph for distribution of one quantitative variable*



- how to construct a histogram:

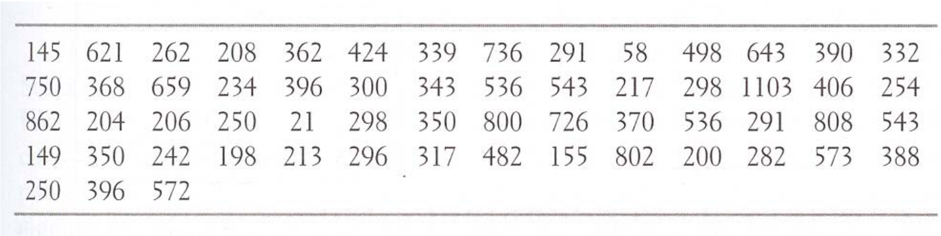
  1. *divide range of data into classes of equal width and count number of observations in each class; be sure to specify classes precisely so that each observation falls into exactly one class*

  2. *label and scale axis and title graph*

  3. *draw a bar that represents the counts in each class*

- **remember**: leave no space between bars; add a break-in-scale symbol (//) on an axis that does not start at 0; 5 classes is a good minimum

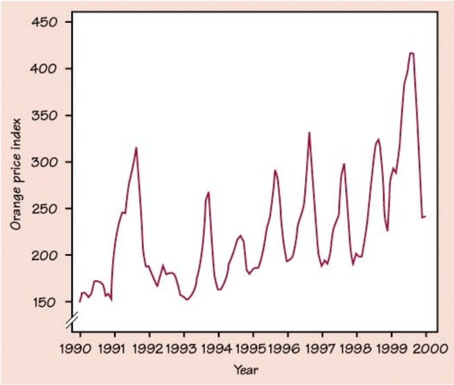
1.14 In 1993, *Forbes* magazine reported the age and salary of the chief executive officer (CEO) of each of the top 59 small businesses. Here are the salary data, rounded to the nearest thousand dollars:



145	621	262	208	362	424	339	736	291	58	498	643	390	332
750	368	659	234	396	300	343	536	543	217	298	1103	406	254
862	204	206	250	21	298	350	800	726	370	536	291	808	543
149	350	242	198	213	296	317	482	155	802	200	282	573	388
250	396	572											

Construct a histogram for this data.

time plots - *plots each observation against the time at which it is measured*



- **trend (p.32) -**

- **seasonal variation (p.32)**

- **remember:** *mark time scale on horizontal axis and variable of interest on vertical axis; connecting points helps show pattern of change over time*

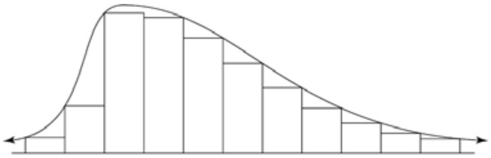
When we describe a graph, we want to discuss the overall pattern of its distribution. Be sure to include the following terms.

- **center (p.12)**

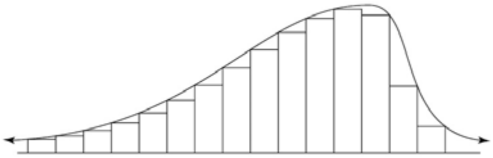
- **spread (p.12)**

- **shape (p.12)**

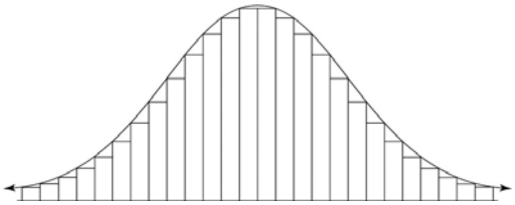
    - **skewed to the right** (**positively skewed**) (p.25)



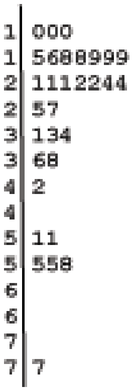
    - **skewed to the left** (**negatively skewed**) (p.25)



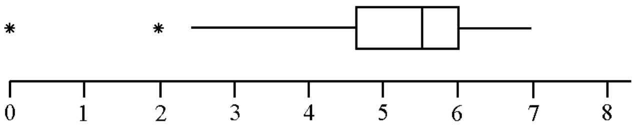
    - **symmetric (p.25)**

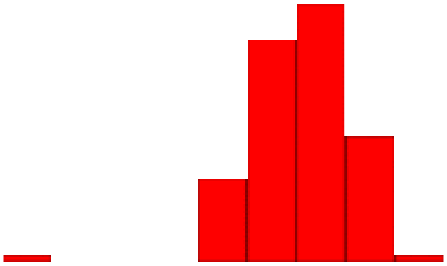


Outlier (p.12) –



```
1 | 000
1 | 5688999
2 | 1112244
2 | 57
3 | 134
3 | 68
4 | 2
4 |
5 | 11
5 | 558
6 |
6 |
7 |
7 | 7
```



