Chapter XVI
Presenting simple descriptive statistics from household survey data

Paul Glewwe
Department of Applied Economics
University of Minnesota
St. Paul, Minnesota, United States of America

Michael Levin
United States Bureau of the Census
Washington, D.C., United States of America

Abstract

The present chapter provides general guidelines for calculating and displaying basic descriptive statistics for household survey data. The analysis is basic in the sense that it consists of the presentation of relatively simple tables and graphs that are easily understandable by a wide audience. The chapter also provides advice on how to put the tables and graphs into a general report intended for widespread dissemination.

Key terms: descriptive statistics, tables, graphs, statistical abstract, dissemination.

A. Introduction

1. The true value of household survey data is realized only when the data are analysed. Data analysis ranges from analyses encompassing very simple summary statistics to extremely complex multivariate analyses. The present chapter serves as an introduction to the next four chapters and, as such, it will focus on basic issues and relatively simple methods. More complex material is presented in the four chapters that follow.

2. Most household survey data can be used in a wide variety of ways to shed light on the phenomena that are the main focus of the survey. In one sense, the starting point for data analysis is basic descriptive statistics such as tables of the means and frequencies of the main variables of interest. Yet, the most fundamental starting point for data analysis lies in the questions that the data were collected to answer. Thus, in almost any household survey, the first task is to set the goals of the survey, and to design the survey questionnaire so that the data collected are suitable for achieving those goals. This implies that survey design and planning for data analysis should be carried out simultaneously before any data are collected. This is explained in more detail in chapter III. The present chapter will focus on many practical aspects of data analysis, assuming that a sensible strategy for data analysis has already been developed following the advice given in chapter III.

3. The organization of this chapter is as follows. Section B reviews types of variables and simple descriptive statistics; section C provides general advice on how to prepare and present basic descriptive statistics from household survey data; and section D makes recommendations on how to prepare a general report (often called a statistical abstract) that disseminates basic results from a household survey to a wide audience. The brief final section offers some concluding remarks.

B. Variables and descriptive statistics

4. Many household surveys collect data on a particular topic or theme, while others collect data on a wide variety of topics. In either case, the data collected can be thought of as a collection of variables, some of which are of interest in isolation, while others are primarily of interest when compared with other variables. Many of the variables will vary at the level of the household, such as the type of dwelling, while others may vary at the level of the individual, such as age and marital status. Some surveys may collect data that vary only at the community level; an example of this is the prices of various goods sold in the local market.[27]

5. The first step in any data analysis is to generate a data set that has all the variables of interest in it. Data analysts can then calculate basic descriptive statistics that let the variables

[27] In most household surveys, the household is defined as a group of individuals who: (a) live in the same dwelling; (b) eat at least one meal together each day; and (c) pool income and other resources for the purchase of goods and services. Some household surveys modify this definition to accommodate local circumstances, but this issue is beyond the scope of this chapter. "Community" is more difficult to define, but for the purposes of this chapter, it can be thought of as a collection of households that live in the same village, town or section of a city. See Frankenberg (2000) for a detailed discussion of the definition of "community".

"speak for themselves". There are a relatively small number of methods of doing so. The present section explains how this is done. It begins with a brief discussion of the different kinds of variables and descriptive statistics, and then discusses methods for presenting data on a single variable, methods for two variables, and methods for three or more variables.

1. Types of variables

6. Household surveys collect data on two types of variables, "categorical" variables and "numerical" variables. Categorical variables are characteristics that are not numbers per se, but categories or types. Examples of categorical variables are dwelling characteristics (floor covering, wall material, type of toilet, etc.), and individual characteristics such as ethnic group, marital status and occupation. In practice, one could assign code numbers to these characteristics, designating one ethnic group as "code 1", another as "code 2", and so on, but this is an arbitrary convention. In contrast, numerical variables are by their very nature numbers. Examples of numerical variables are the number of rooms in a dwelling, the amount of land owned, or the income of a particular household member. Throughout this chapter, the different possible outcomes for categorical variables will be referred to as "categories", while the different possible outcomes for numerical variables will be referred to as "values".

7. When presenting data for either type of variable, it is useful to make another distinction, regarding the number of categories or values that a variable can take. If the number of categories/values is small, say, less than 10, then it is convenient (and informative) to display complete information on the distribution of the variable. However, if the number of values/categories is large, say, more than 10, it is usually best to display only aggregated or summary statistics concerning the distribution of the variable. An example will make this point clear. In one country, the population may consist of a small number of ethnic groups, perhaps only four. For such a country, it is relatively easy to show in a simple table or graph the percentage of the sampled households that belong to each group. Yet, in another country, there may be hundreds of ethnic groups. It would be very tedious to present the percentage of the sampled households that fall into each of, say, 400 different groups. In most cases, it would be simpler and sufficiently informative to aggregate the many different ethnic groups into a small number of broad categories and display the percentage of households that fall into each of these aggregate categories.

8. The example above used a categorical variable, ethnic group, but it also applies to numerical variables. Some numerical variables, such as the number of days a person is ill in the past week, take on only a small number of values and so the entire distribution can be displayed in a simple table or graph. Yet many other numerical variables, such as the number of farm animals owned, can take on a large number of values and thus it is better to present only some summary statistics of the distribution. The main difference in the treatment of categorical and numerical variables arises from how to aggregate when the number of possible values/categories becomes large. For categorical variables, once the decision not to show the whole distribution has been made, one has no choice but to aggregate into broad categories. For numerical variables, it is possible to aggregate into broad categories, but there is also the option of displaying summary statistics such as the mean, the standard deviation, and perhaps the

minimum and maximum values. The following subsection provides a brief review of the most common descriptive statistics.

