Contents

5.6 Percentiles

The median is a measure of the middle, or 50 per cent point, of a distribution. As such, it is a positional measure, indicating the value of the variable where this 50 per cent point is reached. Instead of the 50 per cent point, some other per cent could be requested, not as a measure of the middle of the distribution, but as the position where this alternative percentage of cases has been accounted for. These positional measures, based on the percentage of cases up to and including that value, are termed **percentiles**.

As an example, a researcher may wish to determine the value of income such that only 20 per cent of the population has lower income, with the other 80 per cent having a higher level of income. This would then give a measure of the income level below which the poorest one fifth of the population lie. Standardized tests such as the Graduate Record Examination (GRE) or the Law School Admission Tests (LSAT) give results in percentiles. For example, if a student scores in the 85th percentile on the GRE, this means that 85 per cent of the students who took the test have lower scores, and only 15 per cent of those who took the test have higher scores. For obtaining admission to graduate school, the percentile obtained is likely to be of more importance than the actual score, since graduate schools are interested in admitting students who have the highest scores. If a student is in only the 23rd percentile, that would be an indicator that only 23 per cent of all those who took the test scored lower, while 77 per cent scored higher. This indicates a relatively poor performance on the test.

As in the case of the median, percentiles can be determined only if the variable is measured on a scale which is ordinal, interval or ratio. Where the values of the variable have been grouped into intervals or categories, the method used to determine percentiles is the same as the method used for the median with grouped data. That is, the cumulative frequency or percentage is used, and where the values of the categories are not discrete, integer values, linear interpolation is used to determine the values of the various percentiles.

Definition 5.6.1 The rth **percentile** for a variable is the value of the variable, P_r, such that r per cent of the values for the population or sample are less than or equal to P_r and the other $100 - r$ per cent are greater than or equal to P_r.

For example, the 20th percentile of an income distribution is the value of income, P_{20}, such that 20 per cent of the population has an income less than

or equal to P_{20} and the other 80 per cent of the population has an income greater than or equal to P_{20}. The **median** is the 50th percentile since the median has one half of the cases less than or equal to the median and the other half are greater than or equal to the median. Using this notation, $X_m = P_{50}$.

The method of calculating the percentiles is discussed in the following sections. Only the method for grouped data is shown here.

5.6.1 Percentiles for a Discrete Variable

If a variable has a discrete set of values, and each of these values is given in a frequency distribution, then the determination of percentiles proceeds in the same manner as was used for calculating the median. That is, the cumulative frequency or percentage distribution is obtained, and the value of the variable at which the requested percentage of cases occurs is determined. This is illustrated in the following example.

Example 5.6.1 Percentiles for People per Household

Earlier in this Chapter, in Example **??**, *the distribution of people per household 941 Regina households was given. This frequency distribution and cumulative frequency distribution is given again here in Table 5.1. Suppose the 30th percentile and the 68th percentile are desired. These can be determined as follows.*

As a first step, begin by constructing the cumulative frequency distribution. This gives the frequency of occurrence of the each value of the variable, up to the value of X shown in each row. The 30th percentile is the value of X such that 30 per cent of the cases are less than this, and the other $100 - 30 = 70$ per cent of cases are greater. Since there are 941 cases in total, 30 per cent of this is $(30/100) \times 941 = 282.3$. The 282nd or 283rd value is P_{30}. This means that $X = 2$ is the 30th percentile. That is, for the first value of X, $X = 1$, there are only 155 cases. For $X = 2$, there are another 286 cases, so that the number of cases up to and including $X = 2$ is $155 + 286 = 441$. The 282nd and 283rd values are both exactly at $X = 2$. Thus $P_{30} = 2$. The 68th percentile occurs at the $(68/100) \times 941 = 639.88$ case. That is, the 639th or 640th value is at $X = 4$. By the time all the households with 3 people have been accounted for, there are only 605 households, but once the 223 households with 4 people in them have been included, there are 828 households. The 639th and 640th households are at 4 people per household, so that $P_{68} = 4$.

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 5.1: Frequency and Cumulative Frequency Distribution of Number of People per Household

Example 5.6.2 Percentiles for Percentage Distribution of Attitudes

In Example **??**, *a percentage distribution of attitudes toward immigration into Canada was presented. The percentage and cumulative percentage distribution is given in Table 5.2. Suppose the 23rd, 58th and 92nd percentiles are desired for this distribution.*

Again proceed by first constructing the cumulative percentage distribution. Once this has been done, the appropriate percentiles can be read from the cumulative percentage distribution. The 23rd percentile occurs at attitude level $X = 1$, *because the 31 per cent of cases with the lowest values of* X *occur here. This is more than the lowest 23 per cent, so all these lowest 23 per cent occur at this first value of* X. *Thus* $P_{23} = 1$.

The 58th percentile occurs at $X = 3$, *because attitudes 1, 2, and 3 account for the lowest 59 per cent of values, more than the 58% requested. Thus* $P_{58} = 3$. *Based on the same reasoning, the 92nd percentile can be seen to occur at* $X = 6$. *Only 5 per cent of respondents have larger values of* X *on the attitude scale, and the top two values of* X *include 12% of all the cases, more than the top 8% associated with the 92nd percentile. As a result,* $P_{92} = 6$.

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 5.2: Percentage and Cumulative Percentage Distributions of Attitudes

5.6.2 Percentiles for Grouped Data - Continuous Variable

When the values of the variable have been grouped into intervals, then it is necessary to use linear interpolation to produce an accurate estimate of each percentile. Beginning with a frequency distribution, construct a percentage distribution for the variable, and then a cumulative percentage distribution.

In order to determine the rth percentile of the distribution, locate the interval in which the rth percentile lies, using the cumulative percentages. Then using this interval, the rth percentile, P_r, is

$$\begin{array}{l}\text{Value of the variable at}\\ \text{the lower end of the in-}\\ \text{terval}\end{array} + \left[\begin{array}{l}\dfrac{r - \begin{array}{l}\text{Cumulative per cent at}\\ \text{lower end of the inter-}\\ \text{val}\end{array}}{\begin{array}{l}\text{Per cent of cases in the}\\ \text{interval}\end{array}}\end{array}\right] \left[\begin{array}{l}\text{Interval}\\ \text{width}\end{array}\right]$$

As can be seen by comparing this formula with the formula for the median in Section **??**, the only change is to replace the value of 50 by the value r, where r is the percentile that is desired. Also note that the **real class limits** should be used for the values of the variable at the ends of each interval, and in order to determine the proper interval width.

Example 5.6.3 Percentiles for a Status or Prestige Scale

Table 5.3 presents the distribution of status or prestige for 322 respondents in the Social Studies 203 Labour Force Survey, originally given in Example **??**. *This distribution will be used to find the 75th percentile, P_{75}, and the thirteenth percentile, P_{13}.*

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 5.3: Distribution of Socioeconomic Status

In order to obtain percentiles, begin by converting the frequency distribution into a percentage distribution. This is done in the per cent column of Table 5.3. Then a **cumulative percentage** *column can be calculated, as shown. Recall that the cumulative percentage column gives the per cent of cases that have a value less than or equal to the upper limit of each interval.*

In order to determine the 75th percentile, P_{75}, of socioeconomic status, first find the interval within which this percentile lies. Note that at $X = 40$, 64.3% of the cases have been accounted for, and there are 79.5% of the cases that have a value of X of less than or equal to 45. As a result, the 75th percentile is in the interval 40-45, somewhere between an X of 40 and 45. Since 75% is closer to 79.5% than to 64.3%, one might roughly guess that P_{75} is approximately 43 or 44.

Using straight line interpolation in the interval 40-45 means that it is necessary to go from 64.3% to 75.0%, or $75 - 64.3 = 10.7\%$, out of a total

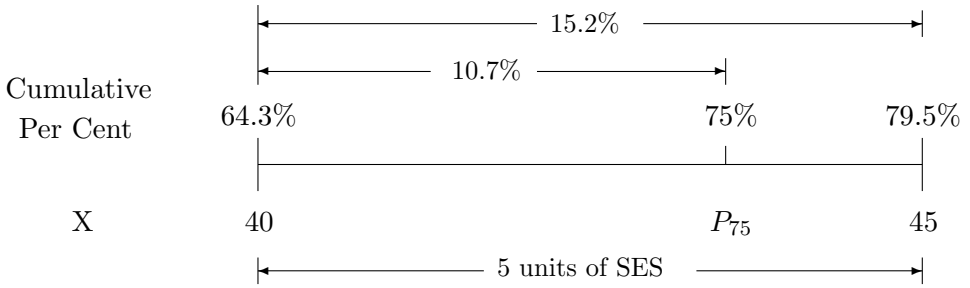

Figure 5.1: 75th Percentile of Socioeconomic Status

distance of 15.2%. As a result, the 75th percentile lies $10.7/15.2 = 0.704$ *of the way between 40 and 45. The value of* P_{75} *is thus*

$$P_{75} = 40 + \left(\frac{75 - 64.3}{15.2}\right) \times 5 = 40 + (0.704 \times 5) = 40 + 3.52 = 43.52$$

Thus $P_{75} = 43.52$*. This is illustrated diagramatically in Figure 5.1.*

The 75th percentile of socioeconomic status might best be reported as a socioeconomic status level of 43.5, or 44. Linear interpolation assumes that the cases in the interval across which the interpolation takes place, are uniformly distributed. This may not be the case, so the 75th percentile as calculated here may not be accurate to more than the nearest integer.

The 13th percentile must be in the interval between 20 and 30, because only 0.6% of the respondents have been accounted for at a socioeconomic status level of 20, but 15.8%, more than 13%, have been accounted for by the time a status level of 30 has been reached. Based on linear interpolation in the 20-30 interval, the 13th percentile is

$$P_{13} = 20 + \left(\frac{13 - 0.6}{15.2}\right) \times 10 = 20 + (0.816 \times 10) = 28.16$$

This could be rounded off so that the 13th percentile, P_{13}*, occurs at a status level of* $X = 28.2$ *or 28. Figure 5.2 gives the diagrammatic representation of this calculation.*

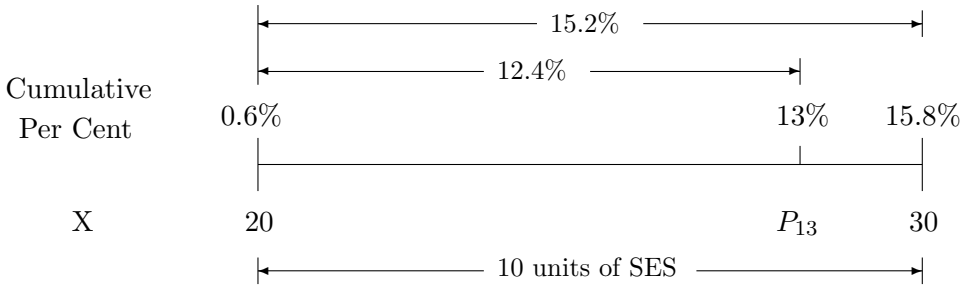

Figure 5.2: 13th Percentile of Socioeconomic Status

5.7 Measures of Variation

The measures of central tendency discussed earlier in this Chapter provide various ways of identifying the centre of a distribution. In addition to the centre, another essential characteristic of distributions is the amount of variation or variability in a distribution. Some distributions have most of their values concentrated near the centre of the distribution so that they have a low degree of variability. Other distributions have values of the variable spread out across a much greater set of values, so that they are more varied. This section presents several measures which can be used to summarize the variation of a distribution.

Example 5.7.1 Grades for Two Students

As an initial example illustrating differences in variation, suppose that two distributions have the same centre but have quite different amounts of variability. Let the grade in per cent for two students in a particular semester be as follows.

Student A	66	69	71	76	82
Student B	60	63	71	79	91

Each student has the same mean grade of approximately 73, and the same median of 71, but the grades of Student B can be seen to be more varied than the grades of Student A. The grades for Student B extend over more values, and each of Student B's grades is farther from the centre than

the corresponding grade for A. For example the second highest grade for A is 76%, about 3 points above the mean, while for Student B the second highest grade is 79%, about 6 points above the mean.

The various measures of variation or variability discussed in the following sections will provide summary measures of the amount of variability in a data set. Distributions having values of the variable which differ considerably, such as the grades for Student B in the above example, will have large values for measures of variation. Distributions with less variability, such as the set of grades of Student A, will have smaller values for the measures of variation.

Measures of variation are likely to be less familiar than measures of the central tendency or the average. While the media and ordinary language use the notion of *average* very commonly, the idea of variation or variability is much less commonly used.

Even though variation is less widely understood, and less intuitive, than is centrality, it is an extremely important concept the social sciences. People differ in their characteristics and behaviour, and a large part of the social sciences is devoted to attempting to understand and explain this variation. In addition, social scientific explanations are developed on the basis of an examination of variability among people, attempting to understand why people differ from each other, both in their innate characteristics, and in the manner in which they develop. Statistics by itself cannot provide explanations of social phenomena, but it can be used to describe variation. Understanding how variation can be described is essential to understanding explanations of differences among people.

The measures of variation discussed in this chapter deal with variables measured on scales which are ordinal or higher level scales. While some measures of variation for scales which are no more than nominal do exist, they are not so commonly used. For variables measured at no more than nominal scale, it is usually advisable to give all the values, either as a list, or as a frequency or percentage distribution.

The measures of variation which will be discussed in the following sections are the range, the interquartile range, the variance and the standard deviation. The methods of calculating these, some of the advantages and disadvantages of each, along with an idea of how these might be interpreted and used, is provided. The last section on variation examines measures of relative variation, discussing how values may differ relative to each other.

5.8 Variation - Positional Measures

Positional measures of variation take two values of the variable and report how far apart these values are. Variables which have greater variation will have values which are farther apart, and variables which have less variation will be closer together. The various positional measures of variation are based on different considerations concerning which position should be considered. The two most common positional measures are the range and the interquartile range.

