Chapter 2. Describing, Exploring, and Comparing Data

2-2 Frequency Distributions

In exercises 1–4, identify the class width, class midpoints, and class boundaries for the given frequency distribution based on Data Set 1 in Appendix B.

1. Class width is 10 units of systolic blood pressure of men. This is the difference between two consecutive lower or upper class boundaries such as 100-90 for the two lower class limits of the two lowest intervals or 159-149= 10 for the two upper limits of the two highest intervals.
Class midpoints are the middle points of the intervals. For each interval, they are determined by finding the average of the upper and lower limits of the interval. For the 90 – 99 interval, the midpoint is (90 + 99)/2 = 189/2 = 94.5. For this example, the midpoints are: 94.5, 104.5, 114.5, 124.5, 134.5, 144.5, and 155.5. Note that the difference between adjacent midpoints is the class width of 10.
Class boundaries represent the point between each interval. These are determined by finding the average of the upper class limit of the lower interval and the lower class limit of the higher interval. In this example, the class boundary between the 90-99 and 100-109 intervals would be (99 + 100)/2= 199/2= 99.5 and the series of class boundaries for this example are: 89.5, 99.5, 109.5, 119.5, 129.5, 139.5, 149.5, and 159.5. Note that all the scores fall in the range of the lower class boundary of the lowest interval and the higher class interval of the highest interval (89.5 to 159.5) and they are all 10 (the class width) apart.

Systolic Blood Pressure of Men	Class Boundaries	Class Midpoints	Frequency
90 – 99	89.5 – 99.5	94.5	1
100 – 109	99.5 – 109.5	104.5	4
110 – 119	109.5 – 119.5	114.5	17
120 – 129	119.5 – 129.5	124.5	12
130 – 139	129.5 – 139.5	134.5	5
140 – 149	139.5 – 149.5	144.5	0
150 – 159	149.5 – 159.5	155.5	1

2. Class width is 20 units of systolic blood pressure of women. This is the difference between two consecutive lower or upper class boundaries such as 100-80 for the two lower class limits of the two lowest intervals or 199-179= 20 for the two upper limits of the two highest intervals.

Class midpoints are the middle points of the intervals. For each interval, they are determined by finding the average of the upper and lower limits of the interval. For the 80 – 99 interval, the midpoint is (80 + 99)/2 = 179/2 = 89.5. For this example, the midpoints are: 89.5, 109.5, 129.5, 149.5, 169.5, and 189.5. Note that the difference between adjacent midpoints is the class width of 20.

Class boundaries represent the point between each interval. These are determined by finding the average of the upper class limit of the lower interval and the lower class limit of the higher interval. In this example, the class boundary between the 80-99 and 100-119 intervals would be (99 + 100)/2= 199/2= 99.5 and the series of class boundaries for this example are: 79.5, 99.5, 119.5, 139.5, 159.5, 179.5, and 199.5. Note that all the scores fall in the range of the lower class boundary of the lowest interval and the higher class interval of the highest interval (79.5 to 199.5) and they are all 20 units of systolic blood pressure (the class width) apart.

Systolic Blood Pressure of Women	Class Boundaries	Class Midpoints	Frequency
80 – 99	79.5 – 99.5	89.5	9
100 – 119	99.5 – 119.5	109.5	24
120 – 139	119.5 – 139.5	129.5	5
140 – 159	139.5 – 159.5	149.5	1
160 – 179	159.5 – 179.5	169.5	0
180 – 199	179.5 – 199.5	189.5	1

3. Class width is 200 units of cholesterol of men. This is the difference between two consecutive lower or upper class boundaries such as 0-200 for the two lower class limits of the two lowest intervals or 1200-1000= 200 for the two upper limits of the two highest intervals.

Class midpoints are the middle points of the intervals. For each interval, they are determined by finding the average of the upper and lower limits of the interval. For the 0 – 199 interval, the midpoint is (0 + 199)/2 = 199/2 = 99.5. For this example, the midpoints are: 99.5, 299.5, 499.5, 699.5, 899.5, 1099.5 and 1299.5. Note that the difference between adjacent midpoints is the class width of 200.

Class boundaries represent the point between each interval. These are determined by finding the average of the upper class limit of the lower interval and the lower class limit of the higher interval. In this example, the class boundary between the 0 – 199 and 200 – 399 intervals would be (199 + 200)/2= 399/2= 199.5 and the series of class boundaries for this example are: -0.5, 199.5, 399.5, 599.5, 799.5, 999.5, 1199.5, and 1399.5. Note that all the scores fall in the range of the lower class boundary of the lowest interval and the higher class interval of the highest interval (199.5 to 1399.5) and they are all 200 units of cholesterol (the class width) apart.

Cholesterol of Men	Class Boundaries	Class Midpoints	Frequency
			13
0 – 199	-0.5 – 199.5	99.5	11
200 – 399	199.5 – 399.5	299.5	5
400 – 599	399.5 – 599.5	499.5	8
600 – 799	599.5 – 799.5	699.5	2
800 – 999	799.5 – 999.5	899.5	0
1000 – 1199	999.5 – 1199.5	1099.5	1
1200 – 1399	1199.5 – 1399.5	1299.5	

4. In this case, the observations are not whole numbers, but are fractional ones, measured to the nearest tenth of a score.
Class width is 6 units of body mass units of women. This is the difference between two consecutive lower or upper class boundaries such as 15.0 – 21.0 for the two lower class limits of the two lowest intervals or 44.9 – 38.9= 6 for the two upper limits of the two highest intervals.
Class midpoints are the middle points of the intervals. For each interval, they are determined by finding the average of the upper and lower limits of the interval. For the 15.0 – 20.9 interval, the midpoint is (15.0 + 20.9)/2 = 35.9/2 = 17.95. For this example, the midpoints are: 17.95, 23.95, 29.95, 35.95, and 41.95. Note that the difference between adjacent midpoints is the class width of 6.
Class boundaries represent the point between each interval. These are determined by finding the average of the upper class limit of the lower interval and the lower class limit of the higher interval. In this example, the class boundary between the 15.0-20.9 and 21.0-26.9 intervals would be (20.9 + 21.0)/2= 41.9/2= 20.95 and the series of class boundaries for this example are: 14.95, 20.95, 26.95, 32.95, 38.95, and 44.95. Note that all the scores fall in the range of the lower class boundary of the lowest interval and the higher class interval of the highest interval (14.95 to 44.95) and they are all 6 units of body mass index (the class width) apart.

Body Mass Index of Women	Class Boundaries	Class Midpoints	Frequency
15.0 – 20.9	14.95 – 20.95	17.95	10
21.0 – 26.9	20.95 – 26.95	23.95	15
27.0 – 32.9	26.95 – 32.95	29.95	11
33.0 – 38.9	32.95 – 38.95	35.95	2
39.0 – 44.9	38.95 – 44.95	41.95	2

In exercises 5–8, construct the relative frequency distributions that correspond to the frequency distribution in the exercise indicated.

The relative frequency is the proportion, usually converted to a percent, of scores in each interval, the equation to be used is:

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

5. For **Systolic Blood Pressure of Men** (Exercise 1), the sum of frequencies (n) is 40
For the 90 – 99 interval, relative frequency = 1/40 * 100% = 0.025 * 100%= 2.5%
For the 120 – 129 interval, relative frequency = 12/40 * 100% = 0.30 * 100%= 30.0%

Systolic Blood Pressure of Men	Frequency	Relative Frequency
90 – 99	1	2.5%
100 – 109	4	10.0%
110 – 119	17	42.5%
120 – 129	12	30.0%
130 – 139	5	12.5%
140 – 149	0	0.0%
150 – 159	1	2.5%

6. For **Systolic Blood Pressure of Women** (Exercise 2), the sum of frequencies (n) is 40
For the 80 – 99 interval, relative frequency = 9/40 * 100% = 0.225 * 100%= 22.5%
For the 100 – 119 interval, relative frequency = 24/40 * 100% = 0.60 * 100%= 60.0%

Systolic Blood Pressure of Women	Frequency	Relative Frequency
80 – 99	9	22.5%
100 – 119	24	60.0%
120 – 139	5	12.5%
140 – 159	1	2.5%
160 – 179	0	0.0%
180 – 199	1	2.5%

7. For **Cholesterol of Men** (Exercise 3), the sum of frequencies (n) is 40.
For the 200 – 399 interval, relative frequency is 11/40 * 100% = 0.275 * 100%= 27.5%
For the 1200 – 1399 interval, relative frequency is 1/40 * 100% = 0.025 * 100= 2.50%

Cholesterol of Men	Frequency	Relative Frequency
0 – 199	13	32.5%
200 – 399	11	27.5%
400 – 599	5	12.5%
600 – 799	8	20.0%
800 – 999	2	5.0%
1000 – 1199	0	0.0%
1200 – 1399	1	2.5%

8. For **Body Mass Index of Women** (Exercise 4), the sum of frequencies (n) is 40
For the 15.0 – 20.9 interval, relative frequency is 10/40 * 100% = 0.25 * 100%= 25.0%
For the 33.0 – 38.9 interval, relative frequency is 2/40 * 100% = 0.05 * 100%= 5.0%

Body Mass Index of Women	Frequency	Relative Frequency
15.0 – 20.9	10	25.0%
21.0 – 26.9	15	37.5%
27.0 – 32.9	11	27.5%
33.0 – 38.9	2	5.0%
39.0 – 44.9	2	5.0%

In exercises 9-12, construct the cumulative frequency distribution that corresponds to the frequency distribution in the exercise indicated.

The cumulative frequency in each interval is the sum of the frequencies in the interval and all of the frequencies in the intervals below (in score terms) that interval. Cumulative frequency in the highest score interval must equal the sum of all frequencies (n).

9. For **Systolic Blood Pressure of Men** (Exercise 5)

Systolic Blood Pressure of Men	Frequency	Working column (not in final table) f in interval + cumulative f up to interval	Cumulative Frequency
90 – 99	1	1 + 0 = 1	1
100 – 109	4	4 + 1 = 5	5
110 – 119	17	17 + 5 = 22	22
120 – 129	12	12 + 22 = 34	34
130 – 139	5	5 + 34 = 39	39
140 – 149	0	0 + 39 =39	39
150 – 159	1	39 + 1 = 40	40

10. For **Systolic Blood Pressure of Women** (Exercise 6)

Systolic Blood Pressure of Women	Frequency	Working column (not in final table) f in interval + cumulative f up to interval	Cumulative Frequency
80 – 99	9	9 + 0 = 9	9
100 – 119	24	24 + 9= 33	33
120 – 139	5	5 + 33 = 38	38
140 – 159	1	1 + 38= 39	39
160 – 179	0	0 + 39= 39	39
180 – 199	1	1 + 39 = 40	40

11. For **Cholesterol of Men** (Exercise 7)

Cholesterol of Men	Frequency	Working column (not in final table) f in interval + cumulative f up to interval	Cumulative Frequency
0 – 199	13	0 + 13 = 13	13
200 – 399	11	11 + 13 = 24	24
400 – 599	5	5 + 24= 29	29
600 – 799	8	8 + 29 = 37	37
800 – 999	2	2 + 37 = 39	39
1000 – 1199	0	0 + 39 = 39	39
1200 – 1399	1	1 + 39 = 40	40

12. For **Body Mass Index of Women** (Exercise 8)

Body Mass Index of Women	Frequency	Working column (not in final table) f in interval + cumulative f up to interval	Cumulative Frequency
15.0 – 20.9	10	10 + 0 = 10	10
21.0 – 26.9	15	15 + 10 = 25	25
27.0 – 32.9	11	11 + 25= 36	36
33.0 – 38.9	2	2 + 36 = 38	38
39.0 – 44.9	2	2 + 38 = 40	40

13. **Bears** Frequency Distribution of Weight of Bears in Pounds (From Data Set 6)

Weight of Bears in Pounds	Frequency
0 – 49	6
50 – 99	10
100 –149	10
150 –199	7
200 – 249	8
250 – 299	2
301 – 349	4
350 – 399	3
400 – 449	3
450 – 499	0
500 – 549	1

14. Body Temperatures (From Data Set 2)

Body Temperature at Midnight of Second Day	Frequency
96.5 – 96.8	1
96.9 – 97.2	8
97.3 – 97.6	14
97.7 – 98.0	22
98.1 – 98.4	19
98.5 – 98.8	32
98.9 – 99.2	6
99.3 – 99.6	4

Two notable things about this frequency distribution are: 1. that the scale of the classes does not include all of the scores since there are four scores that are not included in the range of classes selected and 2. The intervals could have been made shorter (say 3^{o}F) so the distribution could have been spread out more. With this many scores (*n*= 106), more intervals might have given a better detail of the nature of the distribution.

15. Head Circumferences in cm. of Two-Month-Old Babies (From Data Set 4)

Head Circumference in cm.	Frequency of Males	Frequency of Females
34.0 – 35.9	2	1
36.0 – 37.9	0	3
38.0 – 39.9	5	14
40.0 – 41.9	29	27
42.0 – 43.9	14	5

From the comparison of the two frequency distributions, it seems that the head circumference for the male babies is higher than for the female babies. However, we cannot conclude that the difference is statistically significant until we conduct specific tests to test these differences.

16. Very Poplar Trees Weights in kg. (From Data Set 9)

Poplar Weights in kg. for Year 2	Frequency
0.00 - 0.49	7
0.50 - 0.99	15
1.00 - 1.49	10
1.50 - 1.99	7
2.00 - 2.49	1

17. Yeast Cell Counts (From Data Set 11)

Yeast Cell Counts	Frequency
1	20
2	43
3	53
4	86
5	70
6	54
7	37
8	18
9	10
10	5
11	2
12	2

The most common cell count is 4 with a frequency of 86.

18. Number of Classes Using Sturges' guideline**,** n represents the number of data values and x represents the number of classes

$$x = 1 + (\log n)/(\log 2)$$

$$n = 2^{x-1}$$

We will need to find values where there are lower and upper limits associated with determined numbers of data values. If we are looking for values related to having 5 to 12 classes, we need an upper and lower limit for each of these. For 5 this will be 4.5 to 5.5, for 6 this will be 5.5 to 6.5, and so on until the last category would have 12 classes and 11.5 to 12.5 will be used for the limits. The number of values at each limit is determined by:

$n = 2^{x-1}$ where x are the limits.

We can display this in the following table:

Number of classes (x)	Lower and Upper x Limits	Number of Values at Lower and Upper Limits $n = 2^{x-1}$	Number of Values (Rounded)
5	4.5 – 5.5	11.31 – 22.63	11 – 22
6	5.5 – 6.5	22.63 – 45.25	23 – 45
7	6.5 – 7.5	45.25 – 90.51	46 – 90
8	7.5 – 8.5	90.51 – 181.02	91 – 181
9	8.5 – 9.5	181.02 – 362.04	182 – 362
10	9.5 – 10.5	362.04 – 724.08	363 – 724
11	10.5 – 11.5	724.08 – 1448.15	725 – 1448
12	11.5 – 12.5	1448.15 – 2896.31	1449 - 2896

2-3 Visualizing Data

In Exercises 1-4, answer the questions by referring to the SPSS-generated histogram given below. The histogram represents the lengths (mm) of cuckoo eggs found in the nests of other birds (based on data from O.M. Latter and Data and Story Library).

1. **Center** Center is a very general term, but basically it looks for a central value that could be used to represent the entire distribution. One value that could be used is the center of the scale which would be at the average of the low (19.75) and high (24.76) points, which would be (19.75 + 24.76)/2= 22.25 mm of egg length. Another possible value for center would be the value that occurs most often which would be taken as the midpoint of the 22.0 interval. One statistic that is reported in the table is the mean of 22.46 mm. This is one of the descriptive statistics that is often used to represent the center of a continuous distribution.

2. **Variation** For these data, the lowest possible egg length would be the lowest score limit in the lowest interval. That interval is the 19.625 to 19.875 interval and the lowest possible score in that interval is 19.625 mm. The highest possible egg length would be the highest score limit in the highest interval. That interval is the 24.875 to 25.125 interval and the highest possible score in that interval is 25.125 mm. Thus, the range of observed egg values is 19.625 to 25.125.

3. **Percentage** Note the midpoints of the intervals are not all labeled. There are three intervals that have midpoints of 20.75, 21.00, and 21.25 (21.00 is not printed, but it is there). The lower and upper limits of these three intervals are: 20.625 – 20.875, 20.875 – 21.125, and 21.125 – 21.375. 21.125 is the upper limit of the 20.875 to 21.125 interval. Therefore, all of the eggs in that interval and below are less than 21.125 mm. Counting them from the start up though that interval, we have 2 + 2 + 1 + 0 + 6 + 6 = 17 and that equates to a percentage of 17/120 (100%) = 14.2%.

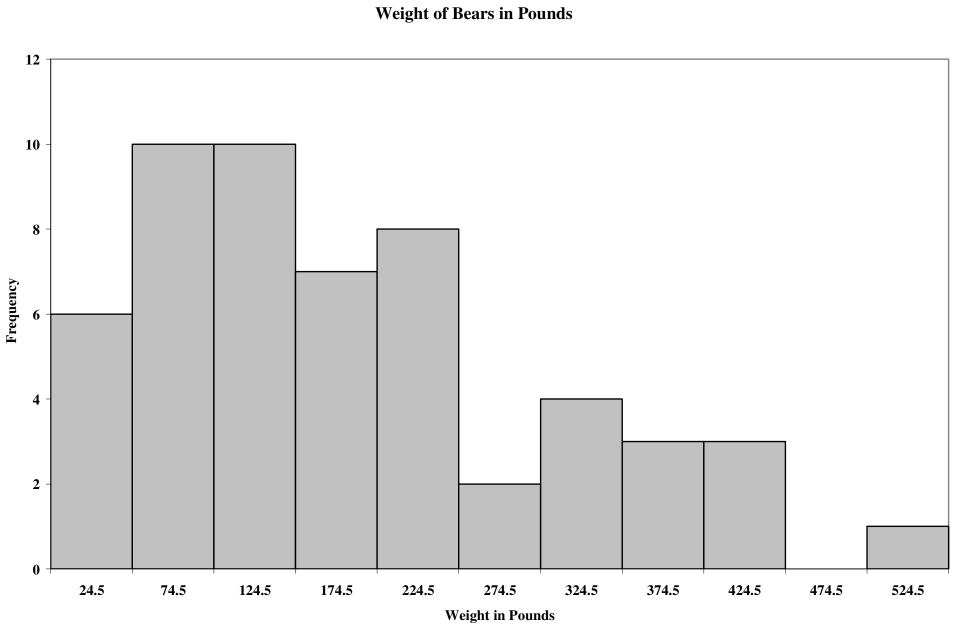
4. **Class width** The easiest way to find this is to find the difference between midpoints of two adjacent intervals. As we saw in exercise three, there were three adjacent intervals with midpoints of 20.75, 21.00, and 21.25. Each of these midpoints is 0.25 mm different from adjacent intervals, the class width is 0.25 mm.

In Exercises 5 and 6, refer to the accompanying pie chart of blood groups for a large sample of people (based on data from the Greater New York Blood Program).

5. **Interpreting Pie Chart** It would appear that about 40% of the observed group has Group A blood and out of 500 in the sample, this would be about 200 people.