2. Simple descriptive statistics

9. Tables and graphs can provide basic information about variables of interest using simple descriptive statistics. These statistics include, but are not limited to, percentage distributions, medians, means, and standard deviations. The present subsection reviews these simple statistics, providing examples using household survey data from Saipan, which belongs to the Commonwealth of the Northern Mariana Islands and from American Samoa.

10. *Percentage distributions.* Household surveys rarely collect data for exactly 100, or 1,000 or 10,000 persons or households. Suppose that one has data on the categories of a categorical variable, such as the number of people in a population that are male and the number that are female, or data on a numerical variable, such as the age in years of the members of the same population. Presenting the numbers of observations that fall into each category is usually not as helpful as showing the percentage of the observations that fall into each category. This is seen by looking at the first three columns of numbers in table XVI.1. Most users would find it more difficult to interpret these results if they were given without percentage distributions. The last three columns in table XVI.1 are much easier to understand if one is interested in the proportion of the population that is male and the proportion that is female for the different age groups. Of course, one may be interested in column percentages, that is to say, the percentage of men and the percentage of women falling into different age groups. This is shown in table XVI.2. (A third possibility is to show percentages that add up to 100 per cent over all age by sex categories in the table, but this is usually of less interest.) Both tables show that percentage distributions can be shown for either categorical or numerical variables.

Table XVI.1. Distribution of population by age and sex, Saipan, Commonwealth of the Northern Mariana Islands, April 2002: row percentages

Broad age group, in years	Numbers			Row percentages		
	Total	Male	Female	Total	Male	Female
Total persons	67 011	29 668	37 343	100.0	44.3	55.7
Less than 15	16 915	8 703	8 212	100.0	51.5	48.5
15 to 29	18 950	5 765	13 184	100.0	30.4	69.6
30 to 44	20 803	9 654	11 149	100.0	46.4	53.6
45 to 59	8 105	4 458	3 648	100.0	55.0	45.0
60 years or over	2 239	1 088	1 150	100.0	48.6	51.4

Source: Round 10 of the Commonwealth of the Northern Mariana Islands Current Labour-force Survey.
Note: Data are from a 10 per cent random sample of households and all persons living in collectives.

Table XVI.2. Distribution of population by age and sex, Saipan, Commonwealth of the Northern Mariana Islands, April 2002: column percentages

Broad age group, in years	Numbers			Column percentages		
	Total	Male	Female	Total	Male	Female
Total persons	67 011	29 668	37 343	100.0	100.0	100.0
Less than 15	16 915	8 703	8 212	25.2	29.3	22.0
15 to 29	18 950	5 765	13 184	28.3	19.4	35.3
30 to 44	20 803	9 654	11 149	31.0	32.5	29.9
45 to 59	8 105	4 458	3 648	12.1	15.0	9.8
60 years or over	2 239	1 088	1 150	3.3	3.7	3.1

Source: Round 10 of the Commonwealth of the Northern Mariana Islands Current Labour-force Survey.
Note: Data are from a 10 per cent random sample of households and all persons living in collectives.

11. It is clear from table XVI.1 that the sex distribution differs across the age groups. This reflects something that cannot be seen in tables XVI.1 and XVI.2, namely that Saipan has many immigrant workers – particularly female workers – employed in its garment factories. While Saipan has slightly more males than females at the youngest ages, the next age group, those 15-29 years, has only 30 males for every 70 females. Age group 30-44 also has more females than males. This is consistent with the fact that most of Saipan's garment workers are women between the ages of 20 and 40. In the next group, those 45-59 years of age, there are more males than females. The column percentages in table XVI.2 show that the largest age group for males was that of 30-44, while the largest age group for females was that of 15-29, the age group of females most likely to work in the garment factories.

12. *Medians.* The two most common statistical measures for numerical variables are means and medians. (By definition, categorical variables are not numerical and thus one cannot calculate means and medians for such variables.) The median is the midpoint of a distribution, while the mean is the arithmetic average of the values. The median is often used for variables such as age and income because it is less sensitive to outliers. As an extreme example, let us assume that there are 99 people in a survey with incomes between $8,000 and $12,000 per year, and symmetrically distributed around $10,000. Thus, the mean and the median would be $10,000. Now suppose one more person with an income of $500,000 during the year is included, then, the mean would be about $15,000 while the median would still be about $10,000. For many income variables, published reports often show both the mean and the median.

13. Returning to the data from Saipan, the median age for the Saipan population was 28.5 years in April 2002, that is to say, half the population was older than 28.5 years and half was younger than 28.5 years. The female median age was lower than the male median age (27.6

versus 30.5), because of the large number of young immigrant females working in the garment factories.

14. *Means and standard deviations.* As noted above, the mean is the arithmetic average of a numerical variable. Means are often calculated for the number of children ever born (to women), income, and other numerical variables. The standard deviation measures the average distance of a numerical variable from the mean of that variable, and thus provides a measure of the dispersion in the distribution of any numerical variable.

15. Table XVI.3 shows medians and means for annual income obtained from the 1995 American Samoa Household Survey. The survey was a 20 per cent random sample of all households in the territory. The fact that household mean income was higher than the median income is not surprising, since some households earned significantly higher wages and derived higher income from other sources. Tongan immigrants are relatively poor, as seen by their low mean and low median income; while the high mean and high median income of "other ethnic groups" indicate that they are relatively well off.

Table XVI.3. Summary statistics for household income by ethnic group, American Samoa, 1994

Annual income	Total	Samoan	Tongan	Other ethnic Groups
Number of households surveyed	8 367	7 332	244	790
Median (United States dollars)	15 715	15 786	7 215	23 072
Mean (United States dollars)	20 670	20 582	8 547	25 260

Source: 1995 American Samoa Household Survey.
Note: Data are an unweighted, 20 per cent random sample of households.

