

Exploring Data

1.1 (a) The individuals are vehicles (or “cars”). (b) The variables are: vehicle type (categorical), transmission type (categorical), number of cylinders (quantitative), city MPG (quantitative), and highway MPG (quantitative).

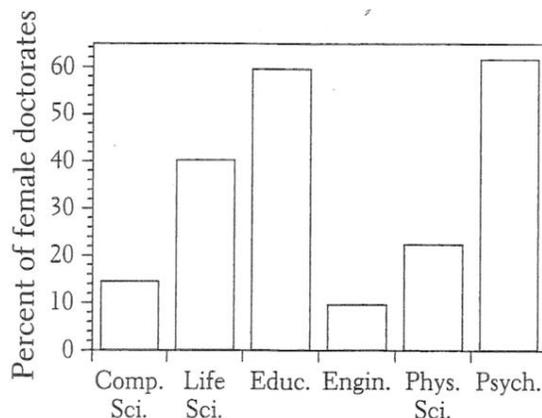
1.2 (a) Categorical. (b) Quantitative. (c) Categorical. (d) Categorical. (e) Quantitative. (f) Quantitative.

1.3 Possible answers (units):

- Number of pages (pages)
- Number of chapters (chapters)
- Number of words (words)
- Weight or mass (pounds, ounces, kilograms . . .)
- Height and/or width and/or thickness (inches, centimeters . . .)
- Volume (cubic inches, cubic centimeters . . .)

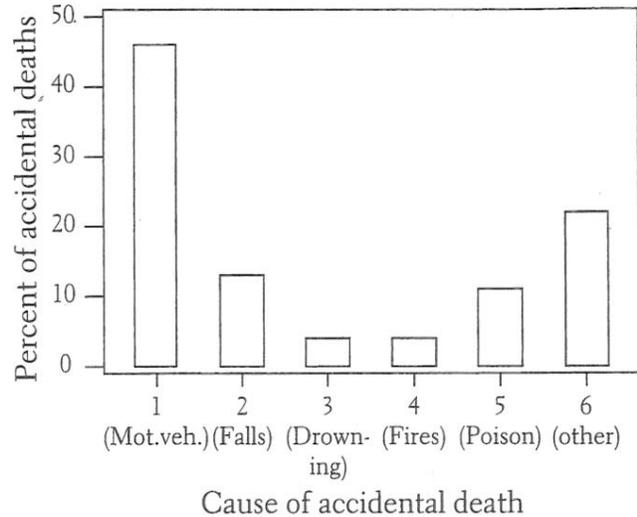
1.4 Possible answers (reasons should be given): unemployment rate, average (mean or median) income, quality/availability of public transportation, number of entertainment and cultural events, housing costs, crime statistics, population, population density, number of automobiles, various measures of air quality, commuting times (or other measures of traffic), parking availability, taxes, quality of schools.

1.5 (a) Shown below. The bars are given in the same order as the data in the table—the most obvious way—but that is not necessary (since the variable is nominal, not ordinal). (b) A pie chart would not be appropriate, since the different entries in the table do not represent parts of a single whole.

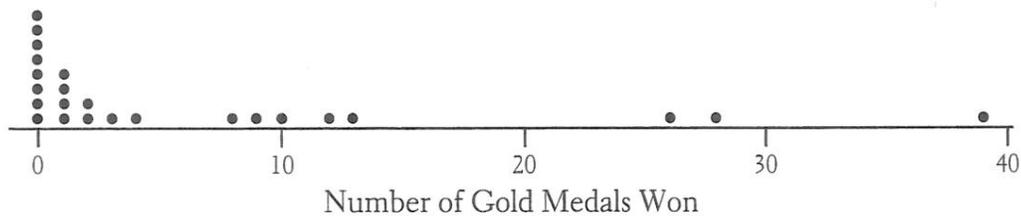


1.6 (a) Below. For example, "Motor Vehicles" is 46% since $\frac{42,340}{92,353} = 0.45846\dots$. The "Other causes" category is needed so that the total is 100%. (b) Below. The bars may be in any order. (c) A pie chart *could* also be used, since the categories represent parts of a whole (all accidental deaths).

Cause	Percent
Motor vehicles	46
Falls	13
Drowning	4
Fires	4
Poisoning	11
Other causes	22

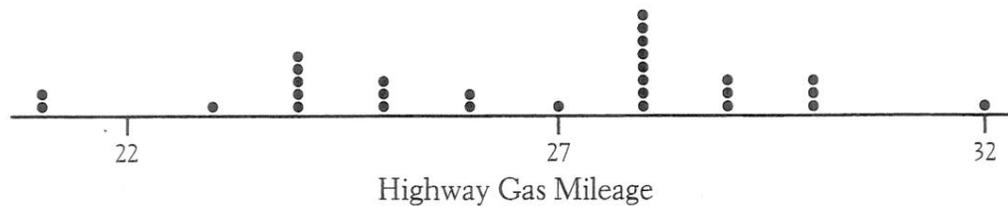


1.7



The distribution has a peak at 0 and a long right tail. There are eight outliers, with the most severe being 26, 28, and 39. The spread is 0 to 39 and the center is 1.

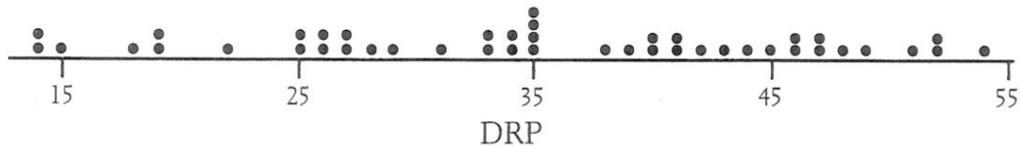
1.8



The distribution is skewed to the left, with a major peak at 28 and a minor peak at 24. The spread is relatively narrow (21 to 32 mpg). The two observations at 21 and the observation at 32 appear to be mild outliers. The center is 28 mpg.

1.9 (a) Stems = thousands, leaves = hundreds. The data have been rounded to the nearest \$100.
 (b) The distribution is skewed strongly to the right, with a peak at the 1 stem. The spread is approximately 19,000 (\$1300 to \$19,300). The center is 45 (\approx \$4500). The observations 182 (\approx \$18,200) and 193 (\approx \$19,300) appear to be outliers.

1.10



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

1.11 (a)

0	399
1	1345677889
2	000123455668888
3	25699
4	1345579
5	0359
6	1
7	0
8	366
9	3

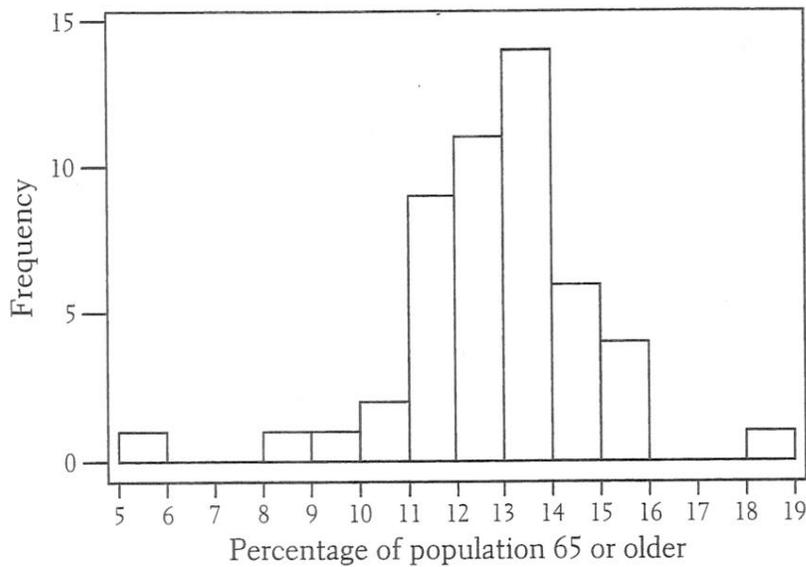
(b)

0	3
0	99
1	134
1	5677889
2	0001234
2	55668888
3	2
3	5699
4	134
4	5579
5	03
5	59
6	1
6	
7	0
7	
8	3
8	66
9	3

Both plots show the general shape of the distribution; however, the split-stem plot may be preferable since it shows more detail.

(c) The distribution is skewed to the right with a peak in the 2 stem(s). The spread is approximately 90 (3 to 93). There are several moderate outliers visible in the split-stem plot; specifically, the five amounts of \$70 or more. While most shoppers spent small to moderate amounts of money, a “cluster” of shoppers spent larger amounts ranging from \$70 to \$93. The center of the distribution is at approximately \$28.