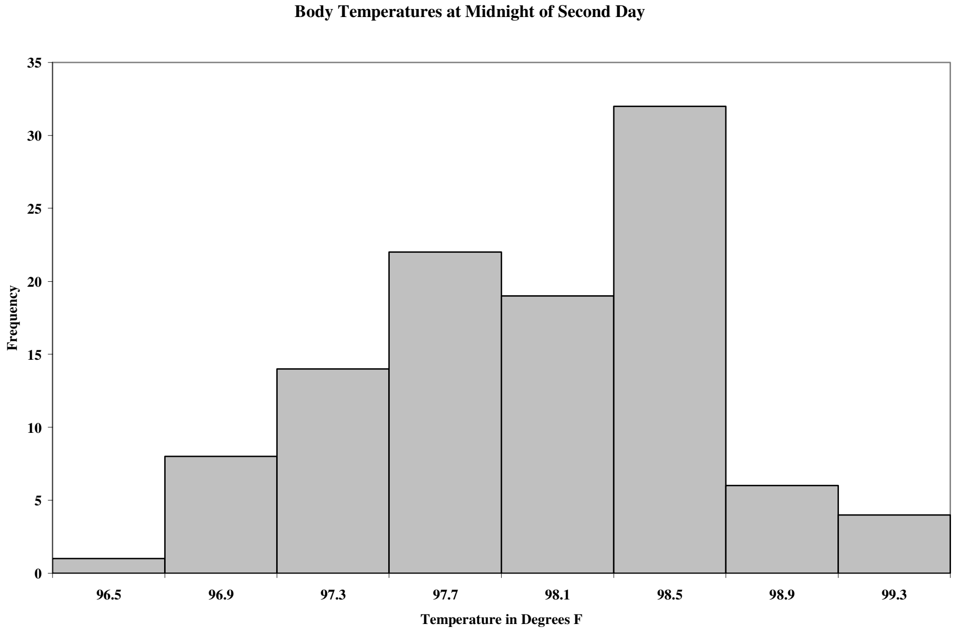
6. **Interpreting Pie Chart** It appears that about 10% of the observed group has Group B blood and out of 500 in the sample, this would be about 50 people.

7. **Bears** Histogram of Weights in Pounds



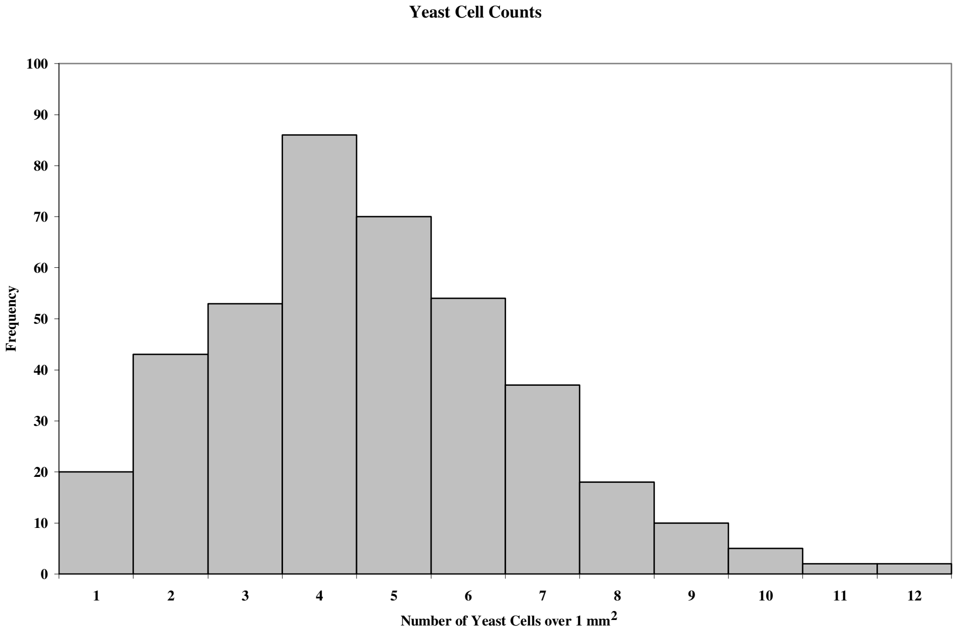
The approximate weight at the center seems to be between 174.5 and 224.5 or about 200 pounds. We could compute the mean and find it is 183 pounds, which would be a useful measure of centrality of the distribution.

8. **Body Temperatures** in Degrees Fahrenheit Histogram (Data Set 2)



98.6^{o}F seems close to where we would expect the mean of this distribution to be. However, there is some variation around that value. This does not appear to be a bell-shaped distribution. It is skewed toward the lower end of the scale, which we would describe as being negatively skewed.

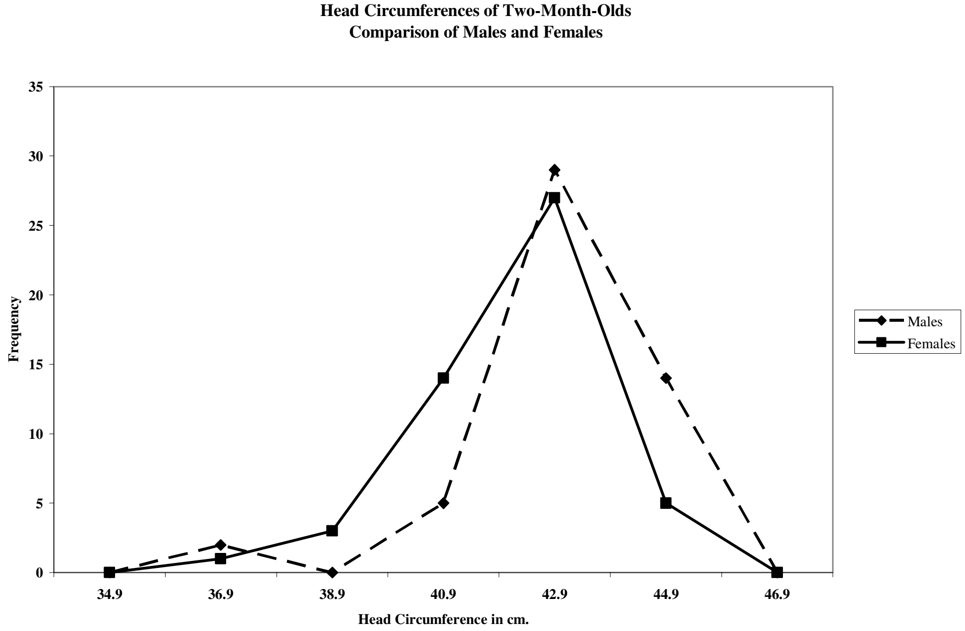
9. Yeast Cell Counts Histogram (Data Set 11)



The distribution seems to depart somewhat from being bell-shaped. It seems to have a reasonable degree of positive skewness. Other specific measures can be used to assess the exact degree of skewness.

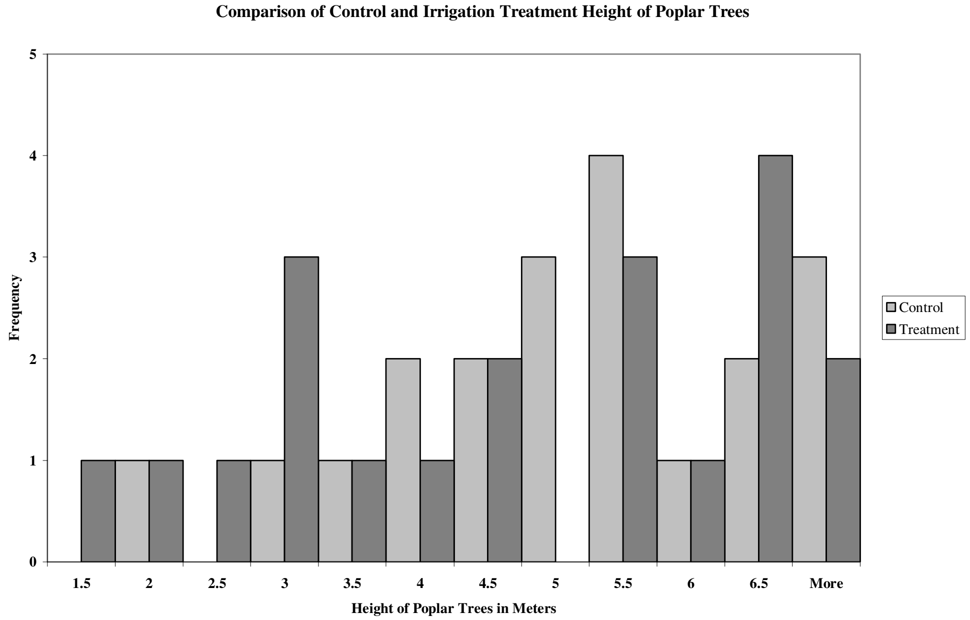
In Exercises 10 and 11, use graphs to compare the two indicated data sets.

10. Head Circumferences (Data Set 4), Frequency Polygons for Males and Females



From the frequency polygon, it would appear that the head circumference for males is slightly higher than that for females. However, to determine statistical significance, we would need to conduct other tests designed to aid in making those decisions.

11. Poplar Trees Histogram comparing Heights (in meters) of Control Group and Irrigation Only Trees (From Data Set 9)



The two distributions do not seem to have much departure from each other and this would not seem to be enough for us to conclude that the distributions are significantly different. There is no sound evidence to conclude that the irrigation treatment had any effect on poplar tree height.

In Exercises 12 and 13, list the original data represented by the given stem-and-leaf plots.

12. To find the original score, the leaf (in units of one) is added to the stem which is in units of tens. The original data displayed in stem-and-leaf plot would be:

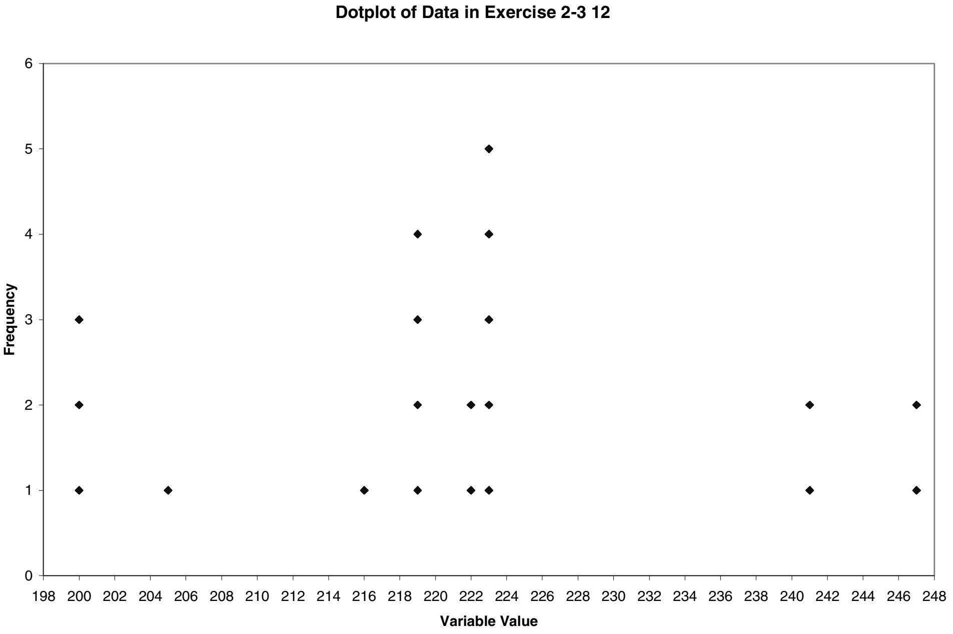
200 200 200 205 216 219 219 219 219 222 222 223 223 223 223 223
241 241 247 247

13. To find the original scores, the leaf in units of tens and ones is added to the stem which is in units of hundreds.

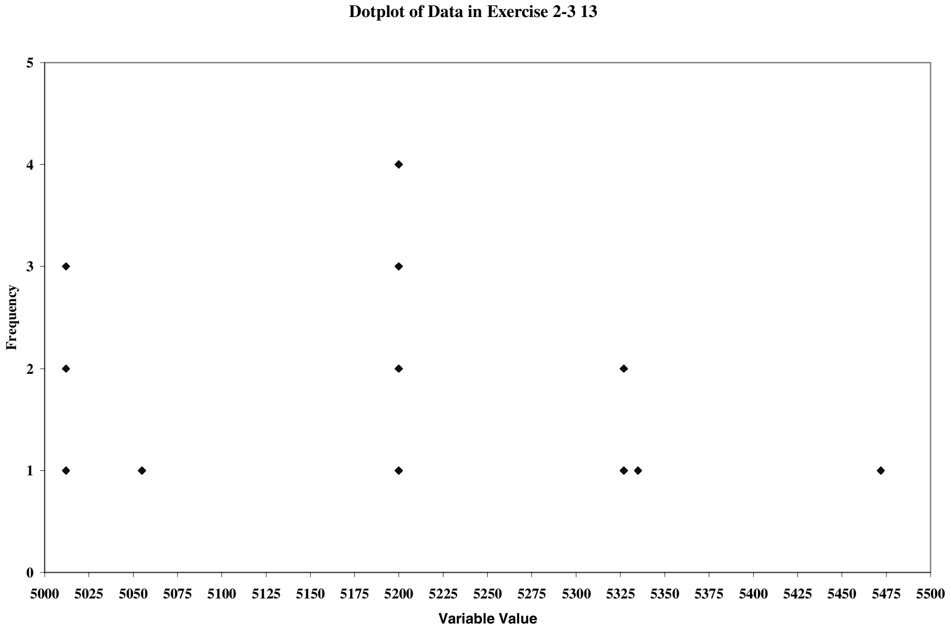
5012 5012 5012 5055 5200 5200 5200 5200 5327 5327 5335 5472

In Exercises 14 and 15, construct the dotplot for the data represented by the stem-and-leaf plot in the given exercise.

14. Dotplot of Exercise 12 Data



15. Dotplot of Exercise 13 Data



16. Construct **Stem-and-Leaf Plot** for Exercise 11.

Data points are:

4.1 2.5 4.5 3.6 2.8 5.1 2.9 5.1 1.2 1.6 6.4 6.8 6.5 6.8 5.8 2.9 5.5 3.3 6.1 6.1

First, put them in order:
1.2 1.6 2.5 2.8 2.9 2.9 3.3 3.6 4.1 4.5 5.1 5.1 5.5 5.8 6.1 6.1 6.4 6.5 6.8 6.8

Stem will be in units of ones and leaf will be in units of tenths, need stems that range from 1 to 6

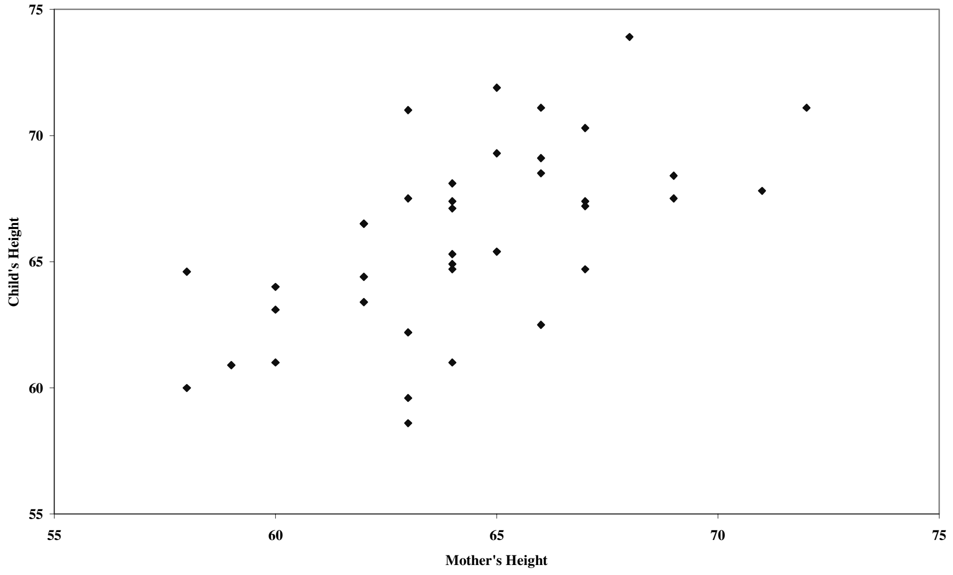
Stem	Leaves
(units)	(tenths)
1.	26
2.	5899
3.	36
4.	15
5.	1158
6.	114588

In Exercises 17-20, use the given paired data from Appendix B to construct a scatter diagram.

17. Parent/Child Height (Data Set 3)

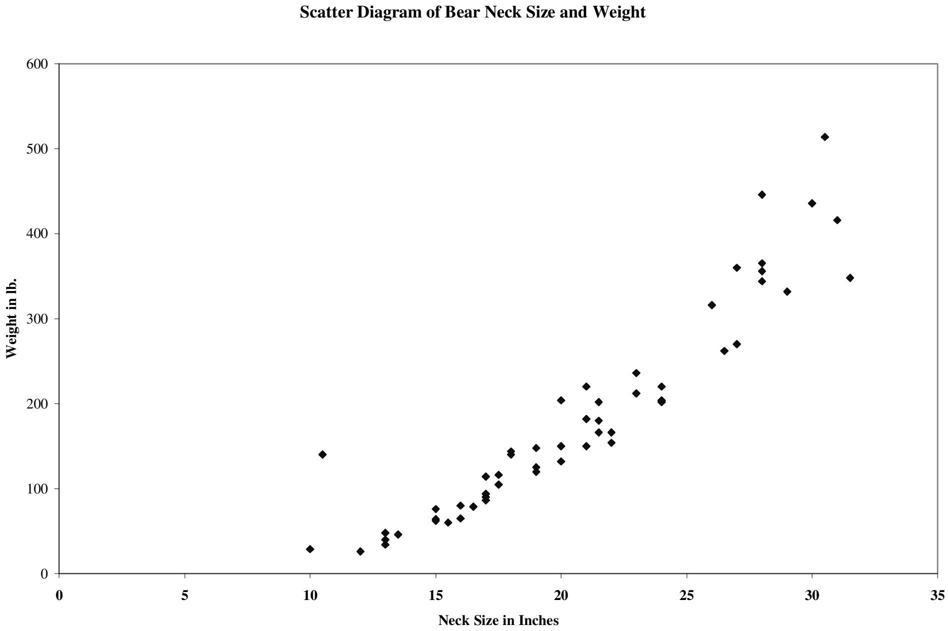
To determine a relationship graphically, we will use a scatter diagram with mother's height on the horizontal (x-axis) scale and the child's height on the vertical (y-axis) scale.

Scatter Diagram of Child and Mother Heights, in Inches



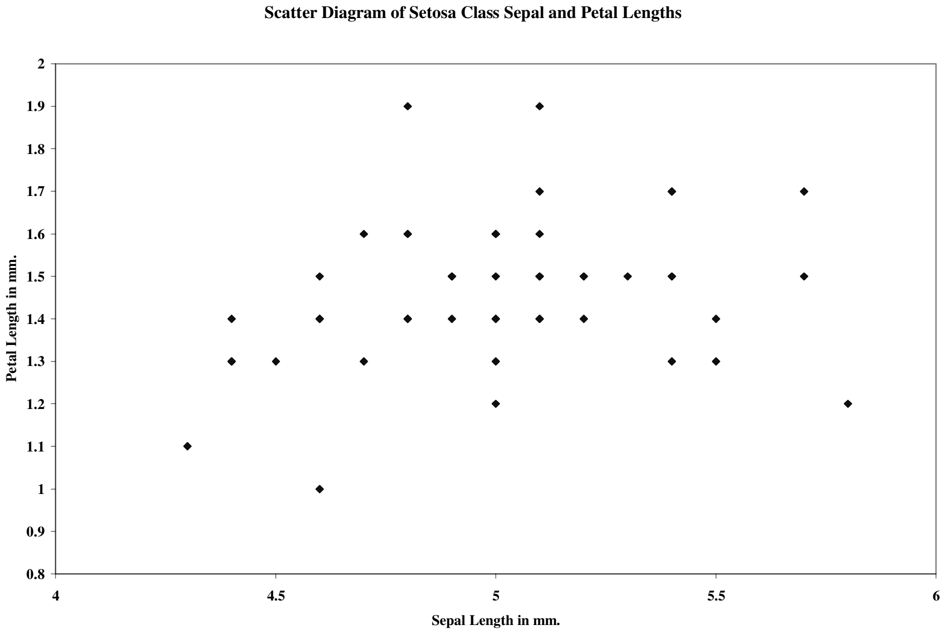
There appears to be a relationship and the nature of the relationship is that low heights on both cluster together and high heights on both cluster together or another way of saying this is that as height on one increases we observe height on the other increasing, also called a direct or positive relationship.

