1981

Basic statistics : a user oriented approach (manuscript)

Makridakis, Spyros

BASIC STATISTICS:

CHAPTER 2

DESCRIPTIVE STATISTICS

2.1 Introduction

Table 2.1 shows the ages of the 230 students who entered INSEAD during the 1980/81 academic year. What can you make out of these numbers? Obviously, the answer will depend upon who you are. Unless you have some objective in mind, you will not bother further examining Table 2.1. But let us assume you are the academic director of INSEAD. Then you may want to know the average age of the students and many other characteristics that the data might reveal. Is the student population of this year older or younger than that of previous years? What is the most common age among students? Are there any students much younger or much older than the average? If there are, it may be that they will not fit in well with the great majority of students. Are the various age groups distributed evenly, or are some ages predominant in the 230 students?

Table 2.1 : Ages of 230 INSEAD Students.

27	25	25	32	30	29	26	28	28	26	21	31	35	25	27	32	28	24	29	31
32	29	27	27	24	24	27	27	26	28	26	26	28	22	29	29	26	28	29	24
28	25	28	33	28	28	29	30	27	26	24	30	25	35	28	26	28	31	30	36
26	29	26	20	31	28	29	28	24	28	26	30	29	29	30	28	31	23	28	30
28	29	27	33	27	23	25	25	30	32	26	30	31	27	24	27	28	26	27	28
25	31	32	22	31	29	26	26	26	27	29	29	28	30	26	30	33	27	28	27
29	24	26	27	28	26	29	24	29	27	29	33	28	28	30	30	27	26	28	29
31	26	26	30	25	31	28	27	25	23	27	31	27	27	26	29	27	28	30	25
26	23	26	28	25	28	26	31	24	27	25	25	27	28	26	26	27	30	26	30
31	32	26	31	25	28	27	28	36	27	28	23	29	27	29	25	30	30	22	26
27	27	30	31	27	26	23	28	28	28	23	29	32	26	25	27	27	29	29	25
27	38	34	34	29	29	25	37	34	29										

Table 2.1 is a rather small collection of numbers. The population of the United States, for instance, is more than 230 million people. Assuming that 920 ages (46 lines of 20 ages each) are printed on a page, it would take 250,000 pages to just list the ages of all people in the United States. Thus, even if you look at each page for a mere 10 seconds, it will take you about 700 hours just to look at all of these pages. These 700 hours mean 70 days (assuming you can work ten hours per day), or 14 uninterrupted five-day weeks. This is a lot of wasted time, since there

is not a lot to be gained from just looking at page after page of numbers for a mere ten seconds per page.

It is the purpose of descriptive statistics to summarize, in a meaningful way, large sets of data. This chapter will explain to you how this can be done and the benefits to be gained by doing so.

Descriptive statistics has been becoming increasingly important as huge volumes of data are being made available - mainly a consequence of today's computers which can store increasingly larger amounts of data at rapidly decreasing cost. Furthermore, existing trends will continue, meaning that even larger quantities of data will be available in the future. This is not a sales pitch from a statistics book but a reality that has been emerging in the last ten years.

You may ask why information about data - such as the 230 million ages of the U.S. population - might be useful to anyone in general, or to you.

Those who work, for instance, pay a portion of their earnings to social security. This is done so that when retired, they can obtain a pension. The social security system - as well as many other pension funds - does not keep the money that one pays now to reimburse him or her with a pension after retirement. Instead, all the money that is being collected now is being used to pay pensions of people who worked and contributed to social security many years ago. This is fine as long as the number of people working and the number of people receiving pensions do not change too much. Serious problems exist, however, if there are big changes as has been the case in the last several years.

There was a big baby boom following World War II. This meant a large labor force and enough receipts by the social security system to pay those at pension age in the 1960's and 1970's. But what is happening as the population is aging? The implication of the baby boom and subsequent reductions in the birth rate is that the ratio of working to retired people will be decreasing. Therefore, not enough revenues will be received by the social security system from those working to pay those who have retired - at least at non-diminishing real levels of pensions.

Descriptive statistics provides the means of analyzing and presenting data in a form that decision or policy makers can use in making better decisions. For instance, if we look at how the ages of the U.S. population change with time, it becomes obvious that the social security system will go bankrupt in a few years if nothing is done. This information should help decision makers take steps to reform social security before it is too late.

The social security system is only one of the many illustrations that requires the analysis of huge volumes of data which then need to be described in a form that can be easily understood so as to be used by decision or policy makers to make better and more intelligent decisions.

In the remainder of this chapter, we will discuss descriptive statistics by using the data of Table 2.1. There is no difference between using 230 pieces of data or 230 million pieces of data. The amount of work (which is almost always done on the computer these days) will be more for larger data sets. However, the methods of analysis and the interpretation of the results will be the same.

A convenient way to describe data is in terms of frequency distributions. Absolute and relative frequency distributions are discussed in Section 2.2, and cumulative versions of these frequency distributions are covered in Section 2.3. The question of the appropriate number and size of classes in frequency distributions is addressed in Section 2.4. Finally, a brief discussion of graphical presentation of data is given in Section 2.5.

2.2 Frequency Distributions

Table 2.2 shows another way of presenting the data of Table 2.1. Instead of having each one of the ages of the 230 students printed, the number of students of each age is given. There is an improvement in the volume of data; instead of having 230 numbers, we now have only 19. Furthermore, even if the number of ages was 230 million and we wanted to put the information in a form similar to Table 2.2, there would be slightly more than 100 entries.

Constructing Table 2.2 is straightforward. From Table 2.1, we simply count the number of students whose age is 20 (no one is less than 20), 21, 22, 23, ..., 35, 36, 37, and 38 (no student is older than 38). This count is shown in the second column of Table 2.2.

