Social Studies 201
October 1-10, 2003

Measures of Variation

Introduction

Measures of variation are statistics that decsribe how varied are values of a variable or members of a sample or population. These measures are not as well understood or as commonly used as are measures of central tendency (mode, median, and mean). But measures of variation are just as important to understanding distributions.

Measures of variation are single numbers, or statistics, that summarize how varied or how dispersed are characteristics of members of a population. The measures of variation that we discuss in this class are as follows (for formal definitions of each, see later parts of these notes):

- **Range**. The range is the maximum value of a variable minus its minimum value.
- **Interquartile range**. The interquartile range is the value of the 75th percentile minus the value of the 25th percentile – the range for the "middle one-half" of a distribution.
- **Standard deviation**. The standard deviation is a type of average of how varied cases are from the mean of a distribution.
- **Variance**. The variance is the square of the standard deviation and is a type of average of the square of the difference of values from the mean of a distribution.
- **Coefficient of relative variation**. The coefficient of relative variation is the value of the standard deviation divided by the mean, an indication of how varied cases are relative to the average of a distribution.

Most of these measures of variation are not used in ordinary conversations or in newspapers and magazines – or they are much less commonly used than are the averages that measure the central tendency of a distribution. One

key to understanding the measures of variation is to use them to compare two distributions. A distribution with a measure of variation that has a large value is more varied or less conentrated than is a distribution with a smaller value of the same measure of variation.

Figures 1 and 2 illustrate this difference. Both are symmetrical distributions with the mean at the same value, in the centre of the distribution. But the distribution of Figure 2 is twice as spread out as the distribution of Figure 1. Most measures of variation for Figure 2 will be double the size of the corresponding measure in Figure 1.

Each of the words **variation**, **dispersion**, and **concentration** can be used to describe the relative variation or spread of the values in the distributions. With respect to variability, the relative situation for the two distributions is stated in Table 1.

Table 1: Variation of distributions of Figures 1 and 2

Figure 1	Figure 2
less varied	more varied
less dispersed	more dispersed
more concentrated	less concentrated

The measures of variation that can be used to describe the distributions are each discussed in the following notes.

1. Range.

The first measure of variation is the range, the set of values over which the variable ranges.

Definition. The range is the maximum or largest value of the variable minus the minimum or smallest value of the variable. Alternatively, the range can be defined as simply a listing of the smallest and largest values.

Example. In the stem and leaf display examined earlier in the semester, the lowest income was $3,563 and the largest income was $110,444. As a result, the range is