3. Presenting descriptive statistics for one variable

16. The simplest case when presenting descriptive statistics from a household survey is that where only one variable is involved. The present subsection explains how this can be produced for both categorical and numerical variables.

17. *Displaying the entire distribution.* Categorical or numerical variables that take a small number of categories or values, say 10 or less, are the simplest to display. A table can be used to show the entire (percentage) distribution of the variable by presenting the frequency of each of the categories or numerical values of the variable. An example of this is given in table XVI.4, which shows the (unweighted) sample frequency counts and percentage distribution for the main sources of lighting among Vietnamese households. Many household surveys require the use of weights to estimate the distribution of a variable in the population, in which case showing the raw sample frequencies may be confusing and thus is not advisable; the use of weights will be discussed in section C below. (The survey from Viet Nam was based on a self-weighting sample

and thus no weights were needed.) A final point is that it is also useful to report the standard errors of the estimated percentage frequencies (see chap. XXI for a detailed discussion of this issue, which is complicated by the use of weights and by other features of the sample design of the survey).

18. In some cases, the number of categories or values taken by a variable may be large, but the major part of the distribution is accounted for by only a few categories or values. In such cases, it may not be necessary to show the frequency of each category or value. One option to prevent the amount of information from taxing the patience of the reader of a table is to combine rare cases into a general "other" category. For example, any category or value with a frequency of less than 1 per cent could go into this category. Indeed, this is what was done in table XVI.4, where "other" includes rare cases such as torches and flashlights. In some cases, there may be other natural groups. For example, in many countries, ethnic and religious groups can be divided into a large number of distinct categories, but there may be a much smaller number of broad groups into which these more precise categories fit. In many cases, it will be sufficient to present figures only for the more general groups. The main exception to this rule concerns categories that may be of particular interest even though they occur rarely. In general, such "special interest but rare" categories could be reported separately, but it is especially important to show standard errors in such instances because the precision of the estimates is lower for rare categories.

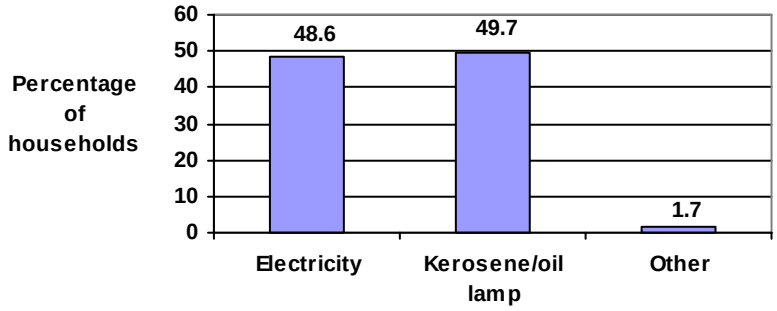
19. In many cases, presentation of data can be made more interesting and more intuitive if it is displayed as a graph or chart instead of as a table. For a single variable that has only a small number of categories or values, a common way to display data graphically is in a column chart or histogram, in which the relative frequency of each category or value is indicated by the height of the column. Figure XVI.1 provides an example of this, using the data presented in table XVI.4. Another common way of displayinig of the relative frequency of the categories or values of a variable is the pie chart, which is a circle showing the relative frequencies in terms of the size of the "slices" of the pie. An example of this is given in figure XVI.2, which also displays the information given in table XVI.4. See Tufte (1983) and Wild and Seber (2000) for detailed advice on how to design effective graphs.

Table XVI.4. Sources of lighting among Vietnamese households, 1992-1993

Method	Number of households	Percentage of households (standard error)
Electricity	2 333	48.6 (0.7)
Kerosene/oil lamp	2 386	49.7 (0.7)
Other	81	1.7 (0.2)
Total households in sample	4 800	100.0

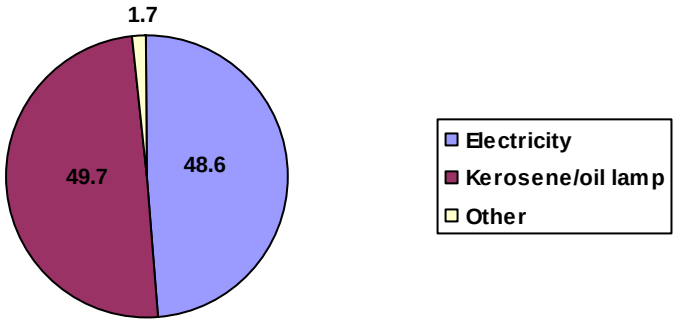
Source: 1992-1993 Viet Nam Living Standards Survey.
Note: Data are unweighted.

Figure XVI.1. Sources of lighting among Vietnamese households, 1992-1993 (column chart)



Source: 1992-1993 Viet Nam Living Standards Survey.
Note: Sample size: 4,800 households.

Figure XVI. 2. Sources of lighting among Vietnamese households, 1992-1993 (pie chart) (Percentage)



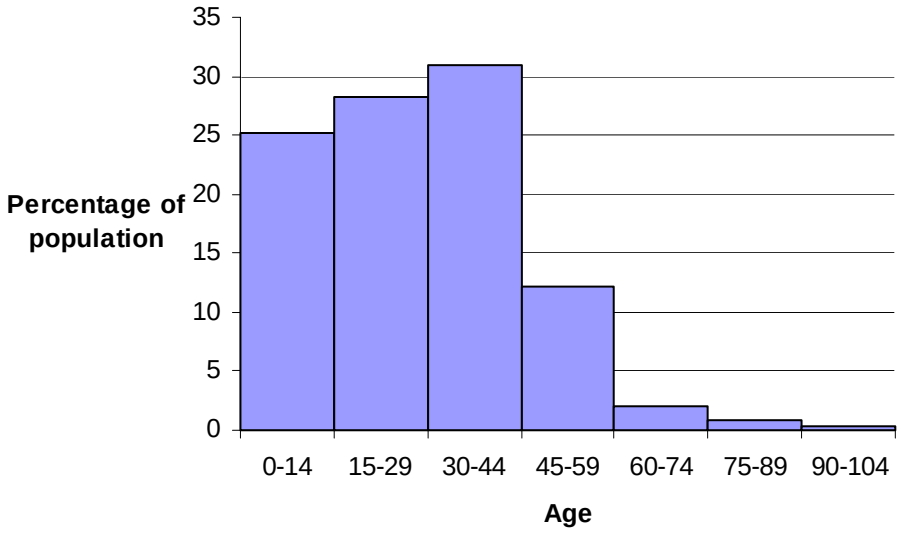
Source: 1992-1993 Viet Nam Living Standards Survey.
Note: Sample size: 4,800 households.

