**Social Studies 201**
**September 13-15, 2004**

## Presenting data and stem-and-leaf display

See text, chapter 4, pp. 87-160.

### Introduction

Statistical analysis primarily deals with issues where it is possible to construct numerical representations of the phenomena being studied. These numerical representations – incomes of individuals, stress levels of workers, grades of students – are termed quantitative data. When phenomena being studied do not have numbers or quantities attached to them – political party preference, sex, ethnicity – statisticians produce numbers by counting the number of people with different characteristics. This part of Module 2 and chapter 4 of the text explain and summarize ways of organizing these different types of quantitative data.

When presenting quantitative data, it is not common to merely list all the values of the variable. Rather, values of the variable are organized into tables, diagrams, graphs, charts, and summary statistics (see Modules 3 and 4 and text, chapter 5 for the latter) to make data readily understood. The aim of a statistical analyst is to present data in a form that properly portrays the characteristics of the sample or population being studied, clearly and accurately illustrating the issues or phenomena being investigated. There is usually not a single correct way of organizing values of a variable contained a particular data set, but there are a number of incorrect or misleading ways to organize the data. This part of Module 2 and chapter 4 of the text provide guidelines concerning useful ways of organizing quantitative data.

Statistical analysts use various techniques to organize and present quantitative data. The rules, procedures, and notation outlined in this part of Module 2 and chapter 4 of the text are generally consistent with the methods other statisticians use when analyzing data sets. These notes deal with the following issues:

- Ways of organizing data – grouping data.
- Conventions about notation.
- Frequency, proportional, and percentage distributions.

- Guidelines concerning tables for distributions.
- Miscellaneous issues – class limits, rounding, and interval width.
- Diagrammatic presentations – histograms.

Issues concerning production and definition of data were discussed in the Module 1 and the first part of Module 2 and are not pursued further here. In this part of Module 2, it will be assumed that the data have been well produced (Module 1 and chapter 2 of the text), and the population and variables clearly defined (first part of Module 2 and chapter 3 of the text). These notes and chapter 4 of the text discuss how to organize and present data obtained by the researcher so the data can be presented in proper statistical form and so the data can be readily examined and analyzed.

**Data sets**

When data are initially obtained from questionnaires, interviews, experiments, administrative sources, or other methods, a statistical analyst encounters a list or array of values of a variable or variables. These values represent characteristics of individuals or objects – the values of a variable or variables taken on by the individuals or objects. If the data are quantitative, these are lists of numbers; if the data are qualitative, these are lists of the words or phrases associated with each response.

In order to analyze such data, it is awkward and time consuming to always have to deal with the list of all possible values. In general, when using these data, the first step of a statistical analyst is to organize them into a form that makes it straightforward to examine and analyze the data set.

This section of the notes examines ways of organizing data. One basic distinction to consider when organizing data is the difference between ungrouped and grouped data. This distinction will be used throughout the course and is as follows (see text, pp. 94-95).

> **Ungrouped data**. A list of all the values of a variable in a data set is referred to as ungrouped data.

That is, ungrouped data comprise simply a list of all the values of one or more variables across the cases of a data set. Examples of ungrouped data are provided in Examples 2.7 and 2.8.

> **Grouped data**. These are data organized into categories or intervals, usually presented in a table or diagram.

Grouped data refer to tables, diagrams, or other presentations of data where values or sets of values are grouped together into categories or intervals.

**Stem-and-Leaf Display**

See text, section 4.5.3, pp. 106-114.

A stem-and-leaf display is an ordering of the values of a variable from the smallest to the largest values. Values are first ordered into broad groupings, such as units of 5 or 10 of the variable. Then, within each group, the individual values are ordered from the smallest to the largest values. The construction of a stem-and-leaf display is an efficient way of organizing a list of numbers, so they can be readily understood and analyzed.

The construction of a stem-and-leaf display provides a relatively quick way of organizing data into categories or intervals when there are larger numbers of cases than an analyst is willing to count or tally. The stem-and-leaf display is most useful with more than twenty cases, but less than one hundred and fifty or two hundred cases. If there are more than two hundred cases, it is time consuming to use this method and is probably best to enter the data into a computer program, using it to organize and analyze the data.

The stem-and-leaf display is useful when the values of the variable are larger than from 0 to 9. But when values of a variable exceed 10 and range, for example, from 5 to 75, the stem and leaf display is likely to be a useful way of organizing the data. For example for ages, hours worked, or grades of respondents, values that range from 0 to 80 or 90, the stem and leaf display is an ideal way to organize the values of the variable into an ordered and grouped list.

