

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).

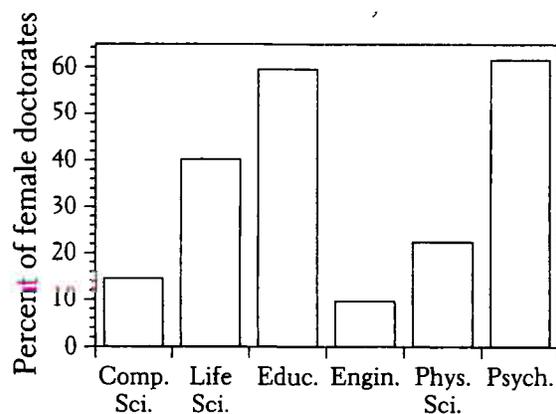
1.2 (a) Categorical. (b) Quantitative. (c) Categorical. (d) Categorical. (e) Quantitative. (f) Quantitative.

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 . . .)

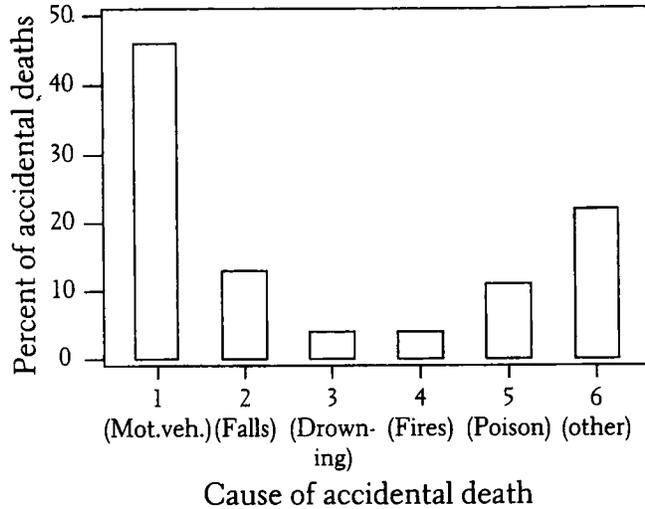
1.4 Possible answers (reasons should be given): unemployment rate, average (mean or median) income, quality/availability of public transportation, number of entertainment and cultural events, housing costs, crime statistics, population, population density, number of automobiles, various measures of air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

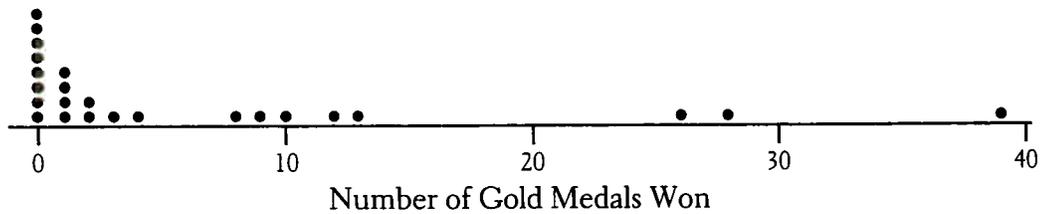


1.6 (a) Below. For example, “Motor Vehicles” is 46% since $\frac{42,340}{92,353} = 0.45846$ The “Other causes” category is needed so that the total is 100%. (b) Below. The bars may be in any order. (c) A pie chart *could* also be used, since the categories represent parts of a whole (all accidental deaths).

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

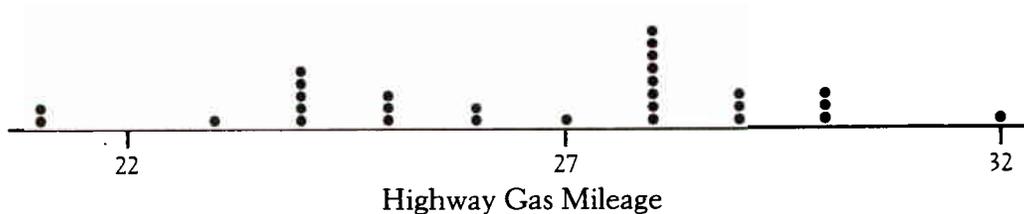


1.7



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.