Figure 1: Less varied frequency distribution

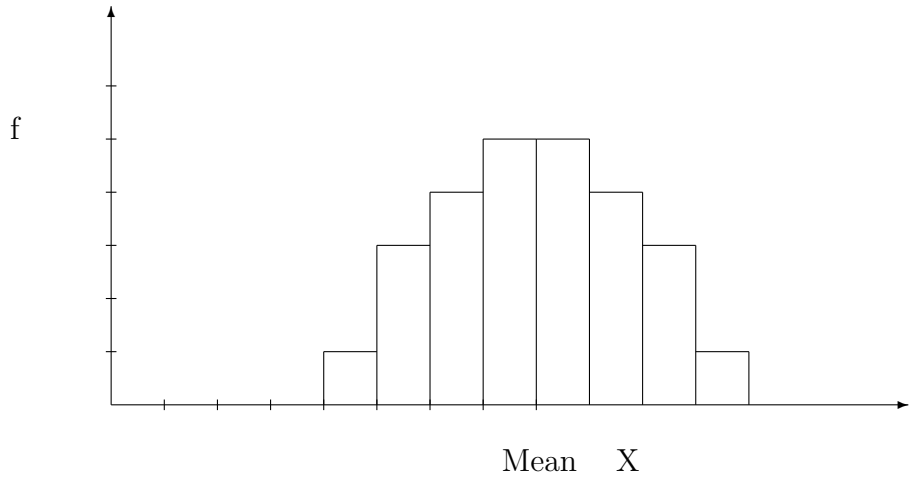
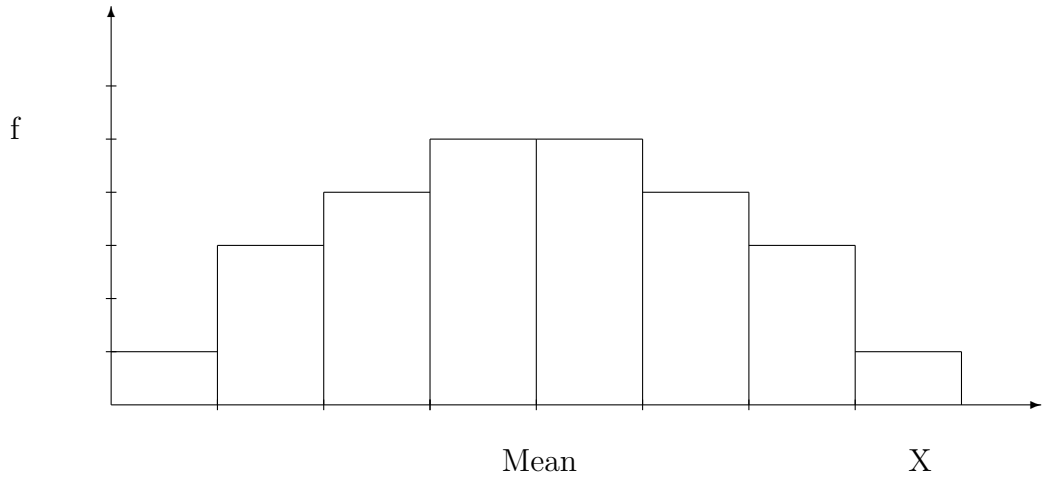


Figure 2: More varied frequency distribution



$$\text{Range} = 110,444 - 3,563 = 106,881$$

or, rounded off to the nearest thousand dollars, $107 thousand dollars. Alternatively, the range could be a listing of the smallest and largest values, that is, the range is from four to one hundred and ten thousand dollars.

Example. For many of the opinion or attitude questions in the data set SSAE98, opinions are measured on a scale from strongly disagree, or 1, to strongly agree, or 5. The range of opinions is $5 - 1 = 4$ or, alternatively stated in words, from strongly disagree to strongly agree.

The range is a simple measure that says nothing about how the values of the variable are spread between the maximum and minimum value. But the range provides a useful first indication of how spread out or concentrated are the values being examined. Imagine two rooms with a number of people in them – in room A the ages of the people have a range from 18 to 24, a range of 6 years; in room B the ages of the people have a range from 15 to 55, a range of 40 years. In these two rooms, it is clear that the people in room A have a fairly similar set of ages (more concentrated or less dispersed), while in room B people are more varied in their ages (less concentrated or more dispersed). As a result, one of the first questions a researcher may ask about a variable in a data set is what is the range of the variable. This does not tell the researcher a great deal about the values of the variable, but provides a useful first indication of the spread or variation of the values.

2. Interquartile Range (IQR).

Definition. The interquartile range, or IQR, is the seventy-fifth percentile minus the twenty-fifth percentile. Alternatively it is the third quartile minus the first quartile, since the third quartile is defined as the seventy-fifth percentile and the first quartile is the twenty-fifth percentile. In symbols,

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

The advantage of using the interquartile range is that it provides an indication of how spread out or concentrated the middle one-half of the distribution is. The 25th percentile, or first quartile, is the value of the variable

such that only 25 per cent or one-quarter or the cases are less than this value. Similarly, the 75th percentile, or third quartile, is the value of the variable such that 75 per cent of the cases are less than this, or only 25 per cent or one-quarter are greater. The difference between these values thus gives a good indication of the range of the values for the fifty per cent of the cases in the middle of the distribution. The IQR eliminates the effect of extreme values – for example, there may be only one very large value, producing a very large range. But the IQR considers only the values from the first to the third quartile.

