Chapter 2

Summarizing and Graphing Data

2-2 Frequency Distributions

1. No. The first class frequency, for example, tells us only that there were 18 pennies with weights in the 2.40-2.49 grams class, but there is no way to tell the exact values of those 18 weights.

2. The sum of the relative frequencies should be 1.00 when proportions are used, and it should be 100% when percentages are used.

3. No. This is not a relative frequency distribution because the sum of the percentages is not 100%. It appears that each respondent was asked to indicate whether he downloaded the four types of material (and so the sum of the percentages could be anywhere from 0% to 400%), and not to place himself in one of the four categories (in which case the table would be a relative frequency distribution and the sum of the percentages would be 100%).

4. The gap in the frequencies suggests the table includes heights from two different populations. Considering the values, it appears that the two populations are elementary students and faculty/staff personnel at the school.

5. a. Class width: subtracting the first two lower class limits, 14−10 = 4.
   b. Class midpoints: the first class midpoint is (10+13)/2 = 11.5, and the others can be obtained by adding the class width to get 11.5, 15.5, 19.5, 23.5, 27.5.
   c. Class boundaries: the boundary between the first and second class is (13+14)/2 = 13.5, and the others can be obtained by adding or subtracting the class width to get 9.5, 13.5, 17.5, 21.5, 25.5, 29.5.

6. a. Class width: subtracting the first two lower class limits, 6−2 = 4.
   b. Class midpoints: the first class midpoint is (2+5)/2 = 3.5, and the others can be obtained by adding the class width to get 3.5, 7.5, 11.5, 15.5.
   c. Class boundaries: the boundary between the first and second class is (5+6)/2 = 5.5, and the others can be obtained by adding or subtracting the class width to get 1.5, 5.5, 9.5, 13.5, 17.5.

7. a. Class width: subtracting the first two lower class limits, 1.00−0.00 = 1.00.
   b. Class midpoints: the first class midpoint is (0.00+0.99)/2 = 0.495, and the others can be obtained by adding the class width to get 0.495, 1.495, 2.495, 3.495, 4.495.
   c. Class boundaries: the boundary between the first and second class is (0.99+1.00)/2 = 0.995, and the others can be obtained by adding or subtracting the class width to get -0.005, 0.995, 1.995, 2.995, 3.995, 4.995.

8. a. Class width: subtracting the first two lower class limits, 1.00−0.00 = 1.00.
   b. Class midpoints: the first class midpoint is (0.00+0.99)/2 = 0.495, and the others can be obtained by adding the class width to get 0.495, 1.495, 2.495, 3.495, 4.495, 5.495
   c. Class boundaries: the boundary between the first and second class is (0.99+1.00)/2 = 0.995, and the others can be obtained by adding or subtracting the class width to get -0.005, 0.995, 1.995, 2.995, 3.995, 4.995, 5.995.

9. a. Strict interpretation: No; because there are more values at the upper end, there is not symmetry.
   b. Loose interpretation: Yes; there is a concentration of frequencies at the middle and a tapering off in both directions.

10. a. Strict interpretation: No; the concentration of values is at the upper end.
    b. Loose interpretation: No; the concentration of values is at the upper end.

11. The requested figure is given below at the left. Obtain each relative frequency by dividing the given frequency by 25, the total number of observations in each table. The "total" line is not necessary.

    The non-filtered cigarettes have much more tar. Yes, the filters appear to be effective in reducing the amount of tar.

Relative Frequency Comparison for #11

	cigarette type	
tar (mg)	non-filtered	filtered
2–5	0%	8%
6–9	0%	8%
10–13	4%	24%
14–17	0%	60%
18–21	60%	0%
22–25	28%	0%
26–29	8%	0%
total	100%	100%

Relative Frequency Comparison for #12

	discard type	
weight (lbs)	metal	plastic
0.00–0.99	8.1%	22.6%
1.00–1.99	41.9%	32.3%
2.00–2.99	24.2%	33.9%
3.00 –3.99	19.4%	6.5%
4.00–4.99	6.5%	3.2%
5.00–5.99	0.0%	1.6%
total	100%	100%

12. The requested figure is given above at the right. Obtain each relative frequency by dividing the given frequency by 62, the total number of observations in each table. The "total" line is not necessary. [Due to rounding, percentages actually sum to 100.1%.]

    While the weights cover approximately the same range, it appears that the weights for the plastic are slightly smaller.

NOTE: For cumulative tables, this manual uses upper class boundaries in the "less than" column. Consider exercise #13, for example, to understand why is done. Conceptually, weights occur on a continuum and the integer values reported are assumed to be the nearest whole number representation of the precise measure. An exact weight of 17.7, for example, would be reported as 18 and fall in the third class. The values in the second class, therefore, are better described as "less than 17.5" (using the upper class boundary) than as "less than 18" (using the lower class limit of the next class). This distinction is crucial in the construction of pictorial representations in the next section. To present a visually simpler table, however, it is common practice to follow the example in the text and use the lower class limit of the next class. Regardless of the "less than" label, the final cumulative frequency must equal the total sample size – and the sum of the cumulative frequency column has no meaning and should never be included.

13. Obtain the cumulative frequency values by adding the given frequencies.

tar (mg) in non-filtered cigarettes	cumulative frequency
less than 13.5	1
less than 17.5	1
less than 21.5	16
less than 25.5	23
less than 29.5	25

14. Obtain the cumulative frequency values by adding the given frequencies.

tar (mg) in filtered cigarettes	cumulative frequency
less than 5.5	2
less than 9.5	4
less than 13.5	10
less than 17.5	25

15. Obtain the relative frequencies by dividing the given frequencies by the total of 2223.

category	relative frequency
male survivors	16.2%
males who dies	62.8%
female survivors	15.5%
females who died	5.5%
	100.0%

16. Obtain the relative frequencies by dividing the given frequencies by the total of 570.

those whose smoking…	relative frequency
continued after the gum	33.5%
stopped after the gum	10.4%
continued after the patch	46.1%
stopped after the patch	10.0%
	100.0%

17. The requested table is given below, with the preliminary Excel output at the right. The frequency distribution of the last digits shows unusually high numbers of 0's and 5's. This is typical for data that have been rounded off to "convenient" values. It appears that the heights were reported and not actually measured.

digit	frequency
0	9
1	2
2	1
3	3
4	1
5	15
6	2
7	0
8	3
9	1
	37

Bin	*Frequency*
0.5	9
1.5	2
2.5	1
3.5	3
4.5	1
5.5	15
6.5	2
7.5	0
8.5	3
9.5	1
More	0

18. The requested table is given below, with the preliminary Excel output at the right. The data are assumed to relate to the 1979 nuclear power plant accident at Three Mile Island. Such data are important because they can be helpful in detecting potentially dangerous situations and in making recommendations for future action.

level of strontium-90	frequency
110–119	2
120–129	2
130–139	5
140–149	9
150–159	13
160–169	6
170–179	2
180–189	1
	40

Bin	*Frequency*
119.5	2
129.5	2
139.5	5
149.5	9
159.5	13
169.5	6
179.5	2
189.5	1
More	0

19. The requested table is given below, with the preliminary Excel output at the right.

nicotine (mg)	frequency
1.0–1.1	14
1.2–1.3	4
1.4–1.5	3
1.6–1.7	3
1.8–1.9	1
	25

Bin	*Frequency*
1.15	14
1.35	4
1.55	3
1.75	3
1.95	1
More	0

20. The requested table is given below, with the preliminary Excel output at the right. The values appear to be lower than the unfiltered ones in exercise #19.

nicotine (mg)	frequency
0.2–0.3	1
0.4–0.5	1
0.6–0.7	1
0.8–0.9	8
1.0–1.1	12
1.2–1.3	2
	25

Bin	*Frequency*
0.35	1
0.55	1
0.75	1
0.95	8
1.15	12
1.35	2
More	0

21. The requested table is given below, with the preliminary Excel output at the right. No, the voltages do not appear to follow a normal distribution – instead of being concentrated near the middle of the distribution, the values appear to be rather evenly distributed.

voltage (volts)	frequency
123.3–123.4	10
123.5–123.6	9
123.7–123.8	10
123.9–124.0	10
124.1–124.2	1
	40

Bin	*Frequency*
123.45	10
123.65	9
123.85	10
124.05	10
124.25	1
More	0

22. The requested table is given below, with the preliminary Excel output at the right. Yes, the voltages do appear to follow a normal distribution – there are many values near the center of the distribution, and the frequencies diminish toward either end. The values appear to be higher than those in exercise #21.

voltage (volts)	frequency
123.9–124.0	2
124.1–124.2	1
124.3–123.4	6
124.5–125.6	9
124.7–124.8	13
124.9–125.0	5
125.1–125.2	4
	40

Bin	*Frequency*
124.05	2
124.25	1
124.45	6
124.65	9
124.85	13
125.05	5
125.25	4
More	0

23. The requested table is given below, with the preliminary Excel output at the right. While over half of the screws are within 0.01 inches of the claimed value (28 of 50 fall between 0.74 and 0.76), there are over twice as many screws below that range as there are above it (15 vs. 7). It appears that there might be a slight tendency to err on the side of making the screws too small.

length (in)	frequency
0.720–0.729	5
0.730–0.739	10
0.740–0.749	11
0.750–0.759	17
0.760–0.769	7
	50

Bin	*Frequency*
0.7295	5
0.7395	10
0.7495	11
0.7595	17
0.7695	7
More	0

24. The requested table is given below, with the preliminary Excel output at the right. Yes, the weights appear to have a distribution that is approximately normal. These weights are considerably higher than the weights in exercise #7.

weight (lbs)	frequency
1.00– 4.99	8
5.00– 8.99	21
9.00–12.99	22
13.00–16.99	8
17.00–20.99	3
	62

Bin	*Frequency*
4.995	8
8.995	21
12.995	22
16.995	8
20.995	3
More	0

25. The requested table is given below, with the preliminary Excel output at the right. The ratings appear to have a distribution that is not normal. While there is a maximum score with progressively smaller frequencies on either side of the maximum, the distribution is definitely not symmetric (i.e., the maximum score is not near the middle, but at the upper end of the distribution).

FICO score	frequency
400–449	1
450–499	1
500–549	5
550–599	8
600–649	12
650–699	16
700–749	19
750–799	27
800–849	10
850–899	1
	100

Bin	*Frequency*
449.5	1
499.5	1
549.5	5
599.5	8
649.5	12
699.5	16
749.5	19
799.5	27
849.5	10
899.5	1
More	0

26. The requested tables are given below, with the preliminary Excel output at the right. In each case the relative frequencies were obtained by dividing the observed frequencies by 36.

REGULAR COKE	
weight (lbs)	relative frequency
0.7900–0.7949	2.8%
0.7950–0.7999	0.0%
0.8000–0.8049	2.8%
0.8050–0.8099	8.3%
0.8100–0.8149	11.1%
0.8150–0.8199	47.2%
0.8200–0.8249	16.7%
0.8250–0.8299	11.1%
	100.0%

DIET COKE	
weight (lbs)	relative frequency
0.7750–0.7799	11.1%
0.7800–0.7849	36.1%
0.7850–0.7899	41.7%
0.7900–0.7949	11.1%
	100.0%

Bin	*Frequency*
0.79495	1
0.79995	0
0.80495	1
0.80995	3
0.81495	4
0.81995	17
0.82495	6
0.82995	4
More	0

Bin	*Frequency*
0.77995	4
0.78495	13
0.78995	15
0.79495	4
More	0

There are two significant differences between the data sets: the weights for Regular Coke are considerably larger than those for Diet Coke, and the weights for Regular Coke cover a much wider range than those for Diet Coke. This suggests that the sweetener in Regular Coke adds weight to the product and does not distribute evenly throughout the product. As the company produces more Regular Coke than Diet Coke, another possibility is that the harder-working machines filling the Regular Coke may not be holding their tolerance as well – and a wider range in volume dispensed might account for the wider range of weights for Regular Coke.