20. *Displaying variables that have many categories or values.* Both categorical and numerical variables often have many possible categories or values. For categorical variables, the only way to avoid presenting highly detailed tables and graphs is to aggregate categories into broad groups and/or combine all rare values into an "other" category, as discussed above. For numerical variables, there are two distinct options.

21. First, one can divide the range of any numerical variable with many values into a small number of intervals and display the information in any of the ways described above for the case where a variable has only a small number of categories or values. For example, this was done for the age variable in tables XVI.1 and XVI.2. This option can also be used in graphs: information on the distribution of a numerical variable that takes many values can be displayed using a graph that shows the frequency with which the variable falls into a small number of categories. One example of such a graph is the histogram, which approximates the density function of the underlying variable. Histograms divide the range of a numerical variable into a relatively small number of "sub-ranges", commonly called bins. Each bin is represented by a

column that has an area proportional to the percentage of the sample that falls in the sub-range corresponding to the bin. Figure XVI.3 does this for the age data in table XVI.2. The first bin is the sub-range from 0 to 14; the next is the sub-range from 15 to 29, and so on.[28] Note that, unlike the column chart in figure XVII.1, there is no distance between the "columns" of the histogram. This is because the horizontal axis in a histogram depicts the range of the variable, and variables typically have no "gaps" in their range.

Figure XVI.3. Age distribution of the population in Saipan, April 2002 (histogram)



Source: Round 10 of the Commonwealth of the Northern Mariana Islands Current Labour-force Survey

22. The second, and perhaps most common, option for displaying a numerical variable that takes many values is to present some summary statistics of its distribution, such as its mean, median, and standard deviation. This can be done only by showing these statistics in a table; it is not possible to show summary statistics for a single numerical variable in a graph. In addition to the mean, median and standard deviation, it is also useful to present the minimum and maximum values, the values of the upper and lower quartiles,[29] and perhaps a measure of skewness. An example of this is given in table XVI.5.

4. Presenting descriptive statistics for two variables.

23. Examination of the relationships between two or more variables often offers much more insight into the underlying topic of interest than examining a single variable in isolation. Yet, at the same time the possibilities for displaying the data increase by an order of magnitude. The

[28] This histogram divides the population aged 60-99 into three groups (60-74, 75-89 and 90-104) each of which spans the same number of years, 15, as the population groups younger than 60. This is done to ensure that the area in each column of the histogram is proportional to the percentage of the population in each age group.

[29] The lower quartile of a distribution is the value for which 25 per cent of the observations are less than the value and 75 per cent are greater than the value, and the upper quartile is the value for which 75 per cent of the observations are lower than the value and 25 per cent are higher than the value.

present subsection describes common methods, distinguishing between variables that have a small number of categories or values and variables that take a large number of values.

24. *Two variables with a small number of categories or values.* The simplest case for displaying the relationship between two variables is that where both variables have a small number of categories or values. In a simple two-way tabulation, the categories or values of one variable can serve as the columns, while the categories or values of the other variable can serve as the rows. An example of this is shown in table XVI.6, which illustrates the use of different types of health service providers in urban and rural areas of Viet Nam. In this example, the columns sum to 100 per cent. As explained above, an alternative would be for the rows to sum to 100 per cent. In the example from Viet Nam, percentage figures that sum to 100 per cent across each row would indicate how the use of each type of health facility was distributed across urban and rural areas of Viet Nam. A third alternative would be for each "cell" of this table to give the frequency (in percentage terms) of the (joint) probability of a visit to a health-care facility by someone in a particular geographical region (urban or rural), in which case the sum of the percentages over all rows and columns would be 100 per cent. This is rarely used, however, since conditional distributions are usually more interesting. In any case, it is good practice to report sufficient data so that any reader can derive all three types of frequencies given the data provided in the table.

Table XVI.5. Summary information on household total expenditures: Viet Nam, 1992-1993 (Thousands of dong per year)

Mean	6 531
Standard deviation	5 375
Median	5 088
Lower quartile	3 364
Upper quartile	7 900
Smallest value	235
Largest value	100 478

Source: 1992-1993 Viet Nam Living Standards Survey.
Note: Sample size: 4,799 households.

