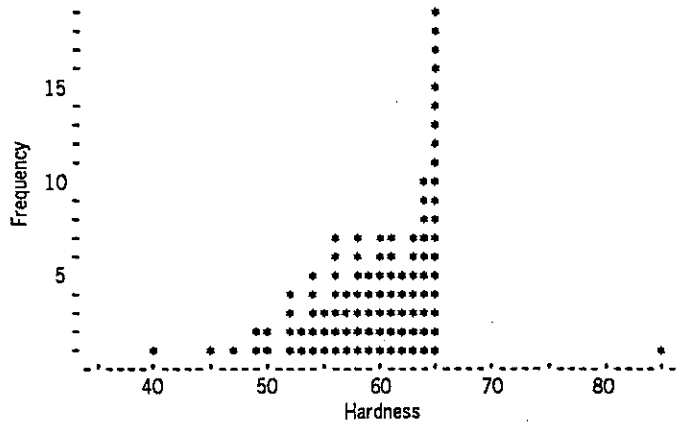
the median is shown as the vertical line near the middle of the box. The range of the distribution on each side is shown by an extended line, so the graph is sometimes called a *box and whisker* plot.

The box plot is not only easy to draw and understand; it also gives a good idea of the center and spread of the distribution. Center and spread are important enough that we will spend the rest of the chapter detailing how they can be measured.

## PROBLEMS

In many of the exercises based on real data, such as Problem 2-2, we have tried to preserve the reality yet keep the arithmetic manageable by using the following device: From the available data (which usually is a whole population, or at least a very large sample) we construct a small sample that closely resembles the population.

2-1 A simple frequency distribution can sometimes provide remarkable clues. For example, this histogram shows the measured hardness of 100 steel coils produced in a steel plant about 1970 (Roberts 1974, p. 73):




Clues in this histogram uncovered several troubles in the plant:

**a.** The employee who measured the hardness of each coil was aware of the maximum hardness the firm's managers would accept. What was she doing wrong, and what observations betrayed her?

**b.** Due to a scheduling error at the steel mill, one coil was made of the wrong kind of steel. Which observation showed this?

2-2 **a.** In a large American university, a random sample of female professors gave the following annual salaries (in thousands of dol-

lars, reconstructed from Katz, 1973). Without sorting into cells, graph the salaries as dots along an X-axis.

9, 12, 8, 10, 16

**b.** Using the same scale, construct a similar graph for the following sample of 25 *male* professors' salaries:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 13 | 11 | 19 | 11 | 22 | 22 | 13 | 11 | 17 | 13 |
| 27 | 14 | 16 | 13 | 24 | 31 | 9 | 12 | 15 | 15 |
| 21 | 18 | 11 | 9 | 13 | | | | | |

In your view, how good is the evidence that, over the whole university, men tended to earn more than women? (This issue will be answered more precisely later.)

**2-3** Graph the data in Problem 2-2 as box plots. (Hints: In **a**, the median is the middle observation. For the quartiles, cut off 1/4 of $5 \simeq 1$ observation from each end. In **b**, the median is the middle observation, with 12 below and 12 above. For the quartiles, cut off 1/4 of $25 \simeq 6$ observations from each end.)

**2-4** Sort the data of Problem 2-2**b** into cells with midpoints of 10, 15, 20, 25, 30, and draw the bar graph.

**2-5** Using the 25 men's salaries plotted in Problem 2-2**b** (where each observation represents 4% of the data), what percentile is a salary of 10 thousand? 20 thousand? 30 thousand?

**2-6** In 1990, the 789 million people in Europe lived in 25 different countries ranked as follows (in millions):

| | | | | | |
|---|---|---|---|---|---|
| Russia | 292 | Yugoslavia | 24 | Bulgaria | 9 |
| West Germany | 59 | East Germany | 17 | Sweden | 8 |
| Italy | 58 | Czechoslovakia | 16 | Austria | 8 |
| Britain | 56 | Holland | 15 | Switzerland | 6 |
| France | 56 | Hungary | 11 | Denmark | 5 |
| Spain | 39 | Portugal | 11 | Finland | 5 |
| Poland | 39 | Belgium | 11 | Norway | 4 |
| Romania | 24 | Greece | 9 | Ireland | 3 |
| | | | | Albania | 3 |

**a.** Graph the relative frequency distribution (with cells centered at 5, 15, 25, . . .).

Note how well this graph emphasizes that one country is overwhelming in size.

