Graphical Displays

TOPIC 1

- Dotplots
- Bar Charts
- Histograms
- Cumulative Frequency Plots
- Stemplots
- Center and Spread
- Clusters and Gaps
- Outliers
- Modes
- Shape

There are a variety of ways to organize and arrange data. Much information can be put into tables, but these arrays of bare figures tend to be spiritless and sometimes even forbidding. Some form of graphical display is often best for seeing patterns and shapes and for presenting an immediate impression of everything about the data. Among the most common visual representations of data are dotplots, bar charts, histograms, and stemplots. It is important to remember that all graphical displays should be clearly labeled, leaving no doubt what the picture represents—AP Statistics scoring guides harshly penalize the lack of titles and labels!

TIP

The first thing to do with data is to draw a picture—always.

DOTPLOTS

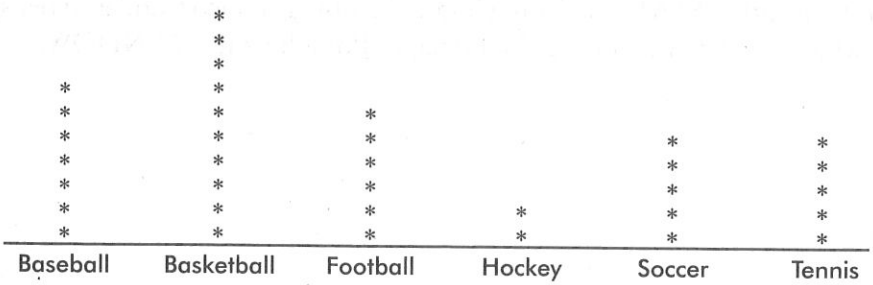
Dotplots and bar charts are particularly useful with regard to *categorical* (or *qualitative*) *variables*, that is, variables that note the category to which each individual belongs. This is in contrast to *quantitative variables*, which take on numerical values.

TIP

Just because a variable has numerical values doesn't necessarily mean that it's quantitative.

EXAMPLE 1.1

Suppose that in a class of 35 students, 10 choose basketball as their favorite sport while 7 pick baseball, 6 pick football, 5 pick tennis, 5 pick soccer, and 2 pick hockey. These data can be displayed in the following *dotplot*:



The *frequency* of each result is indicated by the *number of dots* representing that result.

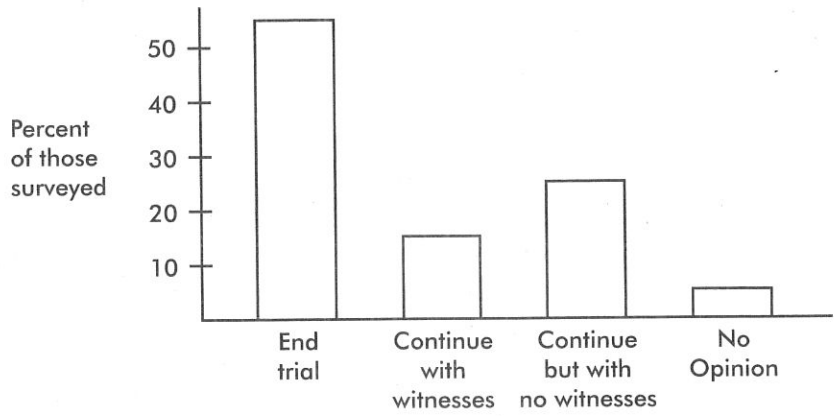
The dotplot can also be drawn with a vertical axis and horizontal rows of dots. In fact, in almost all displays, vertical and horizontal can be switched depending upon which picture seems easier to read or simply which better fits the page.

BAR CHARTS

A common visual display to compare the sizes of categories or groups is the *bar chart.* Sizes can be measured as frequencies or as percents.

EXAMPLE 1.2

In a survey taken during the first week of January 1999, 55% of those surveyed wanted the Clinton impeachment trial to end immediately, 15% wanted it to continue with witnesses, 25% wanted the trial to continue without calling witnesses, and 5% expressed no opinion. These data can be displayed in the following bar chart (or *bar graph*):



The relative frequencies of different results are indicated by the heights of the bars representing these results.

