Contents

Chapter 5

Central Tendency and Variation

5.1 Introduction

In order to examine, present and understand distributions, it is useful to organize the data presented in distributions into summary measures. The centre of a distribution and the variability in a distribution are the most common and often the most useful summary measures which can be provided concerning a distribution. Together these two sets of measures tell a great deal concerning the nature of the distribution, and how the members of the population are spread across the distribution.

The summary measure with which most people are familiar is the average, or in technical terms, the arithmetic mean. For example, an instructor may summarize the set of all grades in a class by calculating the average, or mean, grade. For a student, the grade point average is a measure which summarizes his or her performance. Each of these averages provides a quick idea of where the distribution of the set of grades is centred. This average, or arithmetic mean, is only one of several summary measures which can be used to describe the centre of a set of data. The other measures commonly used are the median and the mode.

Measures of variation are less commonly used than are measures of central tendency. Even so, measures of variation are very important for understanding how similar, or how varied, different members of a population are. The idea of a measure of variation is to present a single number which indicates whether the values of the variable are all quite similar or whether

the values vary considerably. A measure of variation familiar to most people is the range, the numerical difference between the largest and the smallest value in a set of data. For example, if the range of prices for product A is from \$15 to \$25 at different stores, then this product would ordinarily be considered to have a considerable degree of variation from store to store. In contrast, if product B has a range of prices from \$19 to \$21, then this means a considerably smaller variation in price. The range for A is \$10, and for B is \$2, so that the variation in price of A is much greater than the variation in price of product B.

Other measures of variation are the interquartile range, the variance and the standard deviation. Each of these is important in statistical analysis, extensively used in research, but not commonly reported in the media. In order to carry out statistical analysis of data, it is necessary to become familiar with these measures of variation as well.

In addition to measures of central tendency and variation, statisticians use positional measures, indicating the percentage of the cases which are less than a particular value of the variable. Distributions can also be described as symmetrical or asymmetrical, and statisticians have devised various measures of skewness to indicate the extent of asymmetry of a distribution. Various other summary measures will be mentioned in this text, but measures of centrality and variation are the most important statistical measures.

This chapter presents the three common measures of central tendency, showing how to calculate them, and then discussing the different uses for each of these measures. Following this, measures of variation are presented and discussed. Some comments concerning the interpretation and use of these measures are also made.

Measures of Central Tendency. As the name indicates, these measures describe the centre of a distribution. These measures may be termed either **measures of centrality** or **measures of central tendency**. In the case of some data sets, the centre of the distribution may be very clear, and not subject to any question. In this case, each of the measures discussed here will be the same, and it does not matter which is used. In other cases, the centre of a distribution of a set of data may be much less clear cut, meaning that there are different views of which is the appropriate centre of a distribution.

The following sections show how to calculate the three commonly used measures of central tendency: **mode**, **median** and **mean**. Since the man-

ner in which each is calculated differs depending on whether the data is grouped or ungrouped, the method of calculating each measure in each circumstance is shown. Following this, some guidelines concerning when each of the measures is to be used are given.

5.2 The Mode

The mode for a set of data is the most common value of the variable, that value of the variable which occurs most frequently. The mode is likely to be most useful when there is a single value of the variable which stands out, occurring much more frequently than any other value. Where there are several values of a variable, with each occuring frequently and a similar number of times, the mode is less useful.

This statistical use of the word **mode** corresponds to one of the uses of the same word in ordinary language. We sometimes refer to fashion as what is in mode, or in common use. This implies a particular fashion which is more common than others.

It will be seen that the type of scale or measurement is important in determining which measure of central tendency is to be used. For example, the median requires at least an ordinal scale, and the mean requires an interval or ratio scale. In contrast, the mode requires only a nominal scale. Since all scales are at least nominal, the mode of a set of data can always be calculated. That is, regardless of whether a variable is nominal, ordinal, interval or ratio, it will have a mode.

5.2.1 Mode for Ungrouped Data

For a set of data which is ungrouped, the mode is determined by counting the number of times each value of the variable occurs. The mode is then the value of the variable which occurs more frequently than any other value of the variable.

Definition 5.2.1 Mode of Ungrouped Data. For ungrouped data, the **mode** is the value of the variable which occurs most frequently.

When there is more than one value of the variable which occurs most frequently, then there is more than one mode. That is, the mode is not necessarily unique, and a set of data could possibly have two or more modes.

These situations could be referred to as bimodal or trimodal. Further, there may be no unique mode. That is, each value of the variable may occur an equal number of times. In this situation, each value could be considered to be a mode, but this is not particularly useful, so that this situation would ordinarily be reported as having no mode.

Example 5.2.1 Mode in a Small Sample

The example of Table 5.1 was briefly discussed in Chapter 4. This is a small random sample of 7 respondents selected from the data set of Appendix **??**. *For each of the variables, the mode is given in the last line of the table. A short discussion of these modes follows.*

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
Mode	Female	34	NDP	--	40	Middle	37,500

Table 5.1: A Sample of 7 Respondents

For this sample, the mode for each of the variables can be determined by counting the frequency of occurrence of each value. For the variable `SEX`, *there are 5 females and 2 males, so that the mode of* `SEX` *in this sample is female. For the variable* `AGE`, *the value 34 occurs three times, while each of the other ages occurs only once. This means that the modal age of the respondents in this sample is 34. The mode of political preference is NDP, with 5 respondents supporting the NDP and only one respondent supporting each of the other two parties. The mode of gross monthly pay,* `GMP`, *is not unique, with each value occurring only once. The mode of hours worked per week is 40, and of social class is middle class, with these values occurring most frequently.*

Family income, `FIN`*, has a mode of $37,500 since this income occurs twice, and each of the other values of family income occur only once. While a family income of $37,500 is the mode, this may not be a particularly useful measure. This value just happens to occur twice, whereas each of the other values occurs once, only one less time.*

Example 5.2.2 Mode from a Stem and Leaf Display

The ordered stem and leaf display of Table 5.1 was given in Chapter 4. Once such a stem and leaf display has been produced, the mode can easily be determined by counting which value of the variable occurs most frequently. In this example, the value 40 occurs 20 times, a considerably larger number of times than any other value. Thus the mode of total hours worked for this sample of 50 Saskatchewan households is 40 hours worked per week.

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	0	8	8
5	0	0	0	0	0	4																

Figure 5.1: Ordered Stem and Leaf Display for Total Hours of Work

It may also be useful to note that the value which occurs next most frequently is 37 hours per week, occurring 7 times. This could be considered a secondary, although much less distinct, mode. Finally, the value of 50 hours worked per week occurs 5 times, and this might be considered to be a tertiary mode. Reporting these secondary or tertiary modes is not required when reporting the mode, but noting and reporting such secondary and tertiary modes may be useful in understanding the distribution of the variable. In the case of total hours worked, 40 hours per week is by far the most common, with 37 and 50 hours per week also being common hours worked. Together these three values account for $20 + 7 + 5 = 32$ *of the 50 values reported.*

In the stem and leaf display of Example 4.5.3 and Figure 4.5, the mode of family income is 9 thousand dollars, with this value occurring 6 times. The next most common value is 11 thousand dollars, occurring 3 times. None of the other values of family income occur more than twice.

5.2.2 Mode for Grouped Data

When data has been grouped, the mode is determined on the basis of the category or interval which contains the largest number of cases. If the variable is nominal, this is a straightforward process of examining the frequency distribution and determining which category has the largest number of cases.

Where the data is ordinal or interval, more care must be taken in determining the mode, because adjacent values of the variable may be grouped together. This may make some categories occur more frequently than others, as a result of the manner in which the data has been grouped. In the case of such grouped data, the mode occurs at the peak of the histogram, that is, where the density of occurrence of the variable is greatest.

Definition 5.2.2 Mode of Grouped Data - Nominal Scale or Equal Size Intervals. For grouped data where the scale in nominal, the **mode** is the value of the variable which occurs most frequently. Where an ordinal, interval or ratio scale has been grouped into intervals of equal width, the mode is the interval having the greatest frequency of occurrence. In the case of a continuous variable, the mode can be considered to be either the interval with the greatest frequency of occurrence, or the midpoint of this interval.

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

Table 5.2: Provincial Political Preference, 50 Regina Respondents

Example 5.2.3 Political Preference

The distribution of political preference of 50 respondents in Table 5.2 was originally given in Chapter 4. The variable here is political preference and this is measured on a nominal scale. The value of the variable which occurs most frequently is NDP, occurring 24 times, more than the number of times any other value occurs. Thus NDP is the modal value of political preference. Note that this is quite a distinct mode, in that this value occurs in 24 of the 50 cases. With respect to elections, the mode is an important measure, because the party receiving the most votes wins the election. While political preference based on polls is not a sure indicator of how people will vote on election day, the mode of political preference is an important guide in determining the likely winner.

Income in dollars	Per Cent of Individuals of Native Origin	All Origins
0-4,999	30.7	19.0
5,000-9,999	23.5	20.1
10,000-14,999	13.4	13.8
15,000-19,999	9.4	11.3
20,000-24,999	7.4	9.4
25,000 and over	15.6	26.3
Total	100.0	100.0
No. of individuals	253,980	17,121,000

Table 5.3: Distribution of Individuals with Income, Native and All Origins, Canada, 1985

Example 5.2.4 Income Distributions of People of Native and of All Origins

The distribution of income of people of native origin and of all origins in Canada in 1985, in Table 5.3 is taken from Table 3.4 of **The Canadian Fact Book on Poverty 1989** *by David P. Ross and Richard Shillington. The variable 'income of individuals' is measured on an interval or ratio scale. With the exception of the last, open-ended interval, the data has*

been grouped into intervals of equal size, each interval representing $5,000 of income. The distributions in the table are both percentage distributions, so that the interval with the largest percentage will be the modal interval, that is, the interval with the largest number of cases.