5.8.1 The Range

The **range** of a variable is likely to be the only measure of variation which is used by people who are not familiar with Statistics. In ordinary language we sometimes use range to describe limits. For example, we may say that a two year old child has a much more limited range of expression than does a five year old child. When examining data, the notion of range is basically the same as this, focussing on the outer limits of the values of the variable.

Definition 5.8.1 The **range** of a set of values of a variable is the largest value of the variable minus the smallest value of the variable.

For Students A and B, in Example 5.7.1, the ranges are easily determined since the data set is small and the values of the variable are in order. For Student A, the smallest value is 66 and the largest value is 82, so the range is $82 - 66 = 16$. For Student B, the minimum grade is 60 and the maximum grade is 91, so that the range of grades is $91 - 60 = 31$. Based on the range, the grades for Student B are more varied than the grades for Student A.

Alternatively the range may be reported as **the smallest and the largest value**. For Student A, the range could be reported as 66 to 82, and for Student B the range is 60 to 91. This manner of reporting the range gives a little more information, in that the **minimum** and **maximum** values are both reported, although it leaves subtracting these values to those who are examining the data. Either method of reporting the range is acceptable, although the difference between the maximum and minimum values as given in Definition 5.8.1 is more common.

For a variable where the distribution has been grouped into categories or intervals, the range can usually be read from the table. For example, the range of the number of people per household in Table 5.1 is from 1 to 8, or a range of 7.

In Table 5.2, the range of attitudes is from Strongly Disagree to Strongly Agree, a range from 1 to 7, or a range of 6 points on an attitude scale. In the case of an attitude variable such as this, it may make more sense to report the range as the minimum and maximum value, that is, *Strongly Disagree* and *Strongly Agree*. This means more to anyone examining the data, than does a report that the range is 6 points on the attitude scale.

The range of socioeconomic statuses reported in Table 5.3 is 80, from a minimum of 0 to a maximum of 80.

For distributions with open ended intervals, the range cannot accurately be reported. For example, the distribution of hours of work per week for Canadian youth, given in Table **??** has intervals 'less than 10', '11-30', '31-40', '41-50' and '50+'. While the minimum value of hours worked per week has to be 0, the maximum value of hours worked per week for youth cannot be determined from this table. About all that can be done is to report the range as from 0 to '50 and over', although this is not of all that much use.

The range is a useful first measure to examine when encountering a new data set. The range gives a very quick and very rough idea of the set of values. For example, suppose the range of acreages of farms in a particular region of Canada, such as a crop district on the Prairies, is 5300 acres, while in another region, say a county in Prince Edward Island, is 325 acres. These two ranges tell a lot concerning the two areas. They show that farms are a lot larger, and a lot more varied in terms of size, on the Prairies as compared with Prince Edward Island.

Sometimes the range is also useful in indicating whether or not to examine a particular issue, or how to examine that issue. For example, if the price of a product has a range of 0 from store to store, then what has to be explained is the fact that prices do not vary from store to store. If the price has a range of $55 among different stores, then what has to be explained is why the price is so much lower at some locations than in others. As well, in the former case, it is not worthwhile to shop around from store to store to find the cheapest price, while it may be worthwhile in the latter circumstance.

Even though the range is a useful first indicator of the degree of variation of a variable, it is quite limited in terms of the information which it can provide. As a positional measure, the range focusses on only two values, the very smallest value and the very largest value. No other values are considered, and the variation in the remainder of the values is not taken into account. The following measures of variation correct for this weakness in various ways.

In spite of these difficulties, the range is useful, and is often reported. It gives an idea of the set of values to be examined, and provides a quick and rough idea of variation.

5.8.2 The Interquartile Range

Another positional measure, one which corrects for the weakness of the range just mentioned, is the interquartile range. This is defined as follows.

Definition 5.8.2 The **interquartile range** (IQR) of a variable is the seventy fifth percentile minus the twenty fifth percentile. That is

$$IQR = P_{75} - P_{25}$$

The interquartile range is a positional measure in that it takes two values of the variable, P_{75} and P_{25}, and reports the difference between these two values of the variable.

The advantage of the interquartile range over the range is that the IQR examines the range of the middle portion of the distribution. Recall that the 75th percentile of a distribution is the value of the variable such that 75% of the cases lie below this. The 25th percentile is the value of the variable such that only 25% of the cases are below this. The interquartile range, by eliminating the lowest 25 per cent of cases, and the upper 25 per cent of cases, the IQR describes the set of values over which the *middle* half of cases are spread. This gives those analyzing the data a good idea of how varied the cases are for the middle 50 per cent of cases.

The following examples show how to determine the IQR. The first example is that of a discrete, integer valued ordinal scale. The second example is that of a continuous variable on a ratio scale, where interpolation is required in order to obtain the appropriate values of the percentiles.

Example 5.8.1 Explanations of Unemployment

A 1985 survey of Edmonton adults asked these respondents their views concerning various explanations of unemployment. The results of this study are published in H. Krahn, G. S. Lowe, T. T. Hartnagel and J. Tanner, " Explanations of Unemployment in Canada," **International Journal of Comparative Sociology**, *XXVIII, 3-4 (1987), pp, 228-236. The responses*

		Variable 1 Recession and Inflation			Variable 2 Unemployment Insurance		
Attitude	X	f	P	Cum P	f	P	Cum P
Strongly Disagree	1	8	1.9	1.9	54	13.1	13.1
	2	17	4.1	6.0	43	10.4	23.5
	3	17	4.1	10.1	45	10.9	34.5
	4	37	8.9	19.1	43	10.4	44.9
	5	93	22.5	41.5	72	17.5	62.4
	6	133	32.1	73.7	73	17.7	80.1
Strongly Agree	7	109	26.3	100.0	82	19.9	100.0
Total		414	100.0		412	100.0	

Table 5.4: Responses to Explanations of Unemployment

to two of the questions asked in this survey are contained in Table 5.4. Variable 1 refers to responses to the explanation "World wide recession and inflation cause high unemployment," and Variable 2 gives responses to the explanation 'Unemployment is high because unemployment insurance and welfare are too easy to get."

For each variable, respondents were asked how much they agreed or disagreed with each explanation, with responses being given on a 7 point scale, where 1 represents strongly agree and 7 represents strongly disagree. The responses are thus given on a discrete, ordinal level scale. The sample sizes differ slightly for the two variables, because some respondents did not answer one of the two questions.

The distributions are presented in Table 5.4 first as frequency distributions, where f represents the frequency of responses to each question. These are then presented as percentages, P, and cumulative percentage distributions, 'Cum P'.

Note that the range of responses is the same for both explanations. For each variable, the range is 6 points on the attitude scale, from 1 to 7. If the actual distributions are examined though, the differences in variability of responses for these two variables can be seen. For Variable 1, recession and inflation as the cause of unemployment, there is considerable similarity in

response. Very few respondents disagree very strongly with this explanation, and the bulk of responses is in the categories 5, 6 or 7. For Variable 2, responses are much less concentrated on a few values of X. No one value of attitude stands out as having all that many more responses than do other values, although the modal response is strongly agree, value 7. But there are also a considerable number of respondents who strongly disagree that UIC and welfare are too easy to get, and responses are varied across all values of X.

The interquartile range is based on the 25th and 75th percentile. Since this distribution is a ordinal scale with discrete, integer values, the appropriate percentiles occur at these integer values. No interpolation between categories is required. The interquartile range provides a summary measure which shows the variation in these two distributions. For Variable 1, the 75th percentile occurs at attitude value 7, and the 25th percentile at attitude value 5. That is, the cumulative percentage column does not reach 75% until $X = 7$, and it is not until $X = 5$ that the 25% of respondents with the lowest values on the attitude scale are accounted for. Thus,

$$IQR = P_{75} - P_{25} = 7 - 5 = 2$$

For Variable 2, the 75th percentile is at attitude 6 and the 25th percentile at attitude 3. Thus

$$IQR = P_{75} - P_{25} = 6 - 3 = 3$$

As a summary measure then, the IQR shows that Variable 2 has greater variation than does Variable 1. While a difference of one point in the values of the IQR may not seem like much, there are only 7 points on this scale, and an IQR of 3 is actually one and a half times greater than an IQR of 2.

Example 5.8.2 Age Distributions of Inuit and Total Population of Canada

Table 5.5 give age distributions of the Inuit and total population of Canada for 1986. This table is based on data in Statistics Canada, **Canadian Social Trends**, *Winter 1989, page 9. The table can be used to determine the interquartile range for the ages of the Inuit population and of the total population of Canada as follows.*

The variable 'age' in this example has a continuous, ratio level scale. The percentage distributions have already been provided, and the cumulative

Age	Per Cent of: Inuit	Total
Under 15	40	22
15-24	23	18
25-39	20	24
40-54	10	17
55-64	4	9
65 plus	3	10
Total	100%	100%

Table 5.5: Age Distribution of Inuit and Total Population, Canada, 1986

percentages are calculated from these. The values of age are grouped into intervals, and thus linear interpolation must be used to determine the 75th and 25th percentiles. There is also a gap between the endpoints of the intervals, so that the real class limits must be constructed and used. The cumulative percentages and the real class limits are given in Table 5.6. For the Inuit population, the 25th percentile occurs in the first interval, since the youngest 40% of the Inuit population is in that interval. By linear interpolation, P_{25} is

$$P_{25} = -0.5 + \left(\frac{25-0}{40}\right) \times 15$$

$$P_{25} = -0.5 + \left(\frac{25}{40}\right) \times 15 = -0.5 + 9.4 = 8.9$$

The 75th percentile is in the interval 25-39, since there are 63% of the Inuit population of age less than 25, and 83% of age less than 40. Thus P_{75} is

$$P_{75} = 24.5 + \left(\frac{75-63}{20}\right) \times 15$$

$$P_{75} = 24.5 + \left(\frac{12}{20}\right) \times 15 = 24.5 + 9 = 34.5$$

The interquartile range for the Inuit population is

$$IQR = P_{75} - P_{25} = 34.5 - 8.9 = 25.6$$

Age	Real Class Limits	Inuit Population P	Inuit Population Cum P	Total Population P	Total Population Cum P
Under 15	-0.5-14.5	40	40	22	22
15-24	14.5-24.5	23	63	18	40
25-39	24.5-39.5	20	83	24	64
40-54	39.5-54.5	10	93	17	81
55-64	54.5-64.5	4	97	9	90
65 plus	64.5 plus	3	100	10	10
Total		100		100	

Table 5.6: Cumulative Per Cent Distributions, Inuit and Total Population, Canada, 1986

Thus the interquartile range for the Inuit population of Canada in 1986 was 26 years.

For the total population, the method is the same. The 25th percentile occurs in the second interval, 15-24, where the cumulative percentages cross the 25 per cent point. By linear interpolation, P_{25} *is*

$$P_{25} = 14.5 + \left(\frac{25 - 22}{18}\right) \times 10$$

$$P_{25} = 14.5 + \left(\frac{3}{18}\right) \times 10 = 14.5 + 1.7 = 16.2$$

The 75th percentile is in the interval 40-54, since there are 64% of the total population of age less than 40, and over 75% of the total population by the time an age of 54 is reached. Thus P_{75} *is*

$$P_{75} = 39.5 + \left(\frac{75 - 64}{17}\right) \times 15$$

$$P_{75} = 39.5 + \left(\frac{11}{17}\right) \times 15 = 39.5 + 9.7 = 49.2$$

The interquartile range for the total population is

$$IQR = P_{75} - P_{25} = 49.2 - 16.2 = 33.0$$

Based on these interquartile ranges, the distribution of ages of the total population of Canada is more varied than the distribution of the Inuit population of Canada. The interquartile range for the total population is 33 years, and for the Inuit population is 26 years, about 7 years less. This means that the middle half of the Inuit population, in terms of age, is spread across only 26 years of age. In contrast, the middle half of the total population of Canada is between ages 16 and 49, a difference of 33 years.

If the original distributions are examined, it can be seen that the Inuit distribution is more concentrated, and the total population more varied. For the Inuit population, there is a very large concentration of population at the youngest ages, ages less than 15. This single category has 40 per cent of all Inuit in it. For the total population, there are considerable percentages of the population in each age group, with no one interval being an interval in which people are concentrated.

5.8.3 Other Positional Measures of Variation

In addition to the interquartile range, other positional measures of variation could easily be constructed. For example, if a researcher wished to eliminate only the bottom 5 per cent and the top 5 per cent of values of a distribution, then a measure of variation based on the middle 90 percent of the distribution could be constructed. This measure would be the 95th percentile minus the 5th percentile, that is $P_{95} - P_5$. This might be useful for comparing distributions which are very skewed at the ends of the distribution. An income distribution, for example, may have a few incomes of several hundred thousand dollars at the upper end of the income scale. At the lower end, there may be negative incomes among those small business people or farmers whose expenditures exceed receipts in a given year. For purposes of analying the variability of a distribution, a researcher may wish to eliminate both of these extremes so that the variation of incomes for the bulk of the population having more ordinary incomes can be examined.

One example where such a measure has been constructed and used is the Saskatchewan Department of Labour's annual publication **Wages and Working Conditions**. In that publication, the 80% range is used to describe the distribution of wages and salaries. (See Labour Relations Branch, Saskatchewan Human Resources, Labour and Employment, **Wages and Working Conditions by Occupation: Fifteenth Report 1990**, Regina, 1991.)

This survey provides "a summary of results obtained in a survey of busi-

ness establishments operating in the province. The reference month for the survey is October 1990." Data was collected from 964 establishments representing 81,197 employees in 334 occupations. In addition to the **range** of employees' wages, this publication also reports the **80% range**. This is defined on page 220 of the publication as

> The 80% Range gives the lowest and highest wage reported after disregarding 10% of the total number of employees at the lowest wage level and at the highest wage level, i.e., 80% Range represents the middle four-fifths of the employees.

A few examples of the survey results are given in Table 5.7.

