Packet 57 ur 84a

SOC 326

QUANTATIVE METHODS IN SOCIOLOGY

Professor Stockard

University of Oregon

Winter Term 1992

kinko's

the copy center

860 E. 13th
Eugene • 344-7894

Copies:	$3.54
Binding	$1.75
Royalties	$0.00
Permission Handling Charges	$0.00
Total cost of packet:	$5.29

TABLE OF CONTENTS

Jean Stockard - Soc 326
Packet # 57

I. Introduction to Statistics and Computing

In this section some of the basic definitions and instructions needed for understanding the material in the course are presented. First we will examine material relevant to statistics, whether they are computed with the help of machines or by hand; and then we will discuss the basics of using a computer to analyze data.

Uses of Statistics and Basic Definitions

Below the uses of statistics are discussed. Then types of statistics, levels of measurement, arithmetic operations relevant to our work, and, finally, topics related to measurement are briefly discussed. It is assumed that you have had some exposure to most of these topics, so they are reviewed only briefly.

Uses of Statistics

Statistics are a tool. They help social scientists analyze their data. In themselves, statistics can work no wonders. If a sociologist has poor theory or data that are unreliable or invalid, the best statistics in the world can not improve upon these basic problems. Moreover, there are many different statistics, but only certain ones are relevant for a given problem. Researchers, if they are to have useful results, must choose the appropriate statistics for the data and problem.

The problem of choosing appropriate techniques has become compounded with the availability of easy statistical computations with computers. When statistical computations were done by hand they took many hours to complete and one would not embark upon a computation unless one usually was quite sure that it would be useful. Now one can get a myriad of statistics with the push of a button. Only some of those will be appropriate for a statistical problem and the researcher must think very carefully to make the correct choices.

Given these cautions, we may say that statistics do have many uses. They are a most useful means of summarizing the characteristics of large masses of data. They also allow us to describe the incidence of certain events or behaviors, to look at the associations among two or more variables, and to infer from small samples to large populations. Statistics are used by researchers who employ a whole range of data gathering techniques, for statistics may be used with the qualitative data that are often obtained by participant observers as well as the more quantitative data often used by demographers.

You may have heard the saying that one can "lie with statistics." To some extent this is true. However, one can also lie with words. A solid knowledge of sociological methods and social statistics makes it more likely that you will be able to detect such "lies," if or when they occur.

Descriptive and Inferential Statistics

Statistics may be divided into two basic groups: those that describe the characteristics of a sample or population (descriptive statistics) and those that allow us to generalize from a sample to a population (inferential statistics).

To understand this distinction it helps to review the nature of sampling. Remember that a population is the total group of units (people, organizations, cities, etc.) that one is studying. Only rarely does a social scientist study an entire population. Instead, we usually examine only a subset of the population. This subset is referred to as a sample.

Samples may be selected in basically two ways. In one way, called a probability sample, the elements of the sample are selected so that we know the chance that each member of the population has of being included. The simplest type of probability sample is the simple random sample. Other types include the systematic sample, stratified random, and cluster sample. Samples that are not selected in a way in which we know the chance that each member has of being in the population are termed non-probability samples. These include availability samples, quota samples, and theoretical samples.

Descriptive statistics can be used with either probability or non-probability samples. They describe certain characteristics of the sample. Percentages, averages, and measures of association, such as correlation coefficients, are all examples of descriptive measures or statistics. Inferential statistics are used to infer information from a sample to a population. With inferential statistics we can find the probability that certain characteristics in a sample apply to the population. To make accurate inferences we need, however, to have a probability sample, so inferential statistics are only appropriately used with probability samples. While descriptive and inferential statistics have different uses, they are related, for inferences can be made about descriptive statistics--if we have a probability sample. Thus, in this class, we will learn, among other things, how to make inferences about the average characteristics of a population from information about a sample.

Levels of Measurement

You may remember from your research methods classes that when variables are measured they may be measured in different ways. One way of describing the nature of this measurement is to say whether it is qualitative or quantitative--referring to the extent to which numbers may be assigned to the measure or variable. A more exact distinction involves four levels of measurement. These distinctions are very important to understand for they provide the basis of choosing appropriate statistics for a given data set.

The simplest and most all inclusive level is the nominal one. Variables measured on a nominal scale are placed only in categories. Thus the terms nominal and categorical are sometimes used interchangeably. Within this level no order is posited, we cannot say that one category is greater than or less than another. Examples of a nominally measured variable could include religious affiliation, marital status, race, etc. Any variable that has categories that are mutually exclusive and exhaustive is said to be measured on at least a nominal scale.

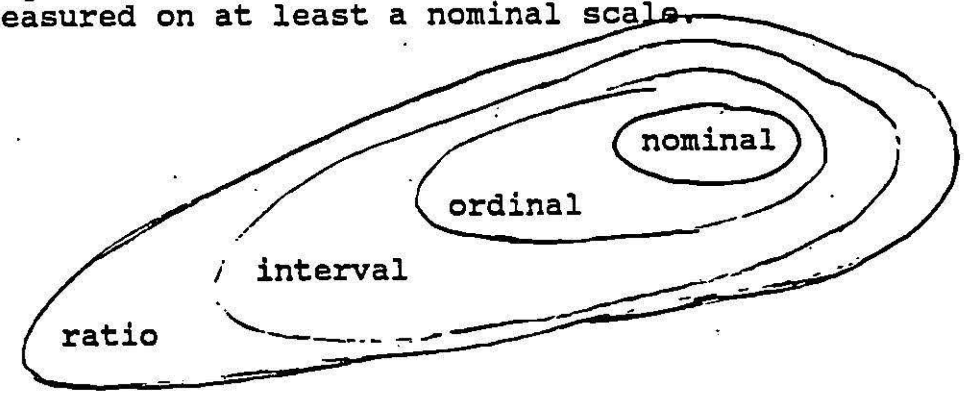


Figure 1-1: Representation of the relative restrictiveness of the four levels of measurement

Variables measured on an ordinal scale are essentially one step up from nominal. The data are still categorical; they have no inherent numerical quality (thus they are still usually referred to as qualitative), but they can be ordered in some fashion. For instance, it is often possible to order religious groups from those that are the most conservative to those that are the most liberal. One can order political groups in the same way. Hair color can be ordered from the most to the least common, etc. Some people claim that practically any variable can be at least ordinal in some theoretical sense.

Interval scales are a step up from ordinal scales, and are the first to be termed quantitative, primarily because arithmetic operations are possible with them. (See more below on this.) An interval scale is like an ordinal scale in the sense that the attributes are ordered. However, with an interval scale we are able to say that the distance between point 1 and point 2 on the scale is the same as that between point 2 and point 3. That is, we can say that there are equal intervals between all points on the scale. Temperature, time, and IQ scores are variables commonly classified as interval.

Ratio scales are the most restrictive. They not only involve ordered categories with equal intervals between them, but there is also a true zero point on the scale. This makes it possible to say that the difference between point 2 and 8, for example, is twice as large as the difference between 2 and 5

(That is, 6 is twice as large as 3). More specifically, we could say that someone who earns $4000/yr. earns twice as much as someone who earns $2000/yr. We cannot say that when it is 80 degrees outside it is twice as hot as when it is 40 degrees, because if we were using different measurement scales (e.g. Celcius or Kelvin) we would have different results than when we used the Farenheit scale. Similarly, grade point averages vary depending on whether we use a four point scale with A=4 or a five point scale with A=1. In each instance the intervals are equal between each letter grade, but the ratios are not.

These examples point to the fact that each level of measurement allows different types of arithmetic relationships or transformations. These in turn specify the types of statistics that can be used. With nominal scales we can employ only matching, or equivalence relations. For instance, if we know that both Mary and John are Catholics, but Beth is not, we can say that Mary and John are in the same category and Beth is in another. Mary and John have equivalent attributes, Beth has a nonequivalent one. ~~With ordinal scales~~ (M=J; M≠B; J≠B).

With ordinal scales we can not only have equivalence relations, we can have ordered relations. Suppose on a scale of political attitudes Mary has the most conservative scores; John has the next more conservative scores; and Beth has the most liberal scores. This tells us that Mary would score highest on a scale of conservatism; John would score lower than Mary, but higher than Beth; and Beth would score lowest (M>J>B and B<J<M).

With interval scales we can have equivalence relations, ordered relations, and also the possibility of adding and multiplying. For instance, we can add up all the high temperatures recorded in a city over a week and compute the average temperature for that week. Similarly, we can compute the average GPA that a student earns in a term. This is possible because the difference between each interval on a temperature scale is equal and the difference between each interval on a grade point scale is equal.

With ratio scales we can not only add and subtract, but we can also discuss ratios. Because there is a meaningful zero we can say that John earns twice as much as Mary or compare the average salaries of whites and blacks as a ratio.

Both the distinction between descriptive and inferential statistics and that between the various levels of measurement will be important, even crucial, in determining which statistics are appropriate for a given problem.

Arithmetic Operations

It is assumed that all students taking this course have taken high school algebra. The following three comments are meant only as a brief review. Students who need a review of basic algebraic definitions and manipulations should consult a textbook.

First, we will often work with rounded numbers or will have to round numbers off to a given point (nearest whole number, nearest ten, etc.). (We will discuss the latter topic more fully in the second part of the course.) When doing computations with rounded numbers, we always round the result to the same point as the original numbers. For instance, if we are doing computations with numbers rounded to the nearest hundredth, the result should be rounded to the nearest hundredth.

e.g. (.36)(.02) = .0072 = .01
or (.36)(.2$\underline{0}$) = .072 = .07

(note that the last significant digit is commonly underlined when it is a zero, to distinguish it from a zero which is not a significant digit.)

The term significant digit refers (as implied above) to how many digits remain in a number that have not been rounded off. That is, it tells us how many of the digits in a number were not rounded off. The chart below illustrates this concept.

Table 1-1

Number	Number of Significant Digits	Rounded to the Nearest ______
1$\underline{0}$	2	whole number
350	2	ten
1400	2	hundred
16$\underline{0}$00	3	hundred
14.$\underline{0}$	3	tenth

Finally, precision refers to how exact our measures are. For instance, a population figure of 43,976 is said to be more precise than a population figure of 44,000. While in areas, such as the physical sciences, very precise measures are both possible and desirable, this is often not the case in the social sciences. In fact the population figure of 44,000 may well be more accurate and thus preferable to the more precise figure.

Measurement Issues

It is assumed that students have had an introduction to the logic involved in measurement in their basic research course.

The following comments then are made only to remind students of important distinctions and concepts.

First, the distinction between discrete and continuous variables can be an important one when working with quantitative variables (those measured on an interval or ratio scale). Discrete variables are those where the values can be actually numbered or counted. Examples could be the number of children in a family, the size of a city or country, etc. We cannot have one-half of a child or one-half of a person. Continuous variables are those whose possible values form a continuum. Examples include age, height, time, etc. We are constantly growing older; people vary along a continuum of height and weight, etc.

Note that we often round continuous variables and treat them as though they were discrete. For instance, we talk about all two years olds, all three year olds, etc. When placing data into tables this is often the preferable step, in order to make the date easier to understand. When doing statistical computations by hand, grouping continuous data also makes them easier to work with. However, as long as our measures are accurate, it is generally best to keep the measures as continuous as possible, especially if one has machines to do the computations.

Second, it is important to briefly discuss measurement error. Measurement error is a very complex topic, well beyond the scope of this course. Here we can only note that errors in measurement do occur. The statistical treatments we will deal with all assume that this measurement error is random. For instance, in measuring income sometimes we may have a high estimate, sometimes our estimate is low--but in the long run these errors balance out. While we know that this is often not the case, the ways of dealing with this error (in a statistical manner) are too complex to be explained until you understand the material given in this course and probably your next statistics course.

Computer Work

Almost all of the statistics we will do this term will be computed with the help of computers. Below we examine the advantages and disadvantages of using the computer, an overview of the SPSS package that we will use, a description of the data file that may be analyzed, and an example of a run using these data.

Computers vs. Hand Computations

Obviously, computers have many advantages over hand computations in doing statistical work. They are much faster and easier to use and they are also much more accurate (assuming the input data and computer programming are correct) than hand computations. Just a relatively few years ago social scientists would spend literally hundreds of hours in data reduction (getting simple frequency counts) and computing the simplest of statistics. They can now accomplish this work in a few minutes.

On the other hand, because it is now so easy to calculate a wealth of statistics at the literal touch of a finger there is a great danger of misusing statistics. Computers cannot decide for you what kind of statistic is appropriate for a given problem or how to interpret a statistic once you have it. The researcher must give a good deal of thought to his or her analysis in order to choose the proper analysis method. Furthermore, we usually code our data when we use machines to analyze it and we must make sure that the measures that the machine is using are comparable to what we really want it to analyze. At all steps of the analysis process the researcher must think very carefully about what is happening. This was true, of course, when computations were done by hand. But, perhaps because it is so easy now to get all kinds of statistics from a machine in just a few minutes, it is especially important to remember how important this planning is now.

Statisical Package for the Social Sciences (SPSS)

In this class you will be using SPSS/PC+ studentware to analyze data. The SPSS package is a very widely used set of computer programs developed for both main frame and personal computers. It is probably the most flexible and widely used program for social scientists. You will be using a version of the program that has been specifically developed for the PC and for student use. The commands that you will be using are similar to those which are used in the mainframe and regular pc version, so it will be relatively easy for you to use other versions of SPSS once you have worked with this package. There are several other computer packages commonly used by social scientists (biomed and SAS are perhaps the most common), and all are relatively easy to learn once you have some familiarity with using a computer for data analysis. The book by Norusis required for the class describes the SPSS/PC+ studentware program in great detail. Classes will also be held to introduce you to the use of the computer package (or software as it is commonly called).

With SPSS we can take a group of data that has been coded and prepared in a form that is readable by the machine (say on cards, tape, or disk) and tell the computer (through ways defined by SPSS) what each of the variables are and where they reside on the cards, tape or disk. This set of data is referred to as our data file or as an SPSS system file, once it has been defined within the SPSS system. A data file is generally arranged so that each case or unit of analysis (people, states, nations, organizations) is in a row or set of rows and each variable is in a different column. The data we will use has already been defined within the SPSS system and is such a data file. (See below.)

Once our data have been defined we can then ask the machine to perform various statistical manipulations with the data. For instance, we might ask the machine to look at a certain variable, tell us how many cases have each attribute of the variable, to compute the percentages associated with these frequencies, and perhaps, if appropriate, to compute some type of average. This would be done with various "tasks" or lines in the program where we define the "procedures" we want the computer to do and the associated statistics. The manuals associated with a given computer program give detailed instructions on how to ask the computer to perform these manipulations.

The Bank Data File

For this class you can use a variety of SPSS system files that have been developed by the SPSS company. One of these includes data on all the employees of a midwestern bank that were hired in 1969, 1970, and 1971. The data were gathered in March, 1977. Data are available on the subjects' sex, race, age, length of employment in the bank, current and beginning salary, educational attainment, and the category of job in which they currently work. The code book for this data set is given below and is similar in format to all codebooks. In the codebook the left-hand column gives the SPSS variable name for each variable. This is the way that the variable is identified in the SPSS system file. Thus, if one wished to analyze the variable regarding job seniority one would ask the computer to look at the variable TIME. If one wanted to look at current salary, one would ask the computer to look at SALNOW.

The right hand column describes each of these variables. For instance, SALBEG, the beginning salary of each employee, is coded as the actual salary, in dollars, at which the employee began work at the bank. SEX is coded with 0 meaning male, and 1 meaning female. Unlike many data sets, the bank data set has not grouped the quantitative data. Because it was possible to actually examine the exact data on salary and age and experience, instead of asking people to report these figures, the actual dollars earned, months worked, or age (in years and fraction of years) are coded.

At the bottom of the page it is noted that N=474. This means that there are 474 people included in the data set. There are no missing data.