An advantage of the stem-and-leaf display, over a count or tally, is that all the original values of the data set continue to be recorded. In the display, none of the original information is lost. Looking ahead to to the example, the stem-and-leaf display of Table 5 displays all the incomes of a sample of Saskatchewan clerical workers originally provided in Table 2. The stem-and-leaf display places the incomes in order from the smallest to the largest income, whereas the original list is not ordered.

A stem-and leaf display is thus an efficient way of organizing some data sets. As long as there are not too many cases, the display provides a way of maintaining all the original information organizing all the values in numerical order, from the smallest to the largest values of the variable. From the stem-and-leaf display, a table or diagram of the data set can be easily constructed, and other forms of statistical analysis can be readily performed.

The best way to learn how to construct a stem-and-leaf display is to work through an example. The following example illustrates how to construct a stem-and-leaf display. for a set of sixty incomes of individuals. A summary list of procedures for constructing a stem-and-leaf display is contained in these notes, following the Example.

**Example 2.10 – Stem-and-leaf display of incomes**. The incomes (in dollars) of sixty Saskatchewan clerical workers are listed, in no particular order, in Table 1. This set of sixty values of the variable income is an example of the type of data that can be obtained from a survey of clerical workers. Because the numbers in the list are not ordered, it is difficult to grasp how these incomes are distributed.

The numbers in Table 1 represent incomes, for the year 2000, of a random sample of sixty Saskatchewan adults who reported their occupations as clerical occupations. These data come from Statistics Canada's Survey of Labour Income and Dynamics.

Table 1: List of income (in dollars) of sixty Saskatchewan clerical workers, 2000

| | | | | | |
|---|---|---|---|---|---|
| 21250 | 39100 | 48825 | 15600 | 9625 | 26200 |
| 34100 | 34925 | 27200 | 24325 | 11875 | 42550 |
| 32000 | 52825 | 16000 | 5250 | 36700 | 39000 |
| 19400 | 9250 | 24300 | 30425 | 33275 | 36325 |
| 46000 | 39025 | 28200 | 27025 | 19100 | 24600 |
| 19000 | 38000 | 27000 | 40050 | 45300 | 29475 |
| 9500 | 24550 | 25625 | 36975 | 35000 | 65850 |
| 10600 | 23350 | 40000 | 46075 | 30250 | 21000 |
| 15125 | 8500 | 31000 | 50100 | 23300 | 12475 |
| 44775 | 16850 | 34000 | 42000 | 16000 | 31000 |

**Source**: Statistics Canada. Survey of Labour Income Dynamics (SLID), 2000: Person file [machine readable data file]. Ottawa, Ontario: Statistics Canada. 7/16/2003.

Use the data in Table 1 to construct a stem-and-leaf display of incomes, organized into groups of ten thousand dollars. First round each value to the nearest thousand dollars, then construct an unordered and an ordered stem-and-leaf display. In words, briefly describe how incomes of these workers are distributed.

**Answer**. The first step is to round the individual incomes to the nearest thousand dollars. This is done in Table 2. The order of the incomes of Tables 1 and 2 are identical, so it should be easy to see how the data are rounded. For example, the first income in Table 1 is \$21,250. Rounded to the nearest thousand dollars, this is closer to \$21,000 than \$22,000, so \$21,250 is rounded to 21 thousand dollars and this is the first entry in Table 2.

Moving down the first column of Table 1, the second entry is \$34,100 and this it closer to 34 than to 35 thousand dollars, so is rounded to 34 in Table 2. Other incomes are rounded in similar fashion and entered into Table 1.

Note that the seventh income in the first column of Table **??** is \$9,500, exactly midway between \$9,000 and \$10,000. As a means of rounding such midpoint values, incomes that are exactly midway between the thousand dollar points are rounded to an even number. That is, \$9,500 is rounded to 10 thousand dollars. It does not matter whether you round to the even or odd number, but be consistent. Adopt a consistent procedure and stick with that across the data set (see bottom of p. 126 and top of p. 127 of the text for a discussion of this).

Table 2: List of income (in thousands of dollars) of sixty Saskatchewan clerical workers, 2000

| | | | | | |
|---|---|---|---|---|---|
| 21 | 39 | 49 | 16 | 10 | 26 |
| 34 | 35 | 27 | 24 | 12 | 43 |
| 32 | 53 | 16 | 5 | 37 | 39 |
| 19 | 9 | 24 | 30 | 33 | 36 |
| 46 | 39 | 28 | 27 | 19 | 25 |
| 19 | 38 | 27 | 40 | 45 | 29 |
| 10 | 25 | 26 | 37 | 35 | 66 |
| 11 | 23 | 40 | 46 | 30 | 21 |
| 15 | 8 | 31 | 50 | 23 | 12 |
| 45 | 17 | 34 | 42 | 16 | 31 |

Using the numbers in Table 2, the first step in constructing a stem-and-leaf display is to determine the minimum and maximum values of the variable. From Table 2, the minimum income is 5, the third entry in the fourth column. The maximum value of income is 66, the fourth last entry in the right column.