For the income distribution of individuals of native origin, the interval which has the largest percentage of cases is $0-4,999 and this is the mode for this distribution. For individuals of all origins in Canada, the mode appears to be the $25,000 and over interval. However, this interval is much wider than the other intervals, and contains such a large percentage of the cases mainly because it is a wide open ended interval. Based on this, the mode for this distribution would best be considered to be the interval $5-9,999, because this interval has the largest percentage of cases of those intervals of equal width.

In each of these two distributions, the mode could alternatively be reported as the respective midpoint of the interval. This would give a mode of approximately $2,500 for individuals of native origin, and $7,500 for individuals of all origins.

Definition 5.2.3 Mode of Grouped Data - Ordinal, Interval or Ratio Scale with Unequal Size Intervals. For grouped data where the scale is ordinal, interval or ratio scale, and where the data has been grouped into intervals of different widths, the **mode** is the interval with the **greatest density of occurrence**. Alternatively, the mode is the value of the variable where the **histogram reaches its peak**.

Example 5.2.5 Canada Youth and Aids Study

The data in Table 5.4 is taken from Figure 6.6 of Alan J. C. King et. al., **Canada Youth and AIDS Study**. *This data come from a 1988 study of 38,000 Canadian Youth. The percentage distributions in Table 5.4 give the distributions of the number of sexual partners of university or college males and females who had had sexual intercourse at the time of the survey. The sample sizes on which these distributions are based is not given in the figure in the publication. The frequency distributions given here are presented with the categories used in the original publication.*

A quick glance at the frequency distributions might give the impression that the mode of the number of sexual partners is 3-5 for males and 1 for

Number of Partners	Per Cent of Individuals Who Are: Male	Female
1	23	36
2	12	17
3-5	29	26
6-10	17	14
11 +	19	7
Total	100	100

Table 5.4: Distribution of Number of Sexual Partners of College and University Respondents, by Gender

females. However, the intervals 3-5 and 6-10 represent more than one value of the variable, while the categories 1 and 2 each represent one value of the variable. For females, the most frequently occurring category is 1, and since that is already an interval representing a single value of the variable, it can be considered to be the mode.

For males, the densities of occurrence for the different intervals must be calculated in order to determine which value of the variable is the mode. The density for each of the first two categories is the percentage of cases in each category, since each of these already represents exactly 1 unit of the variable, 'number of sexual partners'. The interval 3-5 represents 3 values of this variable, 3, 4 and 5. The density of occurrence here is thus $29/3 = 9.7$*. The density of the intervals 6-10 and 11 and over need not be calculated since the percentage of cases on each of these intervals can be seen to be lower than that of the 3-5 interval, and these last two intervals are even wider than the 3-5 interval. As a result of these considerations, the mode of the number of sexual partners for males is also seen to be 1 partner. This category has 23 per cent of the cases, more than the 12% having two partners, and also more than the average of 9.7 per cent of males having each of 3, 4 or 5 partners.*

Example 5.2.6 Examples from Chapter 4

The correct histogram of Figure **??** *gives a peak bar at socioeconomic status of 50-60. The mode of socioeconomic status is 50-60 or 55. Note though that the 40-50 category occurs almost as frequently.*

The two distributions of wives' education Figure **??** *and Figure* **??** *have peaks at 11.5-12.5. The mode of years of education of wives of each income level is 12 years of education.*

The distribution of farm size in Figure **??** *shows an initial peak at the interval 'under 10' acres. However, this is most likely to be an anomaly based on the rather odd grouping that was given in Table* **??**. *Based on this table, one would have to conclude that the mode is under 10 acres. However, the rest of the distribution shows a distinct peak at 240-399 acres, and this would be a more meaningful mode to report. If the interval 'under 10' were to be regrouped with the interval 10-69, this would produce an interval 0-69, of width 70 acres, and with density* $(593 + 1107)/70 = 24.3$ *farms per acre. Based on this, the interval with greatest density is then 240-399, and the modal farm size could be reported to be 240-399, or 320 acres. This would be more meaningful than reporting that the modal farm size in Saskatchewan is under 10 acres.*

The distribution of the number of people per household in Table **??** *might best be reported as yielding two modes. The number of households with 2 people is most common, with 286 households. The number of households with only 1 person, or with three persons is distinctly less than this. However, a secondary peak is reached at 4 persons per household, with 223 households reporting this many people. The distribution has these two distinct peaks, with 2 being the mode, or the primary mode, and with 4 persons per household being a secondary mode. The two distinct peaks can clearly be seen in Figure* **??**.

Summary of the Mode. Since the mode requires only a nominal scale of measurement, and since all variables have at least a nominal scale, the mode can always be determined. All that is required is to count the values of the variable which occur in the sample, and determine which value occurs most frequently. This value is the mode. In the case of grouped data, the mode is either the value of the variable at the peak of the histogram, or the category or interval which occurs with greatest density. In the latter case, the midpoint of the interval with greatest density may be chosen as the median.

The mode is most useful as a measure in two circumstances. First, if

the category which occurs most frequently is all that needs to be known, then the mode gives this. In elections, all that matters is which party or candidate gets the most votes, and in this case the modal candidate is the winner. For purposes of supplying electrical power, power utilities need to be sure that they have sufficient capacity to meed peak power needs. The peak use of power could be considered to be the mode of power use. In circumstances of this sort, the mode is likely to be a very useful measure.

Second, the mode is useful when there is a very distinct peak of a distribution. Where many of the values occur almost an equal number of times, the mode may not be clearly defined, or may not be of much interest. Where a particular value occurs many more times than other values, then it is worthwhile to report this as the mode.

At the same time, the mode has several weaknesses. First, in no way does it take account of values of the variable other than the mode, and how frequently they occur. For example, in Figure **??**, for wives with annual income of less than $5,000 annually, 11 years of education is almost as common as 12 years of education, but the mode is 12. For Figure **??**, the mode is much more clearly 12 years, with 11 years being much less commmon.

Second, in the case where the scale is ordinal, interval or ratio, the mode does not take advantage of the numerical values which the variable takes on. When determining the mode, the variable could just as well be nominal in all cases. Where a variable has numerical values on an ordinal scale, these values can be used to rank the values. If the numerical values have been measured on an interval scale, these values can be added and subtracted. This extra information provided by these numerical values can be used to provide other measures of central tendency. These are discussed in the following sections.

5.3 The Median

The median is a second measure of central tendency, and one which takes advantage of the possibility of ranking or ordering values of the variable from the smallest to the largest, or largest to the smallest value. If this ordering is carried out, the median is the middle value of the distribution, the value such that one half of the cases are on each side of this middle value. The median is thus the centre of the distribution in the sense that it splits the cases in half, with one half less than this centre and one half greater than this centre.

Definition 5.3.1 The **median** of a set of values of a variable is the value of the variable such that one half of the values are less than or equal to this value, and the other half of the values of the variable are greater than or equal to this value.

The median requires an ordinal scale of measurement. Since interval and ratio scales are also ordinal, the median can be determined for any variable which has an ordinal, interval or ratio scale of measurement. For variables which have a scale of measurement no more than nominal, the median cannot be meaningfully determined.

Since the median gives the central value of a set of values, the median is more properly a measure of central tendency than the mode. The mode could occur at one of the extremes of the distribution, if the most common value of the variable occurs near one end of the distribution. By definition, the median always will be at the centre of the values of the variable.

As with the mode, there are different methods of calculating the median depending on whether the data is ungrouped or grouped. These methods are discussed next, with examples of each method being provided.

5.3.1 Median for Ungrouped Data

In order to determine the median value for a distribution, it is necessary to begin with a variable which has an ordinal or higher level of measurement. The method used to determine the median is to begin by taking the values of the variable in the data set, and **ranking these values in order**. These values may be ranked either from the smallest to the largest, or largest to the smallest. It does not matter in which direction the values are ranked. The total number of values in the data set is then counted, and one half of this total is the central value. In order to determine the median, count the ordered values of the variable until you reach one half of the total values. This middle value is the median. If there are an odd number of values, there is a single central or median value of the variable. If there are an even number of total values, there will be two middle values. The median is then reported as either these two values, or the simple average of these two middle values. These precedures should become clearer in the following examples.