27. The requested table is given below, with the preliminary Excel output at the right.

weight (g)	frequency
6.0000–6.0499	2
6.0500–6.0999	3
6.1000–6.1499	10
6.1500–6.1999	8
6.2000–6.2499	6
6.2500–6.2999	7
6.3000–6.3499	3
6.3500–6.3999	1
	40

Bin	*Frequency*
6.04995	2
6.09995	3
6.14995	10
6.19995	8
6.24995	6
6.29995	7
6.34995	3
6.39995	1
More	0

28. The requested table us given below, with the preliminary Excel output at the right. The post-1964 quarters appear to have weights that are lighter (due to their different metallic composition) and spread over a smaller range (due to their fewer years in circulation).

weight (g)	frequency
5.5000–5.5499	3
5.5500–5.5999	9
5.6000–5.6499	11
5.6500–5.6999	9
5.7000–5.7499	7
5.7500–5.7999	1
	40

Bin	*Frequency*
5.54995	3
5.59995	9
5.64995	11
5.69995	9
5.74995	7
5.79995	1
More	0

29. The requested table is given below, with the preliminary Excel output at the right.

blood group	frequency
O	22
A	20
B	5
AB	3
	50

Group	Count	%
A	20	40
AB	3	6
B	5	10
O	22	44

30. The requested table is given below, with the preliminary Excel output at the right.

main cause	frequency
bad track	23
faulty equipment	9
human error	12
other	6
	50

Group	Count	%
E	9	18
H	12	24
O	6	12
T	23	46

31. The frequency distributions including and excluding the outlier are given below, with the preliminary Excel output at the right. In general, an outlier can add several rows to a frequency distribution. Even though most of the added rows have frequency zero, the table tends to suggest that these are possible values – thus distorting the reader's mental image of the distribution.

0.0111 CANS (with the outlier)

weight (lbs)	frequency
200 – 219	6
220 – 239	5
240 – 259	12
260 – 279	36
280 – 299	87
300 – 319	28
320 – 339	0
340 – 359	0
360 – 379	0
380 – 399	0
400 – 419	0
420 – 439	0
440 – 459	0
460 – 479	0
480 – 499	0
500 – 519	1
	175

0.0111 CANS (without the outlier)

weight (lbs)	frequency
200 – 219	6
220 – 239	5
240 – 259	12
260 – 279	36
280 – 299	87
300 – 319	28
	174

Bin	*Frequency*
219.5	6
239.5	5
259.5	12
279.5	36
299.5	87
319.5	28
339.5	0
359.5	0
379.5	0
399.5	0
419.5	0
439.5	0
459.5	0
479.5	0
499.5	0
519.5	1
More	0

32. Let n = the number of data values and let x = the number of classes.
Either (1) solve the given formula $x = 1 + (\log n)/(\log 2)$ for n to get $n = 2^{x-1}$.
or (2) use trial-and-error by entering various values for n.

Use the values x = 5.5, 6.5, 7.5,… to get cut-off values for n shown below.

x	$n = 2^{x-1}$
5.5	22.63
6.5	45.25
7.5	90.51
8.5	181.02
9.5	362.04
10.5	724.04
11.5	1448.15
12.5	2896.31

Assuming n is at least 16, use the cut-off values to complete the table as follows.

n	ideal # of classes
16 – 22	5
23 – 45	6
46 – 90	7
91 – 181	8
182 – 362	9
363 – 724	10
725 – 1448	11
1449 – 2896	12

NOTE: Either the cut-off value method or the trial-and-error method indicates that
for $n < 22.63$, x rounds to 5.
for $22.63 < n < 45.25$, x rounds to 6.
for $45.25 < n < 90.51$, x rounds to 7.
for $90.51 < n < 181.02$, x rounds to 8.
etc.

2-3 Histograms

1. The pulse rate data have been organized into 7 classes. Examining the frequency distribution requires consideration of 14 pieces of information: the 7 class labels, and the 7 class frequencies. The histogram efficiently presents the same information in one visual image and gives all the relevant CVDOT (center, variation, distribution shape, outlier, [time is not relevant for these data]) details in an intuitive format.

2. Not necessarily. Depending on how the potential subjects were approached, a voluntary response sample of health data might fail to be representative of the general population for in the following ways.
   (a) Thinking they might receive free health information, those with health problems might be more likely to volunteer.
   (b) To avoid looking bad when compared with their peers, those with health problems might be less likely to volunteer.
   (c) The pool of potential volunteers may have been approached and/or identified so as to be more homogeneous in some manner (racially, ethnically, etc.) than the general population, and hence the sample data would not reflect the true range of values in the general population.

3. The data set is small enough that the individual numbers can be examined; they do not require summarization in a figure. The data set is not large enough for a histogram to reveal the true nature of the distribution; the histogram will essentially be a repeat of the individual numbers.

4. In ordinary language, "normal" refers to that which is most common; in statistics, "normal" refers to a specific pattern of values. A normal distribution is characterized by a distribution that is approximately bell-shaped (i.e., bunching up in the middle, and tapering off symmetrically at either end). Determining whether a distribution is approximately bell-shaped requires subjective judgment.

NOTE: For exercises 5-8, the following values are used to answer the questions. It appears that the midpoints of the first 3 classes are 5,000 and 10,000 and 15,000. It appears that the heights of the 5 bars are 2, 30, 8, 15, 5.

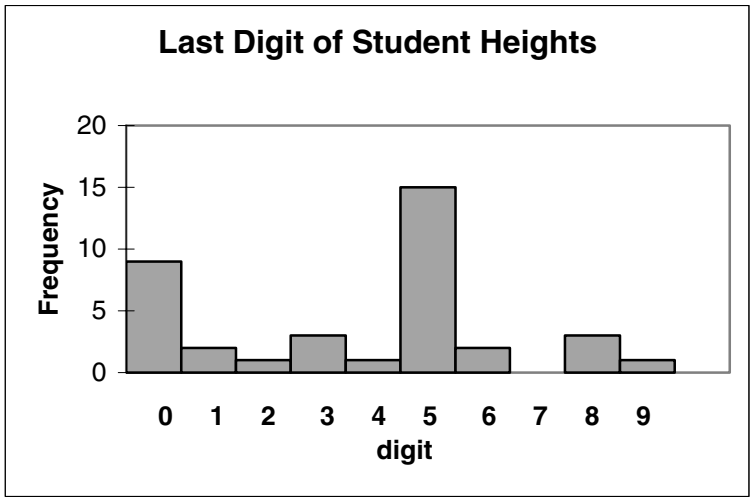
5. a. Adding the heights of all the bars, the total number is 2+30+8+15+5 = 60.
   b. Adding the heights of the two rightmost bars, the number over 20,000 miles is 15+5 = 20.

6. a. Subtracting the first two midpoints, the class width is 10,000−5,000 = 5,000 miles.
   b. The upper class boundary of the first class is the average of the first two class midpoints, (5,000+10,000)/2 = 7500. The lower class boundary of the first class is the upper class boundary minus the class width, 7500−5000 = 2500. While it is unclear whether a reading of exactly 7500 miles would fall into the first or second class, for example, the approximate lower and upper limits of the first class are 2500 miles and 7500 miles.

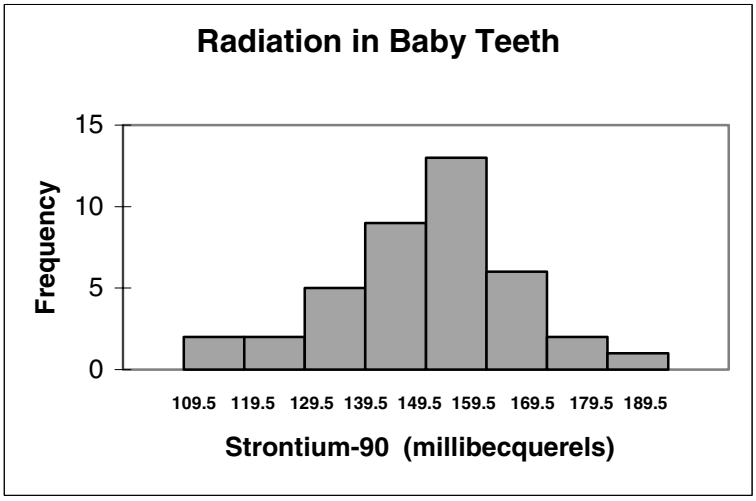
7. a. The minimum possible miles traveled is the lower class boundary associated with the leftmost bar, 2500 miles.
   b. The maximum possible number of miles traveled is the upper class boundary associated with the rightmost bar, 42,500 miles.

8. The histogram appears to include mileage amounts from two different populations. These most likely represent automobiles which are driven in and out of the city each day but are parked during the day (cars belonging to commuters) and automobiles that are driven during the day (taxis, messenger and/or delivery cars).

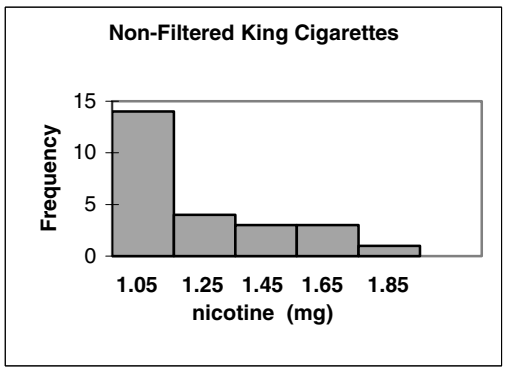
9. The histogram is given below. The digits 0 and 5 occur disproportionately more than the others. This is typical for data that have been rounded off to "convenient" values. It appears that the heights were reported and not actually measured.

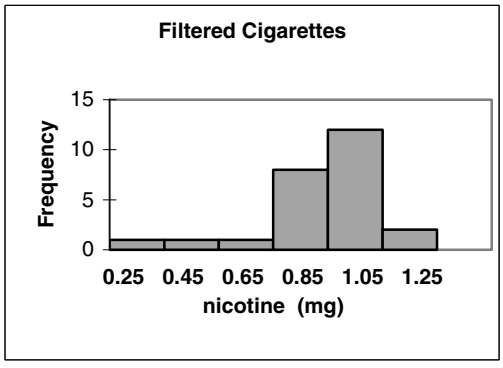


10. The histogram is given below.



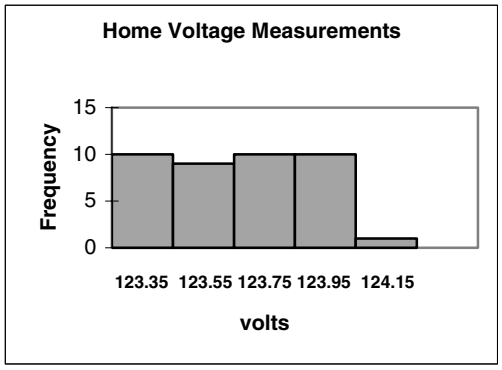
11. The histogram is given below at the left.

