Contents

Chapter 4

Presenting Data

4.1 Introduction

Once the population and the variables have been defined, the researcher obtains data describing the distribution of these variables for some or all members of the population. After this data has been obtained, the researcher then begins to examine the data. This may be done by looking at the raw data itself, or organizing the data into tables, charts, and diagrams, and analyzing these. If the data have already been produced by other researchers, various tables and diagrams will be available from those who have produced this data.

This chapter discusses various methods of organizing data so that the distribution of variables, and the relationships among these variables, can be presented in a straightforward and understandable manner. There are many alternatives which the researcher can choose when presenting the data. There is not necessarily a perfectly correct, or single correct way of presenting the data, although there may be some methods that are clearly incorrect. Rather, there is likely to be a range of acceptable procedures, and choices will have to be made by the researcher concerning the best procedure in the circumstances. The particular method chosen will depend on the type of scale, whether the variable is discrete or continuous, and on the nature of the distribution of the variables themselves. Each of these considerations will become clearer as a variety of examples are examined in this chapter.

In presenting the data, the researcher should attempt to present the data so that the characteristics of the population are illustrated clearly. This should be done in a manner which does not bias the presentation,

either overemphasizing unimportant factors or underemphasizing important factors. Of course, each researcher may have a somewhat different idea of what is important or unimportant, leading to quite different types of data presentation. If data have already been organized into tables and diagrams, there may be little choice but to use the data in the form they are available. When using these data, the analyst should always attempt to determine how the data has been produced, and what problems are associated with the data.

In this chapter, various procedures for presenting data are discussed. Most readers will be familiar with some of these procedures, but may have used them in a somewhat different way than the rules adopted here. Much of the process of presenting data is a matter of learning the set of procedures and guidelines which are ordinarily adopted in Statisics. As a result, this chapter involves learning the conventions concerning data presentation for statistical work.

Chapter Summary. This chapter begins with a short discussion of how raw data can be presented as a listing of data. This is followed by a discussion of frequency distributions, and how to organize raw data into frequency distributions using tallies, stem and leaf displays, or with a computer. A short discussion of frequency distribution tables, notation and diagrams follows. Methods of producing bar charts or histograms are also discussed.

4.2 Organizing Data

Data gathered from surveys, experiments, or administrative records are considered to be **raw data** when first obtained. These data may be lists of numbers, responses on questionnaires, or entries on forms. There may be nonsampling errors in the data, and the research investigators may be able to correct some of these errors before analysing the data. Once the data has been examined in this way, it is still a list of numbers. At this point, the researcher must decide how to present these numbers. Section 4.2.1 discusses the presentation of lists of numbers. Subsequent sections examine how data may be organized into categories or groups.

4.2.1 Small Data Sets

If the set of data is quite small, with the original data in the sample consisting of only a few cases, it may be possible to show all this data as a

list of cases. This may be the best approach in some situations, in other circumstances it may be difficult to analyze the whole data set with a list of all the data, and some summary form of presentation will be required. Several situations illustrating small samples are given here.

City	Population (in thousands)	Number of Break and Enter Offences	Rate of Break and Enter Offences (per 100,000 population)
Edmonton	605	12,778	2,112
Regina	175	3,261	1,863
Saskatoon	177	2,675	1,511
Winnipeg	619	9,073	1,466
Calgary	692	7,691	1,111
Canada	26,602	215,361	1,427

Table 4.1: Break and Enter Offences, Five Prairie Cities and Canada, 1990

Source: Statistics Canada, Canadian Centre for Justice Statistics, **Juristat Service Bulletin**, catalogue 85-002, Volume 12, No. 1, January 1992, Tables 2,4,5.

One situation where a list of all the cases is presented is if there is a comparison of some cities or provinces in Canada. If this is the case, then the list of all the values of the variable, along with the names of these cities (or provinces) may be given. As long as this list is relatively short, it is best to present all the data. This allows anyone examining the data to see the set of data presented.

Example 4.2.1 A Small Population - Prairie Cities

An example of this is presented in Table 4.1. If the population is defined as Prairie cities of 100,000 and over people, then this set of 5 cities represents a complete coverage of these cities, since there are only 5 such cities. The variables are the number of break and enter offences and the rate of break and enter offences. The rates are presented as the number of break and enter offences in 1990 per 100,000 people in each city. Since cities with more

people are likely to have more such offences, it is necessary to compute a rate for such offences. This allows comparison of the number of offences per 100,000 people across the different cities. Such a rate provides a meaningful comparison of the relative incidence of break and enter offences in the different Prairie cities.

The ratio of break and enter offences in any city is calculated by dividing the number of such offences in that city by the city's population. Since the rate is per 100,000, the above ratio is multiplied by 100,000. For example, for Edmonton there are 12,778 offences for 605,000 people, for a ratio of $12,778/605,000 = 0.02112$ *offences per person. This is*

$$\frac{12,778}{605,000} \times 100,000 = 2,112$$

break and enter offences per 100,000 people in Edmonton. The rates for the other cities are calculated in a similar manner.

As can be seen in Table 4.1, in 1990, Edmonton had the largest number of break and enter offences, and also the highest rate. In contrast, Regina has a much lower number of offences, but with the rate of such offences being quite high. The number and rate of break and enter offences can also be seen for the other cities in Table 4.1, and anyone reading this table can get a quick overview of the pattern of break and enter offences in these large Prairie cities. In addition, this table gives the national rate, allowing comparison of these cities with the average level across the country as a whole.

Suppose that a very small sample is selected. This may be a small experiment with only a few subjects, or a small sample. While it is not common to have a sample with a very few cases, say less than 10, the situation does sometimes occur, especially if the researcher has to resort to some nonprobability sampling technique. Graduate students in Psychology often conduct experiments using volunteer subjects. Because of the difficulty of finding volunteers, the number of subjects in such experiments is often quite small. While it is generally advisable to have larger sample sizes, this is difficult or impossible in some circumstances.

Example 4.2.2 A Small Sample with Several Variables

In the example of Table 4.2, a small random sample of 7 respondents is selected from the data set of Appendix **??**. *The complete guide to the*

variables is contained in Appendix **??**. *Here only a few variables concerning these 7 cases are presented. All of these 7 respondents are in the labour force with GMP representing their gross monthly pay in dollars. PARTY represents political preference at the provincial level. CLASS is the social class which the respondent considers himself or herself to be in, where 'U. Middle' represents the upper middle class, and the other values of CLASS are self explanatory. THRS represents the total hours worked at jobs per week, and FIN the family income in dollars.*

Case No.	SEX	AGE	PARTY	GMP	THRS	CLASS	FIN
1	Female	32	NDP	1300	50	U. Middle	42,500
2	Female	33	NDP	1950	25	Working	27,500
3	Male	34	NDP	2500	40	Middle	37,500
4	Female	34	NDP	1850	40	Middle	52,500
5	Male	46	PC	5000	50	Middle	100,000
6	Female	34	NDP	700	16	Working	37,500
7	Female	53	LIB	3125	40	Middle	62,500

Table 4.2: A Sample of 7 Respondents

The data concerning the variables in this sample is certainly complete for this sample of 7 cases, and by spending some time examining this data, one could describe it. For example, the sample contains 5 females and 2 males, with the ages of the respondents being fairly similar to each other for the most part. 5 of the 7 respondents are in their thirties, with only one each in the forties and fifties. Political preference generally seems to be for the NDP with the one person who supports the PCs is a male of very high income.

By spending time examining a data set as in this last example, a more complete description of the sample, and the relationship among the variables in the sample, could be obtained. But it is rather tedious, and difficult, to discern all the patterns associated with this data set. As a result, a method of summarizing this data would be useful. The following sections, and the summary measures of the Chapter 5 show various ways in which this can be carried out.

An additional dificulty involved in small data sets is that generalizing from such data sets to larger populations is very difficult, and perhaps misleading. With a small sample size generalizations of results from the sample to a larger population would be subject to high levels of sampling error. This will be discussed in more detail in Chapter 7.

Trend Data. Another situation where the list of all the data may be presented is when a variable is measured over time. If regular measurements concerning the same variable are made at several different times, then a list of all the values of the variable, and the times when the measurements were made, may prove useful in examining the trend in this variable. Examples of this are variables such as the *unemployment rate* (see Table 2.4), *daily temperature lows* or Statistics Canada's monthly reports of the *consumer price index*. Each of these may be printed as a list or numbers, or may be summarized in a graph or chart, showing the trend.

Example 4.2.3 Trend Data - Monthly Gallup Poll Results

	Party Supported				
Date of Poll	PC	Liberal	NDP	Other	Undecided
July 11-14, 1990	12%	32%	15%	7%	34%
April 4-7, 1990	10%	32%	16%	7%	35%
Jan. 10-13, 1990	16%	34%	16%	5%	29%
Oct. 2-5, 1989	21%	32%	16%	4%	27%
July 5-8, 1989	22%	27%	18%	4%	29%
April 5-8, 1989	28%	28%	17%	2%	25%
Jan. 4-7, 1989	40%	22%	19%	3%	16%
Oct. 1988 Election	43%	32%	20%	5%	

Table 4.3: Distribution of Gallup Poll Respondents, 1989-1990. Per Cent Supporting each Party and Per Cent Undecided

This example comes from the regular Gallup opinion polls, which are produced and released each month by Gallup Canada, Inc. Each month a sample of just over 1,000 Canadian adults is asked a variety of questions.

As part of this set of questions, respondents are asked to state their political preference. The question asked of respondents is:

> If a federal election were held today, which party's candidate do you think you would favor?

Some people have not decided which party they would support, and they are considered to be undecided. The responses of both those who have decided and of undecided, based on this poll, are published each month in the Gallup Canada, Inc. report **The Gallup Report**.

The data in Table 4.3 are the results of the Gallup national opinion polls over the period from January 1989 through to July 1990. The data in this table are the Gallup results from the first month of each quarter over this period. This table includes all respondents, including the undecided respondents. The numbers in Table 4.3 represent the percentage of respondents surveyed who said they would vote for each party, along with the percentage of respondents who said they were uncertain or undecided as to how they would vote. The table also shows the actual percentages supporting each of the major parties in the October 1988 election.

It should be noted that Gallup does not usually report the data in this manner, but usually takes out the undecided respondents. Table 4.3 shows how people actually responded in each of the months shown. The first point to note is how quickly the percentage of undecided respondents jumped after the election. By January, 1989 there were already 16% of respondents who were not sure which party they would vote for. Over the succeeding months, this percentage continued to increase, so that over one-third of all those polled said they were undecided by mid-1990.

Over this period, the trend in the percentage who support each of the political parties can then be examined. Most notable is the decline in the percentage of respondents who said they would vote for the PC party. In January, 1989, in the "honeymoon" period after the solid PC victory in October 1988, the PCs still had the support of 40% of those polled. But after this, the percentage of PC supporters began to drop, and continued to drop through mid-1990, reaching a low point of only 10% of all those polled by April, 1990.

While the Liberal party might ordinarily have been thought to be the beneficiary of this drop in PC support, the level of Liberal support never increased much above the 32% support the Liberals received in the 1988 federal election. The same was true of the NDP over this period. While the Liberals gained a little in some months, the NDP actually slipped slightly

in the percentage of popular support it received. The only category that gained much over this period was the category of Undecided.

Based on this brief set of data and the observations just made, it appears that PC support did slip dramatically after the October, 1988 federal election. But rather than switching to other parties, it appears that former PC supporters interviewed by Gallup merely said they were undecided. From this data, it cannot be proved that these former PC voters moved to the undecided category, but there is at least a strong suggestion that this is what happened. Whether other parties eventually benefit from this decline of PC support, by gaining more seats at a subsequent election remains to be seen. What this data suggests is that when voters are disaffected by a party, they are not so likely to lend support to another party very quickly. Rather, they become undecided, and take some time to decide how they might vote in the next election.

4.2.2 Ungrouped and Grouped Data

When data is first obtained, it may be presented as a list of values of the variable. Such a situation was shown for a number of variables obtained from a sample of 7 people in Table 4.2.

Definition 4.2.1 A list of all the values for the variable in the data set is referred to as **ungrouped data**.

Data is always ungrouped when it is first obtained, because the values of the variable for each member of the sample (or population) are obtained by the researcher. But if there are very many values of the variable, the list of all the values of the variable can be quite long, and may be difficult to absorb and analyze.

In order to present or analyze data, the data are ordinarily organized into categories, or grouped into intervals.

Definition 4.2.2 Grouped data is data which has been organized into categories or grouped into intervals.

When data has been grouped, a list of categories or intervals, along with the number, proportion, or percentage of cases in each category or interval, is usually presented. In Section 4.3 on frequency distributions, all of the tables represent grouped data. Table 4.4 organizes the various political preferences into groups, and Table 4.5 groups 50 respondents into various intervals of

family income. Since most data are obtained for a relatively large number of cases, most data are grouped for purposes of presentation. Methods of grouping data are given in Section 4.5.2 and 4.5.3. These methods can also be used for organizing data in order to analyze the data.

With computers, data are usually left in their original ungrouped or raw form when they are entered into a data file on the computer. Leaving it in this original form means that those analyzing the data have access to all this original data on the computer. Analysis of data ordinarily begins with this complete set of ungrouped data and part of the process of data analysis is the grouping or organization of the data in order to study and meaningfully present the data. A short discussion of how data can be coded and entered on the computer is given in Section 4.5.4.

4.3 Frequency Distributions

The examples in the last section describe data in situations with a relatively small sample of cases. Such data were referred to as ungrouped data. Where there is a larger number of cases, the data are usually organized into various groups or categories. These may be categories that refer to discrete values of the variable (Table 4.4), or intervals that include several possible values of the variable (Table 4.5). Where there are a considerable number of cases for each **value** of the variable (Table 4.4) or where there are many values of the variable for each **interval** (Table 4.5), the data are considered to be grouped data. Such data are best presented by giving the **distribution** of the values of the variable for all of the cases for which there is data.

Definition 4.3.1 A **frequency distribution** gives the values of the variable, along with the number of cases, or frequency of occurrence, of each value of the variable.

The frequency distribution shows how the cases in the data set are distributed across the different values, which values occur more frequently, and which values are less common. This allows anyone examining the table to obtain an idea of what the sample, or the population, looks like, in terms of the variables being examined.

Example 4.3.1 Distribution of Political Preference

A frequency distribution of political preference at the provincial level is given in Table 4.4. These responses are based on the question

If an election were held today, for which party's candidate would you vote?

Political Preference	No. of Respondents
Liberal	6
NDP	24
Progressive Conservative	11
Undecided	8
Would not vote	1
Total	50

Table 4.4: Provincial Political Preference, 50 Regina Respondents

Source: Appendix **??**

The scale of political preference is a nominal and discrete one. Since the scale is a nominal one, the values shown in the frequency distribution are the names of the three major Saskatchewan political parties, along with the undecided response and a response of not voting. While there may be other possible responses to this question, the column Political Preference in Table 4.4 represents the set of all responses for the 50 persons in the sample of Appendix **??***. Examination of the table gives the distribution of the 50 respondents across the various political preferences. The most common preference is NDP, with just under one-half as many respondents indicating a preference for the PCs. The Liberals are supported by only 6 of the 50 respondents surveyed, and those who are undecided who they would support outnumber those who say they would vote Liberal. Finally, one respondent indicated that he or she would not vote.*

Example 4.3.2 Frequency Distribution of Incomes

The sample of 50 Regina respondents in Appendix **??** *gives a list of the values of various variables for these respondents. One of the variables is FIN, the family income of the respondents. This set of 50 incomes has been taken and organized in various intervals in Table 4.5. This table is a frequency distribution table for the family income of the 50 respondents*

Family Income in thousands of dollars	Number of Respondents
0-19	6
20-39	13
40-59	17
60-79	9
80 and over	5
Total	50

Table 4.5: Family Income of 50 Respondents

Source: Appendix **??**

in Appendix **??**. *By examining the table, one can get a quick and fairly accurate idea of the distribution of family incomes for these respondents. There are relatively few (only 6) respondents at the lowest income level, with considerably more respondents at the middle income levels of 20-39 and 40-59 thousand dollars. Altogether, these two intervals contain 30, or over half, of all the respondents. Above $60,000, there are fewer respondents, although 9, or almost one-fifth are in the 60-79 thousand dollar category. But the top category of $80,000 or above family income has only 5 respondents.*

By grouping the data concerning the family incomes of respondents, it is possible to obtain a quick overall picture of how the family incomes of these respondents are distributed. The complete list of values of family income in Appendix **??** *is needed to construct the distribution of Table 4.5, but the list of all 50 cases is too difficult to absorb in its ungrouped form.*

4.4 Notation

4.4.1 Notation for a Frequency Distribution

Where there are relatively few values that the variable takes on, the frequency distribution is presented as a list of all the possible values of the variable. The frequency distribution gives these values, along with the num-

ber of cases that take on each of these values. In order to present these in a systematic manner, algebraic symbols are usually employed.