Table 2.2 : Frequency Distribution of Ages of INSEAD Students.

Age (in years)	Frequency (Number of Students)
20	1
21	1
22	3
23	7
24	10
25	19
26	33
27	35
28	37
29	29
30	20
31	15
32	7
33	4
34	3
35	2
36	2
37	1
38	1
Total	230

Table 2.2 is called a frequency table, or frequency distribution, since it shows the frequency (number) of students of each age. From a descriptive point of view, Table 2.2 makes much more sense than Table 2.1, which just contains numbers (ages). In Table 2.2 these numbers have been organized in a manner that describes much more meaningfully the essence of the data. For instance, you can see at a glance that the youngest student is 20, the oldest is 38, and the most common age is 28.

There are several kinds of frequency distributions. The frequency distribution of Table 2.2 is called an absolute frequency distribution, since the actual number of students of each age is given. Absolute frequency distributions can be easily turned into relative frequency distributions by simply dividing each absolute frequency by the total of 230 students.

Table 2.3 : Relative Frequency Distribution for Ages of INSEAD Students.

Age (in years)	Relative Frequency (Proportion of Students)
20	0.0043
21	0.0043
22	0.0130
23	0.0304
24	0.0435
25	0.0826
26	0.1435
27	0.1522
28	0.1609
29	0.1261
30	0.0870
31	0.0652
32	0.0304
33	0.0174
34	0.0130
35	0.0087
36	0.0087
37	0.0043
38	0.0043
Sum	1

Table 2.3 shows the relative frequency distribution corresponding to Table 2.2. The first relative frequency is $1/230 = 0.0043$, the relative frequency of 28-year-olds is $37/230 = 0.1609$, and so forth. It should be noted that the relative frequencies add up to one. (The actual sum of the second column in Table 2.3 is 0.9998, which differs from 1 only because of rounding error).

The fact that 33 of the students are 26 years old provides some information. However, many people, when faced with this information, would mentally compare the absolute frequency, 33, with the number of students, 230. This comparison reveals that 0.1435, or slightly over 14 percent of the students, are 26 years of age. Proportions and percentages often seem to convey more useful information than do absolute frequencies. An advantage of relative frequency distributions (also true for other relative statistical measures) is that comparisons are easier to make. For example, with the age data, comparisons between different years and across different

schools can be made in terms of relative frequencies. For example, we might ask if the percentage of 30-year-olds is higher at INSEAD than at another school. Absolute frequencies are harder to compare because they depend on the actual number of students of each particular year or institution. This is the major reason that relative measures (such as relative frequency distributions) are very popular in statistics and are usually prefered to absolute measures.

2.3 Cumulative Frequency Distributions

Tables 2.2 and 2.3 show the number of students and the proportion of them at each year of age. Sometimes, however, we might want to know what number of students or what proportion of them is less than a particular age. For example, what proportion of students is 30 years of age or younger? The distributions giving us this type of information are called cumulative distributions.

How cumulative distributions are constructed can be seen in Table 2.4. The basic idea is that at each year the number of students with ages up to this year (including the year itself) are added together. To find the number of students 30 years old or younger, we add: 1+1+3+7+10+19+33+35+37+29+20 = 195. Thus, the relative frequency of students 30 or younger is $^{195}/_{230}$ = 0.8478. The adding is generally done on a year-by-year basis. For instance, 1 student is 20, 1+1 = 2 students are 21 or younger, 2+3 = 5 students are 22 or younger, and so on.

Table 2.4 : Simple and Cumulative Frequency Distributions of Ages of INSEAD Students.

Ages (in years)	Absolute Distribution		Relative Distribution	
	Absolute Number of Students		Proportion of Students	
	Simple distribution	Cumulative distribution	Simple distribution	Cumulative distribution
20	1	1	0.0043	0.0043
21	1	2	0.0043	0.0086
22	3	5	0.0130	0.0217
23	7	12	0.0304	0.0522
24	10	22	0.0435	0.0957
25	19	41	0.0826	0.1783
26	33	74	0.1435	0.3217
27	35	109	0.1522	0.4739
28	37	146	0.1609	0.6348
29	29	175	0.1261	0.7609
30	20	195	0.0870	0.8478
31	15	210	0.0652	0.9130
32	7	217	0.0304	0.9435
33	4	221	0.0174	0.9609
34	3	224	0.0130	0.9739
35	2	226	0.0087	0.9826
36	2	228	0.0087	0.9913
37	1	229	0.0043	0.9957
38	1	230	0.0043	1.000
Total	230		1	

A major advantage of cumulative distributions is the ease by which information of the type "number, or proportion, of students less than 25, older than 27, or between 23 and 31" can be found. Thus, by looking at Table 2.4 we can immediately see that there are 22 students (or 9.57%) who are younger than 25 years of age (that is, 24, 23, 22, 21, 20); we can easily compute that 230 - 109 = 121 (or 52.61%) are older than 27; and we can find out that there are 210 - 5 = 205 (or 89.13%) who are between the ages of 23 and 31 (23 and 31 included).

2.4 Number and Size of Classes

In Tables 2.2 to 2.4 we decided to present the data at yearly intervals. This is a natural breakdown which makes a great deal of sense as far as the data of Table 2.1 are concerned. There may be other cases, however, in which natural breakdowns do not exist. Incomes of people vary from zero to many millions or perhaps even billions of dollars a year. What interval do we choose in this case to construct a frequency distribution? Obviously, not a single dollar. Maybe one hundred dollars, one thousand, or even a larger interval will be more appropriate. But how do we choose?

