American people through the development of programs that utilize natural talents and productive capabilities. The content of the newspaper supports this goal through information of resources, articles on specific tribal arts, and directory of art festivals and workshops. Articles are illustrated with inspiring artwork. Order from the center at 466 Third Street, Niagra Falls, New York 14301.

Owens, Bill, *Documentary Photography* (Danbury, NH: Addison House, 1978).

Techniques for photography are presented in addition to extra topics such as getting the confidence of subjects, the attention of an audience, obtaining grants and getting published. Order from Addison House, Morgan's Run, Danbury, NH 03230.

Shipman, Carl, *Understanding Photography* (Tuscon, AZ: H. P Books, 1974).

A basic guide to photographic methods, covering such topics as camera types, camera adjustments, perspective, light meters, film, color filters, and many others. Order from H. P. Books, P. O. Box 5367, Tuscon, AZ 85703.

Synder, Norman, *The Photography Catalog* (Scranton, PA: Harper and Row, 1976).

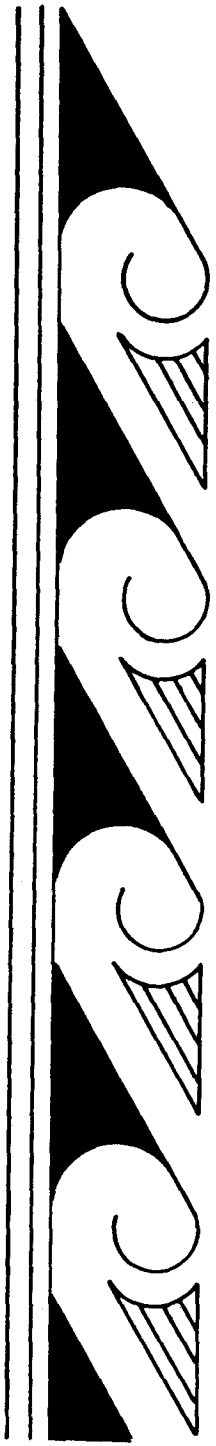
A critical survey of the broad range of available photographic equipment. The evaluations can save time and money in choosing the correct equipment for a project.

Weinstein, Robert A., *Collection, Use, and Care of Historical Photographs* (S. Nashville, TN: American Association for State and Local History, 1977).

This text offers information on the technical and artistic aspects of working with old photographs. Illustrations supplement explanations. Order from American Association for State and Local History, 1400 Eighth Avenue, S. Nashville, TN 37203.

Worth, Sol and John Adair, *Through Navajo Eyes* (Bloomington: Indiana University Press, 1975).

Study of how a group of people structure their view of the world, their reality, through film. Examples of student films illustrate differences in cultural perspectives, as discovered during the film-making process. This is an insightful and valuable text. Films made during the project are available for the Center for Mass Communication, Columbia University Press.



Chapter 6

STATISTICS

Statistics can be useful for describing the characteristics of a community. Such information is considered to be powerful for the development of services and resources for community members. This chapter continues the research process described in earlier chapters for collecting data by showing how different types of data can be coded, tabulated, and displayed for communication to others. Among the calculations you will find described are: frequencies, percentages, cumulative frequencies, means, medians, modes, standard deviations and cross-tabulations. Several ways of displaying data with graphs and charts are shown. For community-based research, data summarized and presented clearly can greatly increase the effectiveness of needs assessments, evaluations, survey efforts and other studies.

Descriptive statistics, or the numberical description of data, can be useful to community efforts in several ways. Certain characteristics about the community or community members can be assigned numbers, summarized, and then described in a way that communicates a total picture of the data collected. Some of the uses of statistics for community-based research are:

* Describing the characteristics of individuals within the community for example, age, blood quantum, tribe, marital status, sex, etc.
* Describing characteristics of families, such as income levels, number of children, type of housing, location, etc.
* Summarizing preferences of community members, as indicated in the results of surveys and needs assessments
* Evaluating the success of certain service delivery approaches
* Determining the cost-effectiveness of programs

Summaries of such socio-economic characteristics can be useful to the community in assessing social services needs, determining priorities for development, and describing community opinion for the structuring of community development.

In this chapter methods are presented for classifying data, determining measures of central tendency (means, modes, medians), representing data by histograms and graphs, and computing variances, ranges, and cross-tabulations. One example is carried through all of these procedures. Although all of these methods of describing data may not be used on one set of data, we are using one set to show some of the possibilities on any given group of data.

THE VARIABLE

In order to be summarized, or analyzed systematically, data must be recorded according to a consistent plan. Once characteristics are defined in the research design, information is gathered repeatedly for each unit, such as individual, family, or community. Characteristics that can assume different values are called **variables**. Variables are small pieces of information that describe each unit in the study. Considered together, they give a larger picture.of the units. Considered separately, they can summarize certain different characteristics of the units in the study. Examples of variables are "client number," "age," "sex," "tribe," "blood quantum," "marital status," "number of children," "religious beliefs," "highest level schooling completed," and "type of educational institution

attended." (Note: These are examples, only some of which a certain project might use. In some communities, data such as "blood quantum" might be seen as information too private to ask. Other communities might see the value of this information in demonstrating eligibility for certain types of funds.)

The "client number," as a variable assigned to each respondent in the study, is used to identify all of the responses or values for one unit in the study. Use of the client number also keeps the identity of the respondent confidential, since names are not used when summarizing the data. The values for "age" might be recorded in years and the values for "number of children" would probably be recorded as a number. A few of these variables and the values they might have are outlined below.

Variable 1. "Blood Quantum"
Values:
0 = No response
1 = Less than ¼
1 = ¼ to ½
3 = Over ½ to ¾
4 = Over ¾ to Full Blood

Variable 2. "Religious Beliefs"
Values:
0 = No response
1 = Traditionalist (Indian religion)
2 = Christian (e. g. Protestant, Catholic, Mormon, etc.)
3 = Christian (Indian church)
4 = Combination Traditionalist and Christian
5 = Other (specify)
6 = None

Variable 3. "Highest Level Schooling Completed"
Values:
0 = No response
1 = No education
2 = Grades 1 to 6
3 = Grades 7 to 9
4 = Grades 10 to 12
5 = High school diploma
6 = Vocational or training school, without HS diploma
7 = Vocational or training school, with HS diploma
8 = College, 1 to 2 years
9 = College, 3 to 4 years
10 = College degree (B.A. or B.S.)
11 = Over 4 years college
12 = Other

Variable 4. "Type of Educational Institution Attended"
Values: 0 = No response
1 = Public
2 = B. I. A. boarding school
3 = B. I. A. day school
4 = Private
5 = Mission
6 = Tribal
7 = Other

In determining the possible values for a variable, culturally important categories are necessary if the results are to be applied for the solution of problems that are important to the community. For example, the values for Variable 1. "Blood Quantum," is decided according to the information generally needed to determine eligibility for services. For Variable 3. "Highest Level Schooling Completed," whether or not a high school diploma was obtained is an important value, as this information could be useful in structuring community educational programs. If the values for this variable were recorded as years of school attendance, this information could be lost, for an individual may attend school for 12 years without obtaining the diploma. Structuring the possible responses or values for a variable is an important and often time-consuming process. Community opinion on the structuring of questions is a valuable resource.