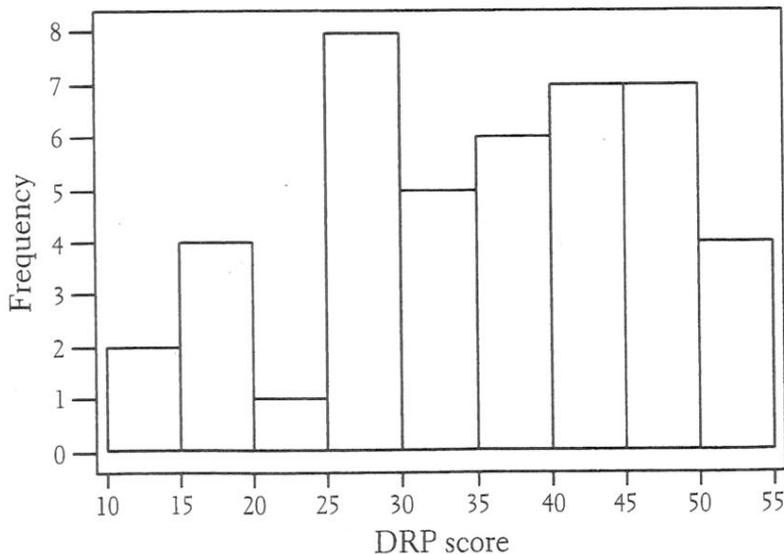
1.12 (a)



Percent	Freq.
5.0-5.9	1
6.0-6.9	0
7.0-7.9	0
8.0-8.9	1
9.0-9.9	1
10.0-10.9	2
11.0-11.9	9
12.0-12.9	11
13.0-13.9	14
14.0-14.9	6
15.0-15.9	4
16.0-16.9	0
17.0-17.9	0
18.0-18.9	1
Total	50

(b) The distribution is slightly skewed to the left with a peak at the class 13.0–13.9. There is one outlier in each tail of the distribution.

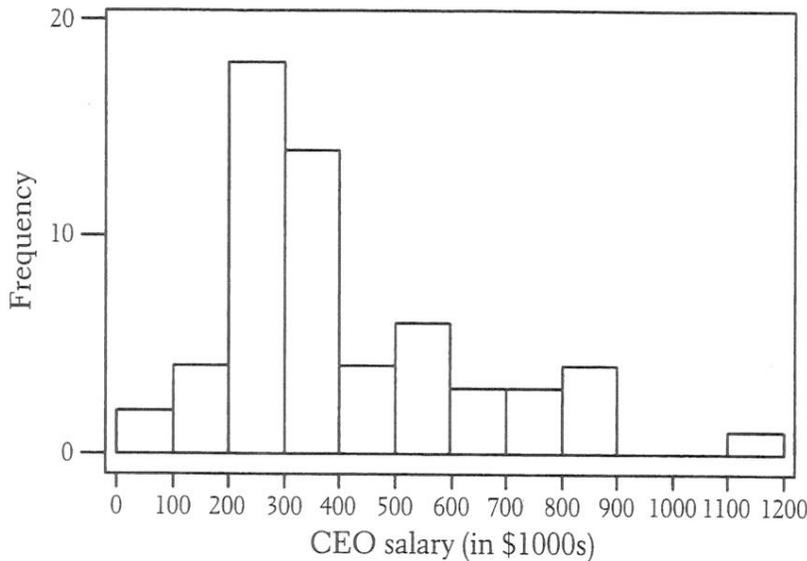
1.13



DRP Score	Freq.
10-14	2
15-19	4
20-24	1
25-29	8
30-34	5
35-39	6
40-44	7
45-49	7
50-54	4
Total	44

The dotplot provides more detail, but the histogram has the advantage of clearly displaying two “clusters” of DRP scores (the classes 25–29 and 40–44, 45–49).

1.14



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

1.15 (b) The distribution is symmetric with a peak at class (chest size) 40. The center is also located at 40. The spread is 15 (33 to 48). Assuming that the sample is representative of all members of the population, the distribution would provide a useful guide to those making clothing for the militiamen. From the frequency table, it is easy to estimate the percentage of all militiamen who have a certain chest size. The production of uniforms can reflect this distribution.

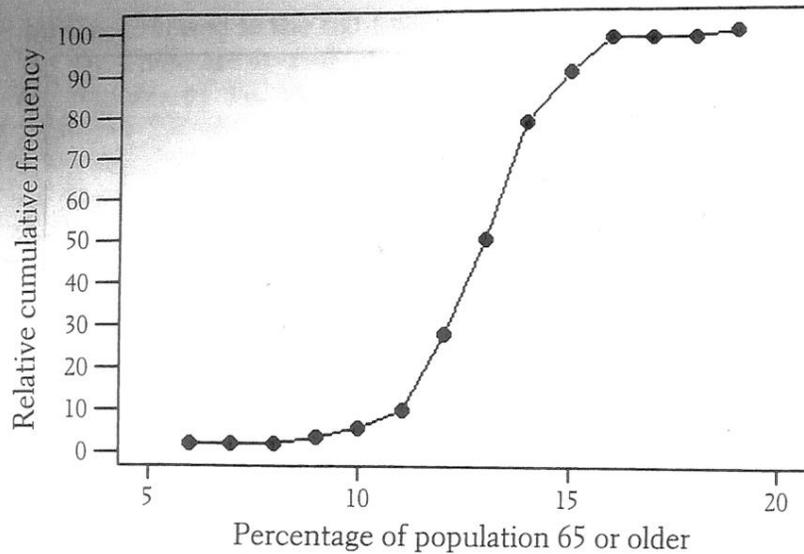
1.16 (a) Roughly symmetric, though it might be viewed as SLIGHTLY skewed to the right. (b) About 15%. (39% of the stocks had a total return less than 10%, while 60% had a return less than 20%. This places the center of the distribution somewhere between 10% and 20%.) (c) The smallest return was between -70% and -60%, while the largest was between 100% and 110%. (d) 23% (1 + 1 + 1 + 1 + 3 + 5 + 11).

1.17 (a) Skewed to the right; center at about 3 (31 less than 3, 11 equal to 3, 23 more than 3); spread: 0 to 10. No outliers. (b) About 23% (15 out of 65 years).

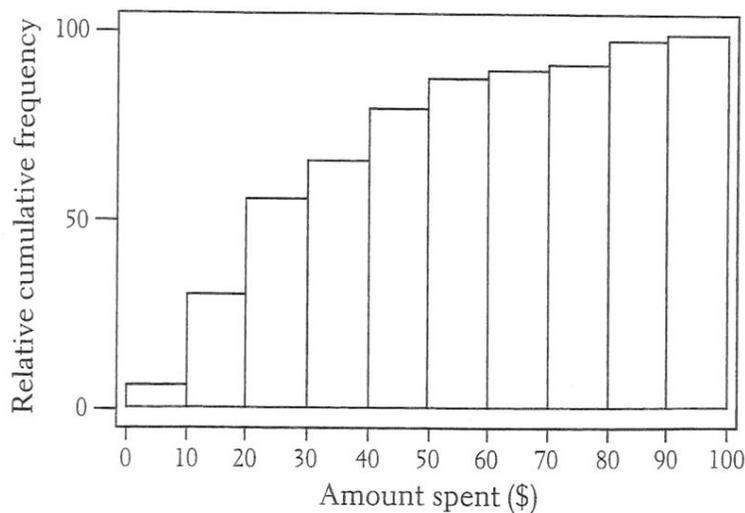
1.18 Lightning histogram: centered at noon (or more accurately, somewhere from 11:30 to 12:30). Spread is from 7 to 17 (or more accurately, 6:30 AM to 17:30, i.e., 5:30 PM). Shakespeare histogram: centered at 4, spread from 1 to 12.

1.19 (a)

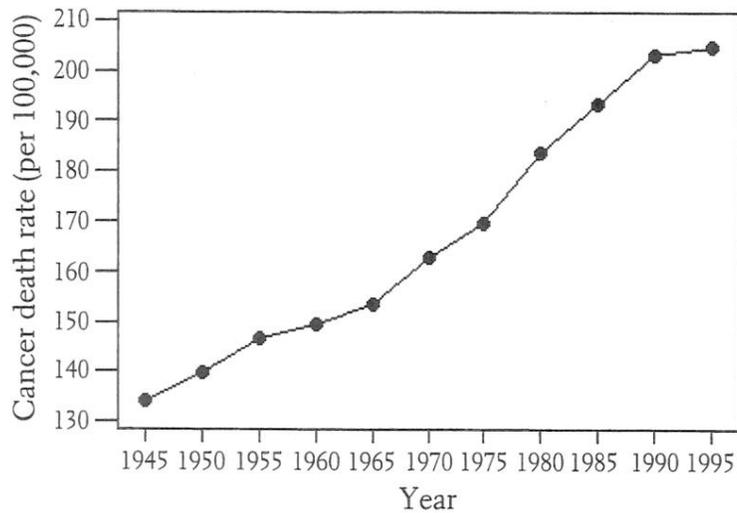
Percent	Cumulative frequency	Relative cumulative frequency	Percent	Cumulative frequency	Relative cumulative frequency
5.0-5.9	1	2%	12.0-12.9	25	50%
6.0-6.9	1	2%	13.0-13.9	39	78%
7.0-7.9	1	2%	14.0-14.9	45	90%
8.0-8.9	2	4%	15.0-15.9	49	98%
9.0-9.9	3	6%	16.0-16.9	49	98%
10.0-10.9	5	10%	17.0-17.9	49	98%
11.0-11.9	14	28%	18.0-18.9	50	100%