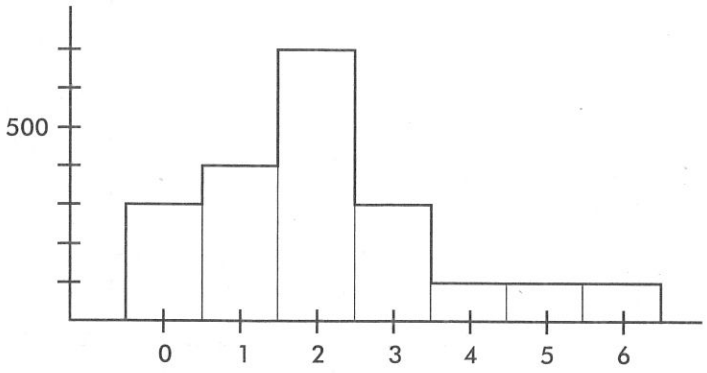
HISTOGRAMS

Histograms, useful for large data sets involving quantitative variables, show counts or percents falling either at certain values or between certain values. While the AP Statistics Exam does not stress construction of histograms, there are often questions on interpreting given histograms.

To construct a histogram using the TI-84, go to STAT → EDIT and put the data in a list, then turn a STATPLOT on, choose the histogram icon under Type, specify the list where the data is, and use ZoomStat and/or adjust the WINDOW.

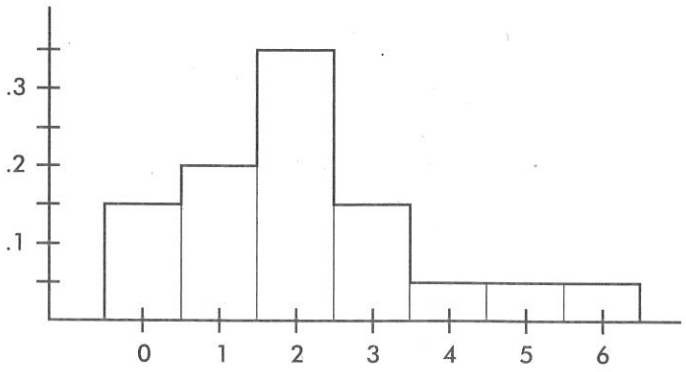
EXAMPLE 1.3

Suppose there are 2000 families in a small town and the distribution of children among them is as follows: 300 families are childless, 400 have one child, 700 have two children, 300 have three, 100 have four, 100 have five, and 100 have six. These data can be displayed in the following histogram:



Sometimes, instead of labeling the vertical axis with frequencies, it is more convenient or more meaningful to use *relative frequencies*, that is, frequencies divided by the total number in the population.

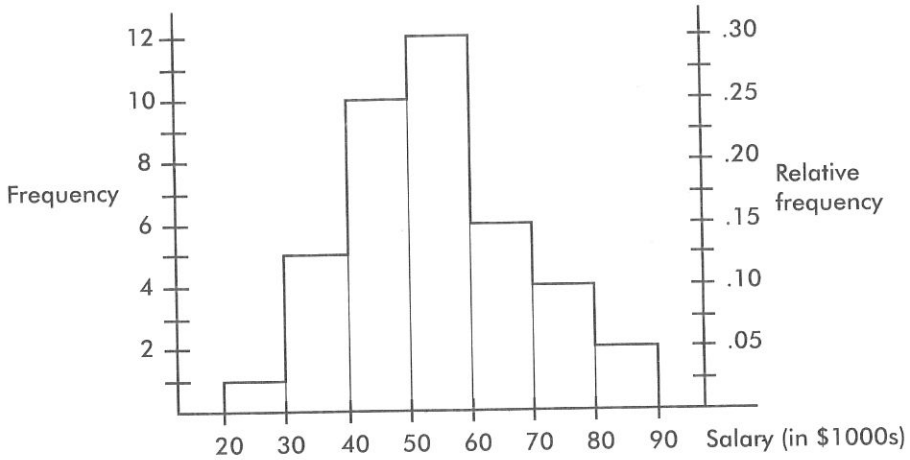
Number of Children	Frequency	Relative Frequency
0	300	300/2000 = .150
1	400	400/2000 = .200
2	700	700/2000 = .350
3	300	300/2000 = .150
4	100	100/2000 = .050
5	100	100/2000 = .050
6	100	100/2000 = .050



Note that the shape of the histogram is the same whether the vertical axis is labeled with frequencies or with relative frequencies. Sometimes we show both frequencies and relative frequencies on the same graph.

EXAMPLE 1.4

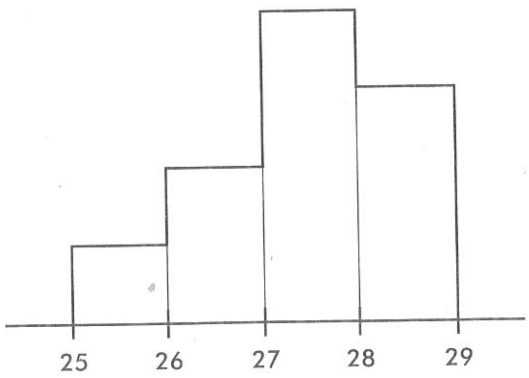
Consider the following histogram displaying 40 salaries paid to the top-level executives of a large company.