QUANTITATIVE AND QUALITATIVE DATA

When considering the use of statistics, many people think of numerical data that is typically included in surveys. How often the community member expresses weariness over being asked the same questions, with "I've already been counted"! Such an expression may mean that the community member does not see any new or culturally important questions being asked. Our discussion of quantitative and qualitative data is intended to show the wide range of questions that can be asked, assigned numerical values, and then, described statistically.

Quantitative data follows an orderly progression, usually thought of in terms of measurable intervals. Examples of quantitative data are such variables as "age," "income," and "number of children." Qualitative data are more abstract representations qualities, conditions, or opinions. Examples of qualitative data would be the variables, "sex," "area of residence," "marital status." The data collected under the question "sex" could be either female or male. These two qualities can be assigned numbers, such as 1 = female, and 2 = male. Assigning numbers to the values makes it possible to numerically count and summarize the data collected for a certain group of individuals. The data collected for "area of residence" might be assigned the values 1 = reservation, 2 = rural, and 3 = urban. Similarly, the values for "marital status" might be assigned as 1 = never married, 2 = married, 3 = widowed, 4 = divorced, 5 = separated, 6 = common law. Another example, "type of healing preferred," might have the values 1 = Indian traditional, 2 = Western or modern, and 3 = a combination of both, for answer choices. With careful thought, the range of answers, or responses, can include all the cultural meaningful possibilities. Creating categories for responses to qualitative variables and assigning these categories numbers, enables the researcher to summarize or describe the data collected. This concept is explained in more detail under the discussion of questionnaire design in Chapter 3, SURVEY RESEARCH.

CALCULATING FREQUENCIES, PERCENTAGES, AND CUMULATIVE PERCENTAGES

The first example of summarizing qualitative data could come typically from the record keeping system of a social service program, from a survey, or from a needs assessment. The information included in this example is the age, given in years, of sixty-one clients. Since the total number of ages in our sample is sixty-one, this is noted as n = 61. In Figure 6.1, the data are presented with the client identification number to the left and then the age of that client to the right. The data appear as they would be taken directly from the records.

Sorting, or arranging this set of data by age (instead of by client number as in Figure 6.1, from the smallest to the largest will create an order to the data. This arrangement increases our ability to visualize the groupings and tendencies (see Figure 6.2). Sorting the data like this is the first step in grouping identical numbers together, and helps in summarizing the number of times that a particular value occurs in the sample.

The count, or the number of times that a particular value occurs, is known as the **frequency**. A table where the data are listed along with the frequency or count for each occurrence, is called a **frequency distribution.** Particularly if the amount of data is large, it may save time to skip the step of ordering the data. An alternative way to construct a frequency distribution is to read through the entire set of data (as in Figure 6.1) and place a tally mark for each time the number occurs. For example, three items would be ///, and six items would be counted 𝍸 /. In Figure 6.3, the set of ages is listed, a tally mark is placed beside each age in years as that age appears in the original list of data, and then the tally marks are added up. The sum of tally marks is then recorded in the third column, under frequency.

Frequency distributions provide useful information as presented in tables, or this information can be used to represent the data by the use of graphs. Once the count is known for each number, the proportion that each number occurs in relation to the whole sample can be calculated. This proportion, known as a **percentage**, is expressed as the number of units in proportion to one hundred. For example, if the total sample were

Figure 6.1 Data Example, "Age"Sorted By Client Number

Client #	Age	Client #	Age
1	22	31	25
2	19	32	30
3	24	33	31
4	36	34	28
5	21	35	33
6	26	36	22
7	33	37	35
8	18	38	23
9	20	39	35
10	40	40	34
11	22	41	23
12	25	42	34
13	28	43	22
14	23	44	36
15	38	45	25
16	32	46	18
17	24	47	26
18	30	48	27
19	25	49	37
20	42	50	31
21	29	51	48
22	42	52	33
23	36	53	20
24	25	54	29
25	32	55	24
26	26	56	21
27	30	57	25
28	26	58	40
29	35	59	30
30	27	60	22
		61	20

N (total number) = 61

Figure 6.2 "AGE" OF SELECTED CLIENTS SORTED FROM SMALLEST TO LARGEST

18	25	31
18	25	32
19	25	32
20	25	33
20	25	33
20	26	33
21	26	34
21	26	34
22	26	35
22	27	35
22	27	35
22	28	36
23	28	36
23	29	36
23	29	37
24	30	38
24	30	40
24	30	40
25	30	42
	31	42
		48

Figure 6.3 FREQUENCY DISTRIBUTION USING TALLY METHOD

Age	Tally	Frequency
18	11	2
19	1	1
20	111	3
21	11	2
22	~~1111~~	5
23	111	3
24	111	3
25	~~1111~~ 1	6
26	1111	4
27	11	2
28	11	2
29	11	2
30	1111	4
31	11	2
32	11	2
33	111	3
34	11	2
35	111	3
36	111	3
37	1	1
38	1	1
40	11	2
42	11	2
48	1	1

61 cases and the age "25" occurred 6 times, then we would say that age "25" comprised 9.8% of the total sample. This percentage is calculated by dividing the count by the number in the total sample and multiplying by 100. In this example---6 divided by 61, times 100 = 9.8% of the total. Another use of the percentage is the **cumulative percentage**, where the cumulative number of counts is used to calculate a percentage. The cummulative percentage enables the researcher to make summarizing statements about the frequencies in a given sample. In Figure 6.4, a table is presented of the ages, frequencies that the ages occur, and the cummulative frequencies for the previously given sample.