In statistical work, variables are usually given algebraic symbols such as X, Y, or Z, that is, capital letters near the end of the alphabet. The number of cases which take on each value is given the symbol f, for **frequency** or **frequency of occurrence** of the variable. In the case of a variable X, the frequency distribution is then a list of the values of the variable X, along with the respective frequencies of occurrence, f of variable X. Finally, the **number of cases** in the sample is usually given the symbol n. That is, there are n members of the population which are selected for the sample, and there are f of these which take on each of the values X. This means that the sum of the frequencies f is n.

Example 4.4.1 Distribution of Number of People per Household

X	f
1	155
2	286
3	164
4	223
5	86
6	21
7	5
8	1
Total	941

Table 4.6: Frequency Distribution of Number of People per Household

Table 4.6 contains a frequency distribution of the number of residents per household for a sample of 941 Regina respondents. The variable is the number of household members and this is a discrete, ratio level scale. If X is defined as the number of household members, the sample size is $n = 941$ and the frequencies of occurrence are the values of f, then the frequency distribution can be presented as in Table 4.6.

This table shows that most respondents live in households with 4 or less members. Of the 941 respondents, only 113 live in households having 5 or

more members per household. In this sample, no household has more than 8 people living in it and there is only one household with exactly 8 members. The most common household sizes are 2 and 4. There are 155 respondents who live alone, and there are almost twice as many respondents (286) who live in a household with someone else.

4.4.2 Indexes

In order to provide a more complete notation, it is useful to place subscripts on the variable X and the frequencies f. These subscripts are used to denote each value of X and f. This notation is introduced here and used in more detail in Chapter 5.

Ungrouped Data. In the situation where a list of values of the variable is presented, each value can be symbolized with a particular indexed value of X. The values of the variable are listed and indexed, so that the **index** is a subscript for X which is used to denote each value.

Suppose the sample size is n, that is, there is a list of n values for the variable X. These n values can be labelled X_1, X_2, X_3, ... X_n. X_1 represents the first value of X, X_2 the second value of X, and so on, until the last value X_n is used to represent the last, or the nth value.

For example, in Table 4.1, suppose that variable X represents the rate of break and enter offences. Let the cities be numbered from 1 to 5 with Edmonton as case 1, Regina as 2, and so on through Calgary as city 5. Then $X_1 = 2,112$, $X_2 = 1,863$, $X_3 = 1,511$, $X_4 = 1,466$ and $X_5 = 1,111$. From the same sample, if variable Y represents population, then Then $Y_1 = 605$, $Y_2 = 175$, $Y_3 = 177$, $Y_4 = 619$ and $Y_5 = 692$.

The notation can be used to shorten the presentation even more. This is done by considering the variable X as taking on the set of values X_i, where $i = 1, 2, 3, ..., n$. That is, there are n values of X in the sample, and these can be considered to be values X_i, where i takes on n integer values beginning at 1 and continuing until all n values are taken into account. In this case i is an **index** which can take on a set of values, and the set of values taken on by the index is $i = 1, 2, 3, ..., n$.

These symbols are used again in Chapter 5, in the formula for the calculation of the mean (average) of the variable.

Grouped Data. In the case of grouped data, the raw or ungrouped data has been organized into a number of categories. Suppose the variable X is

X_i	f_i
$X_1 = 1$	$f_1 = 155$
$X_2 = 2$	$f_2 = 286$
$X_3 = 3$	$f_3 = 164$
$X_4 = 4$	$f_4 = 223$
$X_5 = 5$	$f_5 = 86$
$X_6 = 6$	$f_6 = 21$
$X_7 = 7$	$f_7 = 5$
$X_8 = 8$	$f_8 = 1$
Total	$n = 941$

Table 4.7: Number of People per Household

grouped into k different categories, and these categories are values of the variable which occur in the sample. Let these k values of X be X_1, X_2, X_3, ... X_k.

More generally, the variable X can be considered to take on values X_i, where $i = 1, 2, 3, ..., k$. With this set of values for X, the respective frequencies are f_1, f_2, f_3, ... f_k. That is, there are f_1 cases which take on value X_1, f_2 cases which take on value X_2, and so on, until the last value X_k occurs f_k times. In general, one can say that there are f_i cases which take on value X_i, where $i = 1, 2, 3, ..., k$.

Also note that

$$f_1 + f_2 + f_3 + ...f_k = n$$

That is, the sum of all the frequencies of occurrence f_1, f_2, f_3, ... f_k is the sample size n.

For example, in Table 4.6 there are 8 categories for X, the number of people per household, so that $k = 8$. In this example, $X_1 = 1$, $X_2 = 2$, ... , $X_8 = 8$. Altogether there are $n = 941$ cases in the sample, with $f_1 = 155$, $f_2 = 286$, $f_3 = 164$, ... $f_8 = 1$. All this is summarized in Table 4.7. While it is not common to place all the symbols in a frequency distribution table, this is done in Table 4.7 in order to make clear how the notation using an index is applied.

4.5 Grouping Data

There are various ways in which data can be organized into categories or grouped into intervals. This section first contains a discussion of the categories into which data may be grouped. Then the most common methods of grouping data, constructing a tally or a stem and leaf display are examined. A short discussion of how the data can be entered onto a computer, and then organized into groups, is also given.

Discrete Variable with Few Cases. If the raw or ungrouped data is presented to a researcher to analyze or present, the first step involved is likely to be the grouping of the data into categories or intervals. If there are relatively few values of the variable, as in Table 4.6, then it is a relatively straightforward matter to take the raw data and count the number of cases that take on each value of the variable. This is likely to be the situation where the variable is a discrete variable with relatively few values. The values of the variable can be counted and presented as a frequency distribution table.

In the notation used above, suppose there are k values of the variable X_i, $i = 1, 2, 3, ..., k$. Let the number of cases that take on value X_i be f_i. With the raw data, it is merely a matter of counting the number of cases taking on each value X_i, and presenting these as the f_i. Note that the sum of all the f_i is the total number of cases, or sample size, n.

4.5.1 Organizing Data into Intervals

When the variable takes on a large number of values, or if the variable is a continuous variable, it is likely that the researcher will have to group the data into intervals. In some cases it may be quite apparent which intervals should be used in order to group data, in other cases the researcher will have to decide on the categories to be presented. While there are no strict rules concerning how to group data, or how many intervals to use, this section contains some guidelines concerning some common approaches that are used to decide on the number and size of intervals.

Example 4.5.1 Grouping Income

Table 4.5 presented a set of incomes grouped into intervals of income such as 0-19, 20-39, etc., where the values of the variable are measured in thousands of dollars. Since there are relatively few cases in this sample, only

5 intervals have been used. Because so few intervals have been used, the intervals for income are relatively large intervals, each representing twenty thousand dollars of income. In the case of income, it is more common to use intervals representing 5 or 10 thousand dollars each. This results in more intervals being used, but this also provides a more detailed view of the frequency distribution of income. For example, Table 4.21 gives a distribution of Saskatchewan tax returns, where the returns are grouped into intervals which each represent five or ten thousand dollars of income.

Research Requirements. In some cases, the intervals to be used may be dictated by the requirements of the research. For example, the variable *years of education* might be grouped into intervals such as 0-8, 9-12, 13-15, 16 and over, representing the divisions between elementary school, secondary school, undergraduate work at a university, and postgraduate work. Table 4.32, later in this chapter, gives similar groupings for years of education of Saskatchewan wives.

A related guideline for grouping is to use the intervals that have been commonly used by other researchers. While the intervals that other researchers have used may not be best for subsequent research, using the same intervals does allow the researcher to compare his or her data with that of other researchers. An example of this is contained in Table 4.35 concerning Saskatchewan farm size. The intervals contained in that table would appear to make no sense for current research. Each interval represents a different number of acres, and the intervals do not represent common values such as 0-100, 100-200, etc., or even the quarter section, half section, section division that is common in Saskatchewan farms. However, the intervals in Table 4.35 are the ones that Statistics Canada has used for many years. These are the intervals into which the Census has grouped Saskatchewan farms, and researchers examining Saskatchewan farm size have little choice but to work with these intervals.

Equal Sized Intervals. Another common practice is to decide the number of intervals that is desired, and create this many intervals, with each interval constructed so that all the interval sizes are all the same. That is, each interval represents the same number of values of the variable X. In order to do this, find the smallest value of the variable, and the largest value of the variable, and then divide the difference between the smallest and largest value by the number of intervals desired. This gives the appropriate size for

the equal sized intervals, and then these intervals can easily be constructed.

Example 4.5.2 Hours of Work

Suppose a researcher has the ungrouped data concerning total hours of work of respondents in Table 4.8, and wishes to group this data into 8 intervals. The smallest value of the variable is 3 hours, and the maximum value is 54 hours. In order to take account of the possibility that in a larger sample, some workers might work fewer thean 3 hours, and perhaps some work more than 54 hours, it might be best to let the intervals range from 1 to, say, 64. Then if the researcher wishes to have 8 intervals, each representing 8 hours worked, the intervals could be 1-8, 9-16, 17-24, 25-32, 33-40, 41-48, 49-56 and 57-64.

While the method of equal sized intervals is suggested by many textbooks, this often creates as many problems as it solves. In the example of total hours worked, many people work exactly 40 hours and considerably more work between 35 and 40 hours per week. By using this method of equal sized intervals, all of these people end up being concentrated in the interval 35-40. This creates too many cases in this interval, and not enough cases in other intervals. The result of this method is to obscure the nature of the overall distribution.

The one great advantage of intervals of equal size is that it is much easier to present a bar chart, or histogram, of the frequency distribution, than if there are intervals of different size. Examples of this are presented in the first part of Section 4.8.

Nature of the Distribution. Rather than choosing intervals representing equal amounts of the variable, or always using the same intervals as other researchers have used, it may be preferable to group data into intervals based on the distribution of the cases. This method means that the researcher should attempt to group data into intervals so that not too many cases are placed in any one interval, or too few in others. This requires the researcher to be sensitive to the nature of the distribution of the data, and construct intervals which illustrate the nature of the distribution as clearly as possible.

Example 4.5.3 Hours of Work

In the example of total hours of work per week, it is best to recognize that the great bulk of workers work 35-40 hours per week. The values of the variable between 35 and 40 should be presented with a more detailed set of intervals than are other values of the variables. The set of intervals and categories suggested here is 0-9, 10-19, 20-29, 30-34, 35-39, 40, 41-49, and 50 and over. Note that since there are so many workers who work exactly 40 hours, this value of the variable is given a category all by itself. The last interval is termed an **open ended interval**, *allowing for the possibility of workers who work an extremely large number of hours. Given the relatively few people with hours worked below 30, these categories might be collapsed even more, say as 0-19 and 20-34.*

As noted earlier, there are no strict guidelines for grouping data, and if a stem and leaf display is used (Section 4.5.3), or if the data is entered on a computer (Section 4.5.4), various groupings of data can be tried until the researcher finds one that is acceptable. The method suggested here is to begin by examining the set of data to be grouped. Then select a set of intervals which allow the distribution of values to be seen as clearly as possible. In order to do this, do not select intervals where there are too many cases in any one interval. For values of the variable where cases are densely concentrated, choose narrower intervals, and where there are relatively few cases, widen the intervals, so that there are fewer but wider intervals. This method produces intervals of different size, and creates some problems for presenting meaningful bar charts or histograms of frequency or percentage distributions. These problems are easily solved, however, and examples are presented in Section 4.8, where histograms for intervals are of unequal size are discussed.

4.5.2 Tallies

One commonly used method of grouping data is the construction of a **tally** of the cases. To begin this technique, it is first necessary to pick the intervals into which the data is to be grouped, and then use a tally sheet to count the number of cases which fall into each of the categories or intervals. Once the tally sheet has been completed, then the frequency distribution table can be constructed by counting the number of cases in each category as recorded on the tally sheet.

Example 4.5.4 Grouping Hours of Work in a Tally

Table 4.8 and Figure 4.1 show how the ungrouped data can be grouped by use of a tally. The frequency distribution table based on this tally is then presented in Table 4.9.

The first step in obtaining a tally is to decide the intervals into which the data is to be grouped. In this case, it was decided to use the intervals 0-9, 10-19, 20-29, 30-34, etc. The next step is to go through the list of values of the variable in Table 4.8 and record each value once in the appropriate interval of the tally, as shown in Figure 4.1. The tally sheet in Figure 4.1 groups the number of cases into groups of 5. These can then be easily totalled. This totalling is carried out and presented in Table 4.9. The tally sheet shows that there are 2 cases between 0 and 9, 2 cases between 10 and 19, and so on. These totals for each of the intervals are presented as frequencies of occurrence (f) of the values of the variable in the frequency distribution table of Table 4.9.

50	35	37	37	40
10	40	40	36	35
40	50	35	23	50
30	40	40	37	40
4	40	38	25	40
54	37	40	50	37
48	39	37	40	40
30	40	40	48	40
50	40	40	16	37
3	32	40	40	40

Table 4.8: Total Hours of Work - Ungrouped Data

The summary distribution of Table 4.9 allows anyone examining the data concerning the variable hours of work (X) to get a clearer idea of how the values of X are distributed. While the ungrouped data of Table 4.8 contains all the data, there are too many numbers in that table. By constructing the tally and the table of the frequency distribution, a much better idea of what the distribution is really like can be obtained. From the table of the frequency distribution, the most common value of the variable X is 40 hours, with considerable numbers of people (13) also working between 35 and 39 hours. The number of people in each of the other categories is considerably fewer, although there are 6 people who work 50 or more hours a week at

0-9	11
10-19	11
20-29	11
30-34	111
35-39	$\cancel{1111}$ $\cancel{1111}$ 111
40	$\cancel{1111}$ $\cancel{1111}$ $\cancel{1111}$ $\cancel{1111}$
41-49	11
50+	$\cancel{1111}$ 1

Figure 4.1: Tally for Total Hours of Work

their job or jobs.

The only major problem associated with the method of the tally is that once the tally has been completed, the exact values of the variable X are lost in the tally sheet. That is, the number of cases in each group is known, but the exact values of the variable are not given in the tally, or in the table of the frequency distribution. In the example in Table 4.9, there are 13 cases in the interval 35-39, but when looking at this table, it is not known whether these values are 35, 36, 37, 38 or 39. As a result of this, if the grouping that was used to construct the tally does not seem all that appropriate, and new intervals are deemed necessary, then it is necessary to begin the tally from the beginning again. The stem and leaf method of the next section avoids this problem.

4.5.3 Stem and Leaf Displays

A stem and leaf display is a technique for organizing ungrouped data. The aim is to present a display such that all the data is still present in the display, but the values of the variable are presented in order, from the lowest to the highest value. This technique lists all the values of the variable X, beginning at the smallest value of X and proceeding through to the largest value of X.

Hours of Work (X)	Number of Cases (f)
0-9	2
10-19	2
20-29	2
30-34	3
35-39	13
40	20
41-49	2
50 and over	6
Total	50

Table 4.9: Frequency Distribution of Hours of Work

Where there is more than one case taking on a particular value of X, that value is recorded the same number of times that it occurs in the original list.

A stem and leaf display begins by organizing the values of the variable into groups of 10, usually 0-9, 10-19, 20-29, and so on. This is called the **stem** of the display, and is represented on the display by only the first number of the groups of 10. That is, 0-9 is represented by 0 on the stem, 10-19 is represented by 1 on the stem, 20-29 is represented by 2 on the stem, and so on. These values are placed one above another in a vertical column.