Example 5.3.1 Median in a Small Sample - Odd Number of Values

The small data set of 7 cases presented in Chapter 4 and in Table 5.1 shows how the median can be calculated. Only the variables `AGE` `GMP`, `THRS` *and* `FIN` *are clearly ordinal or higher level scales and thus have meaningful medians. If the variable* `CLASS` *is considered to be ordinal, then the median can also be determined for this variable.*

The ages of the 7 respondents in this data set are 32, 33, 34, 34, 46, 34 and 53. Ranked in order from lowest to highest, these values are 32, 33, 34, 34, 34, 46 and 53. Note that where there is a value which occurs more than once, each occurrence of that value is listed. Since there are a total of 7 values here, with one half of this being $7/2 = 3.5$*, the median is the value such that 3.5 cases are less and 3.5 are greater. This value is the middle 34 of the list and the median for this data set is 34 years of age.*

For `GMP`*, the values in order are 700, 1300, 1850, 1950, 2500, 3125 and 5000. The middle value here is 1950. For* `THRS`*, total hours of work per week, the median is 40, the mdidle value of the set 16, 25, 40, 40, 40, 50, 50. Finally, for family income,* `FIN`*, the median is $42,500.*

If the variable `CLASS`*, social class that the respondent considers himself or herself to be in, is considered to be ordinal, then the median can be determined for this variable. This is the case even though the variable does not have numbers but names. If these names are given in order, they are Upper middle, middle, middle, middle, middle, working and working, where these values have been ranked from the highest to the lowest. For these 7 values, the median is middle class.*

Example 5.3.2 Median in a Small Sample - Even Number of Values

The small data set of 7 cases can be increased by 1 case in order to produce 8 values, an even total number of cases. This is done in Table 5.5. Case 8 with the values as shown in the table has been added to the data set. Here only the variables which are ordinal or higher level scales are presented. The median is given in each case. For the age of respondents, the values of the variable when ranked from low to high, are 32, 33, 34, 34, 34, 39, 46 and 53. With 8 values, the 4th and 5th values are the middle two cases. Since these are both 34, the median is 34. For gross monthly pay, the values in order are 700, 1300, 1850, 1950, 2500, 2690, 3125 and 5000. The median could be reported as the middle two values, $1,950 and $2,500 or, more commonly it would be reported as $2,225, the average of these two values, that is, $(2500 + 1950)/2 = 2225$*. The median for the other variables is determined in a similar manner.*

Case No.	AGE	GMP	THRS	CLASS	FIN
1	32	1300	50	U. Middle	42,500
2	33	1950	25	Working	27,500
3	34	2500	40	Middle	37,500
4	34	1850	40	Middle	52,500
5	46	5000	50	Middle	100,000
6	34	700	16	Working	37,500
7	53	3125	40	Middle	62,500
8	39	2690	37	Working	39,000
Median	34	2225	40	Middle	40,750

Table 5.5: Small Sample of 8 Respondents

Example 5.3.3 Median in a Stem and Leaf Display

If the data set has been organized into a stem and leaf display, the median can be determined quickly. An ordered stem and leaf display presents the data in order, from the smallest value in the data set, to the largest value in the data set. Again the method is to determine the middle value. In the stem and leaf display of Figure 5.1, there are 50 values for the variable, total hours of work per week. The middle values for this are the 25th and 26th values. This value is 40 hours of work per week. This can be quickly determined by counting the number of values of the variable in each row. There are 2 values in the first row, 2 more in each of the second and third rows, and 16 in the third row, for a total of 22 values to this point. In the 4th row, the values in the forties, the first value of 40 is the 23rd value, the second the 24th and the third and fourth values in that row are the 25th and 26th values in the data set. These values are each 40 hours worked per week. As a result, the median hours worked per week for this data set is 40. If the two middle values had been different from each other, the two values could be reported as the median, or the simple average of these two middle values would be reported as the median.

5.3.2 Median for Grouped Data - Discrete Variable

Where a variable measured on an ordinal or higher level scale has already been grouped into categories, it is presented as a frequency distribution. Suppose the variable is labelled X, and the frequencies of occurrence for the different values of X are f. That is, each value of the variable X_i occurs with respective frequency f_i. The categories in the frequency distribution are values of the variable X, and these are ordinarily given in order from the lowest to the highest value of the variable. Since the values of the variable X are already ordered, the median can be determined by finding the value of the variable X which represents the middle case in the data set.

If the total number of cases is n, then the value of the variable X at which the $(n/2)$th case occurs is the median. In finding this value, a **cumulative frequency** column is likely to be useful. The **cumulative frequency** is the number of cases less than or equal to the value of the variable. That is, for each value of the variable X, the cumulative frequency is the sum of the frequencies of occurrence of the variable f for all the values of the variable less than or equal to X. Once the cumulative frequency has been obtained, the value of X at which the $(n/2)$th case occurs is the median. This procedure becomes clearer in the following example.

Example 5.3.4 Median Number of People per Household

In Chapter 4, the distribution of the number of people per household in a sample of 941 Regina households was given in Table **??**. *This table is repeated here in Table 5.6, but with a cumulative frequency column added. For each value of the variable X, the cumulative frequency column gives the sum of the frequencies of occurrence of the variable up to and including that value of X. That is, for $X = 1$, there are 155 cases, so that 155 of the households have 1 or less people in them. Next, there are $155 + 286 = 441$ households with 2 or less people in them, 155 with 1 person and 286 with 2 people. For $X = 3$, there are $155 + 286 + 164 = 605$ households with less than or equal to 3 people in the household. Households with 4 or less people total $155 + 286 + 164 + 223 = 828$, so the cumulative frequency up to $X = 4$ is 828. The remaining cumulative frequencies are determined in the same manner.*

Also note that for each value of X, the cumulative frequency is the previous cumulative frequency plus the number of cases that take on that value of X. For $X = 3$, the cumulative frequency is 605, and this equals $441 + 164$, the previous cumulative frequency of 441, plus 164, the number of

cases which have $X = 3$. Once the cumulative frequency has been obtained,

X	f	Cumulative frequency
1	155	155
2	286	441
3	164	605
4	223	828
5	86	914
6	21	935
7	5	940
8	1	941
Total	941	

Table 5.6: Frequency and Cumulative Frequency Distribution of Number of People per Household

the determination of the median is a relatively straightforward procedure. The total number of cases is $n = 941$, and the median occurs at the middle case, that is, case $941/2 = 470.5$. The value of X here is $X = 3$. That is, there are 441 households having 2 or less people per household, and 605 households having 3 or less people per household. The 470th or 471st household is the median household, and this is one of the 164 households having exactly 3 persons per household. As a result, for this distribution, the median is $X = 3$ persons per household, with one half of the households having 3 or less people per household, and the other half of the households having 3 or more people per household.

As an alternative to using the cumulative frequencies, it may be preferable to construct **cumulative percentages**. The cumulative percentage for any value of the variable X is the percentage of the total cases which have a value less than or equal to X. If the percentages are already given in a percentage distribution, the cumulative percentage column is constructed in a manner exactly analogous to the construction of cumulative frequencies, except that the cumulative percentages are added. Then the median occurs at the half way point, that is, the value of the variable where one half, or

50 per cent, of the cases have been accounted for. The value of the variable which contains this 50 per cent point is the median. This is illustrated in the following example.

Example 5.3.5 Attitudes toward Immigration

The data in Table 5.7 comes from a survey of Alberta residents conducted by the Population Research Laboratory at the University of Alberta. The question asked of a sample of Alberta residents was "whether or not Canada should allow more immigrants". The responses to this question were measured on a 7 point scale with 1 representing strongly disagree, 4 being neutral and 7 being strongly agree. This scale is a discrete, ordinal level scale, with the categories being ranked in order from 1 to 7.

Label	Response X	Per Cent	Cumulative Per Cent
Strongly Disagree	1	31	31
	2	15	46
	3	13	59
Neutral	4	18	77
	5	11	88
	6	7	95
Strongly Agree	7	5	100
Total		100	

Table 5.7: Percentage and Cumulative Percentage Distributions of Attitudes toward Immigration

The data were originally given in percentages, and in order to determine the median, it is not really necessary to know the sample size on which this set of data was based. The cumulative percentage column is constructed by adding the percentages in the per cent column one at a time. For attitude 1, there are 31% of the respondents, so that the cumulative percentage column gives 31% of respondents with a attitude of 1 or less. For attitude 2 or less, there are $31 + 15 = 46$ *per cent of respondents. The per cent of respondents with attitudes of 1, 2 or 3 is* $31 + 15 + 13 = 59$ *per cent. In a similar manner,*

the percentages in the per cent column can be cumulated until all 100% of the respondents are accounted for.

It can be seen that the median attitude occurs at attitude level 3. Moving from 1, the lowest attitude, strongly agree, to 2, there are only 46% of the respondents. By the time category 3 has been included, 59% of the respondents have been accounted for. The median, or 50% point, is at attitude level 3. That is, one half of all respondents have attitude at level 3 or less (1, 2, or 3), and the other half of respondents have attitudes are level 3 or more (3 through 7).

Using the cumulative percentage method may be preferable to using the cumulative frequency method. By definition the median is the value of the variable which has one half, or 50 per cent, of the cases at less than or equal to this value, and the other half, or 50 per cent, at greater than or equal to this value. Once the cumulative percentages have been obtained, the 50 per cent point can clearly be seen in the cumulative percentage column. If the data has originally been presented as a frequency distribution, first construct the percentage distribution, and then from this construct the cumulative percentage column. In the case of a discrete variable, the median is then the value of the variable at which the cumulative percentage first exceeds 50 per cent.

5.3.3 Median for Grouped Data - Continuous Variable

For data which has been measured on an continuous scale, and which has been grouped into categories or intervals, the determination of the median is not as straightforward as in the case of ungrouped data. This section describes how to compute the median for such grouped data on the basis of straight line or linear interpolation.

Definition 5.3.2 The **median** for a variable X is the value of the variable, X_m, such that 50 per cent of the values for the population or sample are less than or equal to X_m and the other 50 per cent are greater than or equal to X_m.

For a continuous variable, it is easiest to see how to use linear interpolation to determine the median on the basis of an example. The following example shows how the median for a continuous socioeconomic status scale can be obtained. Following that, a more general formula is given.

Example 5.3.6 Median for a Status or Prestige Scale

Table 5.8 presents a distribution of status or prestige for 322 Regina respondents from the Social Studies 203 Labour Force Survey. This status or prestige scale is based on the Blishen scale of socioeconomic status which measures the status of the occupations of respondents on the basis of a combination of each respondent's education and income. The scale presented here combines years of education and gross monthly pay to produce a value of the variable X, where X is a measure of socioeconomic status. Since this variable combines two interval level scales, X can also be regarded as having an interval scale. The frequency distribution for X is as shown in Table 5.8. The number of respondents in each category of socioeconomic status is given by f, representing the frequency of occurrence of X. In order

X	f	Per Cent	Cumulative Per Cent
0-20	2	0.6	0.6
20-30	49	15.2	15.8
30-35	85	26.5	42.3
35-40	71	22.0	64.3
40-45	49	15.2	79.5
45-50	29	9.0	88.5
50-60	27	8.4	96.9
60-70	8	2.5	99.4
70-80	2	0.6	100.0
Total	322	100.0	

Table 5.8: Distribution of Socioeconomic Status - 322 Regina Respondents

to obtain the median for this distribution, it is best to begin by converting the frequency distribution into a percentage distribution. This is done in the per cent column of Table 5.8. Then the **cumulative percentages** can be calculated, as shown in the last column. This cumulative percentage column measures the per cent of cases that have a value less than or equal to the upper limit of each interval.