Figure 6.4 PERCENTAGE BREAKDOWN FOR "AGE" OF SELECTED CLIENTS

Age	Frequency	Percentage	Cumulative Percentage
18	2	3.3%	3.3%
19	1	1.6	4.9
20	3	4.9	9.8
21	2	3.3	13.1
22	5	8.2	21.3
23	3	4.9	26.2
24	3	4.9	31.1
25	6	9.8	41.0
26	4	6.6	47.5
27	2	3.3	50.8
28	2	3.3	54.1
29	2	3.3	57.4
30	4	6.6	63.9
31	2	3.3	67.2
32	2	3.3	70.5
33	3	4.9	75.4
34	2	3.3	78.7
35	3	4.9	83.6
36	3	4.9	88.5
37	1	1.6	90.2
38	1	1.6	91.8
40	2	3.3	95.1
42	2	3.3	98.4
48	1	1.6%	100.0%
Totals	61	100.0	100.0

In this example of a frequency breakdown, the age "18" occurs twice, which comprises 3.3% of the total (n = 61). Further down in the table, the age "26" occurs four times, or 6.6% of the total and the cumulative percentage, or all in the sample that are 25 years of age or younger totals

25 or 41%. Similarly, the percentage of those in the sample who are 30 years of age or younger would be 64 percent. Cumulative percentages provide a concise way of describing the cumulative data to certain levels.

GROUPING DATA

For the purposes of examining the data, or presenting the data in a more compact form, the data can be grouped by creating **class intervals**. In other words, classes of data are created by specifying the minimum and maximum, or the intervals, for each class of data. To continue with our example of the data on "age" of 61 clients (Figure 6.5), the ages are grouped into five year intervals. The frequencies are then counted for each class interval, and the **cumulative frequencies** are given.

Figure 6.5 CLASS INTERVALS AND CUMULATIVE FREQUENCIES

Class	Limits (Age in Years)	Frequency	Cumulative Frequency
1	15-20	6	6
2	21-25	19	25
3	26-30	14	39
4	31-35	12	51
5	36-40	7	58
6	41-45	2	60
7	46-50	1	61

Grouping data has both advantages and disadvantages. Data that is grouped into classes can be displayed or presented graphically in a more compact manner. A disadvantage of grouping is that calculations become more difficult or less precise, since the nature of the data becomes less continuous. Methods of dealing with these two considerations are explained later in this chapter. Quantitative data can be collected ungrouped and then grouped later for certain purposes, such as graphic displays or correlations. However, there are certain advantages to creating classes for responses before the data are collected, for the respondent may become more relaxed by the less precise response required. Income level is an example of a variable that sometimes causes tension when the data are elicited, as in a survey effort. Grouping the income intervals gives the respondent a chance to estimate the information. On the other hand, if exact income levels are obtained from kept records, then the use of this more precise data might be desired. Advantages and disadvantages of the different ways of recording data should be taken into consideration during the development of the research design or plan.

The determination of intervals carries cultural information, for more information is gained when data is collected according to a manner appropriate for community needs. For example, in a study of a community where the incomes are known to be low (perhaps by high unemployment rates, health-related problems, and educational levels), the use of $2,000 income intervals may be more appropriate; whereas, if the study concerned tax reform and the income levels of corporation shareholders were being examined, $10,000 income intervals might be more appropriate. In Figure 6.6, a grouping of income levels appropriate for a small rural community is suggested.

Figure 6.6 CLASS INTERVALS FOR INCOME LEVELS

Class	Class Limits
1	Under $2,000
2	$2,000 - $3,999
3	$4,000 - $5,999
4	$6,000 - $7,999
5	$8,000 - $9,999
6	$10,000 - $11,999
7	$12,000 - $13,999
8	$14,000 - $15,999
9	$16,000 - $17,999
10	$18,000 or over

"How many intervals should there be?" is a question often asked by the researcher during the process of designing a questionnaire or data collection system. The spacing of the intervals and the number of intervals is related to the research question, but certain general statements can be made about grouping data. There are generally at least four or five class intervals for a variable, with maximum of fifteen to twenty intervals. When the number of intervals exceeds this amount, then the advantages of grouping the data begin to be lost.

In some instances, it is desirable to group data into intervals that represent a scale, such as smaller to larger, or less to more frequently. This procedure is called **ranking**. Examples of information that may be assigned approximate ranks to the variable "frequency of speaking native language" are (1) speak only English, (2) speak mostly English, (3) speak English and Native language equally, (4) speak mostly Native language, (5) speak only Native language. The sample values presented above for "Blood Quantum" also represent ranked values. Ranking is a way of organizing responses according to some criterion.

One of the pitfalls of ranking data concerns the introduction of bias. For example, the categories for employment could be defined as "unskilled work," "skilled work," "white-collar work," "small business,"

"semi-professional," "professional," and "executive." If these ranked categories were then used in a quantitative way, bias could be introduced to the study due to the value judgements made in setting up the rankings. For example, the values placed on these categories could differ according to different cultural or ethnic views and the incomes reported for the different categories might not necessarily correlate with the ranking from a cultural viewpoint. Or, a ranked employment scale could reflect certain assumptions about the quality of life in certain professions.

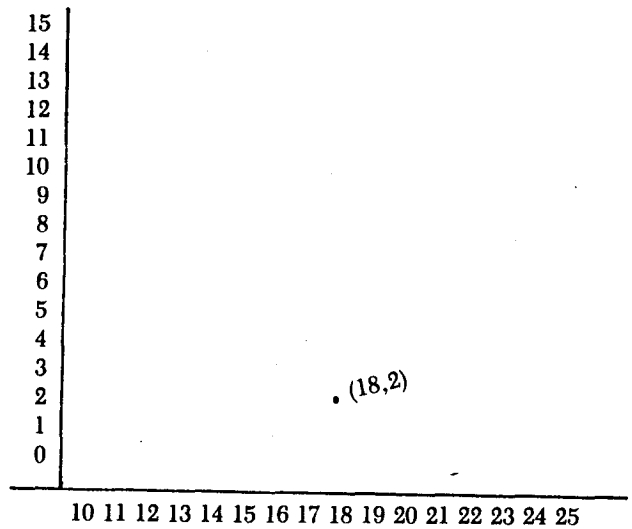
Conflict often arises between the views of the researcher and the views of the community or population being researched, when the categories constructed by the researcher do not correspond with those considered important by the community. For example, in a community where incomes are relatively low, expanded categories within "unskilled work," and "skilled work" might be more descriptive of the community. And other variables relating to the quality of life or job satisfaction might be added to supplement the employment data. The most effective way to reduce bias in the construction of categories is to encourage community participation in the research design. Pilot or sample tests of the instrument can also be useful in obtaining opinions on bias.