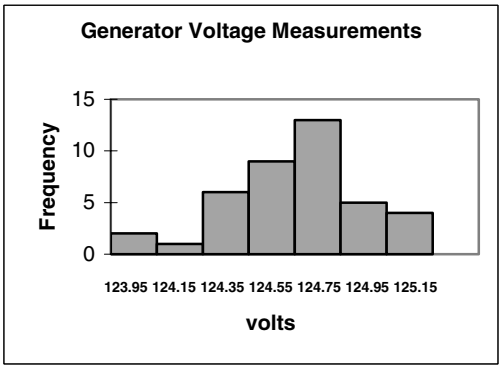




12. The histogram is given above at the right. For a better comparison, the two figures are placed side by side. The nicotine amounts appear to be substantially lower for the filtered cigarettes.

13. The histogram is given below at the left. The labels are the class midpoints. No, the voltages do not appear to follow a normal distribution – instead of being concentrated near the middle of the distribution, the values appear to be rather evenly distributed.

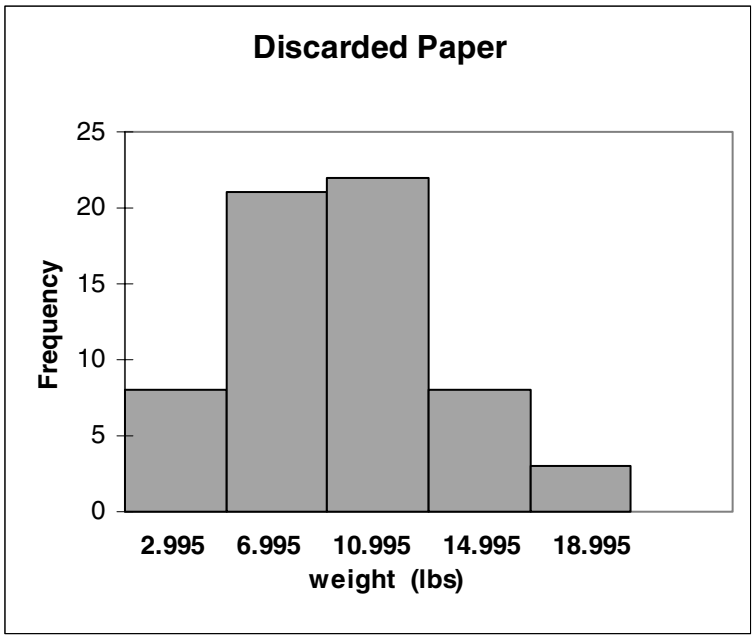




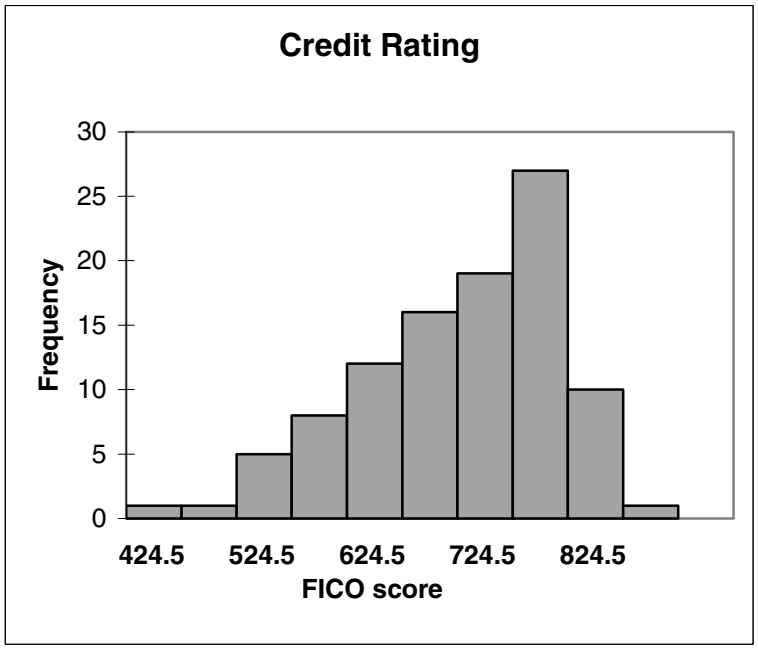
14. The histogram is given above at the right. The labels are the class midpoints. For a better comparison, the two figures are placed side by side. Yes, the voltages do appear to follow a normal distribution – there are many values near the center of the distribution, and the frequencies diminish toward either end. The values appear to be higher than those in exercise #13.

15. The histogram is given below. The labels are the class midpoints. While the 0.75″ label appears reasonably accurate in that all but 5 of the screws were within 0.02″ of that value, it appears that there are slightly more screws below the labeled value than above the labeled value and that the values extended farther below the labeled value.

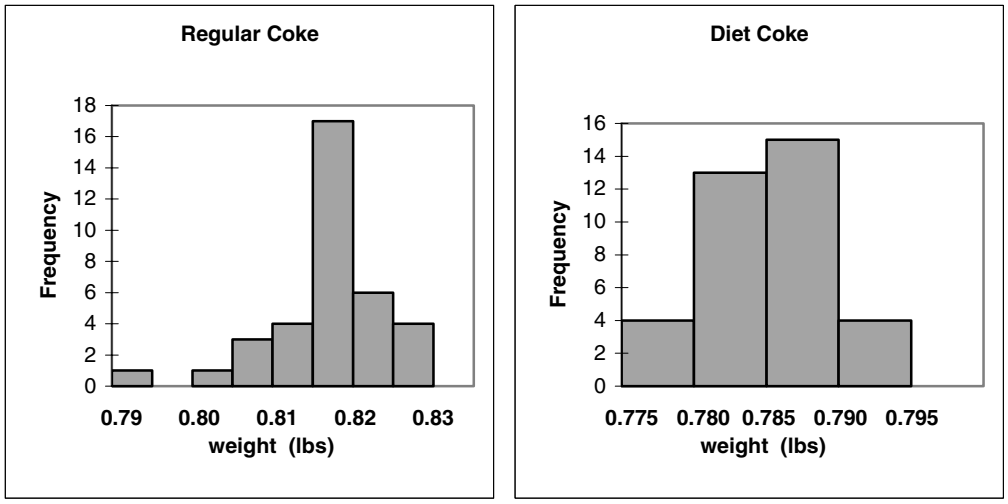
16. The histogram is given below. The labels are the class midpoints. Yes, the weights appear to have a distribution that is approximately normal.



17. The histogram is given below. The labels are the midpoints of every other class. The ratings appear to have a distribution that is not normal. While there is a maximum score with progressively smaller frequencies on either side of the maximum, the distribution is definitely not symmetric (i.e., the maximum score is not near the middle, but at the upper end of the distribution).

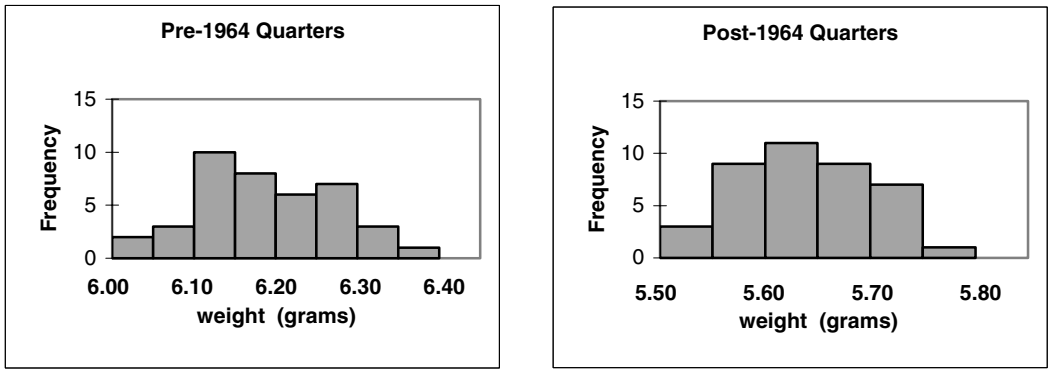


18. The two relative frequency histograms are given below. The true class boundaries are 0.78995,0.79495,0,80495,… (for the regular Coke) and 0.77495,0.77995,0.78495,… (for the diet Coke). The manual presents histograms that communicate the information in an appropriate, though approximate, manner.



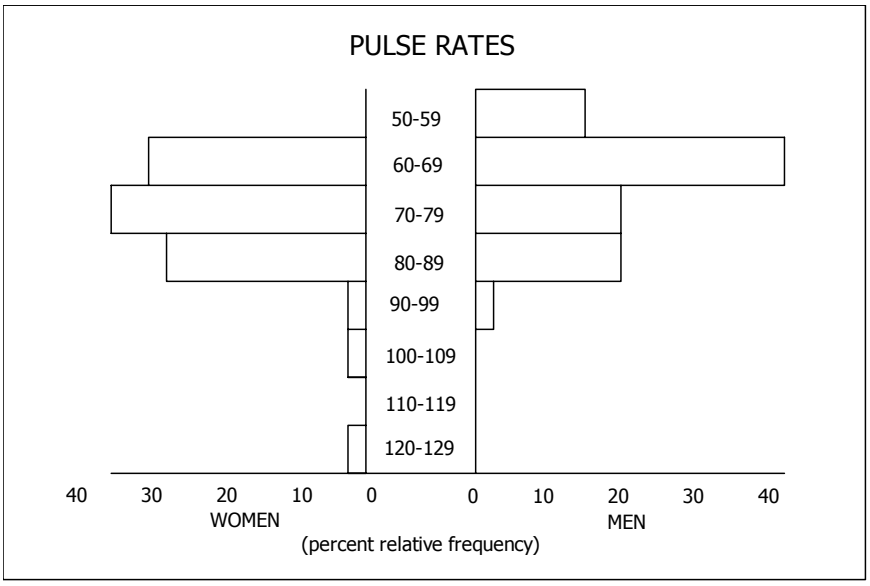
Each set of weights appears to have a distribution that is approximately normal, but there are two significant differences between the two sets: the weights for Regular Coke are considerably larger than those for Diet Coke, and the weights for Regular Coke cover a much wider range than those for Diet Coke. This suggests that the sweetener in Regular Coke adds weight to the product and does not distribute evenly throughout the product. Since the company produces more Regular Coke than Diet Coke, another possibility is that the harder-working machines filling the Regular Coke may not be holding their tolerance as well – and a wider range in volume dispensed might account for the wider range of weights for Regular Coke.

19. The histogram is given below at the left. The true class boundaries are 5.99995, 6.04995, 6.09995, etc. The manual presents a histogram that communicates the information in an appropriate, though approximate, manner.