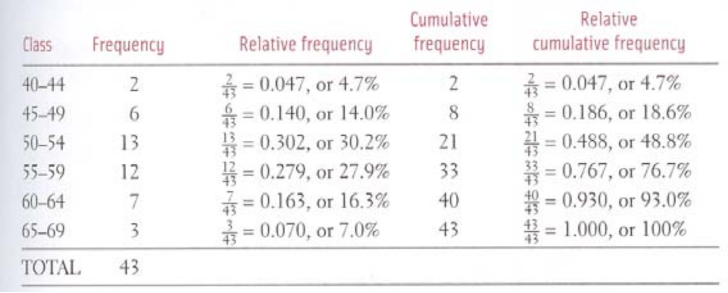


percentile (*p*th percentile) (p.28) –

relative frequency (p.29) –

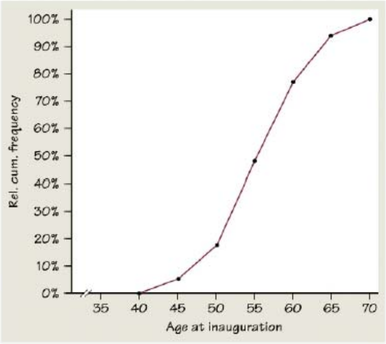
cumulative frequency (p.29) –

relative cumulative frequency (p.29) –

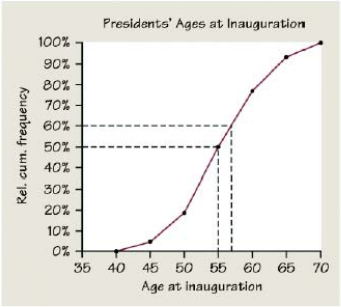


Class	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
40–44	2	$\frac{2}{43}$ = 0.047, or 4.7%	2	$\frac{2}{43}$ = 0.047, or 4.7%
45–49	6	$\frac{6}{43}$ = 0.140, or 14.0%	8	$\frac{8}{43}$ = 0.186, or 18.6%
50–54	13	$\frac{13}{43}$ = 0.302, or 30.2%	21	$\frac{21}{43}$ = 0.488, or 48.8%
55–59	12	$\frac{12}{43}$ = 0.279, or 27.9%	33	$\frac{33}{43}$ = 0.767, or 76.7%
60–64	7	$\frac{7}{43}$ = 0.163, or 16.3%	40	$\frac{40}{43}$ = 0.930, or 93.0%
65–69	3	$\frac{3}{43}$ = 0.070, or 7.0%	43	$\frac{43}{43}$ = 1.000, or 100%
TOTAL	43			

relative cumulative frequency graph/ogive - *graph of relative standing of an individual observation*



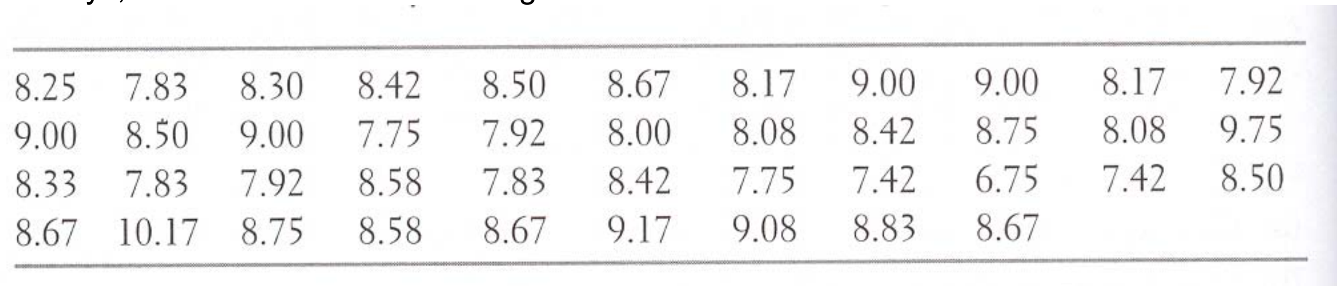
(figure 1.12)



(figure 1.13)

• **remember**: *begin ogive with a point at height 0% at left endpoint of lowest class interval; last point plotted should be at a point of 100%*

1.29 Professor Moore, who lives a few miles outside a college town, records the time he takes to drive to the college each morning. Here are the times (in minutes) for 42 consecutive weekdays, with the dates in order along the rows:



8.25	7.83	8.30	8.42	8.50	8.67	8.17	9.00	9.00	8.17	7.92
9.00	8.50	9.00	7.75	7.92	8.00	8.08	8.42	8.75	8.08	9.75
8.33	7.83	7.92	8.58	7.83	8.42	7.75	7.42	6.75	7.42	8.50
8.67	10.17	8.75	8.58	8.67	9.17	9.08	8.83	8.67		

a. Make a histogram of these drive times. Is the distribution roughly symmetric, clearly skewed, or neither? Are there any clear outliers?

b. Construct an ogive for Professor Moore’s drive times.

c. Use your ogive from *b.* to estimate the center and 90th percentile for the distribution.

d. Use your ogive to estimate the percentile corresponding to a drive time of 8.00 minutes.

1.2a - describing distributions with numbers

measuring center

- **mean** $(\bar{x})$ (read x-bar) (p.37) -

- **median** (***M***) (p.39) –

    1.

    2.

    3.

resistant measure (**of center**) (p.40) -

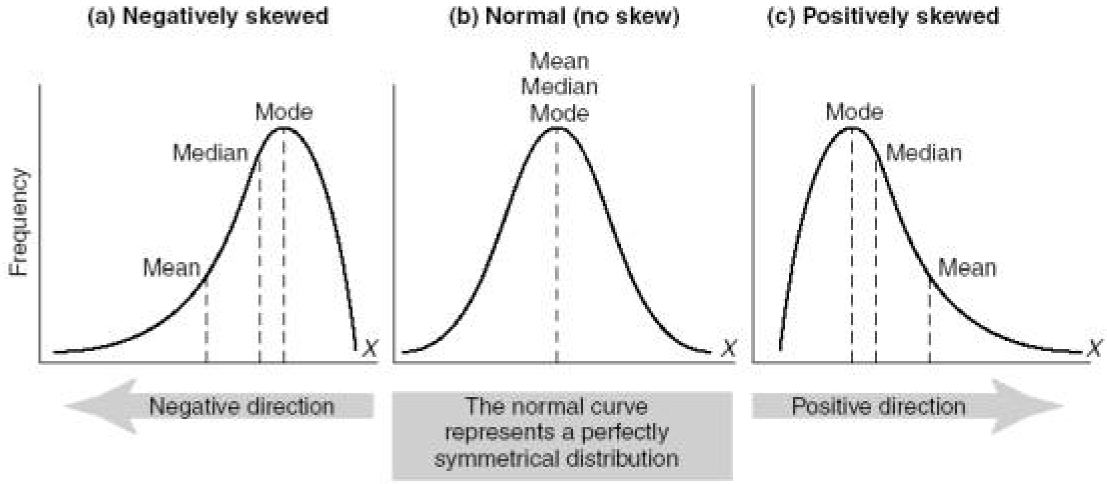


FIGURE 15.6 Examples of normal and skewed distributions

Example 1.32

The Survey of Study Habits and Attitudes (SSHA) is a psychological test that evaluates college students' motivation, study habits, and attitudes toward school. A private college gives the SSHA to a sample of 18 of its incoming first-year women students. Their scores are:

154	109	137	115	152	140	154	178	101
103	126	126	137	165	165	129	200	148

a. Make a stemplot of these data. The pverall shape of the distribution is irregular, as often happens when only a few observations are available. Are there any potential outliers?

b. Find the mean score from the formula for the mean.

c. Find the median of these scores. Which is larger: the median or the mean? Explain why.

measuring spread/variability

- **range** (p.42) -

- **quartiles** (p.42) -

  - **first quartile** (**Q1**) –

  - **third quartile** (**Q3**) –

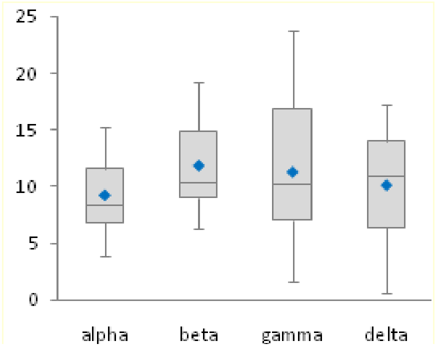
  - “**second quartile**” –

interquartile range (**IQR**) (p.43) -

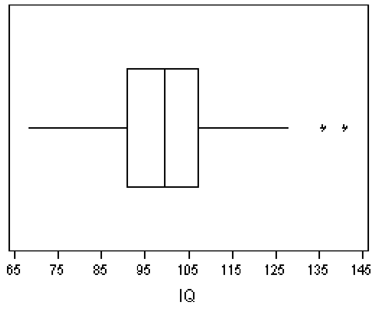
outliers (p.44) –

five number summery (p.44) –

box (and whiskers) plot - *graph of five-number summary of a distribution; best for side by side comparisons since they show less detail than histograms; drawn either horizontally or vertically*



modified boxplot - *plots outliers as isolated points; extend “whiskers” out to largest and/or smallest data points that are not outliers*



