

Social Studies 201

October 23, 2006

Standard deviation and variance for percentage distributions

See text, section 5.10, pp. 259-264.

When working with a percentage distribution, where the percentage of cases taking on each value of the variable is reported, there is a small change in the denominator of the formula for the variance. In the case of the frequency distributions examined so far, the denominator of the formula for the variance and standard deviation used the sample size minus one, $n - 1$. When working with a percentage distribution, the formula is parallel to that used so far, but with the number 100 used in the denominator. That is, it is not $100 - 1 = 99$ per cent in the denominator, but simply 100, representing one hundred per cent of cases. One is not subtracted from one hundred.

Definition. A variable X that takes on values $X_1, X_2, X_3, \dots, X_k$ with respective percentages $P_1, P_2, P_3, \dots, P_k$ has mean

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

where $\Sigma P = 100$. 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 percentages of occurrence. The sum of all percentages of occurrence is one hundred per cent.

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

Alternative formula. The variance of a percentage distribution is

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

The standard deviation is the square root of the variance

$$\text{Standard deviation} = s = \sqrt{s^2}.$$

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

Example – Internet use. Table 1 contains the percentage distributions for infrequent and regular users of the Internet. These data are adapted from p. 5 of Susan Crompton, Jonathon Ellison, and Kathryn Stevenson, “Better things to do or dealt out of the game? Internet dropouts and infrequent users,” from Statistics Canada, *Canadian Social Trends*, Summer 2002.

Table 1: Distribution of time used Internet for infrequent and regular users of the Internet

Number of years used	Percentage who are	
	Infrequent 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

Obtain the variance and standard deviation for infrequent and regular Internet users.

Answer. The variable is length of time used the Internet, measured in years. This is a ratio level measurement so the variance and standard deviation are meaningful measures.

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 2. This table is constructed with the X values, representing the midpoint of the intervals, and percentages P . Then two further columns, for the products PX and the products PX^2 complete the table.

Table 2: 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

Proceed through the table row by row. In the first row, $P = 22$ per cent and $X = 0.25$. The product of these two is $PX = 22 \times 0.25 = 5.5$ and this is the entry in the PX column. Then this PX is multiplied by another X , that is $5.5 \times 0.25 = 1.375$, and this is placed in the PX^2 column. Each of the other rows is completed in a similar manner.

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

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

$$\begin{aligned}
&= \frac{1}{100} \left[741.250 - \frac{207.5^2}{100} \right] \\
&= \frac{1}{100} \left[741.250 - \frac{43,056.25}{100} \right] \\
&= \frac{741.250 - 430.5625}{100} \\
&= \frac{310.6875}{100} \\
&= 3.106875
\end{aligned}$$

and the standard deviation is the square root of this. That 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 3 and following.

Table 3: 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

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

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

$$\begin{aligned}
&= \frac{1}{100} \left[2,319.125 - \frac{392.0^2}{100} \right] \\
&= \frac{1}{100} \left[2,319.125 - \frac{153,664}{100} \right] \\
&= \frac{2,319.125 - 1,536.64}{100} \\
&= \frac{782.485}{100} \\
&= 7.825
\end{aligned}$$

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

Summary. For infrequent users, the mean is 2.1 years and the standard deviation is 1.8 years. For regular users, the mean is 3.9 years and the standard deviation is 2.8 years. It is no surprise that regular users have used the Internet longer than infrequent users. In addition, regular users are more varied in their use of the Internet than are infrequent users. This may be because infrequent users are more clustered at low values of time used the Internet, whereas regular users are more spread across the different intervals, representing a greater variety of length of time used.

Interpretation of the variance and standard deviation

See text, pp. 249-259.

Given that the variance and standard deviation are mathematical constructions that are difficult to calculate and do have an intuitive explanation, it is not always easy to interpret their meaning. Hopefully, the following notes and pp. 249-259 of the text will help you understand and interpret these measures.

- **Useful mathematical constructions.** The variance and standard deviation are statistical measures designed and constructed by statisticians. Since they are useful in statistical analysis, you should be familiar with these measures, even if they are difficult to understand.