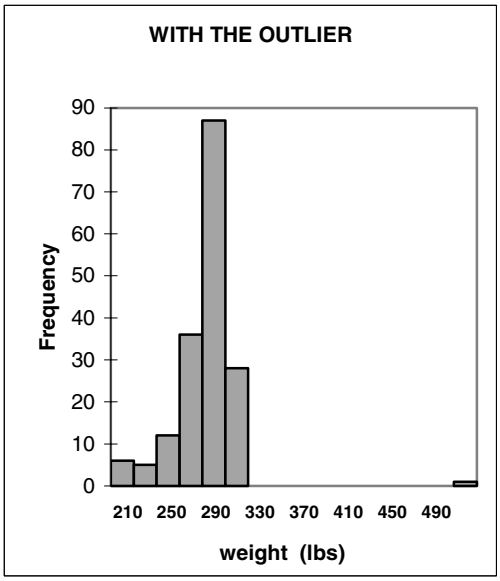


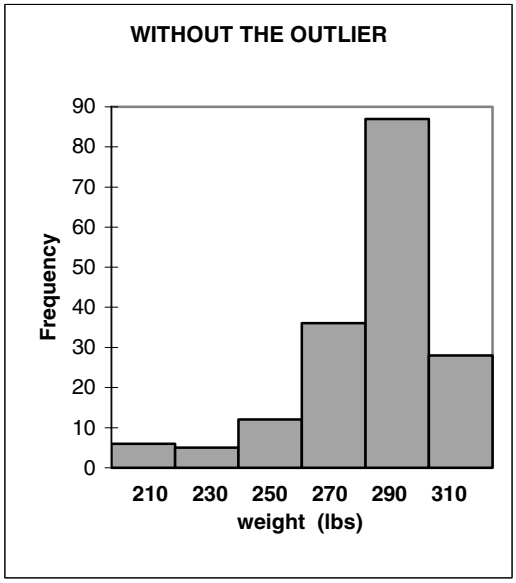
20. The histogram is given above at the right. For a better comparison, the two figures are placed side by side. The true class boundaries are 5.49995, 5.54995, 5.59995, etc. The manual presents a histogram that communicates the information in an appropriate, though approximate, manner. The post-1964 quarters appear to have weights that are lighter (due to their different metallic composition) and spread over a smaller range (due to their fewer years in circulation).

21. The back-to-back relative frequency histograms are given below. The pulse rates of the males tend to be lower than those of the females.



22. The two requested histograms are given below. The true class midpoints are 209.5, 229.5, 249.5,etc. The manual presents a histogram that communicates the information in an appropriate, though approximate, manner. The two figures give very different visual images of the shape of the distribution. An outlier can have a significant effect on the histogram.





2-4 Statistical Graphics

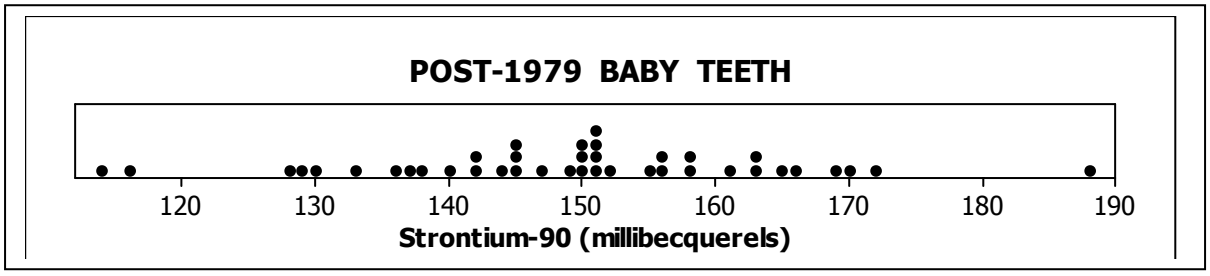
1. The dotplot permits identification of each original value and is easier to construct. The dotplot gives an accurate visual impression of the proportion of the data within any selected range of values; while the polygon is limited to impressions concerning the specified classes (and only the heights at the class midpoints, and not the areas under the lines, give an accurate visual impression of those proportions).

2. A scatterplot requires paired data from two quantitative variables – typically either two pieces of data from each experimental unit (e.g., a child's height and weight), or data from two different sets for which each value from one set may be appropriately associated with a value from the second set (e.g., the weight of the male child and the weight of the female child from mixed-gender twins) . The scatterplot can reveal the nature of the relationship between the two variables.

3. Using relative frequencies allows direct comparison of the two polygons. When two sets of data have different sample sizes, the larger data set will naturally have higher frequencies and direct comparison of the heights of the two polygons does not give meaningful information.

4. Since categories in a Pareto chart are ordered according to frequency, Pareto charts clearly show the relative positions of the categories under investigation. In addition, the Pareto chart is based on height and the pie chart is based on area – and it is easier to compare heights than areas.

5. The dotplot is given below. The Strontium-90 levels appear to have a "spread-out" normal distribution, a wide range of values clustered around 150 and occurring with less frequency at the extremes.

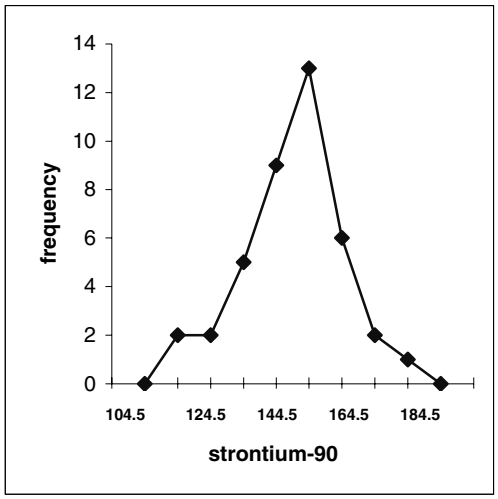


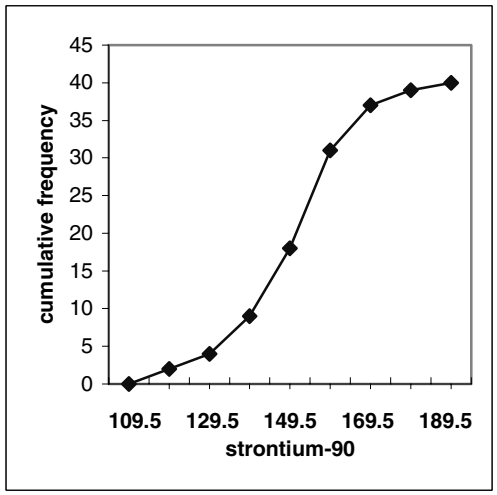
6. The stemplot is given below. The Strontium-90 levels appear to have a normal distribution clustered around 150.

Strontium-90 (mBq)

11	46
12	89
13	03678
14	022455579
15	0001111256688
16	133569
17	02
18	8

7. The frequency polygon is given below at the left.
NOTE: The frequencies are plotted at the class midpoints, which are not integer values. The polygon must begin and end at zero at the midpoints of the adjoining classes that contain no data values.





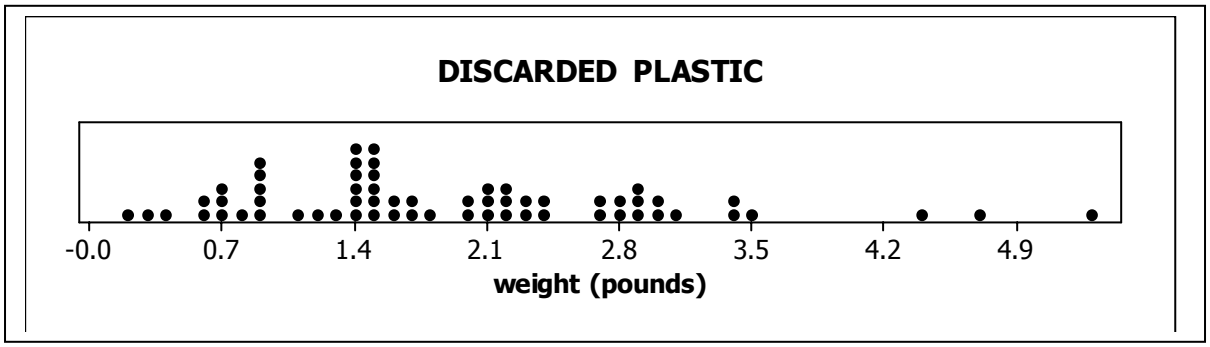
8. The ogive is given above at the right. Using the figure: move up from 150 on the horizontal scale to intersect the graph, then move left to intersect the vertical scale at 18. This indicates there were approximately 18 data values which would have been recorded as being below 150, which agrees with the actual data values.
NOTE: Ogives always begin on the vertical axis at zero and end at n, the total number of data values. All cumulative values are plotted at the upper class boundaries.

9. The stemplot is given below. The weights appear to be approximately normally distributed, except perhaps for the necessary lower truncation at zero.

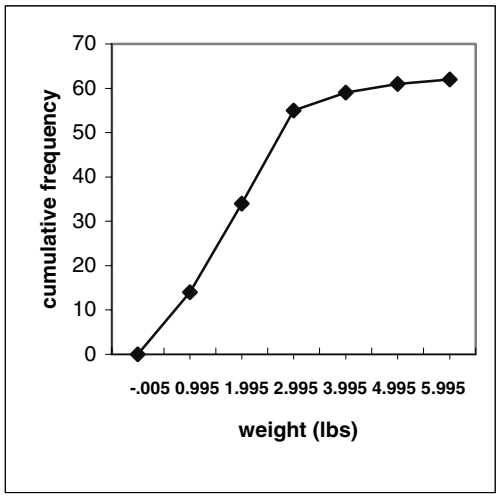
weight (pounds)

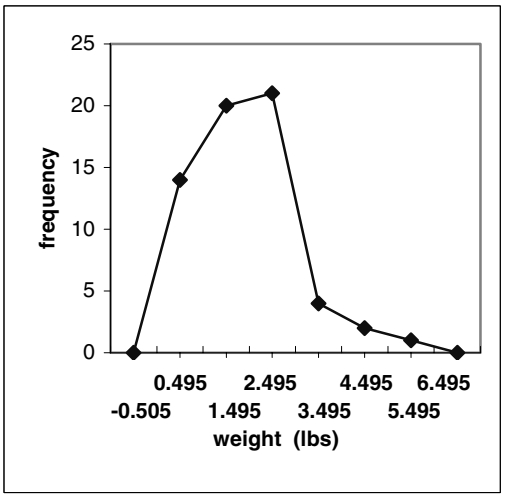
0. | 12356677888999
1. | 1123444444445556678
2. | 001111113334668888999
3. | 9345
4. | 36
5. | 2

10. The dotplot is given below. The weights appear to be approximately normally distributed, except perhaps for the presence of a few high values.



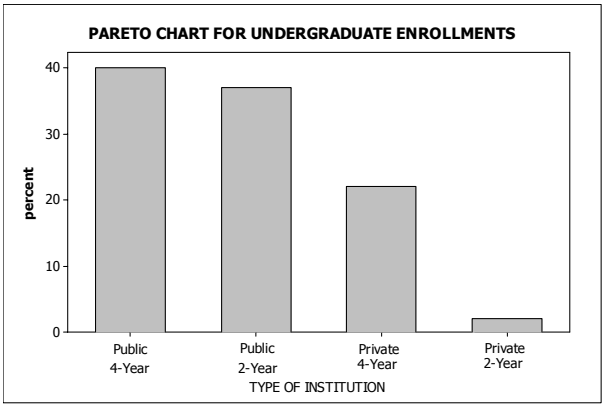
11. The ogive is given below at the left. Using the figure: move up from 4 on the horizontal scale to intersect the graph, then move left to intersect the vertical scale at 59. This indicates there were approximately 59 data values which would have been recorded as being below 4, which agrees with the actual data values.
NOTE: Ogives always begin on the vertical axis at zero and end at n, the total number of data values. All cumulative values are plotted at the upper class boundaries.