- **remember**: *label axis, title graph, scale axis*

1.37 Below you are given all the ages of the presidents at inauguration.

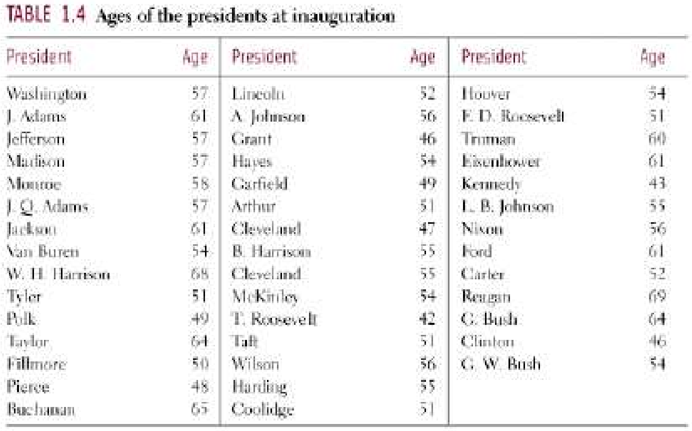
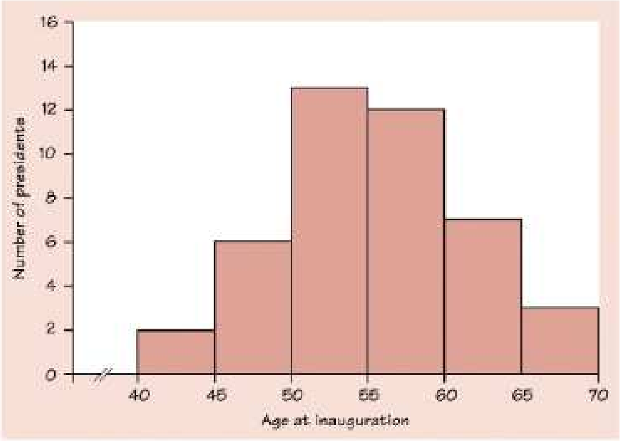


TABLE 1.4 Ages of the presidents at inauguration

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. D. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	61
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. B. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G. Bush	64
Taylor	64	Taft	51	Clinton	46
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55		
Buchanan	65	Coolidge	51		



a. From the shape of the histogram, do you expect the mean to be much less than the median, about the same as the median, or much greater than the median? Explain.

b. Find the five-number summary and verify your expectation from *a*.

c. What is the range of the middle half of the ages of new presidents?

d. Construct by hand a modified boxplot of the ages of new presidents.

e. On your calculator, define Plot1 to be a histogram using the list and Plot2 to be a modified boxplot also using the list. Graph. Is there an outlier? If so, who was it?

1.2b - describing distributions with numbers

measuring spread when we use the median as our measure of center

- range
- interquartile range

measuring spread when we use the mean as our measure of center

• standard deviation (p.49) –

the variance (s^2) of a set of observations is the average of the squares of the deviations of the observations from their mean

$$s^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$$

the standard deviations (*s*) is the square root of the variance (s^2)

$$s = \sqrt{\frac{1}{n-1}\sum(x_i - \overline{x})^2}$$

degrees of freedom (p.50)–

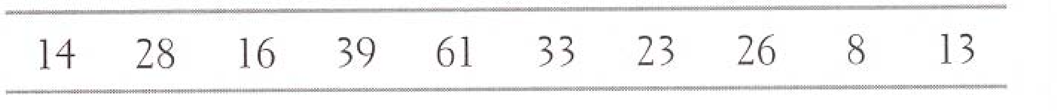
Properties of standard deviations (p.50)

1.

2.

3.