Figure 1-2
Sample of Codebook for Bank Data
Bank Employment

SPSS Variable Name	Description and Code
ID	Identification number of each employee
SALBEG	Beginning salary when hired actual beginning salary is coded (5 digits) 0 -- missing
SEX	Sex of employee 0 -- male 1 -- female 9 -- don't know, missing
TIME	Job seniority, coded in number of months have worked at the bank 0 -- missing
AGE	Employee's age, coded in actual years with two significant decimal points
SALNOW	Current Salary, in actual dollars (5 significant digits)
EDLEVEL	Years of education attained (actual years are coded)
WORK	Years of work experience, with two significant digits beyond the decimal point
JOBCAT	Employment category 1 -- clerical 2 -- office trainee 3 -- security officer 4 -- college trainee 5 -- exempt employee 6 -- MBA trainee 7 -- technical
MINORITY	Minority classification 0 -- white 1 -- nonwhite
SEXRACE	Sex and race classification 1 -- white males 2 -- minority males 3 -- white females 4 -- minority females

N = 474

A Sample Run

You might find it helpful to ask the computer to produce a listing of each of the variables in the file with the number of people holding each attribute and the associated descriptive statistics. You can ask SPSS to produce such output by using the subprogram or procedure FREQUENCIES. The manual gives details on the procedure, but it generally would involve giving the computer instructions like the following.

```
get file = 'bank.sys'.
frequencies  variables = salbeg to sexrace.
```

The first line instructs the computer to access the bank data in what is known as a systems file. This is the part of its memory where it has stored the data. If you have data of your own that you want to use you would need to tell the computer what the data were and how to find them. Note that the line ends in a period. That tells the computer that you are finished with the get file command.

The second line asks the computer to run the procedure "frequencies" and count the number of cases for all of the variables from salbeg to sexrace. Note that ID is not included in the list. That would result in a waste of paper, simply listing each individual case. Other commands can be added to ask the computer to compute various descriptive statistics such as those described in the next section.

II. Descriptive Univariate Statistics

We move now to examining ways of summarizing and describing distributions of single variables. We first discuss the construction of tables that summarize data and then describe graphs that can be used to pictorially represent these data. We then describe various measures of central tendency and finally measures of dispersion.

Tables

Most of our discussion in this section will involve quantitative data (those measured on an interval or ratio scale). The procedures involved with qualitative data are essentially equivalent, but because one cannot "round off" qualitative data or "group" it in the same way one deals with quantitative data, the discussion regarding quantitative data is somewhat more complex and will be the focus of our discussion.

When dealing with masses of quantitative data we usually start with a mass of numbers. For instance, with the bank data we might be interested in the subjects' ages. We could ask the computer to give us a listing of the subjects' ages and we would have a page of computer printout such as that shown on the following pages. Note that the computer has already arranged the numbers in chronological order, and that the computer tells us how many people have each age. One person is 23 years old, 2 people are 23.25 years old, 1 person is 23.33 years old, etc.

Sometimes, we will want to round off the numbers to bring them to a more manageable size. This is especially true if the numbers are quite large or extend to several more decimal points than we desire. For instance, we might be more interested in age to the nearest year, rather than to the hundredth of a year. We would then round 23.25 years to 23 years; 23.58 years would become 24 years, etc. In arithmetic you might have learned that when rounding to the nearest whole number and the original number ends in 5, you automatically round up. Thus 15.5 would become 16, 16.5 would become 17, 17.5 would become 18, etc. Note, however, that this introduces an upward bias. We are always rounding upward. To counteract this upward bias, the convention among social statisticians when rounding to the nearest number is to round to the nearest even number when the original number ends in 5. Thus 14.5 would become 14, 15.5 would become 16, 16.5 would become 16, 17.5 would become 18, etc. This produces somewhat higher groups at each of these even numbers, but it avoids the upward bias present in the other system and is thus more accurate.

Note that we do not always round to the nearest whole number. In fact, with age, in our society, we actually round to the next lower number. One does not become one year of age until living an entire year; one is then considered one year old until

Table 2-1 Output from SPSS Frequencies Run for Age

		Adj	Cum			Adj	Cum			Adj	Cum
Code	Freq	%	%	Code	Freq	%	%	Code	Freq	%	%
23.00	1	0	0	32.00	3	1	50	46.58	2	0	77
23.25	2	0	1	32.08	5	1	51	47.25	1	0	77
23.33	1	0	1	32.17	1	0	51	47.33	2	0	77
23.42	3	1	1	32.25	3	1	52	47.58	2	0	78
23.58	1	0	2	32.33	2	0	53	47.92	1	0	78
23.67	3	1	2	32.50	2	0	53	48.00	1	0	78
23.75	1	0	3	32.67	4	1	54	48.25	1	0	78
24.00	2	0	3	32.83	2	0	54	48.33	1	0	79
24.08	2	0	3	32.92	3	1	55	48.50	1	0	79
24.17	2	0	4	33.08	1	0	55	48.67	1	0	79
24.33	5	1	5	33.33	1	0	55	48.83	1	0	79
24.42	2	0	5	33.42	2	0	56	49.08	1	0	80
24.50	2	0	6	33.50	4	1	57	49.17	1	0	80
24.58	2	0	6	33.67	1	0	57	49.58	1	0	80
24.67	2	0	7	33.75	2	0	57	49.92	1	0	80
24.75	3	1	7	33.83	2	0	58	50.00	1	0	80
24.83	3	1	8	34.00	1	0	58	50.17	1	0	81
24.92	3	1	8	34.17	3	1	58	50.25	2	0	81
25.00	3	1	9	34.25	2	0	59	50.33	1	0	81
25.08	4	1	10	34.33	2	0	59	51.00	1	0	81
25.17	1	0	10	34.50	1	0	59	51.17	1	0	82
25.25	3	1	11	34.58	2	0	60	51.42	2	0	82
25.42	3	1	11	34.67	1	0	60	51.50	3	1	83
25.50	3	1	12	34.75	1	0	60	51.58	2	0	83
25.58	4	1	13	34.83	1	0	61	51.92	1	0	83
25.75	2	0	13	34.92	1	0	61	52.00	2	0	84
25.83	3	1	14	35.17	2	0	61	52.17	1	0	84
25.92	1	0	14	35.25	1	0	61	52.33	1	0	84
26.08	1	0	14	35.33	1	0	62	52.50	1	0	84
26.25	3	1	15	35.42	2	0	62	52.92	1	0	85
26.33	1	0	15	35.58	1	0	62	53.08	1	0	85
26.58	1	0	15	35.67	1	0	62	53.33	1	0	85
26.67	1	0	16	36.00	1	0	63	53.50	1	0	85
26.83	4	1	16	36.92	1	0	63	53.92	3	1	86
26.92	1	0	17	37.08	1	0	63	54.08	1	0	86
27.00	1	0	17	37.17	1	0	63	54.17	2	0	86
27.08	3	1	18	37.50	1	0	64	54.33	1	0	87
27.17	2	0	18	37.83	1	0	64	54.42	1	0	87
27.25	3	1	19	38.00	1	0	64	54.92	1	0	87
27.33	3	1	19	38.17	1	0	64	55.08	1	0	87
27.42	3	1	20	38.42	1	0	64	55.17	1	0	88
27.50	2	0	20	38.50	1	0	65	55.25	2	0	88
27.58	4	1	21	38.67	1	0	65	55.33	1	0	88
27.67	2	0	22	38.92	1	0	65	55.50	1	0	88

Table 2-1 (page 2)

```
27.83    2  0  22   39.00    1  0  65   55.58    3  1  89
28.00    2  0  22   39.33    2  0  66   55.92    1  0  89
28.08    1  0  23   39.42    1  0  66   56.00    1  0  89
28.17    3  1  23   39.50    1  0  66   56.67    2  0  90
28.33    4  1  24   39.67    3  1  67   56.92    1  0  90
28.42    4  1  25   39.75    1  0  67   57.17    1  0  90
28.50    3  1  26   39.83    1  0  67   57.42    1  0  91
28.67    5  1  27   40.08    1  0  67   57.50    1  0  91
28.75    4  1  27   40.17    1  0  68   57.83    2  0  91
28.83    3  1  28   40.33    1  0  68   58.00    1  0  91
29.00    2  0  28   40.50    1  0  68   58.08    1  0  92
29.08    4  1  29   40.58    1  0  68   58.50    1  0  92
29.17    4  1  30   40.67    1  0  68   58.75    1  0  92
29.25    3  1  31   41.00    1  0  69   59.08    2  0  92
29.33    3  1  31   41.17    2  0  69   59.42    1  0  93
29.42    1  0  32   41.67    1  0  69   59.50    1  0  93
29.50    6  1  33   41.92    2  0  70   59.75    1  0  93
29.58    4  1  34   42.08    1  0  70   59.83    3  1  94
29.67    4  1  35   42.17    1  0  70   60.00    1  0  94
29.75    4  1  35   42.33    1  0  70   60.50    3  1  95
29.92    4  1  36   42.42    1  0  70   60.67    3  1  95
30.00    1  0  36   42.58    2  0  71   60.75    1  0  95
30.08    3  1  37   43.25    1  0  71   61.33    1  0  96
30.17    5  1  38   43.33    1  0  71   61.50    1  0  96
30.25    4  1  39   43.42    1  0  72   61.67    2  0  96
30.33    6  1  40   43.67    1  0  72   61.75    1  0  96
30.42    4  1  41   43.92    1  0  72   62.00    1  0  97
30.50    2  0  42   44.00    1  0  72   62.08    1  0  97
30.58    1  0  42   44.42    1  0  72   62.33    1  0  97
30.67    4  1  43   44.50    3  1  73   62.42    1  0  97
30.75    5  1  44   44.58    1  0  73   62.50    1  0  97
30.83    1  0  44   44.67    1  0  73   63.00    1  0  98
30.92    2  0  44   44.83    1  0  74   63.25    1  0  98
31.00    2  0  45   44.92    1  0  74   63.42    1  0  98
31.08    1  0  45   45.17    1  0  74   63.50    1  0  98
31.17    3  1  46   45.50    2  0  74   63.58    1  0  99
31.25    2  0  46   45.67    1  0  75   63.75    2  0  99
31.33    1  0  46   45.92    1  0  75   63.83    1  0  99
31.42    1  0  46   46.00    1  0  75   63.92    1  0  99
31.50    3  1  47   46.17    1  0  75   64.25    2  0 100
31.67    3  1  48   46.25    2  0  76   64.50    1  0 100
31.75    4  1  49   46.42    1  0  76
31.92    5  1  50   46.50    2  0  76


Mean        37.186      Std err      0.541      Median      32.013
Mode        29.500      Std dev     11.787      Variance   138.939
Kurtosis    -0.562      Skewness     0.864      Range       41.500
Minimum     23.000      Maximum     64.500

Valid cases    474      Missing cases    0
```

Table 2-2

Examples of Rounding Rules and Interval Limits

Number	Rounded to the — position	Rounded Value			True limits of Interval		
		nearest	next lower	next higher	nearest	next lower	next higher
181	10	180	180	190	175-185	180-190	180-190
257	100	300	200	300	250-350	200-300	200-300
3191	1000	3000	3000	4000	2500 3500	3000-4000	3000-4000
4.92	.1	4.9	4.9	5.$\underline{0}$	4.85-4.95	4.9-5.0	4.9-5.0
5.019	.01	5.02	5.01	5.02	5.015-5.025	5.01-5.02	5.01-5.02
6 0199	.001	6.02$\underline{0}$	6.019	6.02$\underline{0}$	6 0195-6.025	6.019-6.02$\underline{0}$	6.019-6 02$\underline{0}$
35	10	40	30	40	35-45	30-40	30-40
45	10	40	40	50	35-45	40-50	40-50
55	10	60	50	60	55-65	50-60	50-60

one has lived a total of two years. In the grocery store all prices are rounded to the next higher number. So, if two cans cost $.49 and you buy one you would pay $.25 automatically. If the price is 3 for a dollar and you buy one, you will pay $.34 and not $.33. Table 2-2 illustrates these different rounding rules.

Whether or not one rounds off the numbers one is dealing with, one will then proceed to developing groups or intervals in which to place each of the cases. Suppose that we decided we wanted to group the bank employees into age categories that each included a span of five years. Remembering that we had rounded the ages to the nearest year we could say that we wanted to include all people with ages from 20.51 to 25.49 years (or rounded limits of 21 to 25 years) in the first category. Those from 25.5 to 30.5 (or rounded limits of 26 - 30 years) in the second category, and so on. These categories are displayed in Table 2-3. The rounded limits refer to the rounded numbers that define the ages. The true limits refer to the actual span of ages that is included within each interval. The interval width (i) refers to the total number of years included in each interval. Note that it is the difference between the upper and lower limits of each true interval (i=U-L). The midpoint of each interval is the lower limit of each true interval plus one-half of the interval width (M = L + (1/2)i).

Table 2-3 Intervals & Midpoints for Grouped Age Data Levels

Rounded Limits	True Limits	Interval Width	Midpoint
21-25	20.5-25.5	5	23
26-30	25.5-30.5	5	28
31-35	30.5-35.5	5	33
36-40	35.5-40.5	5	38
41-45	40.5-45.5	5	43
46-50	45.5-50.5	5	48
51-55	50.5-55.5	5	53
56-60	55.5-60.5	5	58
61-65	60.5-65.5	5	63

Now that the intervals are established we can return to the distribution from the computer printout that is in Table 2-1 and actually count up the number of people that fall into each interval. For instance, we can determine that 54 people fall in the first category with ages between 20.51 and 25.49 or 21 and 25 rounded years. (Note that the first interval has a true lower limit that is substantially lower than the lowest age. This was done to allow for age intervals that were evenly spaced at points on the scale that were easy to comprehend.) In the second interval (true limits of 25.5 to 30.5 and rounded limits of 26 to 30) there are 143 people. You may continue this process until you have determined how many people are within each of the intervals. Table 2-4 summarizes these frequency counts and is referred to as the frequency distribution for age for this sample of bank employees.

Table 2-4 Age of Bank Employees

Years	Frequency	"Less than" Cumulative Frequency	"More than" Cumulative Frequency
21-25	54	54	474
26-30	143	197	420
31-35	97	294	277
36-40	28	322	180
41-45	29	351	152
46-50	34	385	123
51-55	33	418	89
56-60	30	448	56
61-65	26	474	26
Total	474		

Table 2-4 also includes two columns that are called the cumulative frequency distributions. The first of these has the "less than" cumulative frequency distribution and tells us how many people are a given age or less. For instance, 54 people are 25 years old or younger; 197 people are 30 years old or younger. The "more than" cumulative frequency distribution tells us how many people are a given age or older. For instance, all 474 employees are at least 21 years old; 420 employees are 26 years old or older. (Note that when reading the less than cumulative distribution we use the upper limit of the interval; when reading the more than cumulative distribution we use the lower limit for a reference point.)

When your sample involves a hundred people (or cases) or more it is best to use percentages rather than raw frequencies. This allows for easy comparisons and is a method of standardization. Table 2-5 is equivalent to Table 2-4 except that the distributions are percentage distributions rather than distributions of the raw frequency data. In reading this table we would know that 11.3% of the employees are between 21 and 25 years of age and that 11.3% are 25 years old or younger. The percentages are given on the computer printout in the columns following the codes and frequencies. The first two columns of percentages (relative and adjusted) give the percentage of cases associated with each code. The cumulative % frequency is a "less than" percentage frequency distribution. When adding these percentages together one should always check to make sure that the computer has rounded the numbers so they do add to 100. If they do not, you will either want to note that fact or redo the computations to make the needed corrections.

Table 2-5 Age of Bank Employees

Years	Frequency %	"Less than" Cumulative Distribution	"More than" Cumulative Distribution
21-25	11.3	11.3 %	100.0 %
21-30	30.2	41.5	88.7
31-35	20.5	62.0	58.5
36-40	5.9	67.9	38.0
41-45	6.1	74.0	32.1
46-50	7.2	81.2	26.0
51-55	7.0	88.2	18.8
56-60	6.3	94.5	11.8
61-65	5.5	100.0%	5.5 %
Total	100%		

n=474

Finally, note the way in which the tables are labeled. Figure 2-1 contains instructions on the elements of a table that is properly constructed. These include labels for the table and each part of it. If percentages, as well as or instead of numbers, are used, you should make sure that enough information is given about the sample size so that the reader can reconstruct the actual numbers of people involved.

Table 2-6 gives yet another example of a frequency distribution. This involves two groups: Native American and non-Native American employees of the Bureau of Indian Affairs. The data examined are the grade level of employment. These grade levels are actually discrete variables, as opposed to the continuous variable of age. Note that when we have discrete variables we simply treat them as though they were continuous. (Some may argue that grade level is ordinal, rather than interval, but the levels correspond to pay increments, and at one time translated directly into dollars, so for the sake of example we will treat these data as measured on an interval scale.) Note too that these data are rounded to the next lower number. A person is in grade four until he or she moves into grade 5.