REPRESENTING DATA BY GRAPHS AND PICTURES

Graphical representation of data is valuable for several reasons. First, the distribution of the data may be clearer to the reader when presented as a picture or a diagram. Characteristics of the data may become clearer than if presented as a table of numbers. And secondly, graphs are an interesting way of presenting data, when they provide a break in a text or a series of tables. This section will cover the use of the **histogram**, the **frequency polygon**, the **ogive**, and the **scattergram**. These are some of the basic ways used to present data visually.

Graphs are generally based upon a rectangular **coordinate system**, that is, the data are plotted on a grid where the vertical direction is called the **y-axis** and the horizontal direction is called the **x-axis**, as illustrated in Figure 6.7, below. Usually, in making a graph, the unit of measurement for the observation (e. g. age in years) would be measured along the horizontal x-axis and the number of times the observation occurred would be measured along the vertical y-axis. When the coordinates are written they are noted in parentheses with the number representing the x-axis first, and the y-axis second. As an example, if the age 18 appeared twice in the sample, then the coordinates of this occurence would be (18,2), as plotted in Figure 6.8 on a coordinate system. (For additional information on coordinate systems, see Gotkin and Goldstein, pp. 65-75.)

Figure 6.7 EXAMPLE OF A COORDINATE SYSTEM



A **histogram** is a representation of the data on a coordinate system, using vertical bars or boxes. The histogram is a particularly common way of displaying grouped data. The y-axis of a histogram generally starts at 0; whereas, the x-axis can start at a later point. It is important to remember that in numbering the x-axis and the y-axis, the intervals used on each axis should be equal. For example, if each unit on the y-axis represents "1," then all the units on the y-axis should be "1," or the same. If the unit on the x-axis represent "5," then all of the units on the x-axis should be the same.

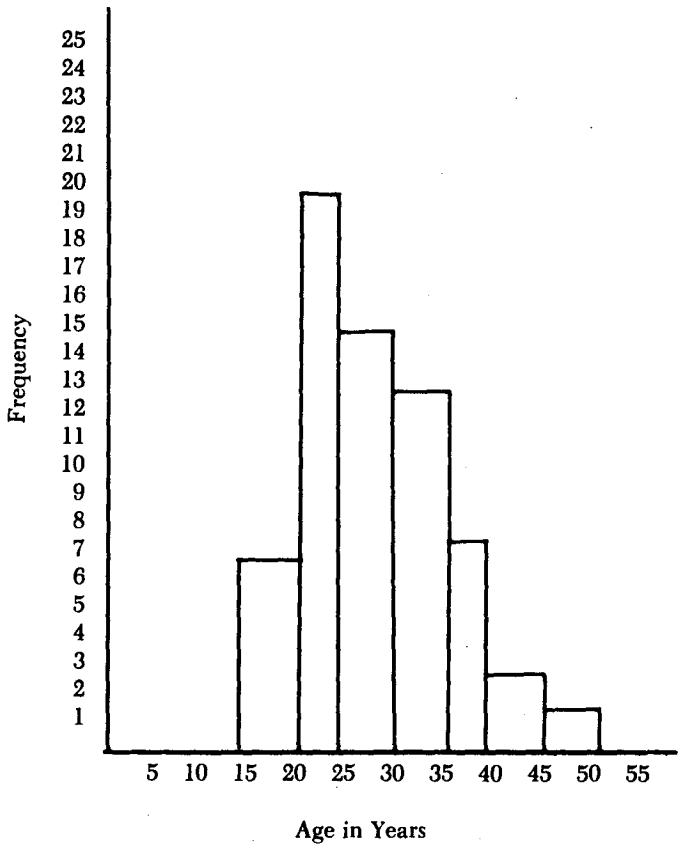
To illustrate the drawing of a histogram, the example given previously on the frequencies of grouped age data will be recorded in bar graph form. The data were presented as follows.

Figure 6.8 GROUP DATA

Class	Class Limits (Age in Years)	Frequency
1		
2	15-20	6
3	21-25	19
4	26-30	14
5	31-35	12
6	36-40	7
7	41-45	2
	46-50	1

Now, this set of data is arranged on a graph to form a histogram in Figure 6.9, with each class represented by a bar on the coordinates. The x-axis represents "age in years" and the y-axis represents the frequency that each class of "age in years" occurs for the sample. To see the contrast between this method of presenting data versus the listing of the individual occurrences, compare the histogram in Figure 6.9 with the original presentation in Figure 6.3 of the frequency distribution of this age sample.

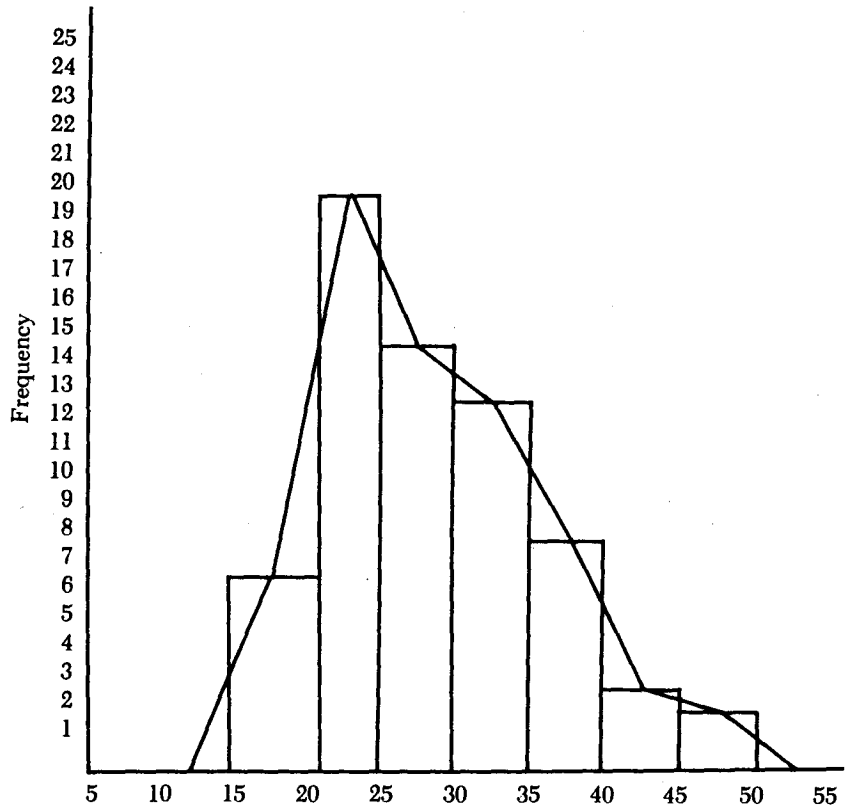
Figure 6.9 HISTOGRAM OF SAMPLE DATA



Another graphical representation of data, the **frequency polygon**, is first plotted by placing a dot at the midpoint of the top of each rectangle. Then, when the dots are connected, a graph is formed. Although this procedure of representing data cuts off some of the area that would be represented by the histogram, it also fills in some spaces that aren't covered by