1.41 New York Yankee Roger Maris held the single-season home run record from 1961 until 1998. Here are Maris's home run counts for his 10 years in the American League:



14	28	16	39	61	33	23	26	8	13

a. Maris's mean number of home runs is $\bar{x}$ = 26.1. Find the standard deviation *s* from its definition.

b. Use your calculator to verify your results. Then use your calculator to find $\bar{x}$ and *s* for the 9 observations that remain when you leave out any outlier(s). How does the "outlier" affect the values of $\bar{x}$ and *s*? Is *s* a resistant measure of spread?

linear transformations (p.53)

changes the original variable *x* into the new variable x_{new} given by an equation of the form:

$$x_{\text{new}} = a + bx$$

adding the constant *a*:

multiplying by the positive constant *b*:

~ *linear transformations do not change the shape of a distribution*

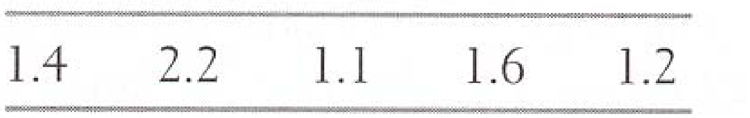
effects of linear transformations (p.55)
:

- linear transformations:

- multiplying each observation by a positive number *b*:

- adding the same number *a* (either positive or negative) to each observation:

1.44 Maria measures the lengths of 5 cockroaches that she finds at school. Here are her results (in inches):



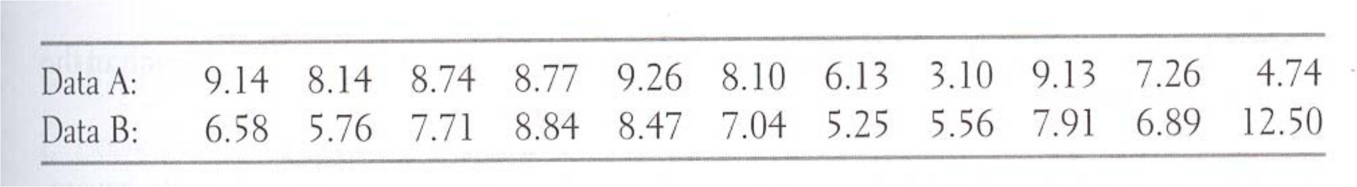
1.4	2.2	1.1	1.6	1.2

a. Find the mean and standard deviation of Maria's measurements.

b. Maria's science teacher is furious to discover that she has measured the cockroach lengths in inches rather than centimeters (There are 2.54 cm in 1 inch). She gives Maria two minutes to report the mean and standard deviation of the 5 cockroaches in centimeters. Maria succeeded. Will you?

c. Considering the 5 cockroaches that Maria found as a small sample from the population of all cockroaches at her school, what would you estimate as the average length of the population of cockroaches? How sure of the estimate are you?

1.54 The mean $\bar{x}$ and standard deviation *s* measure center and spread but are not a complete description of a distribution. Data sets with different shapes can have the same mean and standard deviation. To demonstrate this fact, use your calculator to find $\bar{x}$ and *s* for the following two small data sets. Then make a stem-plot of each and comment on the shape of each distribution.



Data A:	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data B:	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

Name:________________

AP Statistics
Chapter 1 Review Questions

1. Which of the following statements is true?
    i. Two students working with the same set of data may come up with histograms that look different.
    ii. Displaying outliers is less problematic when using histograms than when using stemplots.
    iii. Histograms are more widely used than stemplots or dotplots because histograms display the values of individual observations.

(a) I only
(b) II only
(c) III only
(d) I and II
(e) II and III

2. Which of the following are true statements?
    i. The range of the sample data set is never greater than the ranges of the population.
    ii. The interquartile range is half the distance between the first quartile and the third quartile.
    iii. While the range is affected by outliers, the interquartile range is not.

(a) I only
(b) II only
(c) III only
(d) I and II
(e) I and III

3. On August 23, 1994, the 5-day total returns for equal investments in foreign stocks, U.S. stocks, gold, money market funds, and treasury bonds averages 0.086% (Business Week, September, 5, 1994, p. 97). If the respective returns for foreign stocks, U.S. stocks, gold, and money market funds were +0.70%. -0.11%, +1.34%, and +0.05%, what was the return for treasury bonds?

(a) 0.4132%
(b) 0.43%
(c) -1.55%
(d) -1.636%
(e) -1.77%

4. A 1995 poll by the Program for International Policy asked respondents what percentage of the U.S. budget they thought went to foreign aid. The mean response was 18% and the median was 15%. What do these responses indicate about the shape of the distribution of these responses?

(a) The distribution is skewed to the left.
(b) The distribution is skewed to the right.
(c) The distribution is symmetric around 16.5%.
(d) The distribution is uniform between 15% and 18%
(e) Not enough information.

5. A teacher is teaching two AP Statistics classes. On the final exam, the 10 students in the first class averages 92 while the 25 students in the second class averaged only 83. If the teacher combines the classes, what will the average final exam score be?