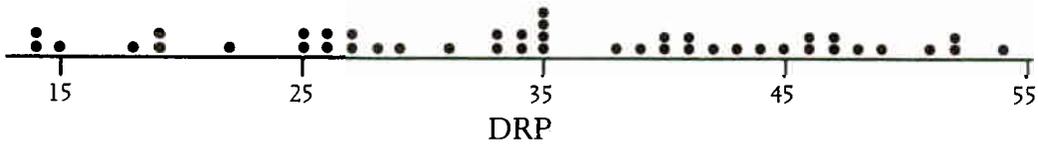
1.8



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

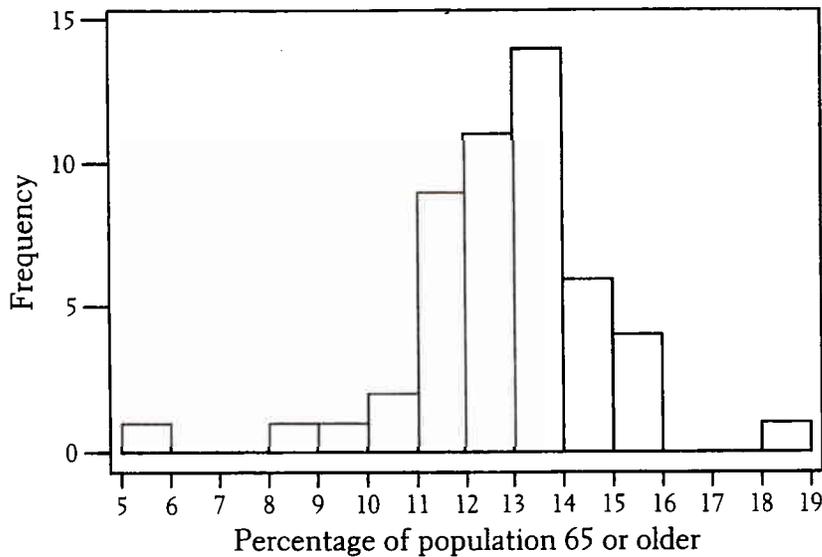
(b)

0	3
0	99
1	134
1	5677889
2	0001234
2	556688888
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.

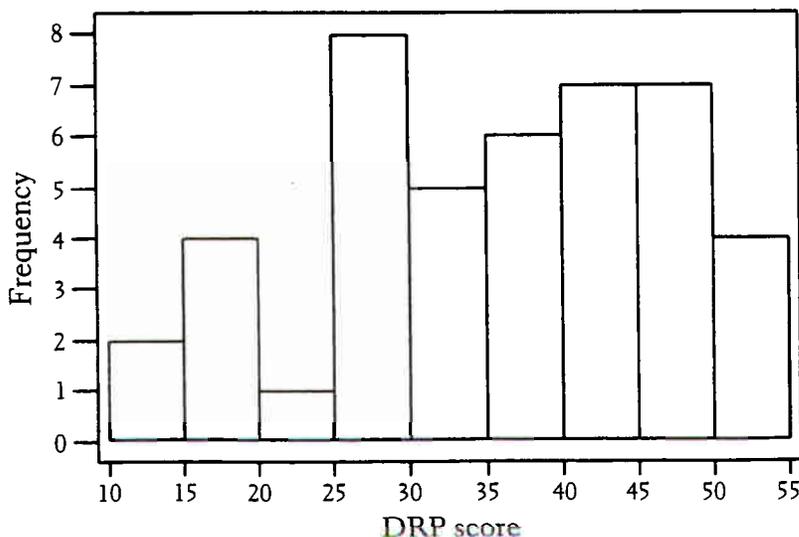
1.12 (a)



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.

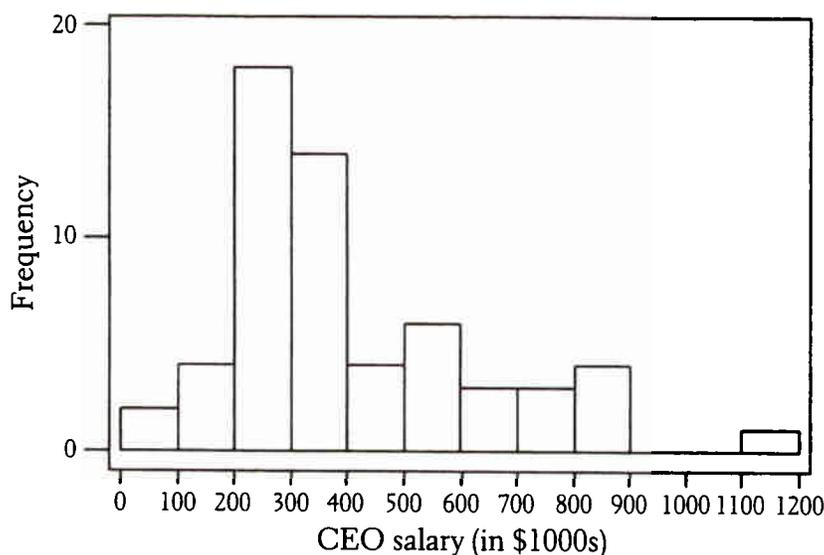
1.13



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.