**b.** Find the median and two quartiles. (Hint: Problem 2-3.)

**c.** Draw the box plot.

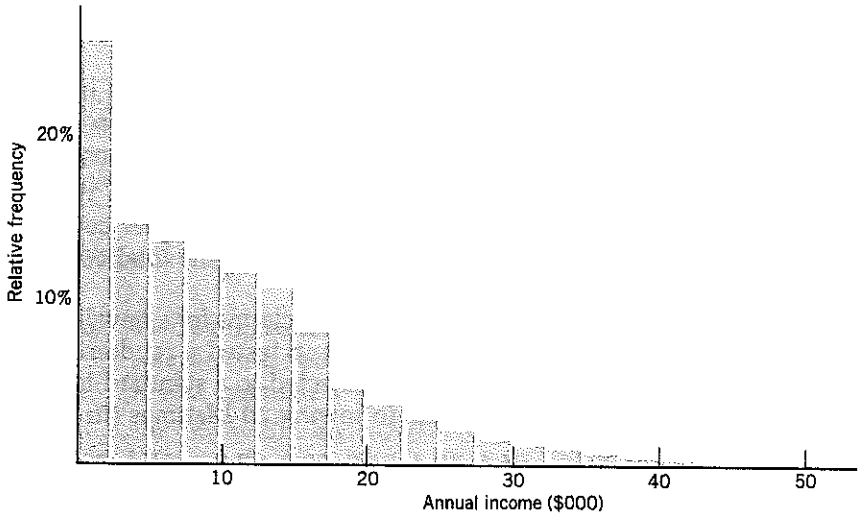



FIGURE 2-5
Incomes of American men, 1975. (Stat. Abst. of U.S., 1980, p. 462)

equally. This is both its advantage and disadvantage: it is the figure of most use to the tax department, since it gives the total income (78 million men × 10 thousand dollars = 780 billion dollars); yet it is not as good a measure of typical income as the median because it can be inflated by a few very large incomes (that is, it lacks the resistance of the median).

To sum up, we can conclude that the mode is the easiest, but the most inadequate measure of center. The median is more useful because it represents a more typical value, as many people understand the term. Finally, the mean is the only central measure that takes into account the total of all the observations. For this reason, the mean is often the most useful value in such fields as engineering, economics, and business. Since it has other advantages as well we will primarily use the mean for the rest of the book.

## PROBLEMS

2-7 One of the most important calculations for business or economics is to find a total. To find the total sales of 200 agents, for example, would we multiply 200 by the mode, median, or mean sales?

2-8 Overheard in a Scottish pub: "When a Scotsman moves from Scotland to England, he improves the average IQ in both places."

**a.** How could this be possible (or is it impossible)?

**b.** What would you hear in an English pub?

2-9 The annual tractor output of a multinational firm in seven different countries was as follows (in thousands):

6, 8, 6, 9, 11, 5, 60,

**a.** Graph the distribution, representing each output as a dot on the X-axis.

**b.** What is the total output? What is the mean? The median? The mode? Mark these three centers on the graph. In view of the skewness, are they in the order you expect?

**c.** For another firm operating in 10 countries, output (in thousands) had a mean of 7.8 per country, a median of 6.5, and a mode of 5.0. What is the total output?

2-10 To see whether the claims for a new long-life battery were justified, a consumers' group tested a random sample of 20 batteries. Each battery was subjected to a standard heavy load until burnout, providing the following 20 lifetimes (in minutes):

65.1 58.4 64.9 76.0 67.8 75.1 76.7 64.2 74.9 77.6
58.0 68.0 73.3 75.4 76.0 59.4 65.4 74.7 76.6 81.3

Using a cell width of 5 minutes,

**a.** Graph the relative frequency distribution.

**b.** Approximately what are the mean and mode? Mark the mean as the balancing point ▲.

2-11 Sort the data of Problem 2-10 into three cells, whose midpoints are 60, 70, and 80 minutes. Then answer the same questions.