(a) 87
(b) 87.5
(c) 88
(d) None of the above
(e) Not enough information

6. A teacher was recording grades fro her class of 32 AP Statistics students. She accidentally recorded one score much too high (she put a "1" in front, so the score was 192 instead of 92, which was the top score). The corrected score was greater than any other grade in the class. Which of the following sample statistics remained the same after the correction was made?

   a. Mean
   b. Standard Deviation
   c. Range
   d. Variance
   e. Interquartile Range

7. During the early part of the 1994 baseball season, many sports fans and baseball players notice that the number of home runs being hit seemed to be unusually large. Here are the data on the number of home runs hit by American and National League teams.

American League	35	40	43	49	51	54	57	58	58	64	68	68	75	77
National League	29	31	42	46	47	48	48	53	55	55	55	63	63	67

(a) Construct back to back stemplots to compare the data.
(b) Calculate numerical summaries of the number of home runs hit in the two leagues.
(c) Are there any outliers in the two sets? Justify your answer numerically.
(d) Write a few sentences comparing the distributions of home runs in the two leagues.

8. Create a data set with five numbers in which

(a) The mean, median and mode are all equal

(b) The mean is greater than the median

(c) The mean is less than the median

(d) The is higher than the median or mean

(e) All the numbers are distinct and the mean and median are both zero

9. Mail-order labs and 1 hour minilabs were compared with regard to price for developing and printing one 24-exposure roll of film. Prices included shipping and handling. Following is a computer output describing the results:

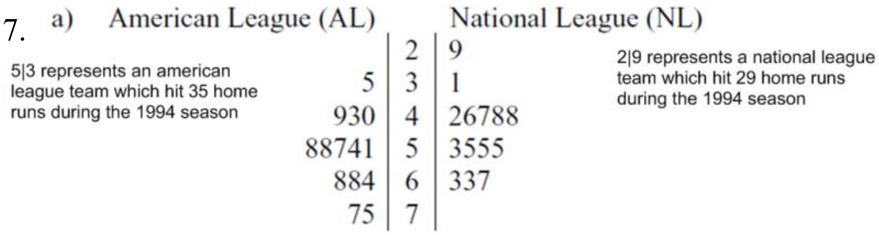
```
Mail-order
Mean = 5.37                 Standard Deviation = 1.92          Min = 3.51
Max = 8.00                  N = 18          Median = 4.77
Quartiles = 3.92, 6.45

Mini-labs
Mean = 10.11                Standard Deviation = 1.32          Min = 5.02
Max = 11.95           N = 15          Median = 10.08
Quartiles = 8.97, 11.51
```

Draw parallel boxplots displaying the distributions. Compare the distributions. Are there any outliers?

Answers to the review sheet

1. a 2. e 3. c 4. b 5. d 6. E

7.



a)

5|3 represents an american league team which hit 35 home runs during the 1994 season

American League (AL)		National League (NL)
	2	9
5	3	1
930	4	26788
88741	5	3555
884	6	337
75	7	

2|9 represents a national league team which hit 29 home runs during the 1994 season

American league
5 number summary
35 49 57.5 68 77

$\bar{x}=56.93$

$s=12.69$

National league
5 number summary
26 46 50.5 55 67

$\bar{x}=50.14$

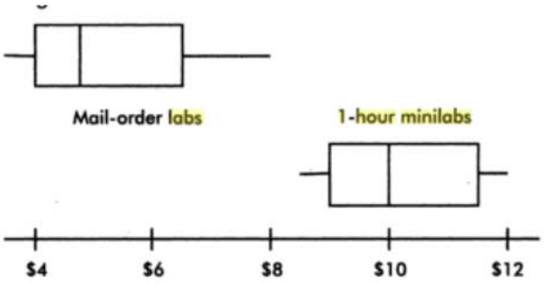
$s=11.13$

Overall, the American League teams hit more home runs than National League teams in 1994. To illustrate this, the median for the AL was 57.5 while the median for the NL was only 50.5. Additionally, the low value for the NL is much lower than in the AL (29 versus 35) and the high value in the AL is much higher than that in the NL (77 versus 67).

8. There are other acceptable answers for this question

a. 3, 4, 4, 4, 5 b. 3, 4, 4, 4, 15
c. 1, 4, 4, 4, 5 d. 1, 2, 3, 4, 4
e. -2, -1, 0, 1, 2

9.

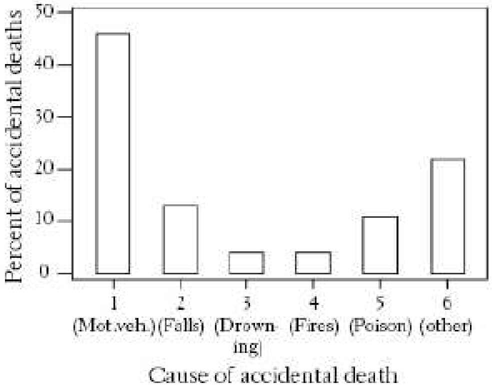


Boxplots show the minimum, maximum, median, and quartile values. The distribution of mail-order lab prices is lower and more spread out than that of prices of 1-hour minilabs. Both are slightly skewed toward the upper end (the skewness can also be noted from the computer output showing the mean to be greater than the median in both cases.)

