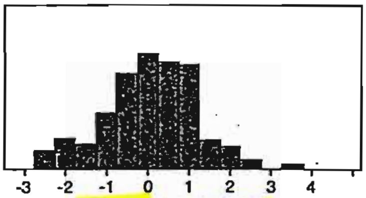
Chapter 3

Constructing and Interpreting Graphical Displays of Distributions of Univariate Data

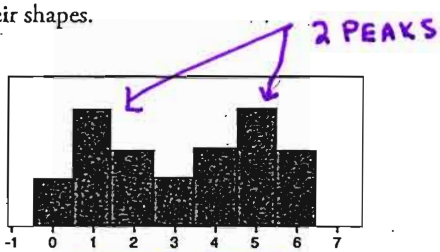
Symmetric graphs appear to have mirror images about their center. If a graph has only one clear peak, it is called **unimodal**; if it has two, it is **bimodal**. Symmetric, unimodal graphs may sometimes be referred to as **mound-shaped**, or **bell-shaped**, because they look like a mound or bell.

A **uniform** distribution is symmetric where the data are distributed fairly evenly across the graph. There are no clear peaks and the data do not seem to cluster in one area or another.

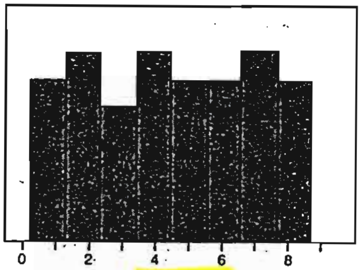
Below are some examples of graphs and their shapes.



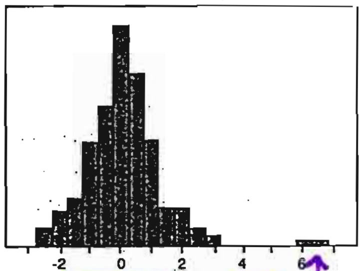
Symmetric and unimodal



Symmetric and bimodal

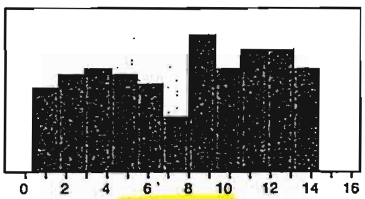


Uniform

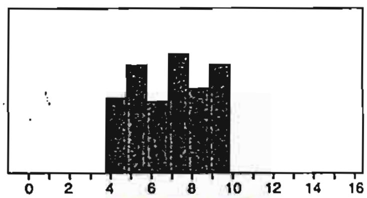


Mound-shaped with outliers

The spread describes the **variability of the distribution of the data.**



Wider Spread



Narrower Spread

Skewed graphs are unimodal graphs that tend to slant—most of the data are clustered on one side of the distribution and "tails" off on the other side. If the tail is on the left, we call the distribution left-skewed. If the tail is on the right, it is right-skewed.

