Chapter 3—Numerically Summarizing Data

To reduce a large body of data to an understandable form that can be quickly grasped, we construct a frequency distribution table for the data and draw the corresponding histogram or frequency curve (as illustrated in Ch. 2). It is useful to simplify the presentation further by defining specific measures that describe important features of the body of data. Two important measures from a set of data are:

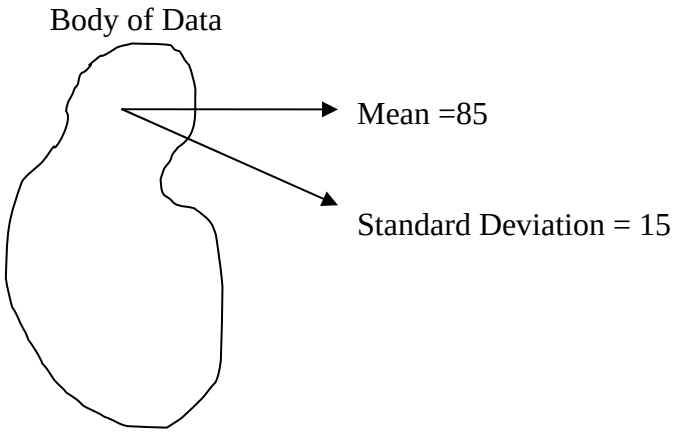
(1) Measure of **Central Tendency**—indicates the center of the body of data.

- Mean
- Median
- Mode

(2) Measure of **Dispersion**—indicates the dispersion (or scatter) about the central point of the data.

- Range
- Variance
- Standard Deviation

Compare the process of characterization of a body of data in statistics to characterizing a person. A person has many characteristics and thus characterization is challenging. For example, a person named John Smith can be characterized by his height, weight, age, hair color, eye color, SAT score, education, income level, hobbies, degree of kindness, etc. By comparison, statistics characterizes a body of data much simpler, using only two measures such as a mean and a standard deviation. In statistics, regardless of the size of the body of data, only two measures are needed to characterize a dataset.



Summation Notation

Summation—is represented by the capital Greek letter Σ (sigma).

X	
x_1	
x_2	X—variable
.	x_i—represents the ith observation (or variate)
.	n—number of observations
.	
x_n	

Definition: $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$

The symbol $\sum_{i=1}^{n} X_i$ is read "the summation of X sub i, where i goes from 1 to n."

Examples of summation:

(1) $\sum_{i=2}^{4} x_i = x_2 + x_3 + x_4$

(2) $\sum_{i=1}^{3} (x_i - 2) = (x_1 - 2) + (x_2 - 2) + (x_3 - 2)$

(3) $\sum_{i=1}^{3} x_i - 2 = x_1 + x_2 + x_3 - 2$

(4) $\sum_{i=1}^{2} x_i^2 = x_1^2 + x_2^2$

(5) $\left[\sum_{i=1}^{2} x_i \right]^2 = [x_1 + x_2]^2 = x_1^2 + 2x_1x_2 + x_2^2$

(6) $\sum_{i=1}^{n} c = n*c$

(7) $\sum_{i=1}^{5} 3 = 5*3 = 15$

3.1 Measures of Central Tendency

Data are provided below for the variable X. We will assume these are observations from the population (N=6).

	X
	x_1=12
	x_2= 5
	7
	10
	3
	11
ΣX_i	48

Unit of measure is tons.

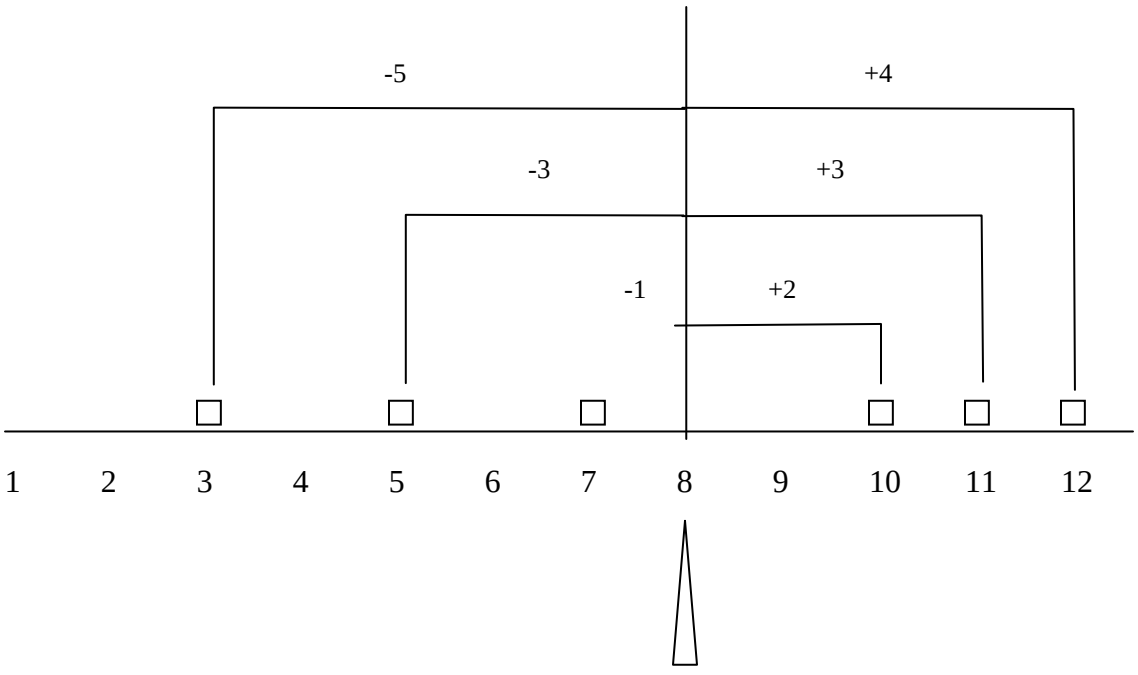
Measures of Central Tendency:

(1) Mean—is commonly referred to as the "average."

Definition: If $x_1, x_2, \cdots, x_N$ are the N observations of a variable from a population, then the population mean, μ, is

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum x_i}{N}$$

In-Class Activity: *"The mean as the Center of Gravity"*—mean as the balancing point (p. 110).