The actual values of the variable are then placed to the right of this stem, as the **leaves** of the display. Each original value of X is placed in its appropriate row. All the 10s are placed in the row labelled 1 on the stem, all the 20s in the 2 row of the stem, and so on. The second, or units, digit of the number is placed in the appropriate row. As a first stage to constructing the display, one proceeds through the list of ungrouped data, placing each value of the variable in the appropriate row. Once this has been completed, the stem and leaf display presents all the data, organized into the groups 0-9, 10-19, 20-29, and so on. The last stage of the process is to order the values in each row from low to high. At the end, the final stem and leaf display then presents all the values of the variable in order, from the smallest to the largest values.

In the stem and leaf display, each value of the variable is placed at equal

distance from each other value. The display then allows one to observe the relative number of cases in each group of 10. The number of cases in each group of 10 is easily counted in this display, and a table of the frequency distribution can be prepared. In addition, since all the original values of the variable are present, but now placed in order, the data can easily be organized into any set of intervals which is desired.

Example 4.5.5 Grouping Hours of Work in a Stem and Leaf Display

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

Figure 4.2: Initial Stem and Leaf Display for Total Hours of Work

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

Figure 4.3: Ordered Stem and Leaf Display for Total Hours of Work

The ungrouped data on hours of work from Table 4.8 is presented in a stem and leaf display in Figure 4.2. The first step involved is to note that the largest value of the variable is 54 and the smallest value is 3. Since the data is to be organized into groups of 10, this means categories 0 (for 0-9), 1 (representing 10-19), 2 (20-29), 3 (30-39), 4 (40-49) and 5 (50-59). These are placed in a vertical column on the left. Then the second, or units, digit of each value of hours worked is placed in the appropriate row of this display. The first value is 50, and this value appears as the first 0 is the 5 row of the display. Proceeding down the first column of the data, the next value encountered is 10, and this appears as a 0 in the 1 row. Subsequent values

are 40 as a 0 in the 4 row, 30 as a 0 in the 3 row, 4 as a 4 in the 0 row, 54 as a 4 in the 5 row.

Proceed systematically proceeds through the table in this manner, until all the 50 values for total hours of work have been entered in the display. This completes the initial stage of the display. At this point, the values are ordered into groups of 10, and by glancing at the rows of this display, one can see which values are most frequent. In Figure 4.2, there are considerably more values in the 40 row, showing that more people work 40-49 hours than any other grouping of 10 hours. The 30 row is also large, meaning that many workers work between 30-39 hours. There are considerably fewer who work less than 30 hours, and not all that many who work 50-59 hours.

The next stage in preparing the display is to order the values within each row. For example, in the first row the values were 4 and 3, representing 4 hours and 3 hours. Placed in order in Figure 4.3, these appear as 3, followed by 4. Each row of the table is ordered in this manner, and the complete set of ordered values is given in Figure 4.3. With this display, the data can now be grouped however the researcher desires. The grouping of Table 4.9 can be constructed and the number of cases in each category can be counted and entered into this table of the frequency distribution. However, the researcher could decide to use somewhat different intervals, and the number of cases in each of these can be easily determined by using the ordered stem and leaf display.

The example of the stem and leaf display shows the flexibility of this approach. None of the information contained in the original list of ungrouped data is lost, and the ordered display has each of these values placed in order. While there are other ways in which the data could be ordered, the stem and leaf display is a relatively quick and efficient way of sorting and ordering a list of ungrouped data.

The display is also useful in determining the various measures of central tendency discussed in the next chapter. The mode, or most common value of the variable, can easily be determined by looking through the stem and leaf display. In Example 4.5.5 the value 40 occurs much more frequently than any other value of hours worked, so that the **mode** of hours worked in 40 hours, the most common number of hours per week that workers work on the job. The **median** is the middle value, so that one half of the values of the variable are less than or equal to this value, and the other have are greater than or equal to it. Since the stem and leaf display orders the values from

low to high, it is only necessary to count from the smallest value through to the point where half the values have been accounted for. In the last example, there are 50 values of total hours worked per week, so that the halfway point is the 25th and 26th values, and both of these are 40. The median hours worked is thus 40, meaning that half of the workers work 40 or less total hours per week, and the other half work 40 or more hours per week.

Example 4.5.6 Grouping Household Income in a Stem and Leaf Display

18	22	9	46	113
49	16	101	11	43
5	85	4	9	76
11	45	33	23	13
53	62	19	14	9
16	24	10	10	9
23	56	21	136	32
36	6	71	68	130
9	64	28	11	25
9	24	50	22	35

Table 4.10: Incomes of 50 Saskatchewan Families, 1988

The list of numbers in Table 4.10 comes from a sample of 50 Saskatchewan households. This sample is, in turn, drawn randomly from a larger sample, the 1989 Survey of Consumer Finances. The Survey of Consumer Finances is an annual survey conducted across Canada by Statistics Canada. It is based on the same sampling methods and procedures that are used in Statistics Canada's Labour Force Survey (see Chapter 2).

The numbers shown in Table 4.10 represent total family income in thousands of dollars, from all income sources, for each of the 50 families in this sample. These incomes refer to incomes in 1988, and are in 1988 dollars. The numbers are first presented as a list of data and then they are organized in a stem-and-leaf display. The initial stem and leaf display is given in Figure 4.4 and in the ordered stem and leaf display is in Figure 4.5. The data are then grouped into a frequency distribution table in Table 4.11. Finally, the data are graphed in a **histogram** *(see Section 4.8) in Figure 4.6.*

0	9	5	4	9	9	9	6	9	9		
1	8	6	1	1	3	9	4	6	0	0	1
2	2	3	4	3	1	8	5	4	2		
3	3	2	6	5							
4	6	9	3	5							
5	3	6	0								
6	2	8	4								
7	6	1									
8	5										
9											
10	1										
11	3										
12											
13	6	0									

Figure 4.4: Unordered Stem and Leaf Display

This example is much the same as the earlier example except that considerably more categories must be used. The smallest income encountered in the initial list of ungrouped values of income is 4 thousand dollars, with the maximum value being 136 thousand dollars. As a result, there have to be 13 rows to the stem and leaf display, in order to account for all the possible values.

When the data has been grouped into the groups of 10 units of family income, the nature of the income distribution can be relatively easily described. From either the table or the diagram, it can be seen that the bulk of families have incomes of less than $30,000, with $10,000 to $19,000 being the single most common interval. As income increases above this point, the number of families declines, at first slowly, and then somewhat more rapidly. Above $40,000, the number of families is relatively small in each of the intervals, with some of the intervals having no families. This sample of 50 families also has several families with over $100,000 annual income. However, these four families are quite spread out over the $100-140 thousand dollar range.

The histogram in Figure 4.6 is fairly typical of the shape of distributions of income. Most individuals or families have low to middle level incomes. As income increases, there tend to be fewer and fewer families. At very high incomes, there are some individuals or families, but relatively few. That is, income distributions tend to peak at low to middle income levels, decline

0	4	5	6	9	9	9	9	9	9		
1	0	0	1	1	1	3	4	6	6	8	9
2	1	2	2	3	3	4	4	5	8		
3	2	3	5	6							
4	3	5	6	9							
5	0	3	6								
6	2	4	8								
7	1	6									
8	5										
9											
10	1										
11	3										
12											
13	0	6									

Figure 4.5: Ordered Stem and Leaf Display

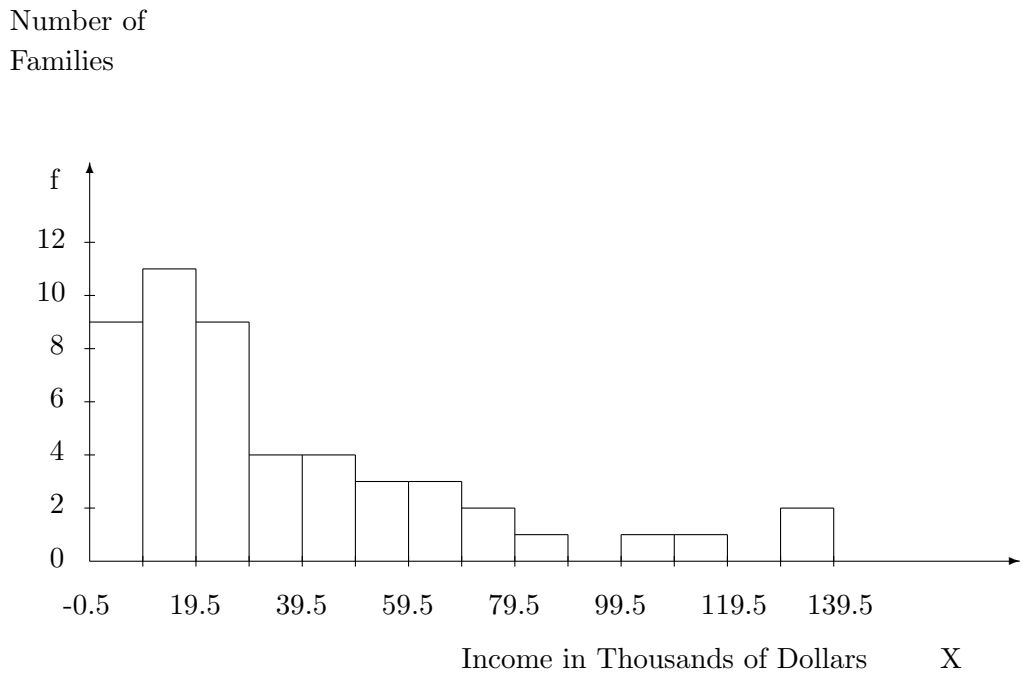

Figure 4.6: Histogram of Income - 50 Saskatchewan Families

Income in Thousands of Dollars	Real Class Limits	f
0-9	-0.5-9.5	9
10-19	9.5-19.5	11
20-29	19.5-29.5	9
30-39	29.5-39.5	4
40-49	39.5-49.5	4
50-59	49.5-59.5	3
60-69	59.5-69.5	3
70-79	69.5-79.5	2
80-89	79.5-89.5	1
90-99	89.5-99.5	0
100-109	99.5-109.5	1
110-119	109.5-119.5	1
120-129	119.5-129.5	0
130-139	129.5-139.5	2
Total		50

Table 4.11: Frequency Distribution of Income, 50 Saskatchewan Households

fairly quickly, but continue for quite a distance to the right, until the very highest income is accounted for.

In this example, there are relatively few cases above $30,000 annual income. In presenting this data, it might be useful to begin to combine intervals, so that a smaller table giving the basic picture of the frequency distribution can be presented. This is done in Table 4.12. There the intervals of $40,000 and over are combined in various ways, so that the distribution can be presented with fewer intervals. While this results in some loss of information in the presentation, the general pattern of the data can be seen more easily with this smaller set of intervals. It is common to combine intervals where there are relatively few cases, so that not too many intervals are used in the presentation. Combining intervals in this manner allows for some economy of presentation, but also presents some problems of presen-

tation. These problems and how to present data of this sort are dealt with in Section 4.8 and the histogram of this distribution is presented in that Section.

Income in Thousands of Dollars	Real Class Limits	f
0-9	-0.5-9.5	9
10-19	9.5-19.5	11
20-29	19.5-29.5	9
30-49	29.5-49.5	8
50-69	49.5-69.5	6
70-99	69.5-99.5	3
100-139	99.5-139.5	4
Total		50

Table 4.12: Frequency Distribution of Income, 50 Saskatchewan Families

4.5.4 Coding Data for Computers

While tallies or stem and leaf displays are useful for small data sets where there are not too many variables, it becomes extremely time consuming and inefficient to classify larger data sets in this manner. Given the widespread availability of computers, it is usually more efficient to code the data so it can be entered onto a computer. Then it can be stored on the computer, and the data analyzed using a statistical program. In addition to being more efficient, such a procedure has the the advantage of allowing the researcher to retain all the original data and, during the the process of data analysis, decide how to group and organize the data. This section contains a very brief description of this process.

The first stage in the process of coding the data is to attach numbers to each of the possible values of the variable. These numbers may be the actual values of the variable, or they may be only numerical codes, attached to different values of the variable in order to keep track of which values are which. For example, a variable such as *age* is quantitative in form and

the codes for the variable are the actual values of age. In a case such as *sex*, arbitrary numerical codes such a 0 for males and 1 for females may be adopted.

Once codes have been attached to all the potential values of a variable, these numerical values are entered onto the computer. The procedure for doing this, and the format to be used, will differ depending on which computer and statistical program is used. Once the data has been entered on the computer, then the statistical program can be used to group the data, and analyze the data. On this basis, the results can then be presented.

Example 4.5.7 Small Sample with Several Variables

In Example 4.2.2, data for the values of several variables for a small random sample of 7 respondents was presented. For Table 4.2 the variables `AGE`, `GMP`, `THRS` *and* `FIN` *are already numerical, and can be entered onto the computer without any alteration. In the case of* `FIN`*, the comma would ordinarily be deleted when entering the values onto the computer, so that 42,500 would be entered as 42500, and so on.*

The other variables are either nominal or ordinal scale, and do not have any numerical values which are naturally attached to them. In the case of each of these variables, it is necessary to attach numerical values to the set of possible outcomes that each of these variables can take on. In addition, it is always useful to keep track of each respondent by entering an identification number or case number for each respondent.

The codes attached to each of the non-numerical values in Table 4.2 is given in Table 4.13. Note that a code of 9 has been designated in some cases for those who did not respond. Also note that for the variable representing political preference, extra codes have been designated for undecided, or support of a political party other than one of the three major parties.

If the coding scheme of Table 4.13 is used, then the data might be entered onto the computer as shown in Table 4.14. The method used here is to enter each case, or each respondent, on a separate line, with the coded values of each of the variables given for each case. Since the computer has to be told how to read this set of numbers, the statistical program will contain a list of the names of the variables, and the column numbers in which each variable is entered.

For example, with the SPSS or SAS program, there would be a set of lines in the program something like:

```
ID 1-2 SEX 4 AGE 6-7 PARTY 9 GMP 11-14 THRS 16-17
CLASS 19 FIN 21-26
```

Variable Name	*Values of Variable*	*Numerical Codes*
SEX		
	Male	*1*
	Female	*2*
PARTY		
	LIB	*1*
	NDP	*2*
	PC	*3*
	OTHER	*4*
	UNDECIDED	*5*
	WOULD NOT VOTE	*6*
	No Response	*9*
CLASS		
	Upper	*1*
	Upper-Middle	*2*
	Middle	*3*
	Working	*4*
	Lower	*5*
	Not Stated	*9*

Table 4.13: Codes for Variables in Example 4.2.2

*This tells the computer that the first two columns are reserved for the identification number of the respondent (*ID*), column 4 for* SEX*, columns 6 and 7 for the age of respondent, and so on. Spaces may be left between the codes entered, as shown here, or spaces may be left out. So long as the list of variables and column numbers for those variables is consistent with this list, then the computer will be able to read the data properly.*

When entering the data on the computer in this manner, no decisions need to be taken concerning how the data is to be grouped. Those decisions can come later, during the stage of data analysis. All that needs to decided when coding and entering the data on the computer is what the potential codes are. When entering the data on the computer, enough space must be left so that all the possible values can be included. In this example, two

```
1 2 32 2 1300 50 2  42500
2 2 33 2 1950 25 4  27500
3 1 34 2 2500 40 3  37500
4 2 34 2 1850 40 3  52500
5 1 46 3 5000 50 3 100000
6 2 34 2  700 16 4  37500
7 2 53 1 3125 40 3  62500
```

Table 4.14: Coded Data or Computer Entry from Example 4.2.2

columns (1-2) have been left for the identification number. This allows up to 99 cases to be entered. If there are 100 cases or more, but less than 1,000, then 3 columns are required for the identification number.

4.6 Proportions and Percentages

Distributions are often presented as proportional or percentage distributions, rather than as frequency distributions. Distributions in this form are sometimes more meaningful, because the table or diagram shows the proportion or percentage of the cases which take on each value of the variable, and these allow the reader to quickly observe the relative number of cases in each category.