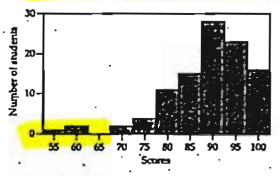
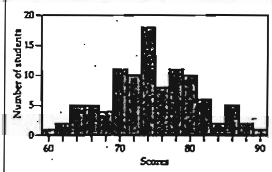
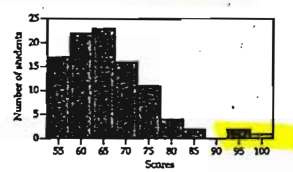
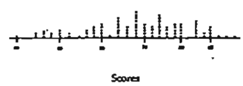
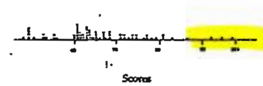
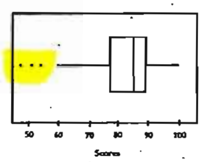
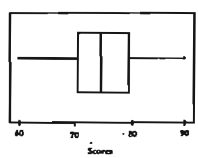
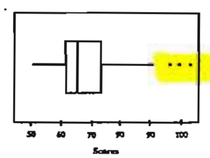
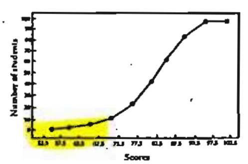
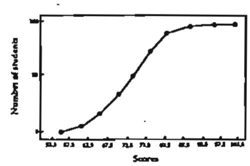
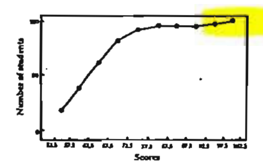
Graph	LEFT-SKEWED OR NEGATIVE SKEW	SYMMETRIC	RIGHT-SKEWED OR POSITIVE SKEW
Histogram			
Dotplot			
Boxplot			
Cumulative Frequency Chart			
Stemplot	4 8 5 2 5 5 6 14 6 889 7 0112233444 7 56667778899 8 00111111222222233334444444 8 5556666667777777777888889999999 9 222233444 9 5555567	6 0 6 22333 6 44555 6 667 6 88899999 7 0000001 7 22222222222333333 7 44444444444455555 7 66677777777 7 88899999 8 00000 8 222222 8 445555 8 67 8 89	5 2333344 5 55556667799 6 0000000001111111111222222233333444 6 55555566666666677777888899 7 0011122224444 8 003 8 7 9 2 9 59

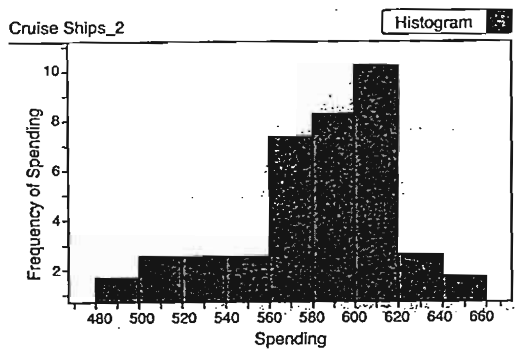
Table 4: Comparison of shapes of different graphs

TEST TIP

Beginning in May 2011, the AP Exam stopped penalizing test takers for incorrect responses to multiple-choice questions. Entering a response for every question—even a wild guess—may help improve your score.

A histogram is an appropriate display for quantitative data. It is used primarily for continuous data, but may be used for discrete data that have a wide spread. The horizontal axis is broken into intervals or **bins**. Histograms are also good for large data sets. The histogram below shows the amount of money spent by passengers on a board ship during a recent cruise to Alaska.

EXAMPLE: Describe the distribution below of the amount spent by passengers on board a ship during a recent cruise to Alaska.



[1] ANSWER: DISTRIBUTION HAS A SHAPE SKEWED TO THE LEFT. IT PEAKS BETWEEN \$600 TO \$620. THE DATA IS CENTERED AROUND \$590 (MEDIAN). UNIMODAL WITH NO OBVIOUS OUTLIERS.

A stemplot, also called stem-and-leaf plot, can be used to display univariate data as well. It is good for small sets of data (about 50 or less) and forms a plot much like a histogram. The stemplot below represents test scores for a class of 32 students.

EXAMPLE: Describe the distribution of test scores for students in the class using the stemplot below.

3	3
4	
5	
6	3 7 9
7	2 2 5 7
8	1 2 6 8 8 8 9 9 9
9	0 0 0 1 3 3 4 5 5 6 7
10	0 0 0 0

Key: 6 | 3 represents a score of 63

[2] ANSWER: THE DISTRIBUTION IS UNIMODAL AND SKEWED TO THE LEFT (OR NEGATIVE). 33% APPEARS TO BE AN OUTLIER. THE DATA IS CENTERED AROUND 89%.

13] DO EXAMPLE AND VERIFY YOUR GRAPH WITH ONE GIVEN

Calculator Tip: INPUT DATA

Many graphs and functions for univariate data can be done on the graphing calculator, but the observations must be first entered into a list. We will use the test scores of the 32 students from a previous example: 33, 63, 67, 69, 72, 72, 75, 77, 81, 82, 86, 88, 88, 88, 89, 89, 89, 90, 90, 90, 91, 93, 93, 94, 95, 95, 96, 97, 100, 100, 100, 100.