(2) Median—middle number (median of a highway divides the highway in half)

Definition: The **median** of a variable is the value that lies in the middle of the data when arranged in ascending order. That is, half the data are below the median and half the data are above the median. We use M to represent the median.

Steps in Computing the Median of a Data Set.

Step 1: Arrange the data in ascending order.
Step 2: Determine the number of observations, N.
Step 3: Determine the observation in the middle of the data set.

- If the number of observations is **ODD**, then the median is the data value that is exactly in the middle of the data set. That is, the median is the observation that lies in the $\frac{N+1}{2}$ position.
- If the number of observations is **EVEN**, then the median is the mean of the two middle observations in the data set. That is, the median is the mean of the data values that lie in the $\frac{N}{2}$ and $\frac{N}{2}+1$ positions.

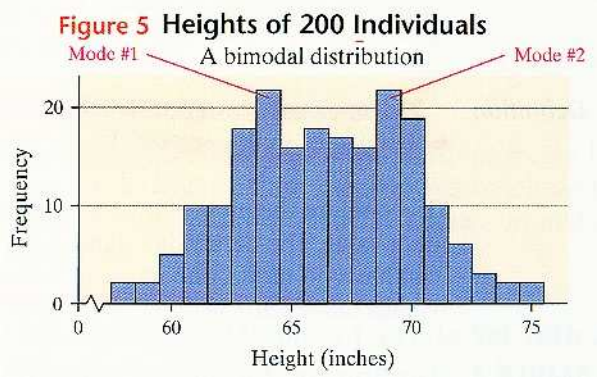
(3) Mode—can be computed for either quantitative or qualitative data.

Definition: The **mode** of a variable is the most frequent observation of the variable that occurs in the data set.

Computing the Mode: Tally the number of observations that occur for each data value. The data value that occurs most often is the mode.

- A data set can have no mode, one mode, or more than one mode.
- If there is no observation that occurs with the most frequency, we say the data has no mode.

Bimodal data—A histogram of heights in which the data set contained both males and females might show two modes. The two modes occur because two genders are mixed in one data set.



N=6	X
	x_1=12
	x_2= 5
	7
	10
	3
	11
ΣX_i	48

Unit of measure is tons.

Mean:

Median:

Mode:

The Shape of the Distribution and the Mean and the Median (p. 113)

#1: 10 20 30 40 50 60 70

mean = 40 and median = 40

#2: 10 20 30 40 50 60 ~~\/~~ 300

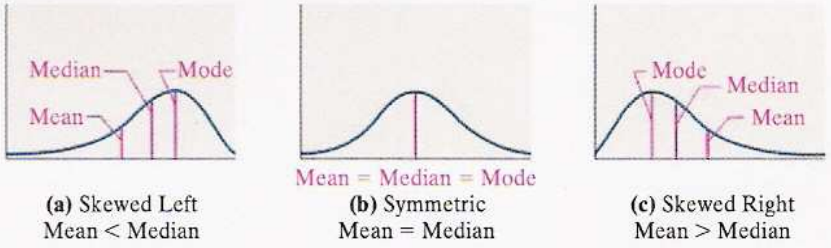
mean = 72.9 and median = 40

Table 5

Relation Between the Mean, Median, and Distribution Shape

Distribution Shape	Mean versus Median
Skewed left	Mean substantially smaller than median
Symmetric	Mean roughly equal to median
Skewed right	Mean substantially larger than median

Figure 5
Mean/median versus skewness



(a) Skewed Left
Mean < Median

(b) Symmetric
Mean = Median

(c) Skewed Right
Mean > Median

Example—Mean and Median Annual Salaries of the 2000-2001 LA Lakers.

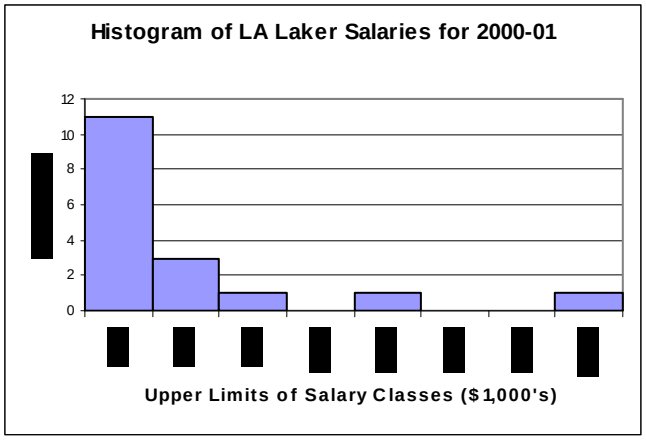
Player	Salary (1,000 $'s)
Shaquille O'Neal	19,290
Kobe Bryant	10,130
Horace Grant	6,500
Robert Horry	4,800
Rick Fox	3,400
Derek Fisher	3,380
Brian Shaw	2,250
Ron Harper	2,200
Greg Foster	1,760
Chuck Person	1,200
Tyronn Lue	870
Devean George	850
Mark Madsen	710
Isaiah Ryder	550
Mike Penberthy	320
Stanislav Medvedenko	320
Sam Jacobson	270

Salary (1,000 $'s)	
Mean	3459
Standard Error	1177
Median	1760
Mode	320
Standard Deviation	4852
Sample Variance	23538774
Kurtosis	7
Skewness	3
Range	19020
Minimum	270
Maximum	19290
Sum	58800
Count	17

This information was calculated in Excel using **Tools/Data Analysis/Descriptive Statistics.**

Frequency Distribution Table:

Class Limits	Frequency
0 - 2500	11
2501 - 5000	3
5001 - 7500	1
7501 - 10000	0
10001 - 12500	1
12501 - 15000	0
15001 - 17500	0
17501 - 20000	1



Example 9—Mean and Median Birth Weight (in grams) of 50 Babies (p. 115)