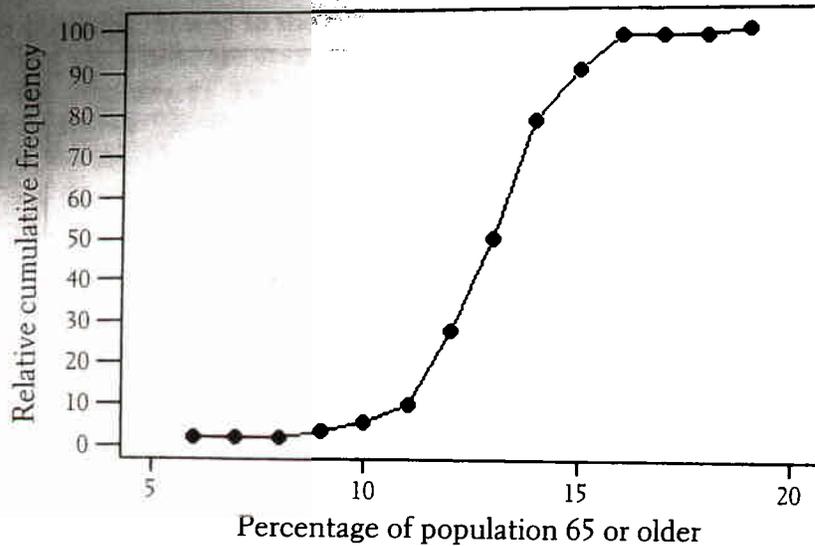
1.16 (a) Roughly symmetric, though it might be viewed as SLIGHTLY skewed to the right. (b) About 15%. (39% of the stocks had a total return less than 10%, while 60% had a return less than 20%. This places the center of the distribution somewhere between 10% and 20%.) (c) The smallest return was between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

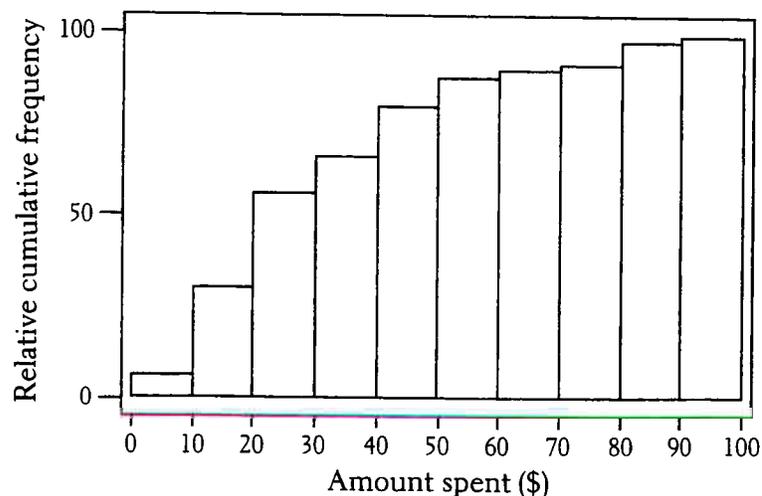
1.18 Lightning histogram: centered at noon (or more accurately, somewhere from 11:30 to 12:30). Spread is from 7 to 17 (or more accurately, 6:30 AM to 17:30, i.e., 5:30 PM). Shakespeare histogram: centered at 4, spread from 1 to 12.

1.19 (a)