From the home screen, press STAT.

```
EDIT CALC TESTS
1:Edit...
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
```

Choose 1:Edit. This brings you into the main list editor.

```
L1      |L2      |L3     1
------- |------- |-------



L1(1)=
```

Enter the data into the list that you are going to use. For now, use L1. Press ENTER after each entry to move to the next line.

```
L1      |L2      |L3     1
 33     |------- |-------
 63     |        |
 67     |        |
 69     |        |
 72     |        |
 72     |        |
 ------ |        |
L1(7)=
```

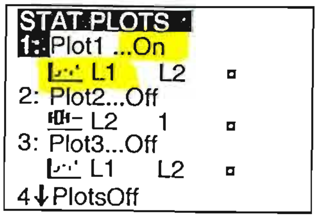
Continue to type in values until the list is complete. Note that the cursor is on the last entry, which in this case is 100, that is the 32nd entry into this list.

```
L1      |L2      |L3     1
 96     |        |
 97     |        |
 100    |        |
 100    |        |
 100    |        |
 100    |        |
L1(32)=100
```

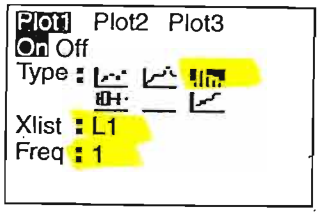
Calculator Tip: CREATE STAT PLOT

3 CONT

Histograms can be made from data in lists. From the home screen, press [2nd] [STAT PLOT], and then choose 1:Plot1....



Turn on the plot by pressing [ENTER] when the cursor is over On. Arrow right to the third type of plot, which is a histogram. Press [ENTER] so that it is highlighted. Arrow down to Xlist: and press [2nd] [L1]. On the Freq: line, enter 1, if it is not already so. (This is the frequency of each value in Xlist and is used if two lists act as a frequency table.)



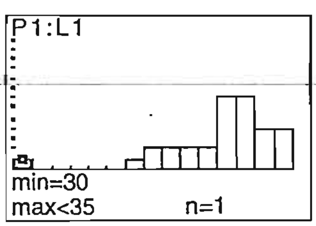
Press [ZOOM] [9] to see the graph. The calculator will sometimes choose bin widths that are hard to use.

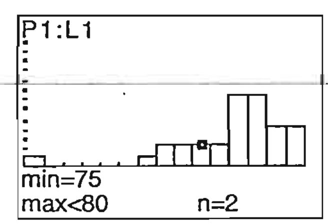
Press [WINDOW]. A bin width of 5 or 10 may make more sense in this situation than the one of 11.16. Change Xmin, Xmax, Xscl, Ymin, Ymax, and Yscl to the values shown below on the right. Xscl is the width of your bins, beginning at Xmin. An appropriate window for these data is shown on the screen below.

```
WINDOW
 Xmin=33
 Xmax=111.16666...
 Xscl=11.166666...
 Ymin=-4.20966
 Ymax=16.38
 Yscl=.1
 Xres=1
```

```
WINDOW
 Xmin=30
 Xmax=105
 Xscl=5
 Ymin=-4
 Ymax=15
 Yscl=1
 Xres=1
```

Press [GRAPH] to view the histogram.