18. Bear Neck/Weight (Data Set 6)



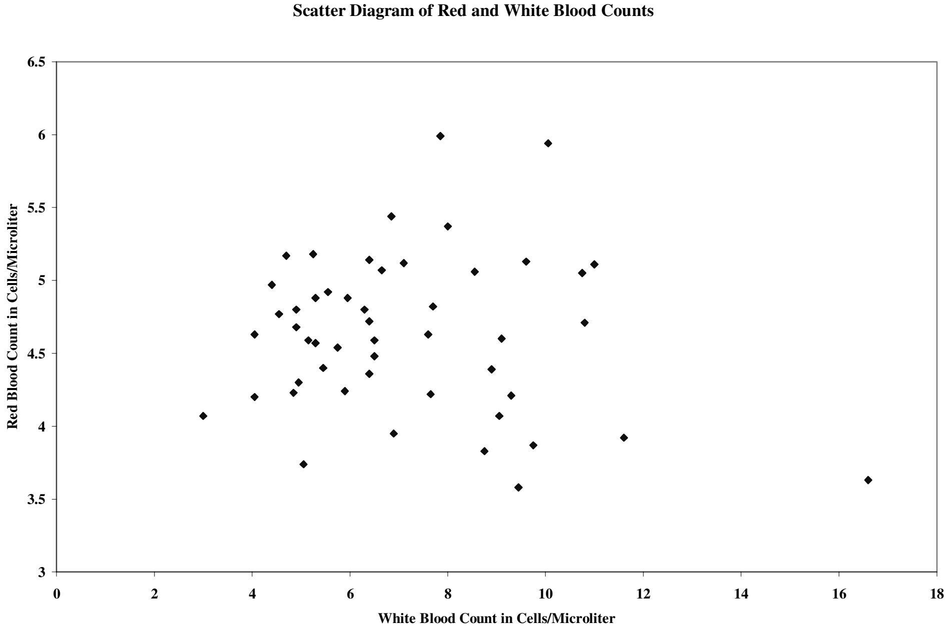
There is a general positive relationship. As neck size increases, we observe an increase in weight.

19. Sepal/Petal Length (Data Set 7)



There does not appear to be a relationship between sepal length and petal length for the Setosa Class. As one of the variables changes in a given direction, there seems to be no corresponding change in the other variable in either direction.

20. Red and White Blood Counts (Data Set 10)



There does not appear to be a relationship between red and white blood count. As one of the variables changes in a given direction, there seems to be no corresponding change in the other variable in either direction.

2-4 Measures of Center

In Exercises 1-8, find the (a) mean, (b) median, (c) mode, and (d) midrange for the given sample data.

The four measures of center used below are:

a. the mean ($\bar{x}$) which we find using the following equation:

Mean: $\bar{x} = \sum x/n$ where $\sum x$ represents the sum of scores

b. the median ($\tilde{x}$) which is determined with the following rules:

If there are an odd number of scores, the median is the score that is exactly in the middle of the ordered set of scores.

If there are an even number of scores, the median is the point exactly between or the average of the two middle scores.

c. the mode, which is the score that occurs most often; if there are two that occur most often that are tied, the distribution is bimodal and two modes are reported; if there are three that occur most often that are tied, the distribution is trimodal and three modes are reported; if no score occurs more often than any others or if there are more than three that occur most often and they are tied, no mode is reported.

d. the midrange is the point exactly between or the average of the lowest and highest scores

1. Tobacco Use in Children's Movies Arranged in order, the scores are:

#	1	2	3	4	5	6
Score	0	0	0	176	223	548

a. Mean: $\bar{x} = \sum x/n = 947/6 = 157.8$

b. Median: $\tilde{x}$ is the midpoint between the two middle scores (3rd and 4th scores) if an even number of scores. This would be (0 + 176)/2= 176/2= 88.0

c. Mode: Mode is score that occurs most often, this is 0 for these data

d. Midrange: Midpoint (or average) between lowest and highest scores, (548 + 0)/2 = 548/2=
274.0

2. Cereal Arranged in order, the scores are

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Score	.03	.07	.09	.13	.13	.17	.24	.30	.39	.43	.43	.44	.45	.47	.47	.48

a. Mean: $\bar{x} = \sum x / n = 4.72/16 = 0.295$

b. Median: $\tilde{x}$ is the midpoint between the two middle scores (8th and 9th scores) if an even number of scores. This would be (0.30 + 0.39)/2= 0.69/2= 0.345

c. Mode: Mode is score that occurs most often, there are three scores that occur most often so the distribution is tri-modal at 0.13, 0.43, and 0.47

d. Midrange: Midpoint (or average) between lowest and highest scores, (0.48 + 0.03)/2 = 0.51/2= 0.255. Is the mean likely to be a good estimate of the mean amount of sugar in each gram of cereal consumed by the population of all Americans who eat cereal? The sample mean is always the best estimate of the population mean. However, since the distribution is skewed (negatively), the median might be a better center estimate of the amount of sugar in each gram of cereal consumed by the population of all Americans who eat cereal.

3. Body Mass Index Arranged in order, the scores are:

#	1	2	3	4	5	6	7	8	9	10	11	12
Score	17.7	19.6	19.6	20.6	21.4	22.0	23.8	24.0	25.2	27.5	28.9	29.1
#	13	14	15									
Score	29.9	33.5	37.7									

a. Mean: $\bar{x} = \sum x / n = 380.5/15 = 25.37$

b. Median: $\tilde{x}$ is the middle score (8th score) if an odd number of scores. This would be 24.0

c. Mode: Mode is score that occurs most often, the only score that occurs more than once is 19.6 so that is the mode

d. Midrange: Midpoint (or average) between lowest and highest scores, (37.7 + 17.7)/2 = 20.0/2= 27.7
Is the mean of this sample reasonably close to the mean of 25.74, which is the mean for all 40 women included in Data Set 1? Yes, they are very close, especially when on considers the range of values to be from about 17 to 38.

4. Drunk Driving Arranged in order, the scores are:

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Score	.12	.13	.14	.16	.16	.16	.17	.17	.17	.18	.21	.24	.24	.27	.29

a. Mean: $\bar{x} = \sum x / n = 2.81/15 = 0.187$

b. Median: $\tilde{x}$ is the middle score (8th score) if an odd number of scores. This would be 0.17

c. Mode: Mode is score that occurs most often. In this case two scores occur three times (0.16 and 0.17) and since these are adjacent to each other in the order, the mode is taken as the midpoint between them, so the mode in this case would be 0.165)

d. Midrange: Midpoint (or average) between lowest and highest scores, (0.29 + 0.12)/2 = 0.41/2= 0.205

5. Motorcycle Fatalities Arranged in order, the scores are:

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Score	14	16	17	18	20	21	23	24	25	27	28	30	31	34	37	38	40	42

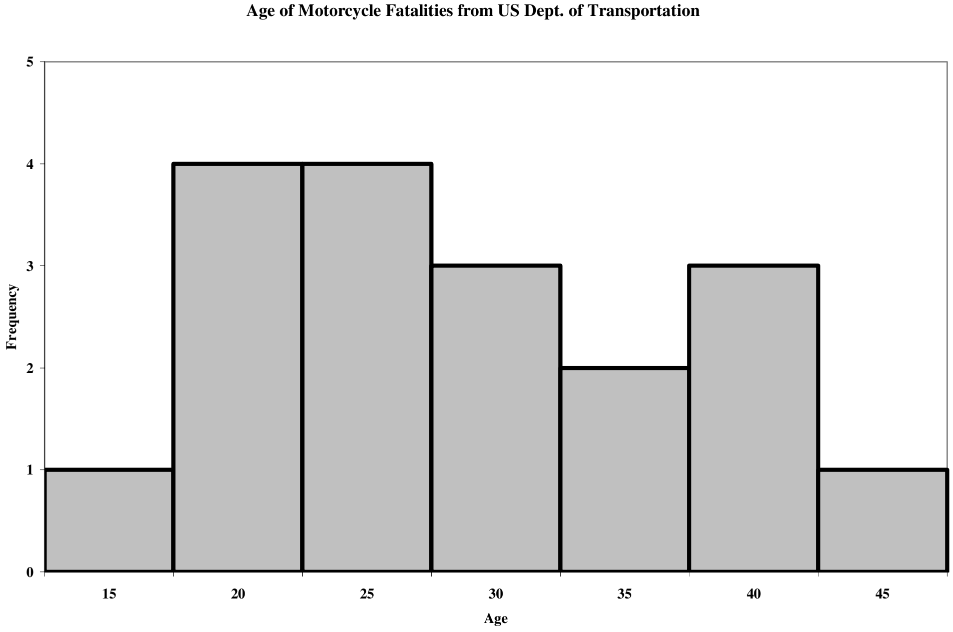
a. Mean: $\bar{x} = \sum x / n = 485/18 = 26.9$

b. Median: $\tilde{x}$ is the average of the middle scores (9th and 10th scores) if an even number of scores. This would be (25 + 27)/2= 52/2= 26.0

c. Mode: Mode is score that occurs most often. In this case no score occurs more than once so there is no mode

d. Midrange: Midpoint (or average) between lowest and highest scores, (14 + 42)/2 = 56/2= 28.0

Do the results support the common belief that such fatalities are incurred by a greater proportion of younger drivers? Following is a histogram of the frequencies:



In the range of ages of fatalities for these data, the distribution is fairly even from ages 14 to 42, with a slightly higher distribution toward the lower range, but not such a distinct difference that would warrant concluding that such fatalities are incurred by a greater proportion of younger drivers. It is difficult to answer this question since we don’t know and might reasonable question whether there would be an equal proportion of motorcycle riders across all the age groups. Also, this data set is relatively small to permit making such a conclusion.

6. **Fruit Flies** Arranged in order, the scores are:

#	1	2	3	4	5	6	7	8	9	10	11
Score	0.64	0.68	0.72	0.76	0.84	0.84	0.84	0.84	0.90	0.90	0.92

a. Mean: $\bar{x} = \sum x / n = 8.88 / 11 = 0.807$

b. Median: $\tilde{x}$ is the middle score (6^{th}) if an odd number of scores. This would be 0.84

c. Mode: Mode is score that occurs most often. In this case the mode is 0.84 since that occurs most often

d. Midrange: Midpoint (or average) between lowest and highest scores, (0.64 + 0.92)/2 = 1.56/2= 0.78

7. **Blood Pressure Measurements** Arranged in order, the scores are:

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Score	120	120	125	130	130	130	130	135	138	140	140	143	144	150

a. Mean: $\bar{x} = \sum x / n = 1875 / 14 = 133.9$

b. Median: $\tilde{x}$ is the midpoint between the two middle scores (7^{th} and 8^{th} scores) if an even number of scores. This would be (130 + 135)/2= 265/2= 132.5

c. Mode: Mode is score that occurs most often. In this case the mode is 130 since that occurs most often (4 times)

d. Midrange: Midpoint (or average) between lowest and highest scores, (120 + 150)/2 = 270/2= 135
What is notable about this data set? The measures of centrality are relatively close, which would indicate the distribution seems relatively symmetrical.

8. Phenotypes of Peas Arranged in order, the scores are:

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Score	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
#	16	17	18	19	20	21	22	23	24	25					
Score	2	2	2	3	3	3	3	3	3	4					

a. Mean: $\bar{x} = \sum x / n = 47 / 25 = 1.88$

b. Median: $\tilde{x}$ is the middle score (13^{th}) if an odd number of scores. This would be 2.0

c. Mode: Mode is score that occurs most often. In this case, 1 occurs more than any other score (11 times) so the mode is 1.0

d. Midrange: Midpoint (or average) between lowest and highest scores, (1 + 4)/2 = 5/2= 2.5
The level of measurement is ordinal since the categories can be ranked or ordered. Measure of center can be obtained for these values. Do the results make sense? While central measures are often used with ordinal data, the fact that they do not have equal intervals would make the mean of questionable use, but the median and mode might be useful.

In Exercises 9-12, find the mean, median, mode, and midrange for each of the samples then compare the two sets of results.

9. Patient Waiting Times

Single Line Central Statistics: Arranged in order, the scores are:

#	1	2	3	4	5	6	7	8	9	10
Score	65	66	67	68	71	73	74	77	77	77

a. Mean: $\bar{x} = \sum x/n = 715/10 = 71.5$

b. Median: $\tilde{x}$ is the midpoint between the two middle scores (5th and 6th scores) if an even number of scores. This would be (71 + 73)/2= 144/2= 72.0

c. Mode: Mode is score that occurs most often. In this case the mode is 77 since that occurs most often

d. Midrange: Midpoint (or average) between lowest and highest scores, (65 + 77)/2 = 142/2= 71.0

Multiple Line Central Statistics: Arranged in order, the scores are:

#	1	2	3	4	5	6	7	8	9	10
Score	42	54	58	62	67	77	77	85	93	100

a. Mean: $\bar{x} = \sum x/n = 715/10 = 71.5$

b. Median: $\tilde{x}$ is the midpoint between the two middle scores (5th and 6th scores) if an even number of scores. This would be (67 + 77)/2= 144/2= 72.0

c. Mode: Mode is score that occurs most often. In this case the mode is 77 since that occurs most often

d. Midrange: Midpoint (or average) between lowest and highest scores, (42 + 100)/2 = 142/2= 71

Comparison of the two distributions relative to central measures: The central measures are exactly the same for both groups. However, there is more spread of the scores from each other for the multiple line group (42 to 100) compared with the single line group (65 to 77).

10. Skull Breadths

4000 B. C. Central Statistics: Arranged in order, the scores are:

#	1	2	3	4	5	6	7	8	9	10	11	12
Score	119	125	126	126	128	128	129	131	131	131	132	138

a. Mean: $\bar{x} = \sum x/n = 1544/12 = 128.7$

b. Median: $\tilde{x}$ is the midpoint between the two middle scores (6th and 7th scores) if an even number of scores. This would be (128 + 129)/2= 257/2= 128.5

c. Mode: Mode is score that occurs most often. In this case the mode is 131 since that occurs most often (3 times)

d. Midrange: Midpoint (or average) between lowest and highest scores, (119 + 138)/2 = 257/2= 128.5

150 A. D. Central Statistics: Arranged in order, the scores are:

#	1	2	3	4	5	6	7	8	9	10	11	12
Score	126	126	129	130	131	133	134	136	137	138	139	141

a. Mean: $\bar{x} = \sum x / n = 1600/12 = 133.3$

b. Median: $\tilde{x}$ is the midpoint between the two middle scores (6th and 7th scores) if an even number of scores. This would be (133 + 134)/2= 267/2= 133.5

c. Mode: Mode is score that occurs most often. In this case the mode is 126 since that occurs most often (2 times)

d. Midrange: Midpoint (or average) between lowest and highest scores, (126 + 141)/2 = 267/2= 133.5

Comparison of the two distributions relative to central measures: The central measures are slightly higher on all central measures except the mode (which tends to be less stable as a measure compared with the others) for the 150 A. D. sample than for the 4000 B.C. group. Head sizes appear to have gotten larger between 4000 B C. and 150 A. D. However, it is not valid to attribute such change to interbreeding with people from other regions. These data provide no basis for making such a conclusion as to any cause for this change.

11. Poplar Trees

Control Group Central Statistics: Arranged in order, the scores are:

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Score	1.9	2.9	3.2	3.6	3.9	4.1	4.1	4.6	4.8	4.9	5.1	5.5	5.5	5.5	6.0
#	16	17	18	19	20										
Score	6.3	6.5	6.8	6.9	6.9										

a. Mean: $\bar{x} = \sum x / n = 99.0/20 = 4.95$

b. Median: $\tilde{x}$ is the midpoint between the two middle scores (10th and 11th scores) if an even number of scores. This would be (4.9 + 5.1)/2= 10.0/2= 5.0

c. Mode: Mode is score that occurs most often. In this case, 5.5 occurs more than any other score (3 times) so the mode is 5.5

d. Midrange: Midpoint (or average) between lowest and highest scores, (1.9 + 6.9)/2 = 8.8/2= 4.4

Irrigation Treatment Group Central Statistics: Arranged in order, the scores are:

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Score	1.2	1.6	2.5	2.8	2.9	2.9	3.3	3.6	4.1	4.5	5.1	5.1	5.5	5.8	6.1
#	16	17	18	19	20										
Score	6.1	6.4	6.5	6.8	6.8										

a. Mean: $\bar{x} = \sum x / n = 89.6/20 = 4.48$

b. Median: $\tilde{x}$ is the midpoint between the two middle scores (10th and 11th scores) if an even number of scores. This would be (4.5 + 5.1)/2= 9.6/2= 4.80

c. Mode: Mode is score that occurs most often. In this case, 4 scores occur the same number of times (twice) so the distribution has four modes with modes at: 2.9, 5.1, 6.1 and 6.8 occurs more than any other score (3 times) so there is no distinct mode

d. Midrange: Midpoint (or average) between lowest and highest scores, (1.2 + 6.8)/2 = 8.0/2= 4.0

Comparison of the two distributions relative to central measures: The central measures are slightly higher on the mean and the median for the control group on the heights (in meters) of the trees.

12. Poplar Trees

The breakout of the stem and leaf plot provides the score data which are arranged in order below:

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Score	3.2	3.9	4.4	5.4	6.3	6.3	6.4	6.7	6.8	6.8	6.9	6.9	6.9	7.1	7.3
#	16	17	18	19	20										
Score	7.3	7.3	7.6	7.7	8.0										

a. Mean: $\bar{x} = \sum x / n = 129.2 / 20 = 6.46$ (Control Group Mean= 4.95)

b. Median: $\tilde{x}$ is the midpoint between the two middle scores (10th and 11th scores) if an even number of scores. This would be (6.8 + 6.9)/2= 13.7/2= 6.85 (Control Group Median= 5.0)

c. Mode: Mode is score that occurs most often (4 times). In this case, 2 scores occur more than any others so there are two modes: 6.9 and 7.3 (Control Group Mode= 5.5)

d. Midrange: Midpoint (or average) between lowest and highest scores, (3.2 + 8.0)/2 = 11.2/2= 5.6 (Control Group Midrange= 4.4)

Comparison with Control Group. All central measures of tree height are higher for the trees that received fertilizer and irrigation when compared with the control group. Without more sophisticated tests we cannot tell whether the difference is statistically significant, that is not likely to be attributable to chance.

In Exercises 13-16, refer to the data set in Appendix B. Use computer software or a calculator to find the means and medians, then compare the results as indicated.

13. Head Circumference Comparison head circumference in cm. of males and females on central measures (Data Set 4)

Statistic	Males	Females
Mean	41.10	40.05
Median	41.10	40.20

There does appear to be a slight difference since both central measures are higher for the males than for the females with males having a 1.05 cm. higher mean circumference.