In order to determine the proportional or percentage distribution, it is first necessary to begin with the frequency distribution. The frequency distribution gives the number of cases, f, which have each value of the variable, or are in each interval. The sum of all the frequencies is the number of cases, or sample size, n.

Definition 4.6.1 The **proportion** of cases in each category or interval is the number of cases in that category or interval divided by the total number of cases n.

If the proportion of cases is calculated for each category or interval, the sum of all the proportions should equal 1. That is, once all the n cases are accounted for, the sum of the proportions is $n/n = 1$.

Percentages are computed by multiplying the respective proportions by 100%

Definition 4.6.2 For any category or interval, the **percentage** of cases is the number of cases in that category or interval divided by the total number of cases (n), and multiplied by 100%.

The sum of all the percentages must total 100%, because once all the cases have been accounted for, this is 100 per cent of the cases. If proportions have not first been calculated, the percentage of cases in any category or interval is the number of cases in that interval, divided by the total number of cases, and multiplied by 100 per cent.

Example 4.6.1 Distribution of Number of People per Household

A distribution of the number of people per household was given in Table 4.6. In Table 4.15 this distribution is shown again and presented as a proportional and a percentage distribution. The column headed p is the proportional distribution and the column headed P is the percentage distribution.

X	f	p	P
1	155	0.165	16.5
2	286	0.305	30.5
3	164	0.174	17.4
4	223	0.237	23.7
5	86	0.091	9.1
6	21	0.022	2.2
7	5	0.005	0.5
8	1	0.001	0.1
Total	941	1.000	100.0

Table 4.15: Frequency, Proportional and Percentage Distributions of Number of People per Household

The first two columns of Table 4.15 are exactly the same as Table 4.6. The frequencies of occurrence of each value of the variable X, number of people per household, are given in the f column. The sum of this column is $n = 941$, the total number of households. The proportions are calculated by dividing the number of cases in each category for X by n. For example, the

proportion of respondents with exactly one person per houshold is $155/941 = 0.165$. *The number of respondents with* $X = 2$ *people per household is* $286/941 = 0.305$. *All the other proportions can be calculated in a similar manner, and when they are added up, the total of the p column is 1.000, meaning that all the cases have been accounted for.*

The percentages can now be computed by multiplying the proportions by 100%. For $X = 1$, *the percentage is* $0.165 \times 100\% = 16.5\%$. *If the proportions have not already been calculated, the percentages can be computed directly. The percentage of respondents with* $X = 2$ *people per household is* $286/941 \times 100\% = 30.5\%$. *For* $X = 3$, *the percentage of cases is* $164/941 \times 100\% = 17.4\%$, *or if the proportion of cases with* $X = 3$ *has already been calculated as 0.174, the percentage is* $0.174 \times 100\% = 17.4\%$. *The sum of the column of percentages is 100%, representing the sum of all the cases.*

Note that **per cent** means **per one hundred**, so that the individual percentages represent the number of cases per one hundred of the total number of cases.

Notation for Proportions and Percentages. In Section 4.4, notation to describe a frequency distribution was discussed. This notation can be adapted for proportional and percentage distributions. Again, let the variable X have k possible values $X_1, X_2, ..., X_k$ with respective frequencies $f_1, f_2, ..., f_k$. This can be alternatively be stated by noting that the variable X takes on values X_i where $i = 1, 2, 3, ..., k$, with respective frequencies f_i, over the same set of k categories. The sum of the frequencies f_i is the sample size n.

The proportions are labelled p, where p_i is the proportion of cases which take on value X_i. The proportions are

$$p_i = \frac{f_i}{n}$$

that is, for each X_i, the proportion is the frequency of occurrence, f_i, of the value of the variable divided by the total number of cases n. The percentages are labelled P, where P_i is the percentage of cases which take on value X_i. The percentages can be determined as follows:

$$P_i = p_i \times 100 = \frac{f_i}{n} \times 100$$

X_i	f_i	p_i	P_i
X_1	f_1	p_1	P_1
X_2	f_2	p_2	P_2
X_3	f_3	p_3	P_3
X_4	f_4	p_4	P_4
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
X_k	f_k	p_k	P_k
Total	n	1.000	100.0

Table 4.16: Notation for Proportions and Percentages

that is, the percentage of cases which take on the value X_i is the proportion of cases which take on this value, multiplied by 100. All of this notation is summarized in Table 4.16.

4.7 Graphing Data

It is often useful to present the data in the form of a diagram, rather than a table. A diagrammatic presentation often allows the nature of the distribution to be easily seen. These diagrams may be pie charts, maps, line charts, bar charts or histograms. In each case, the aim of the presentation is to show the manner in which the variable is distributed, that is, how many cases take on each of the values of the variable. The diagram should present this in as honest and as accurate a form as possible.

Throughout the rest of this textbook, various diagrams of frequency or percentage distributions are presented. While each person presenting data may use a somewhat different form of presentation, there are some guidelines concerning which are the most useful forms of presentation for statistical analysis. This section presents some of the more common forms of diagrams used in Statistics. In presenting data in diagrammatic form, it is important to be aware of the conventions used. Much of this section is concerned with

these conventions. In this textbook, pie charts and maps are not discussed. Most of statistics works with line or bar charts, and these are the types of diagrams discussed in more detail here.

For the various diagrams of frequency distributions, there are some general guidelines which are to be followed when presenting the data. The value of the variable is most commonly shown along the horizontal axis. The X-axis of a graph is used for representing the values of the variable. Since the variable is often given the symbol X, this means that the variable X is measured along the ordinary X-axis.

The vertical, or Y-axis, is most commonly used to represent the frequency of occurrence of the variable. That is, the number of cases, f, which take on each value of the variable, is represented by the vertical, or Y-axis. If the data is presented as a percentage, proportionate, or relative frequency distribution, then the vertical axis is used to represent this instead.

While these rules are not absolutely necessary, these are the conventions normally followed in Statistics, and in most presentation of data. These will be the conventions followed throughout this textbook. The two most common methods of presenting data are line and bar charts. These are discussed in the following sections.

4.7.1 Line and Bar Charts

When graphing distributions in Statistics, the values of the variable X are conventionally placed along the horizontal axis. The frequencies of occurrence f are place on the vertical axis. The lines or bars are drawn vertically at each value of the variable, and the heights of these lines or bars give the frequency of occurrence of the variable for each of the values of the variable.

In the case of a variable which is no more than nominal, there is no meaningful measure of distance between the values of the variable. In this case, the values of the variable have names but these values are not numerical in nature. As a result, the amount of space left between each value of the variable on the horizontal axis is more or less arbitrary. Example 4.7.1 shows how the nominal scale of political preference can be graphed as a line chart.

In contrast, when a variable has an interval or ratio scale, or even an ordinal scale, it is necessary to let equal numerical values of the variable be represented by equal distances along the X axis. Example 4.7.2 gives a line and a bar chart for the distribution of the variable *number of people per household*, a ratio level scale.

If a variable has a relatively few set of values and these values are discrete, then the data may be most clearly presented as a line chart.

Definition 4.7.1 A **line chart** gives the values of the variable along the horizontal axis, and the frequency of occurrence on the vertical axis. A line is drawn vertically at each value of the variable, with the vertical height of the line representing the frequency of occurrence of that value of the variable.

In a line chart, the vertical, or Y, axis gives the frequency of occurrence of the values of the variable. These are the number of cases (f) taking on each value of the variable. The lines are drawn vertically so that the height of each line represents the number of cases taking on each value. The relative height of each line then represents the relative frequency of occurrence of different values of the variable. Values which have higher lines associated with them occur more frequently than values with shorter lines.

A **bar chart** presents vertical bars, rather than vertical lines. As in the case of a line chart, the values of the variable are shown along the horizontal axis, and the frequency of occurrence is shown on the vertical axis.

Definition 4.7.2 A **bar chart** presents a distribution with vertical bars, with the height of the bar at each value of the variable representing the frequency of occurrence of that value of the variable. In the case of a proportional or percentage distribution, the bars have heights corresponding to the respective proportions or percentages.

A bar chart for a discrete variable, of any level of measurement, may be presented with spaces between the bars, as in Example 4.7.1. If the variable is continuous, no spaces should be left between the bars, so that the continuous nature of the variable can be seen in the diagram. Bar charts for continuous variables are usually called histograms, and guidelines for presenting these are discussed in Section 4.8. In Example 4.7.2, the bars are drawn with no spaces betwen the bars, even though the variable is a discrete one.

Example 4.7.1 Distribution of Political Preference

In this example, the frequency distribution of the political preferences of 50 Regina adults are presented first as a line chart in Figure 4.7, and then as bar chart in Figure 4.8. Since political preference has only a nominal scale of measurement, there is no meaningful measure of distance between the different political parties. Even so, these have been spaced equally along

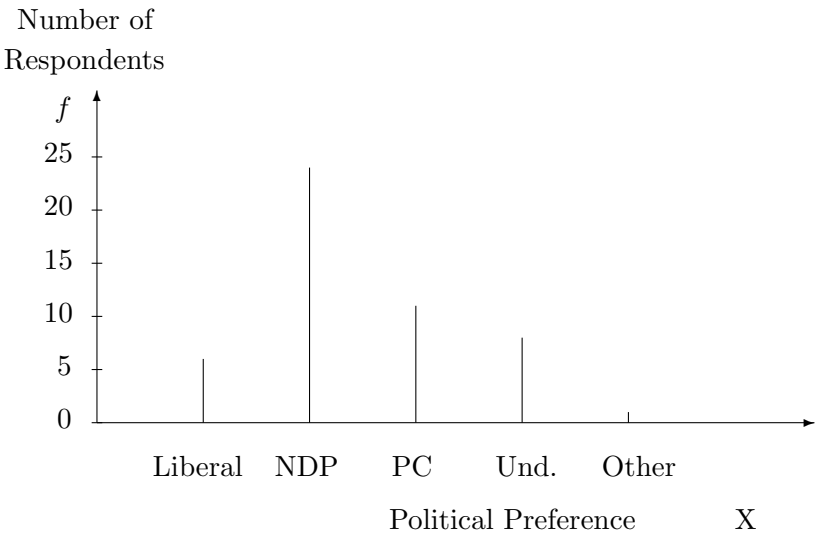

Figure 4.7: Line Chart of Political Preference, 50 Regina Respondents

the X axis. There is also no order in which the different parties must be presented on the diagram, but here they are presented in the same order as in Table 4.4. For each political preference, a vertical line is drawn, with height equal to the frequency of occurrence of that political preference, as given in Table 4.4.

There are 24 respondents who support the NDP, so the line has height $f = 24$*, when* X *takes on the value NDP. For the PCs, there are 11 supporters, so that the height of the line for the PCs is 11. Lines are drawn for each value of* X *until all the values have been accounted for.*

At a glance, anyone examining this diagram would get the idea that the variable is discrete, and would be able to tell the relative support being expressed for each political party. The NDP is clearly the single most common party supported, with the PCs second. There are almost as many undecided as there are PC supporters, and there are relatively few supporters of the Liberals or other parties.

The bar chart in Figure 4.8 presents exactly the same view as does the line chart. Bars of arbitrary width have been drawn vertically at each political preference. The height of the bar represents the frequency of occurrence of each political preference. The bars here are drawn with a space between

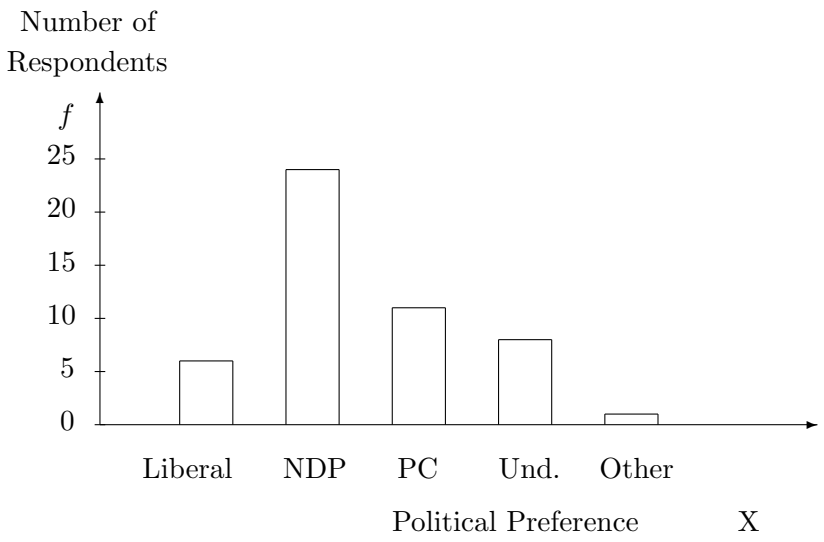

Figure 4.8: Bar Chart of Political Preference, 50 Regina Respondents

them in order to convey the impression that the variable is discrete.

Example 4.7.2 Distribution of Number of People per Household.

The distribution of the number of people per household was given in Table 4.6. If these are presented as a percentage distribution, or a proportional distribution, the results are presented again in Table 4.17. The percentage distributions are graphed as a line chart in Figure 4.9. and as a bar chart in Figure 4.10.

In this example, the variable (X) is a discrete, interval or ratio scale of the number of people per household. The distances between values of X are equal in amount, so that they have been graphed as equal distances between 1 and 2, 2 and 3, and so on.

The vertical (or Y) axis gives the percentage of cases with each of the values of the variable. These are the percentages (P), as they have been calculated in Table 4.17. The proportion of cases taking on each value of the variable are presented first in this table and then multiplied by 100 to

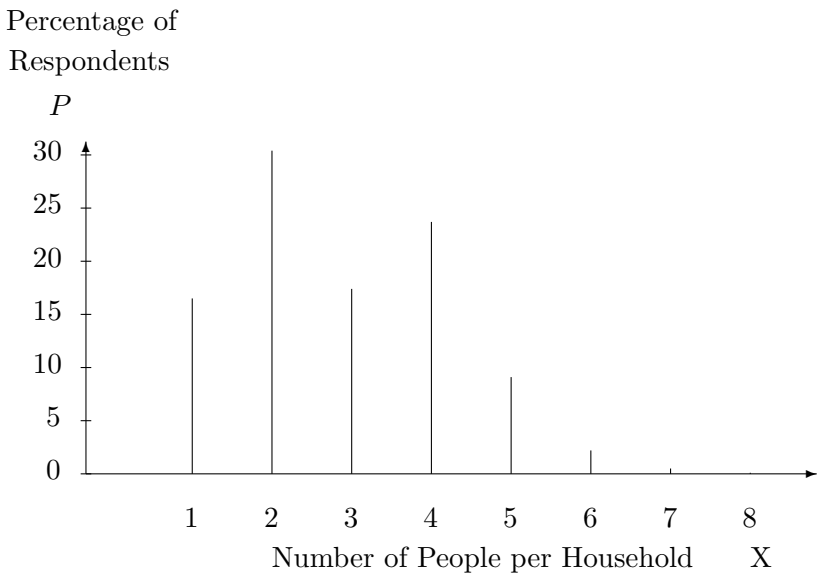

Figure 4.9: Line Chart of Percentage Distribution of Number of People per Household, $n = 941$ Regina Households

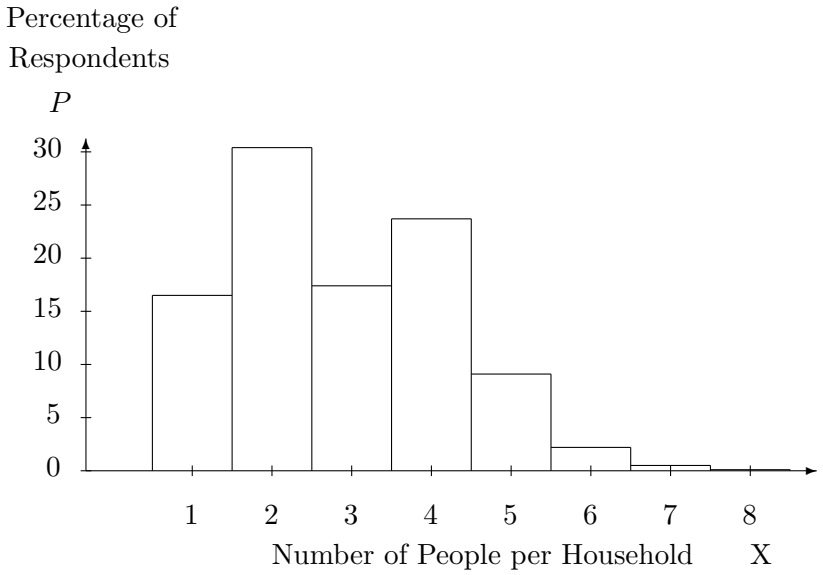

Figure 4.10: Bar Chart of Percentage Distribution of Number of People per Household, $n = 941$ Regina Households

X	f	p	P
1	155	0.165	16.5
2	286	0.305	30.5
3	164	0.174	17.4
4	223	0.237	23.7
5	86	0.091	9.1
6	21	0.022	2.2
7	5	0.005	0.5
8	1	0.001	0.1
Total	941	1.000	100.0

Table 4.17: Distribution of People per Household

determine the percentage distribution. The percentage of the 941 households which take on each value of the variable are presented as vertical lines. The height of each vertical line is the percentage of cases which take on each of these values. The heights of the first four lines are all relatively high, with over 30% of the cases taking on the most common value of $X = 2$ people per household. 4 people per household is the next highest line, representing just under 25% of the cases. The nature of the distribution can be quickly seen by examining this diagram. Most respondents live in households of 4 or less people, with very few having more than 6 household members.