Note how the side-by-side arrangement of data for the two racial/ethnic groups helps in comparisons. (Remember that the lower grades are paid much less.) Most of the native Americans are at the lowest grades. The non-Native Americans are much more spread out and predominate at the higher grades. For instance, over half of the Native-Americans are in grades 3 and 4, but only 9% of the non-Native Americans are at that level. Almost one-fourth of the non-Native Americans are at grades 11 and 12 and one-third of the non-Native Americans are in grades 9 and 10. The comparable figures for Native Americans are 7% and 9% respectively. The cumulative distributions show similar results. 75% of all the Native Americans are at grade 6 or lower, but only 20% of the non-Native Americans fall in that range.

Table 2-6 Grade Level of Native American and Non-Native American Employees of the Bureau of Indian Affairs, 1970

	Native Americans		Non-Native Americans	
		~~(less than)~~(less than)		~~(less than)~~ (less than)
Grade	Frequency	Cum. Freq.	Frequency	Cum. Freq.
1	0.05	0.05	0.04	0.04
2	2.72	2.77	0.34	0.38
3	21.36	24.13	2.64	3.02
4	33.69	57.82	6.19	9.21
5	15.50	73.32	9.14	18.35
6	1.98	75.30	1.61	19.96
7	6.44	81.74	10.14	30.10
8	0.21	81.95	0.21	30.31
9	8.95	90.90	32.82	63.13
10	0.14	91.04	3.40	66.53
11	4.51	95.55	13.79	80.32
12	2.29	97.84	10.35	90.67
13	1.14	98.98	4.88	95.55
14	0.80	99.78	3.57	99.12
15	0.19	99.97	0.79	99.91
16	0.03	100.00	0.06	99.97
17	0.00	100.00	0.03	100.00
Totals	100.00%		100.00%	
n	5853		6697	

Source: Congressional Record, Dec. 14, 1970

In general, when constructing tables with quantitative data one would want about 10 to 15 intervals for easiest understanding. One usually uses equal-sized intervals, unless some of them contain very few people. For instance, there may be very few subjects with very high incomes or very low incomes in a sample and the intervals at these extremes may be made much larger or even open-ended (e.g. $75,000 +) to accommodate these people. Whenever one is comparing two groups, as in Table 2-6, it is important to use the same intervals for both groups so that one has valid comparisons. Also, when one is comparing two or more groups one would always use percentages, rather than raw frequencies, in order to have valid comparisons.

With qualitative data the procedures in table construction are basically the same as those described above, except that one does not have intervals, but instead categories. Table 2-7 gives a hypothetical example of a table with qualitative data, the distribution of religious affiliation for a sample.

Table 2-7 Religious Affiliation of Members of a Hypothetical Community

Religious Affiliation	Percentage
Protestant	55
Catholic	25
Jew	15
Other	5
Total	100
n	375

Graphs

Graphical displays of data are often a very helpful way to summarize and display the information provided in tables. The types of graphs appropriate for data depend on whether one's data are measured on an interval or ratio scale (quantitative data) or an ordinal or nominal scale (qualitative data). We will deal with graphs for both types of data in turn.

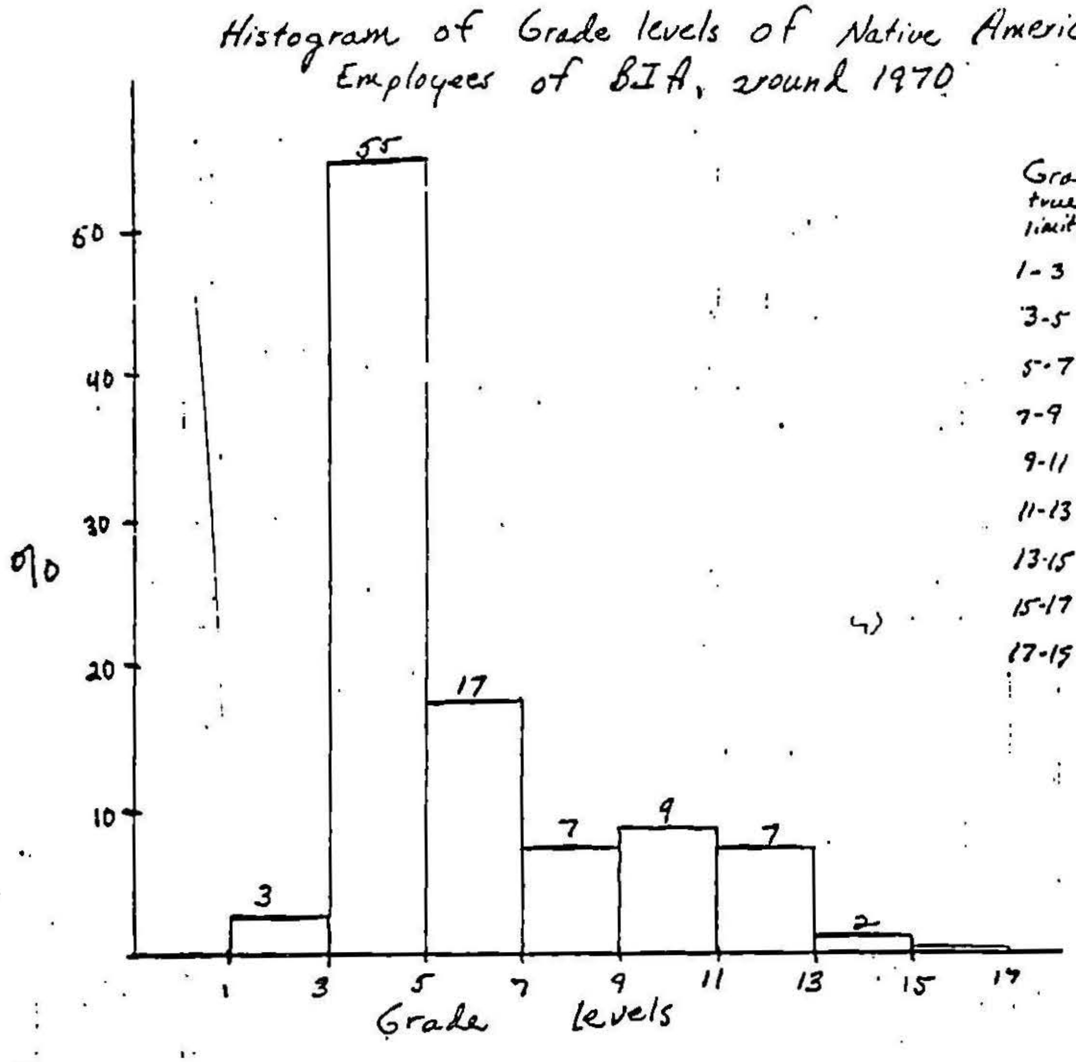
Graphs Appropriate for Quantitative Data

There are three basic graphs that are commonly used to represent quantitative data: histograms, frequency polygons, and ogives (or cumulative frequency graphs). Each of these has a common form in that along the horizontal axis the intervals for the distribution are graphed. These would be the same intervals that one has used in the table displaying the data, except that one would want to make sure that all the intervals were equal in size. That is, if one had doubled the size of some intervals in the table because they contained very few people, one would want to use the actual (uncollapsed) intervals in the graph. Along the vertical axis one plots frequencies or percentages, whichever one wishes to graph. When the sample size is large (over 100) one should use percentages. When comparing several groups percentages would also be more appropriate.

A histogram for data on grade-levels of Native American Employees of the BIA is shown in Figure 2-2. Note that the true limits of each interval are marked along the horizontal axis. Then within the boundaries of each interval a bar is drawn to the height that corresponds with the proportion of people in that interval. Thus, the height of the bar of the histogram for the first interval is at the 3% mark. The height of the bar for the second interval is at the 55% mark, and so on. Note that each bar of the histogram is adjacent to the next. This is because the variable, grade levels, is measured on an interval scale, and we are treating it as though it were continuous. (Intervals are collapsed from those shown in Table 2-6. Percentages used are given in Figure 2-2.)

A frequency polygon of grade levels of Native American employees and of grade levels of non-Native American employees is shown in Figure 2-3. The solid line gives data for the Native-Americans, the broken line gives data for the non-Native Americans. Note that again the base or horizontal axis includes the intervals of the variable grade levels. The percentages are placed along the vertical axis. With the frequency polygon one uses the midpoints of each interval and plots at the midpoint the percentage (or n if using raw data) of people who fall within that interval. Thus, the midpoint of the first interval is 2. For Native Americans the point is plotted to correspond with 2 on the horizontal axis and 3 on the vertical axis, indicating that 3% of the Native Americans fall in that category. For the second interval, the midpoint is 4. Corresponding to this point on the

Figure 2-2

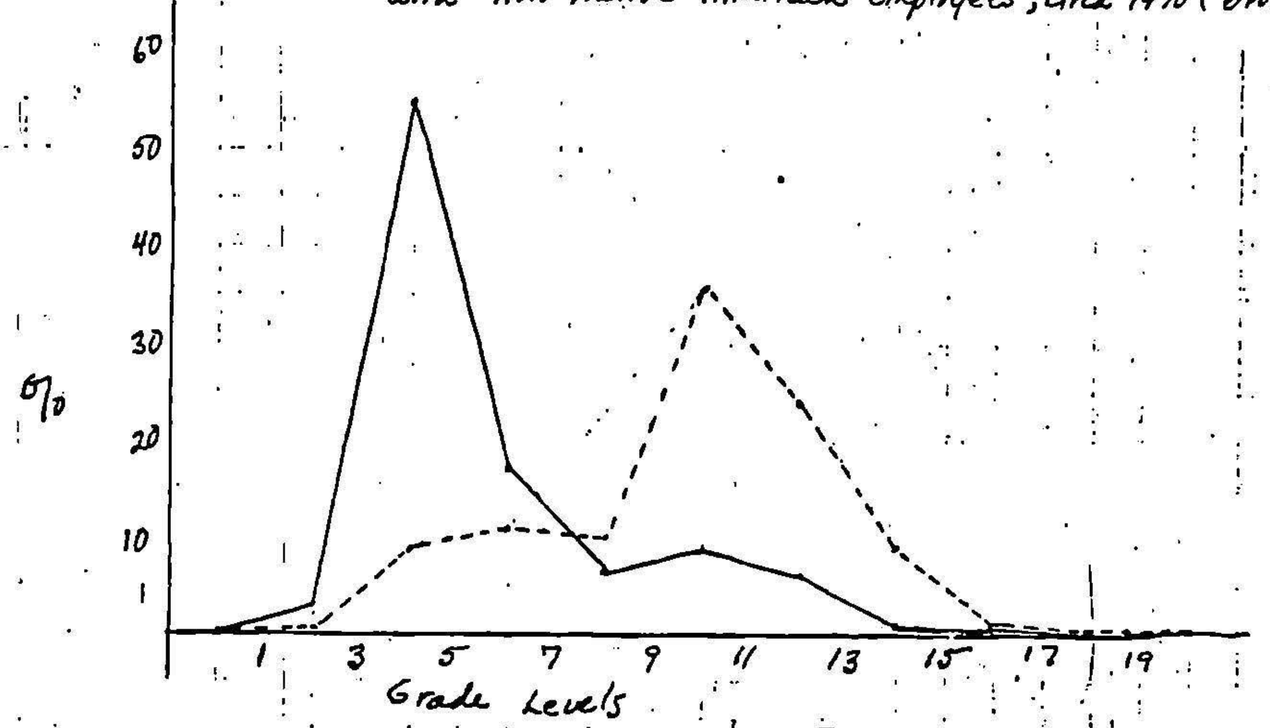


Histogram of Grade levels of Native American Employees of BIA, round 1970

Grade levels			%'s
true limits	rounded limits	NA	$\widetilde{NA}$
1-3	1-2	3	0.60
3-5	3-4	55	9
5-7	5-6	17	11
7-9	7-8	7	10
9-11	9-10	9	36
11-13	11-12	7	24
13-15	13-14	2	9
15-17	15-16	0.22	.86
17-19	17-18	0	0.03

Figure 2-3

Frequency Polygon of Grade Levels of Native American Employees of B.I.A, circa 1970 (solid line) and non-Native American employees, circa 1970 (broken line)



horizontal axis, a point is marked corresponding to 55% on the vertical axis for Native Americans and a point corresponding to 9% on the vertical axis is marked for the non-Native Americans. This process is continued. The points are then connected and the polygons are closed by plotting zero on the vertical axis at the midpoint of the interval that is theoretically below the first interval and the midpoint of the interval that is theoretically above the last interval.

It was mentioned briefly above that if one has uneven intervals in a table, one needs to be careful in transferring these data to a graph to ensure that one does not misrepresent the data. Figure 2-4 illustrates how one could do this. Three intervals are given in the data. The first two have a true interval width of 2 but the third has a true interval width of 4. Because we do not know the actual underlying distribution of these data (if we did we would use the true distribution for this third interval), we simply divide the subjects within the third interval evenly into two intervals the same width as the earlier ones. This is shown in the second table in Figure 2-4. (If the uneven interval had been three times the size of the other ones we would divide it into three parts, etc.) The data with equal intervals are then plotted. A second graph shows how one would incorrectly have graphed the data if one had not divided the subjects up among equal intervals. This incorrect graph shows a much greater proportion of subjects between 4.5 and 8.5 than in actuality are there.

Sometimes one will have data in a table that are open-ended. For instance, we will simply list the first category of income or age as all subjects at or below a certain point (≤5000 dollars, for example). At the upper end we might include all people who make above a certain amount of money (e.g. $50,000+). When graphing these data we clearly cannot continue the graph infinitely, so we must arbitrarily close it. At the lower end we would use zero, or whatever would be appropriate. At the upper end we would simply choose an arbitrary closing amount and then add a footnote to the table indicating that there were people in the last interval who made considerably more money or had considerably higher scores on the variable, but that this could not be represented on the graph.

One final point on graph construction: Sometimes your horizontal axis or interval scale will begin at a point considerably above zero. When drawing a graph for these data, if you wish to include a zero point on the axis, you could include a little break mark to indicate that a number of points were missing, as illustrated in Figure 2-5.

The decision of whether to use a histogram or a frequency polygon is often an esthetic one. For comparative purposes, as in Figure 2-3, the frequency polygon is often better. However, for exact representation of the data, a histogram might be preferable, for all of the data for a given interval are

Figure 2-4

Example of Adjusting uneven Intervals in Constructing a Histogram

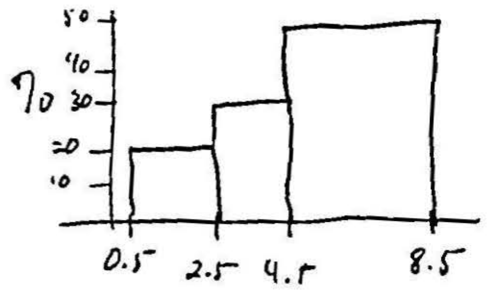
a)

True Interval Limits	%
0.5-2.5	20
2.5-4.5	30
4.5-8.5	50
	100%

b)

Adjusted True Interval Limits	%
0.5-2.5	20
2.5-4.5	30
4.5-6.5	25
6.5-8.5	25

Incorrect Histogram



Correct Histogram

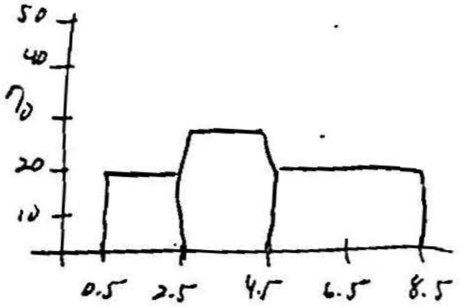
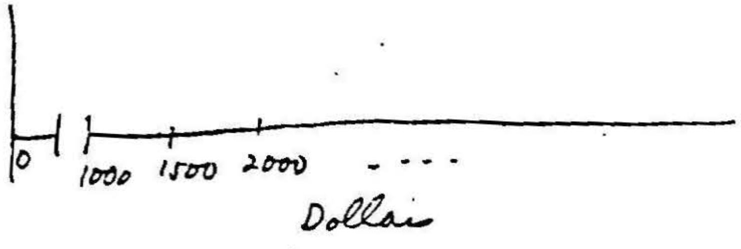