First of all, it must be understood that the more classes (number of rows in Tables 2.2 to 2.4) there are, the more information will be available to us. This is obviously desirable, but there are also cases where too much can cause information overloads. It is much easier to read and comprehend distributions with fewer classes describing our data than distributions with more classes. Furthermore, more classes in the frequency tables usually mean more work in order to construct and more numbers to store. Both of these factors have become less important with the widespread introduction of the computer, but we still do not want too many classes in the frequency distributions. Usually, between 5 and 20 classes are most appropriate; in the final analysis, it will depend upon the individual application.

In our example, suppose we decided that 19 classes were too many and that we would like to have about 10. To achieve this we need to find the largest and smallest ages and then find the range, which is 38 - 20 = 18. Since we want about 10 classes, we need to take an interval of two years ($^{18}/_{10} = 1.8$, or about 2) rather than one year. This assumes that we want the intervals to be of equal size. Although it is not necessary to have intervals of equal size (we might let the intervals cover one year each between 26 and 29 and two years each otherwise), it is desirable insofar as it seems reasonable to do so. Thus, we will let the interval size be two years for all intervals.

Table 2.5 describes the age data using a class interval size of two years. As can be seen, the number of classes (entries) is about half of those of Table 2.4. To construct the distributions shown in Table 2.5, we proceeded as before. First, we counted the absolute number of students in each class (i.e., 20 and 21 years, 22 and 23 years, 24 and 25 years, ..., 34 and 35, 36 and 37, and finally 38 and more years of age); these may then be added up to find the cumulative figures or divided by the total to find the relative frequencies.

Table 2.5 : Frequency Distribution when the Class Interval is Two Years.

Ages (in years)	Absolute Distribution		Relative Distribution	
	Absolute Number of Students		Proportion of Students	
	Simple distribution	Cumulative distribution	Simple distribution	Cumulative distribution
20 - 21	2	2	.0087	0.0087
22 - 23	10	12	.0435	0.0522
24 - 25	29	41	.1261	0.1783
26 - 27	68	109	.2957	0.4739
28 - 29	66	175	.2870	0.7609
30 - 31	35	210	.1522	0.9130
32 - 33	11	221	.0478	0.9609
34 - 35	5	226	.0217	0.9826
36 - 37	3	229	.0130	0.9956
38 - 39	1	230	.0043	1.000
Total	230		1	

It should be re-emphasized that there is little difference between Tables 2.4 and 2.5, apart from the former having more classes than the latter. Having more classes includes advantages (more information available) and disadvantages (possible information overload, more entries to store). It will be up to the person analyzing the data to decide on the number of classes depending upon the specific application, keeping in mind that the number of classes should usually be between 5 and 20.

2.5 Graphical Presentation of Data

Tables 2.2 to 2.5 are frequency distributions describing statistical data. They can be presented to decision or policy makers in the tabular form shown in, say, Table 2.4 or Table 2.5, but the effect might be lost. Most people have difficulties comprehending numbers while they can easily remember, and be impressed by, graphical presentations. It is worthwhile, therefore, to go beyond just summarizing the numbers in a frequency distribution to presenting these frequency distributions in a graphical form. This will most vividly reveal the information in the data and, at the same time, will catch the attention of those to whom the information is presented.

There are many visual or graphical forms that can be used to present descriptive statistical measures such as frequency distributions. Strictly speaking, this is not within the domain of

statistics, but statisticians and others dealing with data use such forms quite often. It is not the purpose of this chapter to dwell upon visual and graphical forms of presenting data. However, some of the most widely used graphical presentations are shown in Figures 2.1 to 2.5.

Figures 2.1 and 2.2 show the simple frequency distributions of Tables 2.4 and 2.5. As can be seen, the vertical axis on the left shows the absolute number of students, while that on the right shows the corresponding proportions. It is easy, therefore, to look at the data at a single glance and know either absolute or relative frequencies. This type of graph is called a histogram. By comparing Figures 2.1 and 2.2, you can see how the choice of intervals affects the histogram in this example.

The same data can be presented in various graphical forms. For instance, Figure 2.3 shows a frequency polygon for the student ages with class intervals of two years. This is obtained by joining the points at the middle of the top of each bar in the histogram given in Figure 2.2. A cumulative histogram and a cumulative frequency polygon for the same data are shown in Figures 2.4 and 2.5.

Figures 2.6, 2.7, and 2.8 show some other common types of charts. Figure 2.6 is called a pie chart because it presents each element of the total as a part of a pie. Figures 2.7 and 2.8 are called bar charts.

It should be noted that, as time passes by, nice graphical designs will become commonplace, since computer graphics are evolving at an astonishing pace. There already exist graphic systems that can be used by non-specialists and can do practically any graphical design desired by the user in a matter of minutes. There is no doubt that in the future computer graphics will become affordable by almost anyone, thus allowing more effective ways of presenting descriptive statistical measures.

Figure 2.1 : Histogram of Student Ages in Proportions.