The first category of socioeconomic status 0-20, has 2 of the 322 respondents, or $(2/322) \times 100\% = 0.6\%$ of respondents. The cumulative percentage of respondents up to $X = 20$ is also 0.6%. For the category 20-30, there

are 49 respondents, or $(49/322) \times 100\% = 15.2\%$ of respondents. Up to a socioeconomic status of 30 there are thus $0.6 + 15.2 = 15.8$ per cent of respondents. For the 30-35 category of socioeconomic status, there are $(85/322) \times 100 = 26.5$ per cent of respondents. For a socioeconomic status of 35 or less, there are thus $15.8 + 26.5 = 42.3$ per cent of respondents.

Determining the Median. Once the cumulative percentages have been obtained, the next step in determining the **median** status level for this distribution is to locate the interval in which the median lies. Then determine the value of X, within that interval, which most closely corresponds to the median value. Since the median is the 50 per cent point, the median must be in the interval 35–40. This can be determined by looking at the cumulative percentage column. By the time status level 35 has been accounted for, there are only 42.3% of the respondents. However, by the time an X value of 40 has been reached, 64.3% of all the cases have been accounted for. Thus, the 50% point must be somewhere in this interval between 35 and 40. As a rough guess, the median value of X is close to a socioeconomic status of 37, closer to 35 than 40, because 50% is closer to 42.3% than it is to 64.3%.

In order to determine a more exact value for the median, the method of **linear or straight line interpolation** can be used. This method assumes that the cases in the interval are evenly or uniformly spread across the interval in which the median lies. In this method, the first step is to determine the proportion of the distance along the interval required to get to the median. This proportion is then converted into a specific value for X.

Begininning with the interval in which the median is located, the proportion of the distance required to reach the median is the distance from the lower end point of the interval to the 50 per cent point, divided by the width of the interval. In percentages, this is the distance $50 - 42.3 = 7.7$ percentage points, divided by 22.0 percentage points, the percentage of cases in the interval. This proportion is

$$\frac{50 - 42.3}{22.0} = \frac{7.7}{22.0} = 0.35.$$

This means that 0.35 of the way along the interval is the point at which 50 per cent of the cases are reached. In terms of values of status, variable X, the interval is 5 units wide. The median then occurs at $0.35 \times 5 = 1.75$ units of status above the lower end point of the interval. Since the interval begins

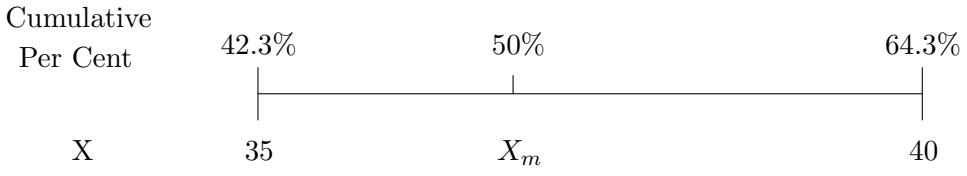

Figure 5.2: Interval containing the Median

at $X = 35$, and the median is 1.75 units of status above this, the median is $35 + 1.75 = 36.75$. All this is summarized in the following formula.

$$\text{Median} = 35 + \left(\frac{50 - 42.3}{22.0}\right) \times 5 = 35 + (0.35 \times 5) = 36.75$$

One half of the respondents thus have status of 36.75 or less and the other half have status of 36.75 or more, based on straight line interpolation.

This method can be illustrated with the help of the diagram of Figure 5.2. Again, first locate the interval in which the median lies. This is the interval from $X = 35$ to $X = 40$ and it can be seen that at $X = 35$, 42.3% of the cases are accounted for. At the upper end of the interval, at $X = 40$, 64.3% of the cases have occurred. Putting all this information on one line, with the values of X below the line and the cumulative percentages above the line, gives Figure 5.2.

The median, X_m, is located about a third of the way between the endpoints of 42.3% and 64.4%. The distance from the lower endpoint to the median can be seen to be $50\% - 42.3\% = 7.7\%$. In percentages, there are 22.0 per cent of the cases in this interval, so that the distance from the lower end point to the median is $7.7/22.0 = 0.35$ of the total distance in the interval. Now, this proportion must be converted into units of status or of X. Since the interval represents 5 units of status, the distance to the median in units of status is $0.35 \times 5 = 1.75$. This means that the median occurs at $35 + 1.75 = 36.75$ units of status. All these calculations can be placed on a similar diagram, as shown in Figure 5.3.

Method for determining the Median of a Continuous Variable. A more general formula for determining the median is as follows. Begin by

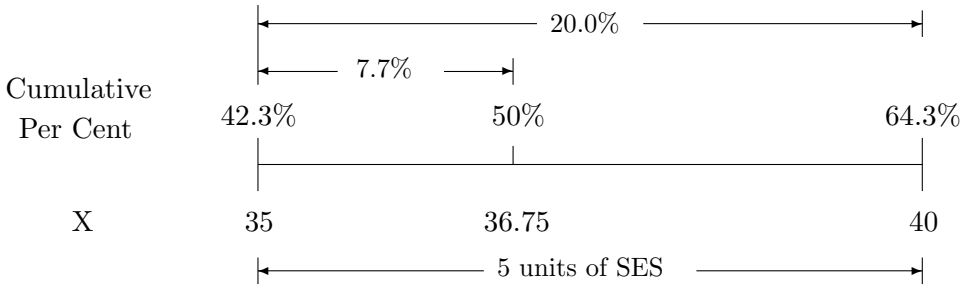

Figure 5.3: Determining the Median

constructing the percentage and the cumulative percentage columns. Then locate the interval in which the median lies. Within this interval, the median is

$$\text{Value of the variable at the lower end of the interval} + \left[\frac{50 - \text{Cumulative per cent at lower end of the interval}}{\text{Per cent of cases in the interval}}\right] \left[\text{Interval width}\right]$$

Example 5.3.7 Hours Worked per Week for Youth

The first and third columns of Table 5.9 come from page 29 of the publication **Canada's Youth: "Ready for Today"**, *by Donald Posterski and Reginald Bibby, prepared by the the Canadian Youth Foundation for the Minister of State for Youth, 1987. The table gives the percentage distribution of the hours worked per week for youth who were in the survey conducted for the study. Note that in this case the real class limits should be used because there is a gap between the upper end of one interval and the lower end of the next interval. Whenever this gap amounts to a significant portion of the value of variable being measured, the real class limits should be used. For this example, the median is in the interval from 10.5 to 30.5 hours worked per week because at 10.5 hours, only 19 per cent of the cases*

Hours Worked Per Week	Real Class Limits	Per Cent	Cumulative Per Cent
<10	0.5-10.5	19	19
11-30	10.5-30.5	36	55
31-40	30.5-40.5	28	83
41-50	40.5-50.5	12	95
50+	50.5+	5	100
Total	100		

Table 5.9: Distribution of Hours Worked Per Week

have been accounted for, but by 30.5 hours, 55% of the cases are accounted for. The median is thus

$$\textit{Median} = 10.5 + \left[\frac{50 - 19}{36}\right] \times 20 = 10.5 + (0.861 \times 20) = 27.7.$$

Based on this method the median hours worked for youth in the survey was 27.7 hours or, rounded to the nearest hour, 28 hours worked per week. The diagrammatic representation of these calculations is given in Figure 5.4.

Summary of the Median. The median provides a measure of the centre of a data set, in that one half of the cases are on each side of the median. For this reason, the median is sometimes referred to as the typical value of the data set. Whether the data is ungrouped or grouped, the median is determined by ranking the cases from the lowest to the highest value of the variable, and then determining the value of the variable in the middle. In the case of grouped data, it is often easiest to do this by constructing a percentage and a cumulative percentage distribution, and using these to determine the median. Where the data have been grouped into intervals, linear interpolation between the real class limits denoting the ends of the intervals is required. Where the variable has integer values, no such interpolation is required. Following the section on the mean, in Section 5.5, some of the properties of the median are discussed.

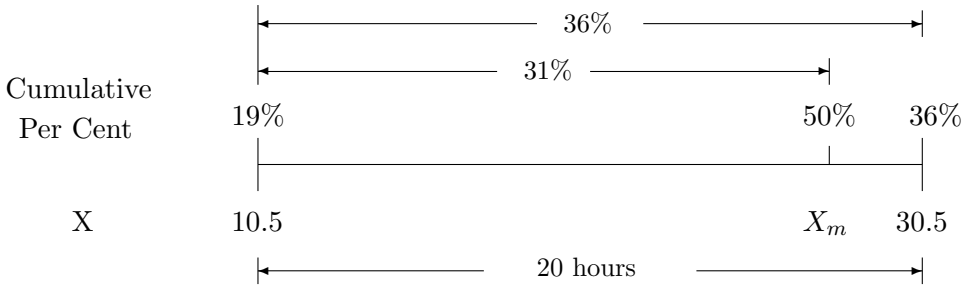

Figure 5.4: Median of Hours Worked for Youths

5.4 The Mean

The most widely used measure of central tendency is the **arithmetic mean**, or more simply as the **mean**. In ordinary usage, this measure is more commonly referred to as the average. Since each of the median, mode and mean are sometimes referred to as the average, it is best to refer to the arithmetic mean as the mean, rather than the average, in order to distinguish it from these other measures. When 'average' is used in this textbook, it is used in a rough sense, to denote the centre of a distribution, and could mean any of the mode, median or mean. Note also that the arithmetic mean is only one of the possible means that can be calculated; other means sometimes used in statistics are the geometric mean and the harmonic mean. In this text, only the arithmetic mean will be used, and it will be referred to as simply the mean.