- (b) • Percentage of states in which percentage of “65 and older” is less than 15% = 90%, since the point (15, 90) lies on the ogive.
 - 40th percentile of distribution $\approx 12.4\%$, since the horizontal line drawn from 40% on the vertical axis intersects the ogive at a point whose horizontal coordinate is approximately 12.4%. Less than 40% of states have 12.4% or less of their population aged 65 or older.
 - Answers vary.
- 1.20 (a) The center corresponds to the 50th percentile. Draw a horizontal line from the value 50 on the vertical axis and determine the point on the ogive where the line intersects the ogive. Then draw a vertical line from this point to the horizontal axis. The line intersects the axis at approximately \$28. Thus, \$28 is the estimate of the center.
- (b) The 20th percentile.
- (c)



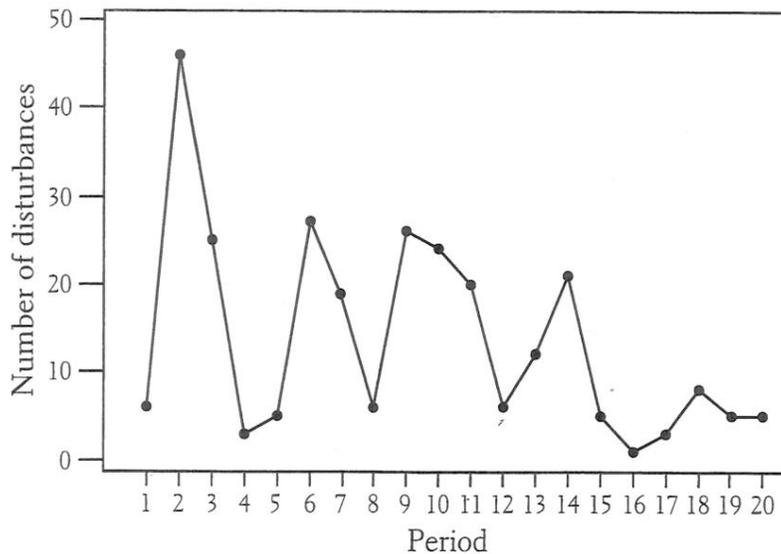
1.21 (a)



The cancer death rate has risen steadily from 1945 to 1995, with the largest increase occurring in the period 1975–1980.

(b) No, the slower rate of increase during the period 1990–1995 suggests that some progress was made during that time (at least in terms of treating the disease effectively). However, we have yet to see a decrease in the death rate, indicating that much work remains to be done in terms of actively preventing the disease.

1.22 (a)



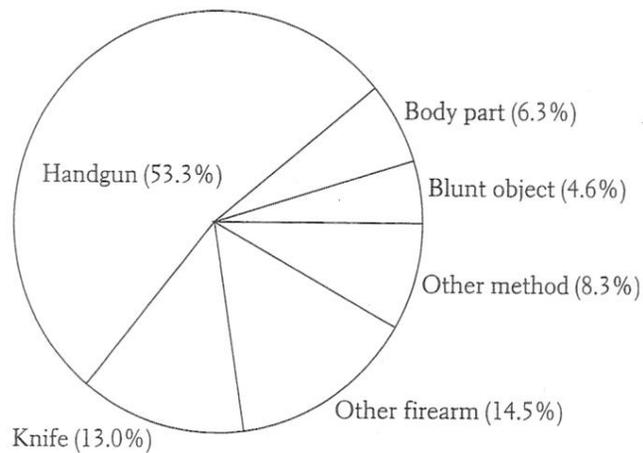
[Note: the periods are numbered consecutively from period 1, Jan.–Mar. 1968, to period 20, Oct.–Dec. 1972, on the horizontal axis.]

(b) The plot shows a decreasing trend—fewer disturbances overall in the later years—and more importantly, there is an apparent cyclic behavior. Looking at the table, the spring and summer months (April through September) generally have the most disturbances—probably for the simple reason that more people are outside during those periods.

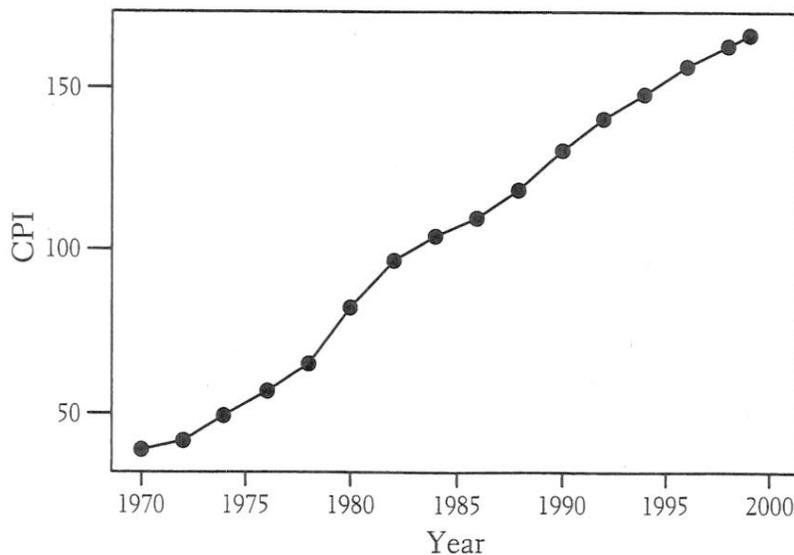
1.23 Gender, party voted for: Categorical
Age, income: Quantitative

1.24 (a) Car makes: a bar chart or pie chart. Car age: a histogram or stemplot. (b) Study time: a histogram or stemplot. Change in study hours: a time plot (average hours studied vs. time). (c) A bar chart or pie chart.

1.25 An "Other Methods" plot is needed because the sum of the percentages for the other categories is less than 100%.



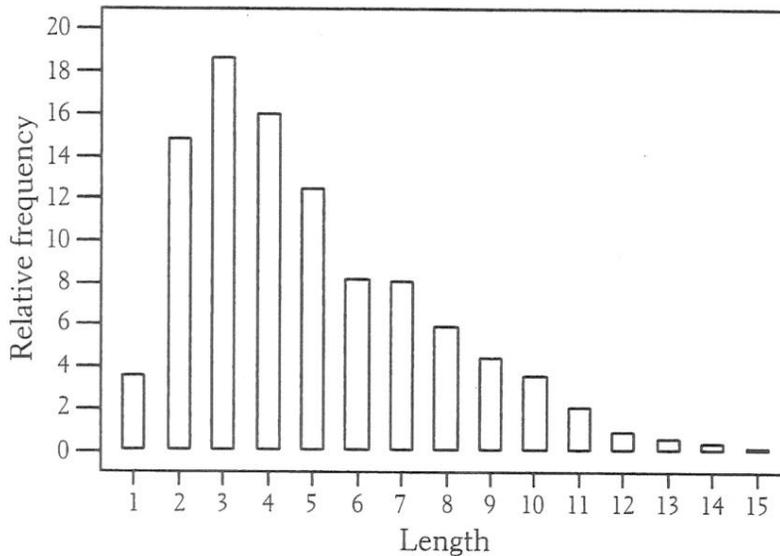
1.26 (a)



(c) Prices rose steadily during this period. There was no reversal of this trend in any of the periods under study.

(d) Prices were rising fastest during the mid- to late 1970s and rising slowest during the early 1970s and the mid-1980s.

1.27 (a)



The distribution is skewed to the right with a single peak. There are no gaps or outliers.

(b) Shakespeare was somewhat more likely to use short words and somewhat less likely to use extremely long words than *Popular Science*. However, the distributions have strongly similar shapes.

1.28

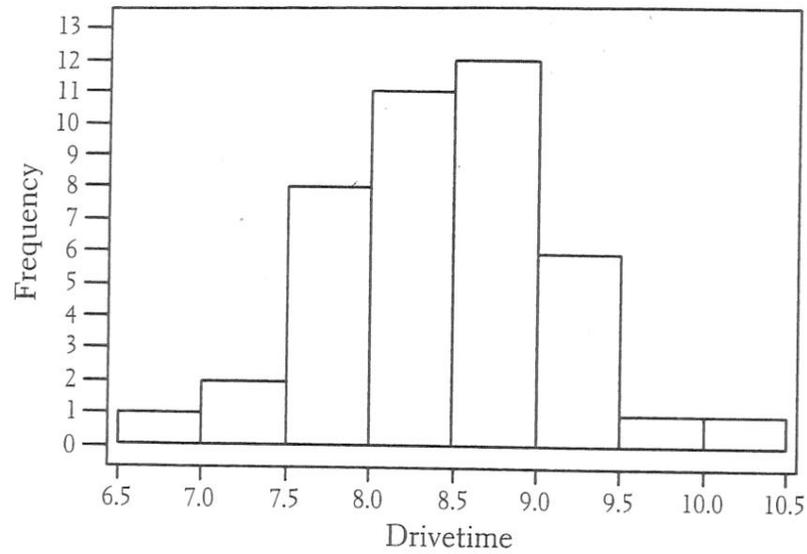
48		8
49		
50		7
51		0
52		6799
53		04469
54		2467
55		03578
56		12358
57		59
58		5

Stem = first two digits Leaf = last digit.