14. Body Mass Index Comparison of BMI of men and women (Data Set 1)

Statistic	Men	Women
Mean	26.00	25.74
Median	26.20	23.90

Yes, there does appear to be a slight difference since both central measures are higher for the men than for the women with men having a 0.26 BMI higher mean difference compared with the women. This may not be a significant difference. We would need to determine this with other methods.

15. Petal Lengths of Irises Comparison of petal lengths of the three classes (Data Set 7)

Statistic	Setosa	Versicolor	Virginica
Mean	1.46	4.26	5.55
Median	1.50	4.35	5.55

No, they do not appear to have the same petal lengths. It is very clear that there are differences in central measures among these three classes of Iris petal lengths. Petal length is much higher for the Versicolor and Virginica classes compared with the Setosa class while the length of the Virginica class seems to be higher than the length of petals for the Versicolor class. They may all be different from each other, but this would need to be tested with more precise methods to decide.

16. Sepal Widths of Irises

Statistic	Setosa	Versicolor	Virginica
Mean	3.42	2.77	2.97
Median	3.40	2.80	3.00

No, they do not appear to have the same sepal widths. There seems to be a difference between the Setosa compared with both the Versicolor and Virginica classes. There may be differences between the widths of the Versicolor and the Virginica classes as well. They may all be different from each other, but this would need to be tested with more precise methods to decide.

In Exercises 17 and 18, find the mean of the data summarized in the given frequency distribution.

17. Mean from a Frequency Distribution The following frequency distribution will be used to find the gouped mean:

Frequency Distribution of Cotinine Levels of Smokers

Serum Cotinine Level (mg/ml)	Frequency (*f*)	Interval Midpoint (*x*)	$f * x$
0 – 99	11	49.5	544.5
100 – 199	12	149.5	1794.0
200 – 299	14	249.5	3493.0
300 – 399	1	349.5	349.5
400 - 499	2	449.5	899.0
Sum	$\Sigma f = 40$		$\Sigma(f * x) = 7080.0$

$$\text{Grouped Mean} = \frac{\sum(f * x)}{\sum f} = \frac{7080.0}{40} = 177.0$$

The mean computed using the actual raw values was 172.5.

The Grouped Mean is close to the actual mean. The grouped mean is an estimate of the actual mean. With the availability of computing equipment, the actual mean would be easy to compute and would always be preferred. However, if the raw scores were not available, this procedure could provide a reasonable estimate.

18. Body Temperatures The following frequency distribution will be used to find the grouped mean:

Frequency Distribution of Body Temperature

Temperature	Frequency (f)	Interval Midpoint (x)	$f * x$
96.5 – 96.8	1	96.65	96.65
96.9 – 97.2	8	97.05	776.40
97.3 – 97.6	14	97.45	1364.30
97.7 – 98.0	22	97.85	2152.70
98.1 – 98.4	19	98.25	1866.75
98.5 – 98.8	32	98.65	3156.80
98.9 – 99.2	6	99.05	594.30
99.3 – 99.6	4	99.45	397.80
Sum	$\Sigma f= 106$		$\Sigma(f * x)= 10405.70$

$$\text{Grouped Mean}= \frac{\sum(f * x)}{\sum f} = \frac{10405.70}{106} = 98.17$$

The mean computed using the actual raw values was 98.6^{o}F.

The Grouped Mean is close to the actual mean, but underestimates it to some extent. The grouped mean is an estimate of the actual mean. With the availability of computing equipment, the actual mean would be easy to compute and would always be preferred. However, if the raw scores were not available this procedure could provide a reasonable estimate.

19. Mean of Means No, this process would not result in an accurate computation of the mean salary of physicians for the country. This process would provide a single mean entry per state in such a way that it would assume the same number of physicians were in each state. Of course this is not true. There would be many more physicians in states like New York and California compared with states like New Hampshire and Wyoming. The actual number of physicians within each state would need to be accounted for in this computation.

20. Trimmed Means

a. The mean is 182.89 based on 54 scores

b. 10% trimmed mean (drop lowest 5 and highest 5 scores, based on 44 scores)= 171.00

c. 20% trimmed mean (drop lowest 11 and highest 11 scores, based on 32 scores)= 159.16

There are large differences among these three means when extreme score are removed. The mean decreases when high extreme scores are removed (as in this example) and it increases when low extreme scores are removed and if these are not distributed approximately equally below and above the mean, bias can occur.

21. Censored Data The mean of the five values, including the one that is censored, would be 3.42. Since the two values of 5 are minimal values and their actual values would be higher than 5, we would conclude that the mean would be 3.42 or higher.

2-5 Measures of Variation

In Exercises 1-8, find the range, variance, and standard deviation for the given sample data (The same data were used in Section 2 – 4 where we found measures of center. Here we find measures of variation).

In the following computations of variance and standard deviation, the equation used to compute the variance includes the following terms:

n represents the sample size

$\sum x$ represents the sum of the scores (the scores are added together)

$(\sum x)^2$ represents the squared value for the sum of scores (the scores are added together and then that value is squared)

$\sum x^2$ represents the sum of the squared scores (each score is squared and then these are all added together)

1. **Tobacco Use in Children's Movies**

. **a.** Range = highest score – lowest score = 548.0 – 0 = 548.0

b. Variance $s^2 = \dfrac{n(\sum x^2) - (\sum x)^2}{n(n-1)} = \dfrac{6(381009) - (947)^2}{6(6-1)} = \dfrac{1389245}{30} = 46308.2$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{46308.2} = 215.2$

2. **Cereal**

a. Range = highest score – lowest score = 0.48 – 0.03 = 0.45

b. Variance $s^2 = \dfrac{n(\sum x^2) - (\sum x)^2}{n(n-1)} = \dfrac{16(1.8144) - (4.72)^2}{16(16-1)} = \dfrac{6.752}{240} = 0.0281$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{0.0281} = 0.168$

Would an amount of sugar of 0.95 be considered "unusual"?
Need to use the mean. From before it is 0.295 (Exercise 4.2)
A sugar amount of 0.95 is not lower than the minimum usual value of mean – 2 s or
0.295 – 2 (0.168)= 0.295 – 0.336 = – 0.041. But, a sugar amount of 0.95 is higher than the maximum usual value of mean + 2 s or 0.295 + 2 (0.168)= 0.295 + 0.336 = 0.631
A sugar amount of 0.95 is NOT between the usual minimum and usual maximum values of
– 0.041 and 0.631. Thus, a sugar amount of 0.95 would be considered unusual.

3. **Body Mass Index**

a. Range = highest score – lowest score = 37.7 – 17.7 = 20.0

b. Variance $s^2 = \dfrac{n(\sum x^2) - (\sum x)^2}{n(n-1)} = \dfrac{15(10101.23) - (380.5)^2}{15(15-1)} = \dfrac{6738.20}{210} = 32.09$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{32.0867} = 5.66$

Would a body mass index of 34.0 be considered "unusual"?
Need to use the mean. From before it is 25.37 (Exercise 4.3)
A body mass index of 34 is not lower than the minimum usual value of mean – 2 s or
25.37 – 2 (5.66) = 25.37 – 11.32 = 14.04 AND a body max index of 34 is not higher than the maximum usual value of mean + 2 s or 25.37 + 2 (5.66) = 25.37 + 11.32 = 36.69
A body mass index of 34 is between the usual minimum and usual maximum values of 20.68 and 36.69. Thus, a body mass of 34 would not be considered unusual.

4. **Drunk Driving**

a. Range = highest score – lowest score = 0.29 – 0.12 = 0.17

b. Variance $s^2 = \dfrac{n(\sum x^2) - (\sum x)^2}{n(n-1)} = \dfrac{15(0.5631) - (2.81)^2}{15(15-1)} = \dfrac{0.5504}{210} = 0.002621$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{0.002621} = 0.051$

When a state wages a campaign to reduce drunk driving, is the campaign intended to lower the standard deviation? Such a campaign would be intended to reduce the variables such as number of crashes related to driving under the influence or the mean blood alcohol concentrations of drivers, not the standard deviation.

5. Motorcycle Fatalities

a. Range = highest score – lowest score = 42 – 14 = 28.0

b. Variance $s^2 = \dfrac{n(\sum x^2) - (\sum x)^2}{n(n-1)} = \dfrac{18(14343) - (485)^2}{18(18-1)} = \dfrac{22949}{306} = 75.00$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{75.00} = 8.66$

How does the variation of these ages compare to the variation of ages of all licensed drivers in the general population? The variation of this group will be lower since the range of 14 – 42 is lower than the typical range of the age of licensed drivers which would probably be from 15 to over 80. One would expect the variance and standard deviation of this group to be lower as well.

6. Fruit Flies

a. Range = highest score – lowest score = 0.92 – 0.64 = 0.28

b. Variance $s^2 = \dfrac{n(\sum x^2) - (\sum x)^2}{n(n-1)} = \dfrac{11(7.2568) - (8.88)^2}{11(11-1)} = \dfrac{0.9704}{110} = 0.008822$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{0.008822} = 0.094$

Would a thorax length of 0.63 mm. be considered “unusual”?
Need to use the mean. From before it is 0.807 (Exercise 2 – 4 .6)
A thorax length of 0.63 mm. is not lower than the minimum usual value of mean – 2 s or
0.807 – 2 (0.094) = 0.807 – 0.188 = 0.619 and a thorax length of 0.63 mm. is not higher than the maximum usual value of mean + 2 s or 0.807 + 2 (0.0.94) = 0.807 + 0.188 = 0.995
A thorax length of 0.63 is between the usual minimum and usual maximum values of
0.619 and 0.995. Thus, a thorax length of 0.63 would not be considered unusual.

7. Blood Pressure Measurements

a. Range = highest score – lowest score = 150 – 120 = 30.0

b. Variance $s^2 = \dfrac{n(\sum x^2) - (\sum x)^2}{n(n-1)} = \dfrac{14(252179) - (1875)^2}{14(14-1)} = \dfrac{14881}{182} = 81.76$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{81.76} = 9.04$

What does this suggest about the accuracy of the readings? They do not appear to be very accurate when one considers that all of these measures were taken on the same patient.

8. Phenotypes of Peas

a. Range = highest score – lowest score = 4 – 1 = 3.0

b. Variance $s^2 = \frac{n(\sum x^2)-(\sum x)^2}{n(n-1)} = \frac{25(109)-(47)^2}{25(25-1)} = \frac{516}{600} = 0.860$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{0.860} = 0.927$

These are nominal data. Central and variation statistics can be computed on any set of numbers. However, since these data are nominal the statistics have no meaning. They are not legitimate since there is a lack of equal intervals and no order to the categories.

In Exercises 9-12, find the range, variance, and standard deviation for each of the two samples, then compare the two sets of results. (The same data were used in Section 2 – 4.)

9. Patient Waiting Times

Single line:

a. Range = highest score – lowest score = 77 – 65 = 12.0

b. Variance $s^2 = \frac{n(\sum x^2)-(\sum x)^2}{n(n-1)} = \frac{10(51327)-(715)^2}{10(10-1)} = \frac{2045}{90} = 22.72$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{22.72} = 4.77$

Multiple line:

a. Range = highest score – lowest score = 100 – 42 = 58.0

b. Variance $s^2 = \frac{n(\sum x^2)-(\sum x)^2}{n(n-1)} = \frac{10(54109)-(715)^2}{10(10-1)} = \frac{29865}{90} = 331.83$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{331.83} = 18.22$

Compare the results. There is clearly higher variation for the multiple line as compared with the single line.

10. Skull Breadths

4000 B.C.:

a. Range = highest score – lowest score = 138 – 119 = 19

b. Variance $s^2 = \frac{n(\sum x^2)-(\sum x)^2}{n(n-1)} = \frac{12(198898)-(1544)^2}{12(12-1)} = \frac{2840}{132} = 21.515$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{21.515} = 4.64$

150 A.D.:

a. Range = highest score – lowest score = 141 – 126 = 15

b. Variance $s^2 = \dfrac{n(\sum x^2) - (\sum x)^2}{n(n-1)} = \dfrac{12(213610) - (1600)^2}{12(12-1)} = \dfrac{3320}{132} = 25.152$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{25.152} = 5.02$

Compare the results. There is higher variation for the 150 A.D. group as compared with the 4000 B.C. group. Even though the range for the B.C. group is higher, the variance and standard deviations for the A.D. group is higher and these are more accurate statistics since they use all of the scores rather than just two of them.

11. Poplar Trees

Control Group:

a. Range = highest score – lowest score = 6.9 – 1.9 = 5.0

b. Variance $s^2 = \dfrac{n(\sum x^2) - (\sum x)^2}{n(n-1)} = \dfrac{20(528.42) - (99)^2}{20(20-1)} = \dfrac{767.40}{380} = 2.019$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{2.019} = 1.42$

Irrigation Group:

a. Range = highest score – lowest score = 6.8 – 1.2 = 5.6

b. Variance $s^2 = \dfrac{n(\sum x^2) - (\sum x)^2}{n(n-1)} = \dfrac{20(461.84) - (89.6)^2}{20(20-1)} = \dfrac{1208.64}{380} = 3.181$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{3.181} = 1.78$

Compare the results. There is slightly higher variation in the heights of the Irrigation Group (s^2= 3.18) compared with the Control Group (s^2= 2.02).

12. Poplar Trees

Fertilizer and Irrigation Treatment:

a. Range = highest score – lowest score = 8.0 – 3.2 = 4.8

b. Variance $s^2 = \dfrac{n(\sum x^2) - (\sum x)^2}{n(n-1)} = \dfrac{20(865.84) - (129.2)^2}{20(20-1)} = \dfrac{642.16}{380} = 1.643$

c. Standard Deviation $s = \sqrt{s^2} = \sqrt{1.643} = 1.28$

Comparison: The variation for the Control Group (s^2= 2.02) is greater than the variation of the Fertilizer and Irrigation Group (s^2 = 1.64).

In Exercises 13 – 16, refer to the data set in Appendix B. Use computer software or a calculator to find the standard deviations, then compare the results.

13. Head Circumference

$s_{\mathrm{F}} = 1.64 \quad s_{\mathrm{M}} = 1.50$

There does not appear to be a substantial difference in the variation.

14. Body Mass Index

$s_M = 3.43 \quad s_W = 6.17$

There does appear to be a difference in that the variation of the women BMI is higher than the variation among the men.

15. Petal Length of Irises

$s_{Setsosa} = 0.174 \qquad s_{Versicolor} = 0.470 \qquad s_{Virginica} = 0.552$

No, they do not appear to have the same variation of petal lengths. It is very clear that there are differences in variances among these three classes of Iris petal lengths. Petal length is much more variable for the Versicolor and Virginica classes when compared with the Setosa class while the variation for the Virginica class seems to be higher than the variation of length of petals for the Versicolor class.

16. Sepal Widths of Irises

$s_{Setosa} = 0.381 \qquad s_{Versicolor} = 0.314 \qquad s_{Virginica} = 0.322$

No, they do not appear to have the same variation of sepal lengths. It is very clear that there are differences in variances among these three classes of Iris petal lengths. Petal length is much more variable for the Versicolor and Virginica classes when compared with the Setosa class while the variation for the Virginica class seems to be higher than the variation of length of sepals for the Versicolor class.

17. Finding Standard Deviation from a Frequency Distribution

Frequency Distribution of Cotinine Levels of Smokers

Serum Cotinine Level (mg/ml)	Frequency (*f*)	Interval Midpoint (*x*)	Interval Midpoint Squared (x^2)	$f * x$	$f * x^2$
0 – 99	11	49.5	2450.25	544.5	26952.75
100 – 199	12	149.5	22350.25	1794.0	268203.00
200 – 299	14	249.5	62250.25	3493.0	871503.50
300 – 399	1	349.5	122150.25	349.5	122150.25
400 – 499	2	449.5	202050.25	899.0	404100.50
Sum	Σf= 40	---	---	$\Sigma(f * x)$= 7080.0	$\Sigma(f * x^2)$= 1692910.00

Find the variance first and then the standard deviation

$$s^2 = \frac{n[\sum(f * x^2)] - [(\sum(f * x)^2]}{n(n-1)} = \frac{40(1692910.00) - (7080)^2}{40(40-1)} = \frac{17590000}{1560} = 11275.64$$

Standard Deviation $s = \sqrt{s^2} = \sqrt{11275.64} = 106.2$

Comparison: In this case, the standard deviation (106.2) is lower when computed as a grouped standard deviation compared with when computed with the original 40 scores (119.5). We make the assumption that the midpoint of the interval is representative of all the scores within the interval. This is not likely to be exactly the case so to the extent that it is not, there will be errors in the estimate and those errors could result in estimates that are lower or higher than the ungrouped estimate. The ungrouped estimate should be used whenever possible, but if the original scores are not available, yet the grouped frequency distribution is, this approach could be used to get an estimate.

18. Body Temperatures

Temperature	Frequency (f)	Interval Midpoint (x)	Interval Midpoint Squared (x^2)	$f * x$	$f * x^2$
96.5 – 96.8	1	96.65	9341.22	96.65	9341.22
96.9 – 97.2	8	97.05	9418.70	776.40	75349.62
97.3 – 97.6	14	97.45	9496.50	1364.30	132951.04
97.7 – 98.0	22	97.85	9574.62	2152.70	210641.70
98.1 – 98.4	19	98.25	9653.06	1866.75	183408.19
98.5 – 98.8	32	98.65	9731.82	3156.80	311418.32
98.9 – 99.2	6	99.05	9810.90	594.30	58865.42
99.3 – 99.6	4	99.45	9890.30	397.80	39561.21
Sum	$\Sigma f = 106$			$\Sigma(f * x) = 10405.70$	$\Sigma(f * x^2) = 1021536.72$

Find the variance first and then the standard deviation

$$s^2 = \frac{n[\sum(f * x^2)] - [(\sum(f * x)^2]}{n(n-1)} = \frac{106(1021536.72) - (10405.70)^2}{106(106-1)} = \frac{4299.83}{11130} = 0.386$$

Standard Deviation $s = \sqrt{s^2} = \sqrt{0.386} = 0.621$

19. Range Rule of Thumb Estimate the standard deviation of all faculty member ages at your college. I work at the University of Alabama at Birmingham. I would estimate the youngest faculty member to be about 25 and the oldest to be about 75. The range is 50 years. Using the range rule of thumb, the estimated standard deviation would be:

$$s \approx \frac{Range}{4} = \frac{50}{4} = 12.5 \text{ years of age}$$

20. Range Rule of Thumb $n = 40$ $\bar{x} = 38.86\,\text{cm}$ $s = 3.78\,\text{cm}$

Minimum "usual" upper leg length would be $\bar{x} - 2s$ = 38.86 – 2(3.78) = 38.86 – 7.56= 31.30 cm

Maximum "usual" upper leg length would be $\bar{x} + 2s$ = 38.86 + 2(3.78) = 38.86 + 7.56= 46.42 cm

Is a length of 47.0 cm considered unusual in this context? Yes, since 47.0 is outside the limits provided using this rule, it would be considered unusual.

21. Empirical Rule Mean of 176 cm and standard deviation of 7 cm in the population.

Approximate percentage of men between:

a. 169 cm and 183 cm

169 cm is one standard deviation (176 – 7= 169) below the mean and 183 cm is one standard deviation (176 + 7= 173) above the mean. The percentage of scores between one standard deviation below the mean and one standard deviation above the mean ($\bar{x} \pm 1s$) is <u>68%</u> in a bell-shaped distribution.

b. 155 cm and 197 cm

155 cm is three standard deviations below the mean [176 – (3 * 7) = 155] and 197 cm is three standard deviations above the mean [176 + (3 * 7) = 197]. The percentage of scores between three standard deviation below the mean and three standard deviation above the mean ($\bar{x} \pm 3s$) is <u>99.7%</u> in a bell-shaped distribution.