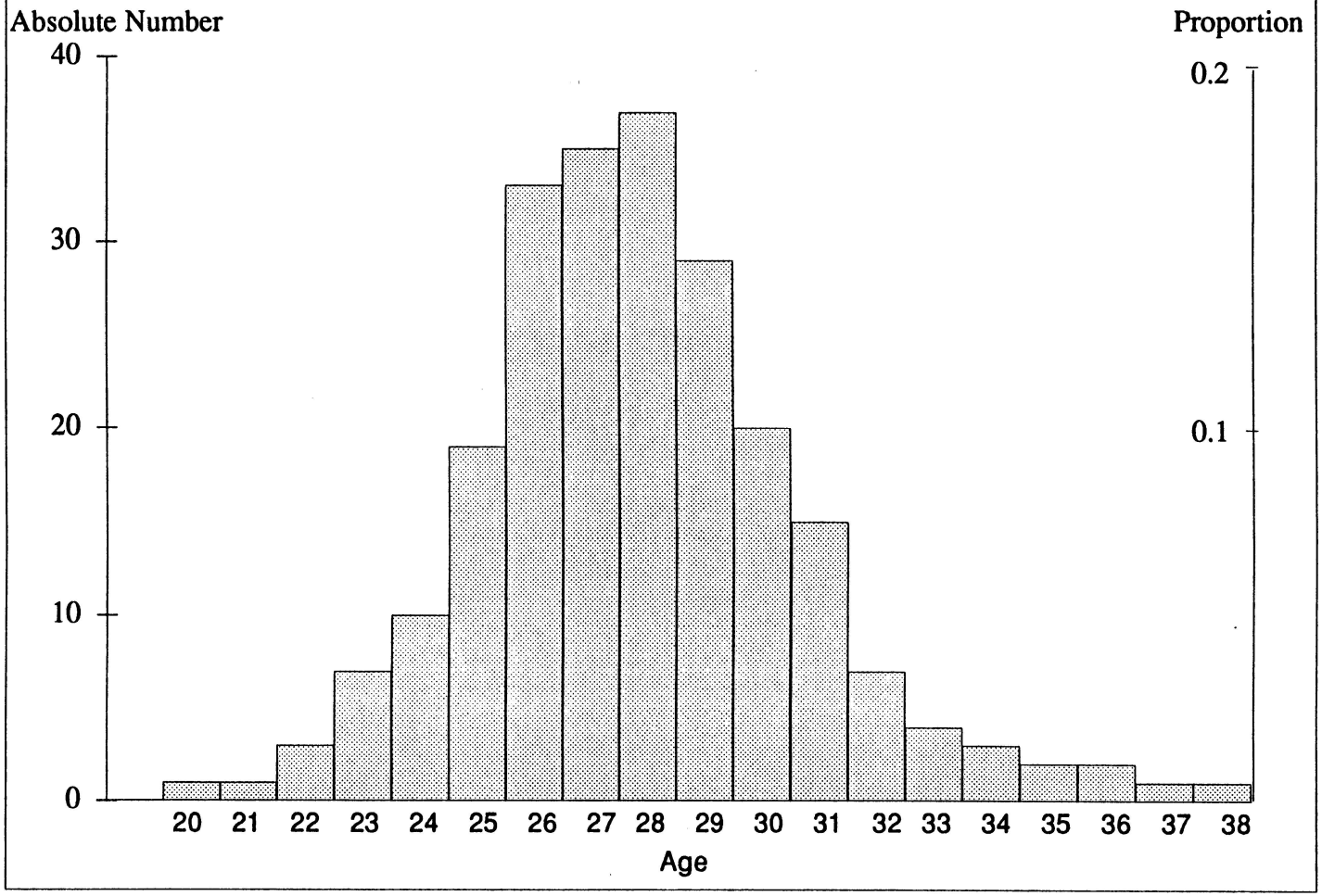


Figure 2.2 : Histogram of Student Ages with Class Interval of Two Years, in Proportions.