What can we learn from this histogram? For example, none of the executives earned more than $90,000 or less than $20,000. Twelve earned between $50,000 and $60,000. Twenty-five percent earned between $40,000 and $50,000. Note how this histogram shows the number of items (salaries) falling *between* certain values, whereas the preceding histogram showed the number of items (families) falling *at* each value. For example, in Example 1.4 we see that ten salaries fell somewhere between $40,000 and $50,000, while in Example 1.3 we see that 700 families had exactly two children.

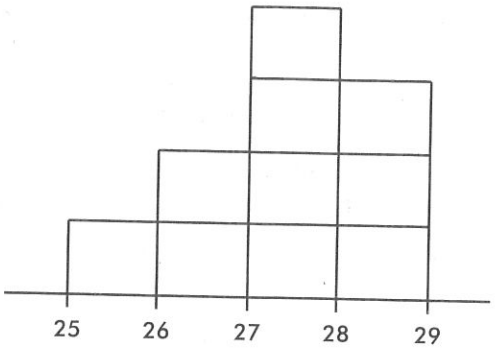
EXAMPLE 1.5

Consider the following histogram where the vertical axis has not been labeled. What can we learn from this histogram?



(continued)

Answer: It is impossible to determine the actual frequencies; however, we can determine the relative frequencies by noting the fraction of the total *area* that is over any interval:



We can divide the area into ten equal portions, and then note that $\frac{1}{10}$ or 10% of the area is above 25–26, 20% is above 26–27, 40% is above 27–28, and 30% is above 28–29.

Although it is usually not possible to divide histograms so nicely into ten equal areas, the principle of relative frequencies corresponding to relative areas still applies.

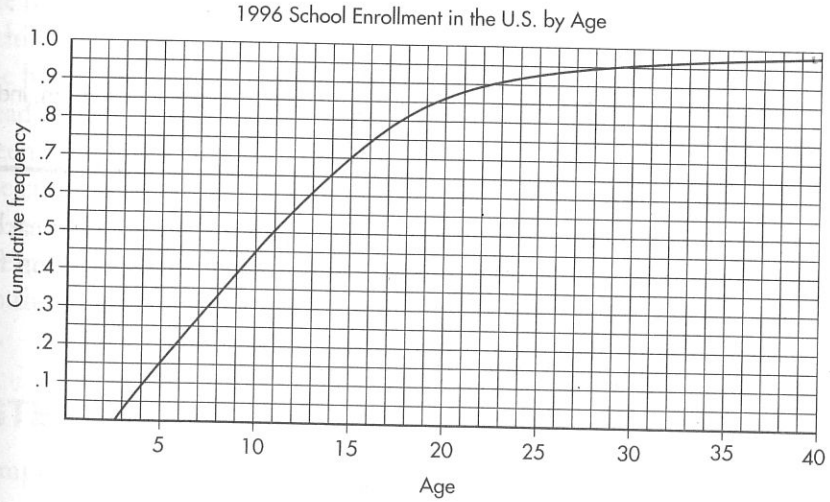
Relative frequencies are the usual choice when comparing distributions of different size populations.

CUMULATIVE FREQUENCY PLOTS

Sometimes we sum frequencies and show the result visually in a *cumulative frequency plot* (also known as an *ogive*).

EXAMPLE 1.6

The following graph shows 1996 school enrollment in the United States by age.



What can we learn from this cumulative frequency plot? For example, going up to the graph from age 5, we see that .15 or 15% of school enrollment is below age 5. Going over to the graph from .5 on the vertical axis, we see that 50% of the school enrollment is below and 50% is above a middle age of 11. Going up from age 30, we see that .95 or 95% of the enrollment is below age 30, and thus 5% is above age 30. Going over from .25 and .75 on the ver-

(continued)

tical axis, we see that the middle 50% of school enrollment is between ages 6 and 7 at the lower end and age 16 at the upper end.

STEMPLOTS

Although a histogram may show how many scores fall into each grouping or interval, the exact values of individual scores are often lost. An alternative pictorial display, called a stemplot, retains this individual information.