The mean is the well known average, that is, the sum of the set of values of the variable, divided by the number of values. For example, if a student obtains 82%, 76%, 69%, 71% and 66% in 5 classes taken in a semester, the mean grade for these 5 classes is

$$\frac{82 + 76 + 69 + 71 + 66}{5} = \frac{364}{5} = 72.8$$

The student's mean grade for the semester is 72.8%, or rounded off to the nearest percentage point, the mean grade for these 5 classes is 73%.

Note that the mean is determined by first calculating **the sum of all the values** of the variable that are in the data set. Then this total of all the values is **divided by the number of cases**. The mean is thus an average based on the notion of splitting this total value equally among all the cases. As another example, suppose 4 people have incomes of 25, 32, 87 and 43 thousand dollars. The total income for these four people is 187 thousand dollars. If this were to be split equally among these 4 people, this would amount to 46.75 thousand dollars each. This value of $46,750 is the mean income for these four people.

This idea of splitting the total equally in the principle on which the calculation of the mean is always based. However, the exact procedure for calculating the mean differs depending on whether the data is ungrouped or grouped. The procedure for each of these cases is outlined in the following sections. The notation introduced in Chapter 4 becomes especially useful here, and some new notation is introduced as well.

5.4.1 Mean for Ungrouped Data

For ungrouped data, the mean is calculated as just shown. That is, the sum of all the values of the variable is determined, and this is divided by the number of values. This can be stated more formally using the notation introduced in Chapter 4.

Definition 5.4.1 If a variable X takes on values

$$X_1, X_2, X_3, \cdots, X_n$$

then the **mean** of this set of values is

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

That is, the sum of all n of the values of X is computed and this total is divided by the number of values, n. In the above example of the five grades, $n = 5$. If the grade for each class is represented by X_i, where $i = 1, 2, 3, 4, 5$, these values are

$$X_1 = 82$$
$$X_2 = 76$$
$$X_3 = 69$$
$$X_4 = 71$$
$$X_5 = 66$$

Using the formula in Definition 5.4.1, the mean grade is given by

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{82 + 76 + 69 + 71 + 66}{5} = \frac{364}{5} = 72.8$$

Bar Notation. The algebraic symbol $\bar{X}$, with the bar on top of the X, denotes the mean value of the variable X. In statistical work, the mean of a variable is usually denoted by a bar on top of the symbol for the variable. For example, if the variable is Y, then the mean value of Y is $\bar{Y}$; the mean of the variable c would be denoted by $\bar{c}$.

Summation Notation. Sums of algebraic values can be represented with a summation notation. The uppercase Greek letter Sigma, written Σ, is commonly used in statistical and algebraic work to denote a sum of values. In this text, this symbol Σ is termed the **summation sign**, rather than sigma. The summation in Definition 5.4.1 can be written as follows:

$$\Sigma X = X_1 + X_2 + X_3 + \cdots + X_n$$

With this notation, the definition of the mean can be more economically written as:

$$\bar{X} = \frac{\Sigma X}{n}$$

The summation notation is quite flexible, and more details can be given on the summation sign by using subscripts. A more complete notation for the sum of the n values

$$X_1, X_2, X_3, \cdots, X_n$$

would be

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \cdots + X_n$$

That is, the set of values over which the summation is to proceed is given as part of the summation sign. The indexes on the summation sign and on

the variable are given in order to denote the set of values to be summed. In this case, below the summation sign is $i = 1$, meaning that the summation is to begin with the first value of the variable X. This value is X_1, that is, beginning the sum with the value of X for $i = 1$. The values of i are all integer values so that after $i = 1$, the next value is X_2, when $i = 2$. This process is continued, with i increasing by 1 in each case. The values to be added are values 1, 2, 3, 4 and so on. Above the summation sign is the index value n, denoting that the sum is to end with value n. This means that all values between $i = 1$ and $i = n$ are to be added, where i increases by 1 in each case.

Using the index notation along with the summation sign, the mean grade in the above example could be written as

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{364}{5} = 72.8$$

As an example of how the summation notation could be used, suppose that a variable Y has values Y_i, where $i = 1, 2, 3, 4, 5, 6$. Suppose that only values 3 through 5 are to be added. This sum could be written

$$\sum_{i=3}^{5} Y_i$$

and this would represent the sum

$$\sum_{i=3}^{5} Y_i = Y_3 + Y_4 + Y_5$$

In the case of the mean, all n values of the variable must be added in order to obtain the total of all values of the variable in the data set. Because all values are added when calculating the mean, the index in often dropped in this text. Unless stated otherwise, in this text it is assumed that the summation occurs across all the values of the variable. Thus the mean may be written in short notation as

$$\bar{X} = \frac{\sum X}{n} \quad \text{or} \quad \bar{X} = \frac{\sum X_i}{n}$$

meaning

$$\bar{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

or in even more detail,

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

The summation notation is used extensively throughout this text. As new uses are encountered, some further instructions concerning the algebra of the summation notation will be given. There are are considerable number of rules concerning the use of the summation sign, but the use of the summation sign allows formulae to be presented in a relatively compact notation.

5.4.2 Mean for Grouped Data

When the data has already been grouped into a frequency distribution, the formula for computing the mean involves using both the values of the variable X, and the frequencies of occurrence f. The basic method is the same as with ungrouped data, to first produce a total of all the values of the variable, and then divide this total by the number of cases, n. In the case of data which has been grouped into a frequency distribution, the mean is defined as follows.

Definition 5.4.2 If a variable X takes on values

$$X_1, X_2, X_3, \cdots, X_k$$

with respective frequencies

$$f_1, f_2, f_3, \cdots, f_k$$

then the **mean** of this set of values is

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + f_3X_3 + \cdots + f_kX_k}{n}$$

where

$$n = f_1 + f_2 + f_3 + \cdots + f_k$$

Using this formula, it is necessary to be careful concerning the order in which various operations are carried out. The numerator of the expression for $\bar{X}$ is

$$f_1X_1 + f_2X_2 + f_3X_3 + \cdots + f_kX_k.$$

This could be written

$$(f_1X_1) + (f_2X_2) + (f_3X_3) + \cdots + (f_kX_k)$$

in order to make clear that each value of f is first multiplied by each value of X, and then this total is summed. That is, multiply f_1 by X_1, f_2 by X_2, and so on until the kth value of the variable, X_k, is multiplied by f_k. Finally, each of these individual products is added to produce the numerator in the above expression.

This formula is sometimes referred to as the **weighted mean**, that is, with each value of the variable X_i being weighted by its respective frequency

of occurrence f_i. By multiplying each value of X_i by each f_i, the total of the values of the variable for all n cases is produced. As in the earlier formula, this total is then divided by the number of cases n. This procedure is illustrated in the following example.

Example 5.4.1 Mean Number of People per Household

A frequency distribution for the number of people per household for 941 Regina respondents was given in Chapter 4 in Tables **??** *and* **??**. *This frequency distribution, along with the calculations for determining the mean is given in Table 5.10. In this table, X represents the number of people per household. There are $f_1 = 155$ respondents with $X_1 = 1$ persons per household, and this results in $f_1X_1 = 155 \times 1 = 155$ people. Following this there are $f_2 = 286$ respondents with $X_2 = 2$ persons per household, and this results in $f_2X_2 = 286 \times 2 = 572$ people. For $X_3 = 3$ people per household, there are $f_3 = 164$ respondents. The total number of people is these households is $f_3X_3 = 164 \times 3 = 492$ people. The same procedure is used through the remainder of the table, until all values of X and f have been included. The sum of the column fX gives the total number of people*

X	f	fX
1	155	155
2	286	572
3	164	492
4	223	892
5	86	430
6	21	126
7	5	35
8	1	8
Total	941	2,410

Table 5.10: Calculations for Mean Number of People per Household

that are in all the households. This is the numerator in the expression for the mean $\bar{X}$. The total number of cases in the data set is the sum of the frequencies of occurrence, and the sum of these f values can be seen to be

$n = 941$ *cases. The mean is thus*

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + f_3X_3 + \cdots + f_kX_k}{n}$$

$$\bar{X} = \frac{155 + 572 + 492 + 892 + 430 + 126 + 35 + 8}{941} = \frac{2,410}{941} = 2.561$$

The mean number of people per household is 2.561, or rounding this off to the nearest tenth of a point, $\bar{X} = 2.6$ *people per household.*

Note in the above example that the mean does not necessarily result in an integer value, even though the variable is a discrete, integer valued variable. In this case of people per household, an individual household must have a discrete value for people per household, that is, an integer value between 1 and 8. However, if the mean or average is desired, then this method can produce any value of the variable. This happens because the total number of people in all of the households is 2,410, and this total value is split equally among all respondents in the data set. Spreading these 2,410 people equally across 941 households produces a mean value of 2.6 people per household.

Summation Notation. Definition 5.4.2 can be restated using summation notation, and this results in a compact presentation of the formula. In order to do this another rule concerning the summation notation is required.