Given that the values range from 5 to 66, when organized into groups of ten, the stem-and-leaf display must provide space for entries from less than ten thousand dollars to those in the sixty thousand category. Since the data are to be organized into groups of ten thousand dollars, the categories are less than 10, 10 to 19, 20 to 29, 30 to 39, 40 to 49, 50 to 59, and 60 to 69 thousand dollars. The format for the stem-and-leaf display is tha of Table 3.

The stem of the stem-and-leaf display is the column to the left of the vertical line. In this display, the entries column on the left represents the tens unit of each value. That is, the first row of the display is labelled 0 and contains all the values less than 10. The second row is labelled 1 and contains all the tens (10, 11, 12, ... 19). The row labelled with 2 contains all the twenties (20, 21, 22, ... , 29). The final row contains all the sixties.

Table 3: Format of stem-and-leaf display

| Stem | Leaves |
|---|---|
| 0 | values of 0 to 9 |
| 1 | values of 10 to 19 |
| 2 | values of 20 to 29 |
| 3 | values of 30 to 39 |
| 4 | values of 40 to 49 |
| 5 | values of 50 to 59 |
| 6 | values of 60 to 69 |

From the numbers in Table 2, what is termed the unordered stem-and-leaf display is constructed. This is presented in Table 4. It is unordered in the sense that the individual values to the right

of the line (the leaves) are not placed in order from low to high. But the broader groupings are ordered – all the values in the first row (incomes less than ten thousand) are less than the values in the second row (incomes from ten to nineteen thousand).

Table 4: Unordered stem-and-leaf display for incomes of clerical workers

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

The entries in Table 4 are obtained from the list of values in Table 2, proceeding down the first column, then down the second column, and so on. Beginning with the first value of 21 in Table 2, this is placed in the 20s row of the stem-and-leaf display. As can be seen, the first value in the 20s row is a 1, and this represents the original value of 21.

The second income of Table 2 is 34, a thirty with a value of 4 as the leaf. The first entry in the 30s row of the stem-and-leaf display is thus a 4, representing 34.

The third income is 32, another 30, with a value of 2 for the leaf. The second entry in the 30s row of the table is a 2, representing this income of 32. Note that the third entry in the 30s row of the stem-and-leaf display is a 9, representing an income of 39. This comes from the first value in the second column of Table 2. That is, after 32 in the first column of the list of incomes, there are no more 30s until the first entry in the second column Table 2.

The fourth income in Table 2 is 19. This is placed in the 10s row of the stem-and-leaf display, as a 9, that is $19 = 10 + 9$. Note that the second entry in the 10s row of the stem-and-leaf display

is also a 9, representing another income of 19 – the sixth value down in the first column of Table 2.

Other entries are obtained in a similar manner, by proceeding through the table systematically until all sixty values of Table 2 have been place in the proper row of the stem-and-leaf display of Table 4.

Before proceeding to construct the ordered stem-and-leaf display, it is best to count the values to make sure you have included them all. There should be sixty entries in Table 4. There are 3 values in the 0s row, 13 in the 10s row, 15 in the 20s row, 17 in the 30s row, 9 in the 40s row, 2 in the 50s row, and 1 in the 60s row. This totals

$$3 + 13 + 15 + 17 + 9 + 2 + 1 = 60$$

so all the sixty values of the original list are included in the stem-and-leaf display.

The next step is to order the values within each row. This produces the ordered stem-and-leaf display. Beginning with the stem-and-leaf display of Table 4, the values in each row are placed in numerical order in Table 5. Remember to include each value, so all sixty values are included in the ordered stem-and-leaf display.

Table 5: Ordered stem-and-leaf display for incomes of clerical workers

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

The first row of Table 5 is obtained from the first row of Table 4, by taking the three values 9, 8, and 5 and putting them in nu-

merical order. In order, from small to large, these are 5, 8, and 9, and this is the order in which they appear in the ordered stem-and-leaf display. Other rows are obtained in the same manner. When ordering the values leaves, representing the values of the incomes, make sure you do not omit any cases. Counting the number of leaves in each row of the ordered display, making sure these match the counts in the unordered display, and that the total of all values listed matches the original number of values.