4.7.2 Some Rules Concerning Rounding of Data

If a discrete variable with relatively few cases is being presented, the set of all values for the variable X is presented. However, if the variable X is continuous, the values of the variables which are presented will most likely be rounded, rather than given to many decimal points. This section contains some guidelines concerning the rounding of numbers. These guidelines are also important in Chapter 5, where the calculation of summary statistics often results in values with a considerable number of decimals.

Rounding to Nearest Value. The usual rule concerning rounding is that a number is rounded to the nearest value. For example, if income is

reported in dollars, then the nearest dollar will be the value reported. If a person's income is $32,436.56, then this would be rounded to $32,437, because the value of $32,436.56 is closer to $32,437 than it is to $32,436. In this method, any value just over $32,436.50 is rounded to $32,437, and any value just under $32,436.50 is rounded to $32,436.

If the income is exactly $32,436.50, then this value could be rounded up to $32,437 or rounded down to $32,436, and it does not usually matter too much which is done. But in order to avoid always rounding up, or always rounding down, it is best to adopt a consistent rule in such circumstances. One common rule is that when the value of the variable falls exactly midway between the two values to which it could be rounded, always round to the **even value**. With this rule, if $32,436.50 is rounded to the nearest dollar, it would be reported as $32,436, the even value, rather than $32,437, the odd value.

Some examples of rounding are given in Table 4.18. This table shows how data may be rounded to several decimals, or may be rounded to the nearest integer, or unit, or even to the nearest ten or 100. In all cases, rounding to the nearest value is the method used. Where the value is midway between the two closest values, it is rounded to the even number. For example, to two decimals, 3.145 is rounded to 3.14, rather than 3.15.

Where the original value is not reported to the accuracy indicated, extra 0s can be added. For example, if a number was exactly 87.5, this could be reported as 87.500 or 87.50. However, these extra 0s should be added only if the original value is exactly 87.5. If the average of exactly 85 and exactly 90 were desired, this is exactly $(85 + 90)/2 = 87.5$, so that it may be reported as 87.50000. However, if the original values were 85 and 90.1, for an average of $(85 + 90.06)/2 = 175.061/2 = 87.53$ and this had already been rounded to 87.5, then extra 0s should not be added.

Significant Figures. Another concept that relates to rounding is the number of **significant figures** in the value. Roughly speaking, the number of significant figures is the number of nonzero digits that appear in the value. Where the number is a decimal, and less than one, then only the digits after the initial zeroes are considered to be significant figures. For example, in the case of 0.004, there is only 1 significant figure (the 4). For 0.026, there are 2 significant figures (the 2 and 6), but for 0.03 there is only one significant figure. For values above zero which have been rounded to the nearest 10 or

Original Value	Rounded to: 3 Decimals	2 Decimals	1 Decimal	Unit	10s	100s
23.1866	23.187	23.19	23.2	23	20	0
768.34552	768.346	768.35	768.3	768	770	800
3.145	3.145	3.14	3.1	3	0	0
0.003798	0.004	0.00	0.0	0	0	0
0.02563	0.026	0.03	0.0	0	0	0
87.5	87.500	87.50	87.5	88	90	100
47,456.3	47,456.300	47,456.30	47,456.3	47,456	47,460	47,500

Table 4.18: Examples of Rounding

100, those zeroes should not be considered significant either. For example, 47,500 has only 3 significant figures (4, 7, and 5), while 87.5 has 3 significant figures.

When reporting data which has been rounded, it is most common to present at least 2 significant figures. For example, when reporting family incomes, the values might be rounded to the nearest $1000 and reported as 37 thousand, 43 thousand, 8 thousand, etc. It would be preferable to report these to at least the nearest 100 dollars, and this would result in 3 significant figures, with for example $37,300, $42,900, $8,100, and so on. Too much accuracy is lost if these are rounded to the nearest $10,000 and reported to only one significant figure as $40,000, $40,000 and $10,000.

For values less than 1, two or three significant figures should also be used. For example if the values are 0.00463, 0.01578, and 0.00034, then it might be best to use all 5 of these decimals, meaning that each value has at least two significant figures. If these values are rounded to 4 decimals, then the values are 0.0046, 0.0158 and 0.0003. The first two values have 2 and 3 significant figures, respectively, while the last value has only 1 significant figure. This is acceptable if not too many values have only 1 significant figure.

While the rules concerning how many significant figures to use when rounding are not exact, there are several general guidelines which can be followed.

1. **Accuracy of the Data**. Be guided by the accuracy of the data as it is originally reported. If the raw, ungrouped data have already been rounded to two significant figures, then any subsequent data is unlikely

to be accurate to more than two significant figures. The data for the 7 respondents in Table 4.2, give several examples. There total hours worked per week, (`THRS`), appears to be rounded to the nearest hour, and there would appear to be two significant figures. (The same would be true of the list of total hours of work for the 50 cases in Table 4.8). Family income, (`FIN`), appears to be rounded to the nearest hundred dollars, and this produces three significant figures for family income.

In contrast, where the data is reported more precisely, then averages, or other summary statistics, can be reported to more decimals. For the 941 respondents in Table 4.15, the values of the variable X, number of people per household, appear to be reported to only 1 significant figure. However, these values are precise. That is, $X = 1$ means that there is exactly one person per household, and in this case $X = 1$ means exactly 1, so that this could be reported as $X = 1.000000$, with any number of decimals. In cases like this, where there has been no rounding, then any number of decimals can later be used, and the results can legitimately be reported to any number of decimals. In Table 4.15, the average or mean number of people per household is 2.8799149. The average can be reported to this many decimals if desired, because the X values are perfectly precise in the original data, and no rounding has occurred.

2. **Use of the Data**. While the average number of people per household can legimately be reported as 2.8799149 based on the data in Table 4.15, it is difficult to imagine why anyone would need the average to be reported to this many decimal places. It would be more common to report the average number of people per household as 2.88, rounded to 2 decimals, or even as 2.9, rounded to 1 decimal (2 significant figures). An average reported with this accuracy would make more sense, because two or, at most three, decimals are likely to satisfy most ordinary purposes.

   While no hard and fast guidelines can be made concerning the number of decimals or significant figures that are needed, be guided by common sense. For most social science research, if family income is reported to the nearest one hundred dollars, this is sufficient accuracy. For tax purposes, Revenue Canada wants to know income, and all its components, to the nearest penny, but for most research purposes, such accuracy is neither needed, nor warranted. In terms of

population of cities, the population to the nearest thousand people is probably sufficient accuracy. For small towns though, population figures might be better reported to the nearest 10 people, or even to the nearest person. For age, the nearest year is sufficient accuracy for many research purposes. However, if more accuracy is desired, birth date can be obtained, and then age could be reported to the nearest day.

3. **Approximations and Non Sampling Errors**. When using data, many approximations are made by researchers. In addition, as noted in Chapter 2, there are many errors involved in the production of data. Because of this, most data is not perfectly accurate.

   For example, in Table 4.3, the percentage of Canadians supporting each political party is reported to the nearest 1 percentage point, to two significant figures. In producing this data, there are likely some non sampling errors, such as the wording of questions, interviewer error, nonresponse, and so on. In addition, there is considerable sampling error. For these reasons, these percentages are certainly not accurate to nearer than 1 percentage point. In fact, Gallup itself notes that the results may be off by as much as ± 4 percentage points, in 19 of 20 samples.

   The income distribution in Table 4.5 may be based on relatively precise data concerning the family income of the 50 respondents in the table. However, this data has been grouped in this table, and the intervals are \$0-19,000 with 6 respondents, \$20-39,000 with 13 respondents, and so on. The original accuracy of the data has been lost as a result of this grouping. For the 6 respondents in the lowest income intervals, it is impossible to know whether they are near the lower end or the upper end of the interval, somewhere in between, or spread all across the interval. As a result, if this table is to be used for anything other than providing a general picture of the nature of the income distribution, the numbers produced are unlikely to be accurate to closer than the nearest thousand dollars, or even closer than the nearest \$10,000.

The above considerations show that there are no hard and fast rules concerning rounding, or use of significant figures. In general, it is best to be guided by **the accuracy of the data, the amount of approximation that has been carried out, the extent of sampling and non sampling error, and the uses of the data.**

In Chapter 5, some guidelines concerning rounding and significant figures will be discussed when calculating averages, and measures of variation. When carrying out such calculations, it is generally advisable to carry a considerable number of significant figures during the calculations, and then round off the results at the end. If there are a series of calculations with the same data, this generally produces more accurate results than if rounding is carried out at each stage of the process. Since you will be carrying out most calculations with a calculator, and calculators can carry many decimals, it is best to use all the decimals produced by the calculator at each stage of the calculation process. When reporting the answer, the rounding can be carried out in line with the accuracy of the data, the approximations made, and the potential uses of the data.

Some Special Cases. While the above are general rules, there are at least two other examples where these rules are commonly violated. These are discussed here.

The variable **age** is not usually rounded to the nearest year when it is reported. In North America, when asked what age you are, most respondents report **age as of last birthday**. This means that most people round their age down, to the age they were at their last birthday. Since this is the manner in which we commonly report age, it may be necessary to imagine that age 8, for example, really represents all ages between 8 years and 0 days to 8 years and 364 days. Some people may round age differently, meaning that age may not be all that accurately reported in many circumstances. For example, young people may round their age up a year or two, to appear older. Old people may round their age up considerably. For example, an 86 year old might say he or she is 90. In order to avoid these problems, a survey could ask birth date. This is likely to be more accurately reported than is age. Then the researcher could compute the exact age of each person. This is likely to be done only in circumstances where the age of the respondent is a very important variable.

When calculating **sample size**, researchers commonly round up the sample size before carrying out the research. For example, if the formula indicating sample size (see Chapter 7), shows that a researcher should collect a sample of 83.2 people, this is likely to be rounded up to 84, or perhaps even up to 85 or 100. This is because a larger sample size generally produces more precise results and because there are likely to be some respondents who do not respond. By boosting the sample size the researcher aims at,

it is hoped that a sample of at least 84 cases, with complete data, will be provided.

Example 4.7.3 Census Reports of Population.

Statistics Canada uses an slightly different method of rounding when reporting population data from the Census. If the Census volumes are examined, you will note that most population figures are reported as numbers ending in 5 or 0. There are at several reasons why Statistics Canada rounds population data in this manner.

First, population figures are not completely accurately determined, even by a Census. As noted in Chapter 2, there are a considerable number of nonsampling errors involved. Some people cannot be found, others do not respond, and there may be some confusion concerning exactly who is a Canadian resident. For these reasons, the population of most areas is not likely to be perfectly accurately reported, and Statistics Canada generally rounds these figures to the nearest 10 people.

Second, many population figures are further broken down into the number of people in each age group, each ethnic group, etc. The determination of the exact age or ethnic group of each person is subject to even more error, and this is an additional reason for rounding.

Third, it is unlikely that there is much need for anyone to know what the exact population total is, except perhaps for some very small communities. If the population of Pictou, or of Moose Jaw, can be determined to the nearest 10 people, this is surely adequate for most uses.

Finally, Statistics Canada carries out what it calls **random rounding**. *Rather than round each population total to 10, the values of population are rounded to either 5 or 10. The income distribution in Table 4.19 gives an example. The number of families in each income group is rounded to either the nearest 5 or 10, and whether the number in each group is reported as ending in 5 or 10 is determined on a random basis. The reason Statistics Canada uses this method is to protect the* **confidentiality** *of respondents. If rounding was not carried out, there might be only one person with a particular characteristic in a small community. For example, it might be reported that there was one person with income over $100,000 per year. If reported in this manner, this would violate the confidentiality that Statistics Canada has assured Candian residents that it will not violate. As a result, a particular town might be reported as having 10 or 5 such people, even though there was only 1. By rounding in such a manner, the total of the*

Variable	No. of Families
Under 5,000	55
5,000- 9,999	85
10,000-14,999	125
15,000-19,999	140
20,000-24,999	115
25,000-29,999	130
30,000-34,999	95
35,000-39,999	100
40,000-49,999	125
50,000 and over	320
Total	1,290

Table 4.19: Random Rounding, Regina Census Tract 12, 1986

column also comes reasonably close to the actual total. If rounding was always to the nearest 10, this might not produce such reliable results.

As a result of these considerations, population figures are never precisely determined, but are accurate to only the nearest 10 people in any category. This is generally precise enough for most research purposes. In contrast, sample data, where the sample size is much smaller, usually report exact results. In the case of samples, other means are used to protect confidentiality.

Rounding and Column Totals. When the data is rounded and presented in a percentage or proportional distribution, the column totals often do not add up properly. Table 4.20 shows how this can occur, and how it can be corrected. There, the distribution of the number of people per household, originally presented in Table 4.15, is presented as a proportional distribution and a percentage distribution. In the case of the proportional distribution, each f value has been divided by the sample size $n = 941$. Even though the sum of the f values is properly reported as 941, the sum of the proportions is 0.999, rather than 1. This sometimes occurs when the values are rounded. In this case, the proportions are rounded to three decimal places (3 significant figures), and because of this rounding, the sum of the proportions is

		Proportion Rounded to:		Percentage Rounded to:	
X	f	3 Decimals	Adjusted	Nearest %	Adjusted
1	155	0.165	0.165	16	16
2	286	0.304	0.305	30	31
3	164	0.174	0.174	17	17
4	223	0.237	0.237	24	24
5	86	0.091	0.091	9	9
6	21	0.022	0.022	2	2
7	5	0.005	0.005	1	1
8	1	0.001	0.001	0	0
Total	941	0.999	1.000	99	100

Table 4.20: Rounding of Distributions of People per Household

0.999. But once all the frequencies have been accounted for, the sum of the proportions should be 1.000. In order to adjust for this, and make the individual values of the proportions sum to the proper total of 1.000, one of the values has been adjusted. In this example, this is the proportion of cases for $X = 2$, adjusted from 0.304 to 0.305. The general rule to follow here is to **adjust the largest value in the column**, in order to produce the proper total. By adjusting the largest value, the smallest relative change in that value occurs. In this case, the proportion of households with 2 people is reported as 0.305, rather than 0.304, a very small change.

The same situation occurs in the case of the percentages. Here the proportions of the third column have been multiplied by 100%, and rounded to two decimals. As often occurs when reporting a percentage distribution, the sum of the percentages is not exactly 100%, even though all the calculations have been properly carried out and the rounding has been rounding to the nearest percentage. Again, an adjusted column has been produced, by adding 1 percentage point to the largest percentage, changing 30% to 31%. The column of adjusted percentages should be the column reported when presenting the data, making it appear more consistent.

After adding a column of percentages that has been rounded to the nearest percentage point, the total should not be off by more than 1, or at most 2, percentage points. If it is off by more than this, then it is likely that there has been some error made in calculating the percentages. However,

if the totals are 98, 99, 101 or 102, the adjustments suggested in the last paragraph should be made.