The distribution is roughly symmetric with one value (4.88) that is somewhat low. The center of the distribution is between 5.4 and 5.5.

Based on the plot, we would estimate the Earth's density to be about halfway between 5.4 and 5.5.

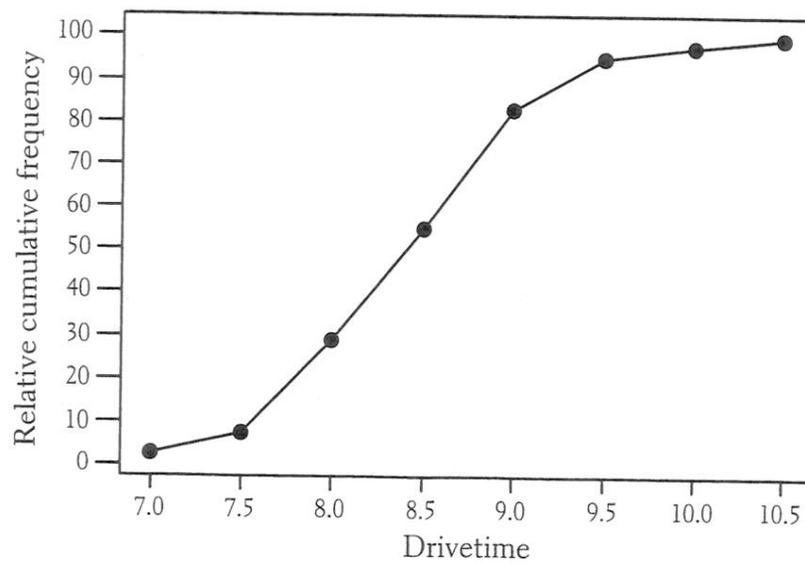
1.29 (a)



The distribution is roughly symmetric with no clear outliers.

(b)

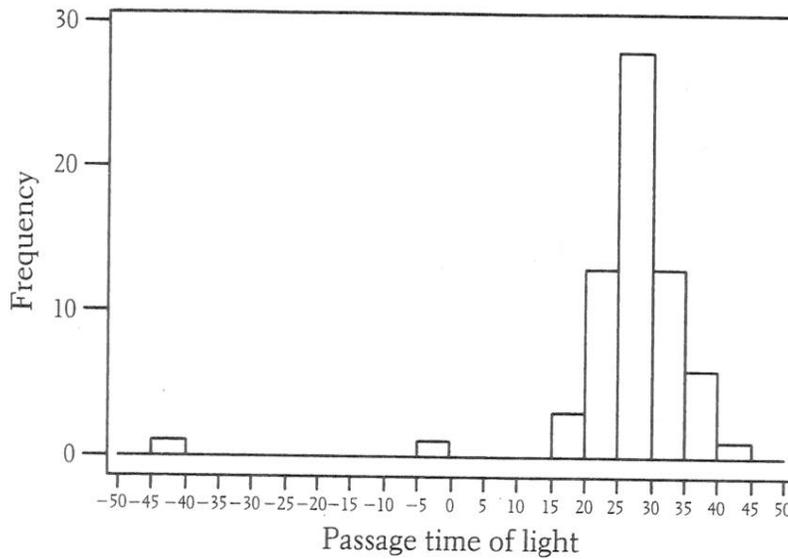
Drivetime	Cum. freq.	Rel. cum. freq.
7.0	1	2.4%
7.5	3	7.1%
8.0	12	28.6%
8.5	23	54.8%
9.0	35	83.3%
9.5	40	95.2%
10.0	41	97.6%
10.5	42	100%



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

(d) $8.0 \approx 28$ th percentile

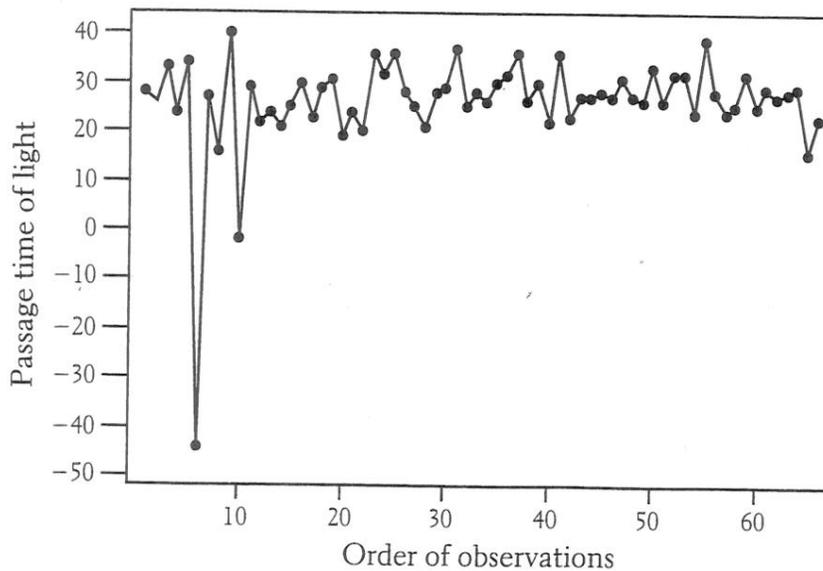
1.30 (a)



A stemplot would have much the same appearance as the histogram, but it would be somewhat less practical, because of the large number of observations with common stems (in particular, the stems 2 and 3).

(b) The histogram is approximately symmetric with two unusually low observations at -44 and -2 . Since these observations are strongly at odds with the general pattern, it is highly likely that they represent observational errors.

(c)



(d) Newcomb's worst measurement errors occurred early in the observation process. As the observations progressed, they became more consistent.

1.31 (a) $n = 14$, $\sum x = 1190$. The mean is $\bar{x} = \frac{\sum x}{n} = \frac{1190}{14} = 85$.

(b) If the 15th score is 0, then $n = 15$, $\sum x = 1190$, and the new mean is

$$\bar{x} = \frac{\sum x}{n} = \frac{1190}{15} = 79.3$$

The fact that this value of \bar{x} is less than 85 indicates the nonresistance property of \bar{x} . The extremely low outlier at 0 pulled the mean below 85.

(c) Minitab splits the decades to show greater detail.

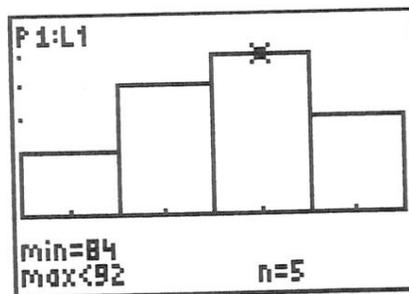
Stem-and-leaf of C1 N = 14
Leaf Unit = 1.0

```

7  4
7  568
8  024
8  67
9  013
9  68

```

And here is a histogram, with the widths of the bars specified to correspond to letter grades: D (68–75), C (76–83), B (84–91), and A (92–100). Both plots show a fairly balanced or symmetric distribution, with the histogram suggesting a slight skewness to the left. (Note that the mean and the median are the same (85).)



Given a rather small data set like this one, the stem plot would normally be preferable. But since we are very interested in letter grades in this case, perhaps the histogram would be most informative.

1.32 (a)

```

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

```

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

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

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

1.33 Since the mean $\bar{x} = \$1.2$ million and the number of players on the team is $n = 25$, the team's annual payroll is

$$(\$1.2 \text{ million}) (25) = \$30 \text{ million}$$

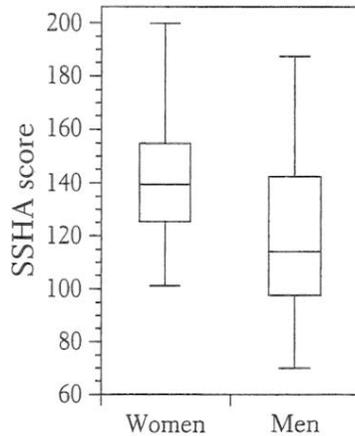
If you knew only the median salary, you would not be able to calculate the total payroll because you cannot determine the *sum* of all 25 values from the median. You can only do so when the arithmetic average of the values is provided.

1.34 $\bar{x} = \frac{\$480,000}{8} = \$60,000$. Seven of the eight employees (everyone but the owner) earned less than the mean. The median is $M = \$22,000$.

A recruiter might try to mislead applicants by telling them the mean salary, rather than the median salary, when the applicants ask about the "average" or "typical" salary. The median is a far more accurate depiction of a "typical" employee's earnings, because it resists the effects of the outlier at \$270,000.

1.35 Mean = \$675,000; median = \$330,000. The mean is nonresistant to the effects of the extremely high incomes in the right tail of the distribution. It will therefore be larger than the median.

1.36 (a)



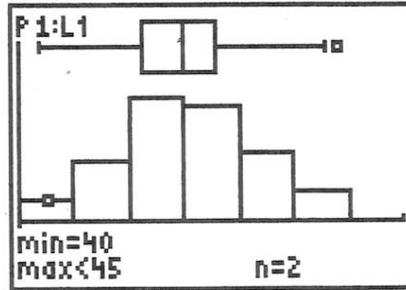
(b)

	\bar{x}	M	Five-number summaries				
Women	141.06	138.5	101	126	138.5	154	200
Men	121.25	114.5	70	98	114.5	143	187

(c) All the displays and descriptions reveal that women generally score higher than men. The men's scores ($IQR = 45$) are more spread out than the women's (even if we don't ignore the outlier). The shapes of the distributions are reasonably similar, with each displaying right skewness.