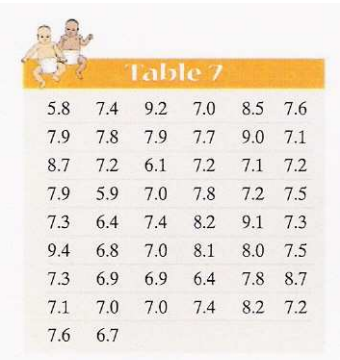
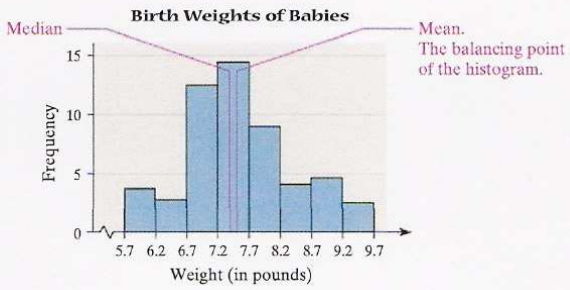


Table 7

5.8	7.4	9.2	7.0	8.5	7.6
7.9	7.8	7.9	7.7	9.0	7.1
8.7	7.2	6.1	7.2	7.1	7.2
7.9	5.9	7.0	7.8	7.2	7.5
7.3	6.4	7.4	8.2	9.1	7.3
9.4	6.8	7.0	8.1	8.0	7.5
7.3	6.9	6.9	6.4	7.8	8.7
7.1	7.0	7.0	7.4	8.2	7.2
7.6	6.7				

Figure 9
Birth weights of 50 randomly selected babies



Relative merits of the Mean, Median, and Mode

Mean—is the most widely used measure of central tendency.

- Takes all measurements into consideration.
- Easy to manipulate with algebra.
- Preferred measure of central tendency for symmetric distributions.

Median—is **resistant** to extreme values.

- For data sets with unusually large or small values relative to the entire data set or when the data are skewed, the median is the preferred measure of central tendency. Examples of the use of the median include *income levels* and *housing prices*, which are skewed right, and *life span (age)*, which is skewed left.
- In economic statistics, it is oftentimes desirable to disregard extreme variates which may be due to unusual circumstances. Median is the preferred measure of central tendency for skewed datasets.

Mode—is resistant to extreme values.

- Can be ambiguous if there are two modes that are not adjacent

3.2 Measures of Dispersion

(1) Range—is the simplest measure of dispersion.

Definition: The **range, R**, of a variable is the difference between the largest data value and the smallest data value.

R = Largest Data Value – Smallest Data Value

(2) Variance—is the average of the squared deviations from the mean.

Definition: The **population variance** of a variable is the sum of the squared deviations about the population mean divided by the number of observations in the population, N. The population variance is symbolically represented by σ^2 (lower case Greek sigma squared).

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N},$$

where $x_1, x_2, \ldots, x_N$ are the N observations in the population and μ is the population mean.

Note: In using this formula, do not round until the last computation. Use as many decimal places as allowed in order to avoid round-off errors.

(3) Standard Deviation—is the most popular measure of dispersion.

Definition: The **population standard deviation**, σ, is obtained by taking the square root of the population variance.

$$\sigma = \sqrt{\sigma^2}$$

Interpretation of variance/standard deviation: The variance (or standard deviation) provides a *quantitative measure* of the amount of variation in a dataset.

	Set #1	Set #2	Set #3
	10	10	5
	10	10	25
	10	10	18
	10	11	40
Mean	10	10.25	22.00
Variance	0	0.19	159.50
Std. deviation	0	0.43	12.63

When all the observations are the *same* (Set #1), the var/SD is 0.
When the observations differ by a *small amount* (Set #2), the var/SD is small.
When the observations differ by a *large amount* (Set #2), the var/SD is large.
General Interpretation: As the dispersion in the observations increases, the var/SD increases.

N=6	X	(xi-μ)	$(xi-\mu)^2$
	12		
	5		
	7		
	10		
	3		
	11		
Σx_i	48	$\Sigma(xi-\mu)$ =	$\Sigma(xi-\mu)^2$ =
μ	48/6=8.0 tons		

Range:

Variance:

Standard Deviation:

The Empirical Rule

If a distribution is roughly bell shaped, then

- Approximately 68% of the data will lie within 1 standard deviation of the mean. That is, approximately 68% of the data lie between $\mu - 1\sigma$ and $\mu + 1\sigma$.
- Approximately 95% of the data will lie within 2 standard deviations of the mean. That is, approximately 95% of the data lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- Approximately 99.7% of the data will lie within 3 standard deviations of the mean. That is, approximately 99.7% of the data lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Figure 13

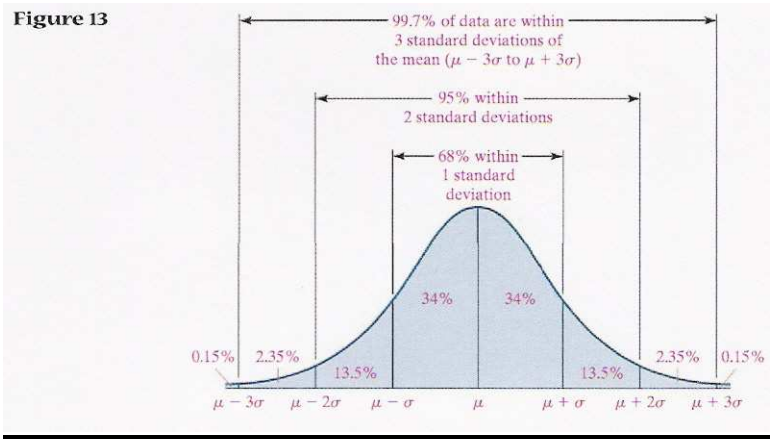
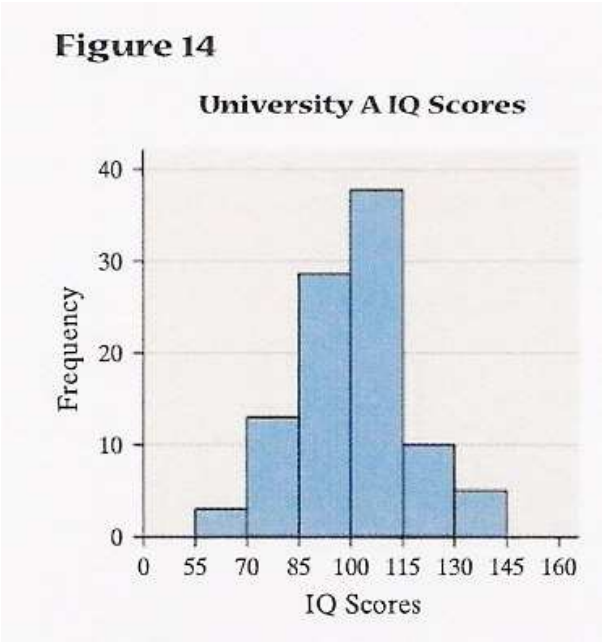


Figure 14



The distribution of IQ scores is approximately bell shaped.