- **Deviations about the mean.** One aspect of the variance and standard deviation that should make sense is that each measure is constructed from differences of individual values from the mean value. Statisticians often refer to these as deviations about the mean. These deviations about the mean are the building blocks for the standard deviation and variance, and should make sense in that they indicate the extent of variation around the mean, or centre, of a distribution.
 - If cases are mostly clustered close to the mean, the deviations about the mean are small, producing a small variance and standard deviation. This indicates a small variability for the distribution.
 - When cases are spread out, with many distant from the mean, the deviations about the mean are large, producing a large variance and standard deviation. This indicates a greater variability for the distribution.
- **Unit.** The variance of a variable X is especially difficult to interpret since it is measured in a unit that is the square of the original units used to measure X . That is, the deviations about the mean are squared and the variance is an “average” of these squared deviations. The variance is extensively used in disciplines such as psychology and agriculture, where experiments are common. For the most part though, the variance is not used in this course, except as a first step in determining the standard deviation. The variance is used in advanced statistical analysis, so if you plan to continue to another course in statistics, plan to become familiar with the variance.

A primary advantage of the standard deviation over the variance is that the standard deviation of a variable X is measured in the same unit as is X . For example, if X represents income in dollars, the standard deviation is an amount of income in dollars. If X represents height of individuals in centimetres, s is measured in centimetres. This means that one aspect of the standard deviation is not too difficult to interpret – it has a familiar unit of measurement.

- **Range divided by four.** One very rough rule of thumb that can be used to obtain an estimate of a standard deviation is

$$\text{Standard deviation} = s \approx \frac{\text{Range}}{4}$$

where \approx means “approximately equal to.” This is not a very precise means of obtaining the standard deviation, but provides a quick and rough means of obtaining the general order of magnitude of a standard deviation.

For example, if the incomes of a group of individuals range from \$20,000 to \$60,000, the range is \$30,000 and the standard deviation is approximately

$$s \approx \text{Range}/4 = 40,000/4 = 10,000.$$

The standard deviation for this group is likely to be near this, say \$15,000 or \$8,500. The standard deviation is extremely unlikely to be, say, \$1,000 or less, or, in the other direction, as high as \$50,000.

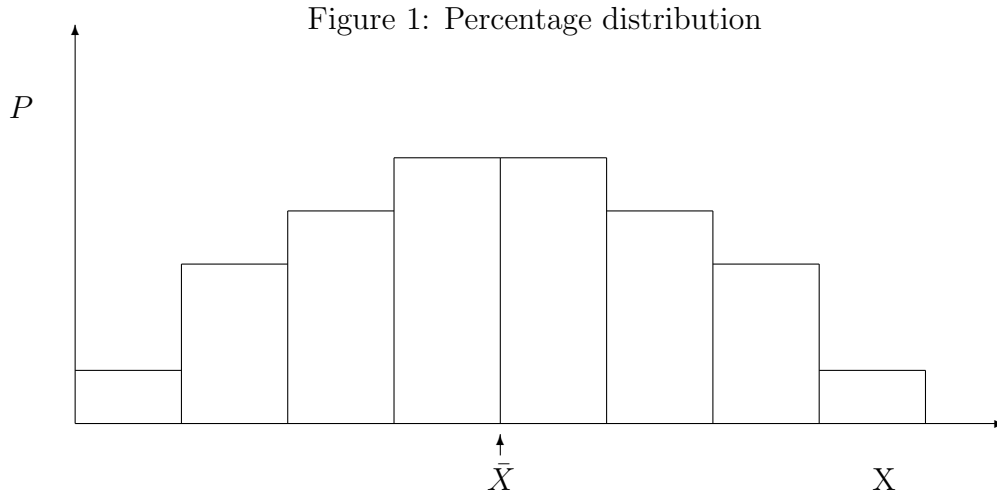
This rule of thumb provides only a very rough check on possible values of the standard deviation. The following discussion concerning percentage of cases within one, two, or three standard deviations of the mean provides a better means of understanding the standard deviation.

Percentage of cases around the mean

See text, pp. 254-259.

One way to understand and interpret the standard deviation is to consider the percentage of cases within one, two, or three standard deviations of the mean. For a distribution with mean \bar{X} and standard deviation s , the following rules of thumb generally hold.

- The interval $(\bar{X} - s, \bar{X} + s)$ usually contains approximately two-thirds, or 67%, of all cases.
- The interval $(\bar{X} - 2s, \bar{X} + 2s)$ usually contains approximately 95% of all cases.
- The interval $(\bar{X} - 3s, \bar{X} + 3s)$ usually contains approximately 99% of all cases.