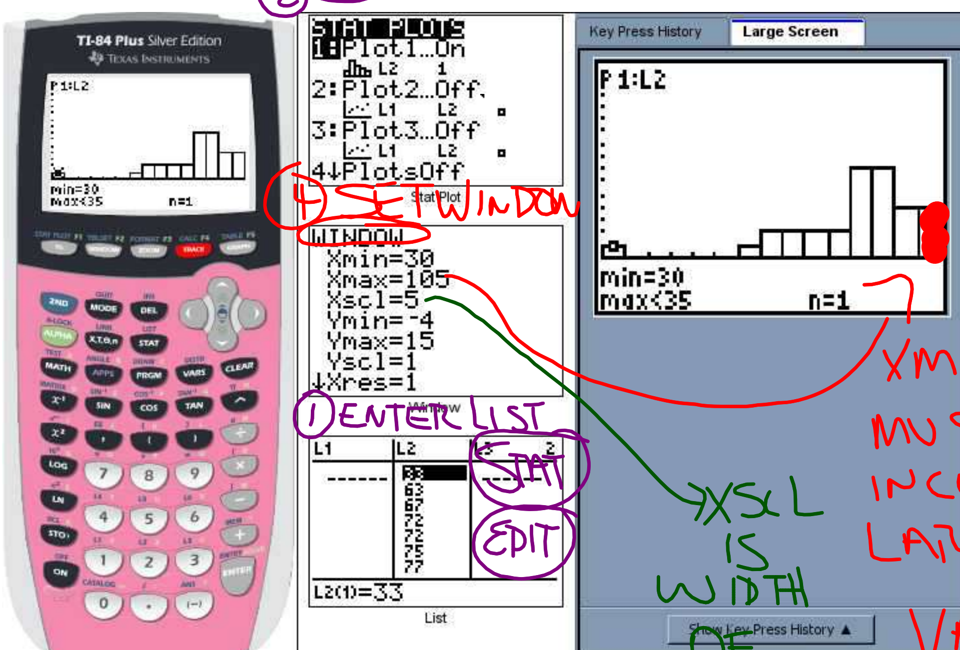


You can use the [TRACE] button and arrows to view the frequencies of each bin.

More notes on prior page 5.

CREATE HISTOGRAM

(2) 2ND STATPLOT



(4) SET WINDOW

(1) ENTER LIST — STAT, EDIT

XSCL IS WIDTH OF BAR

XMAX MUST INCLUDE LARGEST VALUE 100 - 105

(3) ZOOM 9 (STATPLOT) FOR SKETCH OF HISTOGRAM

SECTION 3.4:

Exploring Bivariate Data

Analyzing Patterns in Scatterplots

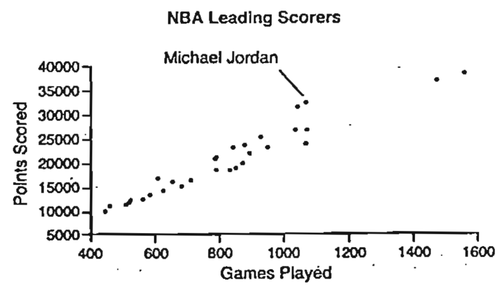
Bivariate data consist of two variables, between which one is typically looking for an association. The variables may be categorical or quantitative; in this section we will focus on quantitative bivariate data.

The two variables under study are referred to as the **explanatory variable** (x) and the **response variable** (y). The explanatory variable *explains* or *predicts* the response variable. The response variable measures the outcomes that have been observed.

EXAMPLE: Data collected from snack foods included the number of grams of fat per serving and the total number of calories in the food. Identify the explanatory and response variables when looking for a relationship between fat grams and calories.