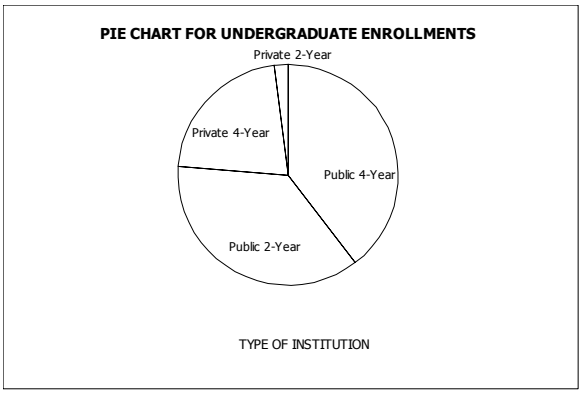




12. The frequency polygon is given above at the right.
NOTE: The frequencies are plotted at the class midpoints, which have one more decimal place than the original data. The polygon must begin and end at zero at the midpoints of the adjoining classes that contain no data values.

13. The Pareto chart is given below at the left.

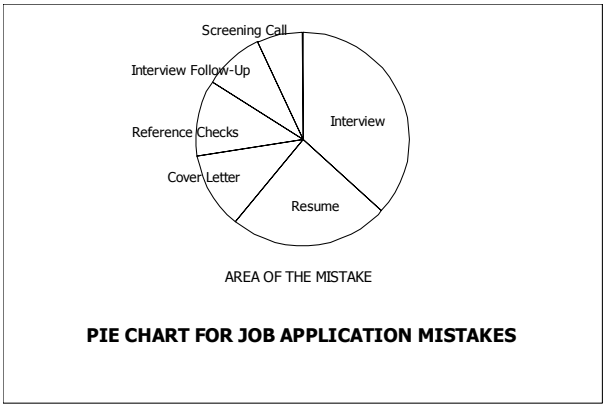


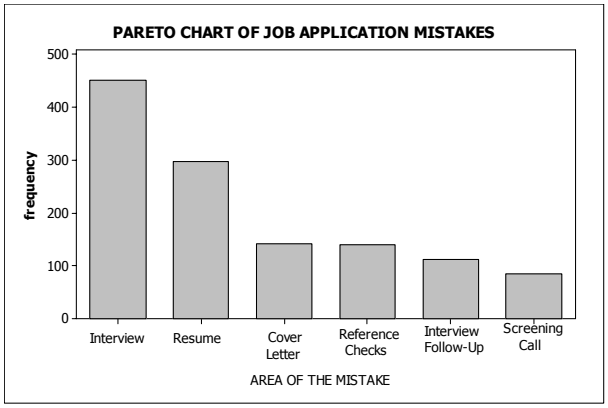


14. The pie chart is given above at the right. The "slices" of the pie may appear in any order and in any position, but their relative sizes must be as shown. The Pareto chart is more effective than the pie chart. While it is clear which bar in the Pareto chart is the tallest, it is not clear which area in the pie chart is the largest.

15. The pie chart is given below at the left. The "slices" of the pie may appear in any order and in any position, but their relative sizes must be as shown. There were 1231 total responses, and the central angle of the pie chart for each category was determined as follows.
Interview: 452/1231 = 36.7%, and 36.7% of 360° is 132°
Resume: 297/1231 = 24.1%, and 24.1% of 360° is 87°

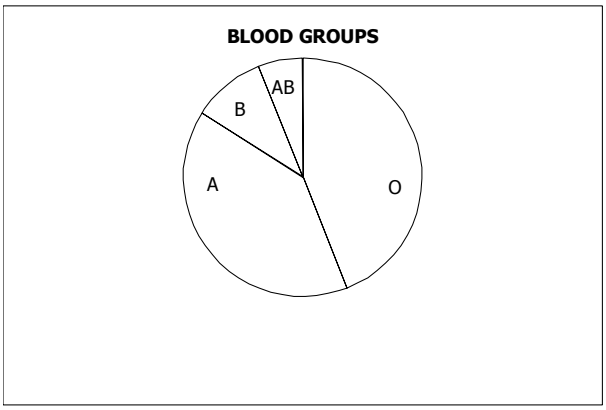
Reference Checks: 143/1231 = 11.6%, and 11.6% of 360° is 42°
Cover Letter: 141/1231 = 11/5%, and 11.5% of 360° is 41°
Interview Follow-Up: 113/1231 = 9.2%, and 9.2% of 360° is 33°
Screening Call: 85/1231 = 6.9%, and 6.9% of 360° is 35°

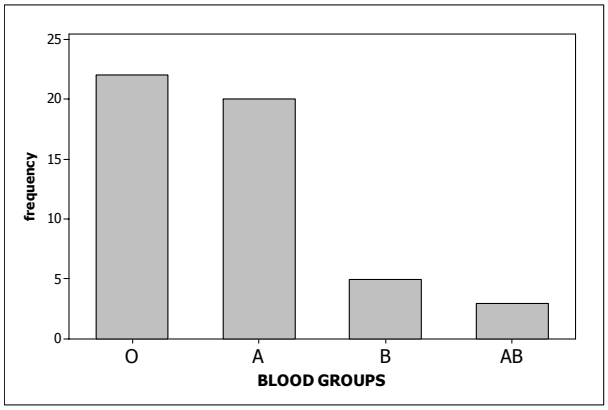




16. The Pareto Chart is given above at the right. The Pareto chart is more effective than the pie chart.

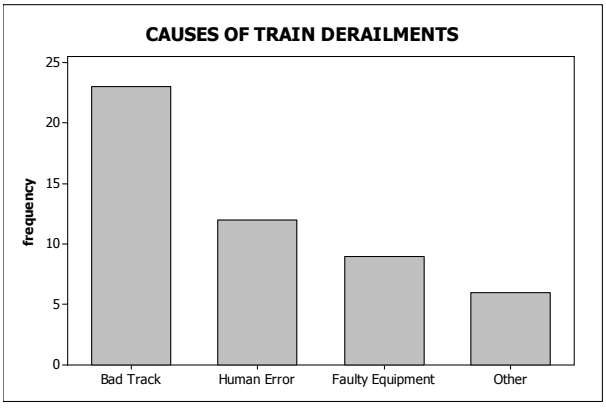
17. The pie chart is given below at the left. The "slices" of the pie may appear in any order and in any position, but their relative sizes must be as shown.

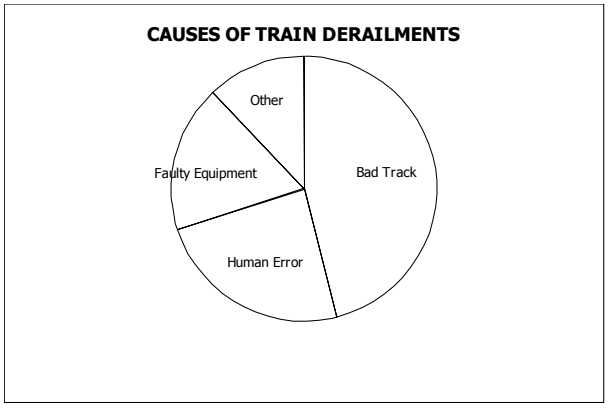




18. The Pareto chart is given above at the right.

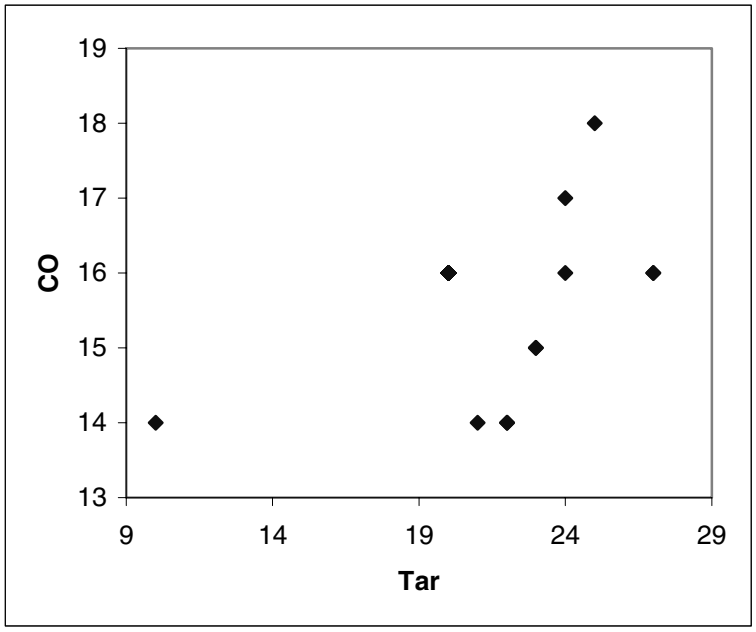
19. The Pareto chart is given below at the left.



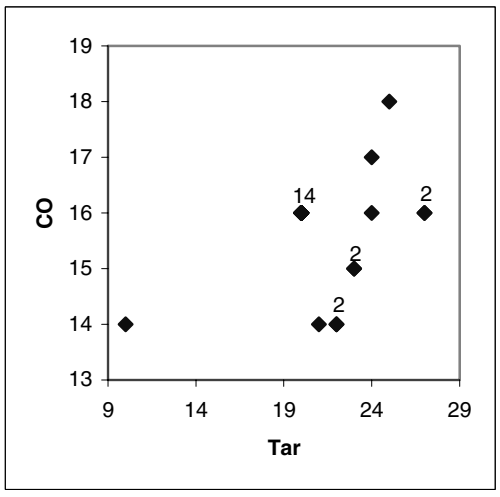


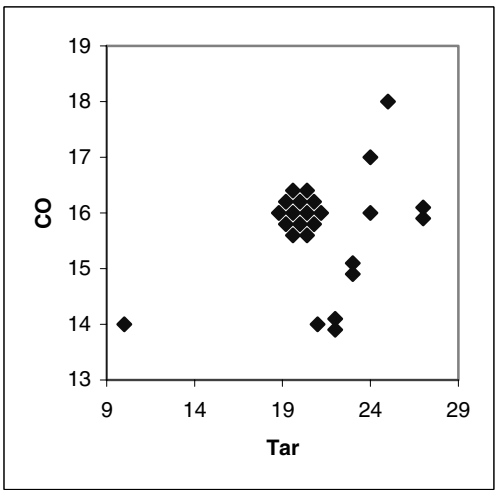
20. The pie chart is given above at the right. The "slices" of the pie may appear in any order and in any position, but their relative sizes must be as shown.

21. The simple, unmodified scatterplot is given below. There appears to be a slight tendency for cigarettes with more tar to also have more CO. NOTE: To make a scatterplot using Excel, the y column must be to the immediate right of the x column. One way to accomplish that in this data set is to delete the nicotine column [but do not save the file using the same name after the nicotine column has been deleted!].