2-12 Summarize the answers to the previous two problems by completing the table below.

| *Grouping* | *Mean* | *Mode* |
|---|---|---|
| Original Data | 70.4 | Not Defined |
| Fine Grouping (Problem 2-10) | | |
| Coarse Grouping (Problem 2-11) | | |

**a.** Why is the mode not a good measure?

**b.** Which gives a closer approximation to the mean of the original data: the coarse or the fine grouping?

2-13 A manufacturer of pocket calculators, bothered by persistently poor quality, tried slowing down the assembly line. The improvement in quality and profitability per calculator seemed to make up for the smaller quantity. In fact, several slower speeds were tried, as follows:

| Speed | Weekly Production | Average Profitability | Median Profitability |
|---|---|---|---|
| Standard | 10,000 | $ .50 | $.50 |
| 20% slower | 8,000 | $ .80 | $.60 |
| 40% slower | 6,000 | $1.00 | $.85 |
| 50% slower | 5,000 | $1.10 | $.90 |

What is the best speed, in order to maximize total profit?

2-14 **a.** Two samples had means of $\bar{X}_1 = 100$ and $\bar{X}_2 = 110$. Then the samples were pooled into one large sample, for which the mean $\bar{X}$ was calculated. What is $\bar{X}$ if the sample sizes are:

**i.** $n_1 = 30, \quad n_2 = 70$?

**ii.** $n_1 = 80, \quad n_2 = 20$?

**iii.** $n_1 = 50, \quad n_2 = 50$?

**iv.** $n_1 = 15, \quad n_2 = 15$?

**b.** Answer true or false; if false, correct it.

We can express the average in general as:

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2} \tag{2-7}$$

2-15 **a.** Calculate the mean of the following 5 numbers:

3, 7, 8, 12, 15

**b.** Calculate the five deviations from the mean, $X - \bar{X}$ (keeping the + or − sign). Then calculate the average of these deviations.

**c.** Write down any set of numbers. Calculate their mean, and then the average deviation from the mean.

**d.** Prove that, for every possible sample of $n$ observations, the average deviation from the mean is exactly zero. Is this also true for deviations from the median?

## 2-3 SPREAD OF A DISTRIBUTION

Although the average may be the most important single statistic, it is also important to know how spread out or varied the observations are. As with measures of center, there are several measures of spread. Two commonly used are the inter-quartile range and the standard deviation. Several others will also be considered because of the way they illuminate these two.

If only $n = 1$ observation were available, this observation would be the sample mean and would give us some idea of the underlying population mean. Since there is no spread in the sample, however, we would have absolutely no idea about the underlying population spread. For example, suppose we observe only one basketball player, and his height is 6′6″. This provides us with an estimate of the *average* height of all players, but no information whatever about how spread out their heights may be (6′4″ to 6′8″? Or 5′ to 8′? From this single observation, we have no clue whatsoever.)

Only to the extent that $n$ exceeds 1 can we get information about the spread. That is, there are essentially only $(n - 1)$ pieces of information for the spread, and this is the appropriate divisor for the variance. Customarily, pieces of information are called *degrees of freedom* (d.f.),[6] and our argument is summarized as:

For the variance, there are $n - 1$ d.f. (degrees of freedom, or pieces of information). (2-16)

## PROBLEMS

2-16 Often the mean is a typical value; but if there is a large standard deviation, it may not be typical at all. In which of the following cases is the mean not typical?