Interpretation of Investment Returns using the Mean and Standard Deviation

Example--Return on Investments (%/yr)

Measures of Central Tendency and Dispersion	Investment #1	Investment #2	Investment #3
Mean	6	6	8
Std Dev.	0	10	25

Notation and Formulas for Mean, Variance, and Standard Deviation (for sections 3.1 & 3.2).

Characteristic	Population	Sample
Mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{SS}{n-1}$
Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

Definitions: **Parameter**—numerical characteristic of a population, e.g., μ and σ.
Statistic—numerical characteristic of a sample, e.g., $\bar{x}$ and s.

SS stands for sum of squared deviations from the mean, meaning $\sum (x_i - \bar{x})^2$. Using algebra, alternative formulas for SS can be derived from the original definition formula given above.

$$SS = \sum x_i^2 - \frac{[\sum x_i]^2}{n}$$

$$SS = \sum x_i^2 - n\bar{x}^2$$

Show how to calculate SS using the formulas to the left, and how to calculate the sample variance using s^2=SS/(n-1).

These formulas are called "working formulas" because they are easier to use in making calculations. All the formulas give the same answers, so you should choose the formula that is easiest to use in a particular situation.

n=6	X	X^2
	12	
	5	
	7	
	10	
	3	
	11	
Σx_i	48	Σxi^2 =
$\bar{x}$	48/6=8	

Caluculate SS using the "working formulas":

$$SS = \sum x_i^{\,2} - \frac{[\sum x_i]^2}{n} =$$

$$SS = \sum x_i^{\,2} - n\bar{x}^2 =$$

3.3--Mean, Variance, and Standard Deviation from Grouped Data.

The Approximate Mean of a Variable from a Frequency Distribution:

Population mean:

$$\mu = \frac{\sum x_i f_i}{\sum f_i} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_h f_h}{f_1 + f_2 + \ldots + f_h}$$

Sample mean:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_h f_h}{f_1 + f_2 + \ldots + f_h}$$

where x_i is the midpoint or value of the ith class.

f_i is the frequency of the ith class

h=number of classes

The Approximate Variance of a Variable from a Frequency Distribution:

Population Variance:

$$\sigma^2 = \frac{\sum_{i=1}^{h} (x_i - \mu)^2 f_i}{\sum_{i=1}^{h} f_i}$$

Sample Variance:

$$s^2 = \frac{\sum_{i=1}^{h} (x_i - \bar{x})^2 f_i}{\sum_{i=1}^{h} f_i - 1}$$

where x_i is the midpoint or value of the ith class.

f_i is the frequency of the ith class

h=number of classes

Table12: Three-Year Rate of Return of Mutual Funds (%).

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

$\bar{x} = 23.4$

$s^2 = 81.49$

$s = 9.03$

	Class Midpoint x_i	Freq. f_i	$x_i f_i$	$\bar{x}_i$	$(x_i - \bar{x}_i)^2 f_i$
10.0 - 14.9	12.5	7	87.5	23.25	808.9375
15.0 - 19.9	17.5	11	192.5	23.25	363.6875
20.0 - 24.9	22.5	8	180	23.25	4.5
25.0 - 29.9	27.5	6	165	23.25	108.375
30.0 - 34.9	32.5	3	97.5	23.25	256.6875
35.0 - 39.9	37.5	3	112.5	23.25	609.1875
40.0 - 44.9	42.5	0	0	23.25	0
45.0 - 49.9	47.5	2	95	23.25	1176.125
		$\sum f_i = 40$	$\sum x_i f_i = 930$		$\sum (x_i - \bar{x})^2 f_i = 3327.5$

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} =$$

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

$$s = \sqrt{s^2} =$$

Simple Mean vs. Weighted Mean

Problem: A mathematics class is divided into two sections, both of which are given the same test. Section 1 (10 students) has a mean score of 62; and section 2 (30 students) has a mean score of 88. Find the mean of the entire class.

Two mean formulas are available—simple mean and weighted mean (adapted from equation (2), p. 144, Text):

Simple Mean:

$$\bar{x}_S = \frac{\bar{x}_1 + \bar{x}_2}{2} = \frac{62 + 88}{2} = 75,$$

Weighted Mean:

$$\bar{x}_W = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{10(62) + 30(88)}{10 + 30} = 81.5$$

where $\bar{x}_S$ =simple mean and $\bar{x}_W$ =weighted mean;

$\bar{x}_1$ and $\bar{x}_2$ are the respective means of sections 1 and 2; and

n_1 and n_2 are the number of students in sections 1 and 2, respectively.

(In comparing this formula to formula (2), p. 117, Text, the weights (w_i) are the number of students in each section, i.e., $w_1=n_1$ and $w_2=n_2$.)

If your objective is to characterize the performance of the entire class, which mean does the best job? The answer is the weighted mean because it gives due ("fair") weight to the individual section means. This is best seen from the weighted mean formula shown below (this formula is algebraically derived from the weighted mean formula above).

$$\bar{x}_W = \frac{n_1}{n_1 + n_2}(\bar{x}_1) + \frac{n_2}{n_1 + n_2}(\bar{x}_2) = \frac{10}{10 + 30}(62) + \frac{30}{10 + 30}(88)$$

$$= \frac{10}{40}(62) + \frac{30}{40}(88)$$

$$= 0.25(62) + 0.75(88) = 81.5$$

This formula shows that the weighted mean is calculated by multiplying each section mean by a weight factor, where the weight factor is the proportion of students in a section ($n_i/\Sigma n_i$). The weight factor for the first section is 0.25, representing the fact that 25% of the entire class is in section 1; and the weight factor for the second section is 0.75, indicating that 75% of the students are in section 2. By comparison, the simple mean for the entire class uses the same weights (0.50) for each section mean:

$$\bar{x}_S = \frac{1}{2}(\bar{x}_1) + \frac{1}{2}(\bar{x}_2) = 0.50(62) + 0.50(88) = 75$$

(Note in calculating the simple mean and weighted mean that the weight factors sum to 1.0.)