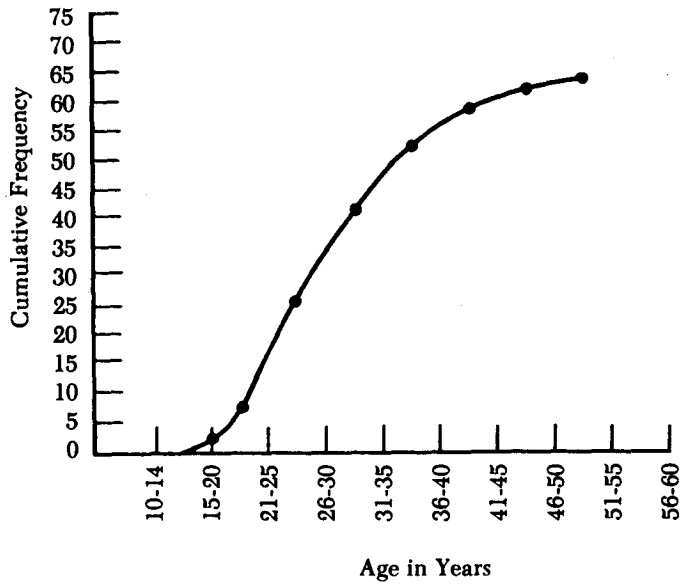
the histogram. Therefore, the total area included under the graph is equal to the histogram. Below, a frequency polygon is drawn over a histogram to illustrate this point.

Figure 6.10 FREQUENCY POLYGON OF SAMPLE DATA OVER HISTOGRAM



The ogive is a graph of a cululative frequency distribution. This type of graph presents additional information, in that the number of observations falling below a certain value are given in addition to the number of observations for each value. To illustrate the ogive, the cumulative frequencies presented in Figure 6.5 are plotted on the graph, thus continuing the sample data set a step further in Figure 6.11, below. To read this graph, one would read the age in years on the horizontal axis and the cumulative frequency on the vertical axis. For example, in the sample there were 39 persons who were of the age thirty or younger, and there were 58 persons of the age forty or younger.

Figure 6.11 OGIVE



The graphic representations shown so far have been concerned with one variable only. One method of displaying the relationship between two variables is the **scatter diagram**. For example, if, for a sample of clients, data were collected on age and income levels, the two variables could be plotted on a graph for each client. The table of income level by age for selected clients, Figure 6.12, will be plotted in a scattergram.

The resulting scatter diagram in Figure 6.13 displays the relationship between these two variables, with the income levels tending to increase as age decreases. Such a pattern might indicate possiblities with further correlations, such as income levels and educational level. The scattergram is a useful visual representation of the relationship between two variables.

MEANS, MEDIANS, AND MODES

Measures of central tendency indicate information about the averages of a given set of data. The most commonly known measure of central tendency is the **mean**, which is often called the "average" in everyday conversation. The mean is calculated by adding together all of the data items or observations, and then dividing that sum by the total number (n) of observations. The mean is explained as:

Figure 6.12 SAMPLE DATA FOR INCOME LEVELS AND AGE OF SELECTED CLIENTS

Client #	AGE	INCOME LEVEL
1	22	$6,000 - $7,999
2	19	$8,000 - $9,999
3	24	$8,000 - $9,999
4	36	$4,000 - $5,999
5	21	$10,000 - $11,999
6	26	$12,000 - $13,999
7	33	$ 6,000 - $ 7,999
8	18	Under $2,000
9	20	$10,000 - $11,999
10	40	$ 4,000 - $ 5,999
11	22	$14,000 - $15,999
12	25	$12,000 - $13,999
13	28	$16,000 - $17,999
14	23	$10,000 - $11,999
15	38	$ 8,000 - $ 9,999
16	32	$ 8,000 - $ 9,999

$$\text{Mean} = \frac{\text{Sum of the observations (or measures)}}{\text{Number of observations}}$$

The symbolism often used to represent the calculation of the mean reads:

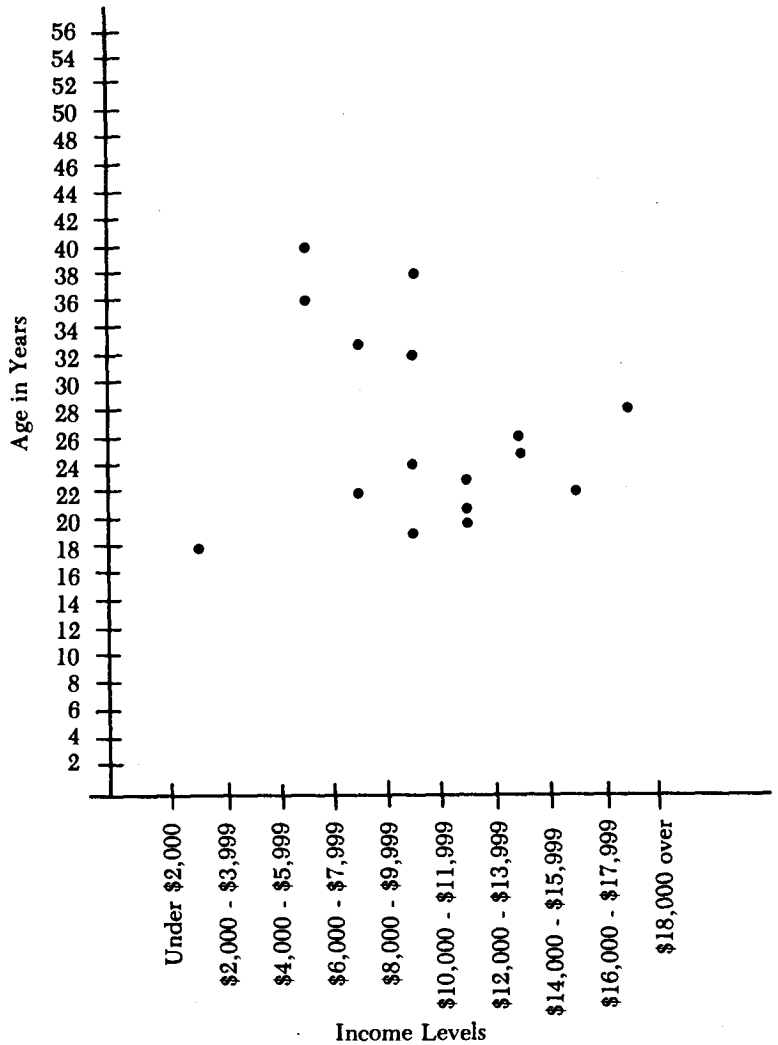
$$M = \frac{\Sigma X}{N}$$

; where M is the mean, Σ describes the operation (summation), X stands for the measures or observations, and N represents the total number of observations.