**a.** My wife and I are very athletic. Between us, we jog an average of 5 miles a day. My wife jogs 10.

**b.** In freeway driving, my car averages 32 miles per gallon.

**c.** Last year my car repairs averaged $48 per month.

**d.** The average statistician has 3.46 children.

**e.** The average fuse time for the army's hand grenades is 4.0 seconds.

**f.** Lake Michigan is a bit deep for swimming. Its average depth is 279 feet.

[6] To see where the phrase "degrees of *freedom*" comes from, consider a sample of $n = 2$ observations, 21 and 15, say. Since $\bar{X} = 18$, the residuals (deviations) are $+3$ and $-3$, the second residual necessarily being just the negative of the first. While the first residual is "free," the second is strictly determined. Hence there is only 1 degree of freedom in the residuals.

Generally, for a sample of size $n$, the first $n - 1$ residuals are free. However, the last residual is strictly determined by the requirement that the sum of all residuals be zero—that is, $\Sigma(X - \bar{X}) = 0$, as shown in (2-9).

2-17 Recall that the women's salaries in Problem 2-2 ranked in order were

8, 9, 10, 12, 16

**a.** Find the range and IQR. (*Hint:* Read them off the box plot in Problem 2-3.)

**b.** Calculate the MAD, MSD, variance, and standard deviation.

2-18 Recall that the 25 men's salaries in Problem 2-2 were:

| $x$ | $f$ |
|---|---|
| 10 | 7 |
| 15 | 10 |
| 20 | 5 |
| 25 | 2 |
| 30 | 1 |

**a.** Find the IQR. (*Hint:* Read it off the box plot in Problem 2-3.)

**b.** Calculate the standard deviation s.

2-19 In a test of the reliability of his machine, a technician repeatedly measured the viscosity of a specimen of crude oil. On each of three days, he took 50 measurements:

| Viscosity | Frequency | | |
|---|---|---|---|
| $x$ | Day 1 | Day 2 | Day 3 |
| 60 | 0 | 1 | 0 |
| 65 | 2 | 7 | 5 |
| 70 | 15 | 22 | 38 |
| 75 | 19 | 18 | 6 |
| 80 | 11 | 2 | 1 |
| 85 | 3 | 0 | 0 |
| | 50 | 50 | 50 |

**a.** Graph the 3 sets of data, side by side. Do you discern any trends from day to day?

**b.** For each of the 3 days, calculate the mean and standard deviation. Do these calculations show the same trends you observed in part **a**?

**c.** For the complete set of 150 observations, calculate the mean and standard deviation. How are they related to the means and standard deviations found in part **b**?

2-20 Suspecting that your company's new scales for ready-mix concrete are registering too heavy, you rent a standard ton weight and weigh

This theorem is proved in Appendix 2-5, and may be interpreted very simply: If the *individual* observations are linearly transformed, then the *mean* observation is transformed in exactly the same way, and the *standard deviation* is changed by the factor $|b|$, with no effect from $a$.

## PROBLEMS

2-21 The temperature inside an experimental solar heater was measured in four different spots (degrees Fahrenheit): 238, 227, 220, 235.

**a.** Calculate the mean and standard deviation.

**b.** If the temperatures had been measured in degrees centigrade, what would the mean and standard deviation be? [Hint: If F and C represent a given temperature in Fahrenheit and Centigrade, then $C = (F - 32) \times 5/9$]

2-22 The altitudes of four mountain states of about the same size are approximately as follows (feet above sea level):

| | |
|---|---|
| Arizona | 4100 |
| Nevada | 5500 |
| Colorado | 6900 |
| Wyoming | 6700 |

**a.** To calculate the mean and standard deviation, it would be natural to just drop the last two zeros. Would this give you the right answer? Why or why not?
Then calculate the mean altitude and standard deviation.

**b.** What are the mean and standard deviation in *yards* above sea level?

2-23 An agricultural experimental station has five square plots, whose lengths (in feet) are:

10, 20, 30, 50, 90

**a.** What is the average length?

**b.** What is the total area? The average area?

**c.** Can you calculate the average area by squaring the average length? Why, or why not?

2-24 The following is the grouped frequency table for the actual weight (in ounces) of 50 "6 ounce" bags of cashews that a supermarket clerk filled from bulk stock.

| *Actual Weight* | *Number of Bags* |
|---|---|
| 5.9 | 2 |
| 6.0 | 16 |
| 6.1 | 22 |
| 6.2 | 10 |

## PROBLEMS

2-26 **a.** In Figure 2-18, considering the stock market since 1965, say, let us see how selecting both the before and after years can show almost anything we want.