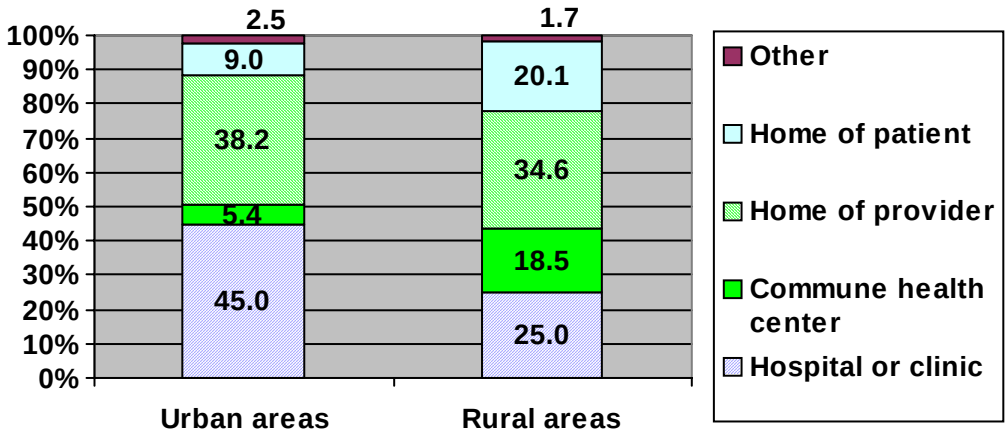
Table XVI.6. Use of health facilities among population (all ages) that visited a health facility in the past four weeks, by urban and rural areas of Viet Nam, in 1992-1993

Place of consultation	Urban areas		Rural areas	
	Frequency	Percentage (std. error)	Frequency	Percentage (std. error)
Hospital or clinic	251	45.0 (2.1)	430	25.0 (1.0)
Commune health centre	30	5.4 (1.0)	318	18.5 (0.9)
Provider's home	213	38.2 (2.1)	595	34.6 (1.1)
Patient's home	50	9.0 (1.2)	376	20.1 (1.0)
Other	14	2.5 (0.7)	29	1.7 (0.3)
Total	558	100.0	1718	100.0

Source: 1992-1993 Viet Nam Living Standards Survey.

25. There are several ways to use graphs to display information on the relationship between two variables that take a small number of values. When showing column or row percentages, one convenient method is to show several vertical columns that sum to 100 per cent. Each column represents a particular value of one of the variables, and the frequency distribution of the other variable is shown as shaded areas of each column. This is shown for the health facility data from Viet Nam in figure XVI.4. Spreadsheet software packages present many other variations that one could use.

Figure XVI.4. Use of health facilities among the population (all ages) that visited a health facility in the past four weeks, by urban and rural areas of Viet Nam, in 1992-1993 (Percentage)



Source: 1992-1993 Viet Nam Living Standards Survey.
Note: Sample size: 2,276.

26. *One variable with a small number of categories/values and a numerical variable with many values.* Another common situation is one where there are two variables. One takes a small number of categories or values (perhaps after aggregating to reduce the number) and the other is a numerical variable that takes many values. Here the most common way to display the data is in terms of the mean of the numerical variable, conditional on each value of the variable that takes a small number of categories or values. One could also add other information, such as the median and the standard deviation. An example of this is seen in table XVI.7, which shows mean household total expenditure levels in Viet Nam in 1992-1993 with households being classified by the seven regions of that country. This could be put into a "profile plot" column graph, where each column (x-axis) represents a region and the lengths of the columns (y-axis) are proportional to the mean incomes for each region.

27. Another option is to transform the continuous variable into a discrete variable by dividing its range into a small number of categories. For example, it is sometimes convenient to divide households into the poorest 20 per cent, the next poorest 20 per cent, and so on, based on household income or expenditures. After this is done, one can use the same methods for displaying data for two discrete variables, as described above. A specific example is to modify figure XVI.4 to show five columns, one for each income quintile.

28. *Two numerical variables with many values.* Statisticians often provide summary information on two numerical variables in terms of their correlation coefficient (the covariance of the two variables divided by the square root of the product of the variances). However, such statistics are often unfamiliar to a general audience. An alternative is to graphically display the data in a scatter-plot that has a dot for each observation. This could show, for example, the extent to which household income is correlated over two periods of time, using observations on the same households in two different surveys (one for each period of time).

Table XVI.7. Total household expenditures by region in Viet Nam, 1992-1993 (Thousands of dong per year)

Region	Mean total expenditures (standard errors in parentheses)
Northern uplands	4 792 (95.5)
Red River delta	5 306 (110.4)
North central	4 708 (107.7)
Central coast	7 280 (234.8)
Central highlands	6 173 (373.7)
South-east	10 786 (398.5)
Mekong Delta	7 801 (167.4)
All Viet Nam	6 531 (77.6)

Source: 1992-1993 Viet Nam Living Standards Survey.
Note: Sample size: 4,799 households.

29. One problem with using scatter-plots is that when the sample size is large, the graph becomes too "crowded" to interpret easily. This can be avoided by drawing a random subsample of the observations (for example, one tenth of the observations) to keep the diagram from becoming too crowded. Another problem with scatter-plots is how to adjust them to account for sampling weights. One simple method is to create duplicate observations, with the sampling weight being the number of duplicates for each observation. This will almost certainly overcrowd the scatter-plot; hence after creating the duplicates, only a random subsample of the observations should be included in the scatter plot.

5. Presenting descriptive statistics for three or more variables

30. In principle, it is possible to display relationships between three or more variables using tables and graphs. Yet, this should be done rarely because it adds additional dimensions that complicate both the understanding of the underlying relationships and the methods for displaying them in simple tables or graphs. In practice, it is sometimes possible to show the descriptive relationships among three variables, but it is almost never feasible to show descriptive relationships among four or more variables.

31. For three variables, the most straightforward approach is to designate one variable as the "conditioning" variable. Either this variable will have a small number of discrete values or, if continuous, it will have to be "discretized" by calculating its distribution over a small number of

intervals over its entire range. After this is done, separate tables or graphs can be constructed for each category or value of this conditioning variable. For example, suppose one is interested in showing the relationship among three variables: the education of the head of household, the income level of the household, and the incidence of child malnutrition. This could be done by generating a separate table or graph of the relationship between income and an indicator of children's nutritional status (such as the incidence of stunting) for each education level. This may show, for example, that the association between income and child nutrition is weaker for households with more educated heads.