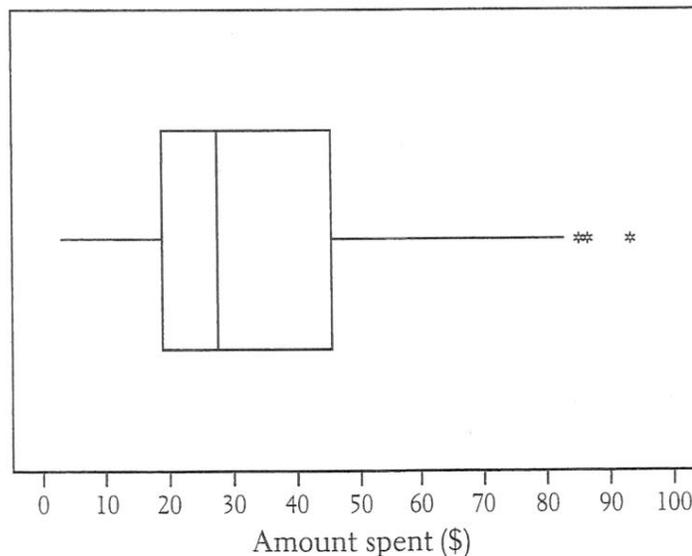
1.37 (a) The mean and median should be approximately equal since the distribution is roughly symmetric.

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



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

- 1.38 Yes, IQR is resistant. Take the data set 1, 2, 3, 4, 5, 6, 7, 8 as an example. In this case the median = 4.5, $Q_1 = 2.5$, $Q_3 = 6.5$, and $IQR = 4$. Changing any "extreme" value (that is, any value outside the interval between Q_1 and Q_3) will have no effect on the IQR. For example, if 8 is changed to 88, IQR will still equal 4.
- 1.39 (a) $\bar{x} = 34.7022$ and $n = 50$: Thus, total amt. spent = $(34.7022)(50) = \$1735.11$.
 (b)



The boxplot indicates the presence of several outliers. According to the $1.5 \times IQR$ rule, these outliers are 85.76, 86.37, and 93.34.

- 1.40 (a) $\bar{x} = 32.4 \div 6 = 5.4$. (b) $\sum(x_i - \bar{x})^2 = (0.2)^2 + (-0.2)^2 + (-0.8)^2 + (-0.5)^2 + (0.3)^2 + (1.0)^2 = 2.06$; $s^2 = 2.06 \div 5 = 0.412$; $s = \sqrt{0.412} = 0.6419$. (c) Yes, they agree: $\bar{x} = 5.4$, $s = 0.6418722614 \approx 0.6419$.

1.41 (a) $\sum(x_i - \bar{x})^2 = (-12.1)^2 + (1.9)^2 + (-10.1)^2 + (12.9)^2 + (34.9)^2 + (6.9)^2 + (-3.1)^2 + (-.1)^2 + (-18.1)^2 + (-13.1)^2 = 2192.9$.

$s^2 = 2192.9/9 = 243.66, s = 15.609$.

(b) Excluding the outlier at 61, we obtain $\bar{x} = 22.2, s = 10.244$. The outlier caused the values of both measures to increase; the increase in s is more substantial. Clearly, s is not a resistant measure of spread.

1.42 The five-number summary is somewhat preferable due to the skewness of the distribution.

Five-number summary: 5.5, 11.5, 12.75, 13.725, 18.3.

1.43 (a) 1, 1, 1, 1. (b) 0, 0, 10, 10. (c) For (a), any set of four identical numbers will have $s = 0$. For (b), the answer is unique; here is a rough description of why. We want to maximize the “spread-out”-ness of the numbers (that is what standard deviation measures), so 0 and 10 seem to be reasonable choices based on that idea. We also want to make each individual squared deviation— $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, (x_3 - \bar{x})^2$, and $(x_4 - \bar{x})^2$ —as large as possible. If we choose 0, 10, 10, 10—or 10, 0, 0, 0—we make the first squared deviation $(7.5)^2$, but the other three are only $(2.5)^2$. Our best choice is two at each extreme, which makes all four squared deviations equal to 5^2 .

1.44 (a) $\bar{x} = 7.5/5 = 1.5, s = .436$.

(b) To obtain \bar{x} and s in centimeters, multiply the results in inches by 2.54: $\bar{x} = 3.81$ cm, $s = 1.107$ cm.

(c) The average cockroach length can be estimated as the mean length of the five sampled cockroaches: that is, 1.5 inches. This is, however, a questionable estimate, because the sample is so small.

1.45 (a) The mean and the median will both increase by \$1000.

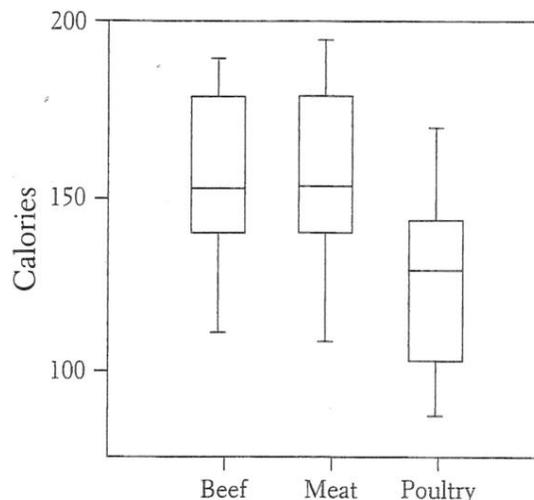
(b) No. Each quartile will increase by \$1000, thus the difference $Q_3 - Q_1$ will remain the same.

(c) No. The standard deviation remains unchanged when the same amount is added to each data value.

1.46 A 5% across-the-board raise will cause both IQR and s to increase. The transformation being applied here is $x^* = 1.05x$, where x = the old salary and x^* = the new salary. Both IQR and s will increase by a factor of 1.05.

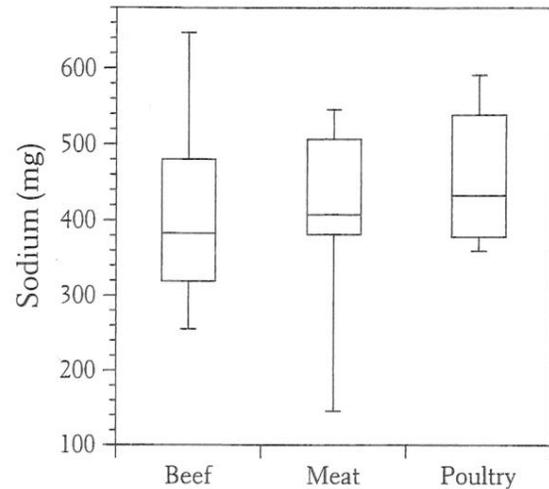
1.47 Calories: There seems to be little difference between beef and meat hot dogs, but poultry hot dogs are generally lower in calories than the other two. In particular, the median number of calories in a poultry hot dog is smaller than the lower quartiles of the other two, and the poultry lower quartile is less than the minimum calories for beef and meat.

Type	Min	Q_1	M	Q_3	Max
Beef	111	140	152.5	178.5	190
Meat	107	138.5	153	180.5	195
Poultry	86	100.5	129	143.5	170

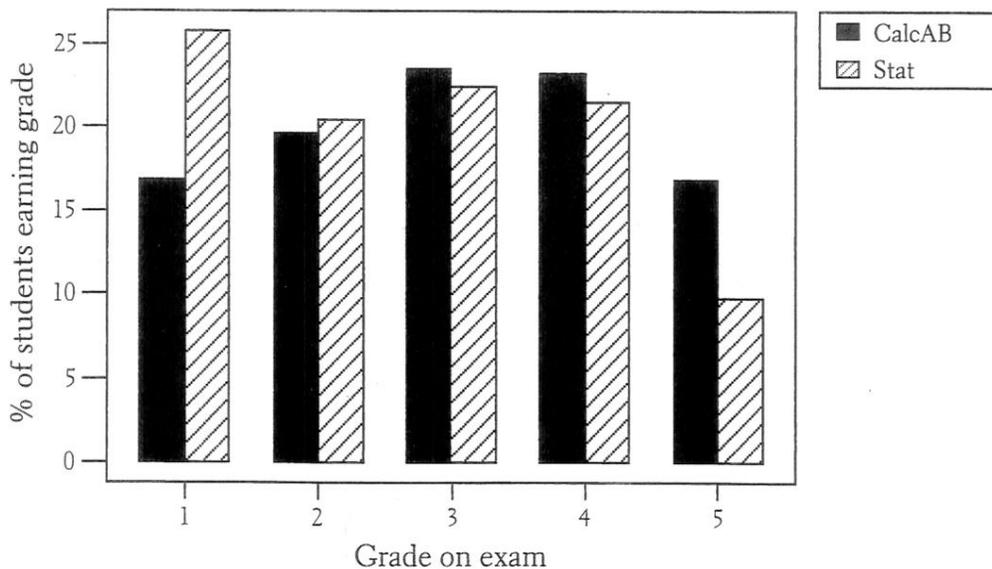


Sodium: Overall, beef hot dogs have less sodium (except for the one with the most sodium: 645 mg). Even if we ignore the low outlier among meat hot dogs, meat holds a slight edge over poultry.