Answers to packet questions:

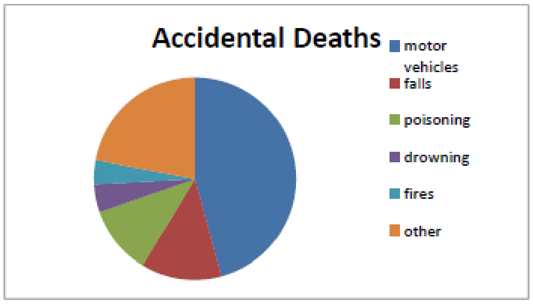
1.6 **ACCIDENTAL DEATHS** In 1997 there were 92,353 deaths from accidents in the United States. Among these were 42,340 deaths from motor vehicle accidents, 11,858 from falls, 10,163 from poisoning, 4051 from drowning, and 3601 from fires.
(a) Find the percent of accidental deaths from each of the causes, rounded to the nearest percent. What percent of deaths were due to other causes? (b) Make a well-labeled bar graph of the distribution of causes of accidental deaths. Be sure to include an "other causes" bar.

Motor Vehicles	= 46 %
Falls	= 13 %
Drowning	= 4 %
Fires	= 4 %
Poisoning	= 11 %
Other causes	= 22 %

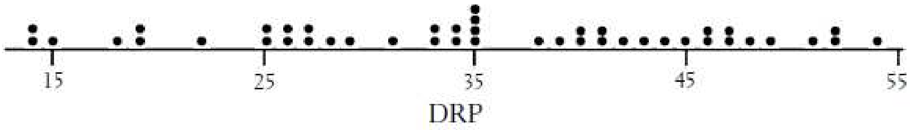


(c) Would it also be correct to use a pie-chart to display these data? If so, construct the pie-chart. If not, explain why not.

A pie chart could also be used, since the categories represent parts of a whole (all accidental deaths).



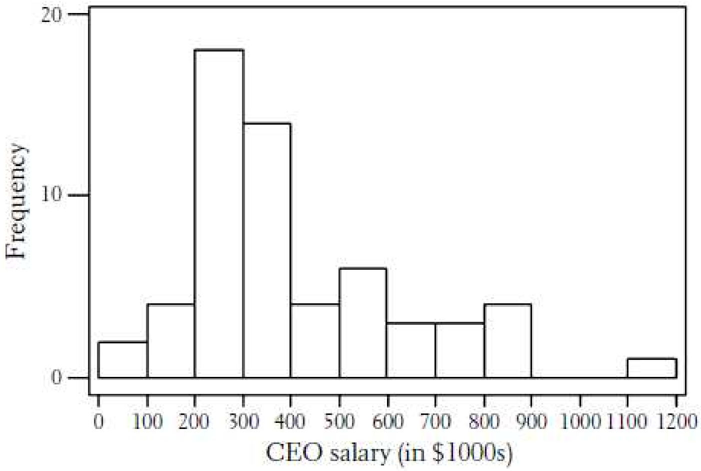
1.10



The center of the distribution is 35, and there are approximately the same number of points to the left and right of the center. There are no major gaps or outliers. The distribution is approximately symmetric.

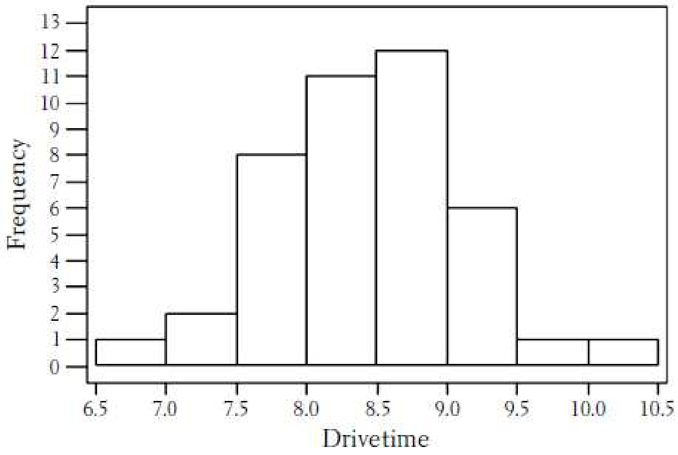
A histogram or a stem and leaf plot would also have been good ways of representing the data for 1.10

1.14



The distribution is skewed to the right with a peak in the 200s class. The spread is approximately 1100 ($21,000 to $1,103,000) and the center is located at 350 ($350,000). There is one outlier in the 1100s class.