**Verbal description**. The main feature of this stem-and-leaf display is that incomes are concentrated between ten thousand and thirty-nine thousand dollars. That is, the great bulk of respondents have incomes within the 10-19, 20-29, and 30-39 thousand dollar categories, with few below ten thousand dollars. There are more respondents in the 30-39 category than in any other category, relatively few with incomes of forty thousand dollars or more, and none with income above sixty-six thousand dollars. A distribution of this type is typical of incomes – relatively few at the lowest income, a majority at low to middle incomes, and then fewer at each successively higher income category.

Note that the verbal description applies to both the unordered and ordered stem-and-leaf display. For later work with distributions of data, the ordered display is more useful than is the unordered display.

This example will be used again in later examples and exercises. These data on distribution of incomes will be used to construct a frequency and percentage distribution and a histogram. In Module 3, this example will be used to obtain measures of centrality.

**Summary of stem-and-leaf procedures**

Beginning with a list of values of a variable, the procedures for obtaining stem-and-leaf displays can be summarized as follows.

1. First make sure that it makes sense to use a stem-and-leaf display. For example, if the list of values includes only values 1-5 or 1-7, with many attitude variables, a count or tally is sufficient. In contrast, if values of a variable extend from 15 to 90, or to even larger values, and there are more than twenty or thirty, but less than two hundred, cases, then a stem-and-leaf display is likely to be useful.

2. Determine the minimum and maximum values in the list. This determines the range of values for the display.

3. Decide on the categories to be used. This is ordinarily units of 10, but it might be units of 5, or some other grouping of units.

4. Contruct a display using the categories identified in the previous note. As the stem, place the category label to the left of a vertical line, leaving space for the leaves to the right of the line.

5. Decide on a systematic way of proceeding through the list of values. That is proceed down the first column, then down the next column, and so on; alternatively proceed across the first row, then across the second row, and so on. It does not matter which procedure you adopt, but make sure you are adopt and use one of these procedures.

6. Using the procedure adopted in the previous note, organize the values from the original list of values, placing the leaf value to the right of the vertical line, in the appropriate row. This produces the unordered stem-and-leaf display.

7. For each row, order the values to the right of the vertical line. This produces the ordered stem-and-leaf display.

8. One final step is to count the number of cases in each row of the display. You should have the same total number of cases as the number of cases in the original list. If the total is not equal to the number of cases, you will have to retrace the above steps.

**Additional notes about stem-and-leaf displays**

1. **Categories other than tens**. Example 2.9 grouped incomes into categories of 10s, 20s, 30s, and so on. Sometimes the categories are subdivided into narrower categories, for example, 10-14, 15-19, 20-24, and 25-29. This is illustrated in Table 6, obtained by splitting some of the rows in Table 5 into two subgroups, from 0 to 4 and from 5 to 9.

Table 6: Ordered stem-and-leaf display for incomes of clerical workers

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

In Table 6, the first row remains the same as earlier, but the second row now represents only the individuals with incomes of 10 to 14. Then the third row, again given stem value 1, since these are still values in the tens, has leaves representing only the individuals from 15 to 19. The rows for the 20s, 30s, and 40s are constructed in the same manner. Since there are so few 50s and 60s, these are left unchanged.

2. **Categories other than 0-90**. Additional categories can be added to a stem-and-leaf display if a variable has values are greater than 100 or less than 0 or greater than 100. For example, rows such as the following could be added to a display.

| Stem | Leaves |
|---|---|
| -1 | values of -10 to -19 |
| -0 | values of 0 to -9 |
| . | |
| . | |
| 10 | values of 100 to 109 |
| 11 | values of 110 to 119 |

3. **All values**. A statistical analyst modifies the groupings to suit the type of data and the needs associated with examining and using the data. Any stem-and-leaf display must include all the original values, with these values placed in order from the minimum to the maximum value. Once you have constructed the display, count the values in the display to make sure you have included all the values and check the order to ensure values are in order.

4. **Interpreting the display**. A quick glance at a stem-and-leaf display gives an overview of the distribution. In Example 2.9, a quick glance demonstrates that there are no cases above the 60 thousand income level, with most cases between ten and fifty thousand dollars. This information would have been difficult to determine using the original list of incomes.

   From the display, a frequency or percentage distribution table or diagram can be readily constructed. Procedures for constructing these are discussed in later sections of this module.

   The values in a stem-and-leaf display are ordered from the minimum to the maximum value. Some of the statistics introduced in Modules 3 and 4 can be readily obtained from the stem-and-leaf display. In particular, the mode (most common value) and median (middle value) of Module 3 can be quickly determined from the display.