Type	Min	Q_1	M	Q_3	Max
Beef	253	320.5	380.5	478	645
Meat	144	379	405	501	545
Poultry	357	379	430	535	588

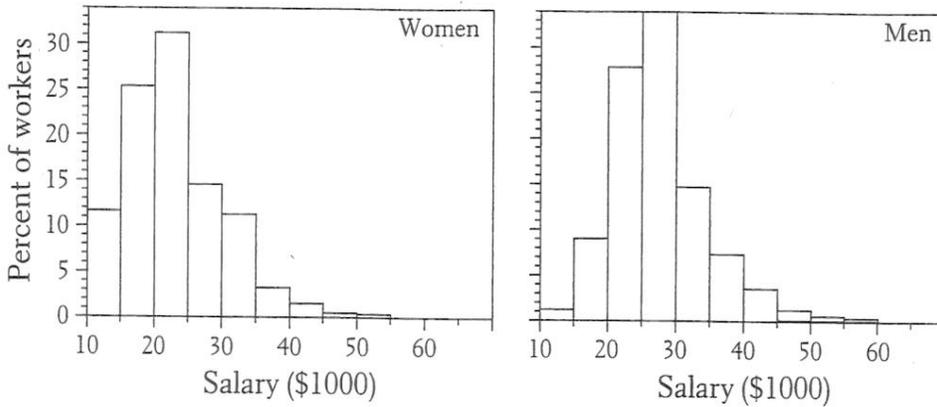


1.48 (a)



(b) The two distributions are roughly similar for grades of 2, 3, and 4. The major differences occur for grades 1 and 5. A considerably larger proportion of Statistics students receive a grade of 1, and a considerably smaller proportion of Statistics students receive a grade of 5. This suggests that the Statistics exam is harder, at least in the sense that students are more likely to get very poor grades on the Statistics exam than on the Calculus AB exam.

1.49 (a)



- Use relative frequency histograms, since there are considerably more men than women.
- (b) The two histograms are both skewed to the right (as income distributions often are). Women's salaries are generally lower than men's.
- (c) The women's total sums to 100.1%, due to roundoff error.

1.50 (a)

V-AA		I-AAA
776	3	8
44	4	1
65	4	677
4	5	1234
8	5	568
20	6	2233444
97766	6	5578
42	7	114
	7	678
	8	
6	8	7
32	9	1
886	9	
	10	
	10	6

(b) The I-AAA point distribution is skewed to the right, while the V-AA distribution is roughly symmetric. The median of the I-AAA distribution is 63.5, and the median of the V-AA distribution is 66; they are roughly equivalent. The I-AAA distribution is slightly more spread out, due in large part to the outlier at 106. The median and quartiles are the best measures for comparison purposes, because one of the distributions is skewed.

V-AA	I-AAA
1	0
8	0
4	1
	1
300	2
55	2
1	3
	3
2	4
7	4
3	5
	5
	6

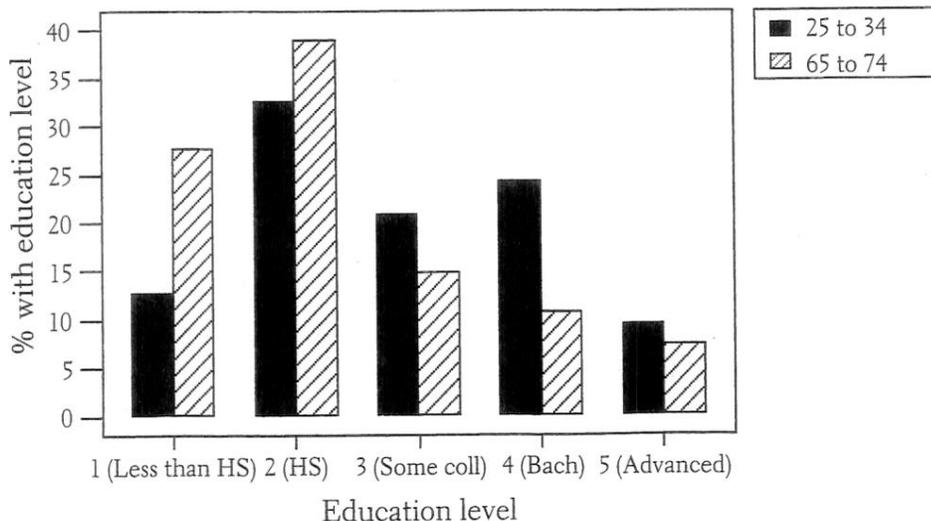
A back-to-back stemplot of the margin of victory distributions reveals that large margins of victory were more likely to occur in the V-AA games than in the I-AAA games (despite the few I-AAA outliers). The medians of the V-AA and I-AAA distributions are 24 and 12.5, respectively, which further suggests that the average margin of victory tended to be larger in the V-AA games.

1.51

10	7
11	
12	
13	5689
14	067
15	3
16	
17	2359
18	2
19	015

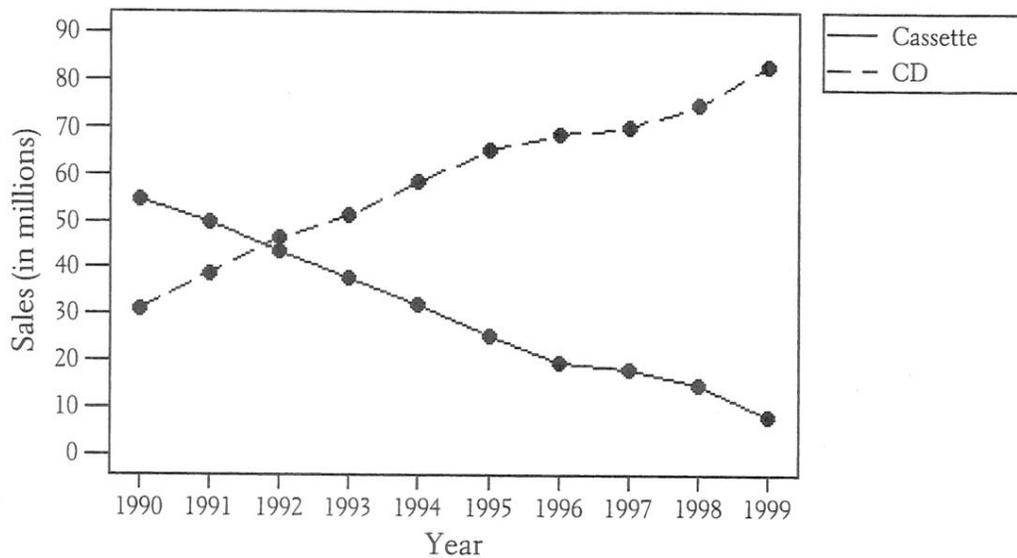
There are two distinct clusters of brands (the distribution has two peaks with a gap between them) and one outlier in the lower tail. The boxplot hid the clusters. The quartiles are roughly at the centers of the two clusters, so much of the *IQR* is the spread between the clusters. Because of this, the $1.5 \times IQR$ rule used in drawing the boxplot did not call attention to the outlier.

1.52



The side-by-side bar graph shows several distinct differences in educational attainment between the two groups. The 65-to-74 age group was more likely to have earned no more than a high school diploma. The 25-to-34 age group was more likely to have gone to college and to have completed a Bachelor's degree. However, the percentages in the "Advanced" group are relatively similar, indicating that those 65-to-74 year-olds who managed to complete college were more likely to earn an advanced degree.

1.53



Over the period 1990–1999, sales of cassettes declined steadily, while sales of CD's increased steadily. The first year during which sales of CD's exceeded sales of cassettes was 1992.

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

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

1.55 (a) $x^* = 746x$, where x = measurement in horsepower and x^* = measurement in watts. The mean, median, IQR, and standard deviation will all be multiplied by 746.