Figure 2-5

Example of Baseline for a Graph w/ Quantitative Data



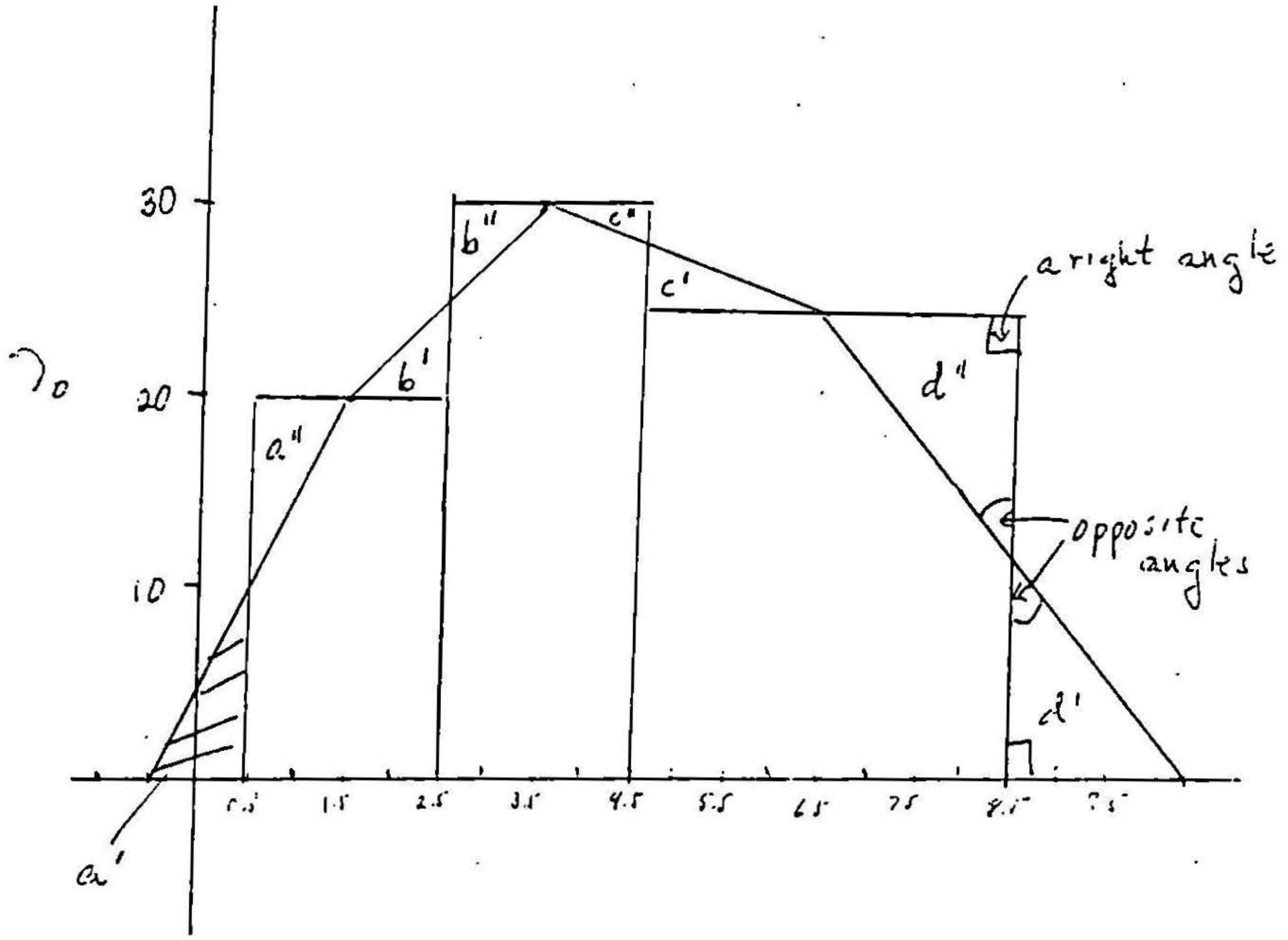
represented within that interval. The data for a given interval within a frequency polygon are actually spread across the area allocated to three intervals.

Nevertheless, the frequency polygon and the histogram both accurately reflect the data in that they both enclose the same amount of area. Consider the histogram drawn in Figure 2-6. This histogram and the associated frequency polygon for the data are produced below in Figure 2-6 superimposed on one another. Note that the polygon and histogram enclose the same area except for several triangles identified by letters. These triangles, however, are congruent to each other and thus hold the same amount of area. Consider the triangles labeled a' and a". The opposite angles are equal, the right angles are equal, and the distance from the base of the histogram bar to the midpoint of each interval is equal (1/2 i). Thus they have at least one equal side and two equal angles. This then implies that they have three equal sides and three equal angles and the two triangles are congruent. The area that is cut out of the histogram by the frequency polygon (a") is added onto the frequency polygon at another place (a'). The same argument could be made for all other pairs of triangles.

The ogive is a graph designed to represent cumulative frequency data. Again the intervals are displayed along the horizontal axis and the percentages (or frequencies if using raw data) are displayed along the vertical axis. One can have ogives for the "less than" and for the "more than" cumulative distributions. Both of these graphs are shown in Figure 2-7 for the data on BIA employees. In plotting points for the ogive one uses the end points of the intervals and one must think about what each distribution means. Consider first the "less than" distribution. 3% of the Native Americans are found at grade 2 or below. Thus, corresponding to grade 3 (the true upper limit of the first interval) the point is plotted at the line corresponding to 3% on the vertical axis. 58% of the Native Americans are in grade 4 or less, so the point is marked at the line corresponding to grade 5 (the true upper limit of this interval) on the horizontal axis and to 58% on the vertical axis. One then continues in this manner until one notes that 100% of the employees are found in grade 14 or lower and plot at the 100% point on the vertical axis at the points corresponding to 15 and to 17 on the horizontal axis.

For the "more than" distribution, the logic is somewhat different. 100% of the employees are in grade one or ~~below~~ above, so we plot a point that corresponds to 1 on the horizontal axis (the lower limit of the first interval) and 100% on the vertical axis. 97% of the subjects are in grade 3 or higher so we plot a point that corresponds to grade 3 on the horizontal axis (the lower limit of the second interval) and 97% on the vertical axis. To complete the graph each of the points plotted is connected.

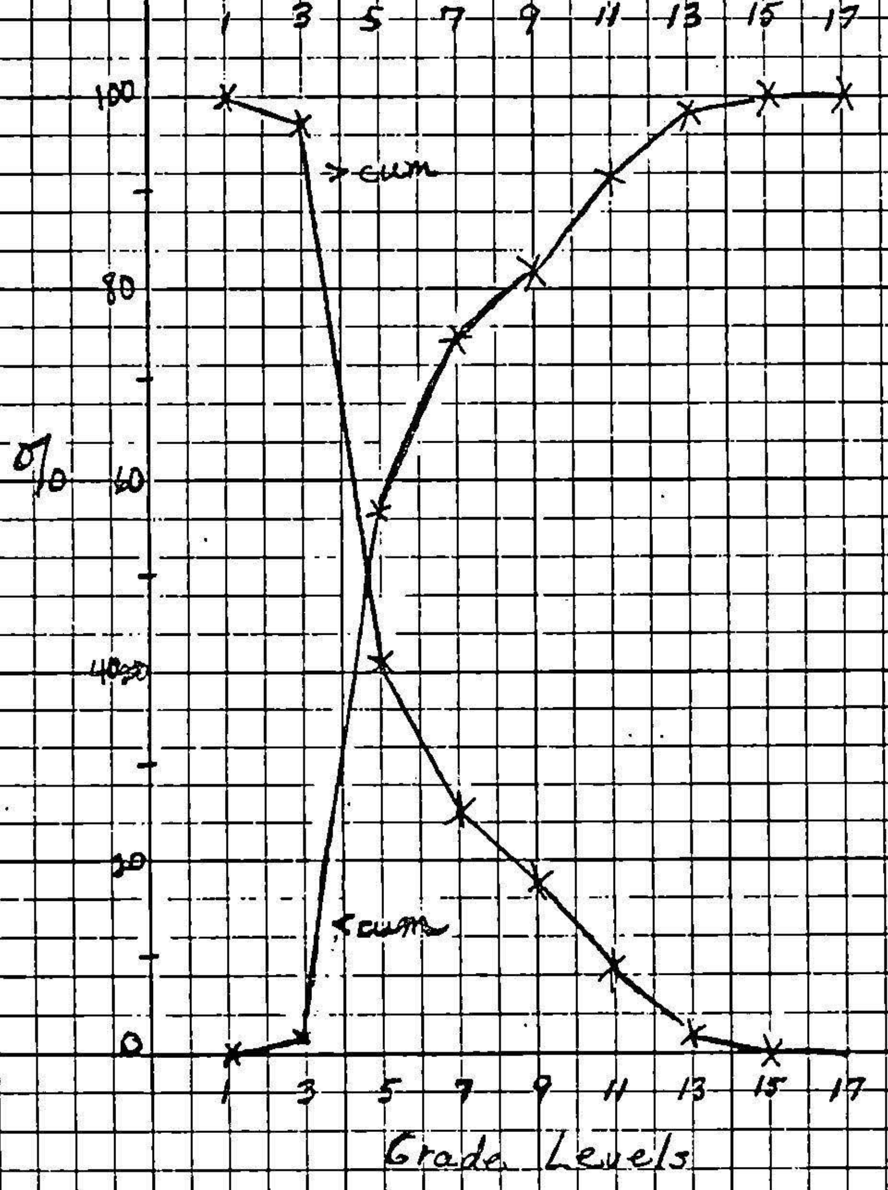
Figure 2-6


30
20
10
ηo
b''
c''
a right angle
c'
b'
a''
d''
opposite angles
d'
a'
0.5
1.5
2.5
3.5
4.5
5.5
6.5
7.5
8.5
9.5

Grade	True Limits	%	< cum	> cum
1-2	1-3	3	3(<3)	100(≥1)
3-4	3-5	55	58(<5)	97(≥3)
5-6	5-7	17	75(<7)	42(≥5)
7-8	7-9	7	82(<9)	25(≥7)
9-10	9-11	9	91(<11)	18(≥9)
11-12	11-13	7	98(<13)	9(≥11)
13-14	13-15	2	100(<15)	2(≥13)
15-16	15-17	0	100(<17)	0(≥15)

Figure 2-7

Ogives Representing Employment Grade Level of Native American BIA Employees



Once one has drawn the appropriate graph for one's data one would then examine it to see how it helps describe the data. For instance, in looking at Figure 2-2, the histogram of grade levels for the Native-American employees, one would note that over half of the employees are found in only two grade levels (those that correspond to aide and janitorial positions). The next highest category involves those in grades 5 and 6, low-level supervisory positions, but relatively few are in the higher level posts and almost none at the highest levels. In looking at Figure 2-3, with the frequency polygons for both racial groups, one could make similar conclusions regarding the Native-Americans and compare their distribution with that of the non-Native American employees. Here you could note the striking lack of overlap or correspondence between the two curves. Most of the Native Americans are at the lower grade levels, most of the non-Native Americans are at the higher grade levels. The two groups of employees appear to be in almost totally different job categories. One could continue with a more detailed examination of these differences, a task which would be good for students to pursue for practice.

In examining the ogive we can see how quickly or how slowly subjects increase or decrease on a certain variable. For instance, in looking at the "less than" distribution in Figure 2-7, we can see that there is a very steep slope, indicating that most of the subjects are included by the very lowest grade levels. The more than distribution also has a very steep slope indicating again that most of the subjects are found at the lowest levels. If one were to graph the ogive for the non-Native Americans (again a profitable exercise for students) one would find that the slope was much less steep, and informative comparisons could be made.

Besides the comparisons noted above, the ogive provides an easy way of finding what proportion (or how many, if using frequencies) of a group fall at or below a certain point. Conversely, we can also find out what point along the distribution or interval scale corresponds to a given percentage or frequency. For example, if we want to know approximately how many subjects have jobs at grade 10 or higher we would locate grade 10 on the horizontal axis and follow that point until we hit the graph. It then appears that about 13% of the subjects are at grade 10 or above. One could also ask what is the point at which we find 50% of the subjects with less than a particular grade and 50% with more. That is, what is the point on the scale that divides the group into two equal parts? One would then find 50% on the vertical axis and follow that line across. Note that this is the point where the "more than" and "less than" graphs cross. It appears that this point corresponds to approximately grade 4.8. If we are interested in the 25% mark, the first quartile, we may follow this line across and find that 25% of the subjects appear to be at grade 3.8 or less (approximately) and that 25% of the subjects appear to be at grade 7 or higher.

students should also continue this exercise on their own until they feel confident in interpreting this graph.

Graphs Appropriate for Qualitative Data

There are a number of graphs that are used with qualitative data. We will focus on bar charts, which are the most common. You may consult various statistics texts for examples of other types. As with the quantitative data, the bar charts are designed to display the data found in the tables in a way that pictorially summarizes the data.

The basic form of the bar chart involves a base line on which the categories of the variables are labelled. Note how the form is different from the histogram. With bar charts there are spaces between each of the categories because we are not dealing with interval data, but with categoric data. The second dimension of the chart involves either percentages or frequencies as with the quantitative data graphs. The length of the bars represents the frequencies or percentages within a given category. The bars may be displayed either vertically or horizontally, depending on the researcher's desires. With ordinal data one would usually want to have the categories in the relevant order. With nominal data one might have an order of the categories that is theoretically important or one might want to display the data in order of frequency of occurrence (e.g. smallest to largest).

Many varieties of bar graphs are possible. It is also possible to use a bar graph to display data for more than one group. Figures 2-8 through 2-10 display the data shown in Table 2-8 on the type of descent system common in three different types of economies. (You might remember from your introductory research methods class that tables are percentaged in categories of the independent variable. We are assuming here that the economy of a society is the independent variable and that the type of descent system that a society adopts depends on the economic system of that society.)

Figure 2-8 is a regular bar graph such as the general case described above, but includes data for the three different types of societies. The first sub-graph includes data for the hunting societies. It is apparent that in these societies matrilineal descent systems are most common, followed by bilateral and then by patrilineal descent systems. The second sub-graph gives the data for societies with a pastoral economy. These are most likely to have patrilineal descent systems, bilateral systems are much less common and matrilineal descent systems are relatively rare. Among agricultural societies matrilineal and patrilineal descent systems are about equally likely to occur and bilateral descent systems appear less frequently. Because we have data for the three types of societies here we can also make comparisons across the three types of societies (among the three categories of the independent variable - type of economy). It is apparent