4.7.3 Class Limits

If a variable is discrete and has a fairly small number of values for the variable, then it may be possible to give a list of all the values of the variable, along with the respective frequencies (f) with which the variable occurs. Examples of this type of frequency distribution were given in Table 4.4 and Table 4.15 In most cases though, when organizing data for presentation, the values of the variable will have to be grouped into categories. Section 4.5 gave some guidelines concerning the grouping of data into categories. When data is grouped into intervals, there are some instances where the exact end points of the interval will need to be constructed. This section discusses these end points or **class limits** of the intervals, and the problems of interpreting these limits.

No Gap Between Intervals. The simplest situation occurs when the variable is grouped into a table where the end of one interval is the same value as the variable takes on at the beginning of the next interval. Such as situation is given in Table 4.21. In this table, the end points of the intervals are \$5,000, \$10,000, \$15,000, \$20,000, and so on. In the case of each of these intervals, the value at the upper end of one interval is the same as the value of the variable at the lower end of the next interval. These end points of the intervals are called the **class limits** of the intervals. If the data is already presented in this form, there is no confusion concerning what are the class limits of each interval, and these class limits can be termed the **real class limits**.

Non Overlapping Intervals. The difficulty with organizing raw data into intervals in this manner is to decide what to do if a value of the variable is exactly equal to the real class limit. Suppose a person has an income of exactly \$15,000. Should this value be placed in the \$10,000-15,000 interval or in the \$15,000-20,000 interval? It is not clear which is the proper interval for this value. If there are many such values, exactly at the real class limit, then perhaps half of these values should fall into the \$10,000-15,000 interval and the other half into \$15,000-20,000 interval.

Since making a decision concerning how to split cases that lie exactly on the dividing line between intervals is an awkward one, it is more common

Income in $	Number of Tax Returns
Under 5,000	380
5,000-10,000	47,890
10,000-15,000	79,630
15,000-20,000	70,660
20,000-30,000	108,500
30,000-40,000	62,970
40,000-50,000	33,600
50,000 and over	29,290
Total	432,910

Table 4.21: Distribution of Saskatchewan Tax Returns, 1988

Source: Saskatchewan Bureau of Statistics, **Economic Review 1991**

to close the first interval at a value slightly less than the value at the lower end of the next interval. This might be done as in Table 4.19. There it is clear than an income of $15,000 should be placed in the $15,000-19,999 interval. For that example, the values 0, $4,999, $5,000, $9,999, $10,000, $14,999, $15,000, and so on are called **apparent class limits**.

If the intervals are constructed in this manner, so that there are not values in common to the two intervals, there is no confusion concerning the interval into which the raw data is to be placed. In this case, suppose the raw data is rounded off to the nearest dollar, and any income up to and including $4,999 will be placed in the first interval, incomes of $5,000 through $9,999 in the second interval and so on.

While it is clear how data should be grouped in the case of these nonoverlapping intervals, another difficulty then emerges. The question that emerges is where the first interval really ends, and where the second one begins. In Table 4.21, does the first interval really end at $4,999, or at $5,000, or somewhere in between? In the example of Table 4.21, it makes little difference which of these limits are used, because the difference between the apparent class limits is only $1, and since the intervals are each several thousand dollars wide, a difference of $1 is insignificant. The following gives examples of

how real class limits can be constructed in cases where the intervals do not overlap.

Real Class Limits in Non-Overlapping Intervals. When intervals do not overlap, but a small gap appears between the upper limit of an interval, and the lower limit of the next interval, it is sometimes useful to define real class limits for these intervals. The class limits that initially appear on the table are called the **apparent class limits**. In the case where the intervals do not overlap, the apparent class limit for the upper end of the first interval is a little less than the apparent class limit at the lower end of the next interval.

Definition 4.7.3 Where there is a gap between the apparent class limits of adjacent intervals, the **real class limit** is the midpoint between the two apparent class limits.

By taking these midpoints at each end of the interval, the real class limits result in a new, slightly wider interval, and one in which there are no spaces or unfilled gaps between the real class limits of the new intervals. The process of constructing these limits will become clearer in the following example.

Example 4.7.4 Real Class Limits for Hours of Work

Hours of work for 60 workers is given in Table 4.22. Here the data is originally presented in intervals 0-9, 10-19, 20-29, 30-39, etc. The end points of these intervals, 0, 9, 10, 19, 20, 29, 30, and so on are called the apparent class limits. Since there appears to be a gap between the intervals, real class limits are constructed which close this gap.

Consider first the interval 10-19. This appears to begin at 10 and end at 19, and if the raw data has been rounded to the nearest hour of work, then all workers who report that they have worked between 10 and 19 hours per week are placed in this interval. But there is a gap between the interval 0-9 and 10-19, and this makes it appear as if there are some values that may be missing. Since hours of work is a measure in terms of time, this variable is continuous, meaning that there should not really be a gap. This gap is closed by constructing the lower real class limit as the midpoint between 9 and 10, that is, 9.5. Similarly, at the upper end, the apparent gap between 19 and 20 is closed by taking the midpoint between these values, that is, 19.5. This makes the real class limits for this first interval 9.5-19.5.

Similarly, the next interval is 20-29, with apparent limits of 20 and 29. Again, if one takes the midpoints at each end, the real class limits are 19.5, midway between 19 and 20, and at the upper end midway between 29 and 30, the real class limit is 29.5. The remainder of the real class limits are similarly constructed.

Hours or Work		
Apparent Class Limits	Real Class Limits	f
0-9	-0.5-9.5	1
10-19	9.5-19.5	5
20-29	19.5-29.5	5
30-39	29.5-39.5	10
40-49	39.5-49.5	25
50-59	49.5-59.5	9
60-69	59.5-69.5	3
70-79	69.5-79.5	1
80-89	79.5-89.5	0
90-99	89.5-99.5	1
Total		60

Table 4.22: Distribution of Hours Worked per Week, 60 Respondents

In the above example, the first interval deserves special attention. This is the interval 0-9, and the real class limits are given as -0.5 to 9.5. The lower real class limit here makes little sense in that it is not possible to work between -0.5 to 0 hours, that is a physical impossibility. The reason for extending the first interval to a point below 0 is for completeness. This will become clearer if the interval width is first defined.

Definition 4.7.4 The **interval width** for an interval is the upper real class limit minus the lower real class limit for an interval.

In this example, each of the other intervals is 10 units wide. For example, the interval 10-19 is represented by 9.5-19.5. The interval width of this interval is $19.5 - 9.5 = 10$ hours. Similarly, the interval 20-29 has real

class limits of 19.5 and 29.5 so it is $29.5 - 19.5 = 10$ hours wide. Each interval here is 10 units of width. Since 0-9 is essentially the same sort of interval as the other intervals, it should also have an interval width of 10 hours. In order to construct a interval of 10 hours of width, in this case it is necessary to make the lower end point -0.5, so that the interval width here is $9.5 - (-0.5) = 10$ hours. (Remember that when a negative number is subtracted, this is equivalent to adding the number, or two minuses are equivalent to a plus). While this procedure makes this first interval look a little odd, it helps with consistency, meaning that each interval has the same interval width.

Example 4.7.5 Alcohol Consumption of Canadians, 1985

Number of Drinks Per Week	Percentage Distribution of Men	Women	Real Class Limits
Less than 1	18.8	30.1	-0.5-0.5
1-6	43.2	52.6	0.5-6.5
7-13	20.3	12.3	6.5-13.5
14 or more	17.7	5.0	13.5 or more
Total	100.0	100.0	

Table 4.23: Alcohol Consumption of Current Drinkers, Aged 15 and Over, 1985

Source: C. McKie and K. Thompson, **Canadian Social Trends**, page 91.

The data in Table 4.23 concerning alcohol consumption patterns of men and women who are current drinkers is taken from **Canadian Social Trends**. *In this table, the data in the first three columns are presented in the form shown in the publication. The apparent class limits here are 1, 6, 7, 13 and 14. In each case, there appears to be a gap between the intervals, although in the case of the first interval, less than 1, there may be no gap between it and the next interval 1-6. If the real class limits are to be constructed here,*

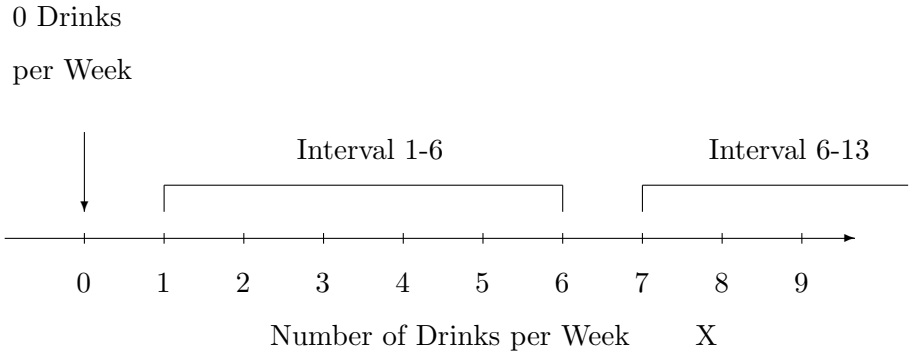

Figure 4.11: Apparent Class Limits for Drinks per WeeK

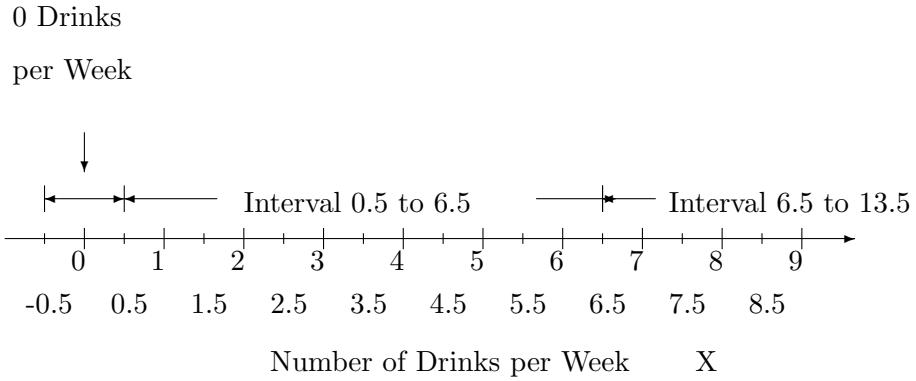

Figure 4.12: Apparent Class Limits for Drinks per WeeK

the same method can be used, splitting the difference between the apparent class limits of the intervals. If the first interval is considered to be 0, then the real class limits for subsequent intervals are 0.5 to 6.5, 6.5 to 13.5, and 13.5 and over. For the first interval, it would make most sense to construct this in the same manner as in the previous example, and running the interval to below 0. That is, the real class limits for this first interval would be -0.5 to 0.5.

The alcohol consumption example also illustrates the meaning of the real class limits in terms of interval width, and what what is accomplishing in

doing this. The interval width of the second interval is 1-6, representing 6 units of the variable, number of drinks per week. By constructing the real class limits, the interval width clearly is $6.5 - 0.5 = 6$. If the apparent class limits had been used, there might have been some confusion concerning whether the interval was $6 - 1 = 5$ units or 6 units. The construction of the real class limits makes this clear. The interval 1 through 6 inclusive, really does represent 6 units of the variable, *number of drinks per week*. Similarly, the interval 7-13 is of width $13.5 - 6.5 = 7$ drinks per week.

The interval less than 1 also represents one unit of the variable. That is, even if this is 0 drinks per week, 0 represents one unit of the variable. The real class limits should reflect this, and have been constructed as -0.5 and 0.5, representing an interval width of $0.5 - (-0.5) = 1$ unit of the variable.

The diagram in Figures 4.11 and 4.12 may also make this process a little clearer. The variable *number of drinks per week* is measured along the horizontal axis, as 0, 1, 2, 3, and so on. Only the range from 0 to 9 drinks per week is shown here so that this section of the values of the variable can be more clearly illustrated. Each integer 0, 1, 2, 3, etc. represents an equal distance along the horizontal axis because the scale is an interval level scale, where each unit of the variable represents an equal extra quantity. That is, each unit of 1 represents one more drink per week. The particular grouping that comes from **Canadian Social Trends** is more or less arbitrary. The value 0 is isolated by itself, but values 1-6 are grouped together. This is followed by another grouping of 7 to 13. Since these are arbitrary groupings, the gaps that appear between the apparent class limits are also arbitrary. The intervals are illustrated in Figure 4.11.

Figure 4.12 gives the corresponding real class limits. All of the space along the X axis has been allocated to one or other of the intervals based on the real class limits. This is done by recognizing that each unit of the variable X, can also be represented by an interval that is one unit wide, with the integer value of the variable in the middle. That is, 1 can be represented by the interval 0.5 to 1.5, the value $X = 2$ by the interval 1.5 to 2.5, and so on. The value 0 can now be seen to be represented by the interval -0.5 to 0.5. In this way, all the space along the X axis can be allocated among one or other of the values of X. Once this has been done, it becomes clear that the interval 1-6 can be represented by all the values of the variable X along the range from 0.5 to 6.5. The real class limits for the interval 1-6 are 0.5 and 6.5. From this diagram, the logic of using the -0.5 to 0.5 interval for 0 can also be seen.

This procedure results in a certain consistency of treatment of the values of X in two senses - rounding and number of values of X. In terms of the latter, the number of values of X in the interval 1-6 is six, that is, 1, 2, 3, 4, 5 and 6. By using the real class limits, the interval width clearly is $6.5 - 0.5 = 6$ units, and this interval width of 6 represents these 6 values of X. Similarly, the interval 7-13 has 7 values of X, 7, 8, 9, 10, 11, 12, 13. This is made clear by the real class limits of 6.5 to 13.5, and the interval width of $13.5 - 6.5 = 7$. The category *less than 1*, or 0 drinks per week, also represents one unit on the axis, or one value of X, even though this value happens to be $X = 0$. Again, it is made clear that this value does represent one unit along the axis by making the real class limits -0.5 and 0.5, an interval width of 1.

With respect to rounding, this process is also consistent. If the values of X are reported rounded off to the nearest integer, then it is likely that respondents rounded the values. That is, if someone drank 1.3 drinks per week, this may be reported as 1. Or if someone had about 8.6 drinks per week, then this may be rounded to 9 drinks per week. When rounding, the usual procedure is to **round to the nearest integer**. When doing this, consider which values of X could result in a value of, say 3 drinks per week. Any value of X just above 2.5, would be rounded to $X = 3$. At the other end, any value of X just under 3.5 would be rounded down to $X = 3$ as well. If the data was reported to more decimals than it is here, the interval 2.5 to 3.5 would be the interval which represents the value of $X = 3$. This can be seen in Figure 4.12. All the values along the continuous X axis are allocated to one or other of the integer values of X. That is, $X = 1$ is represented by any value of just over 1.5, up to 2.5, $X = 2$ is represented by 2.5 to 3.5, and so on. In this way, the construction of the real class limits may be considered to be a process of reversing the rounding procedure. That is, it is a process of considering the range of potential values which could be considered to result in the integer values which are reported in the table.

Example 4.7.6 Graduate Record Examinations

The 1989-90 **Guide to the Graduate Record Examinations** *contains Table 4.24, showing the distribution of test scores in the quantitative portion of the Graduate Record Examination for Sociology majors and for Mathematics majors. The first three columns of this table are taken directly from the* **Guide**, *and the rather odd intervals of 200-290, 300-390, 400-490, etc. are those which are contained in this* **Guide**. *Exactly why the large*

Test Score	Distribution of Test Scores (Per Cent by Discipline) Sociology	Mathematics	Real Class Limits	Interval Width
200-290	5.4	0.1	195-295	100
300-390	16.9	1.2	295-395	100
400-490	28.0	3.9	395-495	100
500-590	27.3	13.4	495-595	100
600-690	15.7	30.6	595-695	100
700-790	6.3	42.1	695-795	100
800	0.4	8.7	795-805	10
Number of Students	2,386	3,890		

Table 4.24: Distribution of Graduate Record Examination Scores

gap is left between the endpoints of the intervals in each case is not clear from the publication, but this is the manner in which they are reported.