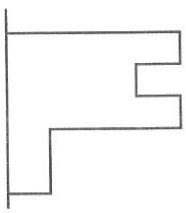
EXAMPLE 1.7

Consider the set {17, 17, 18, 13, 28, 38, 31, 27, 35, 50, 43, 37, 24} of percentages of three-point shots made by Michael Jordan during his 13 years with the Bulls. Let 1, 2, 3, 4, and 5 be placeholders for 10, 20, 30, 40, and 50. List the last digit of each value from the original set after the appropriate placeholder.

The result is a *stemplot* (also called a *stem and leaf display*) of these data:

Stems	Leaves
1	7 7 8 3
2	8 7 4
3	8 1 5 7
4	3
5	0

Drawing a continuous line around the leaves would result in a horizontal histogram:



Note that the stemplot gives the shape of the histogram and, unlike the histogram, indicates the values of the original data.

Sometimes further structure is shown by rearranging the numbers in each row in ascending order. This ordered display shows a second level of information from the original stemplot.

The revised display of the data in Example 1.7 is as follows:

1	3 7 7 8
2	4 7 8
3	1 5 7 8
4	3
5	0

EXAMPLE 1.8

Using a "torsion balance," Henry Cavendish (in 1798) made 29 measurements of Earth's density, obtaining values of 5.5, 5.57, 5.42, 5.61, 5.53, 5.47, 4.88, 5.62, 5.63, 4.07, 5.29, 5.34, 5.26, 5.44, 5.46, 5.55, 5.34, 5.3, 5.36, 5.79, 5.75, 5.29, 5.1, 5.86, 5.58, 5.27, 5.85, 5.65, and 5.39 gm/cm^3.

To the left is a stemplot of this data.

Note that the scale is such that one must multiply each value in the dataset by 0.01 to return the original value. For example, $407 \times 0.01 = 4.07$.

```
40 | 7
41 |
42 |
43 |
44 |
45 |
46 |
47 |
48 | 8
49 |
50 |
51 | 0
52 | 6799
53 | 04469
54 | 2467
55 | 03578
56 | 1235
57 | 59
58 | 56
```

(40 | 7 means 4.07 gm/cm^3)

CENTER AND SPREAD

TIP

Center and spread should always be described together.

Looking at a graphical display, we see that two important aspects of the overall pattern are

1. the *center*, which separates the values (or area under the curve in the case of a histogram) roughly in half, and
2. the *spread*, that is, the scope of the values from smallest to largest.

In the histogram of Example 1.3, the center is 2 children while the spread is from 0 to 6 children.

In the histogram of Example 1.4 the center is between \$50,000 and \$60,000, and the spread is from \$20,000 to \$90,000; in the histogram of Example 1.5, the center is between 27 and 28, and the spread is from 25 to 29.

In the cumulative frequency plot of Example 1.6, the center is between 10 and 11, and the spread is from 3 to 40.

In the stemplot of Example 1.7, the center is 28%, and the spread is from 13% to 50%; in the stemplot of Example 1.8, the center is 5.46, and the spread is from 4.07 to 5.86.

CLUSTERS AND GAPS

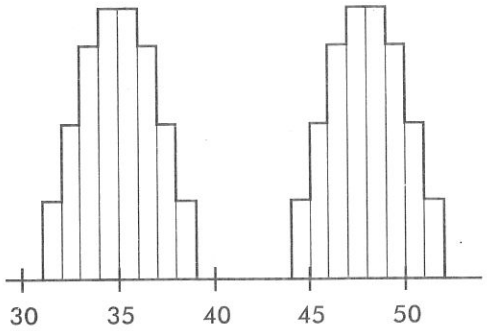
Other important aspects of the overall pattern are

1. *clusters*, which show natural subgroups into which the values fall (for example, the salaries of teachers in Ithaca, NY, fall into three overlapping clusters, one for public school teachers, a higher one for Ithaca College professors, and an even higher one for Cornell University professors), and
2. *gaps*, which show holes where no values fall (for example, the Office of the Dean sends letters to students being put on the honor roll and to those

being put on academic warning for low grades; thus the GPA distribution of students receiving letters from the Dean has a huge middle gap).

EXAMPLE 1.9

Consider the following histogram:

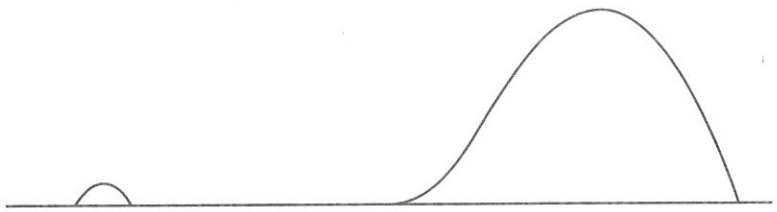


Simply saying that the center of the distribution is around 42 and the spread is from 31 to 52 clearly misses something. The values fall into two distinct clusters with a gap between.