	No. in	100% Range		80% Range			
Occupation	Sample	Low	High	Low	High	Median	Mean
Sales Clerk	247	5.00	12.95	5.25	10.05	6.95	7.28
Assembler	128	6.00	15.71	8.75	15.71	10.35	11.81
Bus Driver	225	7.00	14.57	13.28	14.57	14.57	13.81
Engineer	154	2600	6075	3000	4844	4000	3917
Nurse	1715	2106	3795	2605	3385	3080	3039

Table 5.7: Summary Measures of Wages of Salaries for Various Occupations, Saskatchewan, 1990

The above measures are symmetrical, in that they are based on percentiles which fall an equal distance from the 50 per cent point. Measures of variation could be asymmetrical as well. For example, a measure of variation could be constructed to measure the difference between the 80th and the 5th percentile. This would be a measure which cut off the top 20 per cent, and the lowest 5 per cent of a distribution.

Each of the positional measures of variation provide a very useful view of the variability of a distribution. However, these measures are not usually used except for descriptive purposes. For purposes of statistical inference, and for more statistical analysis, these measures are difficult to manipulate mathematically. As a result, the standard deviation and the variance are the measures of variation more commonly used by statisticians. These measures are discussed in the following section.

5.9 Standard Deviation and Variance

The most commonly used measures of variation are the standard deviation and the variance. Neither is likely to be familiar to those who have not studied Statistics, and the concepts on which each is based are not as intuitive as are the measures of central tendency and variation discussed so far. It is important to be able to understand the standard deviation and variance, since statistical work depends very heavily on these. This section first shows how to calculate these measures in various circumstances, and later discusses various ways of interpreting these measures.

The standard deviation and variance each require an interval or ratio scale of measurement. If a variable has only an ordinal level of measurement, then sometimes this ordinal scale is treated as if it has an interval level scale. The cautions that were mentioned in Example **??** in Section **??**, where the mean of an ordinal scale was calculated, also apply here. The method of determining the standard deviation and the variance in the case of ungrouped data is discussed first. Then the various formulae for calculating these measures using grouped data are presented.

5.9.1 Ungrouped Data

If a list of values of a variable X is given, the standard deviation is calculated by first determining the mean, $\bar{X}$, and then examining how far each value of the variable differs from the mean. These differences of the values from the mean, $(X - \bar{X})$, are often termed **deviations about the mean**. While the manner in which these deviations about the mean are manipulated algebraically depends on the exact formulae for the standard deviation and variance, it is these deviations which form the basis for the standard deviation and the variance. Where the values of the variable are closely concentrated around the mean, these deviations about the mean will be small and the standard deviation will be small. Where the values of the variable are very dispersed, the values of the variable will differ rather considerably from the mean. The deviations about the mean will be much larger and the standard deviation will be larger as well. This can be seen in the example of the grades of Students A and B, first presented in Example 5.7.1.

Example 5.9.1 Deviations About the Mean Grade

As noted earlier, for each of Students A and B, the mean grade is 72.8. Table 5.8 gives the values of the grades, along with the deviation of each

grade about the mean grade.

	Student A		Student B	
	X	$X - \bar{X}$	X	$X - \bar{X}$
	66	-6.8	60	-12.8
	69	-3.8	63	-9.8
	71	-1.8	71	-1.8
	76	3.2	79	6.2
	82	9.2	91	18.2
Total	364	0.0	364	0.0

Table 5.8: Deviations about the Mean for Students A and B

Examining the deviations about the mean in Table 5.8, shows that the deviations about the mean, $X - \bar{X}$, are somewhat smaller values for Student A than for Student B. This is because the grades for Student A are less spread out, and are closer to the mean, than in the case of Student B. Based on these deviations about the mean, the distribution of the grades for Student A can be considered to be less varied than the distribution of grades for Student B.

In Example 5.9.1, examining the list of all deviations about the mean is an awkward procedure. A measure of variation should combine all these deviations about the mean into a single number. The difficulty in doing this can be seen in this same example. Note that the sum of the deviations about the mean, for each of Students A and B in Table 5.8, is 0. The mean is in the centre of the distribution, in the sense that these deviations about the mean total zero. That is, the sum of the distances of the values of $X - \bar{X}$ that lie below the mean value equals the sum of the distances of the values of the values of $X - \bar{X}$ that lie above the mean. The proof for this is given a little later in this Chapter. Since this characteristic of the mean implies that the sum of the deviations about the mean is always 0, this sum cannot be used as a measure of variation.

The technique that statisticians use to deal with this difficulty is to square the deviations about the mean, so that the values of these squares of deviations about about the mean are all positive. Squaring any value

means multiplying the value by itself. Negative deviations about the mean, when squared, become positive, and the square of positive deviations is also positive. The measures of variation called the variance and standard deviation are based on these squares of the deviations about the mean in the following manner.

Definition 5.9.1 If a variable X has n values

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

and the mean of these n values is $\bar{X}$, then the **variance** of this set of values is

$$s^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}$$

The standard deviation of this set of n values of X is

$$s = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}}$$

That is, the deviations about the mean are calculated as in Table 5.8. For each X_i, where $i = 1, 2, \cdots, n$, these deviations about the mean are the values $(X_i - \bar{X})$. Then each of these values is squared, that is, multiplied by itself, producing the n values $(X_i - \bar{X})^2$. All n of these squares of the deviations about the mean are added, and this sum is divided by $n-1$. This produces a measure of variation which is termed the **variance**.

While the variance is used extensively in Statistics, most of the formulae in the next few chapters rely more heavily on the standard deviation. The **standard deviation is the square root of the variance**. That is, once the variance has been calculated, the standard deviation is the number, which when multiplied by itself, results in the value of the variance.

While each of the variance and the standard deviation may be a little difficult to understand at first sight, the standard deviation turns out to be somewhat easier to interpret than is the variance. For this reason, and since the formulae in later chapters are based on the standard deviation, it is this latter measure which will become the main measure of variation used in this textbook. However, the variance is calculated for each example, but mainly as a step in obtaining the value of the standard deviation. That is, the variance is calculated first, and then the square root of the variance is calculated, in order to determine the standard deviation. The grades of Students A and B are used as an example of how to calculate these measures.

Example 5.9.2 Standard Deviation of Grades of Students A and B

The deviations about the mean, and the square of the deviations about the mean are given in Table 5.9. The calculations for the variance and the standard deviation follow this. For Student A,

	Student A			Student B		
	X	$X-\bar{X}$	$(X-\bar{X})^2$	X	$X-\bar{X}$	$(X-\bar{X})^2$
	66	-6.8	46.24	60	-12.8	163.84
	69	-3.8	14.44	63	-9.8	96.04
	71	-1.8	3.24	71	-1.8	3.24
	76	3.2	10.24	79	6.2	38.44
	82	9.2	84.64	91	18.2	331.24
Total	364	0.0	158.80	364	0.0	632.80

Table 5.9: Calculations for Standard Deviation for Students A and B

$$s^2 = \frac{(X_1-\bar{X})^2 + (X_2-\bar{X})^2 + \cdots + (X_n-\bar{X})^2}{n-1}$$

$$s^2 = \frac{158.80}{5-1} = 39.7$$

The variance of these 5 grades is 39.7, and the standard deviation is the square root of this. That is,

$$s = \sqrt{39.7} = 6.301$$

The standard deviation of grades for Student A is 6.3. For Student B, the method is the same, giving

$$s^2 = \frac{632.80}{5-1} = 158.2$$

$$s = \sqrt{158.2} = 12.578$$

The standard deviation of grades for Student B is 12.6. Based on the respective sizes of the two standard deviations, the grades for Student B are approximately twice as spread out as the grades for Student A.

The standard deviation of grades for B being approximately double the standard deviation of grades for A should make some sense. The mean grade for each student is the same, 72.8%. For Student B, each grade is approximately twice as far away from the mean as is the corresponding grade for Student A. For example, for the the lowest grade of 66 for A, this is 6.8 percentage points below the mean for A. For B, the lowest grade is 60, and this is $60 - 72.8 = -12.8$ *percentage points from the mean. For the lowest grade, the deviation about the mean is twice as great for B as for A. A similar statement could be made for all the other grades, so that each deviation about the mean is twice as great for B as for A. All these deviations about the mean are put together into a summary measure, the standard deviation. It thus makes sense that the standard deviation of grades for B is twice as great as the standard deviation of grades for A. While the value of each standard deviation by itself may seem a bit mysterious, the relative sizes of the standard deviations show that the grades for B are twice as varied for B as for A.*

Units for the Standard Deviation. The standard deviation of a variable X has the same units as the units which were used to measure X. In Example 5.9.2, the standard deviation of grades for each student is in units of percentage points. That is, the respective standard deviations are 6.3% and 12.6%. Even though the formula for the standard deviation seems confusing, the standard deviation as a measure of variation is in units which are familiar. As will be noted later in this chapter, this is useful in helping to interpret the meaning of the standard deviation.

The reason the units for s are the same as for X can be seen by examining the formula. The deviations about the mean $(X - \bar{X})$, are in the units of X, since both X and $\bar{X}$ are. These values are squared, producing units which are the squares of the units of X. Then these squares are summed, producing a sum which is measured in the squares of the units of X. In Example 5.9.2, the sum of these squares is in units of 'per cent squared'. Dividing this sum by $n - 1$ does not change these units. This is a difficult unit to understand, and the variance is measured in the square of units of X. But when the square root of the variance is taken, this once again produces a measure in the units in which X was originally measured. Part of the reason the standard deviation is preferred over the variance when working with data is that the standard deviation at least has familiar units. The variance is a useful measure, but being in such strange units, is a little more

difficult to understand.

Summation Notation for Variance and Standard Deviation. Since Definition 5.9.1 works with the sum of a set of values, the definitions of variance and standard deviation can be compactly given with the use of the summation sign. Definition 5.9.1 can be restated as follows:

Definition 5.9.2 If a variable X has n values

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

and the mean of these n values is

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

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

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

The standard deviation of this set of n values of X is

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

The sum of the squares of the deviations about the mean

$$(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2$$

can be expressed compactly as

$$\sum_{i=1}^{n} (X_i - \bar{X})^2$$

or dropping the subscripts and superscripts on the summation sign,

$$\sum (X_i - \bar{X})^2 = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2$$

As with any summation, the first step is to carry out the operations in brackets, and then the algebraic operations to the right of the summation sign. Then these values are added. In this case, the first step is to calculate

the mean $\bar{X}$ of the variable X, and then calculate all the deviations about the mean $(X_i - \bar{X})$. Each of these deviations about the mean are squared, producing the values $(X_i - \bar{X})^2$. Then these values are added, producing the total

$$\sum (X_i - \bar{X})^2$$

which is the value of the numerator in Definitions 5.9.1 and 5.9.2. In order to determine the variance, this sum is divided by $n-1$, producing the variance

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

The standard deviation is then determined by taking the square root of this value, so that

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

An Alternative Formula for the Variance and Standard Deviation. The formulae of Definition 5.9.2 can be reorganized to produce a computationally more efficient formula for the variance and the standard deviation. This is as follows:

Definition 5.9.3 If a variable X has n values

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

and the mean of these n values is

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

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

$$s^2 = \frac{1}{n-1}\left[\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}\right]$$

The standard deviation of this set of n values of X is

$$s = \sqrt{\frac{1}{n-1}\left[\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}\right]}$$

The formulae of Definition 5.9.3 are more efficient than those of Definition 5.9.2 because the latter do not require the computation of the deviations about the mean. Rather, they require only the calculation of the sum of the n values of X, ΣX, and the sum of the squares of the values of X, ΣX^2. These are then entered into the formulae of Definition 5.9.3. An example of the use of these formulae follows, and the proof of the equivalence of the formulae of this definition with those of the earlier definition is given later in the chapter.

Example 5.9.3 Variation in Support for Liberals and NDP in Canada

Each month Gallup Canada surveys approximately 1000 Canadian adults to determine their political preference. Some of the results of these surveys were given in Example **??**. *In Table 5.10 the percentage of decided Canadian adults who support each of the Liberals and NDP over the years 1990 and 1991 is given. The results are reported for each quarter, rather than for each month. This example uses this data set to determine various measures of central tendency and variation, including the variance and standard deviation. The latter are determined using the formulae of Definition 5.9.3.*

	Percentage of Decided Voters Favouring	
Date	Liberals	NDP
December 1991	38	23
September 1991	38	26
June 1991	35	23
March 1991	39	30
December 1990	32	36
September 1990	39	32
June 1990	50	23
March 1990	50	25

Table 5.10: Percentage of Decided Voters Favouring Liberal and NDP, Canada, March 1990-December 1991

In order to used these new formulae, it is necessary to calculate $\sum X$*, the sum of the values of the variable, and* $\sum X^2$*, the sum of the squares of the values of the variable. This is done in Table 5.11. In that table, the percentage support for the Liberals is given the symbol* X*, and the percentage support for the NDP is given the algebraic symbol* Y*, in order that the two can be distinguished. From Table 5.11,*

	Liberals		NDP	
Date	X	X^2	Y	Y^2
Dec. 1991	38	1,444	23	529
Sept. 1991	38	1,444	26	676
June 1991	35	1,225	23	529
March 1991	39	1,521	30	900
Dec. 1990	32	1,024	36	1,296
Sept. 1990	39	1,521	32	1,024
June 1990	50	2,500	23	529
March 1990	50	2,500	25	625
Total	321	13,179	218	6,108

Table 5.11: Calculations of Summations for Liberals and NDP, March 1990 - Dec. 1991

$$\Sigma X = 321$$

and $n = 8$*. Although the mean need not be determined in order to calculate the variance and standard deviation using this formula, the mean value of Liberal support over these months was*

$$\bar{X} = \frac{\Sigma X}{n} = \frac{321}{8} = 40.125$$

so that there was a **mean** *level of 40.1 per cent support for the Liberals over these months. The* **range** *of the Liberal support is* $50 - 32 = 18$ *per cent. For these months,* $\Sigma X = 321$*,* $\Sigma X^2 = 13,179$ *and* $n = 8$*, so that the variance is*

$$s^2 = \frac{1}{n-1}\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right] = \frac{1}{7}\left[13,179 - \frac{321^2}{8}\right]$$

$$s^2 = \frac{1}{7}\left[13,179 - \frac{103,041}{8}\right] = \frac{298.875}{7} = 42.6964$$

The **standard deviation** *is*

$$s = \sqrt{42.6964} = 6.5343$$

or 6.5 percentage points.