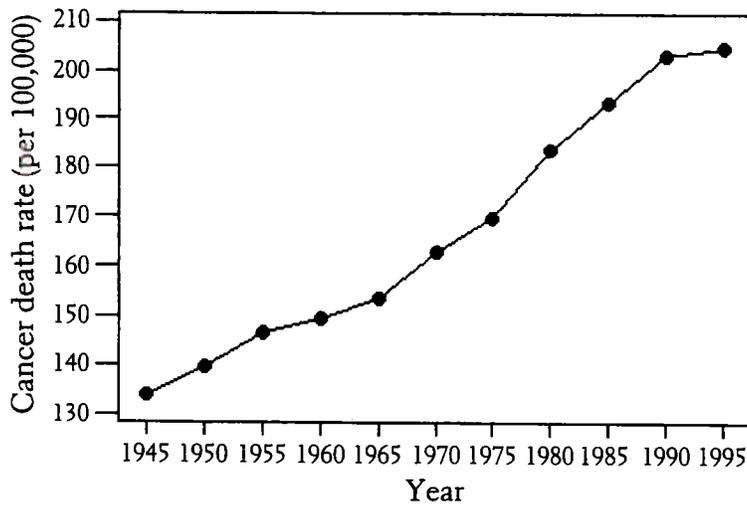
Percent	Cumulative frequency	Relative cumulative frequency	Percent	Cumulative frequency	Relative cumulative frequency
5.0-5.9	1	2%	12.0-12.9	25	50%
6.0-6.9	1	2%	13.0-13.9	39	78%
7.0-7.9	1	2%	14.0-14.9	45	90%
8.0-8.9	2	4%	15.0-15.9	49	98%
9.0-9.9	3	6%	16.0-16.9	49	98%
10.0-10.9	5	10%	17.0-17.9	49	98%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



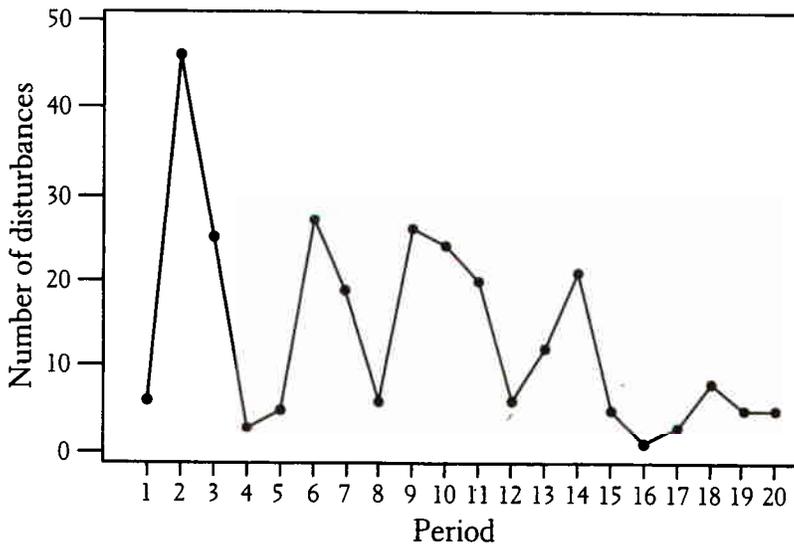
1.21 (a)



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)



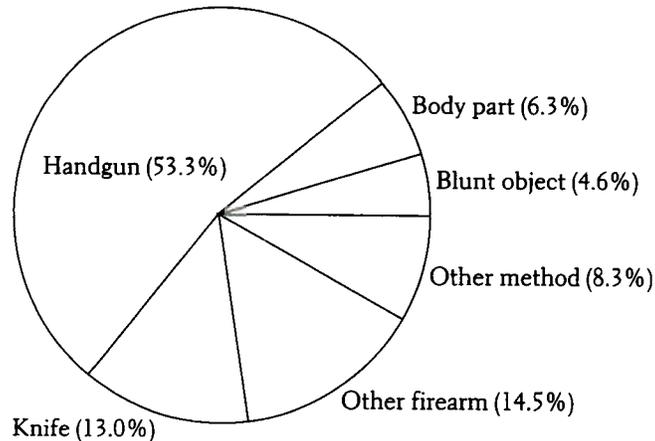
[Note: the periods are numbered consecutively from period 1, Jan.–Mar. 1968, to period 20, Oct.–Dec. 1972, on the horizontal axis.]

(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.