Example. See the answer to problem 4 of Problem Set 2. In this example, the IQR for Blishen scores in Saskatchewan was

$$P_{75} - P_{25} = 51.0 - 28.3 = 22.7$$

and for Ontario was

$$P_{75} - P_{25} = 51.0 - 28.3 = 22.7$$

so the two provinces have similar variation in Blishen scores. While Blishen scores for the two provinces were generally greater for Ontario than Saskatchewan, as indicated by the mean score of 44.3 for Ontario and 37.6 for Saskatchewan, the IQR was very similar for the two provinces. The Ontario distribution can thus be considered to be like the Saskatchewan distribution, just moved to the right about five points. That is, the two distributions are similar, with similar variation, but the Ontario distribution lies about five Blishen points to the right of the Saskatchewan distribution.

3. Standard Deviation and 4. Variance

Unlike the previous two measures of variation, there is no easily understood intuitive explanation for the standard deviation and variance. The idea behind the standard deviation and variance is to examine differences of the values of the variable from the mean and combine them in a particular way. A data set with diverse values of a variable will have large differences of each value from the mean. In contrast, a data set with less diverse values will have smaller differences of each value from the mean.

There are a number of formulae for the standard deviation and variance, depending on how the data are organized. The first formula is for ungrouped

data – a list of values of a variable. Later in these notes there are different formulae for data which are organized or grouped into categories or intervals, or what are referred to as grouped data. Formulas, examples, and explanations for each of these follow.

Ungrouped data – pp. 227 and following.

Definition. A variable X that takes on values $X_1, X_2, X_3, ...X_n$ has mean

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

and the variance is

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

. The standard deviation is the square root of the variance or

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

The procedure for calculating this is

- First sum up all the values of the variable X, divide this by n and obtain the mean of $\bar{X} = \Sigma X/n$.
- Then each individual value of X is subtracted from the mean to obtain the differences about the mean. These are the values $X_1 - \bar{X}$, $X_2 - \bar{X}$, $X_2 - \bar{X}$, ... $X_n - \bar{X}$.
- Multiply each of these differences about the mean by themselves, that is, obtain the squares of each of these differences. This produces the values $(X_1 - \bar{X})^2$, $(X_2 - \bar{X})^2$, $(X_3 - \bar{X})^2$, ... $(X_n - \bar{X})^2$.
- Add all the squares of the differences about the mean to obtain

$$\Sigma(X - \bar{X})^2$$

.

- Divide this sum of the square of the differences about the mean by $n-1$ to obtain the variance s^2, that is,

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

  This is like the mean value of these squares of the differences about the mean, although $n-1$, rather than n, is used in the denominator.

- The standard deviation s is the square root of the variance of the last step, that is,

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

Units for the standard deviation s. Before proceeding to the example, note that the unit for the standard deviation is the same as the unit for the variable X. For example, if a group of five people have ages of 18, 24, 19, 25, and 27 years, the standard deviation will also be measured in years. In this case, the mean of these five ages is 22.6 and the standard deviation is 3.9 years (these can be checked using the formulae above). But the variance is the square of the units, in this case age squared. In this example, the variance is 15.3 years squared. Because "age squared" is an unfamiliar unit with which to work, it is generally preferable to work with the standard deviation – at least it is in a familiar unit. In the example of the five ages, the unit for s is age.

Roughly speaking, the standard deviation can be considered to be a sort of average deviation of the values of the variable from the mean. The exact form of the average is, of course, unusual in that it involves squaring and obtaining a square root. But it is a particular form of average, so a larger standard deviation represents values that differ more from the mean than in the case of a smaller standard deviation. This can be observed in the following example by examining the $X - \bar{X}$ column.

Tabular format. Rather than proceeding directly through the formula or the bulleted items above, most analysts consider it more efficient to obtain the standard deviation using a table or tabular format. In the examples that follow, a tabuler format is used.