From Table 5.11,

$$\Sigma Y = 218$$

and

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{218}{8} = 27.25$$

so that there was a **mean** *level of 27.2 per cent support for the NDP over these months. The* **range** *of the NDP support was* $36 - 23 = 13$ *per cent. For these months,* $\Sigma Y = 218$, $\Sigma Y^2 = 6,108$ *and* $n = 8$, *and the variance is*

$$s^2 = \frac{1}{n-1}\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right] = \frac{1}{7}\left[6,108 - \frac{218^2}{8}\right]$$

$$s^2 = \frac{1}{7}\left[6,108 - \frac{47,524}{8}\right] = \frac{167.500}{7} = 23.92857$$

$$s = \sqrt{23.92857} = 4.8917$$

and thus the **standard deviation** *is 4.9 percentage points.*

Party	n	Median	Mean	Range	S.D.
Liberal	8	38.5	40.1	18	6.5
NDP	8	25.5	27.2	13	4.9

Table 5.12: Summary Measures for Liberals and NDP

These measures are summarized in Table 5.12. and this table provides a summary of the differences in the distribution of Liberal and NDP support. Over these months the average level of support for the Liberals was greater than the average level of support for the NDP, regardless of whether the median or mean is used. The variation in support for the Liberals was also somewhat greater than the variation in support for the NDP. Over these

months, Liberal support varied from a low of 32% of decided respondents, to a high of 50% of the decided respondents in the Gallup poll. The standard deviation was 6.5 percentage points, somewhat greater than the standard deviation of 4.9 percentage points for the NDP. It can also be seen that the range in support for the NDP was from a low of 23% to a high of 36%. In terms of reporting the results, rather than give the detailed list of values of Table 5.10, for most purposes the summary measures of support given in Table 5.12 would be sufficient.

Proofs of Formulae. Earlier in this section, some claims were made concerning the sum of the deviations about the mean, and the equivalence of two different formulae for the variance. This section provides proofs for these claims. If you are not adept at algebra, this section can be skipped, and you can accept the claims as presented. However, in order to develop a better understanding of the formulae, and obtain some practice in the manipulation of summation signs, it is worthwhile to follow these proofs. In some of the summation signs that follow, the subscripts and superscripts are dropped. However, all of the summations are across all n values of the variable X.

In Table 5.8 of Example 5.9.1, the sum of the deviations about the mean was zero. In the case of ungrouped data, it can easily be proved that this must always be the case. The deviations about the mean are $(X_i - \bar{X})$ and there are n of these, from $i = 1$ to $i = n$. This sum can be written,

$$\sum_{i=1}^{n}(X_i - \bar{X}) = (X_1 - \bar{X}) + (X_2 - \bar{X}) + \cdots + (X_n - \bar{X})$$

Grouping together the X_i, and then grouping together the $\bar{X}$ values, this can be written

$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X} = \sum_{i=1}^{n} X_i - n\bar{X}$$

Note that the latter entry $n\bar{X}$ occurs because it is a summation of $\bar{X}$, and $\bar{X}$ is the same for each of the n times it is added, so this sum is $n\bar{X}$. But by definition,

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

and substituting this for $\bar{X}$ in the last expression gives

$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} X_i - n\left[\frac{\sum X_i}{n}\right] = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i = 0$$

That is, for ungrouped data, the sum of the deviations about the mean must always equal zero.

When calculating these deviations about the mean with actual data, the sum of the deviations may not add up to exactly zero, becuase of rounding errors. So long as it adds to 0.2 or -0.1, or a number very close to zero, then this very likely means no errors beyond rounding errors. Where the sum of the deviations about the mean differs by much more than this from zero, then it is likely that some calculating errors have been made.

The equivalence of Definitions 5.9.2 and 5.9.3 is shown as follows. The claim is that

$$s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2 = \frac{1}{n-1}\left[\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right]$$

Since each of the summation part of these expressions are multiplied by the same value $1/(n-1)$, all that has to be shown is that

$$\sum(X_i - \bar{X})^2 = \left[\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}\right]$$

Expanding the square of a difference between two values gives

$$\sum(X_i - \bar{X})^2 = \sum(X_i^2 - 2X_i\bar{X} + \bar{X}^2)$$

Since each part of the expression on the right is either added to or subtracted from the other values, each of the parts of the expression in brackets can be considered to be summed across all n values so that

$$\sum(X_i - \bar{X})^2 = \sum X_i^2 - 2\bar{X}\sum X_i + \sum \bar{X}^2$$

For the last entry on the right, note that

$$\sum \bar{X}^2 = n(\bar{X}^2) = n\left[\frac{\sum X_i}{n}\right]^2$$

Thus

$$\begin{aligned}\sum(X_i-\bar{X})^2 &= \sum X_i^2 - 2\left[\frac{\sum X_i}{n}\right]\left[\sum X_i\right] + n\left[\frac{\sum X_i}{n}\right]^2 \\ &= \sum X_i^2 - 2\frac{\left(\sum X_i\right)^2}{n} + \frac{\left(\sum X_i\right)^2}{n} \\ &= \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}\end{aligned}$$

and this shows the equivalence of the two expressions. While the expressions in the two formulae look quite different, using the formulae in either Definitions 5.9.2 and 5.9.3 will produce the same value of the variance and the standard deviation.

5.9.2 Grouped Data

When a variable has already been grouped into categories, or into intervals, the basic principle for determining the variance and the standard deviation is the same as in the case of ungrouped data. That is, the squares of the deviations about the mean are obtained, and these are used to determine the measures of variation. Much as in the case of the mean for grouped data, values must be weighted by their respective frequencies of occurrence. For the variance and the standard deviation, the squares of the deviations about the mean are multiplied by the respective frequencies of occurrence. Then all of these values are summed. The definitions are as follows.

Definition 5.9.4 If a variable X has k values

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

occurring with respective frequencies

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

and the mean of these k values is

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

where

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

Then the **variance** of this set of values is

$$s^2 = \frac{f_1(X_1 - \bar{X})^2 + f_2(X_2 - \bar{X})^2 + \cdots + f_k(X_k - \bar{X})^2}{n-1}$$

The standard deviation of this set of n values of X_i is

$$s = \sqrt{\frac{f_1(X_1 - \bar{X})^2 + f_2(X_2 - \bar{X})^2 + \cdots + f_k(X_k - \bar{X})^2}{n-1}}$$

All of this can be expressed more compactly with summation notation as follows.

Definition 5.9.5 If a variable X has k values

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

occurring with respective frequencies

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

and the mean of these k values is

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

where

$$n = \sum f_i$$

and the summation is across all k values of X_i. Summing across the same k values, the **variance** is

$$s^2 = \frac{\sum f_i (X_i - \bar{X})^2}{n-1}$$

The standard deviation of this set of k values of X_i is

$$s = \sqrt{\frac{\sum f_i (X_i - \bar{X})^2}{n-1}}$$

The various steps involved in these formulae are as follows:

1. First calculate the mean of the variable X, using the formula for the mean of grouped data in Definition **??**.
2. Subtract each of the k values of X from this mean in order to determine the deviations about the mean.
3. Square each of the deviations about the mean.
4. Multiply each square of the deviation about the mean by its respective frequency of occurrence, f_i.
5. Sum all of the products in (4).
6. The variance is the sum of (5) divided by the sample size minus 1.
7. The standard deviation is the square root of the variance of (6).

These various steps are illustrated in the following example.

Example 5.9.4 Variation in the Number of Children per Family

Statistics Canada's Survey of Consumer Finances for 1987 gives the two distributions in Table 5.13. The data are for families surveyed in the province of Saskatchewan, and the table gives the number of children per family for families which are not in poverty and for families in poverty. Use these distributions to determine the mean, variance and standard deviation of the number of children per family. (Note that there are no families having 6 or 7 children, but one family with 8 children).

The calculations for the mean for the families not in poverty are given in the first three columns of Table 5.14. From this table, the sample size of families not in poverty is $n = \sum f = 3171$ *and* $\sum fX = 1645$ *so that the mean is*

$$\bar{X} = \frac{\sum fX}{n} = \frac{1645}{3171} = 0.52$$

The mean number of children for families not in poverty is 0.52, and in column four of Table 5.14, the deviations of the mean values of X *about this mean are given. These values are squared in column five and then these squares of the deviations about the mean are multiplied by the respective frequencies of occurrence in column six. From this table,* $\sum f(X - \bar{X})^2 =$

Number of Children Per Family	Number of Families Not in Poverty	In Poverty
0	2291	628
1	326	90
2	380	68
3	146	42
4	22	13
5	5	1
8	1	0
Total	3171	842

Table 5.13: Number of Children per Family, Poor and Non-Poor

2847.6384 *so that*

$$\begin{aligned} s^2 &= \frac{\sum f(X-\bar{X})^2}{n-1} \\ &= \frac{2847.6384}{3170} \\ &= 0.8983 \end{aligned}$$

and

$$s = \sqrt{s^2} = \sqrt{0.8983} = 0.9478$$

The standard deviation for families not in poverty is 0.95 children per household, and the variance is 0.90. For families in poverty, the same set of calculations is given in Table 5.15. For those families in poverty, the mean is

$$\bar{X} = \frac{\sum fX}{n} = \frac{409}{842} = 0.49$$

This mean is subtracted from each value of X, and the squares of these deviations are multipled by the respective frequencies. From this table, $n-1 = 842-1 = 841$ and $\sum f(X-\bar{X})^2 = 774.3442$ so that

$$\begin{aligned} s^2 &= \frac{\sum f(X-\bar{X})^2}{n-1} \\ &= \frac{774.3442}{841} \end{aligned}$$

X	f	fX	$(X - \bar{X})$	$(X - \bar{X})^2$	$f(X - \bar{X})^2$
0	2291	0	-0.52	0.2704	619.4864
1	326	326	0.48	0.2304	75.1104
2	380	760	1.48	2.1904	832.3520
3	146	438	2.48	6.1504	897.9584
4	22	88	3.48	12.1104	266.4288
5	5	25	4.48	20.0704	100.3520
8	1	8	7.48	55.9504	55.9504
Total	3171	1645			2847.6384

Table 5.14: Children per Family, Not in Poverty

X	f	fX	$(X - \bar{X})$	$(X - \bar{X})^2$	$f(X - \bar{X})^2$
0	628	0	-0.49	0.2401	150.7828
1	90	90	0.51	0.2601	23.4090
2	68	136	1.51	2.2801	155.0468
3	42	126	2.51	6.3001	264.6042
4	13	52	3.51	12.3201	160.1613
5	1	5	4.51	20.3401	20.3401
8	0	0	7.51	56.4001	0.0000
Total	842	409			774.3442

Table 5.15: Children per Family, Not in Poverty

Measure	Not in Poverty	In Poverty
$\bar{X}$	0.52	0.49
Median	0	0
s^2	0.90	0.92
s	0.95	0.96
n	3171	842

Table 5.16: Summary Measures, Number of Children per Family

$$= \quad 0.9207$$

and

$$s = \sqrt{s^2} = \sqrt{0.9207} = 0.9596$$

The standard deviation for families in poverty is 0.95 children per household, and the variance is 0.92.

Based on these two sets of calculations, Table 5.16 gives summary measures of central tendency and variation for these two distributions. The median is at 0 in each case because over one half of the families have 0 children. What is notable about these two distributions is the similarity in all the measures of central tendency and variation. The median is identical for the two distributions, and to one decimal place, the mean number of children per family is 0.5 in each distribution. The variance and standard deviation for each distribution are also so close to being the same that the two distributions can be considered to have the same variation. Based on these summary measures, the distributions for the number of children per family for families in poverty and for families not in poverty can be considered to be practically identical.

Sum of Deviations about the Mean. With grouped data, note that the sum of the deviations about the mean is not zero. In both Tables 5.14 and 5.15, the sum of the entries in the fourth column is not zero. This is because each of these deviations about the mean occurs a different number of times. If the deviations about the mean are multiplied by their respective frequencies of occurrence, then this sum will be zero. That is, for grouped data,

$$\sum f_i(X_i - \bar{X}) = 0$$

While the calculations for this are not given in Tables 5.14 or 5.15, it is a relatively straightforward procedure to verify this result for either table.

An Alternative Formula for the Variance and Standard Deviation. As in the case of ungrouped data, there are computationally more efficient formulae for the variance and standard deviation. These are presented in the following definition.

Definition 5.9.6 If a variable X has k values

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

occurring with respective frequencies

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

and the mean of these k values is

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

where

$$n = \sum f_i$$

and both of these summations are across all k values of X_i. Summing across the same k values, the **variance** is

$$s^2 = \frac{1}{n-1}\left[\sum f_i(X_i^2) - \frac{(\sum f_i X_i)^2}{n}\right]$$

The standard deviation is the square root of s^2, that is,

$$s = \sqrt{\frac{1}{n-1}\left[\sum f_i(X_i^2) - \frac{(\sum f_i X_i)^2}{n}\right]}$$

While these formulae may look more complex than the earlier formulae, the latter formulae can save considerable time when calculating the variance or the standard deviation. The steps that must be taken in calculating these are as follows:

1. First compute the sample size n by summing the frequencies of occurrence f_i.
2. Multiply each frequency, f_i times its corresponding X value, X_i. This produces the values $(f_i X_i)$.
3. Add the products in (2) to obtain the total $\sum f_i X_i$.
4. Multiply the individual values $(f_i X_i)$ by X_i again to produce the products $f_i X_i^2$.
5. Sum the products in (4) to produce the total $\sum f_i X_i^2$.
6. Square the summation in (3) to obtain $(\sum f_i X_i)^2$.
7. Divide the square of the summation in (6) by n to obtain the value
$$\frac{(\sum f_i X_i)^2}{n}$$
8. Subtract the result in (7) from the result in (5). This gives the value of the expression in the large square brackets in Definition 5.9.6.
9. For the variance, s^2, divide the result of (8) by $n - 1$.
10. For the standard deviation, s, compute the square root of the variance in (9).