The sum

$$f_1X_1 + f_2X_2 + f_3X_3 + \cdots + f_kX_k$$

can be written more compactly as

$$\sum_{i=1}^{k} f_iX_i$$

If the subscripts and superscripts are dropped from the summation sign, this becomes

$$\sum f_iX_i$$

or if all indexes are dropped, this sum can be written simply as

$$\sum fX.$$

That is,

$$\sum f_iX_i = f_1X_1 + f_2X_2 + f_3X_3 + \cdots + f_kX_k$$

When working out this summation, the rule is that the first part of the operation is to multiply each value of f by each value of X, and after each of these products have been calculated, the products are summed. In general, when carrying out summations, the operations to the right of the summation sign are carried out first, and then the results of these operations are summed. In this case, this means that each f_i is first multiplied by each X_i, and then these individual products f_iX_i are added. In order to make clear the order of operations, the products f_iX_i can be bracketed, indicating that these products are computed first. This gives the expression

$$\sum_{i=1}^{k}(f_iX_i) = (f_1X_1) + (f_2X_2) + (f_3X_3) + \cdots + (f_kX_k)$$

Based on this, the definition of the mean can be restated as follows.

Definition 5.4.3 If a variable X takes on values X_i, where $i = 1, 2, \cdots k$, with respective frequencies f_i, $i = 1, 2, \cdots k$, then the **mean** of this set of values is

$$\bar{X} = \frac{\sum_{i=1}^{k}(f_iX_i)}{n}$$

where

$$n = \sum_{i=1}^{k} f_i.$$

If the subscripts are dropped, then this becomes simply

$$\bar{X} = \frac{\sum fX}{n}$$

Percentages and Proportions. If the data has been grouped into either a percentage or proportional distribution, the above definitions can be modified, as follows, to give formulae for the mean based on these distributions.

Definition 5.4.4 If a variable X takes on values X_i, where $i = 1, 2, \cdots k$, with respective percentages P_i, $i = 1, 2, \cdots k$, then the **mean** of this set of values is

$$\bar{X} = \frac{\sum_{i=1}^{k}(P_iX_i)}{100}$$

where

$$\sum_{i=1}^{k} P_i = 100$$

since the sum of all the percentages in a percentage distribution is 100%. If the subscripts are dropped, then this becomes simply

$$\bar{X} = \frac{\sum PX}{100}$$

In the case of a proportional distribution, the percentages P_i are replaced by proportions p_i, $i = 1, 2, \cdots k$, and the **mean** is

$$\bar{X} = \sum_{i=1}^{k} (p_i X_i)$$

where

$$\sum_{i=1}^{k} p_i = 1$$

If the subscripts are dropped, then this becomes simply

$$\bar{X} = \sum pX$$

Note that in the formula for the mean of a percentage distribution, the n in the denominator of the original formula for the mean is replaced by 100, the sum of all the percentages. In the case of the proportional distribution, the sum of all the proportions is 1, so that the expression $\sum pX$ is divided by 1. But dividing by 1 leaves this value unchanged, so that that the expression becomes merely $\bar{X} = \sum pX$ in the case of a proportional distribution.

Data Grouped into Intervals. So long as a variable is discrete, and the values of the variable X are given singly in the table of the distribution (as in Example 5.4.1), then the above formulae are straightforward to apply. In many distributions, though, the data have been grouped so that the categories are intervals which represent a range of values of X. When this is the case, the formulae just presented cannot be applied without deciding which value of X is to be used. The usual procedure is to let the midpoint of each interval be the value of X which will be used in the formula for determining the mean. This is illustrated in the following example.

Example 5.4.2 Mean Years of Education of Saskatchewan Wives

In Chapter 4, Table **??** *gave data concerning the distribution of the years of education completed for Saskatchewan wives earning less than $5,000 annually and for those earning $5,000 or more annually. Table 5.11 gives the calculations for determining the mean years of completed education of the sample of Saskatchewan wives. The calculations for wives earning less than $5,000 annually are in columns 2-4, and for wives with greater earnings in columns 5-7. For these distributions, the variable X is years of education completed, and the f values are as given in the third and sixth columns of Table 5.11.*

Note that in these frequency distributions, some values of years of education completed are grouped together, while others are given as discrete values. For the categories 0-10, 13-17 and 18-22, a value of X must be selected in order to calculate the mean. The value chosen here is the midpoint of each of the intervals. For the interval 0-10, the midpoint is 5, for 13-17 the midpoint is 15, and for 18-22 the midpoint is 20. These values are given in the second and fifth columns as the X values for purposes of calculating the mean. For 11 and 12 years of education, the values of X are 11 and 12, respectively. The fX column is calculated as the product of each of the

Education in Years Completed	Wives Earning <$5,000 Annually			Wives Earning $5,000+ Annually		
	X	f	fX	X	f	fX
0-10	5	41	205	5	18	90
11	11	12	132	11	13	143
12	12	15	180	12	31	372
13-17	15	23	345	15	29	435
18-22	20	6	120	20	14	280
		97	982		105	1,320

Table 5.11: Calculations for Determining Mean Education Level of Saskatchewan Wives, by Earnings

entries of the respective values of X and f as given in columns two and three of the table.

For the wives with lower earnings, the first set of 41 wives have 0-10 years of education, so their total number of years of education is $5 \times 41 = 205$, *assuming that these wives average 5 years of education. For the 12 wives with 11 years of education,* $fX = 11 \times 12 = 132$. *After all the entries in the* fX *column have been calculated, these are added and this produces a total of 982 years of education. Also note that the sum of the* f *column is the sample size, that is,* $n = \sum f = 97$. *The mean is then the total value of 982 years of education divided by 97, the number of wives,*

$$\bar{X} = \frac{\sum fX}{n} = \frac{982}{97} = 10.124$$

The mean number of years of education of wives earning less than $5,000 annually is 10.1 years of education completed.

This mean is best rounded off to the nearest tenth of a year, or even to the nearest year, because the calculation involved the approximation that the midpoint of each interval represents the average number of years of education of wives in that category. This may not be quite accurate, for example, those wives having 0-10 years of education may average a little more than 5 years of education. However, in the absence of more detailed information, choosing the midpoint of each interval seems best.

The calculations for the wives earning $5,000 or more annually use the same X *values since the original categories were the same. Based on the Table 5.11, the mean is*

$$\bar{X} = \frac{\sum fX}{n} = \frac{1,320}{105} = 12.571$$

The mean years of education of these wives with greater earnings is 12.6 years.

Note that the wives earning $5,000 or more have a mean years of completed education that is approximately 2.5 years more than wives earning less than $5,000 annually. When the modes for this distribution were reported in Example 5.2.6, both groups had a mode of 12 years, so the mode did not differentiate these two distributions very well. The mean provides a better measure of central tendency for distinguishing these two distributions, because it shows that wives with greater earnings average considerably more years of education completed than do wives with lower levels of annual earnings.

Real or Apparent Class Limits? When determining the midpoint of each interval, as in Example 5.4.2, it makes no difference whether the real or the apparent class limits are used. Regardless of whether the real of the apparent class limits are used, the midpoint of each interval is the same. This is because the construction of real class limits from apparent class limits means adding an equal amount to each end of the interval. This leaves the midpoint of the interval unchanged.

When the distributions of wives' education was first examined in Tables **??** and **??**, the real class limits for the intervals 13-17 years of education completed were 12.5 to 17.5. The interval 13-17 was extended by 0.5 years on each end to eliminate the gap between the category of $X = 12$ and the interval 13-17. The midpoint of 12.5 to 17.5 is $(12.5 + 17.5)/2 = 30/2 = 15$, and the midpoint of 13-17 is $(13 + 17)/2 = 15$. The other intervals give similar results.

Based on these considerations, it make no difference whether real or apparent class limits are used when calculating the mean. If a distribution is presented with apparent class limits which leave gaps between intervals, the real class limits need not be constructed if only the mean is to be calculated for this distribution. If the median is to be determined, then the real class limits must be used. Of course, if there is no gap between intervals, so that the apparent and real class limits are identical, then these need not be altered for determination of any of the measures of central tendency.

Open Ended Intervals. As noted in Chapter 4, data is often presented so that it has open ended intervals. If the mean is to be determined for such a distribution, some value has to be entered for X for the open ended interval when using the formula for the mean. Exactly what this value should be is not readily apparent from the table of the distribution. About all that can be done in these circumstances is to guess the approximate value for X in the open ended interval. The appropriate value of X for any interval is the mean value of the variable for the cases in that interval. For example, in the case of wives years of completed education in Example 5.4.2, the appropriate value of X for wives who have completed 11 years of education is $X = 11$. For wives who have completed 13-17 years of education, the value used was $X = 15$, and hopefully this approximately represents the mean years of education completed for all the wives who have 13-17 years of education.

About all that can be done in the case of open ended intervals is to

pick a value of X which seems reasonable based on what is known about the distribution of the data. Do not pick a value too high, or too low, but pick a value which you think approximately represents the mean value of the variable for the set of cases in the open ended interval. If the mean is calculated from such a distribution, always make sure that the value of X which is picked is reported along with the mean. Someone else making the same calculation will produce a different mean if the value of X picked is somewhat different.

Examples of calculations of the mean are given below. Study each example and note the rationale given for the value of X used in the case of the open ended intervals. You may disagree with the rationale, but if you do so, make sure you report the value of X which you select. In addition, there may be other slight modifications to the data in the following examples. These will be carried out in order to illustrate a variety of ways in which distributions which are reported in publications can be used to produce reasonably accurate estimates of means.

Example 5.4.3 Mean Income of Individuals of Native and All Origins

Table 5.3 presented percentage distributions of individual incomes of people of native and all origins in Canada for 1985. In this example, a percentage distribution is used to estimate the mean income of people of native origins in Table 5.12. In Table 5.13, the distribution of incomes of individuals of all origins is presented as a proportional distribution, and the mean is calculated on that basis.