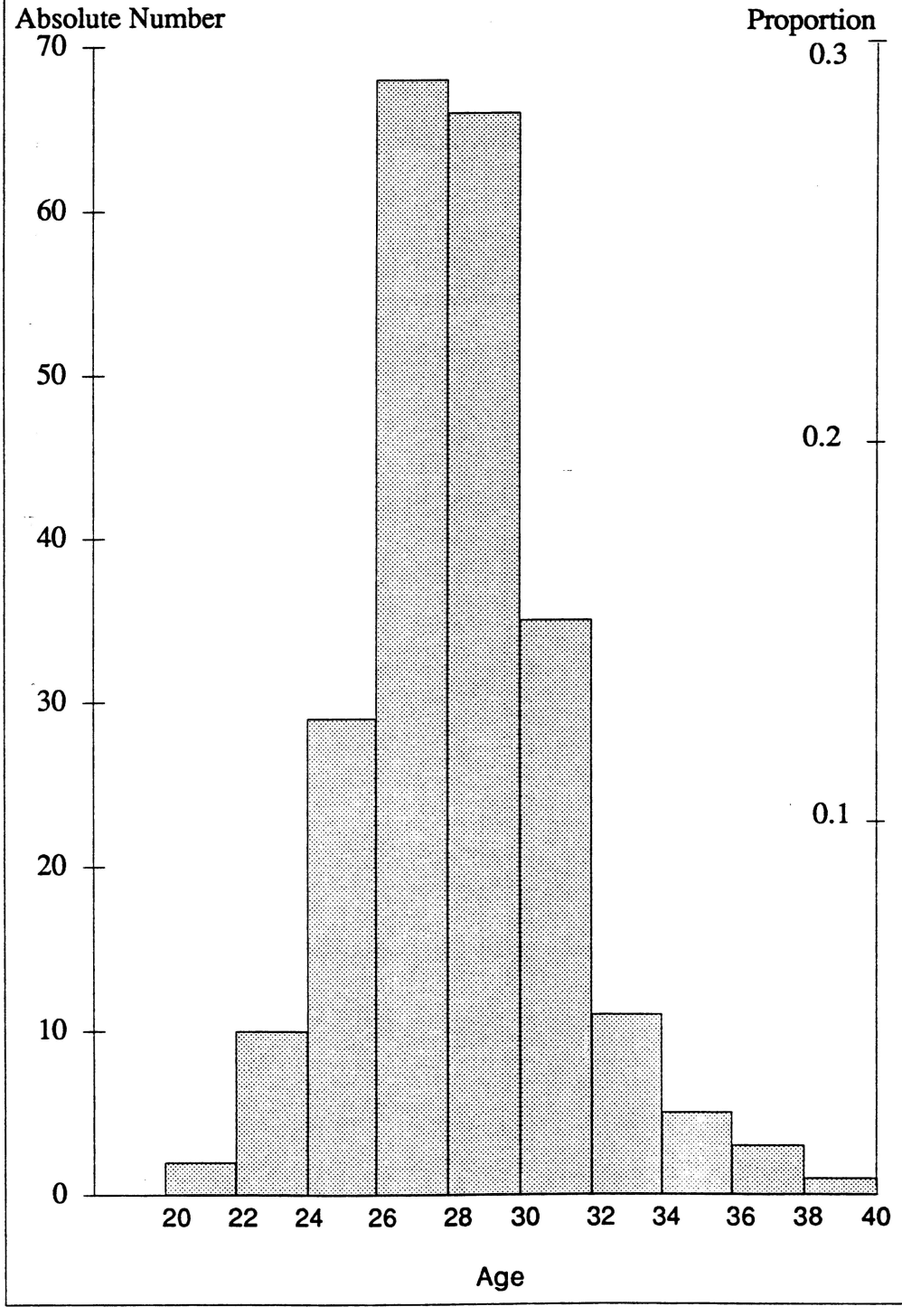


Figure 2.3 : Frequency of Student Ages.

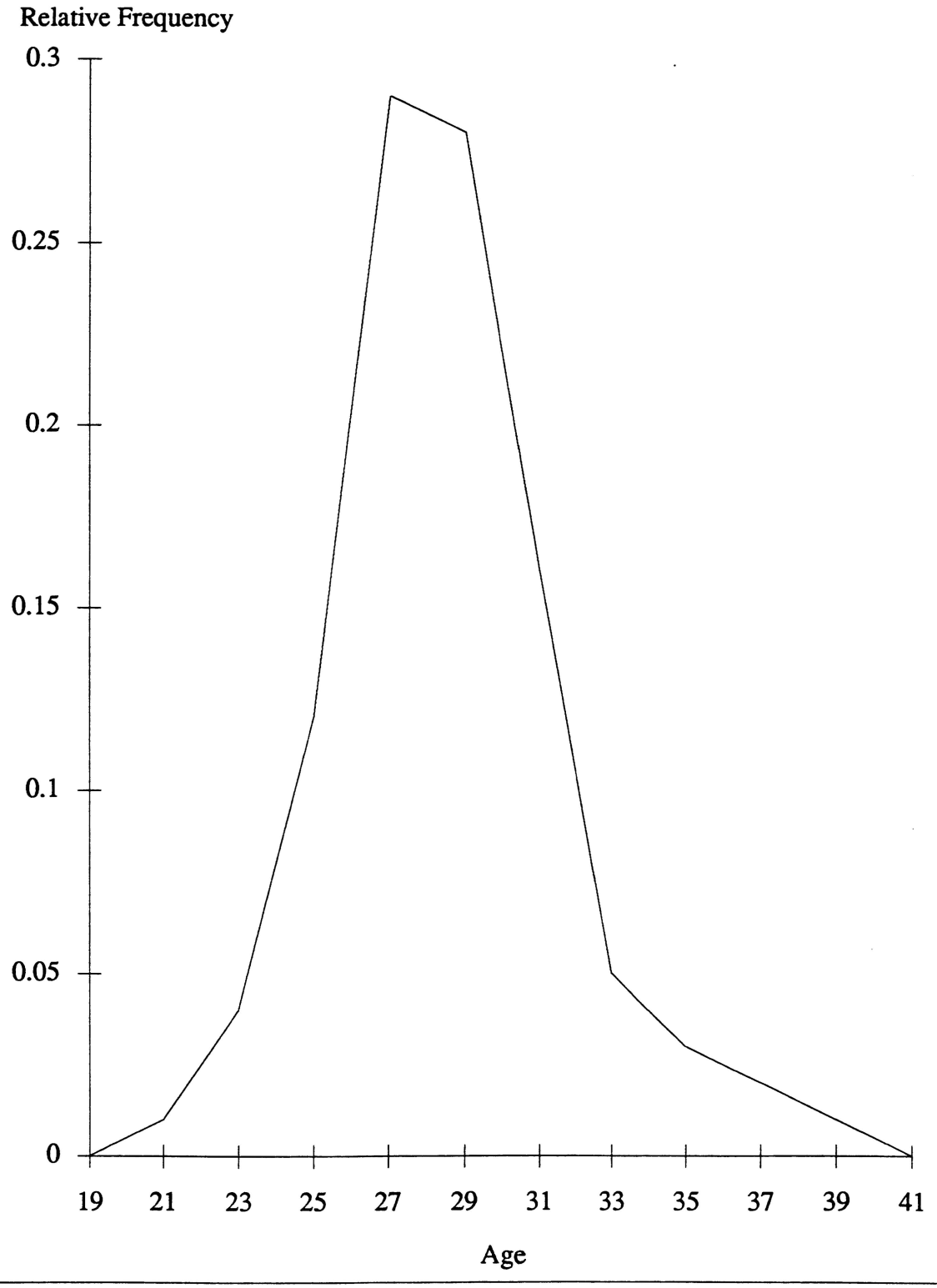


Figure 2.4 : Cumulative Histogram of Student Ages.