22. Coefficient of Variation Data from Exercise 9 related to patient waiting times

Single line: $CV = \frac{s}{\bar{x}} * 100\% = \frac{4.77}{71.5} * 100\% = 6.7\%$

Multiple line: $CV = \frac{s}{\bar{x}} * 100\% = \frac{18.21}{71.5} * 100\% = 25.5\%$

Comparison: The coefficient of variation for the multiple line group is much larger, by a factor of 3.8 times, than the coefficient of variation for the single line group. Relative to units of the mean, there is 6.7% standard deviation units for the single line and 25.5% for the multiple line. Since the means of the two groups are the same, this reflects the relative difference of the two standard deviations. However, most of the time the sample means would not be the same so this conversion of standard deviation into units of the mean will be more useful for comparing different groups on relative standard deviation with units of the original variable removed.

23. Coefficient of Variation
Heights of men: $n = 6 \quad \bar{x} = 69.17 \text{ in} \quad s = 2.14 \text{ in}$

$$CV = \frac{s}{\bar{x}} \bullet 100\% = \frac{2.14}{69.17} \bullet 100\% = 3.09\%$$

Lengths (mm) of cuckoo eggs: $n = 9 \quad \bar{x} = 22.14 \text{ mm} \quad s = 1.13 \text{ mm}$

$$CV = \frac{s}{\bar{x}} \bullet 100\% = \frac{1.13}{22.14} \bullet 100\% = 5.10\%$$

Comparison: The relative variation of the cuckoo eggs lengths is greater by a factor of about 1.7 times, than the relative variation of the men's heights, but this does not appear to be a substantial difference.

24. Equality for All When the standard deviation (s) is equal to zero, all of the scores are equal (have the same value). A standard deviation of zero indicates there is no variability of the scores. Actually all scores will be equal to each other and they will all be equal to the mean.

25. Understanding Units of Measurement Data consisting of longevity times (in days) of fruit flies
Units of standard deviation are days while units of variance are in days squared. This is the primary reason we convert variance to standard deviation since finding the square root also affects the unit of measurement and this puts our measure of variability back on the same continuum as the scores. It's hard for us to comprehend squared days, squared blood pressure, squared minutes, etc.

26. Interpreting Outliers Twenty scores are fairly close together. A new value is added that is an outlier (very far away from the other values). When the new standard deviation is computed including this outlier, the standard deviation will increase even if the score is below the mean. The extent to which it will increase depends on how far away the score is from the mean of the original set of scores and how many scores there are in the original distribution. In general, the further away from the mean the outlier is, the higher the increase in standard deviation. However, the number of scores would be a factor as well. The same outlier as added to this group of 20 will have a smaller effect on the standard deviation if there had been 100 scores rather than 20.

27. Why Divide by *n* – 1? Let the population be: 3, 6, 9
Population Parameters: $n = 3 \quad \mu = 6 \quad \sigma^2 = 6.0 \quad \sigma = 2.45$

a. Population variance (using *n* rather than *n* – 1) for denominator, $\sigma^2 = 6.00$
Nine possible samples of *n* = 2 and their variances as samples:

b. Possible samples of size 2, with replacement*	Sample Mean	**b.** Sample Variances Division by (n – 1)	**c.** Population Variances Division by n	Sample Standard Deviation
3,3	3.0	0.0	0.00	0.00
3,6	4.5	4.5	2.25	2.12
3,9	6.0	18.0	9.00	4.24
6,3	4.5	4.5	2.25	2.12
6,6	6.0	0.0	0.00	0.00
6,9	7.5	4.5	2.25	2.12
9,3	6.0	18.0	9.00	4.24
9,6	7.5	4.5	2.25	2.12
9,9	9.0	0.0	0.00	0.00
Mean	6.0	6.00	3.00	1.88

* With replacement is a critical aspect of sampling. When we pull a score from a distribution, we record it and then return it back to the population for the next sample or when we pull out a sample we record the sample values and then return them back to the population before selecting the next sample. The reason this is done is so that every sampling is done from the same population. Otherwise, we could not have had the samples with scores 3 and 3, 6 and 6, and 9 and 9.

d. The average variance for those samples where n – 1 was used exactly matched the population variance. Thus, the sample variance, on average, is an unbiased estimator of the population variance. An unbiased estimator is one where the average of the sampled statistics will be the population parameter it is designed to estimate. However, notice that the variation is higher for the n – 1 situation. This difference would decrease as the sample size increases.

e. If we find the standard deviation of the sample variances (column e) and then find their average of 1.88, we see that this average is not the same as the population standard deviation of 2.45. Thus, the sample standard deviation is not an unbiased estimator of the population standard deviation. In this case, we would say that the sample standard deviation is negatively biased meaning that the average estimator is lower than the parameter it is being used to estimate.

2-6 Measures of Relative Standing

In Exercises 1-4, express all z scores with two decimal places

1. **Darwin's Height** Darwin's height was 182 cm, population values: $\mu = 176$ cm, $\sigma = 7$ cm

a. Difference between Darwin's height and mean $x = 182 - 176 = 6$ cm

b. Number of standard deviations is 6/7 = 0.86

c. z score $z = \dfrac{x-\mu}{\sigma} = \dfrac{182-176}{7} = 0.86$

d. Usual heights are between -2 and +2 z values. Is Darwin's height usual or unusual?
Darwin's z value of +0.86 is between -2 and +2 z values, so his height would be considered usual.

2. **Heights of Men** Population of adult males, heights μ = 176 cm σ = 7 cm

   a. z score for Danny DeVito with height of 152 cm $z=\dfrac{x-\mu}{\sigma}=\dfrac{152-176}{7}=\dfrac{-24}{7}=-3.43$

   b. z score for Shaquille O'Neal with height of 216 cm $z=\dfrac{x-\mu}{\sigma}=\dfrac{216-176}{7}=\dfrac{40}{7}=5.71$

3. **Pulse Rates of Adults** Population pulse rate is: μ = 72.9 bpm, σ = 12.3 bpm

   a. Difference between pulse rate of 48 and mean, x = 48 – 72.9 = –24.9

   b. How many standard deviations is that? –24.9/12.3 = –2.02

   c. z score $z=\dfrac{x-\mu}{\sigma}=\dfrac{48-72.9}{12.3}=\dfrac{-24.9}{12.3}=-2.02$

   d. If usual is between ± 2z, then the range of pulse rates that are between -2 and +2 z values would be: μ – 2 σ = 72.9 – 2 (12.3)= 72.9 – 24.6= 48.3 and μ + 2 σ = 72.9 + 2 (12.3)= 72.9 + 24.6= 97.5. A pulse rate of 48 would be considered to be unusual since it is outside the range of 48.3 to 97.5. Among the reasons the pulse rate could be this low would be: heredity, young in age, non-smoker, regular exerciser, good diet, low stress job, etc.

4. **Body Temperatures** Population temperatures: μ = 98.20°F, σ = 0.62 °F

   a. z score for temperature of 100 °F $z=\dfrac{x-\mu}{\sigma}=\dfrac{100-98.2}{0.62}=\dfrac{1.80}{0.62}=2.90$

   b. z score for temperature of 96.96 °F $z=\dfrac{x-\mu}{\sigma}=\dfrac{96.96-98.20}{0.62}=\dfrac{-1.24}{0.62}=-2.00$

   c. z score for temperature of 98.20 °F $z=\dfrac{x-\mu}{\sigma}=\dfrac{98.2-98.2}{0.62}=\dfrac{0}{0.62}=0$

In Exercises 5-8, express all z scores with two decimal places. Consider a score to be unusual if its z score is less than –2.00 or greater than +2.00.

5. Heights of Women – Women's height parameters: μ = 63.6 in , σ = 2.5 in

   z score for height of 70 in $z=\dfrac{x-\mu}{\sigma}=\dfrac{70-63.6}{2.5}=\dfrac{6.4}{2.5}=2.56$

   A height of 70 in. is 2.56 standard deviations above the mean and is higher than the +2.00 criterion point for being "usual." Thus, this would be unusual and the Beanstalk Club members may be aptly named.

6. **Length of Pregnancy** Length of pregnancy parameters: μ = 268 days, σ = 15 days

   z score for number of days of 308 $z=\dfrac{x-\mu}{\sigma}=\dfrac{308-268}{15}=\dfrac{40}{15}=2.67$

A value for the number of days of pregnancy at 308 is 2.67 standard deviations above the mean and is higher than the +2.00 criterion point for being "usual." Thus, this would be unusual. While it would have been interesting to see if Dear Abby performed this analysis or provided advice with some other basis, it may be best that we don't try to come of up any causal statement in this case.

7. **Body Temperature** Body temperature parameters: $\mu = 98.20^{o}\text{F}$, $\sigma = 0.62\,^{o}\text{F}$

z score for temperature of 101 ^{o}F $$z = \frac{x-\mu}{\sigma} = \frac{101-98.2}{0.62} = \frac{2.80}{0.62} = 4.52$$

A body temperature of 101 ^{o}F is 4.52 standard deviations above the mean and is higher than the +2.00 criterion point for being "usual." Thus, this would be an unusually high temperature. This would suggest that immediate actions are called for to reduce the temperature for this patient.

8. **Cholesterol Levels** Serum cholesterol parameters: $\mu = 178.1$ mg/100mL, $\sigma = 40.7$ mg/100mL

z score for cholesterol level of 259.0 mg/100mL $$z = \frac{x-\mu}{\sigma} = \frac{259.0-178.1}{40.7} = \frac{80.9}{40.7} = 1.99$$

A cholesterol level of 259 mg/100mL is 1.99 standard deviations above the mean. This is within the ± 2 z values that would be the limits for being "usual." Thus, for his age group, this male is within the usual range and would not be considered unusually high.

9. **Comparing Test Scores** Which is better?

Biology: $x = 85$ $$z = \frac{x-\bar{x}}{s} = \frac{85-90}{10} = \frac{-5}{10} = -0.50$$

Economics: $x = 45$ $$z = \frac{x-\bar{x}}{s} = \frac{45-55}{5} = \frac{-10}{5} = -2.00$$

The range of *z* values is about –3.00 to +3.00 and the order is such that as scores move from the low part of the range (–3.00) through 0 to the upper part of the range (up to about +3.00), scores are increasing. Thus, a *z* of –0.5 is higher than a *z* of –2.0 on that –3.00 to +3.00 continuum. The student performed better, relative to the other students taking the tests, on the biology test than on the economics test.

10. **Comparing Scores** z scores for three students

a. Test with $\bar{x} = 128, s = 34$, score of 144 $$z = \frac{x-\bar{x}}{s} = \frac{144-128}{34} = \frac{16}{34} = 0.47$$

b. Test with $\bar{x} = 86, s = 18$, score of 90 $$z = \frac{x-\bar{x}}{s} = \frac{90-86}{18} = \frac{4}{18} = 0.22$$

c. Test with $\bar{x} = 15, s = 5$, score of 18 $$z = \frac{x-\bar{x}}{s} = \frac{18-15}{5} = \frac{3}{5} = 0.60$$

The highest relative score is the score of 18 on the test with mean of 15 and standard deviation of 5 as reflected in the value of the z score being +0.60, higher than +0.47 and +0.22.

In Exercises 11-14, use the sorted continine levels of smokers listed in Table 2-11. Find the percentile corresponding to the given cotinine level.

It might make this a little easier if we put the order number with the score, as below (we would not usually go to all of this work, we would usually let a computer do all of this, especially if there were a large number of scores):

Score #	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Score	0	1	1	3	17	32	35	44	48	86	87	103	112	121

Score#	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Score	123	130	131	149	164	167	173	173	198	208	210	222	227	234

Score #	29	30	31	32	33	34	35	36	37	38	39	40
Score	245	250	253	265	266	277	284	289	290	313	477	491

The percentile is the percentage of the scores that fall at or below a given score. It is found using:

$$\text{Percentile} = \frac{\text{number of values less than the given score}}{\text{total number of values}} * 100\%$$

11. Value of 149

$$\text{Percentile} = \frac{\text{number of values less than the given score}}{\text{total number of values}} * 100\% = \frac{17}{40} * 100\% = 0.425 * 100\% = 43rd \text{ percentile}$$

12. Value of 210

$$\text{Percentile} = \frac{\text{number of values less than the given score}}{\text{total number of values}} * 100\% = \frac{24}{40} * 100\% = 0.60 * 100\% = 60th \text{ percentile}$$

13. Value of 35

$$\text{Percentile} = \frac{\text{number of values less than the given score}}{\text{total number of values}} * 100\% = \frac{6}{40} * 100\% = 0.15 * 100\% = 15th \text{ percentile}$$

14. Value of 250

$$\text{Percentile} = \frac{\text{number of values less than the given score}}{\text{total number of values}} * 100\% = \frac{29}{40} * 100\% = 0.725 * 100\% = 73rd \text{ percentile}$$

In Exercises 15-22, use the sorted cotinine levels of smokers listed in Table 2-11. Find the indicated percentile or quartile.

In these examples, we find *L,* which is the number that represents the order position of the score when the scores are ordered low to high. *k* is the percentile value and *n* is the number of scores

$$L = \frac{k}{100} * n \text{ provides the order Locator}$$

If the value of *L* is a whole number, the percentile is the average of the score at *L* and the score at *L* + 1

If the value of *L* is not a whole number, the percentile is the value of *L* rounded up to a whole number.

15. Find P_{20}

$$L = \frac{k}{100} * n = \frac{20}{100} * 40 = 0.20 * 40 = 8$$

8 is a whole number, thus:

P_{20} is the average of the 8th and the 9th score or (44 + 48)/2 = 46.0, thus P_{20}= 46.0

16. Find Q_3 this is the 75th percentile

$$L = \frac{k}{100} * n = \frac{75}{100} * 40 = 0.75 * 40 = 30$$

30 is a whole number, thus:

Q_3 is the average of the 30th and 31st scores or (250 + 253)/2= 251.5, thus Q_3= 251.5 (also P_{75})

17. Find P_{75}

$$L = \frac{k}{100} * n = \frac{75}{100} * 40 = 0.75 * 40 = 30$$

30 is a whole number, thus:

P_{75} is the average of the 30th and 31st scores or (250 + 253)/2= 251.5, thus P_{75}= 251.5 (also Q_3)

18. Find Q_2

Q_2 is the median or also P_{50}

$$L = \frac{k}{100} * n = \frac{50}{100} * 40 = 0.50 * 40 = 20$$

20 is a whole number, thus:

Q_2 is the average of the 20th and 21st scores or (167 + 173)/2= 170, thus Q_2= 170 (also the median or P_{50})

19. Find P_{33}

$$L = \frac{k}{100} * n = \frac{33}{100} * 40 = 0.33 * 40 = 13.2$$

13.2 is not a whole number so *L* is the next highest whole number or the 14th number in the order. The 14th number is 121, thus P_{33} is 121.

20. Find P_{21}

$$L = \frac{k}{100} * n = \frac{21}{100} * 40 = 0.21 * 40 = 8.4$$

8.4 is not a whole number so *L* is the next highest whole number or the 9th number in the order. The 9th number is 48, thus P_{21} is 48.

21. Find P_1

$$L = \frac{k}{100} * n = \frac{1}{100} * 40 = 0.01 * 40 = 0.4$$

0.4 is not a whole number so L is the next highest whole number or the 1st number in the order. The 1st number is 0, thus P_1 is 0.

22. Find P_{85}

$$L = \frac{k}{100} * n = \frac{85}{100} * 40 = 0.85 * 40 = 34$$

34 is a whole number, thus P_{85} is the average of the 34th and 35th scores or (277 + 284)/2= 280.5, thus P_{85}= 280.5

23. Continine Levels of Smokers

a. Interquartile Range, distance between Q_1 and Q_3

The interquartile range is $Q_3 - Q_1$.

Finding Q_3 which is the same as P_{75}

$$L = \frac{k}{100} * n = \frac{75}{100} * 40 = 0.75 * 40 = 30$$

30 is a whole number, thus:

Q_3 is the average of the 30th and 31st scores or (250 + 253)/2= 251.5, thus Q_3= 251.5 (also P_{75})

Finding Q_1 which is the same as P_{25}

$$L = \frac{k}{100} * n = \frac{25}{100} * 40 = 0.25 * 40 = 10$$

10 is a whole number, thus Q_1 is the average of the 10th and 11th scores or (86 + 87)/2= 86.5, thus Q_1= 86.5

The interquartile range, IQR= $Q_3 - Q_1$= 251.5 – 86.5= 165.0

b. Midquartile, the point exactly between Q_1 and Q_3

The midquartile is defined as: $$Midquartile = \frac{Q_3 + Q_1}{2}$$

We have found that Q_3= 251.5 and Q_1= 86.5, thus:

$$Midquartile = \frac{Q_3 + Q_1}{2} = \frac{251.5 + 86.5}{2} = \frac{338.0}{2} = 169.0$$

c. 10 – 90 percentile range

Finding P_{10}

$$L = \frac{k}{100} * n = \frac{10}{100} * 40 = 0.10 * 40 = 4$$

P_{10}= Avg. of 4th and 5th scores= (3 + 17)/2= 10.0

Finding P_{90}

$$L = \frac{k}{100} * n = \frac{90}{100} * 40 = 0.90 * 40 = 36$$

P_{90}= Avg. of 36th and 37th scores=

(289 + 290)/2= 579/2= 289.5
10 – 90 percentile range= $P_{90} - P_{10}$= 289.5 – 10.0= 279.5

d. Does $P_{50} = Q_2$? If so, does P_{50} always equal Q_2?
In order to find Q_2, we have to find P_{50}, since 50 must be used for the value of *k* in finding the locator *L* value. Thus P_{50} always equals Q_2. Yes, $P_{50} = Q_2$ and yes, P_{50} always equals Q_2.

e. Does $Q_2 = (Q_1 + Q_3)/2$? If so, does Q_2 always equal $(Q_1 + Q_3)/2$?
In this case Q_1= 86.5 and Q_3= 251.5, $(Q_1 + Q_3)/2$ = (86.5 + 251.5)/2= 169.0 (this is also the value of the midquartile). Q_2 (the median) is 170.0. These values are very close, but they are not the same. They (the midquartile and the median) will be exactly the same if the distribution of scores is symmetrical around the median. To the extent the distribution is skewed, these values will be more different from each other. Actually, because these are so close in this example, we could infer that there is relative symmetry of this distribution.