In both of Tables 5.12 and 5.13, the variable X is income of individuals in 1985 dollars, a ratio scale. In order to make the mean a little easier to calculate, incomes are converted into incomes in thousands of dollars. In addition, the endpoints of each interval have been rounded off to the nearest thousand dollars. For example, the interval 0-4,999 dollars is rounded to 0-5 thousand dollars. This is because there is very little difference between $4,999 and $5,000, and given the other approximations that are being made when the mean is being calculated, this difference is too small to be of any importance in the result. By rounding in such a manner, the midpoints X of the intervals involve less decimal places, and make the calculation a little easier. Based on this rationale, the midpoints are 2.5 thousand dollars for the 0-4,999 interval, 7.5 for the 5,000-9,999 interval, and so on.

For the last, open ended interval a value of 30 thousand dollars is selected in Table 5.12. Hopefully this comes close to approximating the mean income

of all those individuals of native origin who have incomes of $25,000 or more. It might be argued that 27.5 would be a more appropriate X value for this interval, since the list of X values goes from 2.5 to 7.5 to 12.5, increasing by 5 each time. 5 more than 22.5 would be 27.5. However, choosing $X = 27.5$ for this interval is equivalent to assuming that there are very few individuals with an income over $30,000. That is, choosing $X = 27.5$ is equivalent to assuming that the last interval is 25-30 thousand dollars. But in fact there are individuals who have over $30,000 income, and to take into account that some of these may have incomes of considerably more than $30,000, a mean of $30,000 is selected for this interval. Once the X and P columns

Income in Dollars	Income in thousands of Dollars	X	P	PX
0-4,999	0-5	2.5	30.7	76.75
5,000-9,999	5-10	7.5	23.5	176.25
10,000-14,999	10-15	12.5	13.4	167.50
15,000-19,999	15-20	17.5	9.4	164.50
20,000-24,999	20-25	22.5	7.4	166.50
25,000 and over	25+	30.0	15.6	468.00
Total			100.0	1,219.50

Table 5.12: Distribution of Individuals with Income, Native Origin, Canada, 1985

are constructed as shown, the calculation of the mean becomes relatively straightforward. This is a percentage distribution, with each value of P being multiplied by each value of X to produce the PX column. This total is entered into the formula for the mean as follows.

$$\bar{X} = \frac{\sum PX}{100} = \frac{1,219.50}{100} = 12.195$$

The mean is thus 12.195 thousand dollars, or $12,195. Given the approximations that have been made in this calculation, it might be best to round the mean to the nearest $100 and report it as $12,200, for perhaps round it to the nearest thousand dollars and report it as $12,000.

In Table 5.13, values of X are as in Table 5.12, with the exception of the open ended interval. In addition, to illustrate how a mean is calculated for a proportional distribution, the percentages of Table 5.3 have been divided by 100 to give the proportion of individuals of all origins who have each level of income.

For the distribution of incomes of individuals of all origins, a value of $X = 35$ *thousand dollars has been selected for the open ended interval. This is in recognition of the fact that there are over one quarter of all respondents in the open ended interval. Ordinarily this would indicated that the mean income for this interval is considerably more than $25,000. Whether $35,000 is an accurate estimate of this mean income for the open ended interval is not clear, but a value somewhat above the value selected in Table 5.12 should be used in recognition of the fact that there are more individuals in the open ended interval here as compared with the distribution for individuals of native origin. The mean income for individuals of all origins is*

Income in Dollars	Income in thousands of Dollars	X	p	pX
0-4,999	0-5	2.5	0.190	.4750
5,000-9,999	5-10	7.5	0.201	1.5075
10,000-14,999	10-15	12.5	0.138	1.7250
15,000-19,999	15-20	17.5	0.113	1.9775
20,000-24,999	20-25	22.5	0.094	2.1150
25,000 and over	25+	35.0	0.263	9.2050
Total			1.000	17.0050

Table 5.13: Distribution of Individuals with Income, All Origins, Canada, 1985

$$\bar{X} = \sum pX = 17.005$$

In this case the summation is divided by 1, the sum of the proportions. Since dividing any number by 1 does not change the number, the summation of the products of the proportions times the X values is the mean. Thus the mean

for individuals of all origins is $17,005, or rounded to the nearest hundred or thousand dollars, is $17,000.

Based on these results, it can be seen that the mean income for individuals of native origin is approximately $5,000 below the mean income for individuals of all origins. Looking back at the two original distributions of Table 5.3, this is no surprise. The distribution of incomes of individuals of native origin has large percentages of the population at lower incomes. Compared with those of native origin, for individuals of all origins there are larger percentages at some middle income levels, and especially in the highest income category, that of individuals earning $25,000 or more.

Example 5.4.4 Mean Hours Worked Per Week for Youth

Example 5.3.7 gave a table of the percentage distribution of hours worked for Canadian youth. Table 5.14 gives this data again here, with the calculations for the mean also given.

For the open ended interval with less than 10 hours, it seems safe to conclude that this interval is really 0-10, although it might be 1-10, the publication does not say. Assuming it is 0-10, this produces a midpoint of $X = 5$ for this first interval. For the open ended interval of 50 and over hours worked per week, there are relatively few cases, so that a value of not too much above 50 should be chosen for this interval. In addition, hours worked are not likely to be too much greater than 50 for most of these people. Probably most work less than 60 hours. Based on these considerations, a value of $X = 55$ has been selected as representing the mean of the hours worked for the youth in the 50 and over interval. The mean hours worked is given by

$$\bar{X} = \frac{\sum PX}{100} = \frac{2,648.0}{100} = 26.48$$

The mean hours worked per week for these youth is 26.48 hours per week. In view of the approximations involved in calculating this value, the mean should be rounded to at least the nearest tenth or an hour and reported as 26.5 hours. It may be preferable to round to the nearest hour and reported the mean hours worked per week for youth as 26 hours.

Hours Worked Per Week	X	P	PX
<10	5	19	95.0
11-30	20.5	36	738.0
31-40	35.5	28	994.0
41-50	45.5	12	546.0
50+	55	5	275.0
Total		100	2,648.0

Table 5.14: Calculation for Mean Hours Worked Per Week for Canadian Youth

5.5 Uses of Measures of Centrality

Given that there are three measures of centrality or central tendency, choices sometimes must be made concerning which of these measures to calculate and report. This section provides some guidelines concerning which measure to use in various circumstances. Some of the advantages and disadvantages of each of these measures will also be discussed.

Type of Scale. The first consideration involved in deciding on a measure of central tendency is the type of scale which has been used to measure the variable. If the scale is no more than nominal, only the mode can be determined. Where a scale is ordinal, but not interval or ratio, the mode and the median can be meaningfully determined. Only in the case of interval or ratio scales can all three of mode, median and mean be obtained.

In the case of an ordinal scale, sometimes the mean is calculated. For example, a mean may be determined for an ordinal level attitude scale. This is shown in the following exmple.

Example 5.5.1 Mean Attitude Level for a Sample of Alberta Resident

Example 5.3.5 gave the distribution of attitudes concerning whether or not Canada should accept more immigrants, for a sample of Alberta res-

idents. The percentage distribution from that example is given again in Table 5.15. The attitudinal responses X are given on a 7 point scale. Based on the percentages reported in the P column, it can be seen that the mean attitude level of these respondents is

$$\bar{X} = \frac{\sum PX}{100} = \frac{304}{100} = 3.04$$

The mean attitude level is approximately $\bar{X} = 3$, and is one point on the disagree side of neutral. The mean implies that, on average, this sample of Alberta residents tends to disagree a little with the view that Canada should allow more immigrants. Note that the median is also 3, although the modal response is 1, with 31% of the responses begin strong disagreement. The main reason the mean would be used in this example is that the mean is somewhat easier to use in statistical analysis of the type encountered later in the textbook.

	Response		
Label	X	P	PX
Strongly Disagree	1	31	31
	2	15	30
	3	13	39
Neutral	4	18	72
	5	11	55
	6	7	42
Strongly Agree	7	5	35
Total		100	304

Table 5.15: Calculations for Mean Attitude

In the last example, the mean was calculated for an ordinal level attitude scale. In carrying out this calculation, **the assumption is that each unit on the attitude scale represents an equal value of attitude**. That is, the implicit assumption is that the distance between an attitude of 1 and 2 is 1 unit, and this represents the same difference in attitude as does the difference in attitudes between 4 and 5, for example. That is, the ordinal scale is being treated as if it had been measured on an interval scale.

While this is technically illegitimate, this is commonly done in Sociology and Psychology, where ordinal scales are common. So long as the researcher is careful, and does not interpret these results too strongly, this may be done. However, the researcher should always be aware of the implicit assumption being adopted, and should interpret means of this sort with caution.

Symmetric Distributions. If a distribution is **symmetric**, then the median and mean are equal to each other. If the distribution peaks at this same point, as is likely to be the case, then the mode is also equal to the median and the mean. Where these three measures are exactly equal to each other, then it makes no difference which value is reported. Since the mean is most commonly used in Statistics, the mean would ordinarily be the measure reported in such a circumstance.

For distributions which are asymmetric, it is necessary to decide which measure of central tendency to use. Of course, if the scale if interval or ratio, all three measures can be reported, and this is sometimes quite useful. At other times one measure will be preferable over another. The following paragraphs discuss some of the circumstances that lead to selection of a particular measure.