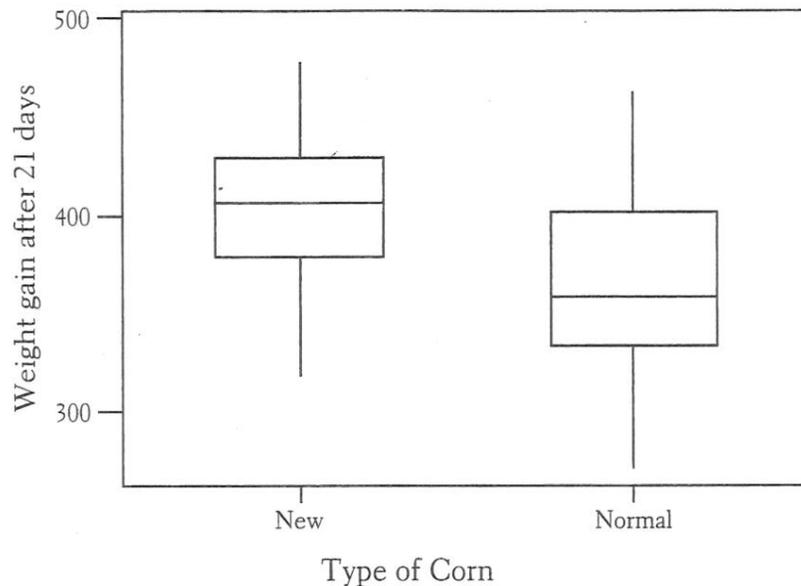
(b) $x^* = (5/9)(x - 32)$, where x = measurement in °F, x^* = measurement in °C. The mean and median will be multiplied by 5/9 and the result reduced by $(5/9)(32) = 160/9$. The IQR and standard deviation will be multiplied by 5/9.

(c) $x^* = x + 10$, where x = original test score, x^* = "curved" test score. The mean and median will increase by 10. The IQR and standard deviation will remain the same.

1.56 Variance is changed by a factor of $2.54^2 = 6.4516$; generally, for a transformation $x_{\text{new}} = a + bx$, the new variance is b^2 times the old variance.

1.57 (a) Five-number summary for normal corn: 272, 337, 358, 400.5, 462. Five-number summary for new corn: 318, 383.5, 406.5, 428.5, 477. The boxplots show that the new corn seems to increase

weight gain—in particular, the median weight gain for new-corn chicks was greater than Q_3 for those that ate normal corn.



(b) Normal corn: $\bar{x} = 366.3$, $s = 50.805$; new corn: $\bar{x} = 402.95$, $s = 42.729$. On the average, the chicks that were fed the new corn gained 36.65 grams more mass (weight) than the other chicks.

(c) Means and standard deviations will all be multiplied by $1/28.35$ in order to convert grams to ounces. Normal corn: $\bar{x} = 12.92$, $s = 1.792$; new corn: $\bar{x} = 14.21$, $s = 1.507$.

1.58 (a) Mean—although incomes are likely to be right-skewed, the city government wants to know about the total tax base. (b) Median—the sociologist is interested in a “typical” family, and wants to lessen the impact of the extremes.

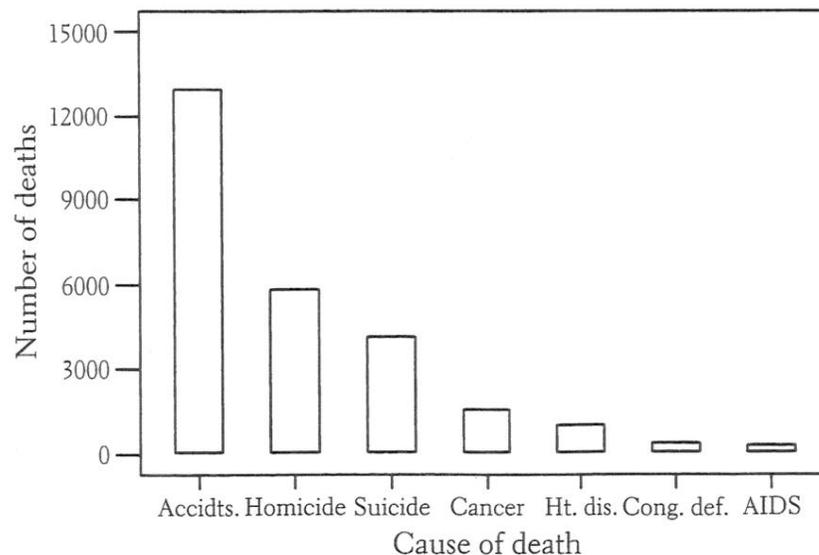
1.59 Possible answers are total profits, number of employees, total value of stock, and total assets.

1.60 (a) All major league baseball players as of opening day, 1998.

(b) Team (categorical), position (categorical), age (quantitative), salary (quantitative).

(c) Age is measured in years, salary in \$1000s (thousands of dollars).

1.61 (a)

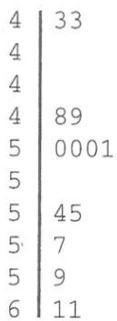


(b) To make a pie chart, you need to know the number of people in the “death from other causes” category. Including “death from other causes” ensures that the category percentages will sum to 100%.

1.62 (a) Since a person cannot choose the day on which he or she has a heart attack, one would expect that all days are “equally likely”—no day is favored over any other. While there is *some* day-to-day variation, this does seem to be supported by the chart.

(b) Monday through Thursday are fairly similar, but there is a pronounced peak on Friday, and lows on Saturday and Sunday. Patients do have some choice about when they leave the hospital, and many probably choose to leave on Friday, perhaps so that they can spend the weekend with the family. Additionally, many hospitals cut back on staffing over the weekend, and they may wish to discharge any patients who are ready to leave before then.

1.63 (a)

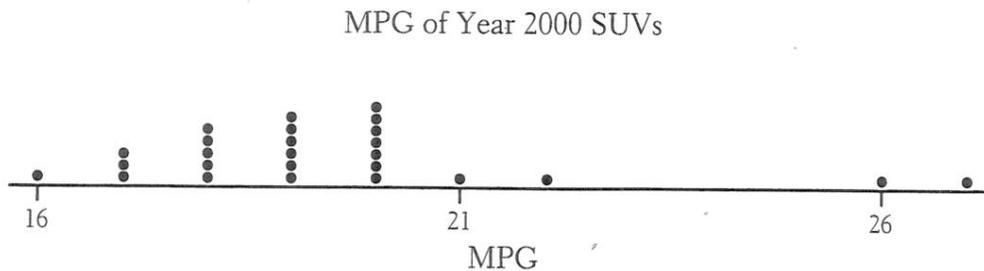


(b) Median = 50.4.

(c) $Q_3 = 57.4$. Landslides occurred in 1956, 1964, 1972, and 1984.

1.64 Slightly skewed to the right, centered at 4.

1.65 (a)



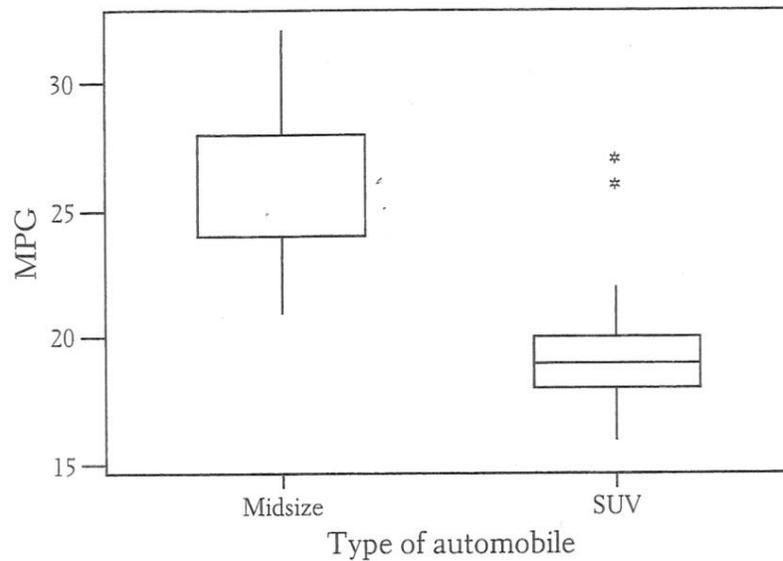
The distribution is skewed to the right with a peak at 20. There are two outliers, at 26 and 27 (Toyota RAV4 and Subaru Forester, respectively).

Variable	N	Mean	Median	StDev
MPG	26	19.500	19.000	2.470

Variable	Minimum	Maximum	Q1	Q3
MPG	16.000	27.000	18.000	20.000

The fact that $\bar{x} > M$ reflects the right skewness of the distribution.

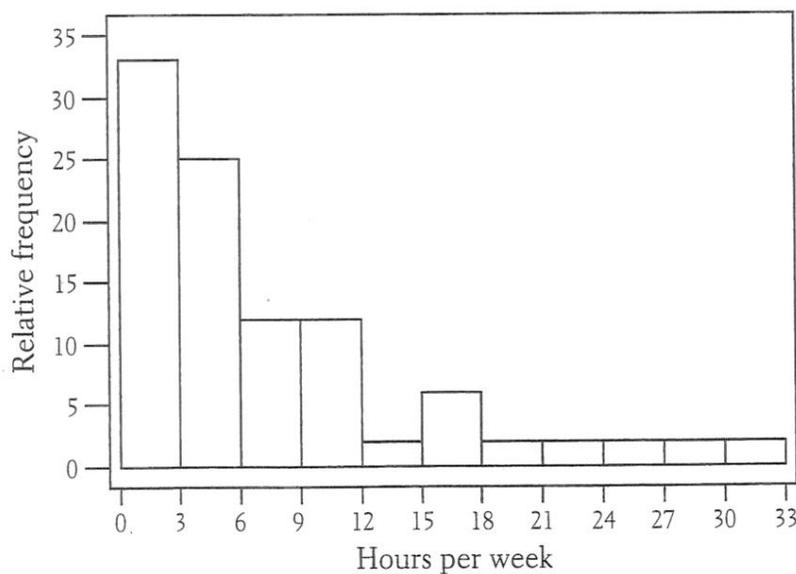
(b)



The boxplots clearly show that the midsize automobiles tend to have higher MPGs than the SUVs. The only SUV observations falling in the “middle” of the midsize distribution are the two outliers at 26 and 27.