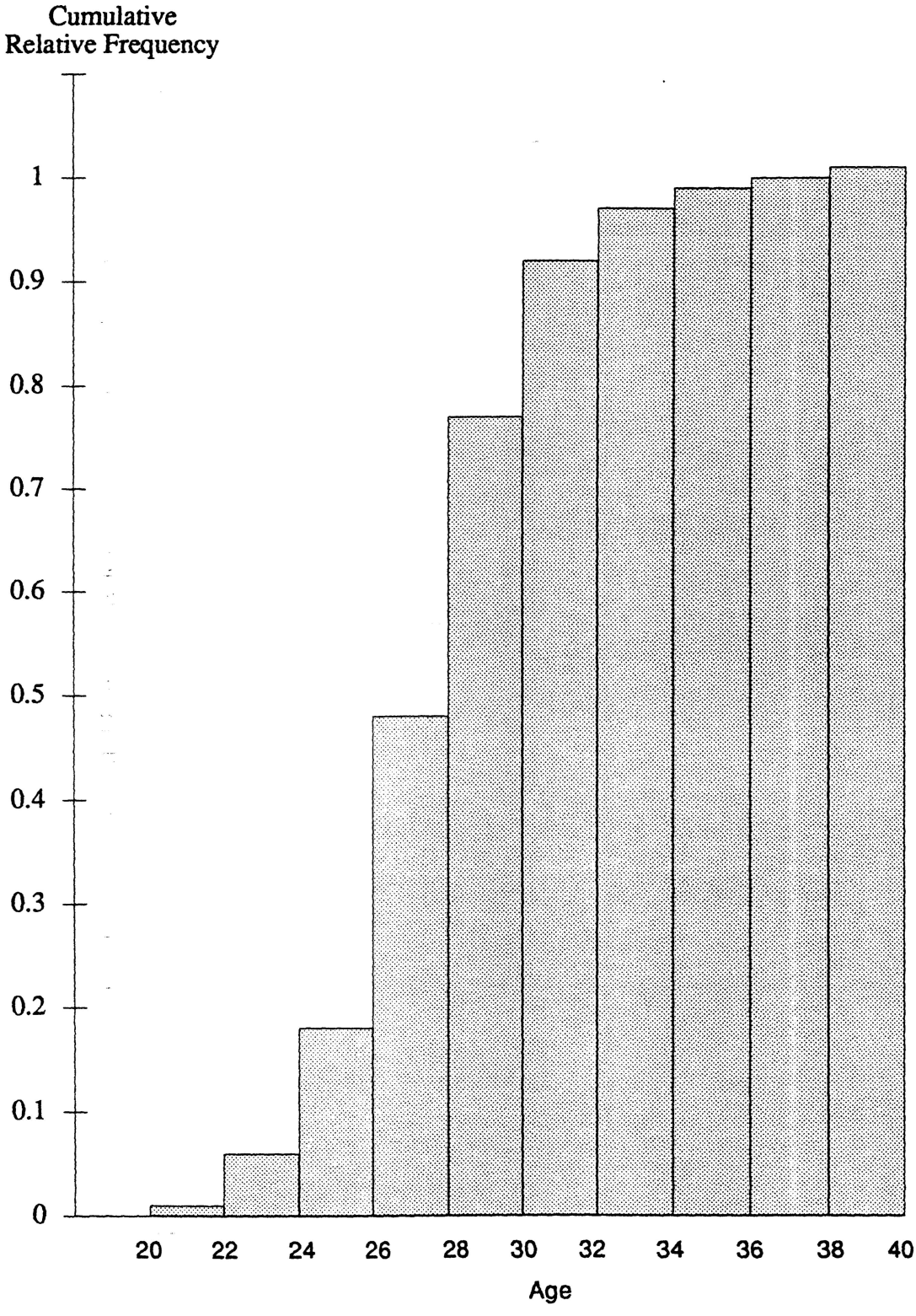


Figure 2.5 : Cumulative Frequency Polygon of Student Ages.