Example – variation in homicides by province. Calculate the variance and standard deviation for each of the two groups of provinces in Table 2.

The numbers of homicides in the ten provinces of Canada are divided into two groups in Table 2 – provinces east of Manitoba and the four western provinces. A quick glance at the numbers in the two areas of Canada shows that the variation in number of homicides by province is much greater in the six eastern provinces than it is in the four western provinces. The range of homicides in the provinces east of Manitoba is from 1 to 170 or $170-1 = 169$. In contrast, the range of homicides in the four western provinces is from 27 to 85 or $85 - 27 = 58$. Just looking at the numbers, the four numbers for Western Canada are closer to each other than are the six numbers for provinces east of Manitoba. We thus expect to find a larger variance and standard deviation for the six eastern provinces than for the four western provinces. The calculations follow the table.

Table 2: Number of homicides in two areas of Canada, 2001

East of Manitoba		Western Provinces	
Province	Number	Province	Number
Nfld	1	Manitoba	34
PEI	2	Sask.	27
NS	9	Alberta	70
NB	8	B.C.	85
Quebec	140		
Ontario	170		

The calculations for determining the standard deviation of the number of homicides is contained in Table 3.

The tabular format is illustrated in Tabe 3. A column is provided for the X values of the variable, then a column for the differences of each X value from the mean, $X - \bar{X}$, and finally a column for the squares of these differences from the mean, $(X - \bar{X})^2$. From the entries into these columns and the sum of these columns, the variance and standard deviation can be relatively easily calculated.

The first step is to calculate the mean for each group. For the six provinces east of Manitoba in Table 3, the sum of the number of homicides is 330 (first

column). The mean number of homicides for these provinces is

$$\Sigma X/n = 330/6 = 55$$

.

The next step is to subtract the mean from each value of the variable. These values are given in the second column of Table 3. For Newfoundland, there was only1 homicide and this is $1 - 55 = -54$ below the mean number of homicides. For Quebec, there were 140 homicides, or $140 - 55 = 85$ homicides above the mean. The other values of the deviations about the mean are calculated in a similar manner.

Table 3: Calculations for Mean and Standard Deviation, Homicides in provinces east of Manitoba

	First Sample	
X	$X - \bar{X}$	$(X - \bar{X})^2$
1	-54	2,916
2	-53	2,809
9	-46	2,116
8	-47	2,209
140	85	7,225
170	115	13,225
330	0.0	30,500

The next step is take the differences about the mean, in the second column, and square each. This results in the squares of the differences from the mean, the $(X - \bar{X})^2$ values, of the third column of Table 3. For example, for Newfoundland, with 54 homicides less than the mean, the squared difference is $54 \times 54 = 2,916$. Quebec has 85 homicides above the mean and the squared difference is $85 \times 85 = 7,225$. Other entries in the third column are calculated in a similar manner. The entries in this column are then added – in this case they total 30,500.

From these calculations, the variance can now be calculated and it is

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1} = \frac{30,500}{6 - 1} = \frac{30,500}{5} = 6,100$$

. The standard deviation is the square root of the variance, that is,

$$s = \sqrt{s^2} = \sqrt{6,100} = 78.102$$

or 78.1 homicides.

The calculations for the four western provinces is contained in Tableřeftab-hsrw01. The procedure is the same, with the mean number of homicides for the four western provinces being

$$\Sigma X/n = 216/4 = 54$$

. Note that this is almost exactly equal to the mean for the eastern provinces.

Table 4: Calculations for Mean and Standard Deviation, Homicides in western Canada

First Sample		
X	$X - \bar{X}$	$(X - \bar{X})^2$
34	-20	400
27	-27	729
70	16	256
85	31	961
216	0.0	2,346

From the third column of the table, the variance is

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1} = \frac{2,346}{4 - 1} = \frac{2,346}{3} = 782$$

. The standard deviation is the square root of the variance, that is,

$$s = \sqrt{s^2} = \sqrt{782} = 27.964$$

or 28.0 homicides.