While there are more steps to this calculation at the final stages, the tables that are needed to obtain these totals are simpler than Tables 5.14 and 5.15. Those tables required 6 columns, with both the deviations about the mean and the squares of these deviations about the mean being required. As can be noted in the following example, the formulae of Definition 5.9.6 require only 4 columns.

Additional Notes Concerning the Formulae in Definition 5.9.6.

1. Note that while the two expressions

$$\sum f_i X_i^2 \text{ and } (\sum f_i X_i)^2$$

may look quite similar, they are different. Make sure you are clear concerning how each is calculated. The former, $\sum f_i X_i^2$, is the frequency of occurrence multiplied by the square of the X value. The latter, $(\sum f_i X_i)^2$, is obtained by multiplying each f by each X, adding all these products, and then squaring this total. This is quite a different value than the former.

2. With respect to point (4) above, note that

$$f_i(X_i^2) = f_i X_i^2 = f_i(X_i X_i) = (f_i X_i)X_i$$

so that the values $f_i X_i^2$ can be obtained by multiplying the values $f_i X_i$, used in calculating the mean, by another X_i.

3. Finally, it must be the case that

$$\sum f_i X_i^2 \geq \frac{(\sum f_i X_i)^2}{n}$$

so that

$$\left[\sum f_i X_i^2 - \frac{(\sum f_i X_i)^2}{n}\right] \geq 0$$

That is, this expression is always positive, and the variance is always a positive number. This is true because the bracketed expression is just another way in which the sum of the squares of the deviations about the mean can be expressed. Since a square of a value is always a positive number, the sum of squares is always positive. The equivalence of the bracketed expression of Definition 5.9.6 with the sum of squares of the deviations about the mean will be shown later in this Chapter.

Example 5.9.5 Age and Sex Distributions of Saskatchewan Suicides, 1985-86

The age distribution of suicides of males and of females in Saskatchewan for the years 1985-86 is given in Table 5.17. This table is drawn from the

Age in Years	Number of Suicides Male	Female
14	1	1
15-19	20	4
20-29	63	9
30-64	101	37
65 and over	25	11
Total	210	62

Table 5.17: Number of Suicides by Age and Sex, Saskatchewan 1985-86

publication **Suicide in Saskatchewan: The Alcohol and Drug Connection 1988**, *Table 3, page 4, produced by the Saskatchewan Alcohol and Drug Abuse Commission (SADAC). For each sex, the variance and standard deviation of the age of suicides are determined as follows.*

In order to determine the standard deviation for each of these distributions, it is necessary to obtain X values for each of the intervals into which the suicides have been grouped. For each of the intervals, 15-19, 20-29 and 30-64, the X values used here are the midpoints of the respective intervals. For the open ended interval, 65 and over, $X = 70$ has been picked as the mean age of the suicides for those aged 65 and over. Based on this, the calculations required for determining the variances and standard deviations are given in Table 5.18. For males, Table 5.18 gives the values $n = 210$ and

$$\sum fX^2 = 389,400.75$$

$$\sum fX = 8,394.5$$

Entering these into the formula for the variance in Definition 5.9.6 gives the following

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left[\sum fX^2 - \frac{\left(\sum fX\right)^2}{n}\right] \\ &= \frac{1}{209}\left[389,400.75 - \frac{(8,394.5)^2}{210}\right] \end{aligned}$$

	Males			Females		
X	f	fX	fX^2	f	fX	fX^2
14	1	14.0	196.00	1	14.0	196.00
17.5	20	340.0	5,780.00	4	68.0	1,156.00
24.5	63	1,543.5	37,815.75	9	220.5	5,402.25
47	101	4,747.0	223,109.00	37	1,739.0	81,733.00
70	25	1,750.0	122,500.00	11	770.0	53,900.00
Total	210	8,394.5	389,400.75	62	2,811.5	142,387.25

Table 5.18: Calculations for Variation in Age of Suicides in Saskatchewan, by Sex, 1985-86

$$\begin{aligned} &= \frac{389,400.75 - 335,560.14}{209} \\ &= \frac{53,840.606}{209} \\ &= 257.61055 \end{aligned}$$

The standard deviation is

$$s = \sqrt{257.61055} = 16.050$$

Thus the standard deviation in the age of suicides for Saskatchewan males was $s = 16.0$ *years in 1985-86.*

For Saskatchewan females, the sample size is $n = 62$ *and from Table 5.18*

$$\sum fX^2 = 142,387.25$$

$$\sum fX = 2,811.5$$

so that the variance is

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left[\sum fX^2 - \frac{(\sum fX)^2}{n}\right] \\ &= \frac{1}{61}\left[142,387.25 - \frac{(2,811.5)^2}{62}\right] \end{aligned}$$

$$\begin{aligned} &= \frac{142,387.25 - 127,492.46}{61} \\ &= \frac{14,894.794}{61} \\ &= 244.17696 \end{aligned}$$

The standard deviation is

$$s = \sqrt{244.17696} = 15.626$$

or $s = 15.6$ *years.*

Table 5.19 summarizes the results, adding some of the measures of central tendency as well. An examination of the two distributions shows that the average age of suicides for males is lower than the average age of female suicides. This is the case whether the median or mean age of suicides is examined. The average is lower for males than for females because there are many more suicides of males at ages 15-19 for males than for females. Even

Measure	Males	Females
Median	36.8	45.6
Mean	40.0	45.3
IQR	30.5	29.3
s^2	257.6	244.2
s	16.0	15.6

Table 5.19: Summary Measures, Age of Suicides, Males and Females

though the distribution of suicides for females has a considerably greater average age, the variation in age of suicides is much the same for males and females. The standard deviations for males and females are almost exactly the same at about 16 years and the variances are almost the same as well. The interquartile range of the age of suicides is also very similar for both sexes.

In summary, the centre of the distribution for males is lower than for females, but the variation for the two distributions is fairly similar, as measured by the variance, standard deviation or IQR.

Proofs of Formulae. When examining the formale for ungrouped data, some proofs of the equivalence of different formulae was given. In the fol-

lowing paragraphs, the same proofs for grouped data are given. Again, this section can be skipped if you do not feel adept at algebra.

In the case of grouped data, the sum of the deviations about the mean, if weighted by the frequencies of occurrence, add to zero. This is shown in the following expressions. Note that all of the summations proceed across all k values into which the data has been grouped.

$$\begin{aligned}\sum_{i=1}^{n} f_i(X_i - \bar{X}) &= f_1(X_1 - \bar{X}) + f_2(X_2 - \bar{X}) + \cdots + f_k(X_k - \bar{X}) \\ &= [f_1X_1 + f_2X_2 + \cdots + f_kX_k] - [f_1\bar{X} + f_2\bar{X} + \cdots + f_k\bar{X}] \\ &= \sum f_iX_i - \bar{X}\sum f_i\end{aligned}$$

For grouped data,

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

where

$$n = \sum f_i$$

so that

$$\begin{aligned}\sum_{i=1}^{n} f_i(X_i - \bar{X}) &= \sum f_iX_i - \left[\frac{\sum f_iX_i}{n}\right]\sum f_i \\ &= \sum f_iX_i - \left[\frac{\sum f_iX_i}{n}\right]n \\ &= \sum f_iX_i - \sum f_iX_i = 0\end{aligned}$$

The equivalence of the formulae of Definitions 5.9.5 and 5.9.6 is shown as follows. The claim is that

$$s^2 = \frac{1}{n-1}\sum f_i(X_i - \bar{X})^2 = \frac{1}{n-1}\left[\sum f_iX^2 - \frac{(\sum f_iX_i)^2}{n}\right]$$

Since each of the summation parts of these expressions is multiplied by the same value $1/(n-1)$, all that has to be shown is that

$$\sum f_i(X_i - \bar{X})^2 = \left[\Sigma f_iX^2 - \frac{(\sum f_iX_i)^2}{n}\right]$$

Expanding the left side gives

$$\begin{aligned}\sum f_i(X_i - \bar{X})^2 &= \sum f_i(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\ &= \sum f_i X_i^2 - 2\bar{X}\sum f_i X_i + \sum f_i(\bar{X}^2)\end{aligned}$$

The middle term can be written as

$$-2\bar{X}\sum f_i X_i = -2\bar{X}n\left[\frac{\sum f_i X_i}{n}\right] = -2n(\bar{X})^2$$

The term on the right becomes

$$\sum f_i(\bar{X}^2) = (\bar{X})^2\sum f_i = n(\bar{X})^2$$

As a result, the original expression becomes

$$\begin{aligned}\sum f_i(X_i - \bar{X})^2 &= \sum f_i X_i^2 - 2\bar{X}\sum f_i X_i + \sum f_i(\bar{X}^2) \\ &= \sum f_i X_i^2 - 2n(\bar{X})^2 + n(\bar{X})^2 \\ &= \sum f_i X_i^2 - n(\bar{X})^2 \\ &= \sum f_i X_i^2 - n\left[\frac{\sum f_i X_i}{n}\right]^2 \\ &= \sum f_i X_i^2 - \frac{\left(\sum f_i X_i\right)^2}{n}\end{aligned}$$

This shows the equivalence of the two expressions.

5.9.3 Interpretation of the Standard Deviation

As noted earlier in this section, there is no easy, intuitive explanation for the standard deviation. The particular formula used to obtain the standard deviations involves sums of squares of the deviations about the mean, an averaging of this sum (i.e. dividing by $n - 1$), and then taking a square root. Once all this has been done, it is difficult to obtain an intuitive idea of this measure of variation. In this section, some comments concerning interpretation of the standard deviation are made. Hopefully these will assist in understanding how this measure can be used.

Units for the Standard Deviation. As noted earlier, the units for the standard deviation are the same units as the units used to measure the variable. This occurs because the deviations about the mean are in the original

units, and these deviations are first squared, and after some manipulation of these squares, a square root is taken. This means that the standard deviation ends up being measured in the units in which the variable X is measured, while the variance is measured in the square of these units.

While this may not be of too much assistance, this means for example, that if distributions of income are measured in dollars, the standard deviation will also be in dollars. Table 5.20 contains summary measures for a number of variables describing Saskatchewan families in 1988. These summary measures are obtained from data in Statistics Canada's Survey of Consumer Finances. This Survey provides data on a variety of income and labour force characteristics of Saskatchewan families. In this table, only those families containing both a husband and a wife are included. All single parent families, and households containing only single people, were excluded when preparing this table.

Table 5.20 gives some idea of the units in which some standard deviations are measured, and also gives an idea of the size of the standard deviation of a number of relatively familiar variables. Family income is measured

Variable	Units	$\bar{X}$	s
Family Income	Dollars	\$40,900	\$25,500
No. of Earners	No. of people	1.70	0.99
No. of Persons	No. of people	3.25	1.29
Husband's Work Yearly	Weeks	38.0	21.5
Age of Husband	Years	47.7	16.2

Table 5.20: Summary Measures of Variables, Saskatchewan, 1988

in dollars, and had a standard deviation of \$25,500 in 1988. The number of earners per family and the number of persons in each family are both measured in numbers of people. Since there are not too many earners, or people, in each family, it can be seen that these standard deviations are relatively small. The standard deviation in the number of weeks worked per year for husbands 21.5 weeks, and the standard deviation of age for these same husbands is 16.2 years.

Size of the Standard Deviation. The size of the standard deviation is apparent once it is calculated or given, as in Table 5.20. But if you

are not familiar with the concept of standard deviation, it is difficult to guess the approximate size of a standard deviation without carrying out the calculations. On rule of thumb which assists is:

As a rule of thumb, the standard deviation is approximately equal to the range divided by 4. That is,

$$s \approx \frac{\text{Range}}{4}$$

This is a very rough rule of thumb but provides some idea of the order of magnitude for a standard deviation. Table 5.21 gives the range, the range divided by 4, and the standard deviation for the variables describing Saskatchewan husband-wife families. It can be seen that in some cases, such as age or number of people, this rule provides a fairly good idea of the approximate size of a standard deviation. In cases such as family income, this rule is not very accurate. However, in the case of family income, if the top 1% of family incomes are eliminated, this produces a range of $132,000 for the bottom 99% of family incomes. Dividing this by 4 gives $33,000, closer to the actual value of $25,500.

Variable	Range	$\frac{\text{Range}}{4}$	s
Family Income	$256,800	$64,200	$25,500
No. of Earners	6	1.5	0.99
No. of Persons	8	2	1.29
Husband's Work Yearly	52	13	21.5
Age of Husband	63	15.75	16.2

Table 5.21: Range and Standard Deviation

While this rule of thumb is useful, it is no more than a very rough rule, and one which can provide some very general idea of the standard deviation. It can provide a very rough check on calculations though. For example, if you had calculated the standard deviation of the number of people per household to be 129, rather than 1.29, a quick check of the range divided by 4 would tell you that 129, or even 12.9, is much too large a number for the standard deviation. In fact, the standard deviation cannot be larger than the range,

so any calculation showing a standard deviation larger than the range must be incorrect.

Relative Size of Standard Deviations. In all of the examples of the standard deviation, two distributions were given, and the sizes of the standard deviations compared. This is where the standard deviations are most useful. Suppose there are several different samples, where the same variable is being measured in each sample. Then the sample whose distribution has larger values for the standard deviation can be considered to be more varied, and the samples with smaller standard deviations can be considered to be less varied. Using some common socioeconomic variables, this is illustrated in the following example.