The apparent class limits in this case are 200, 290, 300, 390, 400, etc. The real class limits of 195, 295, 395, etc. have been constructed in the fourth column of Table 4.24 by taking the midpoint of the gap between each of the apparent limits, and allocating half of this difference to each of the intervals. In terms of the distances along the interval, or the rounding procedure, the argument here is similar to that discussed so far. For example, in order to allocate all the values of X along the line to one or other of the intervals, any value of just over 295 is allocated to the 300-390 category, and any value just under 295 is allocated to the 200-290 category. This is because the values presented in the **Guide** *appear to be rounded to the nearest 10 units in this case.*

The last category of 800 presents a problem in that 800 may be the maximum score. Again, however, if the values of X are rounded to the nearest 10, for consistency it would seem preferable to count the category of 800 as being similar to the other categories. In this case the gap between 790 and 800 is allocated half to each category, but the 800 category is then made to go from 795 to 805, representing 10 units of GRE scores.

Example 4.7.7 Reading Skill Levels in Canada

Reading Skill Level	Per Cent at each Level in: Atlantic Region	Prairie Region	Real Class Limits	Interval Width
1	6	4	0.5-1.5	1
2	13	7	1.5-2.5	1
3	30	19	2.5-3.5	1
4	52	70	3.5-4.5	1

Table 4.25: Percentage Distribution of Reading Skill Levels of Persons Aged 16-69, Atlantic and Prairie Provinces

*Table 4.25 gives percentage distributions of reading skill levels of the adult population of the Atlantic provinces and of the Prairies. Level 1 is the lowest literacy level, and those at this level have difficulty reading printed material. Reading skill levels increase from Level 1 through Level 4, so that those at Level 4 "can meet most everyday reading demands. This is a large and diverse group which exhibits a wide range of reading skills." (***Perception***, Summer 1990, Vol. 14, No. 3, p. 20. The first three columns of this table are taken directly from the article in* **Perception**. *This data is originally taken from a survey of the literacy levels of Canadian, carried out in 1987 by Statistics Canada.*

Int this case, the ordinal variable reading skill is classified into only a discrete set of categories. Since reading skill might be regarded by some as being continuous in nature, for some purposes, it may be useful to construct real class limits. In this case, the real class limits are constructed using the logic presented in Figure 4.12. Each value of reading skill is represented by 1 unit on the scale, and that the real class limits are constructed so that the intervals are each one unit wide. Thus $X = 1$ *can be represented by a range of values from 0.5 to 1.5, producing an interval of 1 unit wide. Each of the other intervals is similarly constructed.*

Uses of Real Class Limits. 6 When presenting data, it is common to classify the data so that the intervals do not overlap, 65or meet each other. This produces apparent class limits which may leave a small gap between the end points of successive intervals. For presentation of the data in a table,

it is best to leave the data in this form. This gives the reader of the table an idea of how the data has been rounded and grouped.

For purposes of producing diagrams, it is sometimes preferable to close the gaps between the intervals. If the variable is continuous, then these gaps should probably be closed when presenting the data in a diagram. This means constructing real class limits, so that the X variable appears to vary continuously. Examples of this are presented in Section 4.8. In addition, any time that the interval widths, or the number of values in the interval, are required, then the construction of the real class limits is a useful way of determining these.

In Chapter 5, the median and percentiles are calculated, and this requires interpolation along the X axis, to determine the exact point at which the particular percentile is reached. In order to carry this out as accurately as possible, it is preferable to construct the real class limits for the intervals. In contrast, if all that is needed is the midpoint of the interval, in order to calculated the mean and standard deviation, then the real class limits need not be considered. That is, the real class limits amount to adding an equal range of values at each end of the interval. This process does not alter the midpoint of the interval.

4.8 Histograms

4.8.1 Introduction

A histogram is a bar chart showing the relative frequency of occurrence of each value or set of values of a variable. The histogram shows which values of the variable occur most frequently and which values occur less frequently. The shape of the histogram is also of interest because it tells how concentrated or dispersed, or how skewed or symmetric, the distribution of the variable is. Histograms also provide a basis for constructing some of the theoretical distributions such as the normal distribution.

The method of constructing a histogram is straightforward if all the intervals into which the data are grouped are of equal size or equal width. This case is described in the next section. Where the data have been grouped into intervals of different width or unequal size, the construction of a histogram involves the calculation of the relative frequency or density of occurrence of the variable. This situation is described in Section 4.8.3 Examples of each type of histogram are given in this section.

Recal that the definition of the width of an interval is the difference

between the upper and lower real class limits for that interval. Alternatively stated, the interval width of any interval is the value of the variable at the upper end of the interval minus the value of the variable at the lower end of the interval. When determining interval width, the real class limits should be used.

4.8.2 Histograms for Intervals of Equal Size

In the case where the data have been grouped into intervals of equal width, one constructs a histogram by drawing bars that are the width of each of the intervals. The heights of the bars are drawn so that they are proportional in height to the frequency of occurrence of the variable in each interval. In the case of a percentage (or proportional) distribution, the height of the bars is the percentage (or proportion) of cases in each interval.

Example 4.8.1 Per Cent Urban in 129 Countries

This example shows how the data in Appendix B can be organized into a histogram. In Appendix B, various characteristics are given for 129 countries. One of the variables given in URB, the **per cent urban***. This refers to the per cent of the population of each country that lives in urban areas. The per cent of the population that is urban is organized into a stem and leaf display, and then presented as a frequency distribution table and histogram.*

The **stem and leaf display** *is given in Tables 4.26 and 4.27 with the resulting frequency distribution table given in Table 4.28, and the histogram in Figure 4.13. Because these numbers are rounded to the nearest integer, the real class limits are used in order to more accurately graph the distribution.*

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

Table 4.26: Unordered Stem and Leaf Display of Per Cent Urban

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

Table 4.27: Ordered Stem and Leaf Display of Per Cent Urban

Per Cent Urban	Real Class Limits	f
10-19	9.5-19.5	4
20-29	19.5-29.5	5
30-39	29.5-39.5	4
40-49	39.5-49.5	11
50-59	49.5-59.5	8
60-69	59.5-69.5	7
70-79	69.5-79.5	10
80-89	79.5-89.5	9
90-99	89.5-99.5	2
Total		60

Table 4.28: Distribution of Per Cent Urban, 60 Countries

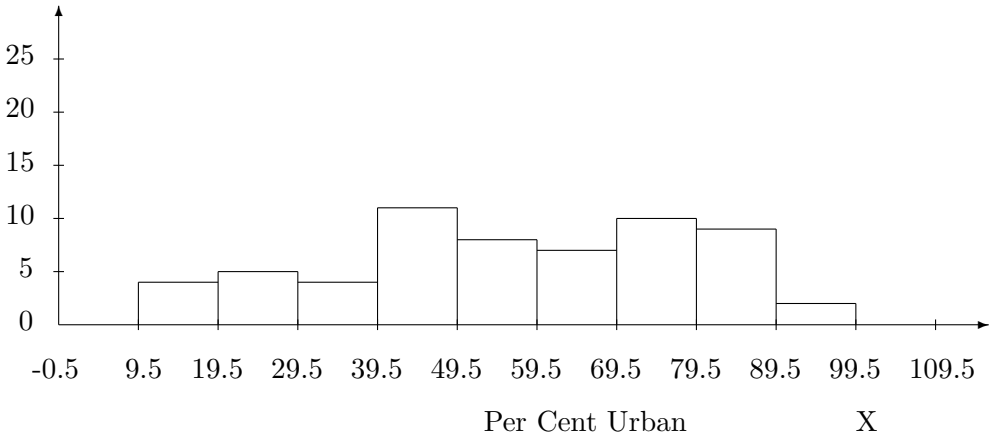

Figure 4.13: Histogram of Per Cent Urban in 129 Countries

Example 4.8.2 Distribution of Socioeconomic Status, Intervals of Equal Size

A sample of 200 Regina labour force members is taken to determine the types of occupations in which they work. The occupations are given a ranking according to the level of prestige or social status accorded to these occupations. The occupational status scale used to rank these occupations is the **Blishen scale of socioeconomic status**. *For Canada as a whole, the values on this scale are based on a combination of average income and education levels associated with each occupation. These values have been applied to the occupations of Regina respondents in this survey. The socioeconomic status scores range from 0 to 100, with 0 indicating an occupation with a very low level of status and with a score of 100 indicating a very high level of socioeconomic status possible for an occupation. It is assumed here that socioeconomic status is measured on an interval level scale. The distribution of scores for the 200 people sampled is given in Table 4.29 and the associated histogram is given in Figure 4.14 Note that in this example, there is no gap between the endpoints of the intervals so that the apparent class limits are also the real class limits.*

Level of Status (X)	No. of Cases (f)	Proportion	Percentage
0-10	0	0.000	0.0
10-20	0	0.000	0.0
20-30	24	0.120	12.0
30-40	44	0.220	22.0
40-50	48	0.240	24.0
50-60	49	0.245	24.5
60-70	23	0.115	11.5
70-80	12	0.060	6.0
80-90	0	0.000	0.0
90-100	0	0.000	0.0
Total	200	1.000	100.0

Table 4.29: Distribution of Socioeconomic Status of 200 Respondents

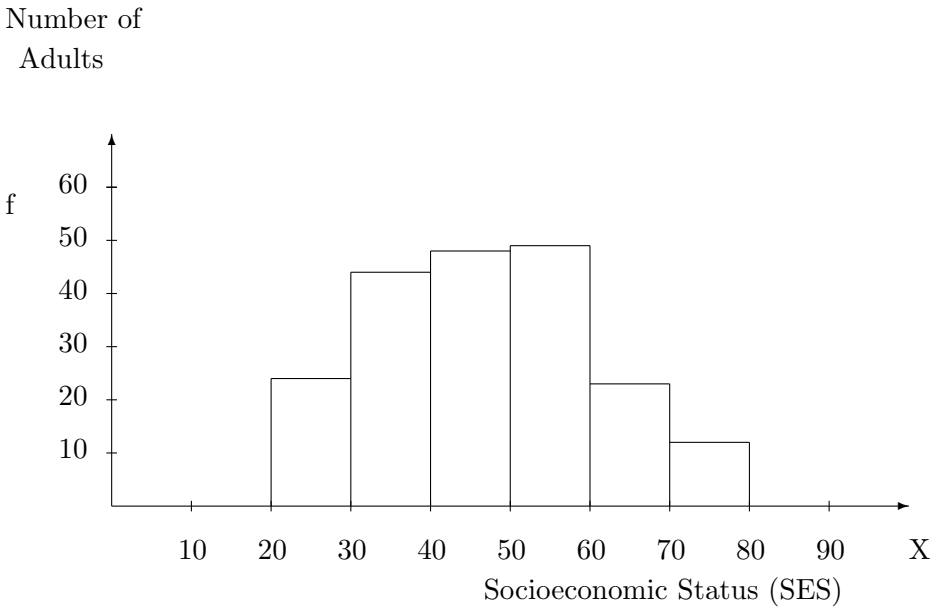

Figure 4.14: Histogram of Socioeconomic Status (SES) of Sample of 200 Regina Adults

The histogram of Figure 4.14 shows that the bulk of the people sampled have levels of socioeconomic status in the middle of the range, from 30 to 60. In this sample, only a few people have occupations with socioeconomic status below 30 or above 70.

4.8.3 Intervals of Unequal Width

Frequently, data is classified into intervals so that the intervals are not all of equal width. For parts of the distribution where there are relatively few cases, the interval widths may have been widened. For other parts of the distribution where there is a greater concentration of cases, the intervals that are used may be much narrower in width. This may have been done in order to provide a more complete description of the population over the more common range of values, or to reduce the size of the table where there are fewer cases occurring. For example, suppose the data contained in Table 4.29 had instead been reported as in Table 4.30 This is the same data, but compressed into fewer intervals.

The data in Table 4.30 measures the same distribution as in Table 4.29

but it is presented in a different form. If the data in Table 4.30 is presented as a histogram without taking account of the different interval sizes, the **incorrect** histogram of Figure 4.15 would result.

Level of Status (X)	No. of Cases (f)
0-20	0
20-40	68
40-50	48
50-60	49
60-80	35
80-100	0
Total	200

Table 4.30: Distribution of Socioeconomic Status

As can be seen in Figure 4.15, this histogram distorts the distribution at the lower and upper ends. It appears that there are more people who have occupations with socioeconomic status between 20 and 40 than there are in the range from 40 to 50 or 50 to 60. In fact, the only reason there are 68 people in total over the range 20 to 40 is that this interval has been created by combining the 2 intervals 20 to 30 and 30 to 40 so that this new interval is twice as wide as the 40 to 50 or 50 to 60 interval. A similar problem appears for the 60 to 80 interval although the situation is not so exaggerated there.

In order to correct for unequal interval widths when constructing a histogram, it is first necessary to compute the **density** or relative frequency of occurrence of the variable.

Definition 4.8.1 The **density** of occurrence of the variable in any interval is defined as the frequency of occurrence of the variable, per unit of the variable in that interval.

The density is calculated by dividing the frequency of occurrence of the variable by the interval width, in units of the variable. The result of

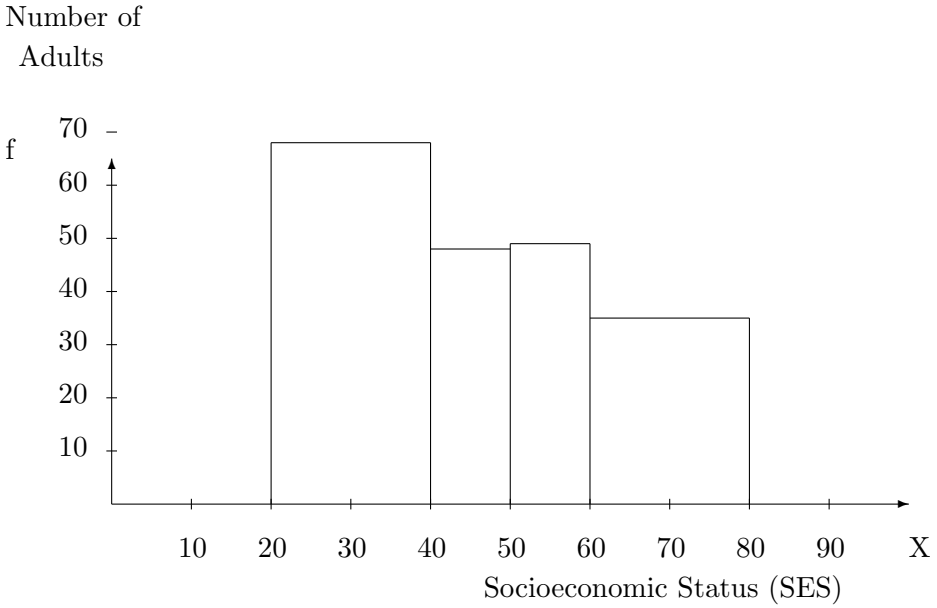

Figure 4.15: Incorrect Histogram of Socioeconomic Status

calculating and graphing these densities is to keep the shape of the histogram much the same, regardless of the way in which the data has been grouped.