In Table 5, these results are summarized. From this summary, the much greater variability for the eastern, as compared with western, provinces is

Table 5: Summary statistics for homicides in two areas of Canada

Measure	Eastern Provinces	Western Provinces
$\bar{X}$	55	54
s^2	6,100	782
s	78.1	28.0
Range	169	58

apparent. The standard deviation and range are each about two and one-half times larger for the eastern than the western provinces. The mean number of homicides is very similar, but variation among the provinces is much greater in the east than the west, at least in terms of number of homicides. Note that the unit for the standard deviation and range are each the number of homicides, whereas the variance has the unit of homicides squared, a difficult unit to comprehend.

Alternative formula for ungrouped data – p. 231.

For ungrouped data, a list of numbers, the following formula may not appear to be computationally more efficient, but if there are a lot of values for the variable, it may be preferable. The variance is

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

and the standard deviation is the square root of this.

In terms of computing this in tabular format, there need only be two columns, an X column and a square of the Xs column, X^2. An example of this follows in Table 6

For the four western provinces,

$$\bar{X} = \frac{\Sigma X}{n} = \frac{216}{4} = 54.$$

and the variance is

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

Table 6: Alternative calculations for variance – 4 western provinces

X	X^2
34	1,156
27	729
70	4,900
85	7,225
216	14,010

$$
\begin{aligned}
&= \frac{1}{4-1}\left(14,010 - \frac{216^2}{4}\right) \\
&= \frac{1}{4-1}\left(14,010 - \frac{46,656^2}{4}\right) \\
&= \frac{14,010 - 11,664}{3} \\
&= \frac{2,346}{3} \\
&= 782
\end{aligned}
$$

and the standard deviation is

$$s = \sqrt{s^2} = \sqrt{782} = 27.964$$

or 28.0 homicides per province. As can be seen, this alternative formula yields the same results as the earlier formula.

Grouped data

When working with data that are grouped into categories or intervals, the variance and standard deviation again involve the sum of squares of differences about the mean. But these are now weighted by the number of times each such difference occurs. The definitions are as follows.

Definition. A variable X that takes on values $X_1, X_2, X_3, ...X_k$ with respective frequencies $f_1, f_2, f_3, ...f_k$ has mean

$$\bar{X} = \Sigma(fX)/n$$

where $n = \Sigma f$. The variance is

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

and the standard deviation is

$$s = \sqrt{s^2}.$$

That is, the squares of the deviations of the X values from the mean are multiplied, or weighted, by the frequencies of occurrence.

The above formula may be computationally inefficient, so in this class we generally use an alternative formula for the variance. This alternative gives exactly the same result and is

$$s^2 = \frac{1}{n - 1}\left(\Sigma f X^2 - \frac{(\Sigma f X)^2}{n}\right).$$

The standard deviation is the square root of the variance

$$s = \sqrt{s^2}.$$

The procedures for calculating the variance and standard deviation in tabular form are as follows:

- First create a table with the values of X in the first column and the frequencies of occurrence f in a second column.
- Create a third column fX with the products of the f and the X from the first two columns entered in this third column.
- Sum the products fX in the third column to obtain the column total ΣfX. Divide this sum by n, the sum of the frequencies in second column, to obtain the mean of $\bar{X} = \Sigma fX/n$.
- Create a fourth column with values of fX^2. That is, the square of the X values multiplied by the frequencies f are entered into the fourth column. But this is just the fX of the third column multiplied by another X (the value in the first column). That is, the fourth column is the entry in the third column multiplied by the entry in the first column. This produces the values of fX^2.

- Sum the values in the fourth column to obtain $\Sigma f X^2$.
- The sums of the fourth column and the third column are then entered into the formula

$$s^2 = \frac{1}{n-1}\left(\Sigma f X^2 - \frac{(\Sigma f X)^2}{n}\right)$$

  and this is the variance.
- The standard deviation s is the square root of the variance of the last step, that is,

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

  .