Referring back to the previous example of quantitative data in Figure 6.2, the mean would be calculated by adding together all of the ages and then dividing that sum by 61, the total number of ages. The mean in this example would be 28.57 years. In notation, the mean is often referred to as $\overline{X}$.

When the data are arranged in order from smallest to largest, as they are presented in Figure 6.2, it is possible to readily determine the middle observation. This number that occurs in the middle of the sample is called the **median**. If the number of observations is even, the median is calculated as the number that is halfway between the two observations occurring in the middle of the sample. In our example, there are 61 observations or ages and the median would be 31st observation. Of the ages given in the example, 27 years would be the 31st observation once the data are ordered from smallest to largest. The median age in this sample would be 27 years.

Figure 6.13 SCATTER DIAGRAM FOR AGE BY INCOME LEVELS



Another measure of central tendency is the mode. The observation in the sample that occurs the most frequently, the **mode**, can be located once the frequencies are determined for all of the observations in a sample. Referring to Figure 6.3, the mode in this sample of ages would be 25 years. Since this age is represented 6 times in the sample, a number of times greater than for any other age, the mode (25 years) is the most frequent age in the distribution.

Means, medians, and modes are useful figures in that a characteristic of a given set of data can be described with one number. This information

can be particularly valuable, when combined with information about the distribution and range of the data.

MEASURING VARIABILITY IN DATA

In addition to knowing about the measures of central tendency or the average in a distribution, it is sometimes useful to know about the dispersion of the data around the average value. Although two sets of data might have the same mean, as reflected in Figure 6.14, the amount that each age varied or deviated from the mean might be different, as in this example. The sum of the ages in each example is equal to 374 years, which divided by 14 (N) gives a mean age of 26.71 years. Yet the list of ages were not identical, some varying more from the mean than others. Each set of ages does not have the same distribution, then, even though the means and the ranges are the same.

Figure 6.14 TWO SETS OF DATA REPRESENTING THE VARIABLE "AGE"

Example 1. Age in Years	Example 2. Age in Years
X	X
17	17
18	19
20	21
20	22
23	22
25	23
27	24
28	26
30	27
31	29
31	30
32	35
32	39
40	40

The variability of a given set of data is based on deviation from the mean. To calculate this variability, or **variance**, from the mean, the following steps would be taken:

1. The deviation of each of the data items (in this case each age) from the mean is calculated by subtracting the mean from each age,
2. Each deviation is then squared, or multiplied by itself,
3. The squares are then summed, or added together, and
4. This sum is then divided by the number of data items (N).

Figure 6.15 THE DEVIATION OF EACH OF X FROM THE MEAN (26.71), THE DEVIATION SQUARED, AND THE VARIANCE

Data From Example 1. (Fig. 6.14)			Data From Example 2		
X	Deviation From $\overline{X}$	$(\text{Deviation})^2$	X	Deviation from $\overline{X}$	$(\text{Deviation})^2$
17	-9.71	94.28	17	-9.71	94.28
18	-8.71	75.86	19	-7.71	59.44
20	-6.71	45.02	21	-5.71	32.60
20	-6.71	45.02	22	-4.71	22.18
23	-3.71	13.76	22	-4.71	22.18
25	-1.71	2.92	23	-3.71	13,76
27	.29	.08	24	-2.71	7.34
28	1.29	1.66	26	- .71	.50
30	3.29	10.82	27	.29	.08
31	4.29	18.40	29	2.29	5.24
31	4.29	18.40	30	3.29	10.82
32	5.29	27.98	35	8.29	68.72
32	5.29	27.98	39	12.29	151.04
40	13.29	176.62	40	13.29	176.62
		558.80 Sum of the squares			664.80 Sum of the squares
		Variance = $\frac{558.80}{14}$ = 39.91			Variance = $\frac{664.80}{14}$ = 47.49

The variance can then be considered as the average of the squared deviations from the mean. The formula for calculating the variance is $\Sigma(X - \overline{X})^2/N$. An alternate, shorter, formula for calculating the variance is $\Sigma X^2/N - \overline{X}^2$. This second formula states that the squares of the data items be summed, this sum is divided by N, and then the square of the mean is subtracted. In summary, the variance is a measure of average dispersion.

Another measure of dispersion is the **standard deviation**. The standard deviation is the square root of the variance, or $s = \sqrt{\text{variance}}$. Rather than expressing the standard deviation as two different formulas, one formula can be used to express the calculation of the standard deviation: $s = \sqrt{\Sigma(X - \overline{X})^2/N}$. For the first example given in Figure 6.15, the standard deviation would be the square root of 39.91, or 6.32. In comparing sets of data, the standard deviations of the sets can give a picture of the distribution in relation to the mean.

The larger the variance or standard deviation, the more the dispersion from the mean. A practical example of how this information could be used is the contrast in the age groups within a certain school grade. If the students in a grade are nearly all the same age, then the variance would be a small number. In contrast, if the ages of students in a particular grade vary quite a bit, there may be socio-economic factors relating to this variance. Gathering and analyzing numerical data can often lead the way to more detailed questions.

CROSSTABULATIONS

In addition to looking at the frequencies of answers to one particular question or variable, it is sometimes valuable to examine the relationship between two variables. For example, certain characteristics of native language speakers might be information useful to the development of a language program. The following example shows the calculation of a crosstabulation of the responses for two variables, "Blood Quantum" by "Speak Native Language."

A crosstabulation is a frequency distribution showing the relationship between two or more variables. Crosstabulation tables, also called contingency tables, provide a way of displaying the joint frequencies. The data used to develop the sample crosstabulation is presented in Figure 6.16 for forty-five respondents. There are four possible responses recorded under the variable "BLOOD QUANTUM," "Less than ¼," "¼ to ½," "Over ½ to ¾," and "Full Blood;" whereas, the responses for the variable "SPEAK NATIVE LANGUAGE" are recorded as "Speak only English," "Speak mostly English," "Speak English and Native language equally," "Speak mostly Native language", and "Speak only Native language."