Example 5.9.6 Variation in Canadian Urban Homicide Rates

Table 5.9.6 contains a variety of socioeconomic variables for the 24 Census Metropolitan Areas in Canada. The observations for these 24 cities were taken at various points in time, as shown in the table. This data is taken from Leslie W. Kennedy, Robert A. Silverman and David R. Forde, "Homicide in Urban Canada: Testing the Impact of Economic Inequality and Social Disorganization," **Canadian Journal of Sociology** *16 (4), Fall 1991, pages 397-410. In the abstract the authors state that*

> Homicide in Canada is regionally distributed, rising from east to west. This study demonstrates a reduction in the regional effect through a convergence in homicide rates between eastern, central, and western Canada in Census Metropolitan Areas (CMAs) with higher levels of inequality and social disorganization.

While Table 5.22 does not show this directly, this table does allow the reader to examine the shifts in the average and variation for some commonly used socioeconomic variables. Using this summary data, it can be seen that there was little change in the variability in homicide rates between 1972-76 when the standard deviation was 0.89, to 1977-81, when the standard deviation was 0.86. However, these values are both about double the standard deviation of 0.43 of 1967-71. This shows that across the 24 CMAs, homicide rates were about twice as varied in the 1970s, as compared with the late 1960s.

With respect to some of the other variable, the authors make the following comments (pages 402-3):

Variable	Year	$\bar{X}$	s	Minimum	Maximum
Homicides per 100,000	1967-71	0.94	0.43	0.09	1.71
	1972-76	1.73	0.89	0.41	4.26
	1977-82	1.77	0.86	0.52	3.77
Unemployment Rate (%)	1971	8.15	2.00	6.03	15.00
	1976	6.82	1.70	3.38	10.48
	1981	7.40	3.14	3.27	15.78
% Males 20-34	1971	17.75	2.16	14.89	24.10
	1976	26.46	1.77	23.00	29.63
	1981	15.42	1.38	12.50	18.55
% Divorced	1971	1.31	0.65	0.09	2.70
	1976	1.99	0.69	0.82	3.53
	1981	2.92	0.70	1.49	4.31

Table 5.22: Summary Statistics, Canadian Census Metropolitan Areas

> ... unemployment, while generally at the same mean rate, increases in variability. This is evident in the increased standard deviation and the greater range in unemployment rates across CMAs. ... the proportion of young males in CMAs increases substantially in 1976, but the variability across cities drops over the ten-year period. Finally, divorce increases in terms of the percentage divorced within CMAs but remains stable in terms of variability across CMAs. These changes may in part be attributed to lack of opportunities for young persons in CMAs, and changes in legal access to divorce.

Each of these statements can be verified by examining the table. Note that the unemployment rate is the percentage of the labour force which is unemployed, the % Divorced is the percentage of the population divorced in the CMAs, and the % Males 20-34 is the "percentage of young males of age 20 through 34 in CMAs." (page 402).

Percentage of Cases Around the Mean. Another useful way to think of the standard deviation and the mean is to ask the question

> What percentage of cases lie within a distance of one standard deviation on each side of the mean?

Since the standard deviation and mean are both measured in the units of the variable X, they can be considered as distances along the horizontal axis. Then the number, or percentage, of the cases in the data set, which lie within a certain distance of the mean can be determined. The following guidelines, while again very rough rules of thumb, generally can be considered to hold for any distribution.

Supppose a variable X is measured for all the cases in a data set, and that the mean value of X for the cases in this data set is $\bar{X}$ and the standard deviation is s. Then

1. The interval from $X - s$ to $X + s$ usually contains about two thirds of all the cases in the data set. Alternatively stated, the interval

   $$(\bar{X} - s, \bar{X} + s)$$

   contains approximately 67% of all the cases in a distribution.

2. The interval from $X - 2s$ to $X + 2s$ usually contains around 95% of all the cases in the data set. That is, the interval

   $$(\bar{X} - 2s, \bar{X} + 2s)$$

   contains approximately 95% of all the cases in a distribution.

3. The interval from $X - 3s$ to $X + 3s$ usually contains 99% or more of all the cases in the data set. That is, the interval

   $$(\bar{X} - 3s, \bar{X} + 3s)$$

   contains approximately 99% of all the cases in a distribution.

Note that the latter point means that very few cases in a distribution are more than 3 standard deviations away from the mean. Point (2) means that the large bulk of cases are within 2 standard deviations of the mean. Point (1) means that well over one half of all the cases are within one standard deviation of the mean. As a result, if the mean and the standard deviation are known, a considerable amount is known about a distribution. This is illustrated in the following example.

Example 5.9.7 Hours Worked per Week

In Chapter 4, the hours of work of 50 Regina labour force members was given in Table **??**. *For this list of 50 values of hours worked per week, the mean hours is* $\bar{X} = 37.0$ *and the standard deviation is* $s = 10.5$. *The interval of one standard deviation on each side of the means is*

$$\bar{X} \pm s = 37.0 \pm 10.5 \quad or \quad (26.5, 47.5)$$

This means that two thirds or more of the hours worked in the data set might be expected to be between 26.5 hours and 47.5 hours worked per week. The ordered stem and leaf display of Figure **??** *provides a quick way of checking to see how many hours per week actually do fall in this range. Counting the number of values between the limits of 26.5 and 47.5 hours per week gives a total of 38 of the 50 respondents having hours worked per week that fall within this range. This is*

$$\frac{38}{50} \times 100\% = 76\%$$

of the cases, more than the 67% expected.

Within two standard deviations of the mean is an interval

$$\bar{X} \pm 2s = 37.0 \pm 21.0 \quad or \quad (16.0, 58.0)$$

This contains all the cases except those three workers who work 3, 4, and 10 hours per week. That is, the interval contains 47 out of 50, or 94% of the cases.

The interval that is three standard deviations on either side of the mean is

$$\bar{X} \pm 3s = 37.0 \pm 31.5 \quad or \quad (5.5, 68.5)$$

This interval contains all but the two workers who work only 3 and 4 hours per week. This is 48 out of 50 or 96% of all the cases, a little less than the 99% expected.

Example 5.9.8 Incomes of 50 Regina Families

A similar example is the stem and leaf display presented in Figure **??** *of Chapter 4. The stem and leaf display there organizes the incomes of 50 Saskatchewan families. The mean and standard deviation for these 50 families are* $\bar{X} = 36.3$ *thousand dollars, and* $s = 32.7$ *thousand dollars. You*

can check the stem and leaf displays to verify that 43 out of the 50 families have incomes within one standard deviation of the mean income, 47 out of 50 are within 2 standard devations, and 49 out of 50 are within 3 standard deviations of the mean.

Example 5.9.9 Distribution of Gross Monthly Pay of 601 Regina Adults

The distribution of gross monthly pay of 601 Regina respondents, given in Table 5.23, is drawn from the Social Studies 203 Regina Labour Force Survey. A histogram for the frequency distribution in Table 5.23 is given in Figure 5.3. A quick examination of Table 5.23 and Figure 5.3 shows that the distribution of gross monthly pay peaks at a fairly low pay level, around $1,500-2,000 per month, and then tails off as one moves to higher income levels. However, there are some individuals with quite high pay levels, so that the distribution goes much further on the right than on the left of the peak income level. Such a distribution is considered to be skewed to the right. Distributions of income and wealth are ordinarily skewed in this manner. However, the rules concerning the percentage of cases within various distances from the mean should still hold in this distribution. In Figure 5.3, a histogram of the frequency distribution is presented. Note that most of the intervals are $500 wide, so that the frequencies of occurrence for these intervals are presented as in Table 5.23. For the $4,000 to $4,999 interval, which represents an interval width of $1,000, the density has been calculated as the frequency of occurrence per $500 of interval width. The interval of $1,000 width is equivalent to two intervals of $500 width, so that the density in this interval is $46/2 = 23$ *cases per $500. The open ended interval is drawn to indicate a considerable number of cases of $5,000 or more, and the proper height of this bar is a guess. Figure 5.3 shows the mean at $2,352, and the intervals around the mean are as follows. The standard deviation is $1,485, so that the interval from one standard deviation less than the mean to one standard deviation greater than the mean is the interval*

$$(\bar{X} - s \ , \ \bar{X} + s)$$

$$(\$2,352 - \$1,485 \ , \ \$2,352 + \$1,485)$$

$$(\$867 \ , \ \$3,837)$$

While the detailed frequency distribution giving the number of adults with each value of gross monthly pay, is not given here, it turns out that there

Gross Monthly Pay ($ per month)	Frequency
Less than 500	45
500-999	51
1,000-1,499	69
1,500-1,999	110
2,000-2,499	77
2,500-2,999	60
3,000-3,499	59
3,500-4,000	52
4,000-4,999	46
5,000 and over	32
Total	601
Mean	$2,352
Standard Deviation	$1,485
Minimum	$50
Maximum	$9,000
Median	$2,000

Table 5.23: Distribution of Gross Monthly Pay, 601 Regina Respondents

are 433 out of the 601 adults who have pay between $867 and $3,837. This is

$$\frac{433}{601} \times 100\% = 0.720 \times 100\% = 72.0\%$$

of all the cases. This is over the two thirds of 67% of the cases that might generally be expected to be within one standard deviation of the mean.

Two standard deviations is $2 \times \$1,485 = \$2,970$ *so that the mean plus or minus $2,970 is the interval*

$$(\$2,352 - \$2,970\ ,\ \$2,352 + \$2,970)$$

$$(\$0\ ,\ \$5,322)$$

Here the lower end of this interval is really less than 0, but since there are no pay levels below 0, the interval is stopped at 0. Going back to the

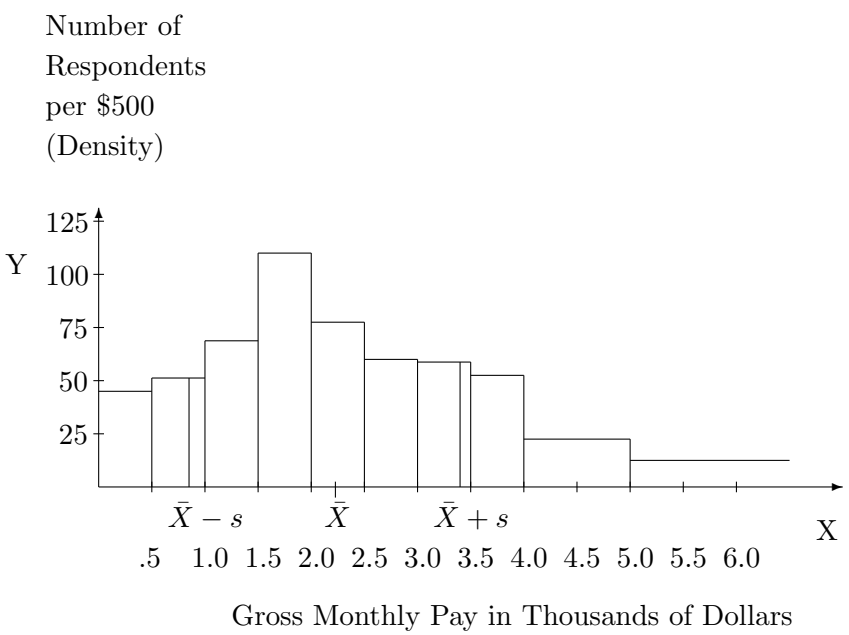

Figure 5.3: Histogram of Distribution of Gross Monthly Pay

original detailed frequency distribution from which this data is drawn, it is found that all but 20 of the Regina adults have pay levels of below $5,322 per month. This means that there are $(581/601) \times 100\% = 96.7\%$ *of cases within two standard deviations of the mean. This is more than the 95% that can generally be expected.*

Within three standard deviations, there are all but 8 cases. That is, between a gross monthly pay of 0 and $2,352 + 3(1,485) = 6,807$ *dollars there are 593 of the 601 cases. This amounts to 98.7% of all cases.*

The mean, and the intervals around the mean are illustrated in Figure 5.3. Remeber that 100% of the cases are in the distribution. It was found that 72% of the cases were in the interval from $867 to $3,837. This means that the area from $867 to $3,837 contains 72% of the area in the bars of the histogram of Figure 5.3. Similarly, the area in the bars of the histogram between 0 and $5,322 contains approximately 97% of the total area in the histogram.

The mean and the standard deviation together provide a great deal of information concerning the distribution. The mean provides an idea of the

centre of a distribution, and the intervals of one, two and three standard deviations around the mean provide a good idea of where the bulk of the cases are. When summarizing a distribution, if the mean, standard deviation and sample size of the sample are reported, this gives those examining the data a considerable amount of information concerning the nature of the distribution.

5.10 Percentage and Proportional Distributions

If the distribution for a variable is given as a percentage distribution then the determination of the mean is straightforward (see Definition **??**) earlier in this Chapter. In the original formula for the mean,

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

the sample size in the denominator is replaced with the value of 100, the sum of the percentages, producing the formula

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

for the mean of a percentage distribution.

When working with the formulae for the variance and the standard deviation, the sample size is reduced by 1, so that $n-1$ is in the denominator, rather than n. This would appear to imply that $100 - 1 = 99$, rather than 100, should be used in the denominator for the variance when working with a percentage distribution. The problem with this approach is that this amounts to always subtracting 1% of cases, rather than 1 case. Rather than do this, if n is not known, or if n is reasonably large, then it is best to use 100 in the denominator. This produces the following definitions for a percentage distribution.

Definition 5.10.1 If a variable X has k values

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

occurring with respective percentages

$$P_1, P_2, P_3, \cdots, P_k$$

and the mean of these k values is

$$\bar{X} = \frac{\sum P_i X_i}{100}$$

where the summation is across all k values of X_i and

$$\sum P_i = 100$$

Summing across the same k values, the **variance** is

$$s^2 = \frac{\sum P_i (X_i - \bar{X})^2}{100}$$

The standard deviation of this set of k values of X is

$$s = \sqrt{\frac{\sum P_i (X_i - \bar{X})^2}{100}}$$

Using the alternative, more computationally efficient formulae,

$$s^2 = \frac{1}{100}\left[\Sigma P_i X_i^2 - \frac{(\Sigma P_i X_i)^2}{100}\right]$$

and

$$s = \sqrt{\frac{1}{100}\left[\Sigma P_i X_i^2 - \frac{(\Sigma P_i X_i)^2}{100}\right]}$$

For a proportional distribution, the percentages, P, are replaced with the proportions, p, and these sum to 1, rather than 100. In the case of a proportional distribution, the standard deviation is

$$s = \sqrt{\left[\Sigma p_i X_i^2 - (\Sigma P_i X_i)^2\right]}$$

If the sample size on which the distribution is based is known and is quite small, it may be best to convert the data back into the actual number of cases in each category and use the original formula. However, if the sample size is very large or the frequency distribution refers to the whole population, then the formulae given in Definition 5.10.1 should be used. The sample size on which the data is based should always be given. Unfortunately, in published data, the sample size is often not given.