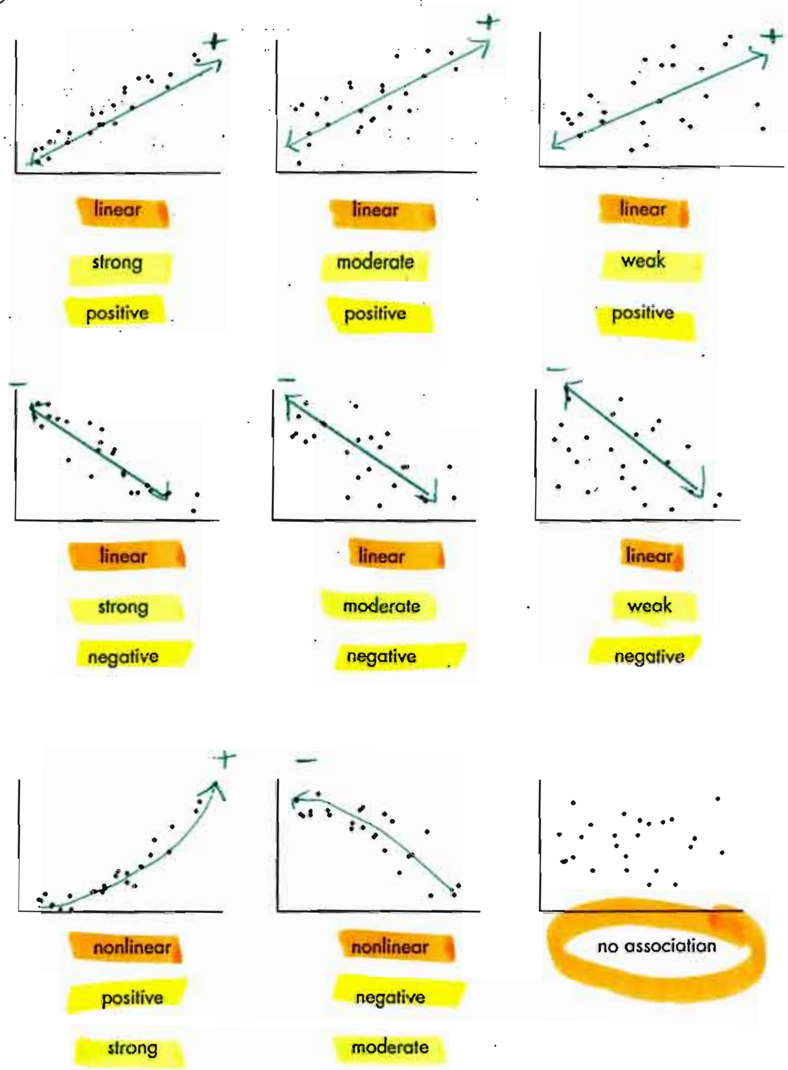
14 ANSWER: EXPLANATORY VARIABLE: GRAMS OF FAT
RESPONSE VARIABLE: CALORIES
* NUMBER OF GRAMS OF FAT WOULD PREDICT CALORIES IN THE SNACK.

Scatterplots are used to visualize quantitative bivariate data. These plots can tell us if and how two variables are related. When examining univariate data we described a distribution's shape, center, spread, and outliers/unusual features. In a scatterplot, we will focus on its shape, direction, and strength, and look for outliers and unusual features. Below is a scatterplot of the top 30 leading scorers in the National Basketball Association (NBA). Each point represents 1 of the 30 players. Michael Jordan, who scored 32,292 points in 1,072 games, is noted.

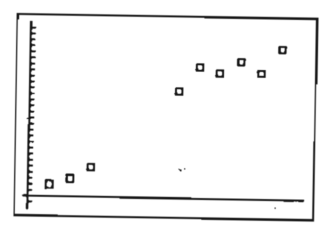


The **shape** of a plot is usually classified as linear or nonlinear (curved). The **direction** of a scatterplot tells what happens to the response variable as the explanatory variable increases. This is the slope of the general pattern of the data. The **strength** describes how tight or spread out the points of a scatterplot are.

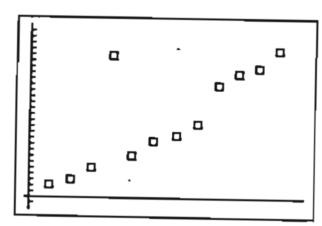
The table below shows comparisons of scatterplots of various shapes, directions, and strengths.



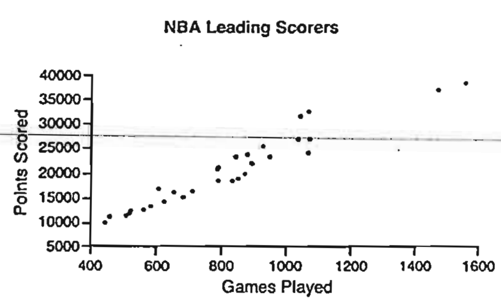
When analyzing a scatterplot it is also a good idea to look for outliers, clusters, or gaps in the data. The scatterplot below has an obvious gap. There is an overall positive, linear association but we should find out the reason for the gap.