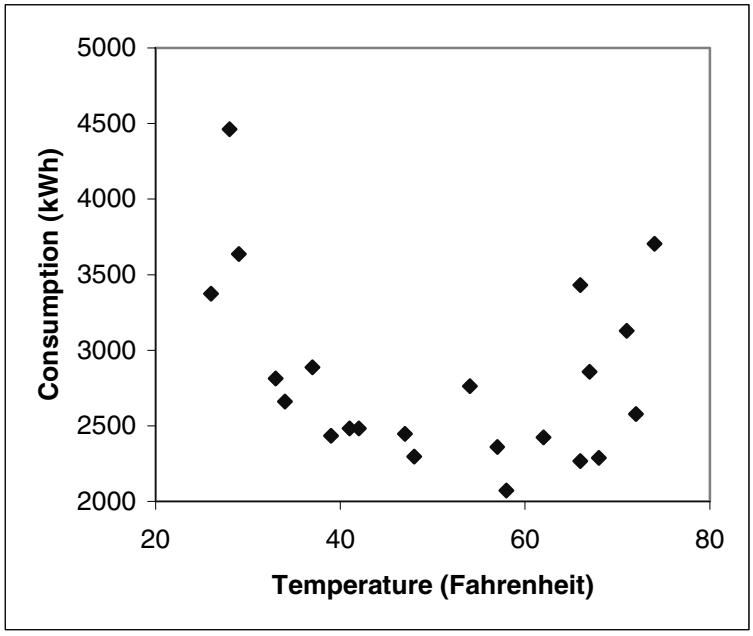


NOTE: The above scatterplot shows only 9 data points, even though there were 25 pairs of tar/CO data points in the original sample. Since the scatterplot actually shows less than half the information contained in the sample, it may not provide an accurate picture of the data. This is caused by duplicate values: the (22,14) and (23,15) and (27,16) each appear 2 times, and the (20,16) pair appears 14 times! Two modifications that adjust for this phenomenon are shown below. The scatterplot on the left inserts numbers to tell how many data points are represented by dots that indicating duplicate values. The scatterplot on the right shows the true number of dots. The same effect can also be obtained by using dots whose size is proportional to the number of duplicate values it represents. The modified scatterplots indicate that there appears to be no relationship between the amounts of tar and CO. Both of the modified sactterplots were constructed in Excel.

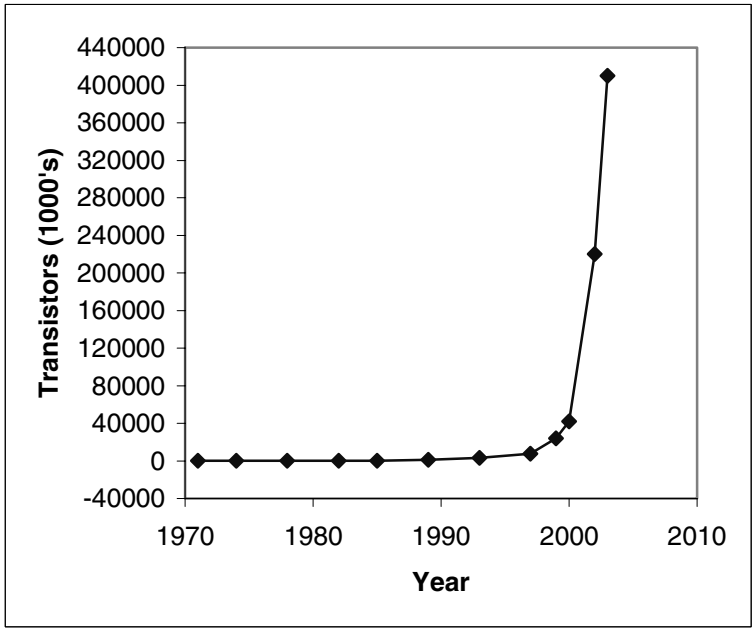




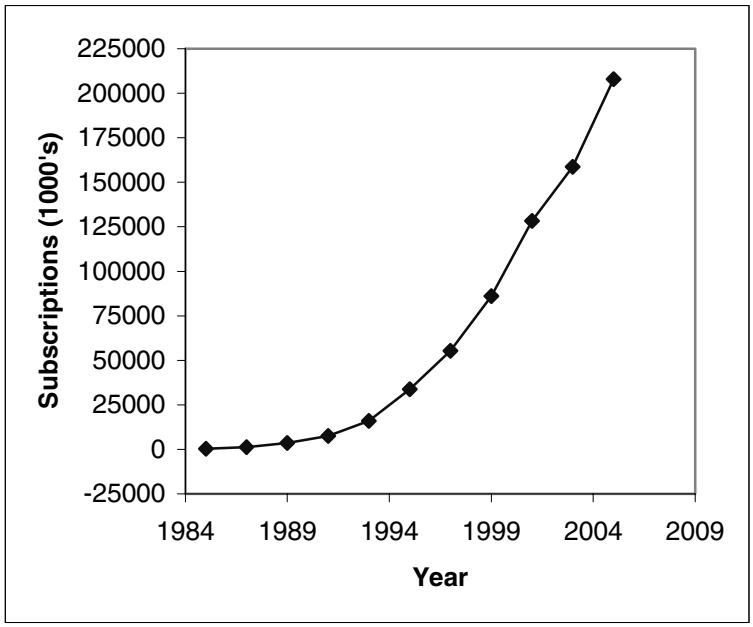
22. The scatterplot is given below. It appears that more energy is used on days when it is very cold (for heating) or very warm (for air conditioning).



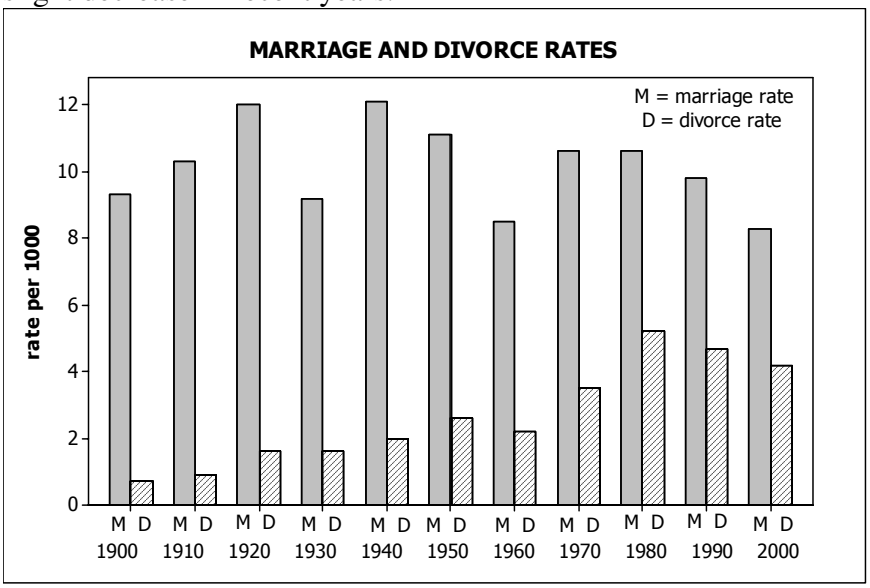
23. The time series graph is given below. Note that the given years are not evenly spaced.



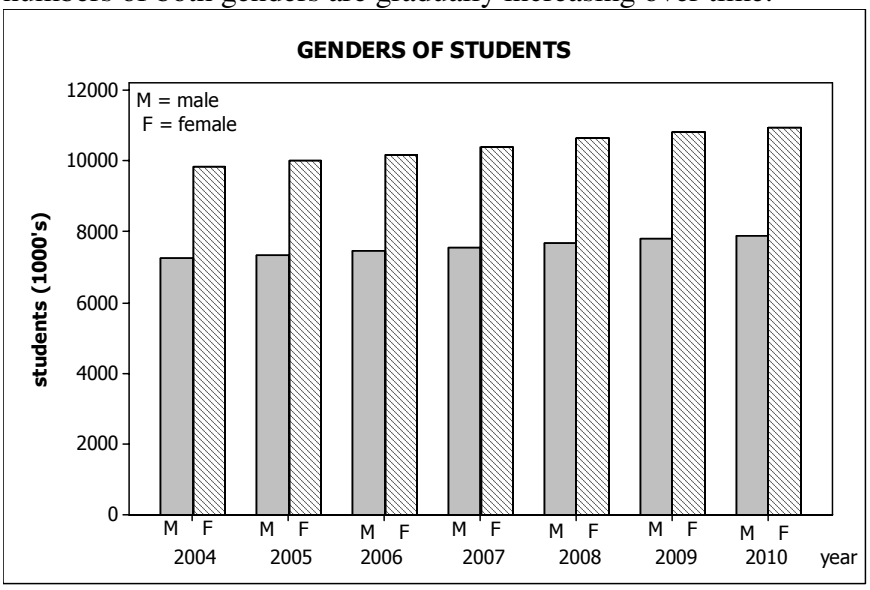
24. The time series graph is given below. The graph does not appear to show linear growth (constant slope) over the entire time period, but it does appear that there was linear growth during certain periods (e.g., since 1999).



25. The multiple bar graph is given below. As the population increases, the numbers of marriages and divorces will automatically increase. To identify any change in marriage and divorce patterns, one needs to examine the rates. This is analogous to using percents (or relative frequencies) instead of frequencies to compare categories for two samples of different sizes. The marriage rate appears to have remained fairly constant, with a possible slight decrease in recent years. The divorce rate appears to have steadily grown, with a possible slight decrease in recent years.



26. The multiple bar graph is given below. The females consistently outnumber the males, and the numbers of both genders are gradually increasing over time.



27. The back-to-back stemplot is given below. The pulse rates for men appear to be lower than the pulse rates of women.

PULSE RATES		
Women		Men
	5	666666
888884444000	6	00000004444444888
66666622222222	7	22222266
88888000000	8	44448888
6	9	6
4	10	
	11	
4	12	

28. a. The next two rows of the expanded stemplot come from the original 70's row as follows.
7 | 22222222
7 | 666666

b. The completed condensed stemplot is as follows.

FEMALE PULSE RATES

stem	leaves
6– 7	000444488888*22222222666666
8– 9	00000088888*6
10–11	4*
12–13	4*

2-5 Critical Thinking: Bad Graphs

1. The illustration uses two-dimensional objects (dollar bills) to represent a one-dimensional variable (purchasing power). If the illustration uses a dollar bill with ½ the original length and ½ the original width to represent ½ the original purchasing power, then the illustration is misleading (because ½ the length and ½ the width translates into ¼ the area and gives the

visual impression of 25% instead of 50%). But if the illustration uses a dollar bill with ½ the area (i.e, with .707 of the original length and .707 of the original width) to represent ½ the original purchasing power, then the illustration conveys the proper visual impression.