(Note: the lack of a “median line” in the midsize boxplot is the result of the median and Q_3 being the same value.)

1.66 (a)



Hrs. per wk.	Rel. freq. (approx)
0-3	.33
3-6	.25
6-9	.12
9-12	.12
12-15	.02
15-18	.06
18-21	.02
21-24	.02
24-27	.02
27-30	.02
30-33	.02
	<u>1</u>

(b) From the ogive: Median ≈ 6 (50th percentile), $Q_1 \approx 2.5$ (25th percentile), $Q_3 \approx 11$ (75th percentile). There are outliers, according to the $1.5 \times IQR$ rule, because values exceeding $Q_3 + (1.5 \times IQR) = 23.75$ clearly exist.

(c) 10 hours \approx 70th percentile.

1.67 (a) $\text{Min} = -34.04$, $Q_1 = -2.95$, $\text{Med} = 3.47$, $Q_3 = 8.45$, $\text{Max} = 58.68$.

(b) The distribution is fairly symmetric, with a single peak in the high single digits (5 to 9). There are no gaps, but four “low” outliers and five “high” outliers are listed separately.

(c) 58.68% of \$1000 is \$586.60. The stock is worth \$1586.50 at the end of the best month. In the worst month, the stock lost $1000(.3404) = \$340.40$, so the \$1000 decreased in worth to $1000 - 340.40 = \$659.60$.

(d) $\text{IQR} = Q_3 - Q_1 = 8.45 - (-2.95) = 11.4$

$$1.5 \times \text{IQR} = 17.1$$

$$Q_1 - (1.5 \times \text{IQR}) = -2.95 - 17.1 = -20.05$$

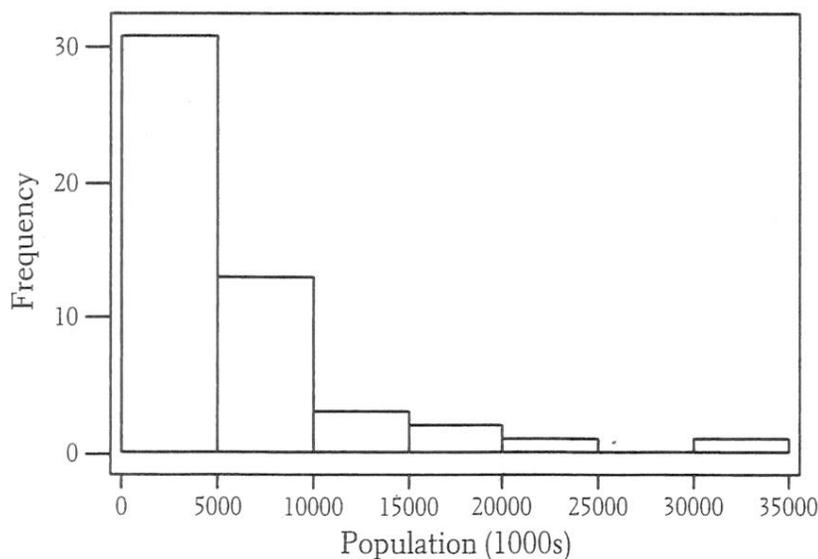
$$Q_3 + (1.5 \times \text{IQR}) = 8.45 + 17.1 = 25.55$$

The four “low” and five “high” values are all outliers according to the criterion. It does appear that SPLUS uses the $1.5 \times \text{IQR}$ criterion to identify outliers.

1.68 The difference in the mean and median indicates that the distribution of awards is skewed sharply to the right—that is, there are some *very* large awards.

1.69 The median—half are traveling faster than you, and half are traveling slower. (Actually, you have found *a* median—it could be that a whole range of speeds, say from 56 mph to 58 mph, might satisfy this condition.)

1.70



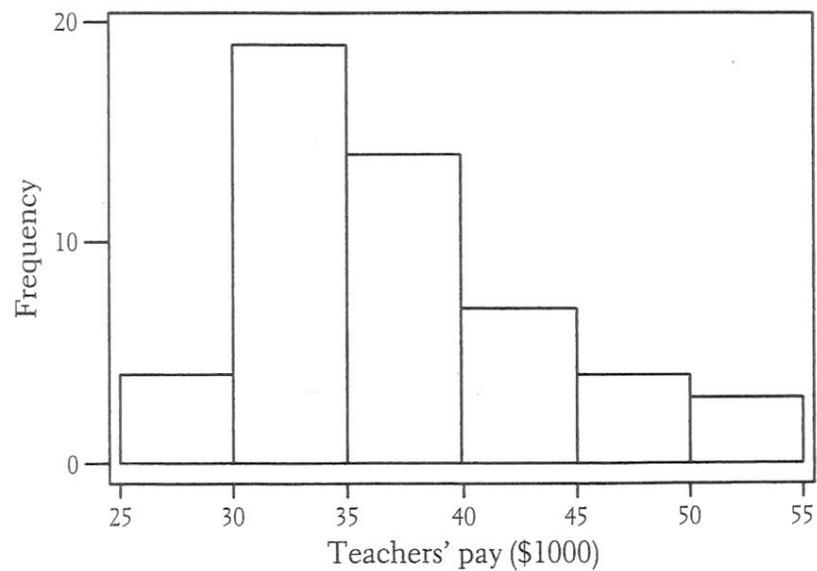
The distribution is strongly skewed right, with a spread of approximately 33,000 and a center located in the first (0–5000) class. The distribution is not surprising inasmuch as the many small to medium-sized U.S. states have small populations and a few states, in particular California, Texas, and New York, have large populations. According to the histogram, California is the lone outlier.

1.71

0	44
0	5556788888999
1	012223
1	68
2	1
2	5
3	244
3	
4	
4	9
5	002233
5	
6	0113
6	5578
7	0002
7	678
8	00

The distribution has two distinct peaks, in the second 0 stem and the first 5 stem. The midpoint is located between the 2 leaf and the first 4 leaf in the first 3 stem. The center does not provide us with information about the distribution's multiple peaks, or, more generally, about the fact that the distribution breaks into several distinct "clusters" of observations.

1.72

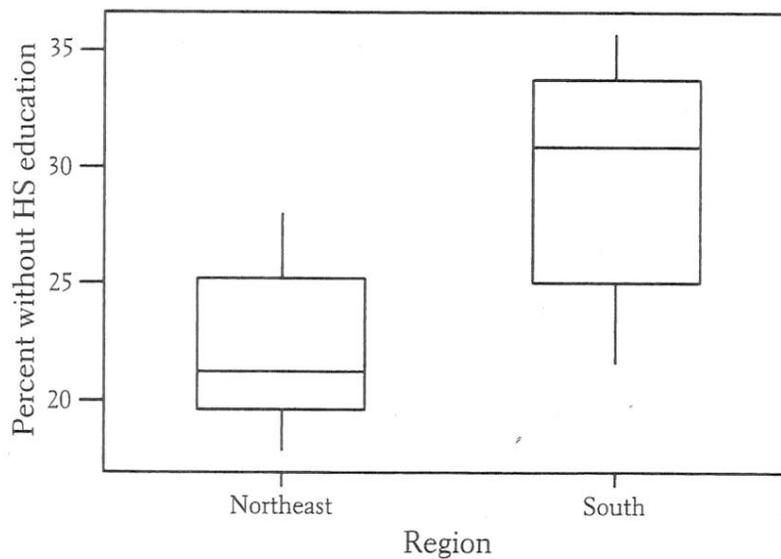


The distribution is skewed to the right. There are no apparent outliers.

1.73 (a)

20.8	Northeast
21.2	Northeast
20.0	Northeast
17.8	Northeast
28.0	Northeast
19.2	Northeast
23.3	Northeast
25.2	Northeast
25.3	Northeast
33.1	South
35.4	South
35.7	South
32.9	South
22.5	South
25.6	South
29.1	South
21.6	South
30.0	South
31.7	South
24.8	South
34.0	South

(b)



Northeast: Five-number summary: 17.8, 19.6, 21.2, 25.25, 28 (IQR = 5.65)
 Mean and standard deviation: $\bar{x} = 22.31$, $s = 3.34$

Southern: Five-number summary: 21.6, 25.2, 30.85, 33.55, 35.7 (IQR = 8.35)
 Mean and standard deviation: $\bar{x} = 29.7$, $s = 4.97$

The percent of individuals without high school diplomas is distinctly higher in the Southern states than it is in the Northeastern states. The first quartile of the Southern states and the third quartile of the Northeastern states are virtually the same. The shapes of the distribution are somewhat different, with the boxplots indicating that the Southern distribution is skewed left and the Northeastern distribution is skewed right.