For example, what period of two years or more would you select to show a gloomy picture? A rosy picture?

**b.** In a sentence or two, give an unbiased summary of the stock market behaviour from 1965 to 1988, somewhere between the two extremes of **a.**

2-27 Criticize this graph (*Washington Post*, Oct. 25, 1978, p. 1, via Wainer, 1984). Then sketch an improved version.

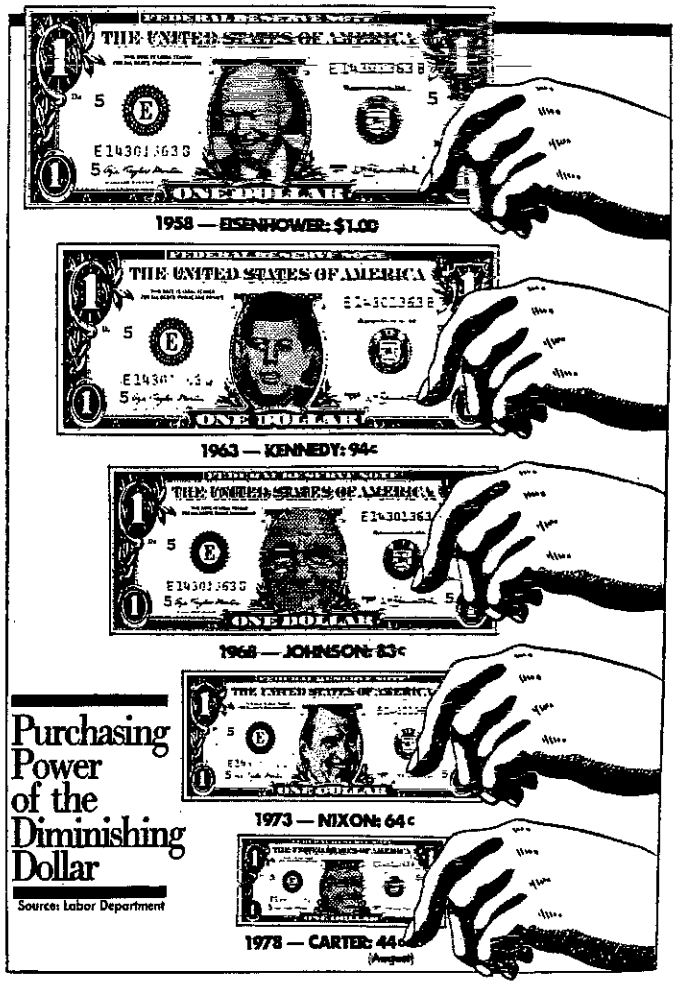
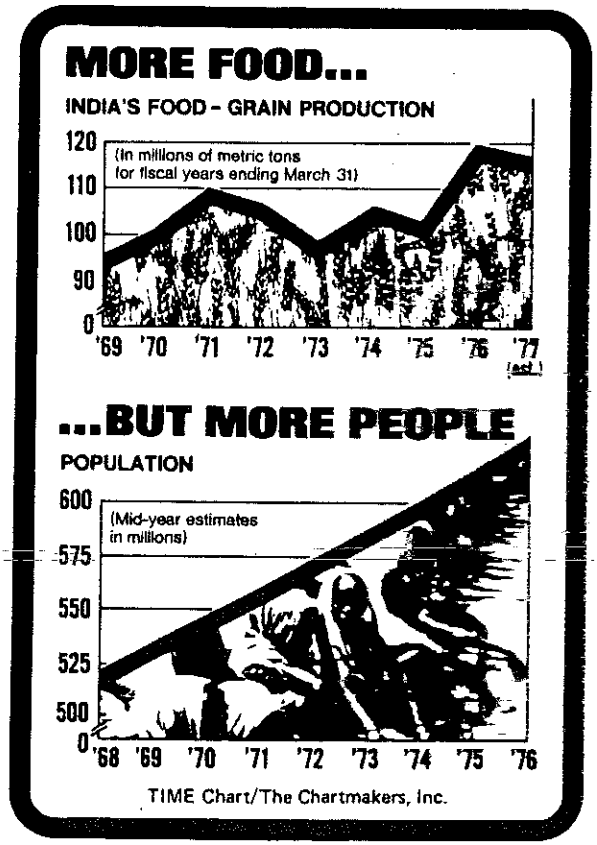