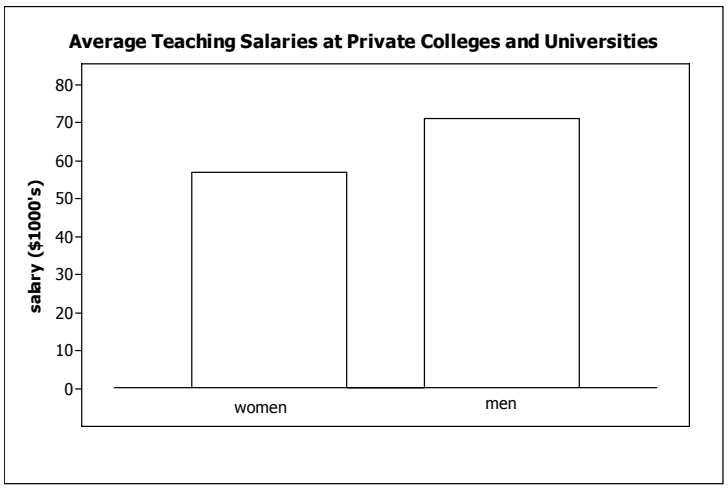
2. No. Since the data comes from a voluntary response sample it may not be representative of the population. Since the sample may not be representative, even sound graphing techniques will not necessarily provide accurate understanding of the population.

3. No. Results should be presented in a way that is fair and objective so that the reader has the reliable information necessary to reach his own conclusion.

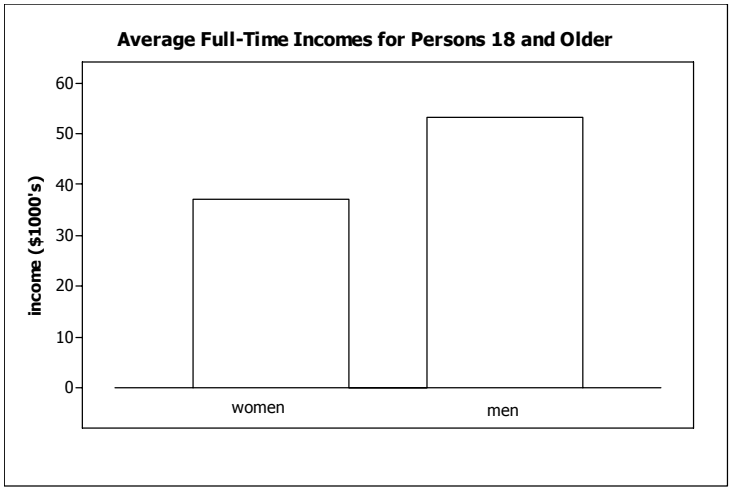
4. No, the resulting graph is not misleading. Since the variable of interest (area) is two-dimensional, it is appropriate to use corresponding two-dimensional figures to make comparisons.

5. No. The illustration uses two-dimensional objects to represent a one-dimensional variable (weight). The average male weight is 172/137 = 1.255 times the average female weight. Making a two-dimensional figure 1.255 times taller and 1.255 times wider increases the area by $(1.255)^2 = 1.58$ and gives a misleading visual impression.

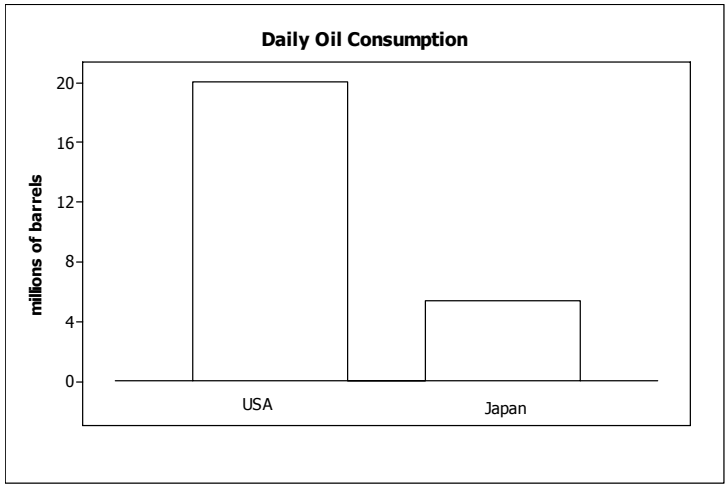
6. The graph creates the impression that men have salaries that are more than twice that of women. The distortion occurs because the vertical scale does not start at zero. A graph that depicts the data fairly is given at the right.



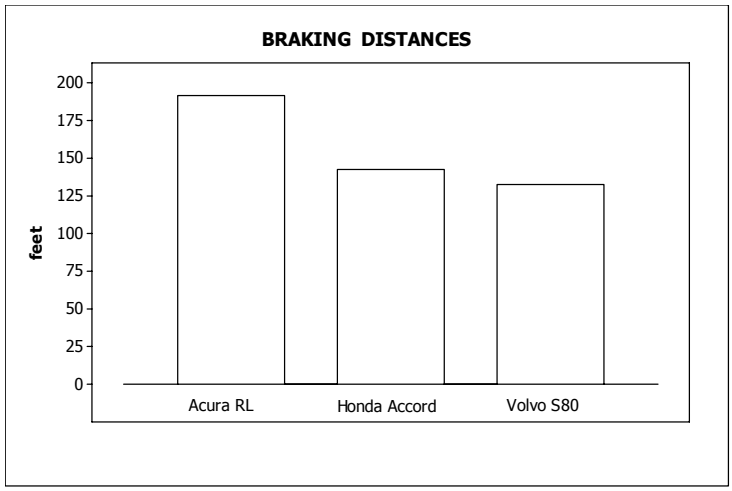
7. The average income for men is about 1.4 times the average income for women. Making the men's pictograph 1.4 times as wide and 1.4 times as high as the women's produces a men's image with $(1.4)^2 = 1.96$ times the area of the women's image. Since it is the area that gives the visual impression in a two-dimensional figure, the men's average income appears to be almost twice that of the women's average income. A graph that depicts the data fairly is given at the right.



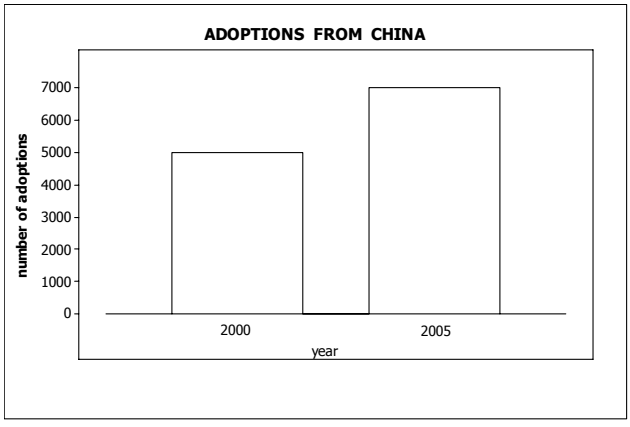
8. The oil consumption for the USA is about 3.7 times the oil consumption for Japan. Making the USA's pictograph 3.7 times larger than Japan's in three dimensions produces an image for the USA with $(3.7)^3 = 50$ times the volume of the image for Japan. Since it is the perceived volume that gives the visual impression in the given figure, the consumption for the USA appears to be 50 times that for Japan. A graph depicting the data fairly is given at the right.



9. The graph in the text makes it appear that the braking distance for the Acura RL is more than twice that of the Volvo S80. The actual difference is about 60 feet, and the Acura RL distances is about 192/133 = 1.44 times that of the Volvo S80. The exaggeration of differences is caused by the fact that the distance scale dies not start at zero. A graph that depicts the data fairly is given at the right.

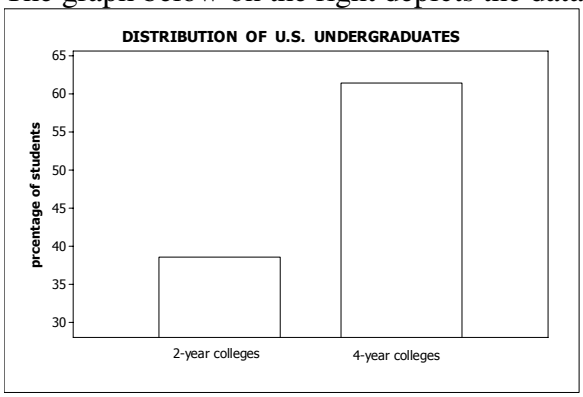


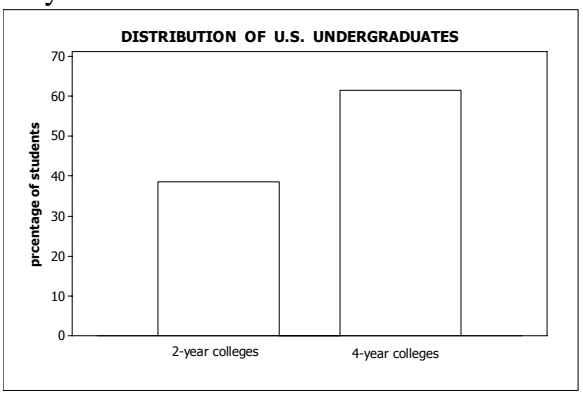
10. The graph given in the text is misleading because it gives the visual impression that the number of adoptions has more than doubled. Not starting the vertical axis at zero exaggerates the differences between categories. A graph that depicts the data fairly is given at the right.



11. The given figure is misleading because the backside of the head is not visible. Categories extending to the backside of the head will not have as much area showing as comparable categories shown at the front of the head. A regular pie chart would give the relative sizes of the categories in as undistorted manner. A better graph would be a bar chart – with the vertical axis starting at 0, and the categories given in order by age. When there is a natural ordering of the categories that can be preserved with a bar chart – but it is hidden in a pie chart, which ends up placing the "first" and "last" categories side by side.

12. For easier comparison, the two graphs are given side by side with the same horizontal scale.
    a. The graph below on the left exaggerates the differences between categories by not starting the vertical scale at zero.
    b. The graph below on the right depicts the data fairly.





Statistical Literacy and Critical Thinking

1. When investigating the distribution of a data set, a histogram is more effective than a frequency distribution. Both figures contain the same information, but the visual impact of the histogram presents that information in a more efficient and more understandable manner.
2. When investigating changes over a period of years, a time series graph would be more effective than a histogram. A histogram would indicate the frequency with which different amounts occurred, but by ignoring the years in which those amounts occurred it would give no information about changes over time.
3. Using two-dimensional figures to compare one-dimensional variables exaggerates differences whenever the areas of the two dimensional figures are not proportional to the amounts being portrayed. Making the height and width proportional to the amounts being portrayed creates a distorted picture because it is area that makes the visual impression on the reader – and a two-fold increase in height and width produces a four-fold increase in area.
4. The highest histogram bars should be near the center, with the heights of the bars diminishing toward each end. The figure should be approximately symmetric.

Chapter Quick Quiz

1. 10 – 0 = 10. The class width may be found by subtracting consecutive lower class limits.
2. Assuming the data represent values reported to the nearest integer, the class boundaries for the first class are -0.5 and 9.5.
3. No. All that can be said is that there are 27 data values somewhere within that class.
4. False. A normal distribution is bell-shaped, with the middle classes having higher frequencies than the classes at the extremes. The distribution for a balanced die will be flat, with each class having about the same frequency.
5. Variation.

6. 52, 52, 59. The 5 to the left of the stem represent the tens digit associated with the ones digits to the right of the stem.
7. Scatterplot. The data is two-dimensional, requiring separate axes for each variable (shoe size and height).
8. True. The vertical scale for the relative frequency histogram will be the values of the frequency histogram divided by the sample size n.
9. A histogram reveals the shape of the distribution of the data.
10. Pareto chart. When there is no natural order for the categories, placing them in the order of their frequencies shows the relative importance without losing the nature of any significant relationships between the categories.