2-7 Exploratory Data Analysis

The five-number summary referred to in the following Exercises are:

1. Minimum score
2. Q_1, the first quartile, also known as P_{25}
3. The median or Q_2, the second quartile, also known as P_{50}
4. Q_3, the third quartile, also known as P_{75}
5. Maximum score

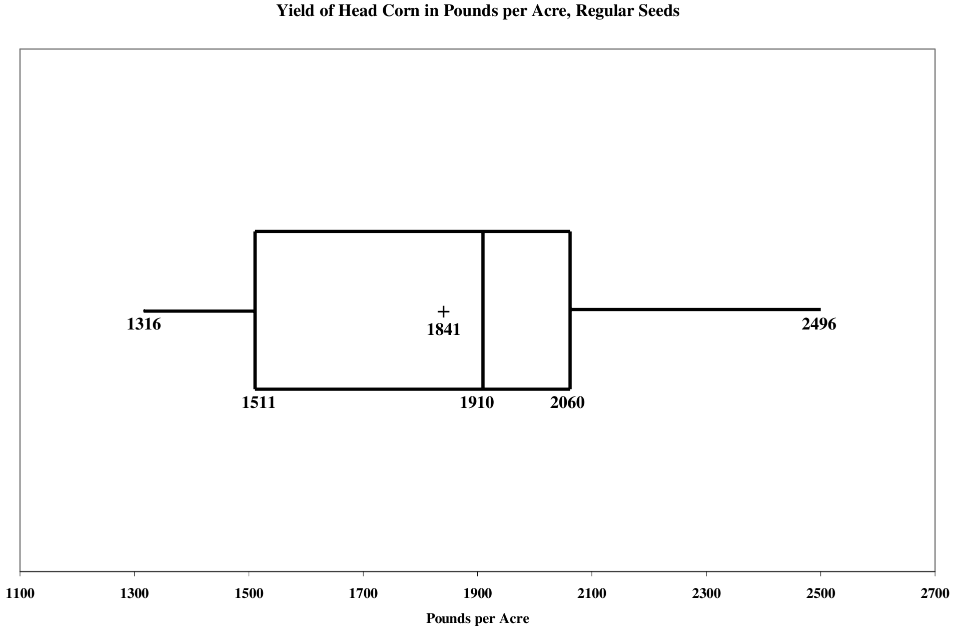
These are used to develop boxplots for exploratory data analysis. Boxplots, also referred to as Box and Whisker Plots, were proposed by a famous modern statistician by the name of John Tukey. They have only recently been incorporated into statistical software packages. As indicated in your text, there is quite a variation on how these are presented by the various packages. Some have the plot going in a vertical direction and some have the plot going in a horizontal direction and there seems to be no evidence that one way is better than another. Some include the mean as a point, usually with the + symbol for the mean on the continuum. Some include points that would be considered outliers with marks to indicate these points, based on rules such as those described in the text. Some boxplots, use variations of the lines that represent the box, such as the boxplots produced by the STATDISK software. All are reasonable ways of presenting the data displayed.

There is no boxplot option in the current version of Excel. However, the boxplots presented in this workbook have been generated using Excel by setting the 5-number plots as a variable and a constant such as 1 for the other variable and then generating a scatterplot. If there is more than one to be plotted on the same graph, the other(s) would be given a different constant than 1. The Excel scatterplot provides the points and the line draw function of Excel is used to draw the box and the whiskers. Outliers are not identified in these boxplots.

1. **Testing Corn Seeds** Regular Corn, *n*= 11, Scores in rank order:

Score #	1	2	3	4	5	6	7	8	9	10	11
Score	1316	1444	1511	1612	1903	1910	1935	1961	2060	2108	2496

Minimum score is 1316
Q_1, the first quartile, also known as P_{25} is the 3rd score, which is 1511
The median or Q_2, the second quartile, also known as P_{50} is the 6th score or 1910 (Mean is 1841)
Q_3, the third quartile, also known as P_{75} is the 9th score, which is 2060
Maximum score is 2496
Following is the boxplot. For the first few of these boxplots, the values will be entered for the 5-number summary so you can see where they are plotted. In addition, the mean is indicated with the + sign. Some boxploting programs include the mean and others don't. It is often useful to see where the mean lies relative to the median which is the middle line of the box. The mean is included in the boxplots presented in this workbook.

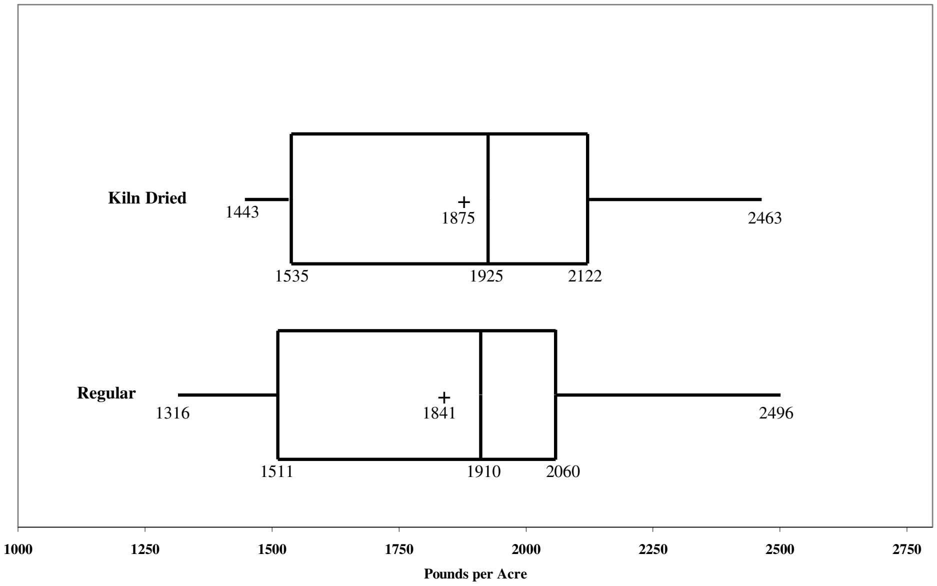


2. **Testing Corn Seeds** Kiln Dried, *n*= 11, Scores in rank order:

Score #	1	2	3	4	5	6	7	8	9	10	11
Score	1443	1482	1535	1542	1915	1925	2009	2011	2122	2180	2463

Minimum score is 1443
Q_1, the first quartile, also known as P_{25} is the 3rd score, which is 1535
The median or Q_2, the second quartile, also known as P_{50} is the 6th score or 1925 (Mean is 1875)
Q_3, the third quartile, also known as P_{75} is the 9th score, which is 2122
Maximum score is 2463
Following is the boxplot with the Regular corn boxplot included for comparison purposes:

Yield of Head Corn in Pounds per Acre, Boxplots Comparing Regular and Kiln Dried Corn

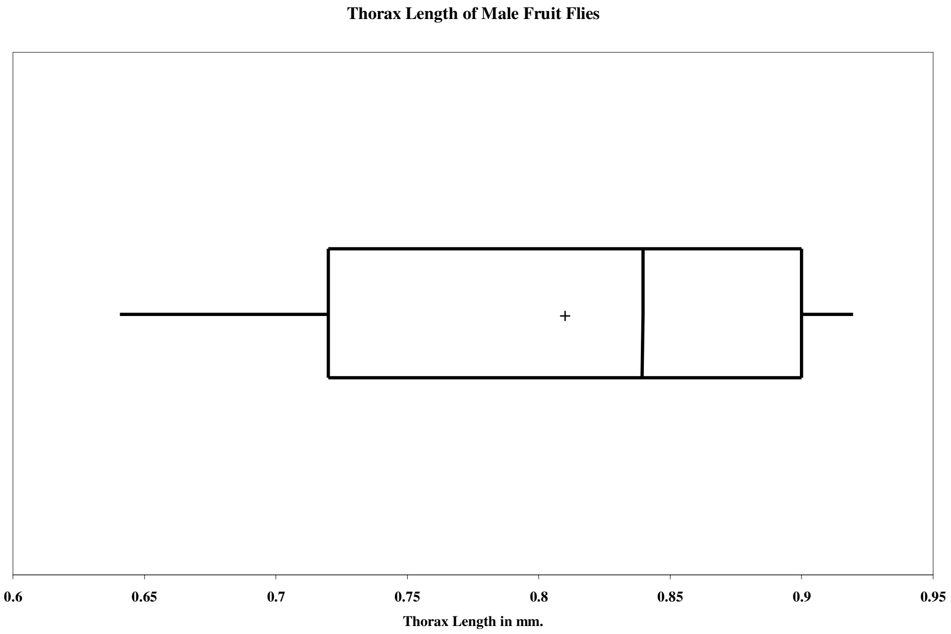


Comparison: There seems to be some difference in these distributions, but it is not probably enough to be considered substantial. The median for the Kiln Dried corn is a little higher and the variability of the Regular corn is slightly higher. These are not substantial differences, especially when one considers that the sample sizes (*n*= 11) are relatively small.

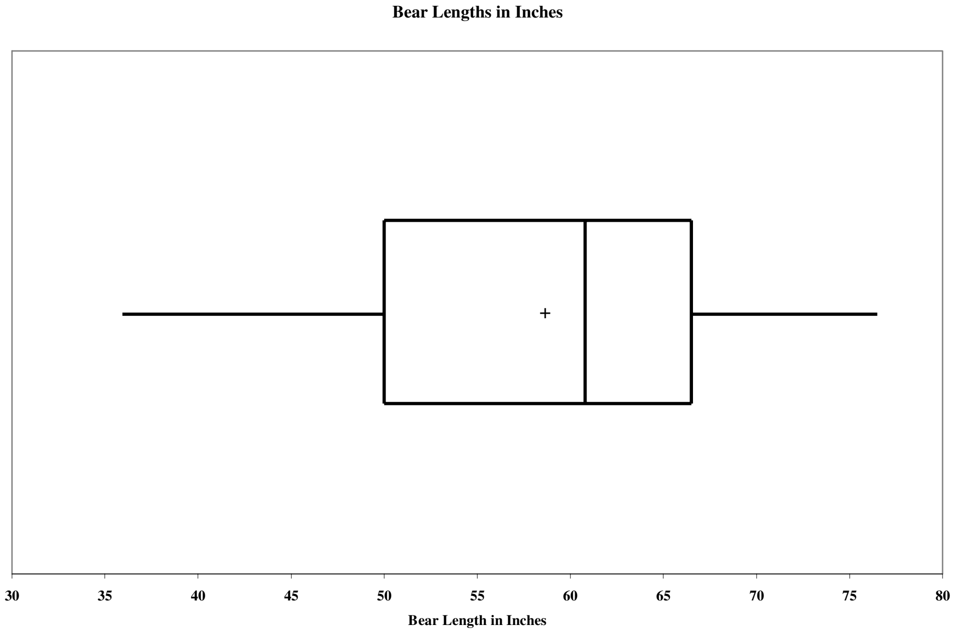
3. Fruit Flies *n*= 11, Scores in rank order:

Score #	1	2	3	4	5	6	7	8	9	10	11
Score	0.64	0.68	0.72	0.76	0.84	0.84	0.84	0.84	0.90	0.90	0.92

Minimum score is 0.64
Q_1, the first quartile, also known as P_{25} is the 3rd score, which is 0.72
The median or Q_2, the second quartile, also known as P_{50} is the 6th score or 0.84 (Mean is 0.81)
Q_3, the third quartile, also known as P_{75} is the 9th score, which is 0.90
Maximum score is 0.92
Following is the boxplot:

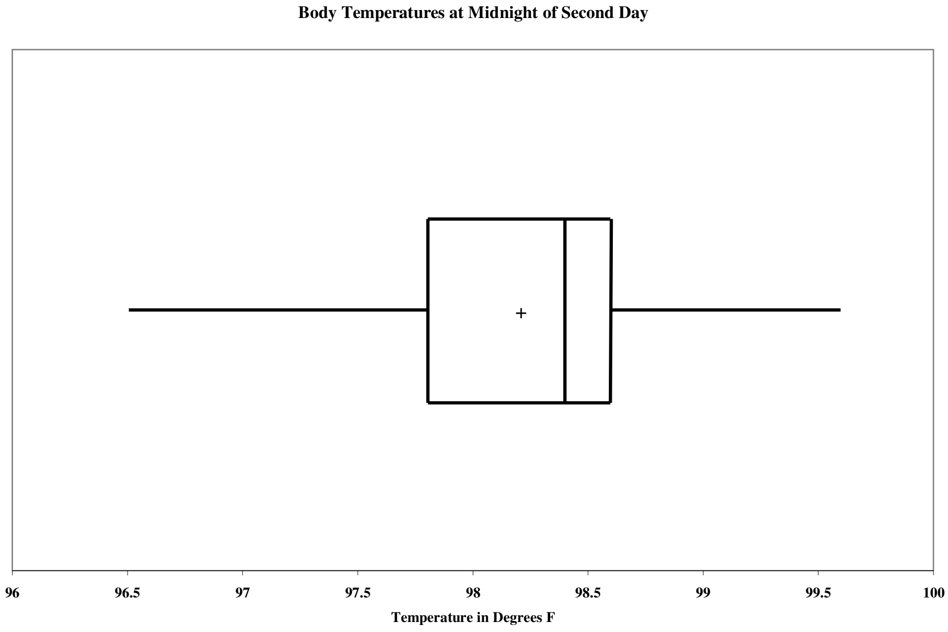


4. **Bear Lengths** (Data Set 6), *n*= 54
Minimum score is 36
Q_1, the first quartile, also known as P_{25} is the 14th score, which is 50
The median or Q_2, the second quartile, also known as P_{50} is the average of the 27th and 28th scores or (60.5 + 61.0)/2= 60.75 (Mean is 58.6)
Q_3, the third quartile, also known as P_{75} is the 41st score, which is 66.5
Maximum score is 76.5
Following is the boxplot:

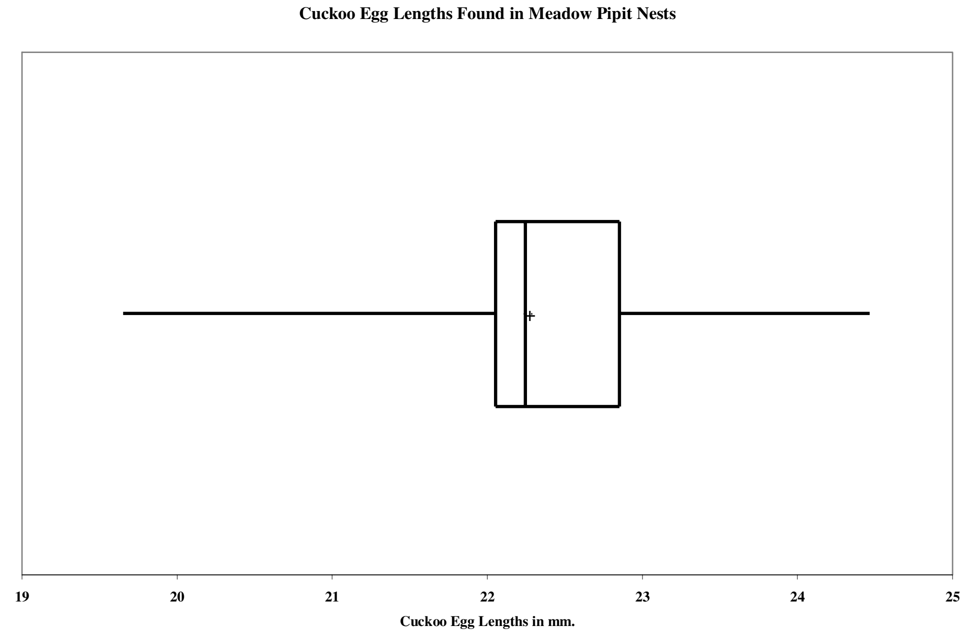


The distribution if bear lengths is not symmetrical. It appears to be skewed in the negative direction. This is also supported by the fact that the mean is lower than the median.

5. **Body Temperatures** (Data Set 2), *n*= 106
Minimum score is 96.5
Q_1, the first quartile, also known as P_{25} is the 27th score, which is 97.8
The median or Q_2, the second quartile, also known as P_{50} is the average of the 53rd and 54th scores or (98.4 + 98.4)/2= 98.4 (Mean is 98.2)
Q_3, the third quartile, also known as P_{75} is the 80th score, which is 98.6
Maximum score is 99.6
These data do not support the common belief that the mean body temperature is 98.6°F. In this distribution 98.6°F is at P_{75}. However, there is no indication that this sample is intended to be representative of the adult population and also these are temperatures taken at midnight. How many subjects needed to be awakened to have their temperatures taken? These factors might make a difference.
Following is the boxplot:



6. **Cuckoo Egg Lengths** (Data Set 8).Meadow pipits, *n*= 45
Minimum score is 19.65
Q_1, the first quartile, also known as P_{25} is the 12th score, which is 22.05
The median or Q_2, the second quartile, also known as P_{50} is the 23rd score or 22.25 (Mean is 22.30)
Q_3, the third quartile, also known as P_{75} is the 34th score, which is 22.85
Maximum score is 24.45
Following is the boxplot:



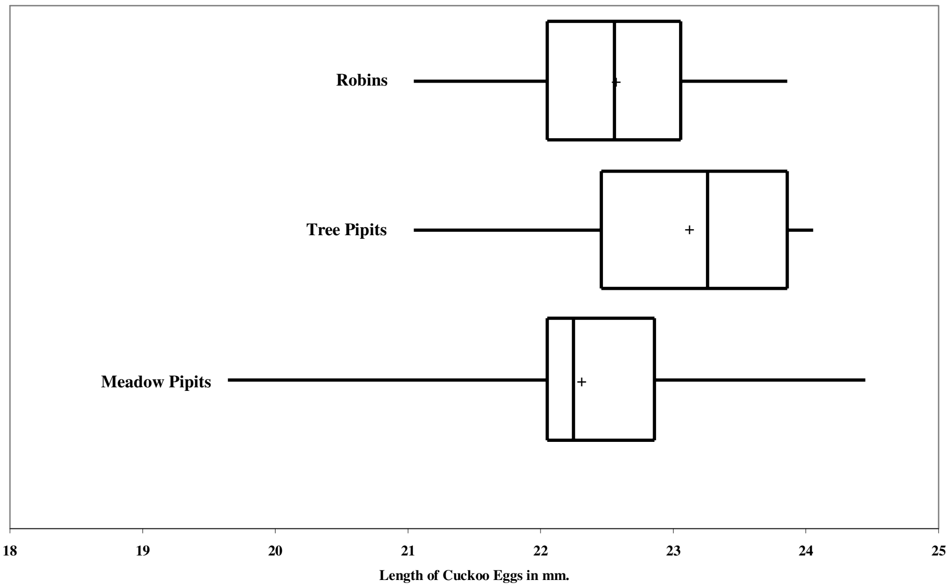
In Exercises 7–16, find the 5-number summaries, construct boxplots and compare the data sets.

7. **Cuckoo Egg Lengths** (Data Set 8)
   Meadow pipits, n= 45
   Minimum score is 19.65
   Q_1, the first quartile, also known as P_{25} is the 12th score, which is 22.05
   The median or Q_2, the second quartile, also known as P_{50} is the 23rd score or 22.25 (Mean is 22.30)
   Q_3, the third quartile, also known as P_{75} is the 34th score, which is 22.85
   Maximum score is 24.45

   Tree pipits, n= 15
   Minimum score is 21.05
   Q_1, the first quartile, also known as P_{25} is the 4th score, which is 22.45
   The median or Q_2, the second quartile, also known as P_{50} is the 8th score or 23.25 (Mean is 23.09)
   Q_3, the third quartile, also known as P_{75} is the 12th score, which is 23.85
   Maximum score is 24.05

   Robins, n= 16
   Minimum score is 21.05
   Q_1, the first quartile, also known as P_{25} is the average of the 4th and 5th scores, which is 22.05
   The median or Q_2, the second quartile, also known as P_{50} is the average of the 8th and 9th scores or 22.55 (Mean is 22.58)
   Q_3, the third quartile, also known as P_{75} is the average of the 12th and 13th scores, which is 23.05
   Maximum score is 23.85
   Boxplots including Meadow Pipits from Exercise 6

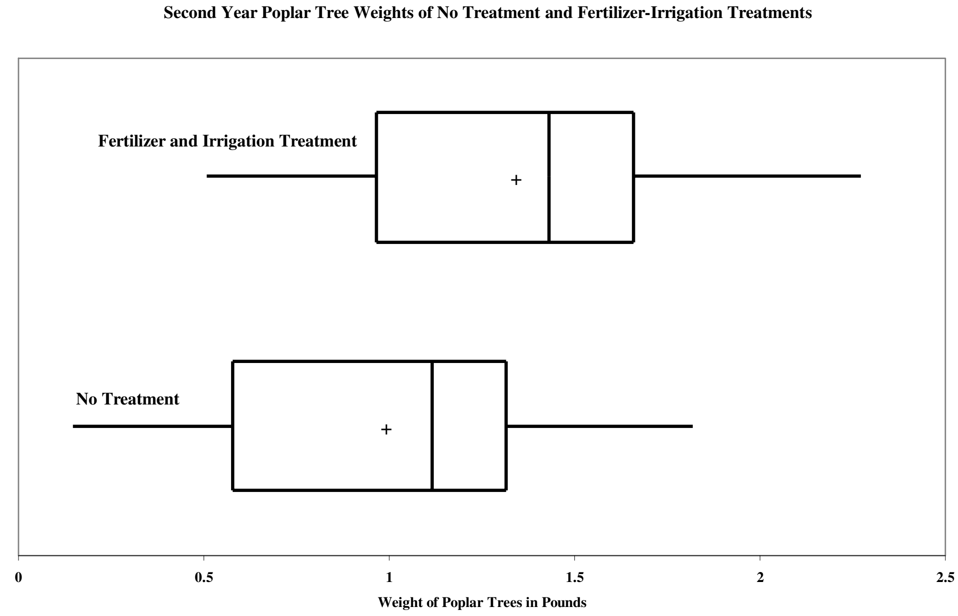
Length of Cuckoo Eggs Found in Three Types of Bird Nests



The length of cuckoo eggs is higher for the Tree Pipits nests and there is more variability of cuckoo eggs found in Meadow Pipits nests. The distribution shape is relatively symmetric for the Robins but is negatively skewed for both the Tree Pipits and the Meadow Pipits.