OUTLIERS

TIP

Pay attention to outliers!



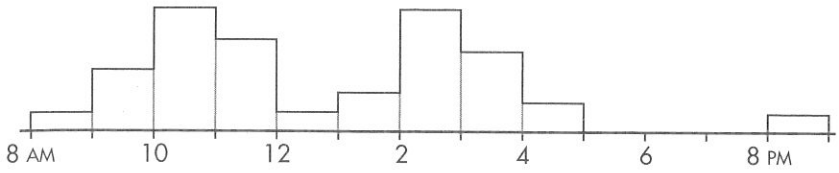
Extreme values, called *outliers*, are found in many distributions. Sometimes they are the result of errors in measurements and deserve scrutiny; however, outliers can also be the result of natural chance variation. Outliers may occur on one side or both sides of a distribution. In the stemplot of Example 1.8, 4.07 is clearly an outlier.

MODES

Some distributions have one or more major peaks, called *modes*. With exactly one or two such peaks, the distribution is said to be *unimodal* or *bimodal*, respectively. But every little bump in the data is not a mode! You should always look at the big picture and decide whether or not two (or more) phenomena are affecting the histogram.

EXAMPLE 1.10

The histogram below shows employee computer usage (number accessing the Internet) at given times at a company main office.



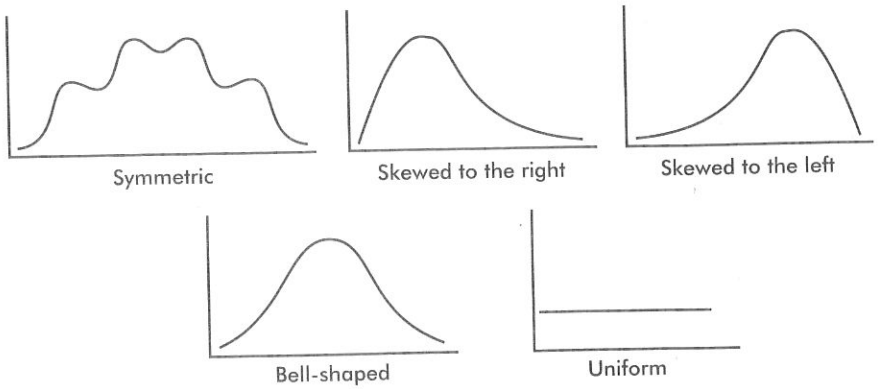
Note that this is a *bimodal* distribution. Computer usage at this company appears heaviest at midmorning and midafternoon, with a dip in usage during the noon lunch hour. There is an evening outlier possibly indicating employees returning after dinner (or perhaps custodial cleanup crews taking an Internet break!).

Note that, as illustrated above, it is usually instructive to look for reasons behind outliers and modes.

SHAPE

Distributions come in an endless variety of shapes; however, certain common patterns are worth special mention:

1. A *symmetric* distribution is one in which the two halves are mirror images of each other. For example, the weights of all people in some organizations fall into symmetric distributions with two mirror-image bumps, one for men's weights and one for women's weights.
2. A distribution is *skewed to the right* if it spreads far and thinly toward the higher values. For example, ages of nonagenarians (people in their 90s) is a distribution with sharply decreasing numbers as one moves from 90-year-olds to 99-year-olds.
3. A distribution is *skewed to the left* if it spreads far and thinly toward the lower values. For example, scores on an easy exam show a distribution bunched at the higher end with few low values.
4. A *bell-shaped* distribution is symmetric with a center mound and two sloping tails. For example, the distribution of IQ scores across the general population is roughly symmetric with a center mound at 100 and two sloping tails.
5. A distribution is *uniform* if its histogram is a horizontal line. For example, tossing a fair die and noting how many spots (pips) appear on top yields a uniform distribution with 1 through 6 all equally likely.



Even when a basic shape is noted, it is important also to note if some of the data deviate from this shape. For example, in the stemplot of Example 1.8, there is an outlier at 4.07 and a value at 4.88, with the remaining values showing a roughly bell-shaped distribution.

Summary

- The three keys to describing a distribution are shape, center, and spread.
- Also consider clusters, gaps, modes, and outliers.
- Look for reasons behind any unusual features.
- A few common shapes arise from symmetric, skewed to the right, skewed to the left, bell-shaped, and uniform distributions.
- For categorical (qualitative) data, dotplots and bar charts give useful displays.
- For quantitative data, histograms, cumulative frequency plots (ogives), and stemplots give useful displays.
- In a histogram, relative area corresponds to relative frequency.