The following example illustrates the use of the alternative formula.

Example – Alcohol Consumption by Income Level

The **1991 General Social Survey - Cycle 6: Health**, conducted by Statistics Canada gives the data in Table 7. Use these data to compute the mean and standard deviation of the number of alcoholic drinks consumed per week for adults in each of the two income groups. Also compute the median number of drinks consumed per week for each distribution. Using these statistics and the data in Table 7, write a short note comparing the two distributions.

The table for the calculations of the mean and standard deviation of the number of alcoholic drinks consumed per week for Saskatchewan adults at each income level is Table 8.

For those of lower income, the mean is

$$\bar{X} = \frac{\Sigma f X}{n} = \frac{864}{302} = 2.861$$

or 2.9 drinks per week. The variance is

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left(\Sigma f X^2 - \frac{(\Sigma f X)^2}{n}\right) \\ &= \frac{1}{301}\left(12,696 - \frac{864^2}{302}\right) \end{aligned}$$

Table 7: Distribution of Saskatchewan Adults by Number of Alcoholic Drinks Consumed per Week and by Personal Income

Number of Drinks per Week	Number of Respondents with Personal Income in Dollars	
	Of Less than \$20,000	Of \$20,000 Plus
None	178	91
1-5	76	77
6-10	28	28
11-15	12	21
16-48	8	18
Total	302	235

$$\begin{aligned} &= \frac{1}{301}\left(12,696 - \frac{746,496}{302}\right) \\ &= \frac{12,696 - 2,471.841}{301} \\ &= \frac{10,224.159}{301} \\ &= 33.967 \end{aligned}$$

and the standard deviation is $s = \sqrt{s^2} = \sqrt{33.967} = 5.828$ or 5.8 drinks per week.

For the higher income group,the tabular format for the calculations is provided in Table 9. From this table, The mean is

$$\bar{X} = \frac{\Sigma fX}{n} = \frac{1,304}{235} = 5.549$$

or 5.5 drinks per week. The variance is

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left(\Sigma fX^2 - \frac{(\Sigma fX)^2}{n}\right) \\ &= \frac{1}{234}\left(24,466 - \frac{1,304^2}{235}\right) \end{aligned}$$

Table 8: Calculations for measures of variation of number of alcoholic drinks consumed per week – Income of less than $20,000

Number of Drinks per Week	X	f	fX	fX^2
None	0	178	0	0
1-5	3	76	228	684
6-10	8	28	224	1,792
11-15	13	12	156	2,028
16-48	32	8	256	8,192
Total		302	864	12,696

$$\begin{aligned} &= \frac{1}{234}\left(24,466 - \frac{1,700,416}{235}\right) \\ &= \frac{24,466 - 7,235.813}{234} \\ &= \frac{17,230.187}{234} \\ &= 73.633 \end{aligned}$$

and the standard deviation is $s = \sqrt{s^2} = \sqrt{73.633} = 8.581$ or 8.6 drinks per week.

A summary description of the differences in variation for these two groups is contained below, following the definition of the coefficient of relative variation.

Percentage distribution – p. 260.

When working with a percentage distribution, where percentages rather than frequencies are reported, there is a slight alteration in the denominator of the formula for the variance. Instead of using $n-1$, as in the above cases, the number 100 is used in the denominator. The definition is as follows.

Definition. A variable X that takes on values $X_1, X_2, X_3, ...X_k$ with respec-

Table 9: Calculations for measures of variation of number of alcoholic drinks consumed per week – Income of $20,000 plus

Number of Drinks per Week	X	f	fX	fX^2
None	0	91	0	0
1-5	3	77	231	693
6-10	8	28	224	1,792
11-15	13	21	273	3,549
16-48	32	18	576	18,432
Total		235	1,304	24,466

tive percentages $P_1, P_2, P_3, ...P_k$ has mean

$$\bar{X} = \Sigma(PX)/100$$

where $100 = \Sigma P$. The variance is

$$s^2 = \frac{\Sigma P(X - \bar{X})^2}{100}$$

and the standard deviation is

$$s = \sqrt{s^2}.$$

That is, the squares of the deviations of the X values from the mean are multiplied, or weighted, by the frequencies of occurrence.