Under what circumstances is a simple mean appropriate?—if both sections of the class have the same number of students. In this case, the simple mean is acceptable because it uses the same weight factors as the weighted mean (0.50 for section 1 and 0.50 for section 2).

3.4 Measures of Position

Z-Scores

Definition: The **z-score** represents the number of standard deviations that a data value is from the mean. It is obtained by subtracting the mean from the data value and dividing the result by the standard deviation. There is both a population z-score and a sample z-score:

$$Population\ Z-Score \qquad z=\frac{x-\mu}{\sigma}$$

$$Sample\ Z-Score \qquad z=\frac{x-\bar{x}}{s}$$

The z-score is unitless; it has mean 0 and standard deviation 1. The z-score will be used in Ch 7 with the normal probability distribution.

Example—Evaluation of students based on GPA.

Let's compare two students, Fred and Tom. Fred has GPA=3.50 and Tom has GPA=2.75. Which student is the better student?—Fred or Tom.

Data are available for GPA's for two universities. Means and standard deviations are calculated for the student GPA's from each university.

	University A	University B
Mean	3.10	2.05
Std. Dev.	0.60	0.40

Fred attends University A and has a GPA of 3.50 (grade of A). Tom attends University B and has a GPA of 2.75 (grade of C+). As candidates for admission to a professional school (e.g., medical, law, pharmacy, etc.), which student would rank the highest?—Fred or Tom. Assume that the overall quality of the education and students are equal at the two universities.

Determining the kth Percentile, P_k.

Step 1: Arrange the data in ascending order.
Step 2: Compute an index i using the formula

$$i = \left(\frac{k}{100}\right)(n+1)$$

where k is the percentile of the data value
n=number of observations in the data set.

Step 3: **(a)** If *i* is an integer, the kth percentile, P_k, is the ith data value.
(b) If i is not an integer, find the mean of the observations on either side of i. This number represents the kth percentile, P_k.

Percentiles

Example of Determining the Percentile of a Data Value Using the PGA Salaries, (the data are two pages below).
Top 130 Golfers on the 2000 PGA Tour, Table 14.

TABLE 14

1. 339,242	20. 425,624	39. 519,740	58. 638,422	77. 877,390	96. 1,263,585	115. 1,968,685
2. 346,569	21. 459,812	40. 527,741	59. 649,674	78. 889,153	97. 1,320,278	116. 2,002,068
3. 379,349	22. 460,024	41. 528,959	60. 660,707	79. 889,381	98. 1,368,888	117. 2,023,465
4. 387,716	23. 461,981	42. 537,105	61. 669,709	80. 896,098	99. 1,384,508	118. 2,025,781
5. 388,341	24. 464,480	43. 538,706	62. 673,387	81. 947,118	100. 1,543,818	119. 2,068,499
6. 391,075	25. 466,345	44. 548,070	63. 700,738	82. 963,974	101. 1,550,592	120. 2,099,943
7. 393,059	26. 466,712	45. 552,795	64. 724,580	83. 964,346	102. 1,557,720	121. 2,169,727
8. 393,316	27. 467,431	46. 564,918	65. 728,635	84. 990,215	103. 1,563,115	122. 2,337,765
9. 397,610	28. 469,590	47. 580,510	66. 731,925	85. 999,460	104. 1,597,139	123. 2,413,345
10. 398,393	29. 482,028	48. 583,605	67. 741,995	86. 1,004,827	105. 1,604,952	124. 2,462,846
11. 402,017	30. 482,744	49. 590,109	68. 747,312	87. 1,040,244	106. 1,631,695	125. 2,547,829
12. 403,982	31. 485,589	50. 597,021	69. 753,709	88. 1,048,166	107. 1,642,221	126. 2,573,835
13. 406,591	32. 493,906	51. 604,199	70. 762,979	89. 1,054,338	108. 1,702,317	127. 3,061,444
14. 414,123	33. 494,307	52. 608,535	71. 774,249	90. 1,063,456	109. 1,747,643	128. 3,469,405
15. 414,509	34. 495,975	53. 610,432	72. 784,754	91. 1,096,131	110. 1,804,433	129. 4,746,457
16. 415,430	35. 498,749	54. 611,209	73. 827,691	92. 1,138,749	111. 1,819,323	130. 9,188,321
17. 417,646	36. 507,308	55. 612,882	74. 838,054	93. 1,142,789	112. 1,842,221	
18. 418,780	37. 511,414	56. 617,242	75. 854,822	94. 1,207,104	113. 1,932,280	
19. 424,309	38. 514,193	57. 631,752	76. 867,372	95. 1,262,535	114. 1,940,519	

Source: Yahoo! Sports

Use the procedure outlined above to calculate the 85^{th} percentile.

$$i = \left(\frac{85}{100}\right)(130+1) = 111.35$$

The 85^{th} percentile is the average of the 111^{th} and 112^{th} observations.