Review Exercises

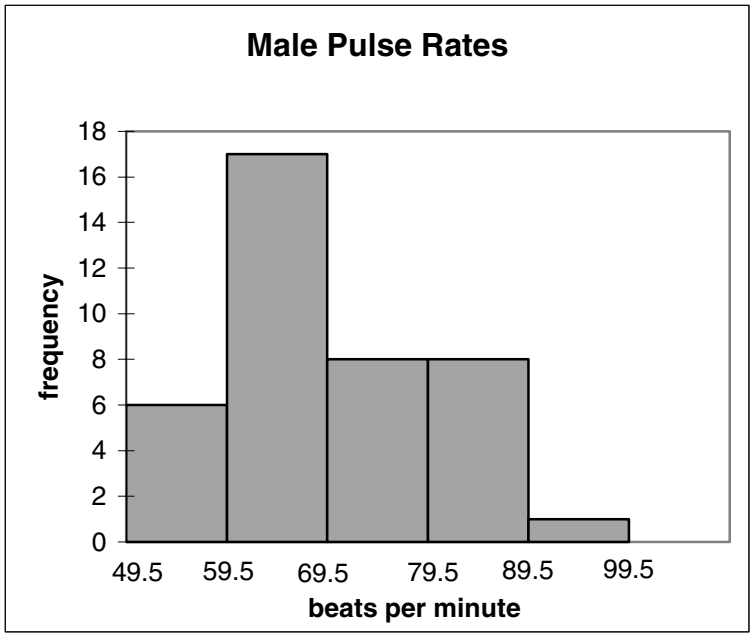
1. The requested frequency distribution is given below, with the preliminary Excel output at the right. The pulse rates for the males appear to be lower than those for the females.

MALE PULSE RATES

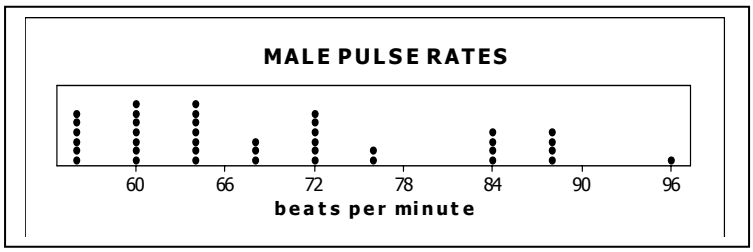
beats per minute	frequency
50–59	6
60–69	17
70–79	8
80–89	8
90–99	1
	40

Bin	*Frequency*
59.5	6
69.5	17
79.5	8
89.5	8
99.5	1
More	0

2. The histogram is given below. The basic shape is similar to the histogram for the females, but the male pulse rates appear to be lower.



3. The dotplot is given below at the left. It shows that the male pulse rates appear to be lower than those for the females.

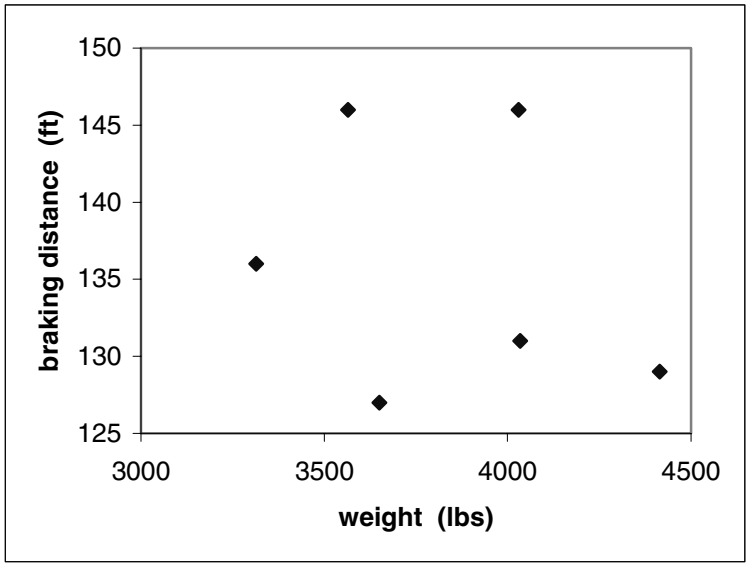


beats per minute

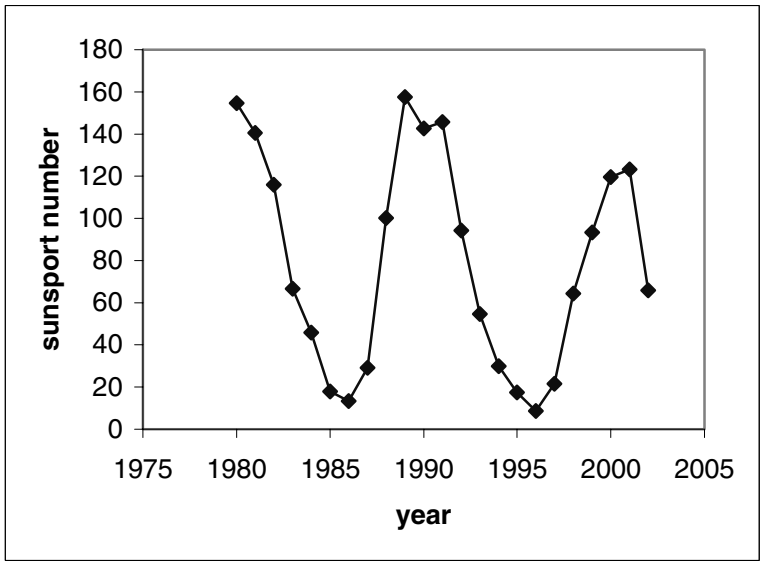
5 | 666666
6 | 00000004444444888
7 | 22222266
8 | 44448888
9 | 6

4. The stemplot is given above at the right. It shows that the male pulse rates appear to be lower than those of the females.

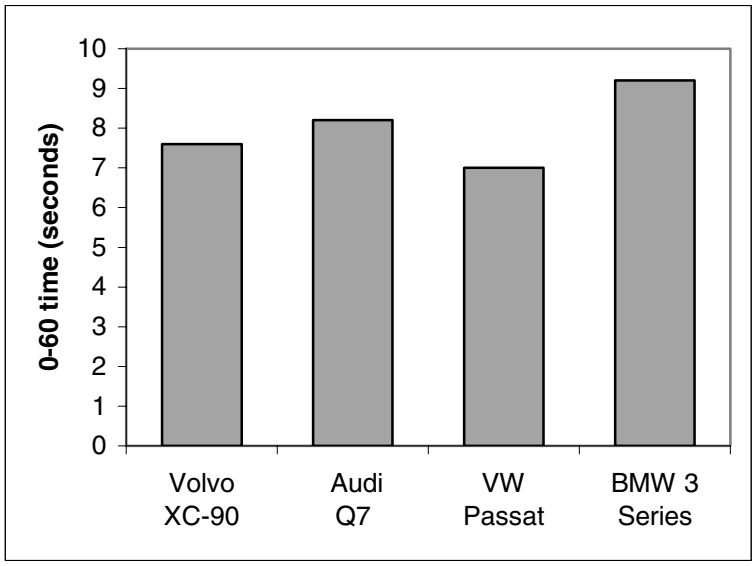
5. The scatterplot is given below. No; based on the figures, there does not appear to be a relationship between the weight of a car and its braking distance.



6. The time-series graph is given below. There appears to be a trend of long-term cycles of approxiamtely 10 years duration.



7. The graph is misleading because the vertical axis does not start at zero, causing it to exaggerate the differences between the categories. A graph that correctly illustrates the acceleration times is given below.



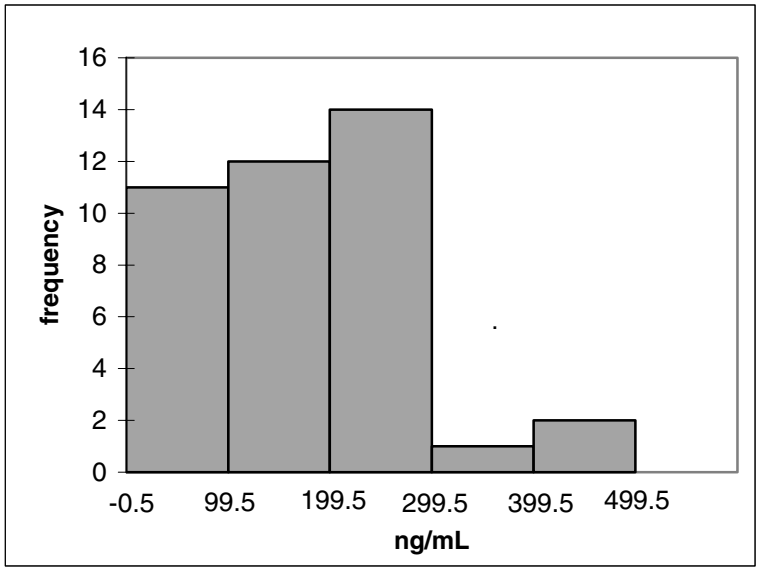
8. a. 25. The difference between the first two lower class limits is 125 – 100 = 25.
   b. 100 and 124. These are the values given in the first row of the table.
   c. 99.5 and 124.5. Values within these boundaries will round to the whole numbers given by the class limits.
   d. No. The distribution is not symmetric; the class with the largest frequency is near the right end of the distribution.

Cumulative Review Exercises

1. Yes. The sum of the relative frequencies is 100%.
2. Nominal. The yes-no-maybe responses are categories only; they do not provide numerical measures of any quantity, nor do they have any natural ordering.
3. The actual numbers of responses are as follows. Note that 884 + 433 + 416 = 1733.
   Yes: (0.51)(1733) = 884. No: (0.25)(1733) = 433. Maybe: (0.24)(1733) = 416
4. Voluntary Response Sample. A voluntary response sample is not likely to be representative of the population of all executives, but of those executives who had strong feelings about the topic and/or had enough free time to respond to such a survey.
5. a. A random sample is a sample in which every member of the population has an equal chance of being selected.
   b. A simple random sample of size n is a sample in which every possible sample of size n has an equal chance of being selected.
   c. Yes, it is a random sample because every person in the population of 300,000,000 has the same chance of being selected. No, it is not a simple random sample because all possible groups of 1000 do not have the same chance of being selected – in fact a group of 1000 composed of the oldest person in the each of the first 1000 of the 300,000 groups has no chance of being selected.

6. a. 100. The difference between the first two lower class limits is 100 – 0 = 100.
   b. -0.5 and 99.5. Values within these boundaries will round to the whole numbers given by the class limits.
   c. 11/40 = 0.275, or 27.5%. The total of the frequencies is 40.
   d. Ratio. Differences between the data values are meaningful and there is a meaningful zero.
   e. Quantitative. The data values are measurements of the cotinine levels.

7. The histogram is given below. Using a strict interpretation of the criteria, the cotinine levels do not appear to be normally distributed – the values appear to be concentrated in the lower portion of the distribution, with very few values in the upper portion.
   NOTE: The histogram bars extend from class boundary to class boundary, but sometimes the approximate labels 0, 100, 200, etc. are preferred over the more cumbersome -0.5, 99.5, 199.5 etc.



8. Statistic. A statistic is a measurement of some characteristic of a sample, while a parameter is a measurement of some characteristic of the entire population.