C. General advice for presenting descriptive statistics

1. Data preparation

32. Before any figures to be put into tables and graphs are generated, the data must be prepared for analysis. This involves three distinct tasks: checking the data to remove observations that may be highly inaccurate; generating complex (derived) variables; and thoroughly documenting the preparation of the "official" data set to be used for all analysis. In all three tasks, extra effort and attention to detail initially may save much time and many resources in the future. The present subsection presents a brief overview of these tasks; for a much more detailed treatment the reader should consult chapter XV.

33. Virtually every household survey, no matter how carefully planned and executed, will have some observations for some variables that do not appear to be credible. These problems range from item non-response (see chap. XI) and other clear errors -- for example, a three-year-old child who is designated as the head of household -- to much less clear cases, such as a household with very high income but an average level of household expenditures. In many cases, the errors are due to inaccurate data entry from paper questionnaires and so the paper questionnaire should be checked first. Such data entry errors can be easily fixed. If the strange data are on the questionnaire itself, there are several options. First, one could change the value of the variable to "missing". If there are only a small number of such cases, those observations can be excluded when calculating any table or graph that uses that variable.[30] If there are a large number of cases, the "missing" values can be calculated as a distinct category of a categorical variable, labelled "not reported" or "not stated". Second, if most of the cases are concentrated in a small number of households, those households could be dropped. Third, if there are many questionable observations for many households for some variables, a decision may have to be made not to present results for that variable.

34. One approach to missing data is to "impute" missing values using one of several methods. Imputation methods assign values to unknown or "not reported" cases, as well as to cases with implausible values. Approaches include the hot deck imputation and nearest neighbour methods, which allow for a "best guess" for a response when none is available. The idea behind these methods is quite simple: households or people that are similar in some

[30] This option has the disadvantage that the sample size will differ slightly for each table. While this could cause confusion, a note at the bottom of each table explaining that a few observations were dropped should provide sufficient clarification.

characteristics are probably also similar in other characteristics. For example, houses in a given rural village are likely to have walls and roofs that are similar to those of houses in other rural areas, as opposed to houses in urban areas. Similarly, most of the people in a household will have the same religion and ethnicity. The survey team must decide on the specific rules to follow in light of the country's demographic, social, economic and housing conditions.

35. While imputation methods are quite useful, they also may have serious problems. The team members responsible for data analysis must decide whether to change missing data on a case-by-case basis or use some kind of imputation method. The effects on the final tabulations must be considered. Imputing 1 or 2 per cent of the cases should have little or no effect on the final results. If about 5 per cent of the cases are missing or inconsistent with other items, imputation should probably still be considered. However, the need to impute a much larger proportion of values, say 10 per cent or more, could very well make the variable unsuitable for use in display and analysis, hence no results should be presented for that variable. Readers should consult chapters VIII and XI and the references therein for further advice on imputation and the handling of missing values.

36. Another aspect of data preparation is calculation of complex (derived) variables. In many household surveys, total household income or total household expenditure, or both, are calculated based on the values of a large number of variables. For example, total expenditure is typically calculated by adding up expenditures on 100 or more specific food and non-food items. While in theory, calculating these variables is straightforward, in practice many problems can arise. For example, in calculating the farm revenues and expenditures of rural households, it is sometimes the case that farm profits are negative. When strange results occur for specific households, it may help to look at each of the components that go into the overall calculation. One or two may stand out as the cause of the problem. Continuing with the example of farm profits, it may be that the price of some purchased input is unusually high. In this case, the profit could be recalculated using an average price.

37. Unfortunately, preparing the data sets when problems arise is more of an art than a science. Decisions will have to be made when it is not clear which choice is the best. Finally, it is important to document the choices made and, more generally, to document the entire process by which the "raw data" are transformed into tables and graphs. The documentation should include a short narrative about the process plus all the computer programs that manipulated and transformed the data.

2. Presentation of results

38. The best way to present basic statistical results will vary according to the type of survey and the audience. However, some general advice can be given that should apply in almost all cases.

39. The most important general piece of advice is to present results clearly. This implies several more specific recommendations. First, all variables must be defined precisely and clearly. For example, when presenting tables and graphs on household "income", the income variable should be either "per capita income" or "total household income", never just "income".

Complex variables such as income and expenditure should be defined clearly in the text and in footnotes to tables and graphs. Does income refer to income before or after taxes? Does it include the value of owner-occupied housing? Does income refer to income per week, per month or per year? This must be completely clear. For many variables, it is very useful to present in the text the wording in the household questionnaire from which the variable has been derived. For example, for data on adult literacy, it should be very clear how this variable has been defined. It may be defined by the number of years the person has attended school, or the person's ability to sign his or her name, or the respondent's statement that he or she can read a newspaper; or it may be based on some kind of test given to the respondent. Different definitions can give very different results.

40. A second specific recommendation regarding clarity is that percentage distributions of discrete variables should be very clear as to whether they are percentages of households or percentages of people (that is to say, of the population). In many cases, these will give different results. In many countries, better-educated individuals have relatively small families. This implies that the proportion of the population living in households with well-educated heads is smaller than the proportion of households that have a well-educated head. A third recommendation regarding clarity is that graphs should show the numbers underlying the graphical shapes. For example, the column chart in figure XVI.1 shows the percentages for each of three sources of lighting among Vietnamese households, and the same is true of the pie chart in figure XVI.2.

41. Finally, there are several other miscellaneous pieces of advice. First, reports should not present huge numbers of tables and a vast array of numbers in each table. Statistical agencies sometimes present hundreds of tables giving minute details that are unlikely to be of interest to most audiences, and a similar point often applies concerning the detail in a given table. Staff preparing reports should discuss the purpose of the various tables that are being prepared, and if little use can be perceived in presenting a particular table or the detailed information in a given table, then the extraneous information should be excluded. Second, estimates of sampling errors should be reported for a selection of the most important variables collected in the survey; in addition, it is highly useful to show the confidence intervals for key variables or indicators. This is an obvious point, but it is often overlooked. It emphasizes the importance of conveying to the reader the degree of precision of the information provided by the household survey. Third, the sample sizes should be given for each table.