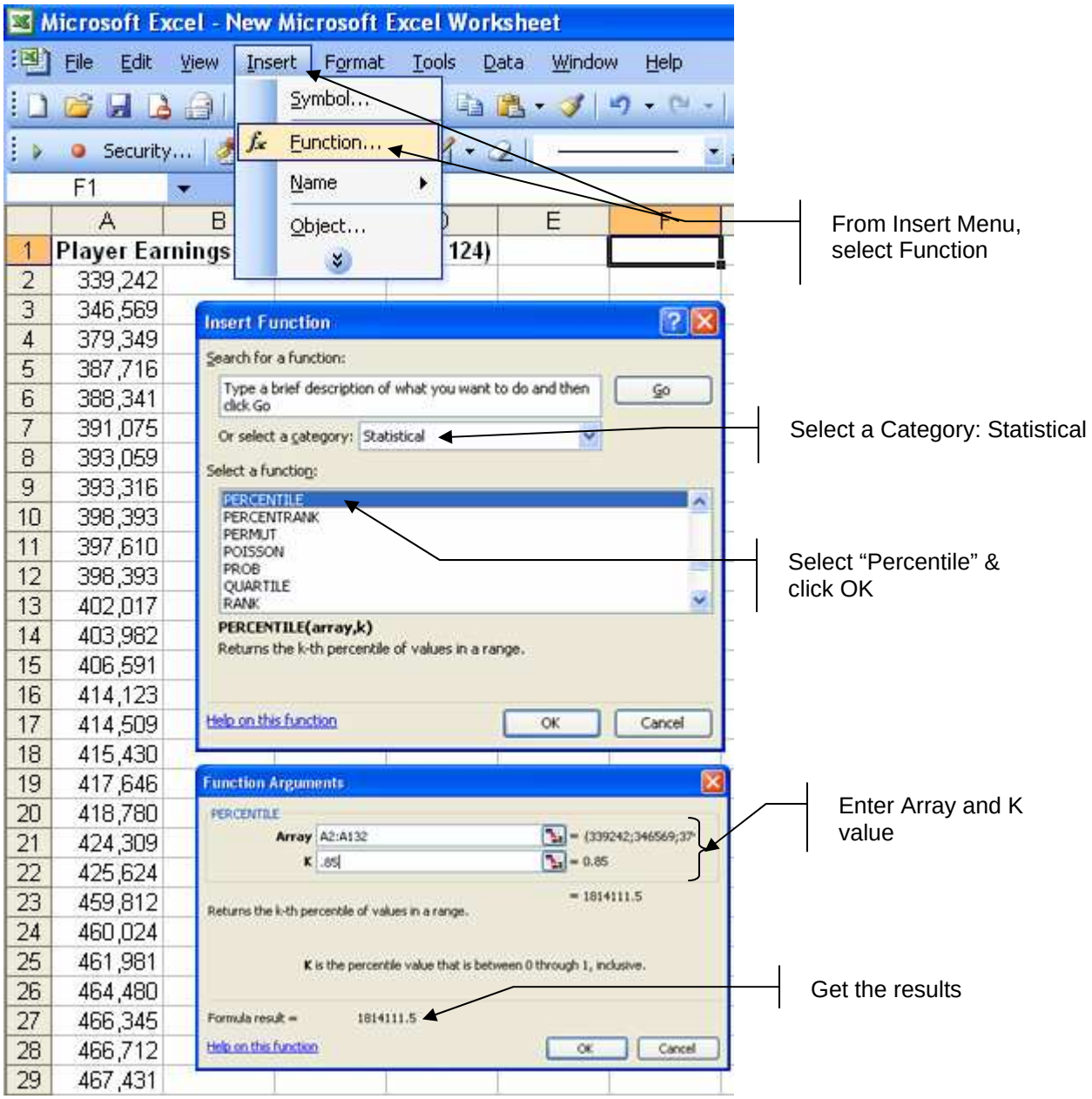
$$P_k = \frac{1{,}819{,}323 + 1{,}842{,}221}{2} = 1{,}830{,}772$$

***Algorithms for the calculation of percentiles use interpolations and different software often uses slightly different rules…So, the answers can vary. Examples of alternative percentile calculation rules, to those used in *Sullivan,* are found in the following texts:

Statistics for Managers Using Microsoft Excel by Levine, Stephan, Krehbiel, Berenson. Pearson-Prentice Hall, 2005. See pp. 110-111.

Introduction to Probability & Statistics (10^{th} ed.) by Mendenhall, Beaver, and Beaver. Duxbury Press, 1999. See pp. 75-76.

2. Use Excel to calculate the 85th percentile (see figure below).



Note that Excel calculates the 85th percentile as $1,814,112 which is different than the $1,819,323 calculated from the percentile procedure in the *Sullivan* text (p. 152). This shows that Excel uses an alternative procedure for calculating percentiles. Do not be disturbed by this difference; the concept of percentile is valid.

.

The 85th percentile is the data value such that approximately 85% of the observations lie below this value

The Five-Number Summary, Boxplots, and Outliers.

The Five-Number Summary

MINIMUM Q_1 M Q_3 MAXIMUM

Drawing a Boxplot

Step 1; Determine the **five-number summary**.

Step 2: Compute the **interquartile range** or **IQR**, which is the difference between the third and first quartiles.

$IQR = Q_3 - Q_1$.

Step 3: Determine the lower and upper fences:

Lower fence = Q_1 -1.5(IQR)

Upper fence = Q_3 + 1.5(IQR)

Step 4: Draw vertical lines at Q_1, M, and Q_3. Enclose these vertical lines in a box.

Step 5: Label the lower and upper fences.

Step 6: Draw a line from Q_1 to the smallest data value that is larger than the lower fence. Draw a line from Q_3 to the largest data value that is smaller than the upper fence.

Step 7: Any data values less than the lower fence or greater than the upper fence are **outliers** and are marked with an asterisk (*).

Table12: Three-Year Rate of Return of Mutual Funds (%).