The scatterplot below has an obvious outlier. An outlier falls outside the general pattern of the data. There could be several possible reasons for the outlier and it merits investigation.



15 EXAMPLE: A scatterplot of the top 30 scorers in NBA history is shown below. Identify the explanatory variable and the response variable. Describe the association between the two variables.



ANSWER:

6 TRY THIS SCATTER PLOT !

Calculator Tip: INPUT BIVARIATE DATA

A scatterplot can be viewed on the graphing calculator. First, the data must be entered into lists. Recall that you access the list editor by pressing [STAT] and then choosing 1:Edit....

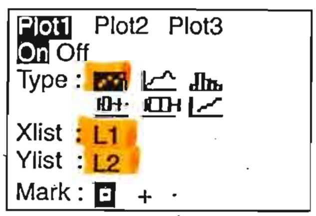
L1	L2	L3 2
2	1	-------
3	3	
5	7	
6	7	
9	11	
10	12	

L2(7) =

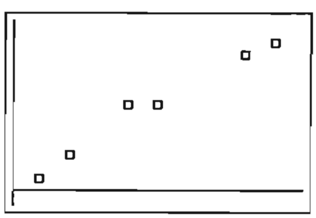
← INPUT DATA

CREATE SCATTER PLOT :

Press [2nd] [STAT PLOT] and choose 1:Plot 1. Turn on the plot, select the scatterplot icon, and enter the appropriate lists for Xlist: and Ylist:.



Press [ZOOM] [9] to see the scatterplot.



← OUTPUT GRAPH

MORE PRACTICE EXAMPLES TO USE WITH TI84

3.4 Displaying Bivariate Numerical Data

A bivariate data set consists of measurements or observations on two variables, x and y. For example, x might be distance from a highway and y the lead content of soil at that distance. When both x and y are numerical variables, each observation consists of a pair of numbers, such as (14, 5.2) or (27.63, 18.9). The first number in a pair is the value of x, and the second number is the value of y.

An unorganized list of bivariate data yields little information about the distribution of either the x values or the y values separately and even less information about how the two variables are related to one another. Just as graphical displays can be used to summarize univariate data, they can also help with bivariate data. The most important graph based on bivariate numerical data is a **scatterplot**.

Example 3.20 Taking Those "Hard" Classes Pays Off

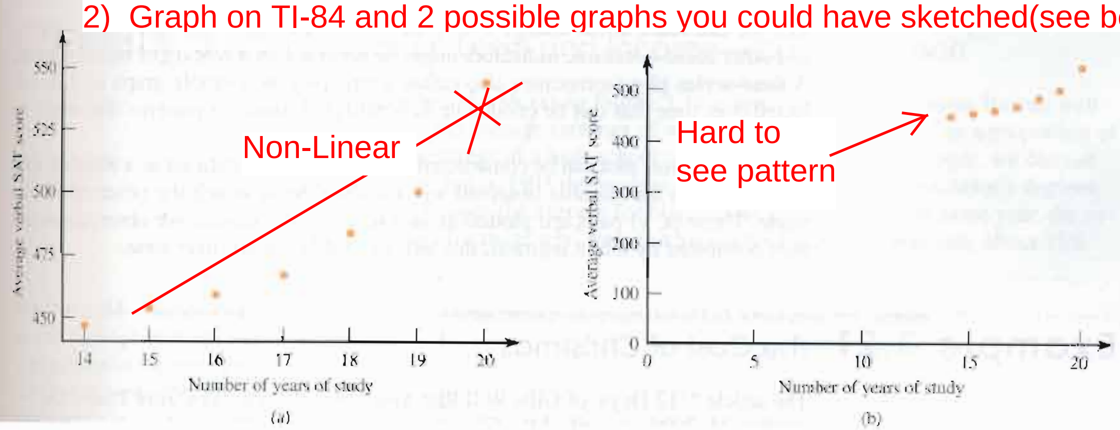
The report titled "2005 College Bound Seniors" (College Board, 2005) included the accompanying table showing the average score on the verbal section of the SAT for groups of high school seniors completing different numbers of years of study in

six core academic subjects (arts and music, English, foreign languages, mathematics, natural sciences, and social sciences and history).