Figure 2-8

Example of Bar Graph for Data in Table 2-8

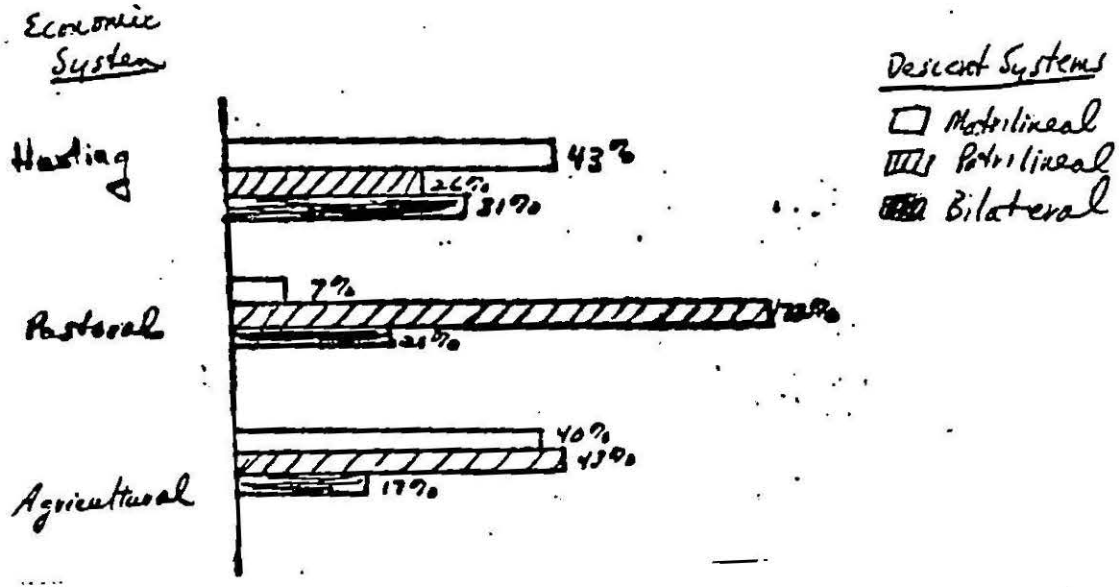


Figure 2-9

Example of a sliding bar graph for data in Table 2-8

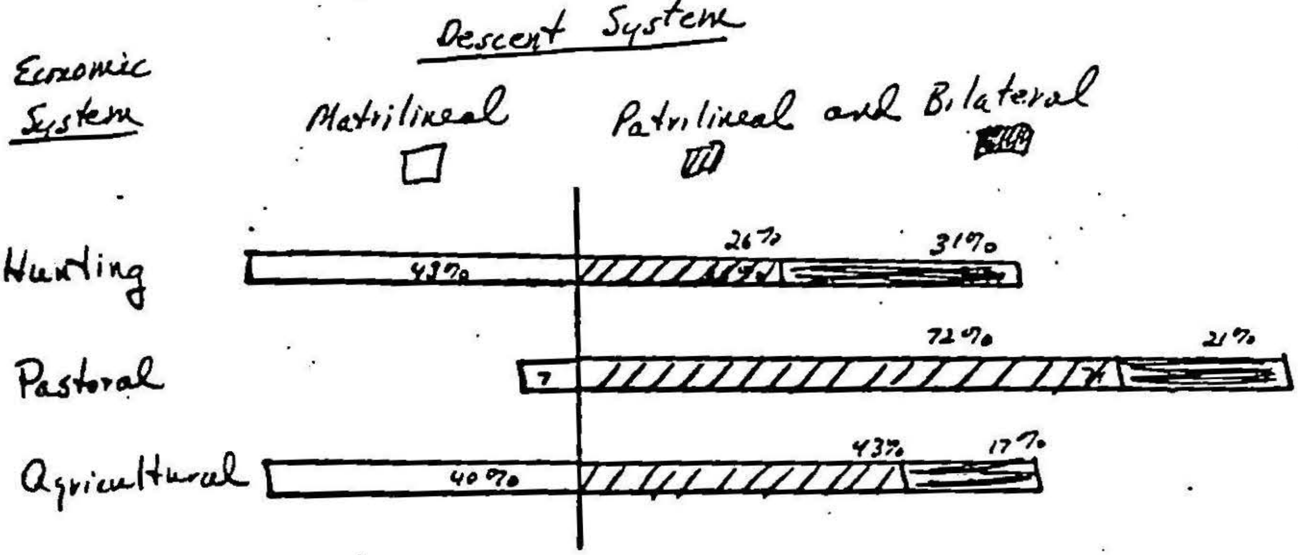
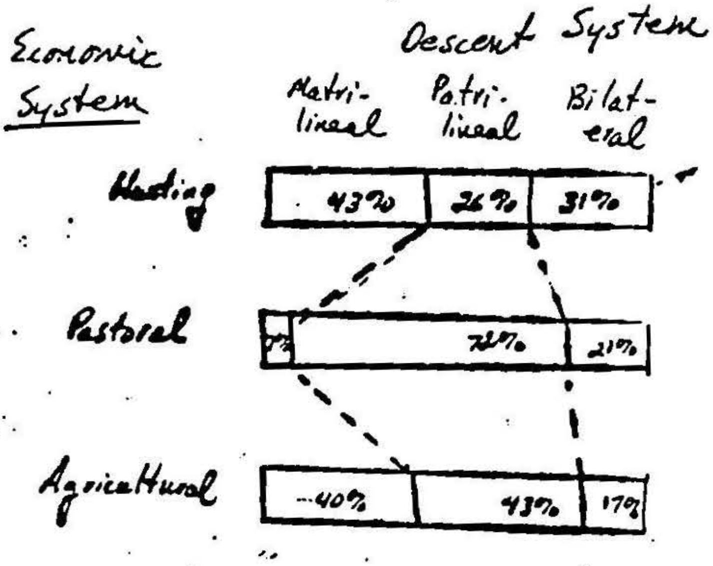


Figure 2-10

Another Example of a Bar Graph with data from Table 2-8



that matrilineal descent systems are about equally likely to occur in hunting and agricultural societies, but only rarely in pastoral societies. Patrilineal descent systems most often occur in pastoral economies, next most often in agricultural economies and least often in hunting societies. Bilateral descent systems occur most often in hunting groups, next most often in pastoral groups and least often in agricultural groups.

Table 2-8 Descent Systems Found in Societies with Different Economic Bases

Economic System	Type of Descent System Matrilineal	Patrilineal	Bilateral	Total
Hunting	43	26	31	100%(70)
Pastoral	7	72	21	100%(14)
Agricultural	40	43	17	100%(110)

(Source: adapted from Mueller, et al, 1977; p. 47)

Figure 2-9 gives a version of a sliding bar graph. This type of graph is most useful when we want to distinguish between two types of attributes of the dependent variable. For instance, in Figure 2-9 we are distinguishing between matrilineal descent systems and the other two types. Within each economy (or subgraph) we have represented the family types on a long bar, all of equal length. These bars are then divided into segments to represent the different family types. Shading is used, as in Figure 2-8 to represent the different types of descent systems. A vertical axis is drawn down the middle of the graph to separate the matrilineal and other descent types. The various graphs are then "slid" to the left or the right to represent the proportion of societies within each group that have matrilineal descent systems. Clearly the pastoral societies are least likely to have this type while the hunting and agricultural societies appear about equally likely to have this type of system. One could have constructed this type of graph with either of the other types of descent systems as the focus of interest, depending on one's theoretical point.

Figure 2-10 gives another way of using bar graphs. Here again the relative representation of descent systems within each society is represented on a bar. A separate bar is drawn for each society. Then to demonstrate the comparisons between the three types of societies dotted lines connect the various categories. These illustrate how the representation of matrilineal types is much larger in hunting and agricultural societies, for instance, than in pastoral societies.

A number of computer packages offer options for graphs. You should be very careful in using these options. They are quite nice if you have the appropriate data and it is coded and input in a way that you want. If not, however, the results are useless and often misleading. Therefore, you should think very carefully before automatically using material that a computer has spewed out in graph form. You also must be very careful when using a graphics program with a micro computer to ensure that the graphs are correctly drawn.

Measures of Central Tendency

While tables and graphs illustrate the dispersion of data and where most subjects or cases tend to be, they do not provide a single summary statistic of the location of most of the people. Measures of central tendency are designed to provide such a summary. Three measures of central tendency are commonly used: the mode, the median, and the mean.

The Mode

The mode is simply the most frequently occurring value or point. We can use the mode when talking about qualitative data if we refer to the modal category. For instance, in Table 2-7 we could say that the modal category is Protestant; it is the category with the greatest number of people. We can simply count the number of cases that have each attribute and find which attribute has the most cases associated with it.

With quantitative data we must go beyond this simple counting procedure and would like to find the value within an interval (assuming that our data have been grouped into intervals) that corresponds to the modal point. There are two ways of doing this. The first is called the crude mode. The crude mode is simply the midpoint of the interval that has the largest number of cases in it. For instance, with the data on BIA employees that is again presented in Table 2-9, the modal interval for Native Americans is that with true limits 3 and 5. The midpoint of this interval is 4.0, and this is the crude mode. Students should verify that they understand this by demonstrating that the crude mode for the non-Native Americans is 10.0.

The second way of computing the mode with grouped data results in what is called the refined mode. The refined mode is an adjusted value that is based on the relative size of the frequencies in intervals adjacent to the modal interval. It is based on the idea that the true place of greatest density (the true location of the mode in an interval) will be closer to the interval with a higher frequency. The larger one adjacent interval is than the other, the more that the mode will be shifted toward that larger interval. The formula for the refined mode is given below in equation 2-1.

$$\text{Refined Mode} = L + \left(\frac{D_1}{D_1 + D_2} \times i \right) \qquad (2-1)$$

where L = the true lower limit of the modal interval;
D_1 = the difference between the frequency in the modal interval and the frequency (number or % of cases) in the next lower interval;
D_2 = the difference between the frequency in the modal interval and the frequency in the next higher interval; and
i = the width of the interval.

Computations in Table 2-9 show that for the Native Americans the refined mode = 4.16. For the non-Native Americans the refined mode is equal to 10.37.

Examining Formula 2-1 more closely it may be seen that when $D_1 = D_2$, that is when the two adjacent intervals have the same number of cases, the refined mode equals the crude mode. In this case we would add one-half of the interval width (i) to the lower limit of the interval, thus being at the midpoint of the interval.

If the adjacent lower interval had more people than the adjacent higher interval, D_1 would be less than D_2. That is, the size of the next lower interval would be closer to the modal interval than would the next higher interval. When D_1 is less than D_2, $D_1/(D_1 + D_2)$ is less than one-half and the refined mode would be smaller than the crude mode (i.e. not as large as the midpoint of the interval). When, however, the next higher interval has more cases, $D_1 / (D_1 + D_2)$ would be greater than 1/2 and the refined mode would be larger than the crude mode.

Graphically the mode appears as the high point of the graph. On the frequency polygon, the mode would be the highest point, the scale point that corresponds to the highest frequency or percentage found in any category of the data. Sometimes there will be more than one high point. We say then that a distribution is bi-modal if there are two high points or trimodal if there are three. This can result if there are basic divisions within the group. For instance, if we were to graph the grade level of all BIA employees, combining the two groups in Table 2-9, we might well have a bi-modal distribution. This, however, would be because the two racial groups have very different job level distributions.

Table 2-9 Example of Computing Mode and Median with BIA Data

Grade Levels		Percentage		Cumulative %	
Rounded Limits	True Limits	Native Americans	non-Native Americans	Native Americans	non-Native Americans
1-2	1-3	3	0	3	0
3-4	3-5	55	9	58	9
5-6	5-7	17	11	75	20
7-8	7-9	7	10	82	30
9-10	9-11	9	36	91	66
11-12	11-13	7	24	98	90
13-14	13-15	2	9	100	99
15-16	15-17	0	1	100	100
	Total	100%	100%		

Native Americans

crude mode = 4.0 = 3+1

$$\text{refined mode} = 3+\left[\frac{(55-3)}{(55-3)+(55-17)} \times 2\right]$$

$$= 4.16$$

$$\text{Median} = 3.0+\left[\frac{\frac{100}{2} - 3.0}{55} \times 2\right]$$

$$= 3.0 + \left[\frac{47}{55} \times 2\right]$$

$$= 3.0 + 1.71 = 4.71$$

non-Native Americans

crude mode = 10.0 = 9+1

$$\text{refined mode} = 9.0+\left[\frac{(36-10)}{(36-10)+(36-24)} \times 2\right]$$

$$= 10.37$$

$$\text{Median} = 9.0+\left[\frac{\frac{100}{2} - 30}{36} \times 2\right]$$

$$= 9.0 + \left[\frac{50-30}{36} \times 2\right]$$

$$= 9.0 + 1.11 = 10.11$$

The SPSS subprogram FREQUENCIES gives the mode as one of its statistics. This is not a refined mode or a crude mode, for the program assumes that the actual values are given as input, at least for this statistic. If you have data that are coded in intervals and input in such a way you would probably want to compute the refined or crude mode yourself. Also, if you have bimodal (or multi-modal) data, SPSS will not tell you this. Instead, it will automatically assign the mode to the lowest value on your scale or variable that has the highest frequency. (Say you are studying age and 35 people fall at ages 29, 39, and 49, SPSS will report only 29 as the mode. You will have to inspect the data to find the other modes.)

The mode has certain advantages. It can be used with qualitative data. It is easy to calculate and it can be easily related to a graph. However, the mode does have certain disadvantages. It generally cannot be used in further calculations. While this is often not a problem with qualitative data, it can be a real disadvantage with quantitative data. The mode is also unstable and can be greatly influenced by how large the intervals are in a data set. Third, the mode is nonspecific. We don't know "how modal" a certain point is. We know from the mode what value most often occurs, but we don't know if this point occurs twice as often as all others, or just a tiny bit more often.

The Median

The median is a position average and is defined simply as the point in the distribution where one-half of the cases are above and one-half are below. It is strictly suitable only for variables measured on an interval or ratio scale, but it is sometimes used with variables measured on an ordinal scale. With an ordinal scale, however, we can only talk about the median category, the category in which the median is found.

To compute the median with ungrouped data we simply arrange the data in order from the smallest to the largest and then take the middle case. If there are an even number of cases, as in Table 2-10 below, this would be the point halfway between the two middle points, as shown. If there are an odd number of cases, as in Table 2-11, we would use the point exactly in the middle, as shown. Table 2-12 gives an example with ordinal data. Here the median category is that of mild support.

Very often we don't have ungrouped data, we have data that have been grouped into intervals. Here we can find the median interval by examining the cumulative frequency distribution. But, as with the mode, we still must determine the point within that interval where the median falls. To do this we assume that the cases are evenly spread throughout the interval (note how this differs from the assumption involved in computing the refined mode where we assume they are more grouped toward the

Table 2-10 Example of Computing Median With an Even Number of Cases

Ages of People Referred to Clinic

5 cases	6	13	5 cases
	7	17	
	8	19	
	9	21	
	11	22	

$$\text{Median} = \frac{11 + 13}{2}$$

$$= \frac{24}{2} = 12$$

Table 2-11 Example of Computing Median with an Odd Number of Cases

Ages of People Referred to Clinic

4 cases	6		13	4 cases
	7	11	17	
	8	↑ median value	19	
	9		21	

Median = 11

Table 2-12 Degree of Support Respondents Report for President

	%	
Highly Supportive	20	
Mildly Supportive	40	←Median Category
Neutral	10	40%
Mildly Unsupportive	10	
Highly Unsupportive	20	
Total	100%	

adjacent interval with more cases). We then see how far we need to go within that interval to get to the median point. For instance, in Table 2-13 below, an imaginary distribution, there are 189 cases in all. The median case would be 189/2 = 94.5, or between the 94th and 95th case. We can see from examining the cumulative frequency distribution that this occurs in the interval with the true limits of 4,950 and 5,950. There are 51 cases in this interval and at the beginning of the interval we have 81 cases. To get to the 94.5th case we must go 13.5 cases beyond the lower limit of the interval. Since there are 51 cases in all in the interval we must go through 13.5/51 cases or about 26.5% of the total interval. The interval here is 1000 wide, so 26.5% of 1000 is 265. If we add 265 to the lower limit of the interval we have 4950 + 265 = 5215, and this is the median.

Table 2-13 Imaginary Income Data

True Limits	Frequency	Cumulative Frequency
1,950-2,950	17	17
2,950-3,950	26	43
3,950-4,950	38	81
4,950-5,950	51	132
5,950-6,950	36	168
6,950-67,950	21	189

In general, the formula for the median is

$$\text{Median} = L + \left[\frac{N/2 - cf}{f} \times i\right] \qquad (2\text{-}2)$$

where L is the true lower limit of the interval containing the median, N/2 is one-half of the total sample size; cf is the cumulative frequency at the beginning of the median interval; f is the frequency in the median interval; and i is the width of the interval.

For the example above,

$$\text{Median} = 4950 + \left[\frac{(94.5 - 81)}{51} \times 1000\right] = 5215 \qquad (2\text{-}3)$$

A procedure just like that outlined above is used with percentages except that we are looking for the 50th percentile (or N/2 = 50). Table 2-9 gives an example of finding the median for the BIA data. Students should work through these examples to make sure they are familiar with the procedure. If you have discrete data you simply, as before, treat it as though it were continuous. (A good example would be data on family size.)

The formula for a median can also be used to compute other position measures. The most common ones are quartiles (25%, 75% points), deciles (10%, 20%,...), and centiles (1%, 2%, etc.). While you can use the cumulative frequency graph (ogive) to approximate these positions you can use a variation of the median formula to get the exact value. All one does is alter the N/2 part of formula 2-2. For instance, if one is interested in the first quartile, the 25% point, one would want to look at N/4 instead of N/2. For the third quartile, the 75% point, one would want to look at 3N/4 instead of N/2. For the third decile one would examine 3N/10, and so on. Below, examples of computing various other positions are given using the BIA data.

1st quartile — Native Americans — non-Native Americans

$$Q_1 = L + \left[\frac{N/4 - cf}{f} \times i\right] \qquad 3.0 + \left[\frac{25-3}{55} \times 2\right] \qquad 7.0 + \left[\frac{25-20}{10} \times 2\right] \qquad (2\text{-}4)$$

$$= 3.8 \qquad\qquad = 8.0$$

3rd quartile

$$Q_3 = L + \left[\frac{\frac{3N}{4} - cf}{f} \times i\right] \qquad 5.0 + \left[\frac{75-58}{17} \times 2\right] \qquad 11.0 + \left[\frac{75-66}{24} \times 2\right] \qquad (2\text{-}5)$$

$$= 7.0 \qquad\qquad = 11.75$$

1st Decile

$$O_1 = L + \left[\frac{\frac{N}{10} - cf}{f} \times i\right] \qquad 3.0 + \left[\frac{0-3}{55} \times 2\right] = 3.25 \qquad 5.0 + \left[\frac{10-9}{11} \times 2\right]$$

$$= 5.18$$

Position measures such as the above are commonly used in comparisons of individuals (e.g. SAT scores, GRE's, height and weight percentile placements for children, etc.)