Beginning with the histogram of Figure 1, the mean is at the centre of the distribution. There are one hundred per cent of cases in the distribution, within the boxes of Figure 1. How large would the standard deviation be for this distribution?

The rules of thumb, Figures 2 and 3, and the discussion that follows, should be helpful in understanding and determining an approximate size for the standard deviation.

In Figure 2, an area equivalent to two-thirds of the total area in the distribution is highlighted with grid lines. From the mean, at the centre of this symmetrical distribution, to obtain two-thirds of the distribution, it is necessary to go a certain distance to the left of centre and the same distance to the right of centre. From the rule of thumb for one standard deviation, the distance to the left and right of centre is approximately one standard deviation. The size of the standard deviation is illustrated on the upper right of the figure, just above the distribution. This is a distance, which when subtracted from \bar{X} and added to \bar{X} gives approximately two-thirds of the area in the distribution. That is, the distance from $\bar{X} - s$ to $\bar{X} + s$ is associated with approximately two-thirds of the cases in this distribution.

By proceeding to two standard deviations on either side of the mean, approximately ninety-five per cent of the distribution is included. This is illustrated in Figure 3.

Figure 2: Diagrammatic illustration of two-thirds of cases within one standard deviation of the mean

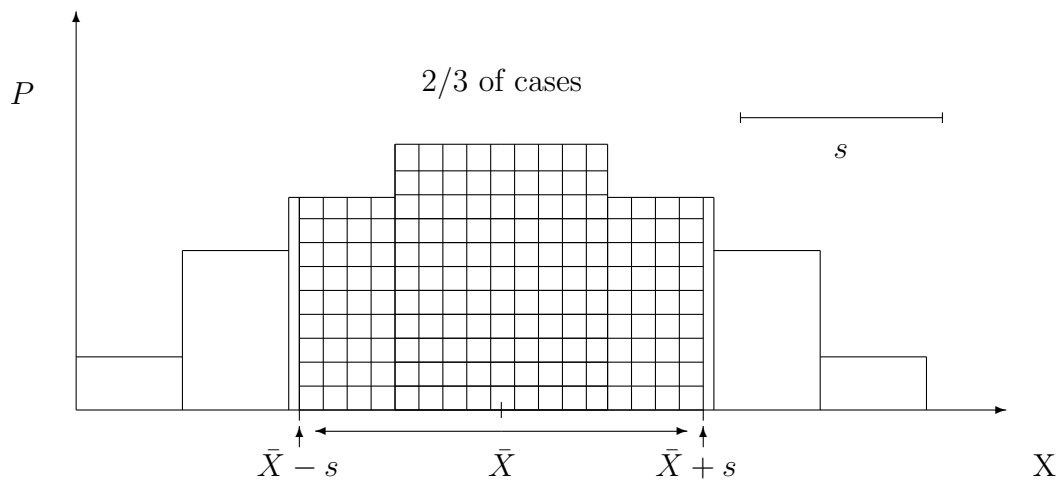
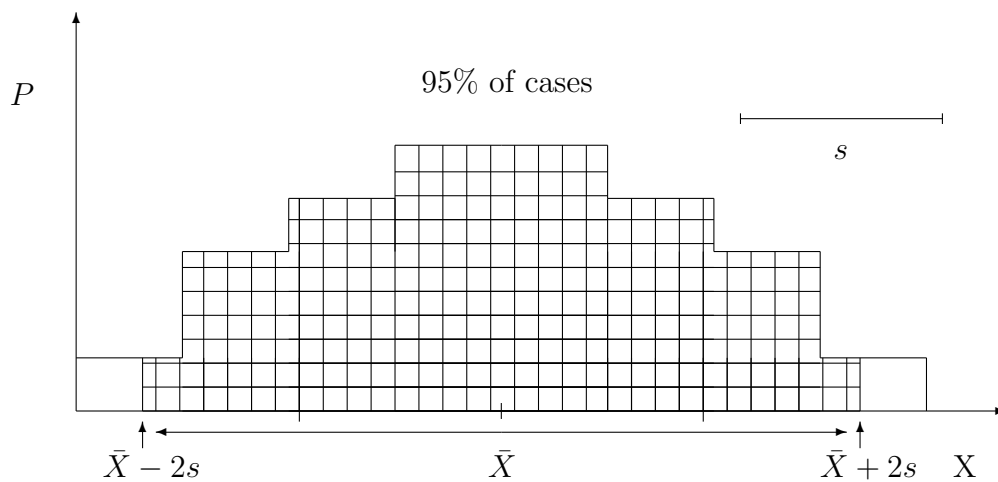