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

Five-number summary:	MINIMUM	Q1	M	Q3	MAXIMUM
	10.8	17.05	22.05	28.8	47.7

Interquartile range: IQR = Q1 – Q3 = 28.8 – 17.05 = **11.75**

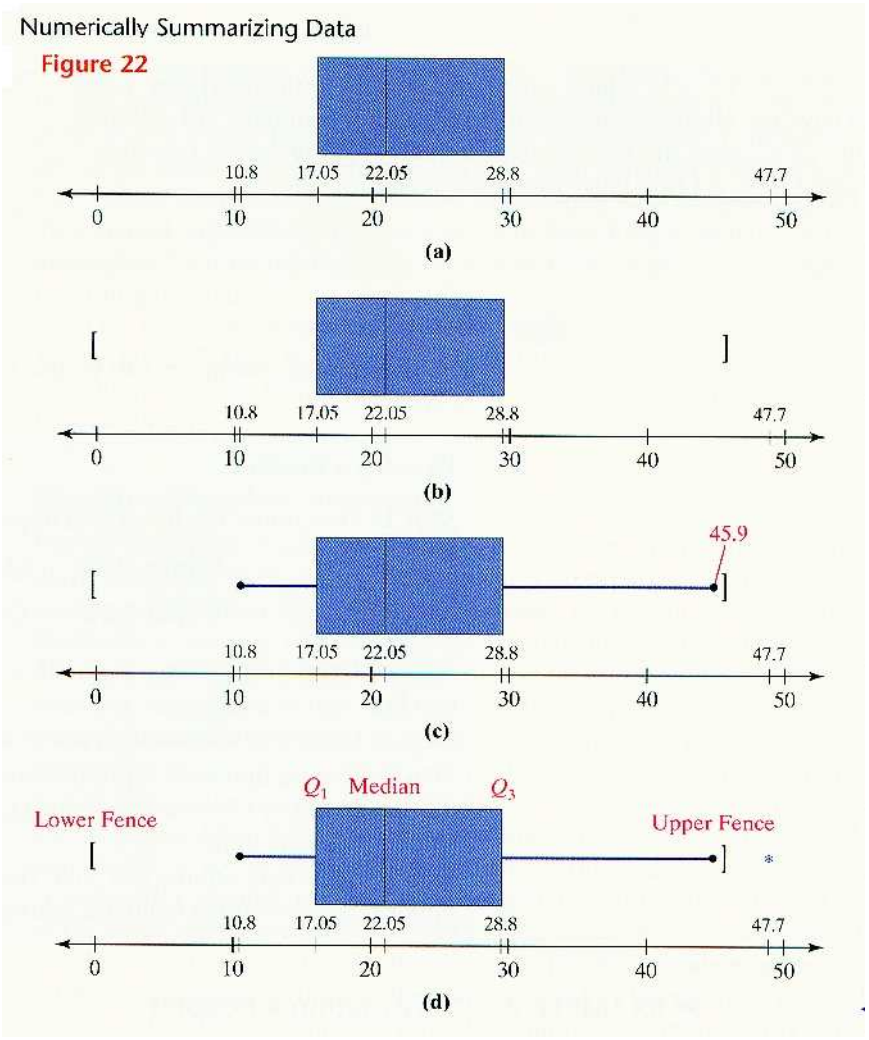
Lower fence = Q1 – 1.5(IQR) = 17.05 – 1.5(11.75) = **-0.575**

Upper fence = Q3 + 1.5(IQR) = 28.8 + 1.5(11.75) = **46.425**

Smallest data value that is larger than the lower fence is **10.8.**

Largest data value that is less than the upper fence is **45.9.**

Outliers: Any data value that is lower than the lower fence or greater than the upper fence is marked with an *. There is one upper outlier, **47.7**.



Note: Do not show fences (upper or lower) on a boxplot.

Drawing Boxplots—is a tedious and time-consuming task. An Excel macro program is available to draw boxplots

Distribution Shape Based on Boxplot (p. 162)

1. If the median is near the center of the box and each of the horizontal lines is approximately equal length, then the distribution is roughly symmetric.
2. If the median is to the left of the center of the box or the right line is substantially longer than the left line, the distribution is skewed right.
3. If the median is to the right of the center of the box or the left line is substantially longer than the right line, the distribution is skewed left.

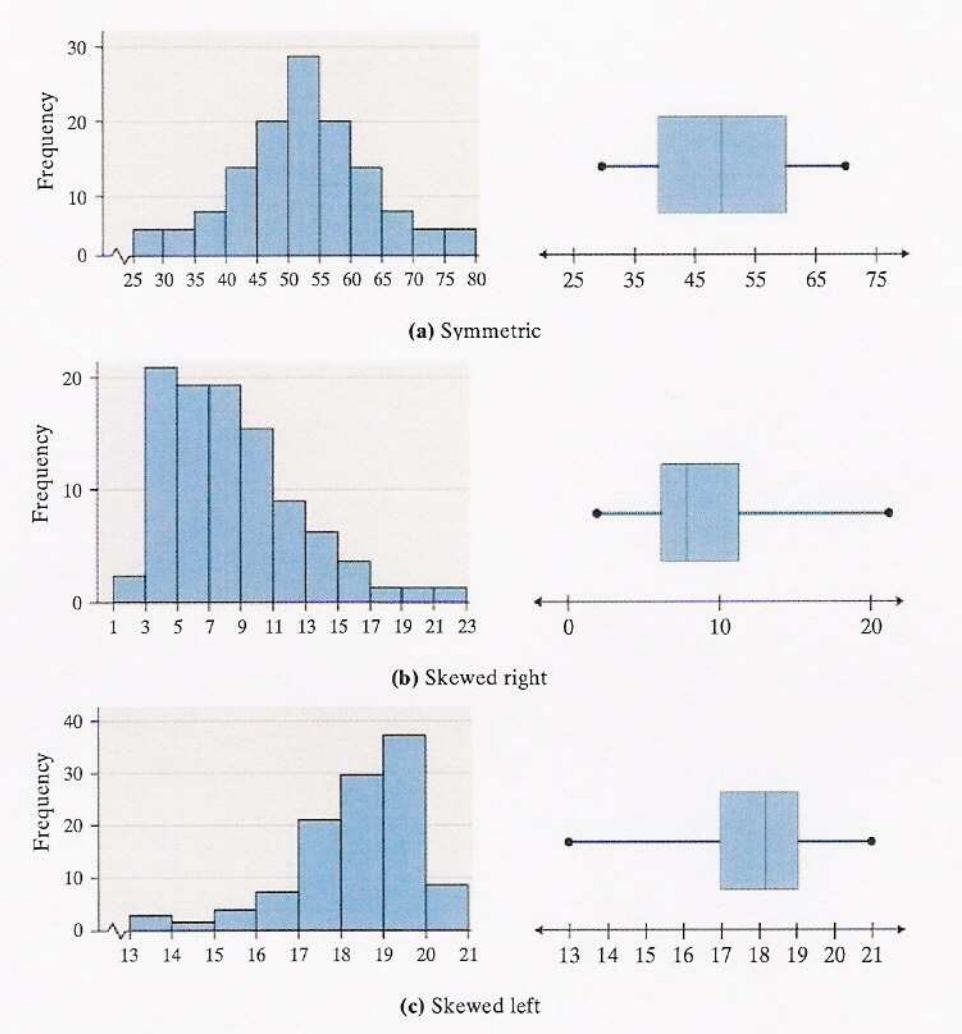
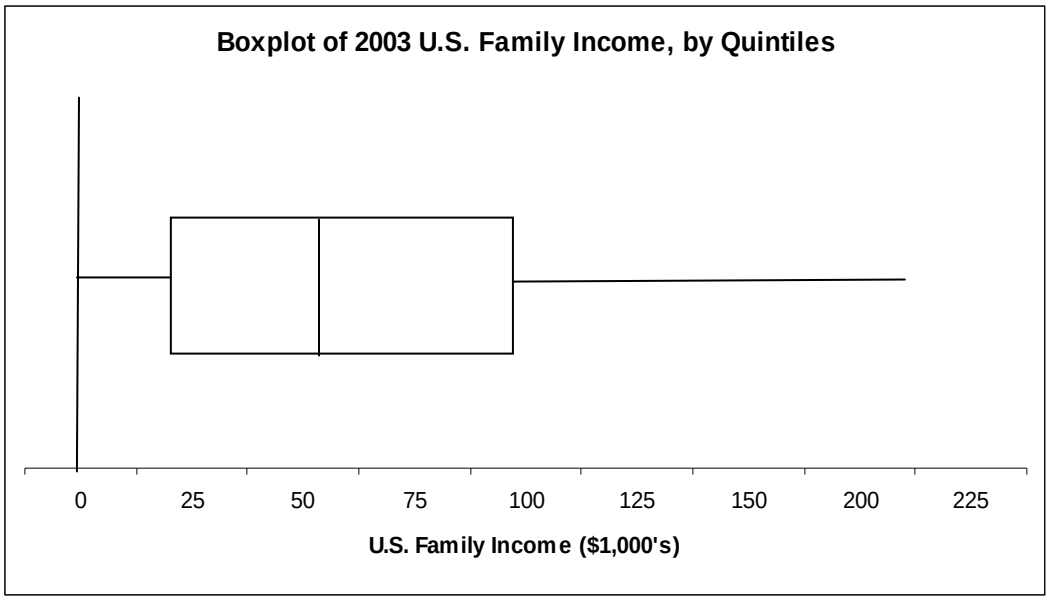