Example 5.10.1 Distribution of Family Income in Canada, 1984

The following data in Table 5.24 comes from Statistics Canada's Survey of Consumer Finances for 1984. This table gives the percentage distribution of the income of families in Canada for 1984. The calculations for the standard deviation of family income are given in the table, and in the formulae which follow. Note that the table has been set up so that the incomes are in thousands of dollars. The midpoint of each interval has been selected as the appropriate X value in each case, with \$65,000 selected as the appropriate mean value of the \$45,000 and over income interval.

Income in \$'000s	Per Cent of Families P_i	X_i	P_iX_i	$P_iX_i^2$
0-10	7.1	5.0	35.50	177.500
10-15	9.7	12.5	121.25	1,515.625
15-20	9.7	17.5	169.75	2,970.625
20-25	9.0	22.5	202.50	4,556.250
25-30	10.1	27.5	277.75	7,638.125
30-35	10.1	32.5	328.25	10,668.125
35-45	17.6	40.0	704.00	28,160.000
45 plus	26.7	65.0	1,735.50	112,807.500
Total	100.0		3,574.50	168,493.750

Table 5.24: Distribution of Family Income, Canada, 1984

From Table 5.24,

$$\Sigma P_i X_i = 3,574.50$$

$$\Sigma P_i X_i^2 = 168,493.750$$

Thus the standard deviation is

$$s = \sqrt{\frac{1}{100}\left[\Sigma P_i X_i^2 - \frac{(\Sigma P_i X_i)^2}{100}\right]}$$

$$
\begin{aligned}
&= \sqrt{\frac{1}{100}\left[168,493.750 - \frac{(3,574.50)^2}{100}\right]} \\
&= \sqrt{\frac{168,493.750 - 127,770.500}{100}} \\
&= \sqrt{\frac{40,723.248}{100}} \\
&= \sqrt{407.232} = 20.180
\end{aligned}
$$

The standard deviation of family income for 1984 in Canada is estimated to be $20,180 based on the above table and formula.

5.11 Measures of Relative Variation

All of the measures of variation discussed so far are measures based on the units in which the variable is itself measured. For example, the range, interquartile range or standard deviation of the heights of a group of people would be expressed as so many inches, or in centimetres if the metric system were used. This has the advantage of giving these measures in familiar units. In the case of the standard deviation, intervals around the mean can be constructed, and the percentage of cases in each of these intervals can be determined. All of the measures of variation discussed so far can be considered to be **measures of absolute variation**.

A different approach can be taken by considering how varied the distribution is, where the measure of variation is considered as being relative to another measure. These measures of variation are considered to be **measures of relative variation**. The idea of such measure can be obtained by considering as an example, the problem of how to compare the variation in heights of children with the variation in heights of adults.

Suppose, for example, that we were to compare the variation in heights of children of age 3 with the variation in heights of adults. It is likely that the standard deviation of heights of children be a fairly small number. This is because the heights of children are relatively small numbers and the deviations in heights of individual children about the mean height for all children of age 3 are not large numbers. For adults, the standard deviation of height is likely to be a larger number because the numbers expressing heights of adults are larger in absolute value. Because of this, the numbers expressing deviation of adult heights about the mean height for all adults are also likely to be larger. The result is that the standard deviation of adult

heights is likely to be larger than the standard deviation of the heights of children. That is, in absolute terms, the variation in heights of adults is likely to be greater than the variation in heights of children. The main reason that the standard deviation of adult heights exceeds the standard deviation of heights of children is that adults are taller on average than are children. The above reasoning suggests that it might be useful to examine the variation in height, relative to the average height.

One measure which results from the above reasoning is based on the standard deviation divided by the mean. This is defined as follows.

Definition 5.11.1 The **coefficient of relative variation** (CRV) or simply the **coefficient of variation** is defined as

$$CRV = \frac{s}{\bar{X}} \times 100$$

Sometimes the CRV is defined as simply the ratio of the standard deviation to the mean, that is, as $s/\bar{X}$. In this text, the first of the two definitions will be used.

In the case of the heights of children and adults, the CRV, determined as the ratio standard deviation to the mean, multiplied by 100, may be much the same size for both children and adults. This is because for both children and adults, there is likely to be a similar degree of variation of height relative to their mean height.

The CRV is useful for two major reasons. First, sometimes there will be two distributions which describe the distribution of similar variables but these variables are measured in different units. If this is the case, then the standard deviations are not directly comparable, whereas the coefficients of relative variation are comparable. That is, each standard deviation is measured in the units in which the variable has been measured. Two standard deviations in two different units cannot be directly compared, in order to determine which standard deviation represents greater variation. But the two CRVs can be directly compared.

Second, when a variable X has larger numbers than another variable Y, it may be the case that X also has a larger standard deviation than does Y. This does not mean that the distribution of X is inherently more dispersed than that of Y. The larger standard deviation may just reflect the fact that with larger numbers, the data is more spread out in an absolute sense. But in relative terms, there is little difference in variation relative to the mean.

The coefficient of relative variation is a number, with no units. This occurs because the CRV is the ratio of two other numbers, both of which are measured in the same units. The CRV is thus dimensionless and can be meaningfully compared for any two different distributions.

Example 5.11.1 Attitudes Measured on Two Different Scales

Two sample surveys of adults, one taken in Regina and the other in Edmonton, asked similar questions concerning whether or not it is too easy to get welfare assistance. The Regina survey asked the question, "Is it too easy to get welfare assistance?" Respondents were asked to give their responses as one of "Strongly agree, somewhat agree, somewhat disagree, or strongly disagree." The Edmonton survey made the statement, "Unemployment is high because unemployment insurance and welfare are too easy to get," and asked respondents to give their response to this statement on a 7 point scale, where 1 is strongly disagree and 7 is strongly agree.

While the two questions are not really the same, they are both likely to reflect some of the underlying views of respondents with respect to unemployment insurance and welfare. The percentage distributions of responses are given in the first parts of Tables 5.25 and 5.26 and the calulations required for the standard deviation in the last two columns of each table. For these distributions, determine the standard deviation and coefficient of relative variation.

Response	X	f	fX	fX^2
Strongly Agree	1	257	257	257
Somewhat Agree	2	239	478	956
Somewhat Disagree	3	177	531	1,593
Strongly Disagree	4	88	352	1,408
Total		760	1,618	4,214

Table 5.25: Responses to Regina Question

For each table, it is necessary to compute the mean and standard deviation. For the Regina survey, the results are as follows.

$$\sum fX^2 = 4,214$$

Response	X	f	fX	fX^2
Strongly Disagree	1	50	50	50
	2	42	84	168
	3	42	126	378
Neutral	4	41	164	656
	5	65	325	1,625
	6	71	426	2,556
Strongly Agree	7	74	518	3,626
Total		385	1,693	9,059

Table 5.26: Responses from Edmonton Survey

$$\sum fX = 1,618$$

and $n = 761$. *Entering these values into the formula for the variance in Definition 5.9.6 gives the following*

$$\bar{X} = \frac{\sum fX}{n} = \frac{1,618}{761} = 2.126$$

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left[\sum fX^2 - \frac{\left(\sum fX\right)^2}{n}\right] \\ &= \frac{1}{760}\left[4,214 - \frac{(1,618)^2}{761}\right] \\ &= \frac{4,214 - 3,440.1104}{760} \\ &= \frac{773.88962}{760} \\ &= 257.61055 \end{aligned}$$

The standard deviation is

$$s = \sqrt{257.61055} = 1.0183$$

and the coefficient of relative variation is

$$CRV = \frac{s}{\bar{X}} \times 100 = \frac{1.0183}{2.126} \times 100 = 0.47893 \times 100 = 47.893$$

For the Edmonton survey the results are

$$\sum fX^2 = 9,059$$

$$\sum fX = 1,618$$

and $n = 385$. *Entering these values into the formula for the variance in Definition 5.9.6 gives the following*

$$\bar{X} = \frac{\sum fX}{n} = \frac{1,693}{385} = 4.397$$

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left[\sum fX^2 - \frac{\left(\sum fX\right)^2}{n}\right] \\ &= \frac{1}{384}\left[9,059 - \frac{(1,693)^2}{385}\right] \\ &= \frac{9,059 - 7,444.8026}{384} \\ &= \frac{1,614.1974}{384} \\ &= 4.20364 \end{aligned}$$

The standard deviation is

$$s = \sqrt{4.20364} = 2.0503$$

and the coefficient of relative variation is

$$CRV = \frac{s}{\bar{X}} \times 100 = \frac{2.0503}{4.397} \times 100 = 0.46625 \times 100 = 46.625$$

All of these results are summarized in Table 5.27. In each case, the values have been rounded to 2 significant figures. While the computations here are accurate, the scale for each variable is an ordinal scale, and yet it is being treated as an interval scale. As a result, the values of the various summary measures appear more accurate than they really are.

If the range, interquartile range, standard deviation or variance is examined in these two distributions, it appears as if the variation in responses in Edmonton is considerably greater than the variation in responses in Regina. The absolute variation for Edmonton is approximately double that for Regina. But the main reason the variation for Edmonton appears greater

Measure	Regina	Edmonton
Mean	2.1	4.4
s^2	1.0	4.2
s	1.0	2.1
CRV	48	47
Range	4	7
IQR	2	3
IQR/Median	1	0.6

Table 5.27: Summary Measures, Attitude Questions, Regina and Edmonton

than for Regina, is that attitudes are measured on two different scales. The 7 point attitude scale for Edmonton has built into it a much greater range of attitudes than does the 4 point scale of the Regina survey. If only the measures of absolute variation are examined, the impression that would be taken from these two distributions is that Edmonton adults were much more varied in their responses than were Regina adults.

Measures of relative variation give quite a different picture. The CRV for Regina ends up being slightly greater than the CRV for Edmonton, although the two are practically identical. Another possible measure of relative variation, the IQR divided by the median, actually shows a lower value for Edmonton than it does for Regina. What produces this approximately equal relative variation for the two cities is the considerably larger average for Edmonton than for Regina. This larger mean occurs because the scale of attitudes is allowed to take on considerably more values in Edmonton than it does in Regina.

For comparing attitudes in these two cities, all the measures are useful. But since the scales are so different in these two surveys, the measures of relative variation are superior to the measures of absolute variation here. That is, the measures of relative variation correct for the differences in the scale, and show that once the scale differences are takin into account, the variability of attitudes in the two cities is very similar.

Example 5.11.2 Relative Variation in Canadian Urban Homicide Rates

In Example 5.9.6, the standard deviation of homicide rates was used as a measure of absolute variation. In Table 5.22 the distribution of homicide rates across 24 Canadian cities was shown to have a larger standard deviation and range in the 1970s than in the late 1960s. However, if measures of relative variation are constructed, as in Table 5.28, it appears as if there was very little shift in the relative variability in urban homicide rates over the periods shown. The CRV changes very little, although it is slightly larger in the years 1972-1976 than in other periods. The Range has been divided by the mean, in order to construct a measure of the relative range, that is, the range relative to the average homicide rate. Again, this shows less change than the Range, and this measure of relative variation is very similar for 1977-82 and 1967-71. The reason for the small shift in the relative

Variable	Year	$\bar{X}$	s	CRV	Range/$\bar{X}$
Homicides per 100,000	1967-71	0.94	0.43	46	1.72
	1972-76	1.73	0.89	51	2.23
	1977-82	1.77	0.86	49	1.84

Table 5.28: Summary Measures of Relative Variation in Canadian Urban Homicide Rates

variation is that homicide rates across the country increased considerably in the 1970s. The larger values of homicide rates produced larger differences among cities in their homicide rates. But relative to the typical, or average, homicide rate, the variation in homicide rates across the 24 cities changed little. You can compute the CRV and the Range divided by the mean for the other socioeconomic variables in Table 5.22 to see whether the same conclusion holds for these other variables. If it does, then this casts some doubt on the authors' conclusions in Example 5.9.6. However, in some cases it appears that both relative and absolute measures of variation give similar conclusions.

Example 5.11.3 Income Inequality

One situation where the CRV may be used is when data in dollars is to be compared over different years. As we all know, inflation erodes the value of the dollar each year, so it does not make a great deal of sense to compare figures in dollars in two different years unless one first corrects for changes in the value of the dollar over those years. One way of doing this is to construct income in constant or real dollars (see Example 5.11.4). When looking at variation, another method is to examine the changes in the CRV.

	Value in Current Dollars		
Year	Mean	s	CRV
1954	2374	2400	101.1
1961	3110	2727	87.7
1969	4713	4860	103.1
1971	5389	6479	120.2
1973	6383	6352	99.5

Table 5.29: Measures of Income Inequality, Canada, Selected Years

In Table 5.29, drawn from Statistics Canada, **Income Inequality: Statistical Methodology and Canadian Illustrations***, 13-559, page 78, it can be seen that the degree of variation in incomes, as measured by the standard deviation, increased dramatically from one year to the next for the years shown. Based on this, one would be tempted to conclude that there had been close to a tripling in the inequality of incomes, over the period shown. Such a conclusion is incorrect because the value of the dollar also changed dramatically over this period, although one cannot tell by how much, based on these figures. In this latter connection, it might be noted that mean income rose partly because of inflation and partly because incomes rose in real terms.*

In order to get a more accurate idea of whether the distribution of incomes is more or less equal, examine the CRV column. There, it can be seen that, relative to the mean, the standard deviation sometimes declined and sometimes rose. For the years from 1954 to 1961, there was a decline in CRV,

indicating a decline in the inequality of incomes. From 1961 through 1971, it appears that there was a gradual increase in the inequality of incomes and after 1971, incomes again became slightly more equally distributed.