The table in Figure 6.17 displays the responses for each of the variables. Then, the answers are totaled for each of the joint categories. For exam-

ple, there were two respondents who answered that they are "Less than ¼" Indian and "Speak only English." Reading down to the next row, there was one person who responded as "¼ to ½" for 'Blood Quantum" and 'Speak only English" for the language variable. Also under the "¼ to ½" category for "Blood Quantum" there were five respondents who answered that they "Speak mostly English." Then, reading down to the next row, there were nine respondents who answered "Over ½ to ¾" under "Blood Quantum" and "Speak mostly English" under the language variable. Under the same blood quantum, seven responded "Speak English and Native language equally," with one person who responded "Speak mostly Native language." For the response "Full Blood," there were six who answered "Speak English and Native language," and nine who answered "Speak mostly Native language," and five who answered "Speak only Native language." The totals are then recorded for the rows and for the columns. Instead of counts, the percentages for the rows or columns are sometimes given. An example of this type of table is presented under the chapter on COMPUTERS.

To calculate a crosstabulation for three variables, there would be a table like the one in Figure 6.17 for each response possible in the third variable. The crosstabulation is a valuable method of tabulating and displaying the number of times, or frequency, that the responses for two or more variables occur together.

SAMPLING

In every research study, it is necesary to define the individuals or units included in the study. This process is an important part of the research design, as well as a necessary part of the discussion of the findings or results. The entire group represented by the findings of a study is known as the **population**. The population can be a community, a town, a city, a reservation, or a geographical area defined by some known characteristic. When a study is about a small group, such as a small town or a rural community, it is often possible to include all of the individuals or families in the study. Data collected from an entire population is often called a census. The information gathered when the entire population is included in the study is likely to represent that group more accurately than if only some of the population is reached through the study.

With the use of **descriptive statistics**, the researcher is describing numerically certain characteristics represented by the data collected. When the population in the study is too large to reach every individual, a **sample** or a part of the population may be actually reached by the study. In using descriptive statistics, it is important to remember that the descriptions of data refer only to the sample, and are not used to **infer** or reach conclusions about characteristics of the whole population. This is the main difference between **descriptive statistics** and **inferential statistics**. It

Figure 6.16 DATA FOR VARIABLES, "BLOOD QUANTUM" and "SPEAK NATIVE LANGUAGE"
(Preparation for Crosstabulation)

Respondent #	Blood Quantum	Speak Native Language
1	Over ½ to ¾	Speak English and Native language equally
2	Full Blood	Speak mostly Native language
3	¼ to ½	Speak only English
4	Less than ¼	Speak only English
5	Full Blood	Speak only Native language
6	Over ½ to ¾	Speak mostly English
7	¼ to ½	Speak mostly English
8	Full Blood	Speak mostly Native language
9	Over ½ to ¾	Speak English and Native language equally
10	Full Blood	Speak English and Native language equally
11	Over ½ to ¾	Speak mostly Native language
12	¼ to ½	Speak mostly English
13	Over ½ to ¾	Speak English and Native language equally
14	Over ½ to ¾	Speak mostly English
15	Full Blood	Speak mostly Native language
16	¼ to ½	Speak mostly English
17	Full Blood	Speak English and Native language equally
18	Over ½ to ¾	Speak English and Naitve language equally
19	Full Blood	Speak only Native language
20	Over ½ to ¾	Speak mostly English
21	Full Blood	Speak mostly Native language
22	Full Blood	Speak English and Native language equally
23	Over ½ to ¾	Speak mostly English
24	¼ to ½	Speak only English
25	Full Blood	Speak English and Native language equally
26	Over ½ to ¾	Speak mostly English
27	Over ½ to ¾	Speak English and Native language equally
28	Full Blood	Speak only Native language
29	¼ to ½	Speak mostly English
30	Over ½ to ¾	Speak mostly English
31	Less than ¼	Speak only English
32	Full Blood	Speak mostly Native language
33	Full Blood	Speak only Native language
34	Over ½ to ¾	Speak mostly English
35	Full Blood	Speak English and Native language equally
36	Over ½ to ¾	Speak mostly English
37	Over ½ to ¾	Speak English and Native language equally
38	Full Blood	Speak mostly Native language
39	Full Blood	Speak only Native language
40	Over ½ to ¾	Speak mostly English
41	Full Blood	Speak mostly Native language
42	Full Blood	Speak English and Native language equally
43	Over ½ to ¾	Speak English and Native language equally
44	Full Blood	Speak mostly Native language
45	Full Blood	Speak mostly Native language

Figure 6.17 CROSSTABULATION OF BLOOD QUANTUM BY SPEAK NATIVE LANGUAGE

BLOOD QUANTUM	SPEAK NATIVE LANGUAGE					
	Speak only English	Speak mostly English	Speak English and Native language equally	Speak mostly Native language	Speak only Native language	Totals
Less than ¼	2					2
¼ to ½	1	5				6
Over ½ to ¾		9	7	1		17
Full Blood			6	9	5	20
Totals	3	14	13	10	5	45

is easy to see why obtaining a sample that represents the differences in individuals is very important for an accurate study.

One of the more commonly heard about and sometimes misunderstood methods of sampling is the **random sample**. The idea behind obtaining a random sample is that each individual in the population has an equal chance of being selected. To select a random sample, it is necessary to know the location of all those included in the population. In certain areas, such as reservations or rural areas, it is usually possible to find out the location of all residents, through tribal or county records. In an urban situation, the task of finding the location of all residents becomes more complex. The way researchers tend to determine a random sample for an urban area is through the use of the listings in a phone book. For minority populations, this may not be an accurate way of determining the population, since it is dependent upon the assumption that all households have telephones. Cultural bias can easily affect the sampling method; therefore, the sampling criteria chosen are important.

What can a researcher do when neither an exhaustive sample (all the members in the population) nor a random sample can be chosen? It is difficult to find the answer to this question in the majority of statistics books. The best solution to gaining a representative sample is to include all segments of the population. Some of the variables to remember in considering differences include:

* Age
* Sex
* Geographical location
* Tribe
* Clan
* Religion
* Income Level

The adequate sample represents the different segments of population proportionately, while eliminating the expense and time involved in reaching every person in the population.