3. What constitutes a good table

42. The present subsection offers specific advice about preparing tables that present information from a household survey. When preparing tables and graphs, the following general principle applies: the information the tables include should be sufficient to enable the user to interpret them correctly without having to consult the text of the report. This is highly important because many users of reports photocopy tables and later use them without reference to the accompanying text.

43. The advice given below is general in nature. For any survey, the survey team must decide which conventions are most appropriate. Once the conventions are chosen, they should

be very strictly followed. However, in some cases, divergence from the conventions may be necessary to illustrate specific points or to display specific types of statistical analyses. A final point regarding this subsection is that almost all of these guidelines for tables also apply to graphs.

44. The various parts of a good table are included in table XVI.6. Each table should contain: a clear title; geographical designators (when appropriate); column headers; stub (row) titles; the data source; and any notes that are relevant.

45. *Title.* The title should provide a succinct description of the table. This description should include: (a) the table number; (b) the population or other universe under consideration (including the unit of analysis, such as households or individuals; (c) an indication of what appears in the rows; (d) an indication of what appears in the columns; (e) the country or region covered by the survey; and (f) the year(s) of the survey.

46. Regarding the table number, most statistical reports number their tables consecutively, starting with table XVI.1, and continuing through to the last table. Sometimes countries use letters and numbers for different tables sets, for example, H01, H02, etc., for housing tables, and P01, P02, etc., for population tables. While this procedure is simple and straightforward, it has the disadvantage that reports become locked into the numbering, making additions or deletions very cumbersome.

47. The universe is the population or housing base covered by the table. If all of the population is included in the table, then the universe can be omitted from the title: the total population is assumed. In contrast, if a table encompasses a subpopulation such as persons in the labour force, and the potential labour force is defined as persons aged 10 years or over, then the title might contain the phrase " Population aged 10 years or over".

48. The title of table XVI.6 also includes an indication of what appears in the rows and what appears in the columns of the table. In particular, it states that the table presents information on types of health facilities used (the rows) and shows this information separately for urban and rural areas (the columns). Including the country or region in the title makes the geographical universe immediately apparent. This feature is most important for researchers comparing results between countries. Obviously, the country statistical office collecting the data will know its own country name; but persons using tables from different countries may need this information in order to distinguish between the countries.

49. Finally, the year(s) of the survey should be in the title to make the time frame immediately apparent. Sometimes, a country's national statistics agency may want to show data from two or more different surveys in the same table. Then two dates may appear, for example "1990 and 2000" or "1980 through 2000". The survey team must make a decision about whether it wants to write out a series of dates (for example, "1980, 1990 and 2000", rather than the simpler, but less complete, "1980 through 2000"); once the decision has been made, however, the country should always follow its decision.

50. *Geographical designators.* Whenever the same table is repeated for lower levels of geography, each table should have a geographical designator to clarify which table applies to which geographical region. For example, if table XVI.6 were repeated for each of Viet Nam's seven regions, the name of the region could appear in parentheses in a second line immediately below the title of the table. "Non-geographical" designators could also be used. For example, a table might be repeated for major ethnic groups or nationalities.

51. *Column headers*. Each column of a table must be labelled with a "header". Column headers can have more than one "level"; for example, in table XVI.6, the header for the first two columns is designated as "Urban areas" and the header for the last two is designated as "Rural areas"; and within both urban and rural areas, there are separate headers for the frequency of observations and for the percentage distribution of those observations. Another point pertains to columns of "totals" or "sums", such as the first column of table XVI.3. The survey team should choose a convention with respect to where these columns will be placed. Traditionally, the total comes last, with all of the attributes shown first across the columns. However, if a table continues for multiple pages, with many columns of information, the survey team may prefer to have the total first (at the left) for the series of columns. When the total appears first, any user will immediately know the total for that series of columns, without having to page through all of the table.

52. Column headers and their associated columns of data should be spaced to minimize blank space on the page. Spacing of columns needs to take into account the number of digits in the maximum figures to appear in the columns, the number of letters in the names of the attributes appearing in the columns, and the total number of "spaces" allowed by the particular font being used. The font used is very important, and should be chosen early in the tabulation process.

53. *Stub (row) titles.* The survey team must also determine conventions to be used for stub (row) headings and titles. Stub "headings" should be left justified and only one variable should be listed on each line. Stub headings should consist of the names of variables displayed in the row. Stubs may include subcategories (nested variables). For example, a stub "group" may have two separate rows, one for male and one for female. Some conventions need to be established to distinguish between the different stub groups; the convention usually involves different indentation for different "levels" of variables.

54. *Precision of numbers*. Many tables suffer from presenting too many significant digits. When percentages are shown, it is almost always sufficient to include only one digit beyond the decimal point; presenting two or more digits rarely provides useful information and has three disadvantages: it distracts the reader, wastes space, and conveys a false sense of precision. Numbers with four or more digits rarely need any decimal point at all. When large numbers are displayed, they should appear in "thousands" or "millions," so that no numbers of more than four or five digits appear.

55. *Source.* The source of the data should appear as the complete name of the survey, usually at the bottom of the table (as seen in table XVI.6). However, sometimes tabulations display more than one survey for a country, or surveys from more than one country. When this happens, the information in the sources becomes more important. The date should be included along with

the name of the survey. If the source is a published report, it is useful to distinguish between the date of publication of the report and the year of data collection. For example, a country might have collected data in 1990, but published the data in 1992. Hence, the source might read "1990 Fertility Survey, 1992" with 1992 indicating the date of publication.