FIGURE 2-19
Napoleon's Russian Campaign (Marey 1885, via Tufte, 1983, p. 41).

2-28




**a.** On the basis of a quick glance, underline the correct choice: India is producing more grain, but its population is growing [slower, faster]. This means that the average Indian is eating [more, less].

**b.** Now take the time to calculate the numbers to confirm your answers to **a**. What do you find?

**c.** Write a few words stating why the graph is appropriate, or else criticize it and redraw it appropriately.

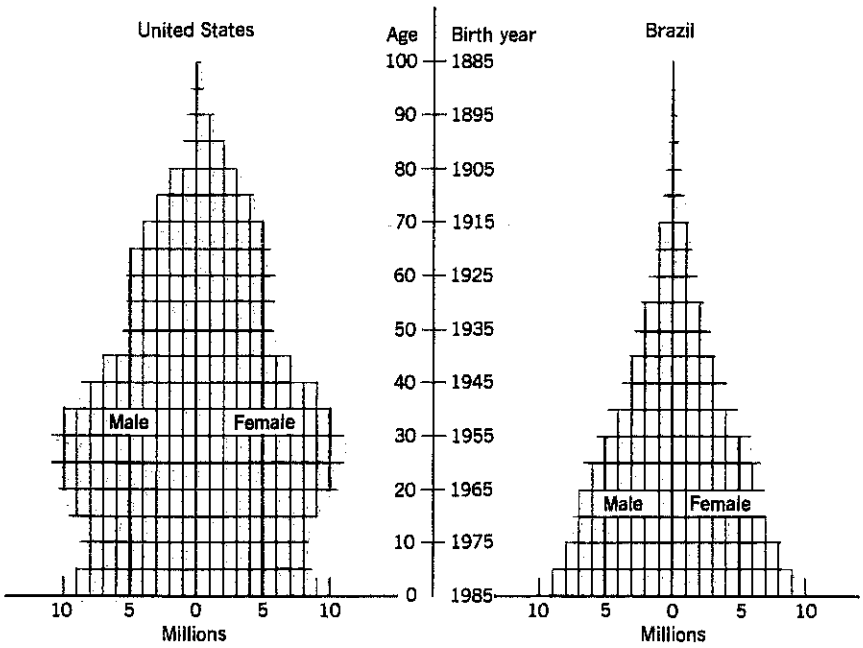
2-29 As well as showing the terrible cruelty of war, the graph of Napoleon's Russian Campaign in Figure 2-19 gives a great deal of historical detail. For example:

**a.** What river crossing cost Napoleon half his men? About how cold was it then? (This was a battle made famous in Tolstoy's *War and Peace*.)

**b.** Of the 10,000 French soldiers who finally made it back to Poland, only a fraction had actually returned from Moscow. Where did the others come from?

c. Of all the soldiers who started out for Moscow, what proportion made it there? That is, what was the chance that a soldier would make it to Moscow? What was the chance that a soldier returning from Moscow would make it back to Poland?

**2-30** As another example of how effective good graphs can be, these two "population pyramids" compare the United States and Brazilian populations in 1985 (Stat. Abst. of U.S., 1987, p. 18). If we examine them carefully, they will give us great insight into the past, and even into the future. First let us begin with the most obvious points:




a. Approximately how many times larger is the U.S. population than Brazil's?

b. Aside from sheer size, what is the main difference between the U.S. and Brazilian population?

c. Both pyramids taper off at the top. Why?

d. The population "pyramid" for the United States, like other developed countries, is rather like a carved table leg. It shows there was a "baby boom" that peaked about ____, producing a large number of people (the "baby boomers") aged ____ in 1985.