Example 5.11.4 Income Distributions

The question of whether relative or absolute differences in income present the best picture of income distribution has always been a matter of some debate. One of the arguments presented by those who favour programs of government assistance to help the poor, has been been that poverty should be measured on a relative scale. In absolute terms, many of the poor in North America may have considerable income when compared with the poor in third world countries. But there are many poor people in Canada and the United States, when one compares these lowest income people in our society with the typical or socially determined standards that exist in North America.

Depending on which view one takes, different pictures of the variation of incomes among families in Canada can be presented. Table 5.30 presents measures of variation of family income in Canada for the years 1973, 1984, 1986 and 1987. The data have been corrected for price changes over these years by converting the data for each year into constant 1986 dollars.

An examination of Table 5.30 shows that, in absolute terms, the gap between better off and less well off families has become somewhat greater. The standard deviation of family income and the interquartile range for family income have each increased considerably over these years. There may be some increase in relative inequality because the CRV does increase by about 12% between 1973 and 1986, although it declines again slightly in 1987. However, the relative disparity between rich and poor does not appear to have increased as dramatically as the absolute gap over the years shown here.

The middle 50 per cent of families, as measured by the interquartile range, are spread over a greater distance in the 1980s than they were in 1973. However, if one looks at the ratio of the third quartile (or 75th percentile) to the first quartile (25th percentile), then it appears that the relative gap is little different. In fact, this ratio declines between 1984 and 1987, so that it is not much above the level of 1973.

Which of the two approaches to take is a matter of judgment. The absolute gap between rich and poor is likely greater in 1984, 1986 and 1987 than in 1973. But in relative terms, it appears as if the poor were not much,

Measure	1973	1984	1986	1987
s	21,592	26,299	28,085	28,170
$\overline{X}$	34,980	38,722	40,371	41,788
CRV	61.9	67.9	69.6	67.4
P_{75}	45,082	50,003	52,039	53,544
P_{25}	20,490	20,759	22,167	23,107
IQR	24,592	29,244	29,872	30,437
P_{75}/P_{25}	2.20	2.41	2.35	2.32

Table 5.30: Measures of Family Income Variation, Canada

if any, poorer, relative to the better off, than they were in 1973. On the other hand, it is fairly clear that the poorer families, while not relatively all that much worse off, certainly were not able to close the gap between rich and poor in the country. It should also be remembered that this is only one set of data and one specific set of measures and more detailed study is certainly warranted on the basis that these measures give conflicting evidence concerning what happened over this period of time.

Notes on Data in this Example. All of the percentiles, IQR, s and $\bar{X}$ are measured in 1986 dollars. The data for this example comes from Statistics Canada's Survey of Consumer Finances. The data is obtained from the economic families data tapes for 1973, 1984, 1986 and 1987. All data refers to what Statistics Canada defines as economic families.

Example 5.11.5 Number of Children per Family

In the 1950s the birth rate in Canada rose considerably, producing more children per family. From the early 1960s through to the 1980s, the birth rate has fallen, producing fewer children per family. The summary data showing the mean number of children per family is contained in Table 5.31. This data is based on tables from the Census of Canada for the years shown. It might be noted that the birth rate is not the only factor that has produced these changes in the number of children per family. Changes in the mortality rate of children and, more importantly, changes in the age at which children leave

the family residence and become independent of the family, both influence the number of children per family.

Table 5.31 shows that in years when the average number of children per family was larger, the standard deviation of the number of children per family was usually greater. In years when the mean is lower, the standard deviation is usually lower. This may reflect the fact that when there are few children per family, say 0, 1 or 2, there is little room for absolute variation in the number of children per family. In contrast, when there are more children per family, say 2 through 5 or so, there is greater room for absolute variation in the number of children per family.

Looking at the absolute variation in number of children per family, as measured by the standard deviation, it appears as if the variation of the number of children per family fluctuates considerably. In particular, the increase in the birth rate and the increased number of children per family in the 1950s is accompanied by an increase in variation from 1951 to 1961.

The coefficient of relative variation presents a somewhat different picture, with a continued decline in the CRV over the whole period, with the exception of 1966. It would appear, based on the CRV, that the general trend toward reduced variation in fertility continued even during the baby boom period of the 1950s and early 1960s. This lends support to the following analysis by A. Romaniuc in **Fertility in Canada: From Baby Boom to Baby Bust**, *Statistics Canada, Catalogue 91-524E, 1984, page 14:*

> In the 1930s there was a polarization of couples into two groups, those with relatively large numbers of children, and those with only one or no children. ... Today there has been an overall adjustment toward significantly lower childbearing targets. ... The regional variations in the birth rate have narrowed significantly in comparison to the situation before World War II. ... a greater homogeneity [in fertility] is expected throughout Canada in the years to come ...

Either the standard deviation or the coefficient of relative variation to get an idea of the amount of variation in the data. In some cases, the two measures present the same picture, in other cases, a somewhat different view emerges. If the latter is true, it is necessary to decide which of the two measures gives the best idea of the degree of variation in the data. If only the absolute difference among values matter, then the standard deviation is most appropriate. If the view is taken that different values are best measured

Year	Mean	s	CRV
1941	1.86	2.09	112.4
1951	1.69	1.87	110.7
1961	1.89	1.94	102.6
1966	1.91	1.88	98.4
1971	1.75	1.75	100.0
1981	1.37	1.28	93.4
1986	1.27	1.17	92.1

Table 5.31: Number of Children per Family, Canada, Selected Years

in relative terms, say relative to a typical value such as the mean, then the CRV is most appropriate.

5.12 Statistics and Parameters

In this chapter, various measures of the central tendency and variation of a distribution have been presented. These have usually been referred to as summary measures of the distributions being examined. When discussing these summary measures, no distinction was made between samples and populations. The summary measures can be used as a means of describing whole populations, or they can be used to provide summary descriptions of samples. For the most part, the definition and use of these measures is the same, regardless of whether the measures summarize a population or sample distribution. However, there are a few definitional differences in the case of the mean and the standard deviation.

The manner in which these summary measures are used also depends to some extent on whether they describe samples or whole populations. In Chapters 7 and following, summary measures based on samples are used to derive inferences concerning the comparable summary measures for whole populations. For example, the mean of a sample will be used to make some statements concerning the likely values of the mean of the whole population.

In order to deal with these differences, statisticians make a distinction between statistics and parameters. Statistics refer to characteristics of samples, and parameters refer to characteristics of whole populations.

Definition 5.12.1 A **statistic** is a summary measure of a sample distribution, that is, a summary measure which is used to describe the distribution of data from a sample.

If data has been obtained from a sample, measures such as the mean, the range, the median or the coefficient of relative variation are all statistics which can be used to describe the distribution of the sample. The formulae for the standard deviation and the variance in Section 5.7 are the proper formulae for these measures for sample data.

Definition 5.12.2 A **parameter** is a summary measure of a population distribution.

The true mean income for all Canadian families, the range or interquartile range of grades for all students at a University, or the variance of heights of all Manitoba children of age 3, are all examples of parameters. These are the same summary measures as have been described in this Chapter, except that the data on which they are based is the set of all data from the whole population.

Note that when statistics were defined, nothing was said concerning how good or how bad the sample is. While researchers always hope to have a representative sample, sometimes the sample is not so representative. One of the questions which emerges is how representative the statistics from samples are of the population as a whole. One way in which a sample could be defined as being **representative** is if statistics from the sample are very close in value to the corresponding parameters from the population.

Sometimes parameters are defined as summary measures which describe theoretical or mathematical distributions, such as the binomial or normal distribution. These will be considered in Chapter 6. For this reason, summary measures which describe the characteristics of whole populations are sometimes referred to as **population values** or **true values**, or even true population values.

Since most of Statistics uses the summary measures of central tendency and variation, the distinction between statistics and parameters is an important one. Inferential Statistics is concerned with estimating or making hypotheses about parameters or population values. Statistics obtained from samples of the population are used to make these estimates or test these hypotheses. In order to keep these two types of summary measures distinct, the following notation is used.

Summary Measure	Statistic	Parameter
Mean	$\bar{X}$	μ
Standard Deviation	s	σ
Variance	s^2	σ^2
Proportion	$\hat{p}$	p
Number of Cases	n	N

Table 5.32: Notation for Statistics and Parameters

Notation for Statistics and Parameters. In order to distinguish statistics and parameters, different symbols are used for the two. In general in statistical work, statistics are given our ordinary Roman letters, such a $\bar{X}$ or s, while letters in the Greek alphabet are often used to refer to parameters. The symbols generally used for the common summary measures are given in Table 5.32. The symbols such as $\bar{X}$, s and s^2 are used to denote the mean, standard deviaton and variance, respectively, of sample data. These were defined earlier in the chapter. The sample size for a sample is usually referred to as n, and the population size is N. The symbols p and $\hat{p}$ are discussed later.

The mean of a whole population is usually given the symbol μ. This is the Greek letter *mu*, pronounced "mew." When referring to the mean of a theoretical distribution, in Chapter 6, this same symbol μ will be used to denote the mean of this distribution. For example, the mean of the normal curve will be given the symbol μ.

As a parameter, the standard deviation is given the symbol σ, the lowercase Greek *sigma*. Remember that the summation sign $\sum$ can also be named sigma; $\sum$ is the uppercase Greek sigma. Since the variance is the square of the standard deviation, the variance for a whole population is given the symbol sigma squared, σ^2. These may be used to refer to the standard deviation and variance of mathematical distributions. In Chapter 6, the normal curve has standard deviation σ.

Proportions have not been discussed so far in this text, except as proportional distributions. However, proportions as summary measures are very useful in statistical work. A proportion is the fraction of cases with a particular characteristic. For example, in studies of aging, researchers are

likely to be interested in the fraction of people who are over age 65, or the proportion of these who are over age 80. Observers of the political scene will be interested in the proportion of voters who support each Canadian political party. The symbol most commonly used to describe the proportion of a population is p. That is, p is generally considered as the parameter. Since this is already a Roman letter, the manner in which the corresponding characteristic of the sample is described is to place the symbol ˆ on top of the p, to produce the symbol $\hat{p}$. Statisticians usually refer to the symbol $\hat{p}$ as *p hat*.

The convention of placing a hat, ˆ, over another symbol is a common way of distinguishing statistics and parameters in statistics. For example, the mean of a sample could be referred to as $\hat{\mu}$, or the standard deviation of a sample could be $\hat{\sigma}$. While this is not commonly done with μ and σ, it is a common practice with other measures. In statistical work generally, if any algebraic symbol has ˆ on it, this undoubtedly means that the symbol refers to the characteristic of a sample, and is a statistic.

Computation of μ, σ and σ^2. μ is computed in exactly the same manner as is $\bar{X}$. That is μ is the sum of all the values of the variable, where these values are summed across all members of the population, and then this total is divided by the number of members of the population. If X is the variable, and there are N members of a population, then

$$\mu = \frac{\sum X}{N} \quad \text{or} \quad \mu = \frac{\sum fX}{N}$$

depending on whether the data is ungrouped or has been grouped into categories where the values of X occur with frequencies f.

The standard deviation and variance are defined a little differently in the case of sample and the population. When the variance was first introduced, it was presented as the sum of the squares of the deviations about the mean, divided by the sample size minus one. That is, if the sample size is n, the denominator of the expression for the variance is the value $n - 1$. In contrast, the variance for a whole population has the population size in the denominator. If the data is ungrouped, and there are N members of the population, then the variance of the population is defined as

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

and in the case of grouped data, where each value of X_i occurs with frequency f_i, and $\sum f_i = N$,

$$\sigma^2 = \frac{\sum f_i(X_i - \mu)^2}{N}$$

The standard deviation of a population is the square root of the population variance, so that there would be an N, rather than $n-1$ in the denominator under the square root sign.

The standard deviation and variance have different formulae in the case of a sample for mathematical reasons. By using $n-1$ in the denominator, s^2 as an estimate of the true variance, σ^2, is better than if n is used in the denominator. This will be discussed briefly near the end of Chapter 6.

Using a Calculator. Some calculators contain built in formula for the standard deviation or variance. The values of the variable are entered into the calculator, and then a simple pressing of the buttion gives the value of the standard deviation or variance. If you use a calculator in this manner, make sure you know whether the calculator has n or $n-1$ built into its formula. Most calculators use $n-1$, but some use n. Some calculators have a button for each. If the latter is the case, then you will generally use the button indicating $n-1$, since most data you will be working with are based on samples. Also note that these calculator methods usually work only for ungrouped data, where you can enter a list of the actual values of the variable. Few calculators have built in formulae for calculating the mean and standard deviation for grouped data.

5.13 Conclusion

This chapter presented various measures of central tendency and variation. While all of these measures are useful, the mean as a measure of the centre of a distribution, and the standard deviation as a measure of variation, are most important in the remainder of the text. The mean and standard deviation are by far the most commonly used measures, most of the inferential statistics in Chapters 7 and following is concerned with estimating or making hypotheses about the true mean of a population μ. In doing this, the sample mean $\bar{X}$ and the standard deviations of both sample and population, s and σ respectively, are also essential. Thus it is important to become familiar with these measures, both in terms of how to calculate them, and also how to interpret them.

This chapter completes the first section of the textbook, descriptive statistics. The following chapter is concerned with probability, some mathematical probability distributions, and with the behaviour of random samples from a population. All of these involve principles of probability. Chapter 6 may seem to be quite different from the discussion of distributions and the characteristics of these distributions, subjects which have occupied most of the text to this point. Near the end of Chapter 6, the principles of probability are used to consider the mean and standard deviation of probability distributions. In Chapter 7, these probability distributions, along with the descriptive statistics of these first few chapters, are both used to discuss inferential statistics, that is, estimation and hypothesis testing.