Figure 3: Diagrammatic illustration of 95% of cases within two standard deviations of the mean



These are very rough rules of thumb, and should not be relied on to obtain exact values of the standard deviation. But the rules should provide a way of understanding the standard deviation. Beginning from a distribution, if you consider the middle two-thirds, this represents approximately one standard deviation on each side of the mean. Alternatively, starting with a standard deviation, if you are given the value of s and \bar{X} , then you have a good idea of the range of the middle two-thirds of the distribution, that is from $\bar{X} - s$ to $\bar{X} + s$. In this case, if you construct the interval within two standard deviations of the mean, from $\bar{X} - 2s$ to $\bar{X} + 2s$, approximately ninety-five per cent of cases will be within this interval.

While not illustrated here, if you proceed from $\bar{X} - 3s$ to $\bar{X} + 3s$, three standard deviations on either side of the mean, you will often account for ninety-nine per cent or more of all cases. Cases more than three standard deviations are unusual. For example, for someone to have a height more than three standard deviations above the mean height would be unusual – professional basketball players might have this characteristic, but less than one per cent of members of a population would be more than three standard deviations either above or below the mean.

Coefficient of Relative Variation (CRV)

See text, section 5.11, pp. 262-273.

The measures of variation discussed so far are all in the units (or units squared) of the variable X . In contrast, the coefficient of relative variation (CRV) is a measure that has no unit of measure, so is independent of the unit used to measure a specific variable. The CRV can thus be used to compare variation of variables that are measured in different units. In addition, when typical values of a distribution differ greatly in size, this may affect the size of the variance and standard deviation unduly. The coefficient of relative variation provides a way of comparing distributions with different magnitudes for the variable.

Definition. The coefficient of relative variation (CRV) is the standard deviation divided by the mean, and multiplied by 100. That is,

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

The CRV is easy to calculate, at least once the mean and standard deviation are available. The CRV is simply the standard deviation divided by the mean, and multiplied by 100. In some statistical analysis this ratio of s to \bar{X} is the CRV; in this course the ratio is multiplied by 100.

Example – Hypothetical data. Consider a hypothetical group of children and adults. Suppose the heights of a group of children aged 8 years have a mean of 120 cm. and a standard deviation of 4 cm. Also suppose there is a group of adults aged 30 years with mean height of 170 cm. and a standard deviation of 6 cm. From the respective standard deviations, adults are more varied in height than are the children. But the inherent variability of heights of the two groups is likely to be similar. That is, some children are short and some are tall; some adults are short and some are tall. Children grow into adults, so the lower variability for children may be mostly because children are short and the numbers expressing differences in height are smaller than among adults.

A correction for this problem is to obtain the coefficient of relative variation. For children, the CRV is $(4/120) \times 100 = 3.3$. For adults, the CRV is $(6/170) \times 100 = 3.5$. These two CRVs are very similar, demonstrating that the different variability for the two groups occurs primarily because the heights of children are small numbers and the heights of adults are larger numbers. Once each standard deviation is measured relative to its mean, the relative variability for the two groups is very similar.

This example is hypothetical, not using actual data. Hopefully the example illustrates a useful application for the CRV.

Example – Internet use. For the example of Internet use, the summary statistics for infrequent users gave a mean of 2.075 years and a standard deviation of 3.107 years. The CRV is

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

for infrequent users.

The CRV for regular users, using the mean and standard deviation calculated earlier, is

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

Table 4: 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

The summary measures are shown in Table 4. From this it can be seen that the standard deviation and CRV give different pictures of the variability for 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 length of time used the Internet, 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, and using the CRV, there is only a small difference in relative variation of the distributions. In fact, regular users have a slightly lower relative 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, summary statistics, including CRVs are contained in Table 5.

NOTE – The statistics here come from a different example than that of the October 18 notes.

Table 5: 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

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

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

On average, higher income adults consume considerably more alcohol than do lower income adults, and those with higher income are also more varied in their consumption patterns than are those with lower income. 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. Compared with those of lower incomes, 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.

Conclusion to Part I of the course

This completes the first part of the course and the first part of the text, chapters 1-5. Before proceeding to Part II, read the concluding pages of Part I of the text, pp. 273-278, and the following notes.