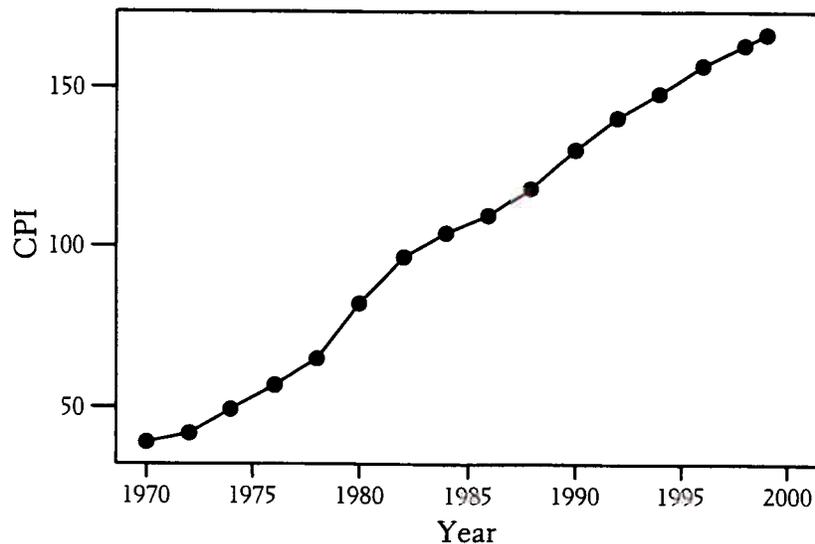
1.23 Gender, party voted for: Categorical
Age, income: Quantitative

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%.



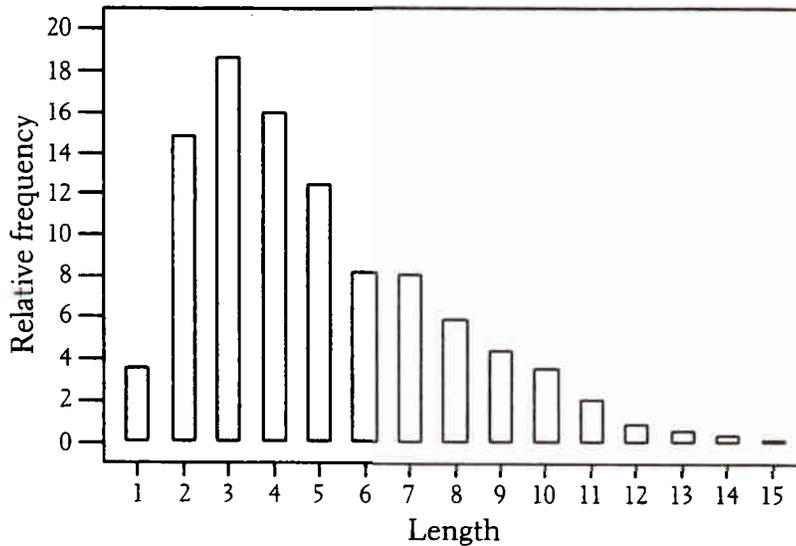
1.26 (a)



(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

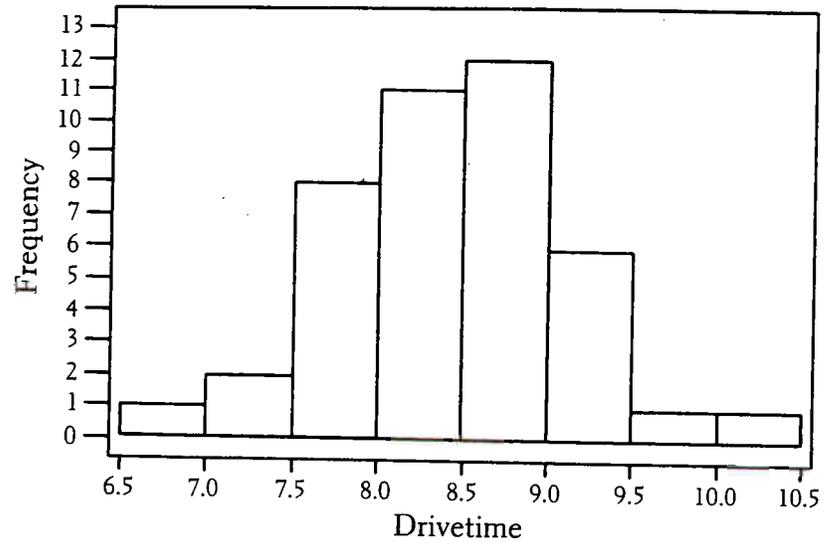
48	8
49	
50	7
51	0
52	6799
53	04469
54	2467
55	03578
56	12358
57	59
58	5

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.