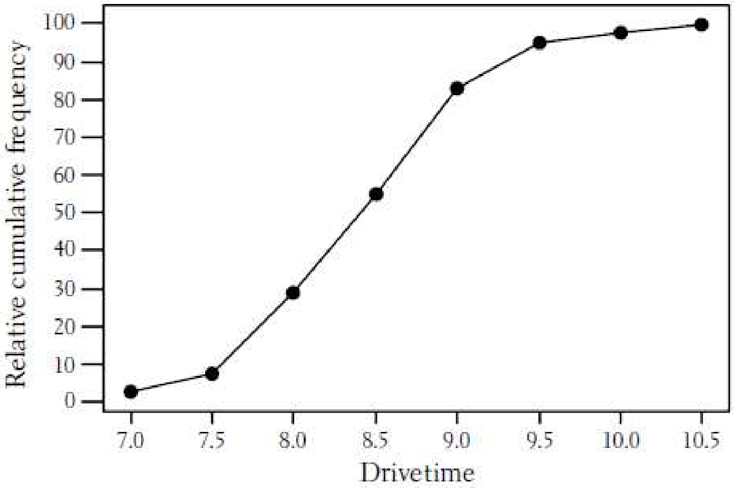
1.29 (a)



The distribution is roughly symmetric with no clear outliers.

(b)

Drivetime	Cum. freq.	Rel. cum. freq.
7.0	1	2.4%
7.5	3	7.1%
8.0	12	28.6%
8.5	23	54.8%
9.0	35	83.3%
9.5	40	95.2%
10.0	41	97.6%
10.5	42	100%



(c) Center ≈ 8.5, 90th percentile ≈ 9.4

(d) 8.0 ≈ 28th percentile

1.32 (a)

```
10 | 139
11 | 5
12 | 669
13 | 77
14 | 08
15 | 244
16 | 55
17 | 8
18 |
19 |
20 | 0
```

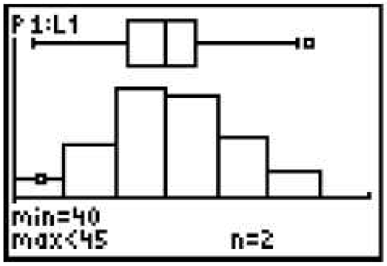
200 is a potential outlier. The center is approximately 140. The spread (excluding 200) is 77.

(b) $\overline{x} = 2539/18 = 141.058$.

(c) Median = average of ninth and tenth scores = 138.5. The mean is larger than the median because of the outlier at 200, which pulls the mean towards the long right tail of the distribution.

1.37 (a) The mean and median should be approximately equal since the distribution is roughly symmetric.

(b) Five-number summary: 42, 51, 55, 58, 69 $\bar{x} = 2357/43 = 54.8$
As expected, median and $\bar{x}$ are very similar.
(c) Between Q_1 and Q_3: 51 to 58.
(e)



The point 69 is an outlier; this is Ronald Reagan's age on inauguration day. W. H. Harrison was 68, but that is not an outlier according to the 1.5(IQR) test.

1.41 (a) $\sum(x_i - \bar{x})^2 = (-12.1)^2 + (1.9)^2 + (-10.1)^2 + (12.9)^2 + (34.9)^2 + (6.9)^2 + (-3.1)^2 + (-.1)^2 + (-18.1)^2 + (-13.1)^2 = 2192.9.$
$s^2 = 2192.9/9 = 243.66$, $s = 15.609$.
(b) Excluding the outlier at 61, we obtain $\bar{x} = 22.2$, $s = 10.244$. The outlier caused the values of both measures to increase; the increase in s is more substantial. Clearly, s is not a resistant measure of spread.

1.44 (a) $\bar{x} = 7.5/5 = 1.5$, $s = .436$.
(b) To obtain $\bar{x}$ and s in centimeters, multiply the results in inches by 2.54: $\bar{x} = 3.81$ cm, $s = 1.107$ cm.
(c) The average cockroach length can be estimated as the mean length of the five sampled cockroaches: that is, 1.5 inches. This is, however, a questionable estimate, because the sample is so small.

1.54 The means and standard deviations are basically the same. For Set A, $\bar{x} \approx 7.501$ and $s \approx 2.032$, while for Set B, $\bar{x} \approx 7.501$ and $s \approx 2.031$. Set A is skewed to the left, while Set B has a high outlier.

Set A		Set B	
3	1	5	257
4	7	6	58
5		7	079
6	1	8	48
7	2	9	
8	1177	10	
9	112	11	
		12	5