Recap – measures of variation

Measures of variation are just as essential as are measures of centrality and position when describing samples, populations, and distributions. The notion of “average,” associated with terms such as “the average person” or “the majority” are embedded in our language. In contrast, measures of variation are less well understood and less commonly used in ordinary language and in much statistical reporting. Much of the presentation of statistics in the media concentrates on averages, and ignores variation. But each individual is different and there is diversity across members of a population – measures of variation are summary measures that describe this variation. While the measures we have examined in this module may seem distant from the concept that people differ, statisticians have found the measures examined here to be useful in describing and analyzing variety in populations.

To conclude the discussion of variation, a short summary of the measures, along with their advantages and disadvantages, is contained in the following bulleted items.

- **Range.** Maximum minus minimum value of a data set. Useful as a first indication of variation. Does not consider variation of cases between minimum and maximum values, so has limited use.
- **Interquartile range.** Seventy-fifth percentile minus twenty-fifth percentile. Often more useful than the range in that it indicates the range for the middle fifty per cent of cases, eliminating the extremes of a distribution. Its weakness is that it is a positional measure and does not consider the value for each individual case in a distribution.
- **Variance (s^2).** Average of the squares of the deviations of each individual value about the mean. The value of each case is taken into consideration in constructing this measure. Its main defect is that it is measured in an unfamiliar and difficult to interpret unit (square of the units of the variable).

- **Standard deviation** (s). Square root of the variance, or a sort of “average” or “standard” amount of variation of each case about the mean. This is the most widely used measure of variation in statistical analysis and provides an excellent means of summarizing the extent of dispersion or concentration of values. It is in the same unit as the variable, so is relatively easy to use. Its main defect is that it is difficult to calculate, can be difficult to interpret, and is a mathematical construction that has no intuitive explanation.
- **Coefficient of relative variation (CRV)**. This is the standard deviation divided by the mean, and multiplied by 100. This measure is useful in comparing distributions with different units or with quite different values of a variable. It is easy to calculate once the standard deviation and mean have been obtained. Like the standard deviation, it is a mathematical construction and can be difficult to understand and interpret.

All the above measures are useful in statistical analysis and provide a means of comparing the variability among data sets, samples, distributions, and populations. Different forms of analysis use different measures. Psychologists commonly use the variance and conduct what are termed analysis of variance. For other social sciences, and for this course, the standard deviation is the primary measure of variation used.

Statistics and Parameters

See text, section 5.12, pp. 273-277.

Statistics are summary measures of centrality or variation calculated from data obtained from a sample or experiment. That is, measures such as the median, mean, interquartile range, and standard deviation that are obtained from data are statistics describing the distribution of data from a sample or experiment.

In later sections of the course, we need a term to refer to the corresponding characteristics of whole populations and theoretical distributions. These are termed parameters, sometimes termed population values. Parameters are measures of centrality or variation obtained from a theoretical distribution or a whole population. That is, the mean of a theoretical distribution or a whole population is referred to as a parameter but the mean of cases in a sample (\bar{X}) is referred to as a statistic.

In order to refer to parameters or population values, we need a set of symbols. The symbols commonly used by statisticians, and used through the remainder of this course, are given in Table 6. Greek letters are used as symbols for parameters, so when you encounter a Greek letter as a symbol in statistics, this ordinarily means a characteristic of a theoretical distribution or a whole population.

Table 6: Notation for statistics and parameters

Statistical measure	Statistic	Parameter
Mean	\bar{X}	μ
Standard deviation	s	σ
Variance	s^2	σ^2
Number of cases	n	N

The Greek letter commonly used to denote the mean of a theoretical distribution or population is μ , a Greek equivalent of an English “m.” This is call “mu” and pronounced “mew.”

The Greek letter σ is used as the symbol to indicate the standard deviation as a parameter. This is labelled and pronounced “sigma” and cor-

responds to the English “s.” The variance of a theoretical distribution or whole population is referred to as σ^2 , and referred to in verbal usage as “sigma squared.” You may recall that the symbol Σ , used in summation notation, is also termed sigma. In order to avoid confusion, in this course the Greek symbol σ will be referred to as “sigma,” the symbol σ^2 will be termed “sigma squared,” and the symbol Σ will be referred to as the summation sign.

Finally, the number of cases in a sample or experiment will be referred to as n , that is, lower-case letter “n.” The upper-case letter “N” will be used to indicate the size of the whole population. That is, n is sample size and N is population size.

Last edited October 18, 2006.