Sample size is another factor to consider in planning the research methodology. Again, there are no set rules to an adequate sample size. The usual recommendation is to include the largest sample possible. Larger samples are necessary when small differences are expected in the groups being studied or when groups are divided into subgroups for the purposes of comparison. In general, a survey effort tries to include 10% to 15% of the population. In correlational research, showing relationships between variables, it is considered adequate to include a minimum of about 30 individuals or cases per group. One detail to consider in deciding on a sample size is attrition, or loss of participants in the sample. For example, if a study compares clients before and after a treatment program, the sample may be reduced due to

drop-outs. Or, a certain number of those selected as respondents may refuse to participate in the study. The selected sample size, then, should allow for an expected number for dropping out or for failure to respond.

Sampling in rural areas presents an unusual situation for selecting participants, due to the distances involved. One survey effort on the Navajo reservation[1] developed a technique for survey sampling, by using maps and dividing the area into grids. Then, a certain number of families were selected from each section of the total area. With highly accurate area maps now available, area sampling procedures can be utilized by the community-based research study.

Some of the more common mistakes in sampling include:

* Failure to define the research population before selecting the sample
* Using volunteer subjects rather than a representative sampling approach
* Using a sample size that is too small to allow for comparison of subgroups
* Not allowing for attrition, or loss of cases in the sample

Regardless of the type of sampling method chosen, the details of the sampling should be described in the methodology section of the research report. This enables the person reading the research study to understand and interpret the research findings more clearly. Such a section generally includes a description of the characteristics of the population, the type of sampling used, the reasons for selecting the sampling method, and the proportion of the population reached. The positive benefits of good sampling are efficiency in time spent, with accuracy in the representativeness of a study.

BEYOND DESCRIPTIVE STATISTICS

In this chapter, the basic techniques and limitations of descriptive statistics are discussed. As a further step, the community-based researcher may desire to draw conclusions or infer about populations based on a smaller sample. A knowledge of elementary probability and some basic math skills are usually necessary for understanding the techniques of inferential statistics.

Another advantage to the researcher in the use of inferential statistics is the variety of techniques available to compare two or more variables. The following is a partial list of common techniques: Significance Tests, Chi-square, Bivariate Correlation Analysis, and Regression Analysis. In the bibliography following this chapter, several books are listed that explain these techniques.

Statistics can be a powerful tool in summarizing and presenting information, when the resulting numbers are interpreted and related to the research question. One of the more common misuses of statistics occurs when tables upon tables of numbers are presented, without any explanation of their meaning. Such a presentation leaves the interpretation to the imagination of the reader; and especially if the reader of the study is not very familiar with the topic or the ethnic group, the interpretations may not be correct. The discussion in the research report, then, is an important part of the total presentation. The steps taken in defining the problem, choosing the methodology for data analysis, gathering the data, analyzing the data, and summarizing the results are all a part of the effective research report.

NOTES

1. A description of the survey methods is reported in John Hubbard, Anita Muneta, and Thomas Stewart, "Survey Sampling on the Navajo Reservation," *Human Organization*, Vol. 38, No. 2, pp. 187-189.

ADDITIONAL SOURCES

Blalock, Hubert M., *Social Statistics* (New York: McGraw-Hill, 1972).

An intermediate level text, this book is intended to explain procedures that are appropriate for social science research. The content includes an explanation of the place of statistics in the research process, different types of measurement (nominal, ordinal, and interval scales), measures of central tendency, measures of dispersion, probability, analysis of variance, correlation and regression, analysis of covariance, and different sampling techniques.

Comrey, Andrew L., *Elementary Statistics: A Problem Solving Approach* (Homewood, IL: The Dorsey Press, 1975).

This guide presents problems and explains the solutions for techniques including data representation, measures of central tendency, measures of variability, the normal curve, regression analysis and correlation, testing hypotheses, analysis of variance, and chi-square.

Dixon, Wilfrid J., and Frank J. Massey, Jr., *Introduction to Statistical Analysis* (New York: McGraw-Hill Book Company, 1969).

This is a beginner level text including basic descriptive statistical techniques (histograms, percentiles, means, variance), sampling, statistical inference (single and two populations), analysis of variance, regression, correlation, probability, and nonparametric statistics.

Fitz-Gibbon, Carol Taylor and Lynn Lyons Morris, *How to Calculate Statistics* (Beverly Hills: Sage Publications, 1978).

Contains the principal methods utilized in analyzing questions in evaluation research, such as measures of central tendency, statistical test, and correlation. Presentations are in workbook format, at beginner to intermediate level.

Gotkin, Lassar and Leo Goldstein, *Descriptive Statistics: A Programmed Textbook* (New York: John Wiley & Sons, 1964).

The programmed textbook approach offers an opportunity for the learner to respond after concepts are explained, and then to obtain feedback as to the accuracy of the response by glancing at the answers. Concepts covered include population and sample, variables, data arrangement, data presentation, measures of central tendency, measures of dispersion (range, variance, standard deviation).

Hayslett, H. T., *Statistics Made Simple* (Garden City, NY: Doubleday & Company, 1968).

An excellent, well-illustrated introductory guide, covering pictorial descriptions of data, measures of location, measures of variance, elementary probability, normal distributions, statistical hypotheses, correlation and regression, confidence limits, non-parametric statistics, and the analysis of variance.

Kimble, Gregory A., *How to Use (and Misuse) Statistics* (Englewood Cliffs, NJ: Prentice Hall, 1978).

A readable introductory guide to graphic representations of data, frequency distributions, probability, normal curves, sampling, and correlation. The author gives practical examples, showing applications for the use of statistics and common errors made in the use of statistics.

Koosis, Donald J., *Statistics* (New York: John Wiley & Sons, 1972).

A self-teaching guide that presents a programmed approach to learning statistics; that is, questions and problems are presented as exercises, with the answers given after the problem. Topics covered include frequency distributions, measures of central tendency, measures of variability, samples, estimating, hypothesis testing, difference between means, relations between two sets of measures, chi-square. A good introductory text for the person learning on their own.

Langly, Russel, *Practical Statistics Simply Explained* (New York: Dover Publications, 1968).

The book emphasizes statistical inference, covering the nature of probability, sampling, averages and scatter, the design of investigations, and significance tests. Practical aspects of the text include a discussion of how numbers and the presentation of data can be misleading, as well as examples of the applications of techniques.

Moroney, M. J., *Facts From Figures* (Baltimore: Penguin Books, 1968).

This introductory to intermediate text is easily readable and covers probability, sampling, correlation, ranking methods, analysis of variation, normal distribution, and other techniques.