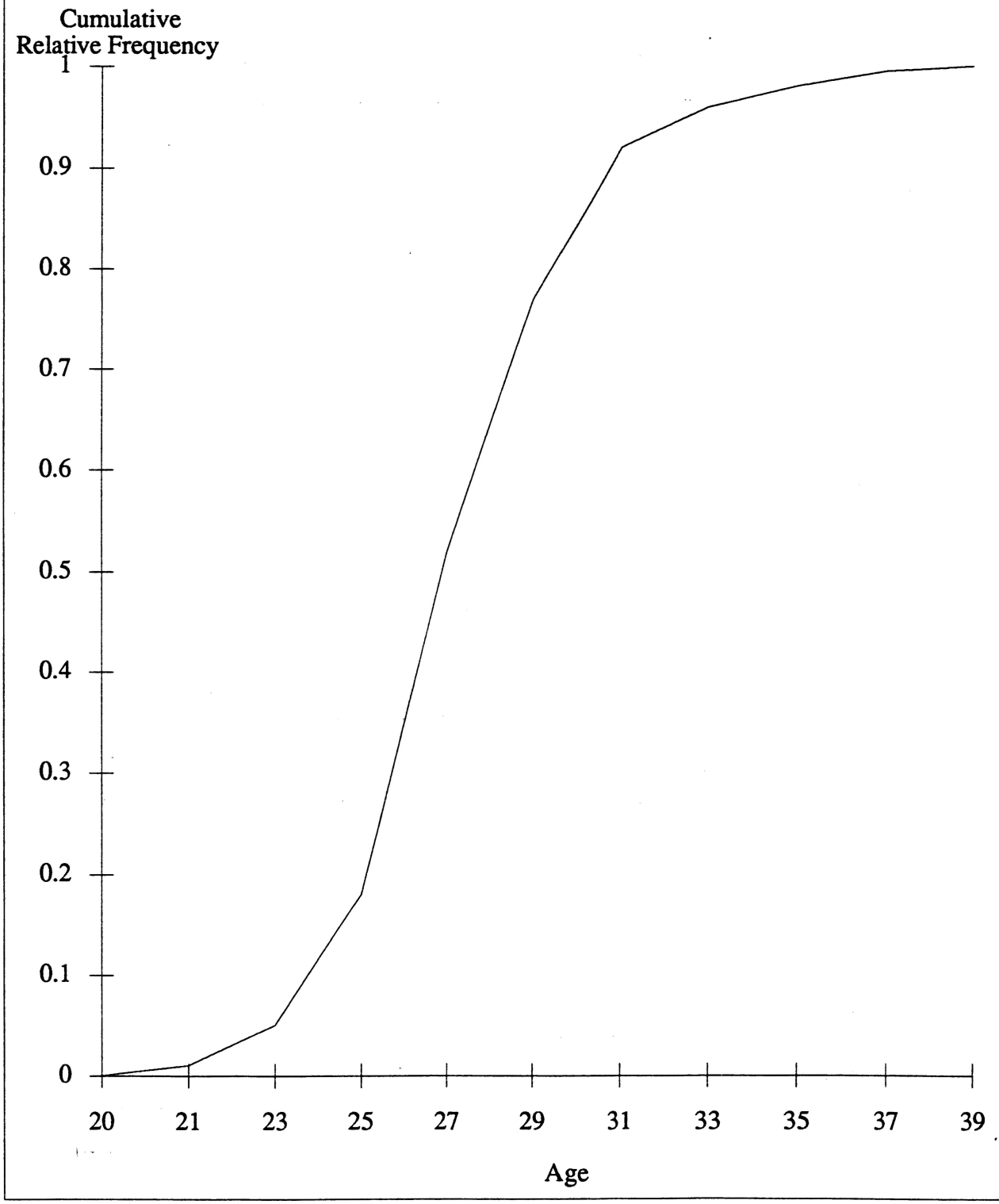
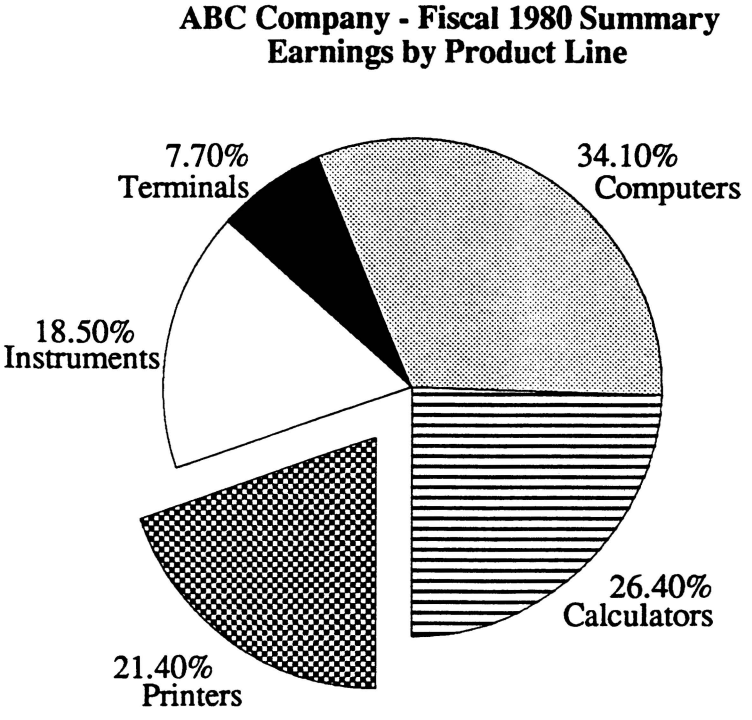