Figure 3.32(a), we let MINITAB select the scale for both axes. Figure 3.32(b) was obtained by specifying that the axes would intersect at the point (0, 0). The second plot does not make effective use of space. It is more crowded than the first plot, and such crowding can make it more difficult to see the general nature of any relationship. For example, it can be more difficult to spot curvature in a crowded plot.

Years of Study	Average Verbal SAT Score
14	446
15	453
16	459
17	467
18	484
19	501
20	545

1) Use TI-84 and enter the above data into 2 lists
2) Graph on TI-84 and 2 possible graphs you could have sketched(see below)



▪ Time-Series Plots

Data sets often consist of measurements collected over time at regular intervals so that we can learn about change over time. For example, stock prices, sales figures, and other socio-economic indicators might be recorded on a weekly or monthly basis. A **time-series plot** (sometimes also called a time plot) is a simple graph of data collected over time that can be invaluable in identifying trends or patterns that might be of interest. A time-series plot can be constructed by thinking of the data set as a bivariate data set, where *y* is the variable observed and *x* is the time at which the observation was made. These (*x*, *y*) pairs are plotted as in a scatterplot. Consecutive observations are then connected by a line segment; this aids in spotting trends over time.

▪ Example 3.23 Life Expectancy over Time

The article "Americans Living Longer Than Ever" (*San Luis Obispo Tribune*, September 13, 2002) included a time-series plot that showed how life expectancy at birth for people living in the United States has changed over time. The plot was based on the following data from "The Vital Statistics Report," published by the Center for Disease Control:

Year	Life Expectancy at Birth (years)
1940	62.9
1950	68.2
1960	69.7
1970	70.8
1980	73.7
1990	75.4
2000	76.9

Use TI-84 to Sketch Time Series
* Draw line for time series

A time-series plot of these data is shown in Figure 3.34. From this plot, the upward trend is clear, providing justification for the article headline.

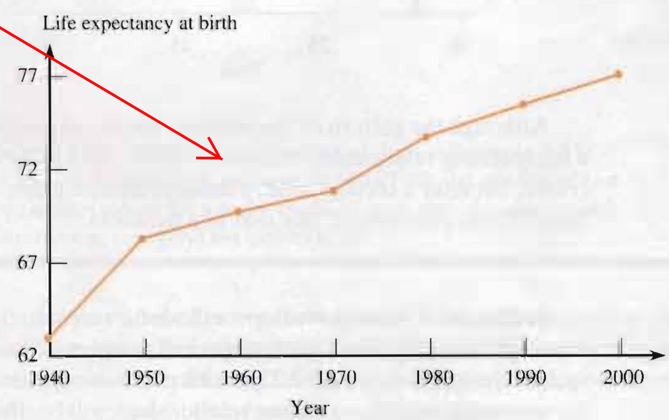


Figure 3.34 Time-series plot for the life expectancy at birth data of Example 3.23.

3.41 ● The National Telecommunications and Information Administration published a report titled "Falling Through the Net: Toward Digital Inclusion" (U.S. Department of Commerce, October 2000) that included the following information on access to computers in the home:

Year	Percentage of Households with a Computer
1985	8.2
1990	15.0
1994	22.8
1995	24.1
1998	36.6
1999	42.1
2000	51.0

a. Construct a time-series plot for these data. Be careful—the observations are not equally spaced in time. The points in the plot should not be equally spaced along the x axis.

b. Comment on any trend over time.

Use TI-84 to Sketch Time Series - See next slide for solution

3.41 Solution

* a) Below is the graph from the TI84

* b) Comments: At first, starting in the late 1980's, the increase in computer ownership was low. Then in 1995, the increase has been more rapid.
This is an example of a non-linear relationship.