Position measures have certain disadvantages as well as advantages. They cannot be used in algebraic manipulations and thus have limited utility for use in more advanced statistical manipulations. The median however is quite stable. It is not affected much by extremes and is usable with open-ended data. It is commonly used in describing income distributions because it is so unaffected by extreme cases. Graphically, the median is the point where the less than and more than cumulative frequency distributions cross (See Figure 2-7).

The median is part of the output given on the subprogram FREQUENCIES by SPSS. When computing the median SPSS assumes that data are grouped into intervals with an interval width of 1. It then uses the type of formula described above to find the median point with the interval.

Very often you will have data that have been grouped and have been coded with these groups. In Table 2-14 are the codes in the National Opinion Research Survey data for income that is self-reported. Note that the categories are quite large. Also, note that they have been coded. If SPSS were to report the mode for this data it would give the value as 9. If it were to report the median, it would give the value as 6.61. Clearly, these values are not correct. One solution would be to recode the data within the computer (a minor procedure) to reflect the midpoints of each interval (1 would become $500; 2 would become $2,000; etc.). The mode would then be given as "12,500"in the SPSS output. The median would then be computed within the interval of one dollar around the value of $7500. In deciding what step to take, you would have to consider what purpose these various statistics would have for you. To have the most accurate results you should compute the median by hand using the full interval width of $1000.

Table 2-14 Example of Income Data from an NORC Survey

41. Did you earn <u>any</u> income from (JOB DESCRIBED IN Q. 11) in 1973?

Yes . . . . . . ( )
No . . . . ( )

A. <u>IF YES</u>: In which of these groups did your earnings from (JOB IN Q. 11), for the last year--1973 fall? That is, before taxes or other deductions. Just tell me the letter.

	COLS. 38-39	
RESPONSE	PUNCH	N
Under $1,000	01	69
$ 1,000 to 2,999	02	116
$ 3,000 to 3,999	03	49
$ 4,000 to 4,999	04	67
$ 5,000 to 5,999	05	64
$ 6,000 to 6,999	06	48
$ 7,000 to 7,999	07	57
$ 8,000 to 8,999	08	89
$10,000 to 14,999	09	155
$15,000 to 19,000	10	60
$20,000 to 24,999	11	30
$25,000 or over	12	35
Refused	13	37
Don't know	98	15
Not applicable	BK	593

The Mean

The arithmetic mean or the arithmetic average is probably the most common measure of central tendency. It is usable only with variables measured on an interval or a ratio scale. Conceptually we should see the mean as the arithmetic average. If we think of all the cases in a distribution as spread out along a graph, such as a frequency polygon, the mean would be the center of gravity, the place along the base line that would be the balancing point for the distribution.

The formula for the mean is simply:

$$\bar{X} = \frac{\Sigma X_i}{n} \qquad (2\text{-}8)$$

where n = the size of the sample,
X is the mean,
X_i refers to each individual value of S, and
ΣX_i refers to the sum of all of the values of X_i

The mean is used in many advanced statistics and its usefulness derives from the fact that it is the "center of gravity" of a distribution. More specifically, the mean is the only value from which the sum of all deviations of scores will balance out or equal zero. That is, if we examine the deviations of all scores in a distribution from the mean and add up these deviations, we will find that the sum equals zero. This means that the sum of the deviations of scores around the mean is lower than the sum of the deviations would be around any other value.

Table 2-15 illustrates this quality of the mean. Note that the mean of the distribution is 11. The median of the distribution is 9. The sum of the deviations around the mean is zero. The sum of the deviations around the median is 12. Students may try substituting other numbers and will discover that only the mean will produce the sum of zero in adding deviations.

Table 2-15 Example of Computing Deviation Around the Mean

	Ages Referred to Clinic	$X-\bar{X} = x$	X - Md
	6	6-11 = -5	6-9 = -3
	7	7-11 = -4	7-9 = -2
	8	8-11 = -3	8-9 = -1
	10	10-11 = -1	10-9 = 1
	16	16-11 = 5	16-9 = 7
	19	19-11 = 8	19-9 = 10
Totals	66	0	+ 12

$\bar{X} = \frac{66}{6} = 11$

Median = 9

Table 2-15 showed how one would compute the mean if there were only one case with each value. If there is more than one case with a value in a distribution, as in Table 2-16, the computation of the mean is again quite simple. We simply multiply the frequency (or number) of cases with each value times that value and add up all the products. For instance, in Table 2-16 below, instead of adding 6+6+7+7+7+....we add 2(6) + 3(7) +....

The general formula is

$$\bar{X} = \frac{\Sigma f X_i}{n} \qquad (2\text{-}9)$$

where $\bar{X}$ is the mean,
f_i is the frequency associated with each value,
X_i is each value of the variable X
and n is the sample size.

Table 2-16 Example of Computing Mean with Grouped Data

X	Frequency (f)	fx
6	2	12
7	3	21
9	1	9
10	3	30
12	2	24
15	1	15
Total	12	111

$$\bar{X} = \frac{\Sigma f X_i}{n} = \frac{111}{12} = 9.25$$

If we have discrete data rather than continuous data we simply assume that our data are continuous and proceed as above.

If our data are grouped into intervals we use the same procedure as in Table 2-16, but we use the midpoint of the interval in computing the mean. The relevant formula is given below:

$$\bar{X} = \frac{\Sigma f X_i}{n} \qquad (2\text{-}10)$$

$\bar{X}$ is the mean,
f is the number of cases in each interval,
N is the sample size, and
X_i is the midpoint of the interval.

Table 2-17 gives an example of computing the mean with the grouped data on the job levels of BIA employees.

Table 2-17 Computation of Mean for BIA Data

Grade Level	Frequencies (%)		fx	
Midpoint of	Native	non-Native	Native	non-Native
Intervals	Americans	Americans	Americans	Americans
2	3	0	6	0
4	55	9	220	36
6	17	11	102	66
8	7	10	56	80
10	9	36	90	360
12	7	24	84	288
14	2	9	28	126
16	0	1	0	16
Totals	100	100	586	972

Native Americans

$$\bar{X} = \frac{\Sigma f X_i}{n} = \frac{586}{100} = 5.86$$

non-Native Americans $\bar{X} = \frac{\Sigma f X_i}{n} = \frac{972}{100} = 9.72$

Before the days of computers and inexpensive calculators with memories we used fairly complex methods of computing the mean with grouped data. These methods were designed to reduce errors when using large numbers and doing lengthy hand calculations such as multiplying frequencies by interval midpoints. Now that we have very cheap calculators with extensive memories these older techniques are not all that useful. To compute a mean with a calculator you could simply use the actual midpoint of the interval and formula 2-10 given above. SPSS uses formula 2-10 in computing the mean also.

As long as you have submitted the actual raw data into the computer there will be no problem with SPSS using formula 2-10. However, if you have put in your data coded in some manner, such as the NORC data on income shown in Table 2-14, you must be careful in interpreting the results. With the codes given in Table 2-14, the computer would tell you that the mean for the data is 6.17. You would want to instead tell the computer to regard each code as the midpoint of the interval. You could do this with a RECODE command, as in RECODE VAR22 (1

If we know the mean, median, and mode for a set of quantitative data we can draw a rough diagram of the frequency distribution or frequency polygon. We know that the mode represents the highest point of the graph, the median represents the halfway point, and the mean is the center of gravity. Because the mean is more affected by extreme points than the median is, we can tell the nature of skew (unevenness) in the distribution by examining their relative values. If the mean is greater than median, the distribution has a positive skew, as in Figure 2-11. If the mean is smaller than the median, the distribution has a negative skew as in Figure 2-12. If the mode, median, and mean are equal, we have a symmetrical distribution, as in Figure 2-13. Finally, Figure 2-14 illustrates the situation where two distributions have identical means, but unequal modes and medians. This illustrates the importance of examining all three measures of central tendency when you have the appropriate level of measurement and the usefulness of graphing data.

Figure 2-11 Example of a Positively Skewed Distribution

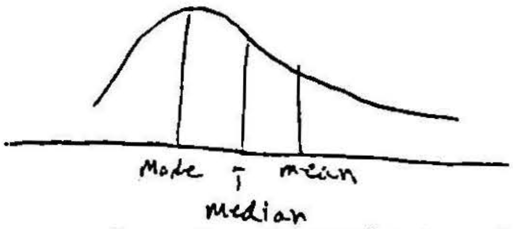


Figure 2-12 Example of a Negatively Skewed Distribution

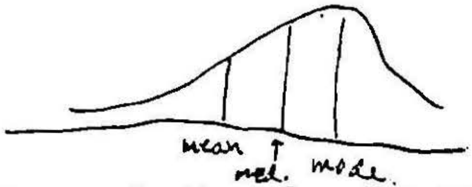


Figure 2-13 Example of a Symmetrical Distribution

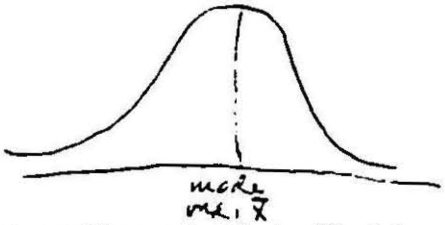
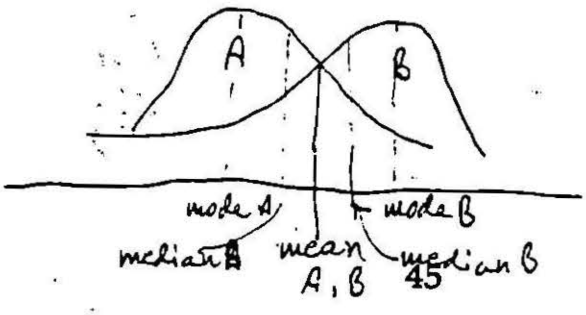


Figure 2-14 Example of Distributions with Equal Means and Unequal Medians and Modes



= 500   2 = 1500) . . . . The machine would then use these recoded values in computing the mean and would tell you that $6684 was the mean.

Sometimes you will want to combine the means from several groups. How you combine these means depends on your purpose, what you want to accomplish. You might want to have the average (mean) of the groups. That is, if you are looking at the average GPA's of students in various schools and college in the university, you might want to know the average GPA of these schools. Your unit of analysis is the school or college. Then you would simply add up the averages for each of these schools and compute the average of these averages. This is shown in part a of Table 2-18.

Table 2-18 Combining Means from Several Groups

School or College	Mean GPA= $\frac{\Sigma f X_i}{n_i}$	ni	$n_i \bar{X}_i = \Sigma f_i X_i$
Journalism	2.9	30	(30)(2.9) = 87
P.E.	2.8	40	(40)(2.8) = 112
Education	2.7	60	(60)(2.7) = 162
AAA	3.2	40	(40)(3.2) = 128
CAS	3.10	100	(100)(3.1)= 310
Totals	14.7	270	799

a) $\bar{X} = \frac{14.7}{5} = 2.9$ (unit of analysis is the school or college)

B) $\bar{X} = \frac{799}{270} = 2.96 = 3.0$ (unit of analysis is the individual)

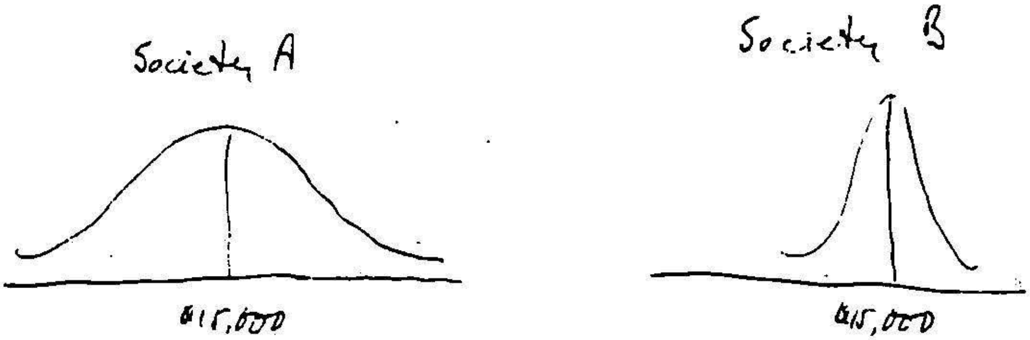
The Mean, Median or Mode?

Finally, how do we decide which measure of central tendency to use? We would want to consider the level of measurement of our data, for some are appropriate for some types of data only. We would also want to consider what we want to know about our data. We would also want to consider the shape of our data. If we have a lot of extreme values then the mean might be a less accurate summary measure of the central tendency than the median, for it is more affected by extreme values. If we have a flat distribution, with no clear modal value, the mode might be very misleading. Finally, if we want to make further arithmetic calculations, the mean is usually the most useful statistic to have. Note that computer programs commonly give all three statistics, so the researcher must decide which ones to report.

Measures of Dispersion

To this point we have been discussing measures of central tendency, statistics that describe where most people are. However, we aren't always interested in these "central" points. Sometimes we might be interested in the furthest ranges - e.g. How much money do the richest people make? How poor are the poorest people? Or we might be interested in how spread out a distribution is. Consider the two income distributions graphed in Figure 2-15 below. In society A the mean income is $15,000 and in society B the mean income is also $15,000. But in society A people are much more spread out around the mean than in society B. Which society would you rather take your chance of living in? Your decision would be much more informed if you knew not just the central tendency of the distribution but also had some idea of its dispersion. That is what we will look at now. We will first look at a measure of dispersion appropriate for qualitative data; then explore measures useful with quantitative data: the range, average deviation, variance and standard deviation; and finally examine a measure that incorporates both measures of central tendency and measures of dispersion, the coefficient of relative variation.

Figure 2-15 Hypothetical Income Distributions in Two Societies



The Index of Qualitative Variation

Because qualitative variables have no magnitude associated with them, they are categoric, we cannot examine dispersion as the amount of distance from a set measure of central tendency (as we will do below). Instead, we look at how variable -- or how different -- are the cases in a given data set on the variable of interest. Consider the distribution of the hypothetical sample in Table 2-19 below. In part a the cases are distributed evenly among the four religious categories. In part b of Table 2-19, the cases are all within one category of the religious affiliation variable. The subjects are much more diverse or varied in their religious affiliation in part a of the table than in part b. We would say then that the variation for subjects in part a is greater than the variation for subjects in part b. In fact, since the subjects are equally distributed among the four categories in part a, they show as much diversity as they possibly could. That is, their diversity is at a maximum. Since the subjects in part b are all grouped into one category, they show the least diversity that they possibly could and we would say that their diversity is at a minimum.

Table 2-19 Hypothetical Data on Religious Affiliation of 3 Samples

Religious Affiliation	a	b	c
Protestant	25	100	40
Catholic	25	0	30
Jew	25	0	20
Other	25	0	10
Totals	100%	100%	100%

The Index of Qualitative Variation (IQV) has the very nice quality of reporting this amount of diversity in a proportion. When a measured variable has the maximum variation or diversity possible, the IQV = 1.00. When the variable shows no diversity whatsoever, the IQV = 0.

To compute the IQV one determines how many differences - or how diverse - a set of cases could possibly be. That is, one computes the maximum number of differences among cases within a data set. This is called S_m. One then examines the actual variation in one's data set. This is called the observed differences and is called S_o. The IQV is then the ratio of these observed differences to the maximum possible number of differences:

$$IQV = S_o / S_m \qquad (2\text{-}11)$$

To compute the number of observed differences one multiplies every category frequency by every other category frequency and sums these products. This is represented by the formula:

$$S_o = \sum_{i=1}^{k} \sum_{j=1}^{k} N_i N_j \quad i \neq j \qquad (2\text{-}12)$$

where N_i = number of cases in the i^{th} category
N_j = number of cases in the j^{th} category
and k = the number of categories

For part a of Table 2-19, $S_o = (25)(25) + (25)(25) + (25)(25) + (25)(25) + (25)(25) + (25)(25) = 3750$

For part b of Table 2-19, $S_o = 100(0) + 100(0) + 100(0) + 0(0) + (0)(0) + (0)(0) = 0$

For part c of Table 2-19, $S_o = (40)(30) + (40)(20) + 40(10) + (30)(20) + (30)(10) + (20)(10) = 3500$

To compute the maximum number of differences one uses the formula

$$S_m = \frac{k}{2}(k-1)\bar{N}^2, \text{ where } \bar{N} = \frac{N}{k} \qquad (2-13)$$

For part a of Table 2-19, $S_m =$

$$\frac{4}{2}(4-1)(25)^2 = (2)(3)(625) = 3750$$

For part b of Table 2-19 $S_m = 2(3)(625) = 3750$

For part c of Table 2-19 $S_m = (2)(3)(625) = 3750$

The IQV's for these various tables are as follows:

for part a of Table 2-19 IQV = 3750/3750 = 1.00

for part b of Table 2-19 IQV = 0/3750 = 0

for part c of Table 2-19 IQV = 3500/3750 = .93

Note that $S_m = S_o$ for part a of Table 2-19. This is as it should be because we knew that those data were as diverse as they could possibly be. For part b, $S_o = 0$, for there is no diversity. S_o for part c is between those for parts a and b.