8. Poplar Trees (Data Set 9)
Second-year weights of trees given no treatment, *n*= 10
Minimum score is 0.13
Q_1, the first quartile, also known as P_{25} is the 3rd score, which is 0.56
The median or Q_2, the second quartile, also known as P_{50} is the average of the 5th and 6th scores or 1.10 (Mean is 0.97)
Q_3, the third quartile, also known as P_{75} is the 8th score, which is 1.30
Maximum score is 1.80

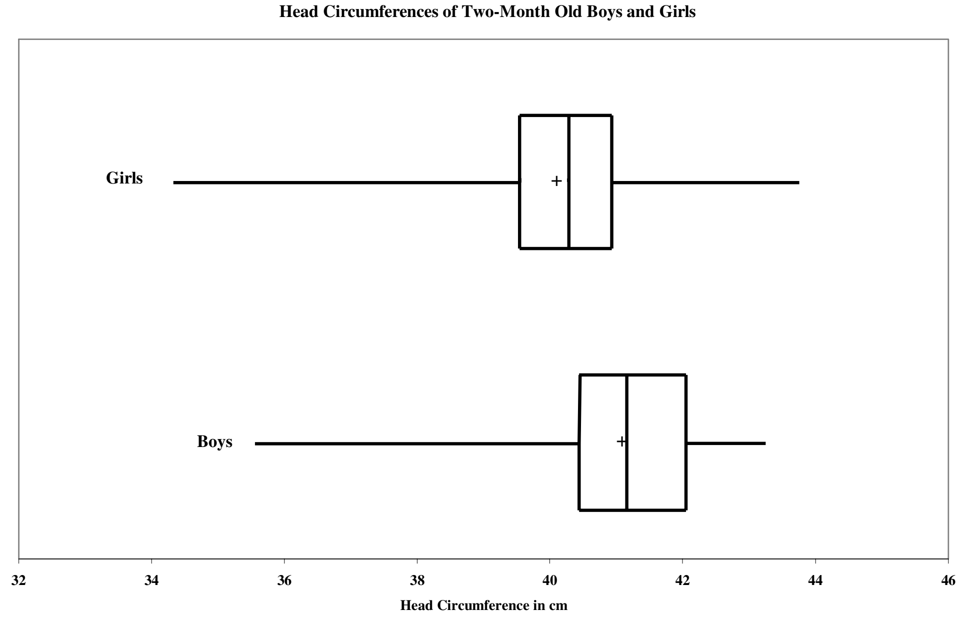
Second-year weights treated with fertilizer and irrigation, *n*= 10
Minimum score is 0.49
Q_1, the first quartile, also known as P_{25} is the 3rd score, which is 0.95
The median or Q_2, the second quartile, also known as P_{50} is the average of the 5th and 6th scores or 1.43 (Mean is 1.35)
Q_3, the third quartile, also known as P_{75} is the 8th score, which is 1.64
Maximum score is 2.25
Boxplots:



Clearly, there is a difference in the average weight of the poplar trees and it is in favor of the trees that were treated with fertilizer and irrigation. However, the degree of variability seems to be about the same between the trees in the two different treatments and both distributions appear to have some degree of negative skewness which is also confirmed by the fact that the means are lower than the medians.

9. **Head Circumference** (Data Set 4)
Boys, *n*= 50
Minimum score is 35.5
Q_1, the first quartile, also known as P_{25} is the 13th score, which is 40.4
The median or Q_2, the second quartile, also known as P_{50} is the average of the 25th and 26th scores or 41.1 (Mean is 41.1)
Q_3, the third quartile, also known as P_{75} is the 38th score, which is 42.0
Maximum score is 43.2

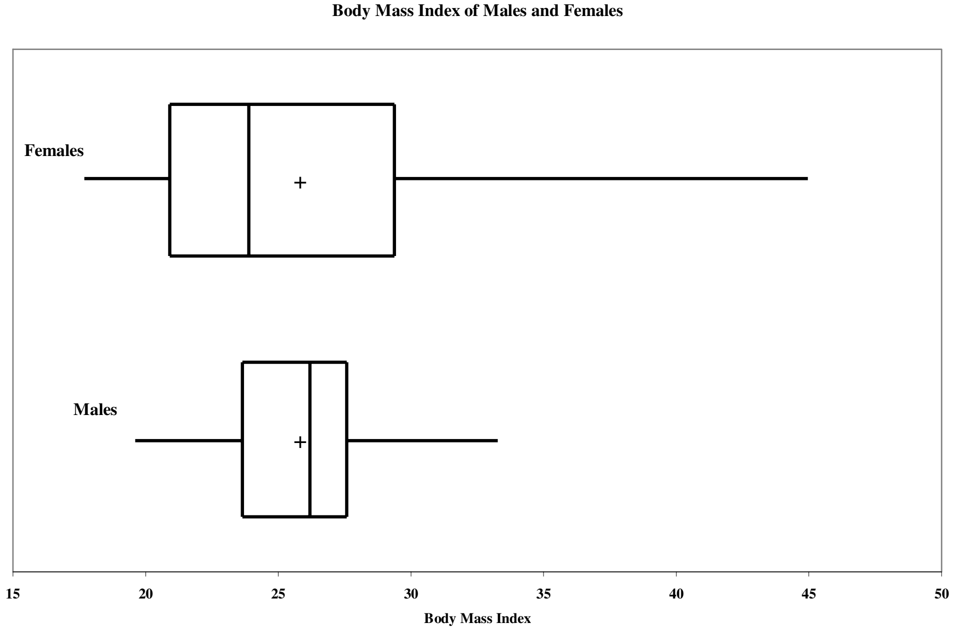
Girls, *n*= 50
Minimum score is 34.3
Q_1, the first quartile, also known as P_{25} is the 13th score, which is 39.5
The median or Q_2, the second quartile, also known as P_{50} is the average of the 25th and 26th scores or 40.2 (Mean is 40.1)
Q_3, the third quartile, also known as P_{75} is the 38th score, which is 40.9
Maximum score is 43.7
Boxplots:



The head circumference of the two-year old boys appears to be higher than that for the girls. The variability for the girls appears to be slightly higher than for the boys.

10. Body Mass Indexes (Data Set 1)
Males, *n*= 40
Minimum score is 19.60
Q_1, the first quartile, also known as P_{25} is the average of the 10th and 11th scores, which is 23.65
The median or Q_2, the second quartile, also known as P_{50} is the average of the 20th and 21st scores or 26.20
(Mean is 26.00)
Q_3, the third quartile, also known as P_{75} is the average of the 30th and 31st scores, which is 27.60
Maximum score is 33.20

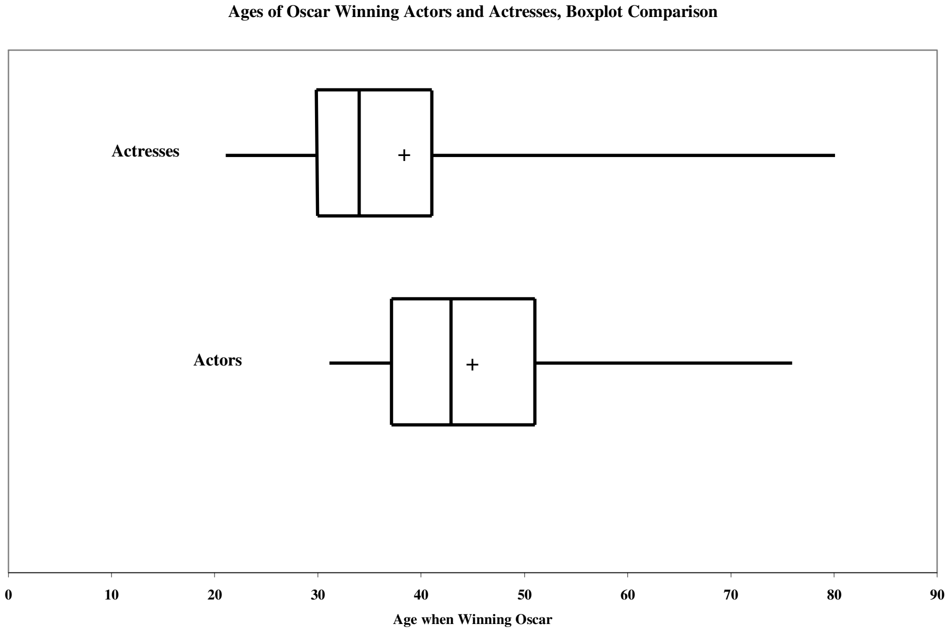
Females, *n*= 40
Minimum score is 17.70
Q_1, the first quartile, also known as P_{25} is the average of the 10th and 11th scores, which is 20.95
The median or Q_2, the second quartile, also known as P_{50} is the average of the 20th and 21st scores or 23.90
(Mean is 25.74)
Q_3, the third quartile, also known as P_{75} is the average of the 30th and 31st scores, which is 29.40
Maximum score is 44.90
Boxplots:



When looking at the medians, the average BMI for the males appears to be higher than that for the females. However when looking at the means, they are very close together. Clearly, the female group is positively skewed and that relates to the means being close together when the medians are not since the mean is being influenced in a positive direction by the extreme scores in the female group. The variability of the female BMI values is clearly higher than that for the males.

11. Ages of Oscar-Winning Actors and Actresses
Actors, *n*= 39
Minimum score is 31
Q_1, the first quartile, also known as P_{25} is the 10th score, which is 37
The median or Q_2, the second quartile, also known as P_{50} is the 20th score or 43 (Mean is 44.8)
Q_3, the third quartile, also known as P_{75} is the 30th score, which is 51
Maximum score is 76

Actresses, *n*= 39
Minimum score is 21
Q_1, the first quartile, also known as P_{25} is the 10th score, which is 30
The median or Q_2, the second quartile, also known as P_{50} is the 20th score or 34 (Mean is 38.1)
Q_3, the third quartile, also known as P_{75} is the 30th score, which is 41
Maximum score is 80
Boxplots:

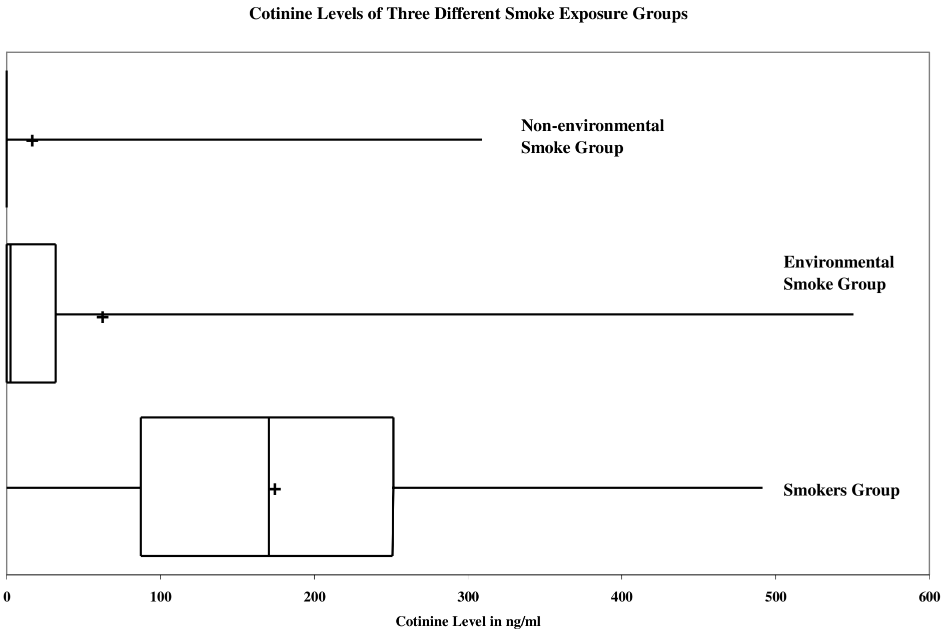


The average age for actresses when they won the Oscar was lower than the age of actors. There appears to be more variability of the actress ages compared with the actor ages and both distributions appear to have a degree of positive skewness.

12. Cotinine Levels (From Table 2-1)
Smokers, *n*= 40
Minimum score is 0.0
Q_1, the first quartile, also known as P_{25} is the average of the 10th and 11th scores, which is 86.5
The median or Q_2, the second quartile, also known as P_{50} is the average of the 20th and 21st scores or 170.0 (Mean is 172.5)
Q_3, the third quartile, also known as P_{75} is the average of the 30th and 31st scores, which is 251.5
Maximum score is 491.0

Environmental Tobacco Smoke (ETS), *n*=40
Minimum score is 0.0
Q_1, the first quartile, also known as P_{25} is the average of the 10th and 11th scores, which is 1.0
The median or Q_2, the second quartile, also known as P_{50} is the average of the 20th and 21st scores or 1.5 (Mean is 60.6)
Q_3, the third quartile, also known as P_{75} is the average of the 30th and 31st scores, which is 32.0
Maximum score is 551.0

Non-Environmental Exposure to Tobacco Smoke (NOETS), *n*= 40
Minimum score is 0.0
Q_1, the first quartile, also known as P_{25} is the average of the 10th and 11th scores, which is 0.0
The median or Q_2, the second quartile, also known as P_{50} is the average of the 20th and 21st scores or 0.0 (Mean is 16.4)
Q_3, the third quartile, also known as P_{75} is the average of the 30th and 31st scores, which is 0.0
Maximum score is 309.0
Boxplots:

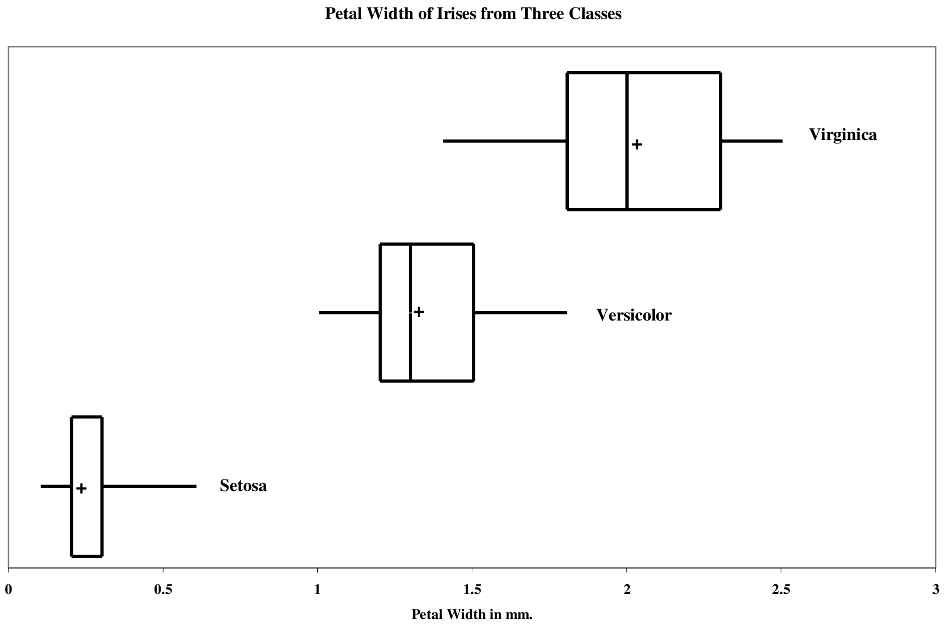


These boxplots clearly indicate that the highest average cotinine levels are for smokers and they also have the greatest variability of cotinine levels. The averages as well as variances for both the group with non-environmental smoke present and the group with environmental smoke present are low when compared with the smokers group. In addition, the distributions of non-environmental smoke exposure and environmental smoke exposure are highly positively skewed with the smoker distribution have some, but nearly as much, positive skewness.

13. Petal Widths of Irises (Data Set 7)
Setosa, *n*= 50
Minimum score is 0.1
Q_1, the first quartile, also known as P_{25} is the 13th score, which is 0.2
The median or Q_2, the second quartile, also known as P_{50} is the average of the 25th and 26th scores or 0.2 (Mean is 0.24)
Q_3, the third quartile, also known as P_{75} is the 38th score, which is 0.3
Maximum score is 0.6

Versicolor, *n*= 50
Minimum score is 1.0
Q_1, the first quartile, also known as P_{25} is the 13th score, which is 1.2
The median or Q_2, the second quartile, also known as P_{50} is the average of the 25th and 26th scores or 1.3 (Mean is 1.326)
Q_3, the third quartile, also known as P_{75} is the 38th score, which is 1.5
Maximum score is 1.8

Virginica, *n*= 50
Minimum score is 1.4
Q_1, the first quartile, also known as P_{25} is the 13th score, which is 1.8
The median or Q_2, the second quartile, also known as P_{50} is the average of the 25th and 26th scores or 2.0 (Mean is 2.03)
Q_3, the third quartile, also known as P_{75} is the 38th score, which is 2.3
Maximum score is 2.5
Boxplots:



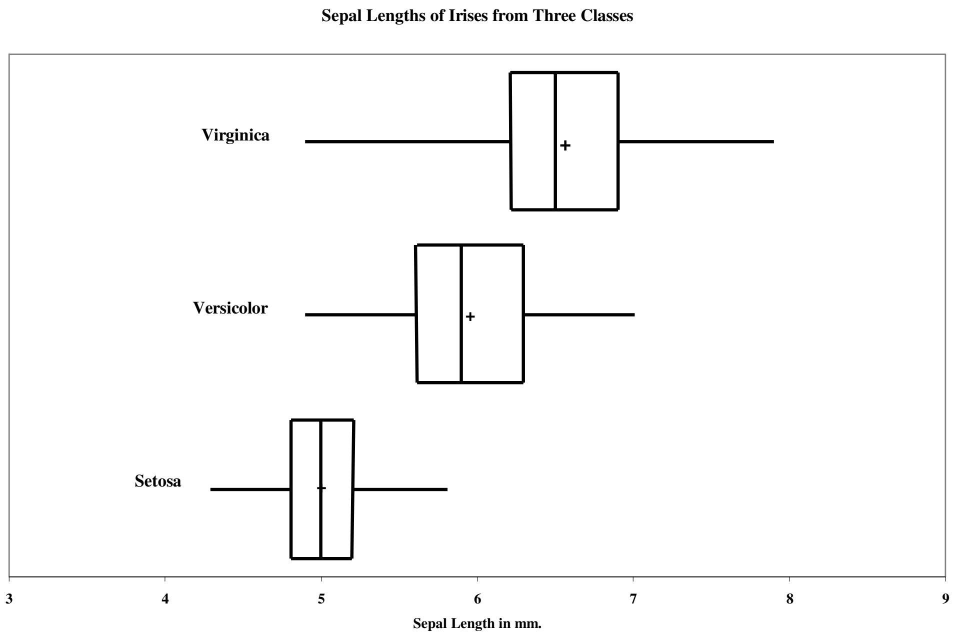
There are very clear differences in the average length of the petal widths among these three types of irises. The highest length is for the Virginica, the second highest length is for the Versicolor, and the lowest length is for the Setosa. The lowest variation of petal width was for Setosa, Veriscolor had the next highest variation, and the Virginica had the highest variation of petal lengths.