Figure 2.6. A Pie Chart



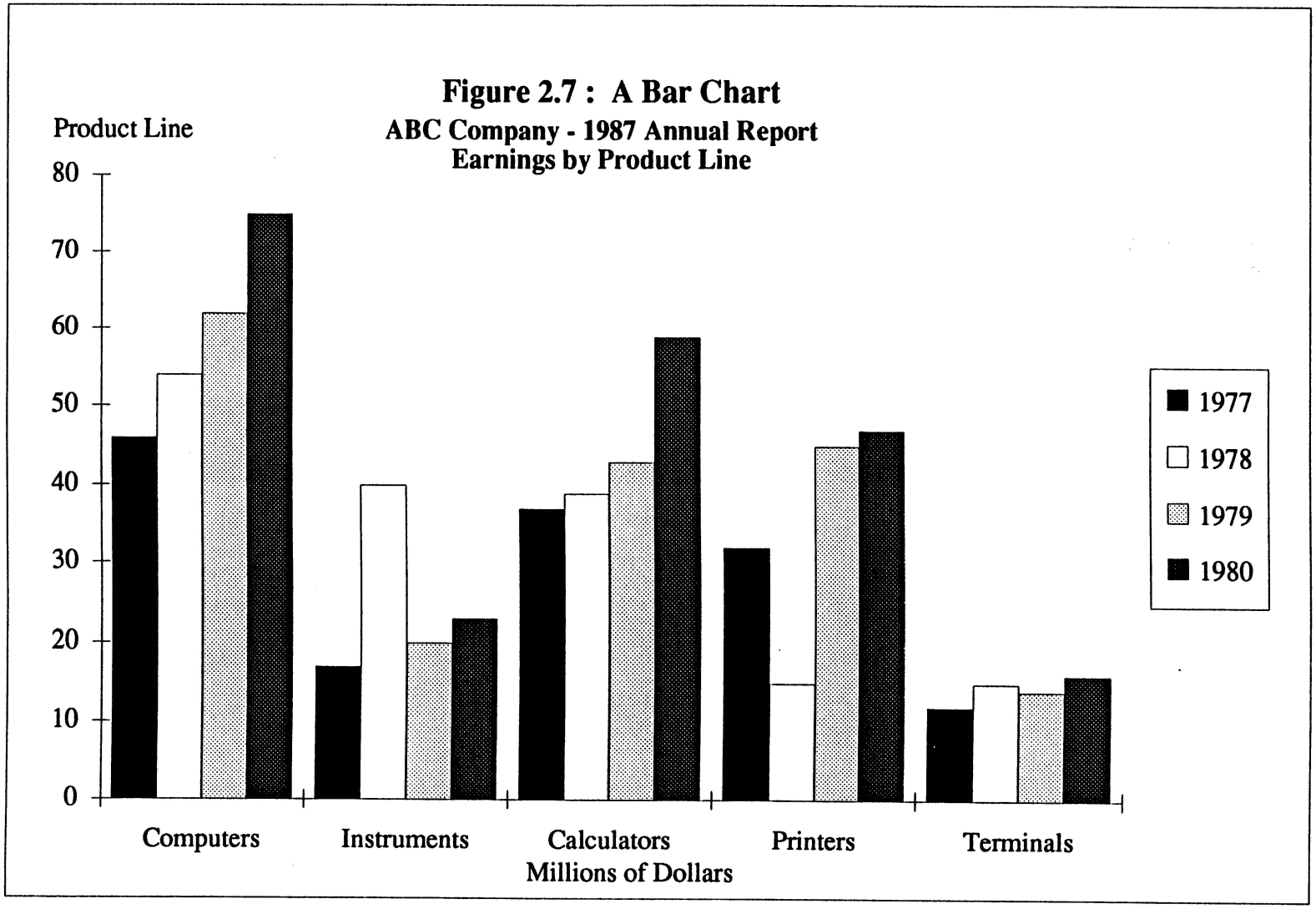


Figure 2.7 : A Bar Chart
ABC Company - 1987 Annual Report
Earnings by Product Line
Product Line
80
70
60
50
40
30
20
10
0
1977
1978
1979
1980
Computers
Instruments
Calculators
Printers
Terminals
Millions of Dollars

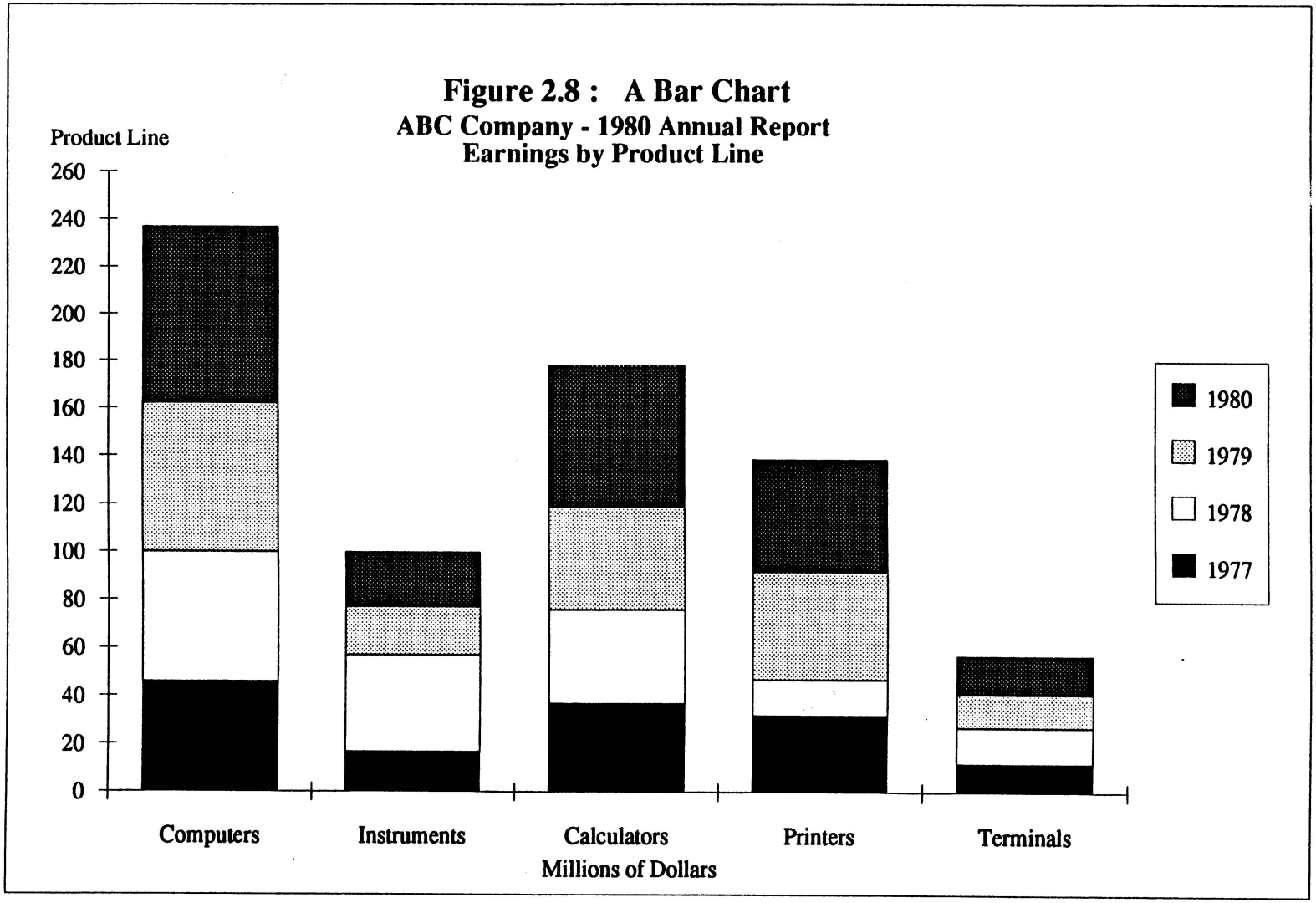


Figure 2.8 : A Bar Chart
ABC Company - 1980 Annual Report
Earnings by Product Line
Product Line
260
240
220
200
180
160
140
120
100
80
60
40
20
0
1980
1979
1978
1977
Computers
Instruments
Calculators
Printers
Terminals
Millions of Dollars