The IQV can be used nicely for comparative purposes. Mueller, et al (1978) give an example in computing the relative amount of racial homogeneity in two communities. The numbers of whites and blacks in Indianapolis and Louisville in 1970 are shown in Table 2-20 below. The IQV for each city is also computed and it may be seen that they are quite similar in the amount of homogeneity.

If one has data that are given in proportions rather than in raw frequencies one can simply compute the IQV using the proportions rather than the frequencies, as shown with the data from Table 2-19.

Table 2-20 Racial Composition of Indianapolis and Louisville, 1970

	Number of Whites	Number of Blacks
Indianapolis	967,710	137,364
Louisville	724,120	100,683

For Indianapolis

$S_o = (967,710)(137,364) = 13,292,851$
$S_m = (553,537)(553,537) = 30,640,321$
$IQV = S_o/S_m = .434$

For Louisville

$S_o = (724,120)(100,683) = 7,290,657$
$S_m = \underline{2} \quad (2-1)(412,402)^2 = 17,007,499$
$IQV = S_o/S_m = .429$

The Range

While the IQV is suitable for qualitative data the range is suited for quantitative data (and in a limited sense to data measured on an ordinal scale). The range is simply the smallest interval that encompasses all values. For instance, in Table 2-1, the ages of the bank employees range from 23 to 64.5. This is a total range of 41.5 years. The SPSS computer printout gives the minimum value of this range (23.0), the maximum value (64.50) and the total range (41.5 years). It assumes that we are dealing with quantitative data.

If we have data measured on an ordinal scale we can discuss its range in a theoretical sense. For instance, we may say that political organizations in a community range from the John Birch Society on the far right to a neo-Maoist organization on the left. This is a theoretical range, however, not a mathematical one; so it cannot be regarded as a statistic and is not used in computations.

There are, of course, many problems with the range as a statistic. It is crude, inexact and gives no hint as to the distribution of values between the extremes. We have no idea if the minimum and maximum are erratic cases or actually not that atypical. To counteract these problems you might want to report some type of intermediate range. These would use the position measures discussed earlier in conjunction with the computation of the median. For instance, you might report the interquartile range, the first and third quartile of a set of data (3.8 and 7.0 for the BIA data for Native Americans). You might also report the middle 80% range (from C10 to C90) (3.25 to 10.78) for Native Americans.

Sometimes we might be interested in how the range is affected with changes in a frequency distribution. Say we are examining the distribution of incomes within a population. Suppose the minimum is $3000; the maximum is $15,000; and the range is $12,000. If everyone earns $1000 more then the range is unchanged, even though the minimum and maximum both are increased. If only the poor people earn more, the range would become smaller; if only the rich earn more, the range would become larger. This illustrates how the range can be useful in a limited sense.

Averaged Deviations

The most common way of measuring dispersion within a frequency distribution is to examine the deviations of scores from a measure of central tendency. There are three types of these measures and each will be considered below. They all involve summing the deviations of the scores from the mean or median and then averaging these deviations.

The Average Deviation -- As noted above, the sum of deviations of scores around the mean equals zero. However, if we ignore the sign of these deviations and simply look at the absolute difference of scores from the measure of central tendency, the sum of deviations or absolute deviations around the median is smaller than the sum of absolute deviations around the mean. This is illustrated in Table 2-21 below with data from Table 2-15.

Table 2-21 Example of Computing Absolute Deviations Around Mean and Median

X	|X-Md|	|X-X|
6	3	5
7	2	4
8	1	3
10	1	1
16	7	5
19	10	8
	24	26

The average deviation around the median (AD_{med}) is simply the average of these absolute deviations of scores around the median. For the data in Table 2-21, $AD_{med} = 24/6 = 4.0$

In general,

$$AD_{med} = \frac{\Sigma|X - Md|}{n} \quad \text{where X is a score, Md is Median, and n is the sample size} \quad (2\text{-}13)$$

This is also referred to as the median deviation. The value can simply be interpreted as the average distance of values in the distribution from the median.

One could also compute the average deviation of scores from the mean, but because this value is consistently larger than the AD_{med}, it is seldom used. In fact, even though the AD_{med} has a very nice intuitive interpretation it is seldom reported in the literature and is not commonly provided by computer programs, including SPSS.

The AD can also be computed for grouped data. Table 2-22 gives the computation of the AD_{med} for the BIA data. Note that the general formula is:

$$AD_{med} = \frac{\Sigma f_i |X_i - Md|}{N} \quad \text{where fi is frequency of an interval, Xi is midpoint of that interval, Md is the median, and N is the sample size} \quad (2\text{-}14)$$

Table 2-22 Competition of Average Deviation from Median for BIA Data

	NA	non NA	freq.		f|X - Md|	
X_i	$\|X_i - Md\|$	$\|X_i - Md\|$	NA	non NA	NA	non NA
2	2.7	8.1	3	0	8.1	0
4	0.7	6.1	55	9	38.5	54.9
6	1.3	4.1	17	11	22.1	45.1
8	3.3	2.1	7	10	23.1	21.0
18	5.3	0.1	9	36	47.7	3.6
12	7.3	1.9	7	24	51.1	45.6
14	9.3	3.9	2	9	14.6	35.1
16	11.3	5.9	0	1	0.0	5.9
			100	100	205.2	211.2

med NA = 4.7
med non NA = 10.1

for Native Americans $AD_{med} = \frac{205.2}{100} = 2.05$

for non-Native Americans $AD_{med} = \frac{211.2}{100} = 2.11$

The Variance -- Much more common than the average deviation is the variance. The variance involves deviations around the mean. However, because the deviations around the mean sum to zero, it is necessary to somehow get rid of the negative signs. This is done by squaring each of the deviations. The variance is then computed by averaging these squared deviations. It is defined as follows:

$$\sigma^2 = \frac{\Sigma(X-\bar{X})^2}{N} \quad \text{where} \quad \begin{array}{l}\bar{X} = \text{mean} \\ N = \text{population size}\end{array} \qquad (2\text{-}15)$$

Note that we have used the Greek letter σ^2 in defining the variance. This indicates that the value is for the population. In talking about the sample we use the roman letter s^2.

The variance does not have an easy intuitive interpretation. It is the average of the squared deviations of scores around the mean, but this does not seem to mean much on an intuitive level, especially when you realize that we are talking about squared units. Table 2-23 gives the computations for the variance for the BIA data. Note that this says that the variance for Native Americans is 8.18 squared grade levels; the variance for non-Native American employees is 8.08 squared grade levels.

The Standard Deviation -- The standard deviation is a translation of the variance into units that are more easily understood. The standard deviation is simply the square root of the variance:

$$\sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} \qquad (2\text{-}16)$$

for grouped data:

$$\sigma = \sqrt{\frac{\Sigma f_i(X_i-\bar{X})^2}{N}} \qquad (2\text{-}17)$$

Table 2-23 Computations of Variance and Standard Deviation for BIA Data

	Frequencies		$(X_i - \bar{X})$		$(X_i - \bar{X})^2$		$f(X_i - \bar{X})^2$	
Xi	NA	non NA	NA	non NA	NA	non NA	NA	non NA
2	3	0	-3.9	-7.7	15.2	59.3	45.6	0
4	55	9	-1.9	-5.7	3.6	32.5	198.6	292.4
6	17	11	0.1	-3.7	.01	13.7	.2	150.6
8	7	10	2.1	-1.7	4.4	2.9	30.9	28.9
10	9	36	4.1	0.3	16.8	0.1	151.3	3.2
12	7	24	6	2.3	37.2	5.3	260.5	127.0
14	2	9	8.1	4.3	65.6	18.5	131.2	166.4
16	0	1	10.1	6.3		39.7	0	39.7
	100	100					818.3	808.2

$\bar{X}_{na} = 5.9$ for NA $\sigma^2 = 818.3/100 = 8.18$ $\sigma = 2.86$

$\bar{X}_{non NA} = 9.7$

for non NA $\sigma^2 = 808.2/100 = 8.08$ $\sigma = 2.84$

To eliminate rounding errors, hand computations should generally use a computing formula.

For the BIA data, the standard deviation for the native Americans is 2.86; for the non-Native Americans it is 2.84. Note again, however, that the standard deviation does not really have an easy intuitive definition. It is the square root of the average of the squared deviations of scores around the mean. By comparing the standard deviation of the native Americans and non-Native Americans we can see that they are essentially equally diverse. They have approximately equal standard deviations.

As noted above, we have used the Greek letters above in defining the standard deviation and the variance. This is because the values and formulas differ slightly if we are describing a population or a sample. Simply because we are taking a sample from a population any sample is less variable than the population it comes from. When the sample is small compared to the population this difference can be substantial, but with very large samples it is quite small. The formulas for the standard deviation and the variance of a sample take this into account, however, by altering the denominator to be n-1 (or one less than the sample size) rather than n. For small samples this will produce greater differences between the formulas for σ and s than for larger samples. Some texts call this formula $\hat{\sigma}$ instead of s. You should understand the logic and look for the formula as it is defined. The formulas for the sample values of the standard deviation and variance are given below.

$$s^2 = \hat{\sigma}^2 = \frac{\Sigma(X-\bar{X})^2}{n-1} \qquad (2\text{-}18)$$

$$s = \hat{\sigma} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} \qquad (2\text{-}19)$$

The SPSS program assumes the data it is given are from a sample and uses the formulas given directly above in its computations. Sometimes the value of the variance is too large for the computer to print (it has too many digits). When this happens you can compute it by simply squaring the value of the standard deviation. Just as with the measures of central tendency you must be careful in how you submit data to the computer for the results with the standard deviation and variance to be accurate. Your best bet is to simply recode the values, as with the income data in Table 2-14 from the NORC study, to the midpoints of the intervals. If you had used the unrecoded data the computer would give you a much smaller value as the standard deviation for these data then if you had recoded to the midpoint of each interval.

<u>The Coefficient of Relative Variation</u> -- The full utility of the standard deviation will only become clear after we discuss the normal distribution in the next section. The standard deviation and the average deviation, however, both have a nice descriptive use in the Coefficient of Relative Variation, a measure that is used with ratio data. It is necessary to have data measured on a ratio scale when using the CRV because it involves looking at the relative size of the measure of dispersion and the measure of central tendency. If the size of the intervals were arbitrary (that is, if there were no true zero point), this ratio would be meaningless.

The form of the CRV is simply the measure of dispersion divided by the measure of central tendency. For the median

$$\text{CRV} = \text{AD}_{\text{med}}/\text{Med} \qquad (2\text{-}20)$$

and for the mean

$$\text{CRV} = S/\bar{X} \qquad (2\text{-}21)$$

The CRV is used to compare the deviations of a group to the average for that group. You might remember that while the native American and non-Native American employees of the BIA have very dissimilar measures of central tendency in grade level, the measures of dispersion are quite similar. The CRVs for these data are given in Table 2-24 below.

It appears that the CRV for Native Americans is substantially larger than the CRV for non-Native Americans. This indicates that not only do the non-Native Americans have a larger mean, but that relative to this mean they vary much less.

Table 2-24 Computation of CRVs for BIA Data

	Native American	non-Native American
CRV median	2.05/4.71 = 0.44	2.11/10.11 = 0.21
CRV mean	2.86/5.86 = 0.49	2.84/9.72 = 0.29

Another example is given by Mueller et al, 1978. This involves the homicide rates in the New England and South Atlantic states. The AD for the New England states is .78, while the AD for the South Atlantic states is 3.60, suggesting that the states of the northeast are much more homogeneous since their average divergence from the median is so much smaller. However, once we look at this average deviation relative to the median the picture changes. The median homicide rate for the New England states is 2.75, while that for the South Atlantic States is much larger, 12.15. The CRV's are computed below.

New England States:	CRV = .78/2.75 = .284
South Atlantic States:	CRV = 3.60/12.15 = .296

It is now apparent that relative to their respective medians, the two groups of states do not differ markedly in their relative variation.

Yet, another example of the use of the CRV is in Tables 2-25 and 2-26. These are taken from Christopher Jencks' book <u>Inequality</u> (1972). The first shows the coefficients of variation for education (years of regular schooling completed) for various groups of cohorts of individuals in the United States. The second gives the coefficients of variation for income. Note that the CRV's are much smaller for education than for income, a central point in Jencks' analysis.

It must be mentioned again that the CRV is only usable when we have ratio data. It involves computing ratios and this can only be done when we have a true zero point, when those ratios would make sense.

Summary

We have examined a number of ways of describing univariate distributions: frequency distributions displayed in tables, graphs of the data, measures of central tendency, and measures of dispersion. We have noted which forms or statistics are appropriate for variables measured on different levels. We have also cautioned students on the use of computers and calculators and their output.

We have used one example throughout this chapter -- the grade levels of employees of the Bureau of Indian Affairs in 1970. We have assumed that this variable is measured on a ratio scale (although this is admittedly stretching it unless we translate the grades into dollars earned, the original reason for setting up the grade limits). The frequency distribution for both Native American and non-Native American employees is given in Table 2-6. Relevant graphs are given in Figures 2-2, 2-3, 2-7. Statistics for these data are computed throughout the text and are summarized in Table 2-27. Note that all of these results suggest that Native Americans are employed at much lower grade levels than non-Native Americans, even though it is the policy of the Bureau (and has been for many years) to give Native Americans employee preference in hiring. All of the measures of central tendency are much lower for the Native Americans than for the non-Native Americans. The range for the Native Americans is slightly smaller although the average deviation, variance and standard deviations are almost equal. However, the coefficients of relative variation are strikingly different, with that for the non-Native Americans being much less. This suggests that, relative to their means, the non-Native Americans actually have much less variation than the Native Americans.

Table 2-27 Summary of Measures of Central Tendency and Dispersion for BIA Data

Measure	Native Americans	non-Native Americans
Mode		
Crude	4.0	10.00
Refined	4.16	10.37
Median	4.71	10.11
Mean	5.86	9.72
Minimum*	1	1
Maximum*	16	17
Range*	15	16
Average Deviation (median)	2.05	2.11
Variance	8.18	8.08
Standard Deviation	2.86	2.84
CRV Median	0.44	0.21
CRV Mean	0.49	0.29

*Computed from data with interval lengths of 1 grade. All others computed from data with interval widths of 2 grades.

Packet 119

SOC 326

QUANTATIVE METHODS IN SOCIOLOGY

Professor Stockard

University of Oregon

Winter Term 1992

up-84
b

kinko's
the copy center
860 E. 13th
Eugene • 344-7894

Copies:	$3.70
Binding	$0.00
Royalties	$0.00
Permission Handling Charges	$0.00
Total cost of packet:	$3.70

TABLE OF CONTETS

Jean Stockard - Packet 119

III. The Normal Distribution

In this section we examine the characteristics of the normal distribution. The normal distribution is a special frequency distribution that has very useful mathematical properties. It is symmetrical, that is both sides of the distribution are identical. This means that half the cases are above the mean and half the cases are below the mean. It is bell shaped, indicating that most of the cases are at the mean and relatively fewer are at the extremes. It is infinite; that is the distribution keeps going out on either side infinitely. It is also unimodal; the mean, the mode, and the median are all the same value. Even though all normal curves share these characteristics, not all normal curves look alike. Some are relatively short and wide, others are taller and narrower. Some are more peaked, while others are more flat. Figures 3-1 and 3-2 give examples of the normal curve. In the first example three normal curves are shown. They all have the same standard deviation, but different means. In the second example the distributions are also both normal. They have the same mean, but they have different standard deviations.