Type of Issue Being Examined. Measures of central tendency sometimes lend themselves toward a particular use. The **mode** is most useful if a researcher is interested in the most common value of the variable in a data set, and this most common value is clearly identifiable. In the case of an election result, the party or the candidate receiving the largest number of votes is the winner of the election. In this case, all that matters is which is the modal party or candidate. In the case of traffic patterns at an intersection or on a road, the capacity of the road will be an important consideration. This capacity, or peak use, can be considered to be the mode of use, and this must be adequate to handle traffic needs.

In the social sciences there are sometimes situations where a most common value is clearly identifiable, and does become important. In the last example of attitudes toward immigration, Example 5.5.1, the *strongly disagree* category stands out as having considerably more respondents than does any other category. Since almost one-third of respondents strongly disagree that Canada should allow more immigrants, this shows an antagonism toward immigration that is likely to make its voice felt. Overall, the mean and median of 3 indicates a broad range of responses, with an average re-

sponse being slightly disagree, not an indication of the serious antagonism toward immigration that the modal response shows. In this example, it would probably be worthwhile to report both the median and the mode of attitudes.

The **median** is the middle value of a set of data, and is used to indicate the value which splits the data set so that one half of the values are less than or equal to the median and one half are greater than or equal to the median. Since the median depends only on where the centre of the distribution is, and does not depend on the values on either side of the centre, it is often considered to be the best measure of the centre. It is sometimes referred to as a *typical* value of the variable, especially in distributions where the bulk of cases are fairly near the centre.

One situation where the median is commonly used is to indicate the centre of the age distributions of populations. Populations which have high birth rates and relatively low death rates, as in many Asian countries, are relatively young populations, with a median age of under 20 years. For example, in Pakistan the median age of the population was 17.5 years in 1985. This means that in 1985, half of the population of Pakistan was less than 17.5 years old, the other half of the Pakistani population was older than 17.5 years. By contrast, in Canada, with an even lower death rate, and much lower birth rate, has a population with a median age of approximately 31 years. Roger Sauvé, in **Canadian People Patterns**, pp 18-19, notes that "the median age will rise to about 37 by the year 2000 and to 49 by the years 2036." If these forecasts prove correct, by 2036 one half of all people living in Canada would be 49 or older, while the other half would be less than 49 years of age.

The median is also used to report the middle value of income distributions. As will be noted below, the mean may be an unreliable and misleading indicator of typical income levels in some circumstances. The median income of a population tends to be an income which can be regarded as more typical of a population than is the mean income. Sometimes definitions of poverty are given on the basis of the median income. For example, if the median income is relatively typical, researchers might regard as poor anyone with an income of less than half the median level of income.

Some of the differences in the mean and median can be seen in Table 5.16. In this table, the mean and median income for Canada, calculated in dollars of 1990, is given. This data comes from Statistics Canada, **Income Distributions by Size in Canada 1990**, Ottawa, 1991, catalogue number 13-207, Table 1, page 45. The severe effect of the recession of the

Year	Mean	Median
1980	49,843	45,876
1983	47,056	41,753
1986	49,276	44,052
1989	52,471	46,576
1990	51,633	46,049

Table 5.16: Mean and Median Family Income in 1990 Dollars, Canada, Selected Years

early 1980s on Canadian family incomes can be seen by comparing either the mean or median of 1983 with that of 1980. Note though that median family income declined by over $4,000 in these three years, while the mean shows a decline of approximately $2,800. Comparing 1990 with 1980 shows that mean family income increased by approximately $2,000 over the decade. In contrast, the median income for 1990 is very little more than it was in 1980. Over the decade, mean income may have increased because the income of some higher income groups rose, while the incomes of poorer or middle income groups changed little. Given the small increase in median family income between 1980 and 1990, the table suggests that this is what may have happened in the latter part of the 1980s, although considerably more evidence would be required before we could be sure this is what happened.

The **mean** is the most commonly used measure of central tendency, at least for interval and ratio level scales. The mean is the total of all values of the variable, divided by the number of cases. As a description of the centre of the distribution, the mean is equivalent to assuming an equal division of the total across all members of the data set. For example, when comparing countries, income per capita of each country is often reported. This is the mean income of each country, where the total income has been divided by the number of people in that country. While income is unequally distributed within each country, income per capita provides an idea of the *average* level of living in each country, or an idea of the relative wealth of each country.

Where this notion of equal division of a total is a useful way of describing populations, the mean is likely to be the measure of central tendency used. The difficulty with the mean is that it does hide the variation within each of these populations. For example, women are often said to perform better

than men on some tests of verbal abilities, but less well than men on tests of mathematical abilities. While the exact meaning of this is not always made clear, most likely this means that women have lower mean scores on tests of mathematical abilities than do men, while on tests of verbal abilities women have higher mean scores than men. But this hides the variation in test scores for each group, and does not mean that all men are better at mathematics, and all women are better at verbal tasks. In each case there is a distribution of test scores, and the nature of each distribution, and the manner in which they differ must be understood. In this case, the means may mislead more than help understanding the differences between groups.

If a total value of a variable is to be estimated, knowledge of the mean may assist in this. Suppose that a representative sample of people are selected from a population, and the mean income of these people is found to be \$17,300 per year. If the total income of all members of the population is to be estimated, then this can be done by multiplying the mean income for the sample times the number of people in the population. This produces a reasonable estimate of the total value of income for the population. This can be seen be looking at the formula for the mean. If the sample has sample size n, the mean income of the sample is

$$\bar{X} = \frac{\sum X}{n}$$

so that

$$\sum X = \bar{X} \times n$$

is the total value of X for all the cases in the sample. Analogously, if N is the size of the population, the total value of X for the whole population is

$$\sum X = \bar{X} \times N$$

If $\bar{X}$ is known from the sample, and if the size of the population N is known, then multiplying $\bar{X}$ by N gives an estimate of the total value of the variable for the whole population. If the mean income of a sample of people in a small city is \$17,300, and there are 9,300 people in the city, then the total income for the whole city is $17,300 \times 9,300 = 160,890,000$.

Type of Distribution. A further consideration when deciding on the measure of central tendency is the nature of the distribution. In particular, in an asymmetric distribution, the mean may provide a misleading measure of the centre of a distribution. This is illustrated in the following example.

Example 5.5.2 Distribution of Salaries of Employees

Suppose that 7 employees of a firm have salaries of 25, 27, 22, 31, 135, 37 and 31 thousand dollars annually. The mean income for these 7 employees is

$$\bar{X} = \frac{\sum X}{n} = \frac{25 + 27 + 22 + 31 + 135 + 37 + 31}{7} = \frac{308}{7} = 44$$

or \$44,000. As can be seen, no employee has an income of very close to \$44,000, and the mean is the total income of \$308 thousand, split equally across all 7 employees. But in fact, one employee is very well paid, and the other 6 employees are paid at a much lower rate. The mean pay is somewhere in between these two set of values.

In this example, the values are placed in order from the lowest to the highest value are 22, 25, 27, 31, 31, 37 and 135. The median is the fourth value, and is equal to 31. Here the median provides a more typical value, closer to what might be considered the centre of the distribution.

It is also worthwhile to note what happens when the single large income is removed from the list of values. In the case of the median, the values from low to high are 22, 25, 27, 31, 31, and 37. The two middle values are 27 and 31 so the median is $(27+31)/2 = 29$. *The mean income for the 6 employees is*

$$\bar{X} = \frac{\sum X}{n} = \frac{25 + 27 + 22 + 31 + 37 + 31}{6} = \frac{173}{6} = 28.833$$

or \$28,800. In the case of these six employees, where the set of values of the variable contains no extreme values, the median and mean are very similar, and either measure provides a good measure of the centre of the distribution.

What is notable is the extent to which the mean is sensitive to a single, or a few, extreme values. A few very large or very small values can have a considerable influence on the mean value. The median is not at all sensitive to these extreme values, changing only slightly if a few very large or very small values are added to the data set. As can be noted from the last example, which of the two measures is most desirable depends on how the researcher wishes to include these extreme values. If the total value of the variable is to be split equally across all values when reporting the measure of central tendency, then the mean is an appropriate measure. If the researcher wishes to report a more typical value, closer to what most people would consider the centre of a distribution, then the median is a preferable measure.

Open Ended Intervals. When the frequency distribution is organized so there are open ended intervals, this creates a problem for calculating the mean. As noted earlier, this involved making an educated guess concerning the mean value of X for the cases in the open ended interval.

An open ended interval does not prevent obtaining a fairly accurate calculation of the median, provided that the median is not in the open ended interval. This is one advantage of the median over the mean. If the frequency or percentage distribution table has open ended intervals, this presents no barrier to reasonably accurately computing the median. This is because the median is a positional measure, measuring only the position at which 50 per cent of the cases have been accounted for. The exact size of the other values of the variable has no real influence on the median, all that influences the median is the ranking of the cases. Percentiles, discussed in the next section, have the same characteristic as the median. That is, percentiles are positional measures, and can be determined reasonably accurately even when there are open ended intervals.

Other Statistical Uses. Regardless of the various advantages and disadvantages of the three measures of central tendency, most statistical work is carried out using the mean. There is a long historical tradition of using the mean, and when the average is reported, most people assume this is the mean. For example, grade point averages, average temperatures, and average incomes are usually calculated using the mean. Most of the formulae and statistical methods used later in this text are based on the mean as the measure of average. Statisticians have investigated the characteristics of the mean in great detail. In addition, if new research is being carried out, and it is to be compared with earlier research work, then there is little choice but to report the mean, since that is likely what previous researchers have used.

In spite of the extensive use of the mean, it is best to consider which measure of central tendency is most appropriate in each circumstance. For many uses, the median may provide a superior measure of the centre of a distribution. Where this is the case, it may be best to report both the median and mean, the former because it gives a better idea of the centre, and the latter because it is so widely used, and because so many statistical formulae are based on the mean.