It also shows the first "baby bust" that peaked about ____, and a much larger one about ____.

e. The baby bust helps to explain some important economic and social phenomena, such as (underline the correct choice):

Over the next two decades, the average age of Americans will very likely [decrease, increase].

In about 30 or 40 years, there will likely be a large [increase, decrease] in the social security burden to fund retirement, as the relative number of people under 65 paying for it shows a large [increase, decrease].

**2-31** From the U.S. population pyramid in Problem 2-30, we can find some interesting clues about growth rates. For simplicity, we will concentrate on just the female half of the pyramid.

**a.** The average woman has her children around age 25, so this can roughly be taken as the age between generations. Roughly estimate the population size for these female generations:

generation 1, the bar for age 50 to 55
generation 2, the bar for age 25 to 30
generation 3, the bar for age 0 to 5

**b.** By what factor did the female population grow in the past, from generation 1 to 2? Assume immigration was negligible, also mortality until age 50 or 60. Then this factor is roughly the "Net Reproduction Rate," NRR (detailed in Haupt and Kane 1978, for example).

Similarly find the present NRR, from generation 2 to 3.

**c.** If the present NRR continues, project the size in 25 years of the next generation of female children aged 0–5.

Then project another generation later, and another, and another. Graph your answer for all four generations (100 years).

**d.** Do you think **c** is an accurate prediction of what will happen in the next 100 years? Or is it more useful as a policy tool to suggest what will happen *if* present trends continue?

**2-32** For the U.S. population pyramid in Problem 2-30, let us compare the two halves—the males and females.

**a.** What is the ratio of males to females for babies (age 0–5)? For the old (over 75)?

What does this imply about the mortality rate for men compared to women?

**b.** About what age does the sex ratio become 1.00?

## CHAPTER 2 SUMMARY

**2-1** Data can be easily understood by graphing the frequency distribution. Alternatively, the box plot gives less detail but three useful summaries—the median and two quartiles.

**a.** How much does the average share cost John ("straight" average)?

**b.** Susan York buys under a different plan. Every quarter she sets aside $800 to buy as many stocks as she can at the going price. How much does the average share cost Susan ("dollar-cost" average)?

**c.** Which gives the lower average price—straight averaging or dollar-cost averaging?

**d.** Next year, Susan wants to *sell* a little stock every quarter, and wants the price per share to be relatively *high*. Now what should her strategy be—sell off 20 shares per quarter, or $160 worth per quarter?

**2-36** In the summer of 1983, several dozen economic forecasters gave widely varying predictions of the inflation rate for the next year, as follows:

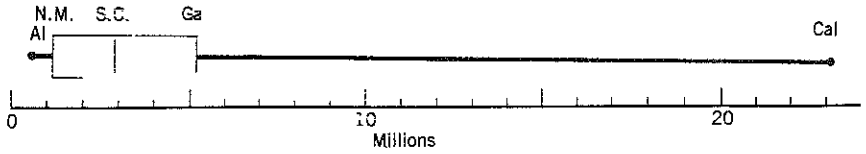
| *Inflation Rate (range of forecast)* | *Proportion of Forecasters* |
|---|---|
| 2–4% | .12 |
| 4–6% | .60 |
| 6–8% | .23 |
| 8–10% | .05 |

*AMSTAT News*, Nov. 1983, p. 6

**a.** What is the average forecast? How does it compare to the inflation rate of 3.9% that actually occurred?

**b.** What is the standard deviation?

**2-37** The 1980 U.S. population of 222 million was widely distributed among the 50 states—from Alaska (0.4 million) to California (23 million) as shown by this box plot:




Roughly estimate:

**a.** The median state population.

**b.** The mean state population.

**c.** The IQR.

**d.** The approximate percentile ranking of Louisiana (population 4.0 million).