56. *Notes*. Notes provide immediate information with which to properly interpret the results shown in the table. For example, the notes to tables XVI.1 and XVI.2 indicate that the sampled population includes all persons living in either individual dwellings or collectives. In addition to notes at the bottom of a table, a series of definitions and explanations might appear in the text accompanying the tables. The text would include the definitions of the characteristics, for example, it would indicate that the birthplace referred to the mother's living quarters just prior to going to the hospital to deliver, rather than to the hospital location. The text might also include explanations regarding how the data were obtained or are to be used. For example, if the date of birth and age were both collected, but date of birth superseded age when they were inconsistent, this information might assist certain users, like demographers, in assessing the best method of interpreting the data.

4. Use of weights

57. The present subsection provides a brief overview of the use of weights when producing tables and graphs using household survey data. For much more detailed treatment, see chapters II, VI, XIX, XX and XXI and the references therein.

58. With respect to survey weighting, the simplest type of household survey sample design is the "self-weighted" type. In such a case, no weights need actually be used in the analysis because each household in the population has the same probability of being selected in the sample. The 1992-1993 Viet Nam Living Standards Survey used in several of the examples in this chapter was such a survey. Yet, variation in response rates across different types of households usually implies that weights should be calculated to correct for such variation. More importantly, most household surveys are not self-weighted because they draw disproportionately large samples for some parts of the population that are of particular interest. For these surveys, weights must be used to reflect the differential probabilities of selection in order to properly calculate unbiased estimates of the characteristics of interest to the survey.

59. Accurate weights must incorporate three components. The first encompasses the "base weights" or "design weights". These account for variation in the selection probabilities across different groups of households (that is to say, when the sample is not self-weighting) as stipulated by the survey's initial sample design. The second component is adjustment for variation in non-response rates. For example, in many developing countries, wealthier households are less likely to agree to be interviewed than are middle-income and lower-income households. The base weights need to be "inflated" by the inverse of the response rate for all groups of households. Finally, in some cases, there may be "post-stratification adjustments". The rationale for post-stratification is that an independent data source, such as a census, may provide more precise estimates of the distribution of the population by age, sex and ethnic group. If the survey estimates of these distributions do not closely correspond to those of the independent source, the survey data may be re-weighted to force the two distributions to agree.

For a more detailed account of the second and third components, see Lundström and Särndal (1999).

D. Preparing a general report (abstract) for a household survey

60. Most household surveys first disseminate their results by publishing a general report which contains a modest amount of detail on all of the information collected in the survey. Such reports usually have much wider circulation than do more specialized reports that make full use of certain aspects of the data. These general reports are sometimes called "statistical abstracts". The present section provides some specific recommendations for producing these reports, based on Grosh and Muñoz (1996).

1. Content

61. The main material in any general statistical report is a large number of tables and graphs. They should reflect all of the main kinds of information collected in the survey; in-depth analysis of more narrow topics should be left to more focused special reports. A small amount of text should accompany the tables, just enough to clarify the type of information in those tables. There is no need to draw particular policy conclusions, although possible interpretations can be suggested as fruitful areas for future research.

62. The most basic information can be broken down by geographical regions, by sex and, perhaps, by age. If the survey contains income or expenditure data, they can also be broken down by income or expenditure groups. In some countries, there will be large differences across these different groups, and the nature of these differences can be explored further in additional tables. In other countries, some of these differences will not be very large, so there will be no need to present more detail.

63. In addition to the results from the household survey data, the general report should have several pages describing the survey itself, including the sample size and the design of the sample, the date of the survey's start and the date of its termination, and some detail on how the data were collected. The questionnaire or questionnaires used should be included as an annex to the main report.

2. Process

64. A good general statistical report is produced by a team of people, several of whom will ideally have had experience on previous reports. Some team members will focus on the technical aspects of generating tables and graphs, while others will mainly be responsible for the content and the text accompanying the tables. The more technically-oriented team members can choose the statistical software with which they are most familiar, since most statistical software are able to produce the figures needed for the tables and graphs. However, estimation of standard errors will likely require software specifically designed for that purpose, since household survey sample designs are virtually always too complex to be handled properly by standard statistical software packages (see chapter XXI for a discussion of these issues).

65. The team members responsible for the content should meet with experts in government agencies regarding the topics to be included in the report. This will ensure that the tables and graphs present the data in a form most useful to those agencies. It might prove useful as well to consult international aid agencies, which could also find the data useful in planning their programmes (see chap. III for a more general discussion of how to form an effective survey team).

E. Concluding comments

66. This chapter has provided an introduction to the presentation of simple descriptive statistics using household survey data. The treatment has been very general, and undertaken at a very basic level. As much of what has been presented constitutes little more than common sense, data analysts should use their own common sense when facing particular issues regarding the analysis of their surveys. More sophisticated methods can also be used to analyse household survey data, some of which are discussed in later chapters. All things considered, the data analysis for any given household survey will have to be tailored to the main topics and objectives of the survey, and researchers will have to consult specialized books and journals to obtain guidance on issues specific to those topics.

References

Frankenberg, Elizabeth (2000). Community and price data. In *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study,* M. Grosh and P. Glewwe, eds. New York: Oxford University Press, for the World Bank.

Grosh, Margaret, and Juan Muñoz (1996). *A Manual for Planning and Implementing the Living Standards Measurement Study Survey*". Living Standards Measurement Study Working Paper, No. 126. Washington, D.C.: World Bank.

Lundström, S., and C. E. Särndal (1999). Calibration as a standard method for the treatment of non-response in sample surveys", *Journal of Official Statistics,* vol. 13, No. 2, pp. 305-327.

Tufte, Edward (1983). *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press.

Wild, C. J., and G. A. F. Seber (2000). *Chance Encounters: A First Course in Data Analysis and Inference.* New York: Wiley.