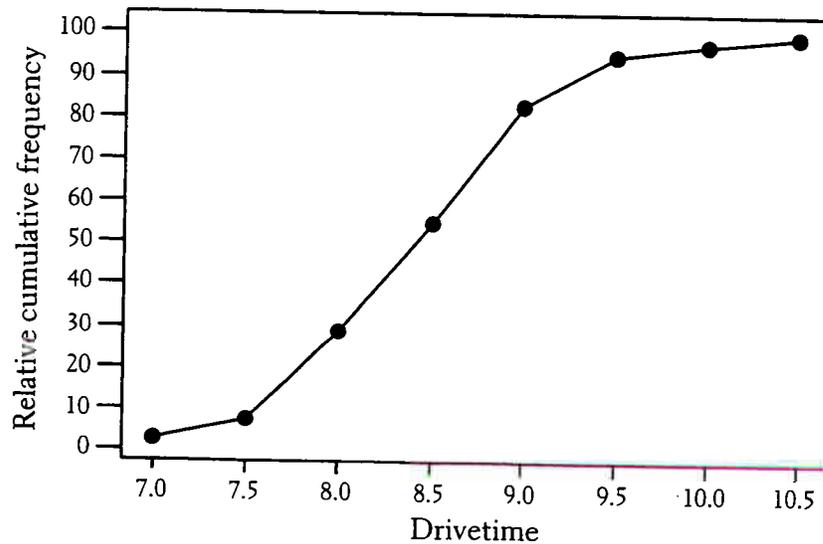
1.29 (a)



The distribution is roughly symmetric with no clear outliers.

(b)

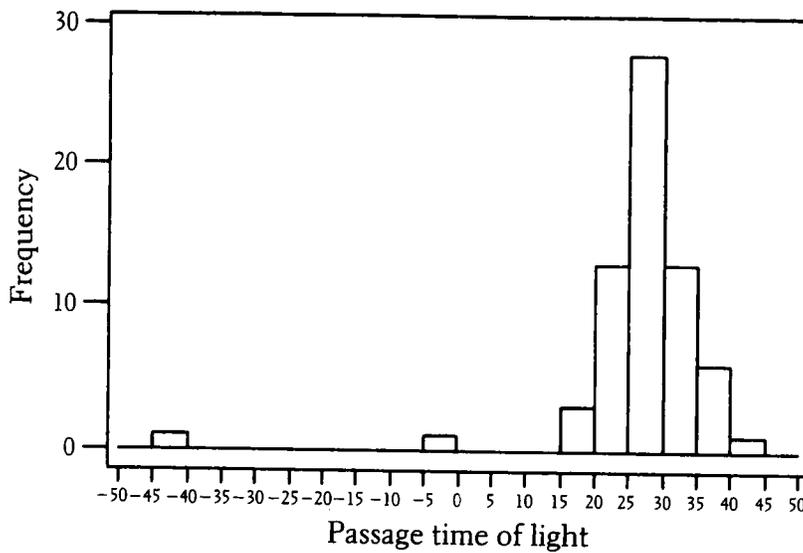
Drivetime	Cum. freq.	Rel. cum. freq.
7.0	1	2.4%
7.5	3	7.1%
8.0	12	28.6%
8.5	23	54.8%
9.0	35	83.3%
9.5	40	95.2%
10.0	41	97.6%
10.5	42	100%



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

(d) $8.0 \approx 28$ th percentile

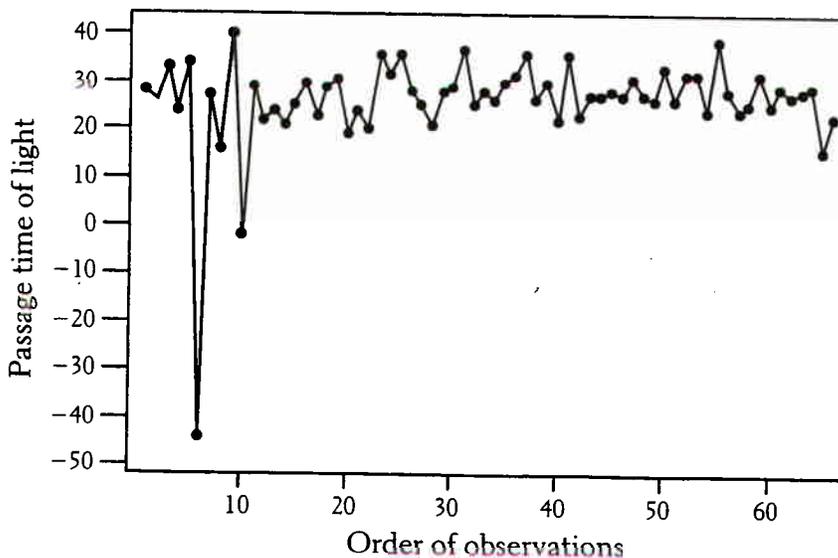
1.30 (a)



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.