Figure 23 p. 163

U.S. Family Income, by Quintile (Q)	
Q_1 (20%)	24,117
Q_2 (40%)	42,057
Median	52,680
Q_3 (60%)	65,000
Q_4 (80%)	98,200

Quintiles divide data into fifths, or five equal parts. The U.S. Census Bureau reports income data in quintiles (1/5's), not quartiles (1/4's).

Source: http://www.census.gov/hhes/www/income/histinc/f01ar.html



Technology Insert[i]

How to execute concepts using Microsoft® Excel

3.1 & 3.2--"Manual" Calculation of Measures of Central Tendency and Dispersion.

Example--Pulse Rates (in beats per minute).

1. Mean, Median, Mode, Range, Variance, and Standard Deviation.

	B	C	D	E	F
	Student Name	**X=Pulse**	**Mean**	**(X-Mean)**	**(X-Mean)^2**
9	Perpectual	76	72.2	3.8	14.27
10	Megan	60	72.2	-12.2	149.38
11	Jeff	60	72.2	-12.2	149.38
12	Clarice	81	72.2	8.8	77.05
13	Crystal	72	72.2	-0.2	0.05
14	Janette	80	72.2	7.8	60.49
15	Kevin	80	72.2	7.8	60.49
16	Tammy	68	72.2	-4.2	17.83
17	Kathy	73	72.2	0.8	0.60
18					

	Manual Calculations:			**Excel Statistical Functions:**		
20	ΣX	650	=SUM(C9:C17)			
21	No of obs	9	=COUNT(C9:C17)			
22	**Mean**	**72.2**	=C20/C21	**Mean**	**72.2**	=AVERAGE(C9:C17)
23						
24	(N+1)/2 or (n+1)/2	5	=(C21+1)/2			
25	**Median** a/	**73**	=SMALL(C9:C17,C24)	**Median** b/	**73.0**	=MEDIAN(C9:C17)
26						
27						
28	Because manual calculation of the mode involves several steps					
29	and is time consuming, use the Excel **MODE** function. Interpret the MODE as the "most frequent observation."			**Mode** c/	**60.0**	=MODE(C9:C17)
30						
31	Max	81	=MAX(C9:C17)			
32	Min	60	=MIN(C9:C17)	A range function is not available in Excel, so		
33	**Range**	**21**	=C31-C32	you must perform the calculations manually.		
34						
35	Σ(X-Mean)	0.0	=SUM(E9:E17)			
36	Σ(X-Mean)^2	529.56	=SUM(F9:F17)			
37	No of obs	9	=COUNT(F9:F17)			
38	**Variance_Pop**	**58.84**	=C36/C37	**VARP**	**58.84**	=VARP(C9:C17)
39	**Std Dev_Pop**	**7.67**	=C38^0.5	**STDEVP**	**7.67**	=STDEVP(C9:C17)
40	**Variance_Sample**	**66.19**	=C36/(C37-1)	**VAR**	**66.19**	=VAR(C9:C17)
41	**Std Dev_Sample**	**8.14**	=C40^0.5	**STDEV**	**8.14**	=STDEV(C9:C17)

Notes:

a**/ Calculating the Median:**

When the number of observations (n) is **ODD:** (see text, pp. 110-11)

Calculate (n +1)/2 to find the mid-point position in the ordered dataset.

Use the **SMALL** function to find the data value in the (n+1)/2 position.

When the number of observations (n) is **EVEN:** (see Example 3, p. 111))

Calculate the median as the average of the data values in the n/2 and (n/2)+1 positions.

b/ **Calculating the median with the Excel Statistical function, MEDIAN.**

From the Excel menubar use **Insert/Function. S**elect the category Statistical and then select the function, **MEDIAN**.

c/ **Calculating the mode with the Excel Statistical function, MODE.**

From the Excel menubar use **Insert/Function**. Select the category Statistical and then select the function, **MODE**. Be reminded that the **MODE** calculates a single mode. For bimodal or multimodal data, you will need to eliminate the single-mode observations and re-run the **MODE** function to check for additional modes.

[i] The concepts presented in the Technology Insert may be formally demonstrated as time permits.