The above formula may be computationally inefficient, so in this class we generally use an alternative formula for the variance. This alternative gives exactly the same result and is

$$s^2 = \frac{1}{100}\left(\Sigma PX^2 - \frac{(\Sigma PX)^2}{100}\right).$$

The standard deviation is the square root of the variance

$$s = \sqrt{s^2}.$$

The tabular format for calculating the variance and standard deviation is exactly the same as earlier, with P replacing f and 100, the sum of the percentages, replacing $n-1$. An example follows.

Example – Internet use, from first midterm examination.

From the first midterm, Table 10 contains the percentage distributions for infrequent and regular users.

Table 10: Distribution of time used internet for infrequent and regular users of the internet

No. of years used	Percentage who are Infreqent users	Regular users
<0.5	22	6
0.5-1	18	8
1-4	49	46
4-7	10	31
7-15	1	9
Total	100	100

The table for the calculations of the mean and standard deviation of the length of time used the internet for Saskatchewan adults who are infrequent users is Table 11.

For infrequent users, the mean is

$$\bar{X} = \frac{\Sigma PX}{n} = \frac{207.5}{100} = 2.075$$

or 2.1 years. The variance is

$$\begin{aligned} s^2 &= \frac{1}{100}\left(\Sigma PX^2 - \frac{(\Sigma PX)^2}{100}\right) \\ &= \frac{1}{100}\left(741.250 - \frac{207.5^2}{100}\right) \end{aligned}$$

Table 11: Calculations for measures of variation of number of years used internet – infrequent users

Number of years used	X	P	PX	PX^2
<0.5	0.25	22	5.5	1.375
0.5-1	0.75	18	13.5	10.125
1-4	2.5	49	122.5	306.250
4-7	5.5	10	55.0	302.500
7-15	11.0	1	11.0	121.000
Total		100	207.5	741.250

$$\begin{aligned} &= \frac{1}{100}\left(741.250 - \frac{43,056.25}{100}\right) \\ &= \frac{741.250 - 430.5625}{100} \\ &= \frac{310.6875}{100} \\ &= 3.106875 \end{aligned}$$

and the standard deviation is $s = \sqrt{s^2} = \sqrt{3.106875} = 1.763$ or 1.8 years.

The table and calculations of the mean and standard deviation of the length of time used the internet for Saskatchewan adults who are regular users is Table 12 and following.

For infrequent users, the mean is

$$\bar{X} = \frac{\Sigma PX}{n} = \frac{392.0}{100} = 3.920$$

or 2.1 years. The variance is

$$\begin{aligned} s^2 &= \frac{1}{100}\left(\Sigma PX^2 - \frac{(\Sigma PX)^2}{100}\right) \\ &= \frac{1}{100}\left(2,319.125 - \frac{392.0^2}{100}\right) \end{aligned}$$

Table 12: Calculations for measures of variation of number of years used internet – regular users

Number of years used	X	P	PX	PX^2
<0.5	0.25	6	1.5	0.375
0.5-1	0.75	8	6.0	4.500
1-4	2.5	46	115.0	287.500
4-7	5.5	31	170.5	937.750
7-15	11.0	9	99.0	1,089.000
Total		100	392.0	2,319.125

$$\begin{aligned} &= \frac{1}{100}\left(2,319.125 - \frac{153,664}{100}\right) \\ &= \frac{2,319.125 - 1,536.64}{100} \\ &= \frac{782.485}{100} \\ &= 7.825 \end{aligned}$$

and the standard deviation is $s = \sqrt{s^2} = \sqrt{7.825} = 2.797$ or 2.8 years.