Example 4.8.3 Distribution of Socioeconomic Status, Intervals of Unequal Size

The densities for the data in Table 4.30 are calculated in Table 4.31. In order to calculate the density for an interval, determine the interval width and divide the number of cases in that interval by the interval width. For example, for the interval 20 to 40, the interval width is 20 units of status and there were 68 people in this interval. This is a density of $68/20 = 3.40$ *people per unit of socioeconomic status. By calculating these densities for each interval, one can meaningfully compare the relative frequencies of occurrence of the variable across all values of the variable. The correct histogram is graphed in Figure 4.16.*

The correct histogram in Figure 4.16 has a shape that is very similar to the more detailed histogram of Figure 4.14. The gross distortions that were introduced in the incorrect histogram, Figure 4.15 have now disappeared. The only real difference in Figures 4.16 and Figure 4.14 is that the latter

X	f	Interval Width	Density
0-20	0	20	$0/20 = 0.00$
20-40	68	20	$68/20 = 3.40$
40-50	48	10	$48/10 = 4.80$
50-60	49	10	$49/10 = 4.90$
60-80	35	20	$35/20 = 1.75$
80-100	0	20	$0/20 = 0.00$
Total	200		

Table 4.31: Frequencies and Densities for Distribution of Socioeconomic Status

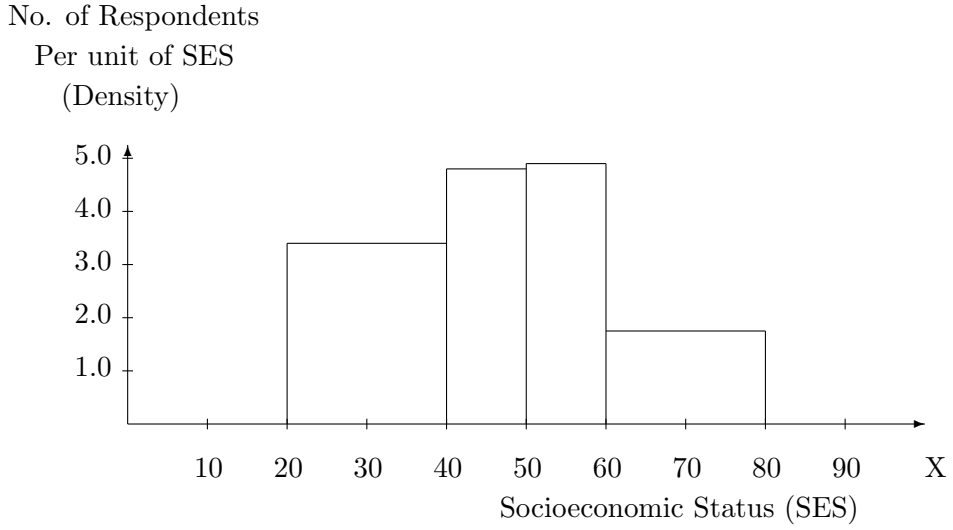

Figure 4.16: Correct Histogram of Socioeconomic Status

histogram shows a little less detail than does the first, and the units on the vertical axis are different. By using wider intervals in Tables 4.30 and 4.31, less detail is presented. The result of this is that the intervals at the lower and upper end are wider than in Figure 4.14. However, they now have the correct height.

In terms of the difference in units on the vertical axis, note that this axis is now measured in frequency per single unit of socioeconomic status. In Figure 4.14 the frequencies on the vertical axis were really frequencies per 10 units of socioeconomic status, since they were frequencies for each interval of 10 units. If the densities of Figure 4.16 are multiplied by 10, the original units of Figure 4.14 can be restored.

The aim of calculating and graphing the frequency distribution in terms of densities is to present a picture of the data so that the intervals into which the data have been grouped have a minimal effect on the shape of the histogram. Even if wider intervals are used, the general shape of the histogram should be unchanged from where narrower intervals are used. This is because the same data is being used and the phenomenon being described is the same.

Example 4.8.4 Education Level of Saskatchewan Wives

A sample of wives is taken from Statistics Canada's annual Survey of Consumer Finances. The sample is divided into those wives who earn less than $5,000 per year, and those who have earnings of $5,000 plus. The sample contains 202 wives, with earnings levels of these wives and the number of years of education completed by these wives as given in Table 4.32.

Education in Years Completed	Number of Wives Earning <$5,000	$5,000 plus
0-10	41	18
11	12	13
12	15	31
13-17	23	29
18-22	6	14
	97	105

Table 4.32: Education Level of Saskatchewan Wives, by level of Annual Earnings

For this problem, one must calculate the densities of occurrence of the number of wives per year of education completed. This is done in Tables 4.33 and 4.34 and the results are graphed in Figures 4.17 and 4.18. The real class limits are also used, because of the gaps between the categories into which the data was originally grouped.

Based on Figures 4.17 and 4.18 one can quickly get an idea of the distribution of education levels for each of the two groups. If one had looked only at the table, one might get the impression the lowest education levels were most common, especially for wives with lower earnings. Fully 41 of the 97 wives with earnings under $5,000 per year had less than 11 years of education. However, once one calculates the density for this group, it is apparent that the relatively large number of wives with this low education level is at least partly a consequence of the wide interval, 0-10 years. When the densities are calculated, it is seen that 11 and 12 years of education is by far the most common for both sets of wives.

At the same time, some differences between the two distributions can be

Education in Years Completed Apparent Limits	Real Limits	Width	Number of Wives	Density
0-10	-0.5 -10.5	11	41	3.73
11	10.5-11.5	1	12	12
12	11.5-12.5	1	15	15
13-17	12.5-17.5	5	23	4.6
18-22	18.5-22.5	4	6	1.5
Total			97	

Table 4.33: Education Level of Saskatchewan Wives, Annual Earnings <$5,000

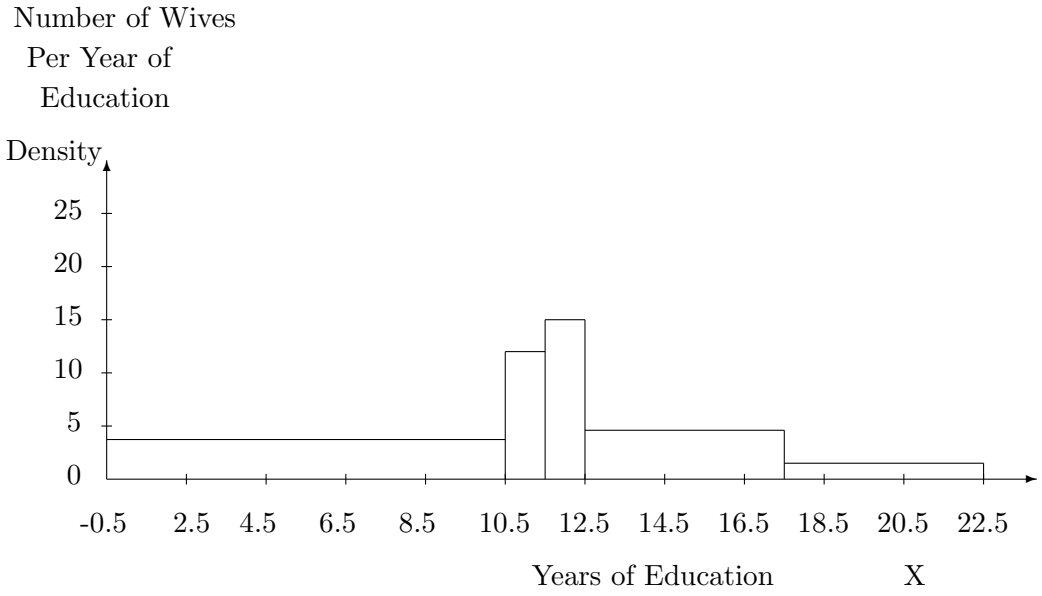

Figure 4.17: Histogram of Years of Education of Wives Earning less than $5,000 Annually

Education in Years Completed Apparent Limits	Real Limits	Width	Number of Wives	Density
0-10	-0.5 -10.5	11	18	1.64
11	10.5-11.5	1	13	13
12	11.5-12.5	1	31	31
13-17	12.5-17.5	5	29	5.8
18-22	18.5-22.5	4	14	3.5
Total			105	

Table 4.34: Education Level of Saskatchewan Wives, Annual Earnings $5,000 Plus

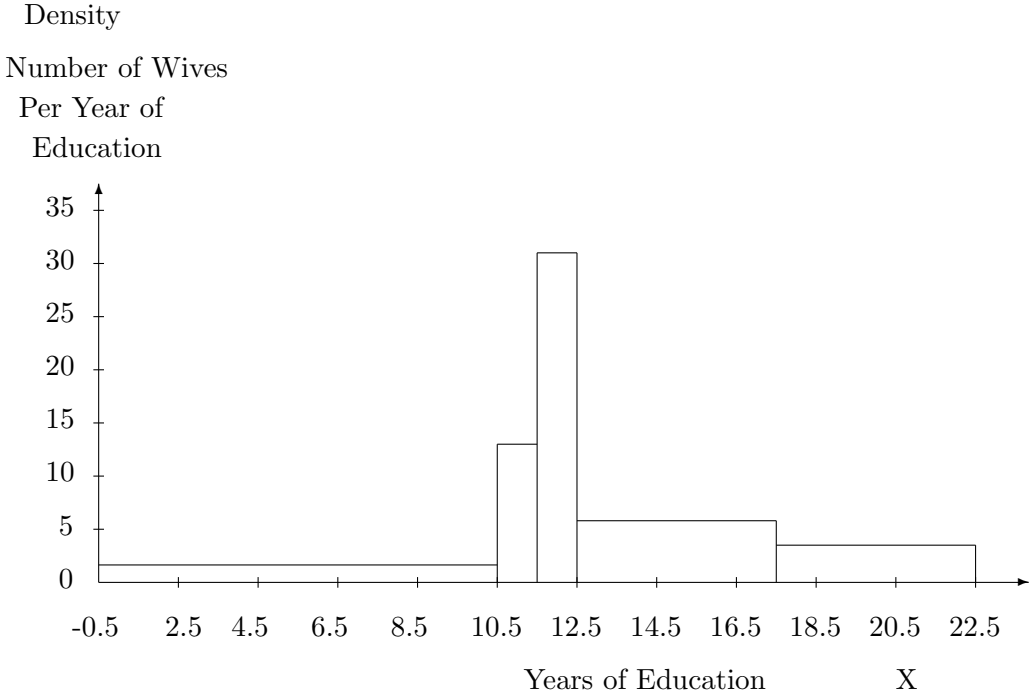

Figure 4.18: Histogram of Years of Education of Wives Earning $5,000 or more Annually

seen. The wives with lower earnings tend to have somewhat lower levels of education than the wives with higher earnings. The density for 0 to 10 years of schooling is lower for the wives with lower earnings. On the other hand, the densities for the wives with higher earnings is greater than for those with lower earnings. While 12 years of schooling is by far the most common single number of years of schooling for each group, those with higher earnings are considerably more likely to have 12 than 11 years. For the lower earnings group of wives, those having 11 years of education are almost as common as those having 12 years. The densities for the 13 and over years of education is distinctly greater for the high than the lower earnings wives.

Example 4.8.5 Farm Size in Saskatchewan

The distribution of farm size for census farms in Saskatchewan from the 1986 Census of Agriculture is given in Table 4.35. The first two columns of Table 4.35 are taken from Government of Saskatchewan, Bureau of Statistics, **Economic Review 1990**, page 21, Table 35. The rather odd set of intervals into which the Census groups farm sizes means that densities must be calculated before drawing the histogram. The histogram is presented in Figure 4.19. Note that in this example, the difference of 1 acre between intervals has been ignored because it is such a small amount compared to the size of farms and the width of the intervals. The real class limits are thus taken to be 10, 70, 240, etc.

The intervals that Statistics Canada has used for grouping farm size do not seem particularly appropriate, and it is not clear why such intervals have been used. However, these same intervals have now been used for several decades and given that researchers and analysts like to compare farm sizes in different years, it seems that Statistics Canada is unlikely to change this grouping. Given a table with this grouping, one must calculate the densities in order to present a reasonably accurate histogram showing the distribution of farm size in Saskatchewan.

The last interval, 1,600 acres and over, is an example of an **open ended interval**. No information on the distribution of farm sizes in this interval is available from the publication containing the table, other than the fact that there are 10,541 farms of 1,600 acres or more. All that can be done in graphing this is to approximate the correct height of the bar for this last interval. Since we do not know how wide the interval is, the density cannot be accurately calculated. In addition, most of the farms in this interval are not likely to be too much larger than 1,600 acres. Finally, the number of farms is considerably fewer as one moves to larger and larger farm size.

Farm Size in Acres (X)	No. of Farms (f)	Interval Width	Density
Under 10	593	10	59.3
10-69	1,107	60	18.4
70-239	7,017	170	41.3
240-399	7,505	160	46.9
400-559	6,514	160	40.7
560-759	7,939	200	39.7
760-1,119	12,323	360	34.2
1,120-1,599	9,892	480	20.6
1,600 plus	10,541		
Total	63,431		
Average Size	1,036		

Table 4.35: Distribution of Saskatchewan Farm Size, 1986

Given the above considerations, the best that can be done is to examine the rest of the histogram and guess at which height the bar for the open ended interval should be drawn. One usually draws it at a height that seems reasonable, given the behaviour of the distribution in the intervals just below this. Often one does not close the bar on the right in order to graphically indicate that this bar is based on an open ended interval. One could also label the interval and state how many cases this interval contains.

The histogram in Figure 4.19 gives one a fairly clear idea of the distribution of farm sizes in the province, apart from one oddity. Apart from the very first interval of under 10 acres, the histogram takes on the distribution that one might expect. That is, the smallest farms are not all that frequent in occurrence, with the relative occurrence of farms being greatest in the range of about 100 to 1,000 acres. In this range, the mode is in the 240-399 range. As one moves to larger and larger farm size, the densities decline, showing that while there are a lot of larger farms (over 1,000 acres) in the province, in relative terms there are not as many as there are at lower acreage levels. This part of the diagram is the usual sort of distribution one encounters for distributions of farm size, not so dense at the lowest end, more dense in the middle where the bulk of the farms lie, and tailing off toward lower densities

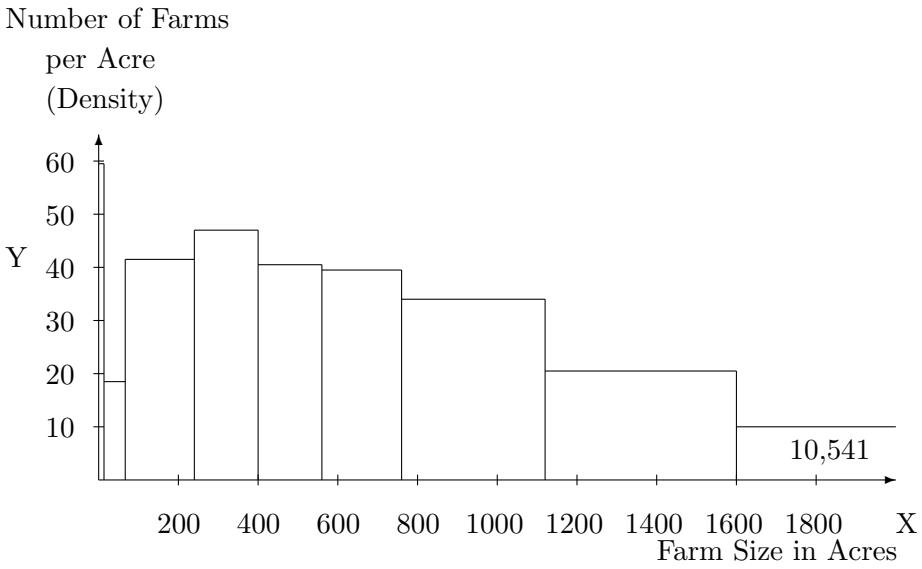

Figure 4.19: Histogram of Saskatchewan Farm Size, 1986

at the upper level. The larger farms account for a considerable portion of the total acreage, because they are so large, but there are relatively few of these largest farms.

The one oddity is that the interval under 10 acres is, in relative terms, the one with the greatest density. Why this occurs is not clear, but per acre of farm size, there are actually more farms in this lowest acreage interval than there are in any other interval. Whether these very small farms can really be considered as farms of the same type as larger farms, is not clear from this table. Most likely, these farms would be considerably different in nature than larger farms. These smallest farms might produce some vegetables or livestock, but are unlikely to be similar to the larger grain farms of the province.

4.9 Conclusion

This chapter has discussed various ways that data can be grouped in order that it can be meaningfully examined. It is often difficult to examine and absorb the raw, or ungrouped, data that a researcher begins with. These data can be grouped using a tally or a stem and leaf display. The resulting data can then be organized into a frequency distribution table or diagram.

This chapter has presented various guidelines concerning how this can be carried out so that the data can be analyzed.

The frequency distribution tables and diagrams of this chapter form the basis for the the statistical analysis of the later chapters of the textbook. Chapter 5 begins this process by examining various summary measures of frequency distributions.