14. Sepal Lengths of Irises (Data Set 7)
Setosa, *n*= 50

Minimum score is 4.3
Q_1, the first quartile, also known as P_{25} is the 13th score, which is 4.8
The median or Q_2, the second quartile, also known as P_{50} is the average of the 25th and 26th scores or 5.0 (Mean is 5.01)
Q_3, the third quartile, also known as P_{75} is the 38th score, which is 5.2
Maximum score is 5.8

Versicolor, *n*= 50
Minimum score is 4.9
Q_1, the first quartile, also known as P_{25} is the 13th score, which is 5.6
The median or Q_2, the second quartile, also known as P_{50} is the average of the 25th and 26th scores or 5.9 (Mean is 5.94)
Q_3, the third quartile, also known as P_{75} is the 38th score, which is 6.3
Maximum score is 7.0

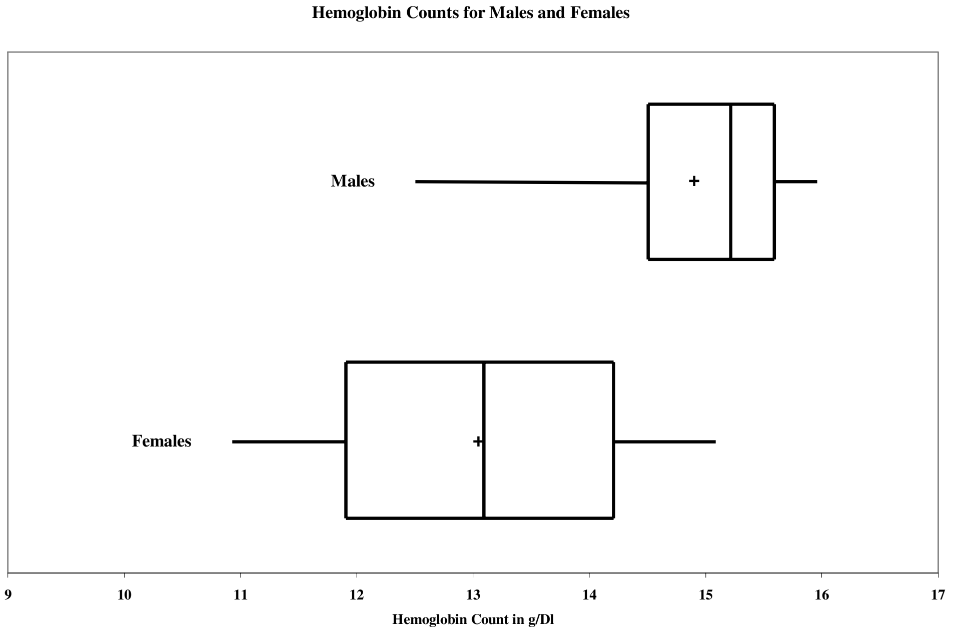
Viginica, *n*= 50
Minimum score is 4.9
Q_1, the first quartile, also known as P_{25} is the 13th score, which is 6.2
The median or Q_2, the second quartile, also known as P_{50} is the average of the 25th and 26th scores or 6.5 (Mean is 6.59)
Q_3, the third quartile, also known as P_{75} is the 38th score, which is 6.9
Maximum score is 7.9
Boxplots:



Clearly, there are differences in sepal length of irises among the three types. The longest is the Virginica, next longest is the Veriscolor, and the shortest is the Setosa. The variability is highest for the Virginica, followed by the Versicolor, and lowest for the Setosa. All three distributions were relatively symmetrical.

15. Hemoglobin Counts (Data Set 10)
Females, n= 27
Minimum score is 10.95
Q_1, the first quartile, also known as P_{25} is the 7th score, which is 11.90
The median or Q_2, the second quartile, also known as P_{50} is the 14th score or 13.10 (Mean is 13.05)
Q_3, the third quartile, also known as P_{75} is the 21st score, which is 14.20
Maximum score is 15.10

Males, n= 23
Minimum score is 12.50
Q_1, the first quartile, also known as P_{25} is the 6th score, which is 14.50
The median or Q_2, the second quartile, also known as P_{50} is the 12th score or 15.20 (Mean is 14.91)
Q_3, the third quartile, also known as P_{75} is the 18th score, which is 15.60
Maximum score is 15.95
Boxplots:

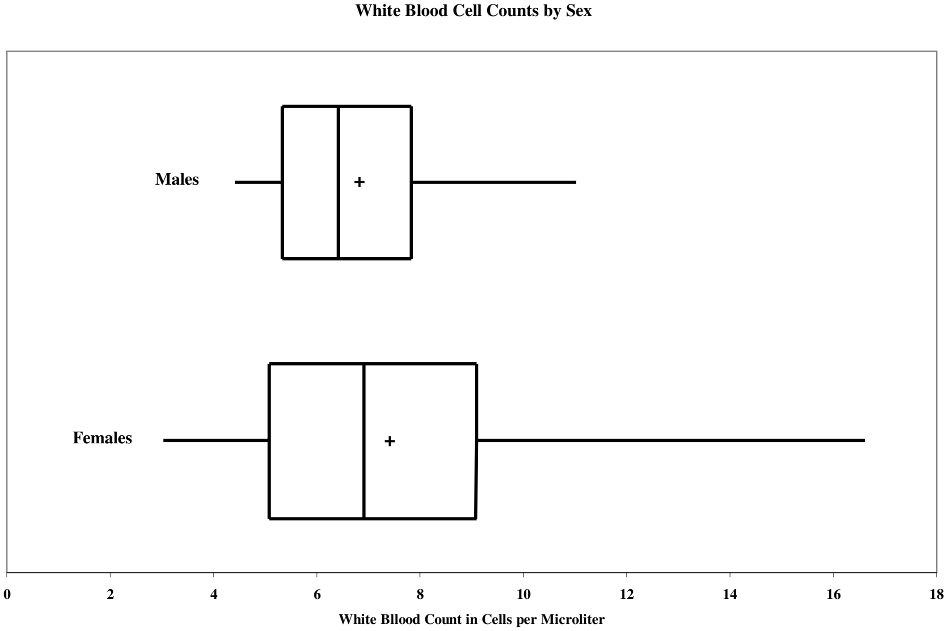


The average hemoglobin count is higher for the male group than it is for the female group. The female group appears to be more variable than the male group. The female distribution appears to be relatively symmetrical while the male distribution appears to be negatively skewed.

16. White Blood Cell Counts (Data Set 10)

Females, *n*= 27
Minimum score is 3.00
Q_1, the first quartile, also known as P_{25} is the 7th score, which is 5.05
The median or Q_2, the second quartile, also known as P_{50} is the 14th score or 6.90 (Mean is 7.44)
Q_3, the third quartile, also known as P_{75} is the 21st score, which is 9.10
Maximum score is 16.60

Males, *n*= 23
Minimum score is 4.40
Q_1, the first quartile, also known as P_{25} is the 6th score, which is 5.30
The median or Q_2, the second quartile, also known as P_{50} is the 12th score or 6.40 (Mean is 6.79)
Q_3, the third quartile, also known as P_{75} is the 18th score, which is 7.85
Maximum score is 11.00
Boxplots:



The average white cell count appears slightly higher for the female group than it is for the male group. The female group appears to be more variable than the male group. Both distributions appear to be positively skewed.

Review Exercises

1. **Tree Heights**, *n*= 20

a. Mean: $\bar{x} = \sum x / n = 90.7/20 = 4.54$

b. Median: $\tilde{x}$ is the midpoint between the two middle scores (10th and 11th scores) if an even number of scores. This would be (3.9 + 4.0)/2= 7.9/2= 3.95

c. Mode: Mode is score that occurs most often. In this case, 3 scores occur more than any others (1.8, 3.7, and 5.1) so the distribution would be called trimodal with modes at 1.8, 3.7, and 5.1

d. Midrange: Midpoint (or average) between lowest and highest scores, (1.8 + 13.7)/2 = 15.5/2= 7.75

e. Range = highest score – lowest score = 13.7 – 1.8 = 11.9

f. Standard deviation

$$s = \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n(n-1)}} = \sqrt{\frac{20(544.85) - (90.7)^2}{20(20-1)}} = \sqrt{\frac{2670.51}{380}} = \sqrt{7.027} = 2.65$$

g. Variance $s^2 = 2.65^2 = 7.03$

h. Q_1 is the first quartile, also known as P_{25}, $L = \frac{k}{100} * n = \frac{25}{100} * 20 = 0.25 * 20 = 5$

Q_1 is the average of the 5th and 6th scores, which is (3.1 + 3.4)/2= 6.5/2= 3.25

i. Q_3, the third quartile, also known as P_{75}, $L = \frac{k}{100} * n = \frac{75}{100} * 20 = 0.75 * 20 = 15$

Q_3, is the average of the 15th and 16th scores, which is (5.1 + 5.2)/2= 10.3/2= 5.15

j. P_{10} is the 10th percentile, k is the percentile (10), $L = \frac{k}{100} * n = \frac{10}{100} * 20 = 0.10 * 20 = 2$

Since L is a whole number, in this case 2, P_{10} is taken as the average of the 2nd and 3rd scores or (1.8 + 1.9)/2= 1.85

2. Circumference of 13.7 ft.

a. z score $z = \frac{x - \bar{x}}{s} = \frac{13.7 - 4.54}{2.65} = \frac{9.16}{2.65} = 3.46$

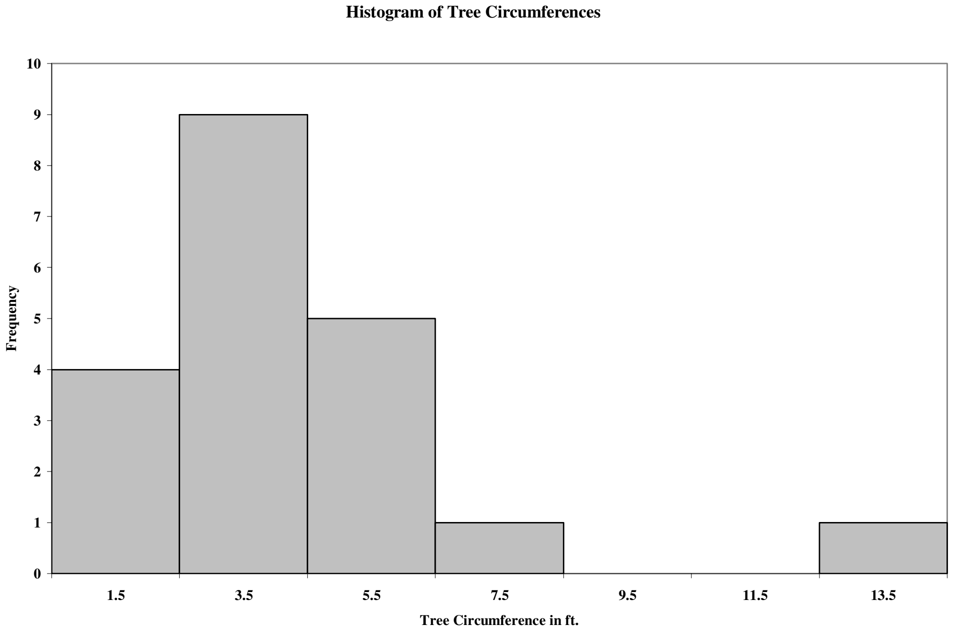
b. This would be considered "unusual" since it falls out of the range of the $\bar{x} \pm 2s$ or has a z value lower than -2 or above +2.

c. Any value below or above 2 standard deviations from the mean would be considered "unusual" In this case, that would be 4.54 ± 2(2.65) = 4.54 ± 5.30 or -0.76 to 9.84. Of course, a value of -0.76 or any negative circumference cannot exist, so any circumference above 9.84 would be considered "unusual"

3. Frequency Distribution

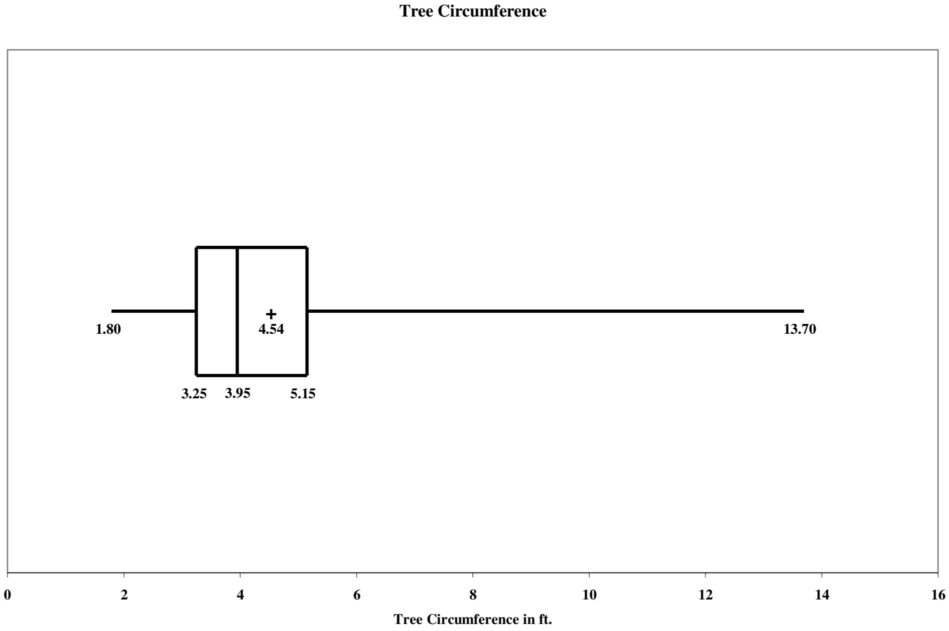
Tree Circumference in ft.	Frequency
1.0 – 2.9	4
3.0 – 4.9	9
5.0 – 6.9	5
7.0 – 8.9	1
9.0 – 10.9	0
11.0 – 12.9	0
13.0 – 14.9	1

4. Histogram



The distribution of this sample of tree circumferences is positively skewed.

5. Boxplot
5-number summary
Minimum score is 1.80
Q_1, the first quartile, also known as P_{25} is the average of the 5th and 6th scores, which is 3.25
The median or Q_2, the second quartile, also known as P_{50} is the average of the 10th and 11th scores or 3.95 (Mean is 4.54)
Q_3, the third quartile, also known as P_{75} is the average of the 15th and 16th scores, which is 5.15
Maximum score is 13.70
Boxplot:



Cumulative Review Exercises

1. Tree Measurements

a. The measures are <u>continuous</u> since the unit of measurement of feet can take on fractional values.

b. The level of measurement is <u>ratio</u> since there is a natural zero or beginning point on the scale. There are no negative values on a ratio scale.

2. a. The <u>mode</u> is the only value that can be used to represent a nominal level of measurement value. Nominal scale variables do not have order needed for finding the median and midrange and it does not have order and equal intervals that would be needed to find the mean.

b. This would be a sample of <u>convenience</u> since the possible homes that could be included in the sample do not have an equal chance of being selected to be included in the sample.

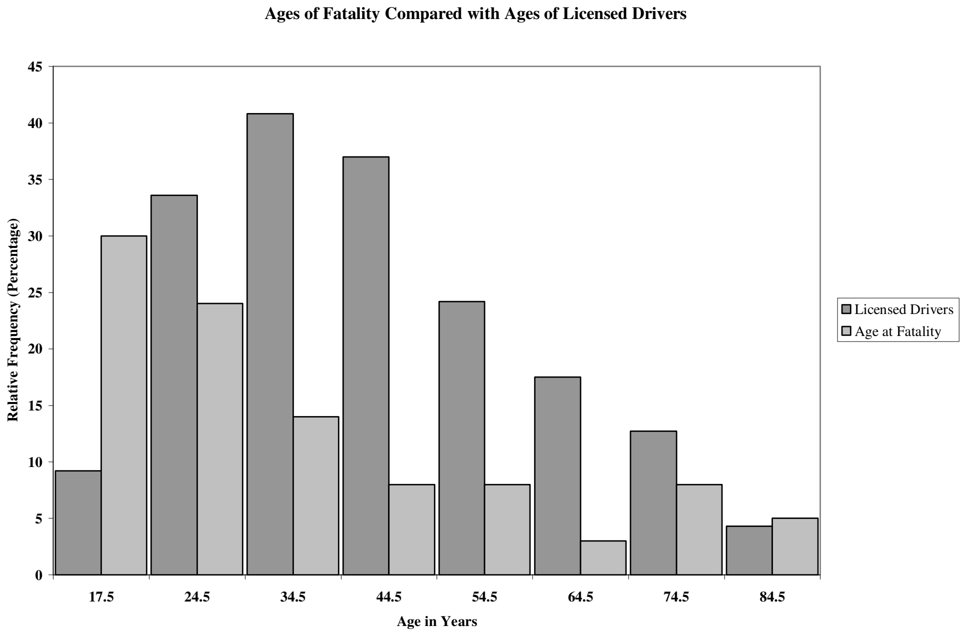
c. This would be a <u>cluster</u> sample since clusters, in the form of polling places, are selected at random and all the voters coming out of the polling places are surveyed.

d. The primary statistic of importance in this situation would be the <u>standard deviation</u>. This measures the variance of the fertilizer in the sticks. The desire would be to decrease this value which would mean there is more consistency or homogeneity of the amount of fertilizer in the sticks while keeping the mean at an acceptable level..

Critical Thinking

Age at Fatality	Relative Frequency of Licensed Drivers (millions)	Relative Frequency of Age of Drivers Killed in Car Crashes (Randomly selected)
16 – 19	9.2	30.0
20 – 29	33.6	24.0
30 – 39	40.8	14.0
40 – 49	37.0	8.0
50 – 59	24.2	8.0
60 – 69	17.5	3.0
70 – 79	12.7	8.0
80 – 89	4.3	5.0

The age category of 16 – 19 clearly has a substantially higher percentage of fatalities in the sample of 100 than in the group of licensed drivers. It would seem justified to charge higher premiums for the 16 – 19 age group. The 80 – 89 age group has a slightly higher percentage for the ages at fatality. When one considers the absolute number of drivers involved, this may be another group that should be charged higher premiums. The following graph displays this comparison:


Ages of Fatality Compared with Ages of Licensed Drivers
Relative Frequency (Percentage)
0
5
10
15
20
25
30
35
40
45
17.5
24.5
34.5
44.5
54.5
64.5
74.5
84.5
Age in Years
Licensed Drivers
Age at Fatality