5. Coefficient of Relative Variation – p. 262.

The measures of variation discussed so far are all in the units (or units squared) of the variable X. In order to obtain a measure of variation that is independent of the units of measurement, the coefficient of relative variation can be used. This latter measure can be used to compare variation of variables that are measured in different units. In addition, where the typical values of a distribution differ greatly in size, this may affect the size of the variance and standard deviation unduly, and the coefficient of relative variation provides a way of comparing distributions with quite different magnitudes for the variable.

Ceofficient of Relative Variation (CRV). The coefficient of relative vari-

ation is the standard deviation divided by the mean, and multiplied by 100. That is,

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

Example – Internet use. In the above example of internet use, for infrequent users the mean was 2.075 years and the standard deviation was 3.107 years. The CRV is

$$\text{CRV} = \frac{s}{\bar{X}} \times 100 = \frac{2.075}{1.753} \times 100 = 85.963.$$

For regular users, using the mean and standard deviation calculated earlier,

$$\text{CRV} = \frac{s}{\bar{X}} \times 100 = \frac{3.920}{2.797} \times 100 = 71.352.$$

The summary measures are shown in Table 13. From this it can be seen that the standard deviation and CRV give somewhat different pictures of the variability of the two distributions. In terms of years of use, there is no doubt that the regular users are more varied, with a standard deviation of 2.8 years of use, as compared with a standard deviation of only 1.8 years for infrequent users. But in terms of average lenght of use, infrequent users have used the internet for only about one-half as long (mean of 2.1 years) as regular users (mean of 3.9 years). Relative to these different means, there is only a small difference in relative variation of the distributions, at least using the CRV. In fact, in relative terms regular users have a slightly lower variation (CRV of 71.4) than infrequent users (CRV of 85.9).

Example – Alcohol consumption for different income levels

For the example of alcohol consumption for low and high income Saskatchewan residents, the CRVs are as follows. For the lower of the two income groups

$$\text{CRV} = \frac{s}{\bar{X}} \times 100 = \frac{5.828}{2.861} \times 100 = 2.037 \times 100 = 203.7$$

and for the upper income group

$$\text{CRV} = \frac{s}{\bar{X}} \times 100 = \frac{8.581}{5.549} \times 100 = 1.546 \times 100 = 154.6$$

Table 13: Summary Statistics for internet use

Measure	Infrequent users	Regular users
$\bar{X}$	2.1	3.9
s^2	3.1	7.8
s	1.8	2.8
CRV	85.9	71.4

Table 14: Summary Statistics for Alcoholic Drinks Consumed

Measure	Low Income	High Income
$\bar{X}$	2.9	5.5
s^2	34.0	73.6
s	5.8	8.6
CRV	203.7	154.6

Table 14 summarizes these values. The higher income adults consume considerably more alcohol than do the lower income adults, and the higher income are also more varied in their consumption patterns. Well over one-half of low income adults report no alcohol consumption while just over one-third of high income adults report no consumption. The mean consumption for the high income adults is about double that for the low income, and the median is approximately two drinks more per week. Since the low income adults are so heavily concentrated at the 0 level, the standard deviation for the low income group is relatively low. More of the higher income group tends to be spread out across all the intervals, producing a larger standard deviation. In relative terms though, it is the lower income group that is more varied – the variation in their alcohol consumption habits is relatively large when compared with the low mean consumption. In contrast, the higher income group is more varied in terms of their alcohol consumption but, since their mean consumption level is greater than those of lower incomes, in relative terms, their consumption is less varied.

Conclusion to CRV. Measures of relative variation are especially useful when comparing two distributions measured in different units. The example on p. 269 of the text illustrates how the differing value of the dollar each year, as a result of inflation, can provide misleading indication of variability of income, if no correction is made for the changing value. The CRV provides one means of obtaining this correction.

Other issues in measuring variation

- Make sure you read the section on interpretation, p. 249 and following. These pages give some guidelines about interpreting the standard deviation.
- The distinction between statistics and parameters, p. 273 and following, and the chart on p. 276, will be important for later parts of the course.

Notes last revised on October 15, 2003