Figure 3-1

Normal Distributions with Unequal Means and Equal Standard Deviations

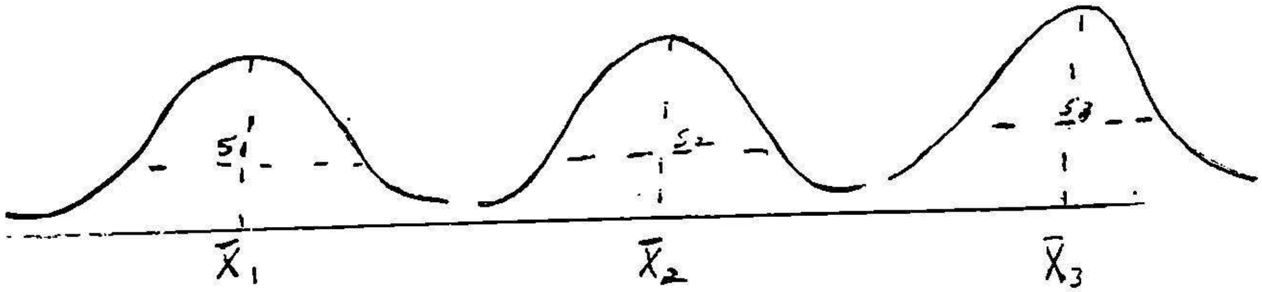
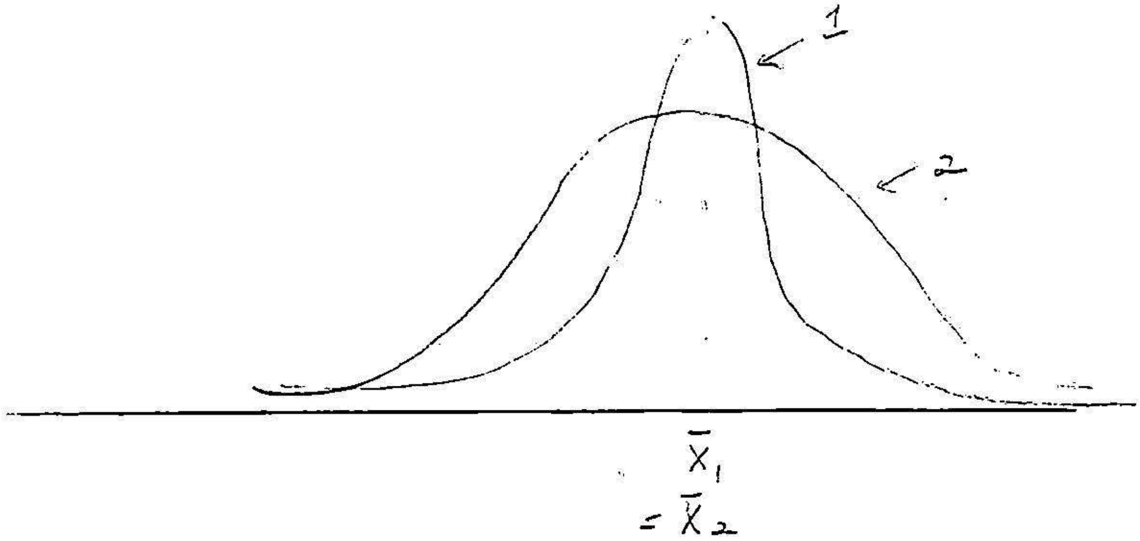


Figure 3-2

Normal Distributions with Equal Means and Unequal Standard Deviations



Normal distributions are most commonly approximated in natural situations. For instance, shoe size, height, weight, gestation periods, and other biological phenomena generally tend to assume a shape like a normal distribution. Other distributions tend to approach the normal one, but, most importantly, theoretical distributions used in statistical inferences are often normally distributed. Many of our statistics are based on the properties of the normal curve.

The most important aspect of the normal curve involves the area under or enclosed by the curve. Regardless of what the mean or the standard deviation is, the proportion of the area under the curve between the mean and a given distance in standard deviation units from the mean is constant. In other words, we could mark out the distance from the mean in standard deviation units, as is done in Table 3-2, and know what proportion of the area under the curve is in each part. Obviously, half of the area is above the mean and half is below the mean. About 34% of the area is between the mean and one standard deviation on each side. Or about 68% of the area is between one standard deviation above and one standard deviation below the mean. About 95% of the area is between two standard deviations above and below the mean.

Because this is standard within all normal distributions, we can compute the area under the curve and corresponding information for any normal distribution (examples are given below). This is done by using standard tables statisticians have developed that tell what proportion of the area under the curve is between the mean and any standard deviation unit from the mean. An example is Table 3-1, the table of the normal distribution which was handed out in class. Another example is Table A in Appendix C of Elifson, et al (pp. 463-465). To use this table for any normal distribution you need only convert your normal distribution to equal the one where the mean is zero and the standard deviation is one. Some tables, like the one in Elifson, also give the proportion of area found under the curve beyond a given standard deviation unit from the mean. Note that the two values (the proportion of the area between the mean and a given standard deviation unit and the proportion of area beyond a given standard deviation unit) must sum to .50. This is because half of the area under the normal curve lies on each side of the mean.

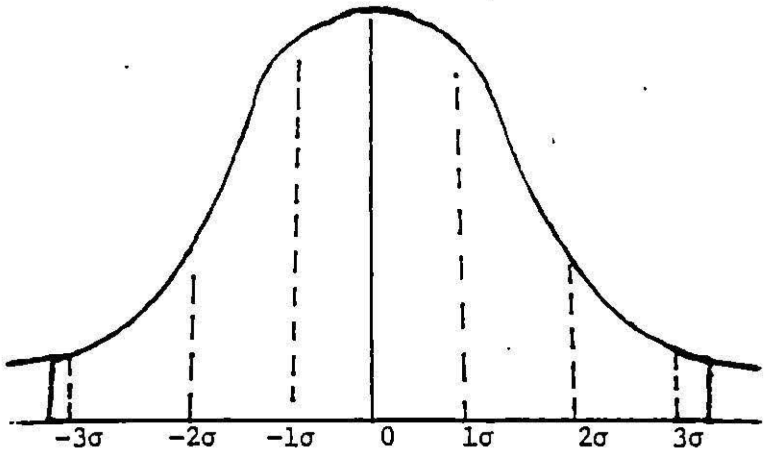
Part one of Table 3-2 illustrates the use of this table with a normal distribution. For instance we know from the properties of the normal distribution that the area on one side of the mean of zero is 50% of the total distribution (lines a and b). Suppose we were interested in the proportion of area under the normal curve between the mean and one standard deviation above the mean. To find what value corresponds to this area we look down the left hand column of Table 3-1 until we find 1.0, corresponding to 1 standard deviation unit from the mean. We then move to the next column to the right headed .00. (The columns headed by two decimal points [.00, .01, .02, ...] are used when finding the area under the curve at a point in standard deviation units measured to the nearest hundredth.) The value here is .3413, indicating that the area from the mean (0) to one standard deviation above the mean includes 34.13% of the total area (line c). Remembering that 50% of the area lies below the mean we can say that below 1 standard deviation above the mean there is 50% + 34.13% = 84.13% of the total area under the curve (line d).

Again looking at Table 3-1 we can see that between the mean and two standard deviations above the mean we have .4772 of the total area (line e). If we remember that one-half of the area is below the mean we can easily calculate that .9772 of the total area falls below two standard deviation units above the mean (line f). Then combining information in lines c and e we can tell that between one standard deviation and two standard deviations above the mean is .1359 of the area (line g). Line i looks at the corresponding area below the mean. If we remember that the normal distribution is symmetrical, we can compute that .8185 of the total area is between one standard deviation below the mean and two standard deviations above the mean (line h).

Part two of Table 3-2 illustrates how one finds the proportion of area under a normal curve when the mean is not

Table 3-2

Examples of Using the Normal Curve Table (3-1)

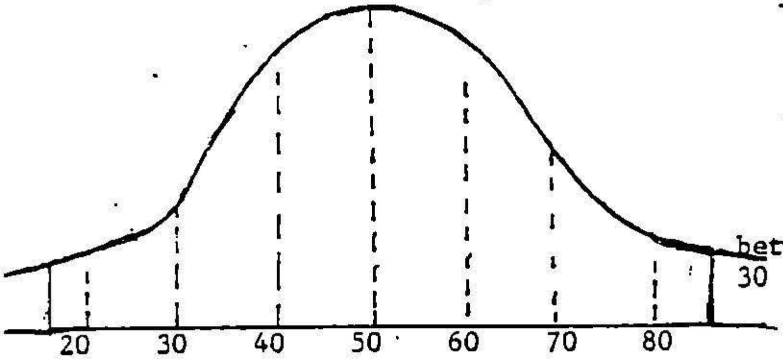


N(0,1); X = 0,σ =1, a normal distribution

from __ to __ there is	of the distribution
a) −∞ to 0	.5000
b) 0 to + ∞	.5000
c) 0 to 1 σ	.3413
d) −∞ to 1 σ	.8413
e) 0 to 2 σ	.4772
f) −∞ to 2 σ	.9772
g) 1σ to 2 σ	.4772 - .3413 = .1359
h) -1σ to 2 σ	.3413 + .4772 = .8185
i) -2σ to-1 σ	.4772 - .3413 = .1359

X = 50;σ = 10, normal distribution

$$z = \frac{x - \bar{x}}{\sigma}$$



X	z	Proportion of area under curve to that point X
60	1.0	.5000 + .3413 = .8413
65	1.5	.5000 + .4332 = .9332
40	-1.0	.5000 - .3413 = .1587
50	0.0	.5000 ± 0 = .5000
25	-2.5	.5000 − .4938 = .0062
between 30 & 70	between -2.0 & +2.0	[.5000 + .4772] - [.5000 + .4772] = .9544

equal to zero and the standard deviation is not equal to one. In the example the mean is 50 and the standard deviation is 10. To transform this distribution to one where it is N(0,1) we compute z-scores. This is a simple transformation that simply moves the mean of the distribution along to zero and stretches or compresses the standard deviation so that it is equal to one. The z transformation is simply

$$z = (X - \bar{X})/s \quad \text{or} \quad (X-\mu)/\sigma \qquad (3\text{-}1)$$

You may see in part_b of Table 3-2 that when the mean (50) is substituted for $\bar{X}$ in the z-transformation the z-score equals zero. When 40, one standard deviation below the mean is substituted, z = -1. When 60, one standard deviation above the mean is substituted, z = +1. The chart in part b of Table 3-2 gives the z-score for various values of X and then shows how one would compute the proportion of area under the curve up to that value of X.

For instance, when X (the score under consideration) equals 60, the corresponding z-score is (60-50)/10 = +1.0. We can then refer to Table 3-1 and note that between the mean and one standard deviation above the mean there is .3413 of the total area. Since we know that .5000 of the area is below the mean, we can say that .5000+ .3413 = .8413 of the area under the curve is at or below the score of 60. As another instance, consider X = 40. Here z = (40-50)/10 = -1.0 or one standard deviation unit below the mean. We know that between the mean and one standard deviation below the mean there is .3413 of the total area. Since there is .5000 of the total area below the mean, below one standard deviation below the mean, there must be .5000 - .3413 or .1587 of the total area. Students should work through remaining examples to assure they understand the procedures involved.

So far we have only talked about "scores" and in rather abstract terms. Suppose instead, again considering part b of Table 3-2, that the scores represent the number of items on a test that students had correctly answered. Assume also that there were many students involved and that the distribution of scores was N (50, 10) (normally distributed with a mean of 50 and a standard deviation of 10). The computations in part b of Table 3-2 would then tell us that 84.13% of the students had scores of 60 or lower, 93% of the students had scores of 65 or lower, etc. In addition, 95% of the students had scores between 30 and 70.

Very few actual frequency distributions that sociologists work with are normally distributed. Yet, understanding the characteristics of the normal curve can help in interpretations of the standard deviation for all types of distributions. For instance, suppose one was interested in studying the distribution of income within a population and that one knew that the mean was $20,000 and the standard deviation was $3000. You could then know that if this distribution were shaped like a normal distribution, approximately 64% of the cases in the population would have incomes between $17,000 and $23,000 (+ or - one standard deviation from the mean). Similarly, approximately 96% of the cases would have incomes between $14,000 and $26,000 (+ or - two standard deviations from the mean). Similarly, you could compute a z-score to find that an income of $21,500 was .5 standard deviations above the mean ($z= (X-\bar{X}) /s$ = ($21,500 - $20,000) / $3,000 = 1500/3000 = .5). Then you could consult the table of the normal curve to determine that, if the distribution were shaped like a normal curve, .3085 or 31% of the cases would have incomes higher than this value and 69% would have incomes lower than this value.

IV. Bivariate Statistics Appropriate for Qualitative Variables

The work in the last two sections has generally focused on measures appropriate only for variables measured on at least an interval scale. However, as we have noted earlier, many variables used in the social sciences are qualitative in nature and are measured on only a nominal or ordinal level. In addition, the previous work has looked only at univariate distributions. It has involved looking at only one variable at a time. In this section we begin to look at how two variables are associated with each other, or go together.

We first review the basic rules involved in precentaging tables and displaying and interpreting data regarding the relationship between two qualitative variables and then examine two measures of association that can describe these relationships.

Developing and Interpreting Bivariate Tables

Suppose we were interested in the relation between subjects' religious preference and their political party identification. Our theoretical and substantive readings had led us to conclude that religious preference has an influence on the type of political party with which people identify. One would then say that political party identification is the dependent variable and that religious preference is the independent variable. Each of these variables may be said to be measured on a nominal scale: Party Identification with three attributes (Democrat, Independent, and Republican) and Religous Preference with four attributes (Protestant, Catholic, Jewish, and Other).

The most appropriate way to display these data would be a table that was percentaged to show the relationship between the two variables. Such a table is Table 4-1. Note that the table includes a title, subheads, and notes as described in an earlier section. Note also that the percentages are computed within categories of the independent variable and that the percentage distribution for the total group on the dependent variable is also included. This distribution for the total group is referred to as the marginal distribution. The distributions within each category of the independent variable are called the conditional distributions (conditional upon the categories of the independent variable). The entire table, the whole set of columns and comparisons, is referred to as the joint distribution of religious preference and party affiliation.

Table 4-1
Percentage Crosstabulation of Religious Preference and Party Identification

Party Identification	Religious Preferences				
	Protestant	Catholic	Jewish	Other	Total
Democratic	42.5%	53.5%	61.8%	30.6%	44.8
Independent	31.1	32.6	29.4	56.5	33.2
Republican	26.5	13.9	8.8	12.9	22.0
TOTAL	100.1%*	100.0%	100.0%	100.0%	100.0%
N	998	368	34	108	1508

* Does not add to 100% due to rounding
Source: 1977 General Social Survey

The fact that the table is percentaged within categories of the independent variable is very important. This is necessary to allow for meaningful comparisons across the categories of the independent variable, to tell what kind of effect the independent variable has on the dependent variable. By percentaging within each category of the independent variable the data within each of those categories is standardized. In Table 4-1 we are interested in the effect that religious preference, the independent varible, has on political party identification, the dependent variable. In reading or interpreting percentaged tables such as this, we compare the percentages across the dependent variable, whether the dependent variable is placed in the columns or the rows of the table. Another way of describing this is to say that we compare the marginal distribution of the dependent variable with the conditional distributions.

In examining Table 4-1 we would first look at the marginal distribution (the total figures), noting that close to half (44%) of the respondents are Democrats, about one-third (33.2%) are Independents, and only slightly more than one-fifth (22%) are Republicans. We would then look at the conditional distributions, comparing these to the marginal distributions. We see then that religious preference (the independent variable) appears to be related to political identification (the dependent variable). With regard to democratic affiliation, those with a Jewish preference are most likely to be in this group, followed by Catholics. Those in the Protestant category are slightly less likely than those in the total group to be Democrats and those with the "Other" religious preference are least likely to be

Democrats. Those with the "Other" preference are much more likely than Protestants, Catholics, or Jews to identify with the Independents; Protestants, Catholics and Jews have about equal tendencies to identify with the Independents. Finally, Protestants are much more likely than those who prefer the other religious groups to identify with the Republicans. Jews are least likely to identify with the Republicans; but, Catholics, Jews and those with an "other" preference all identify with the Republicans far less often than the total group. Note that in these comparisons we are essentially reading across or comparing the values of the dependent variable in the categories of the independent variable.

The information contained in a simple bivariate table will usually be insufficient to answer a research question. We will want to introduce one or more control variables to further examine the relationship apparent in the bivariate (or zero-order) table. Table 4-2 illustrates the introduction of the control variable "annual family income" (dichotomized as below $20,000 annually and $20,000 and above). Note that in this table the original bivariate table is reproduced for those from families earning less than $20,000 annually and for those with a family income of $20,000 and higher. Note also that percentages are computed within each category of the independent variable within each sub-table.

Here we would first compare the marginal distributions in the two partial tables. We would note that those with lower family incomes are much more likely to be Democrats, those with higher incomes are much more likely to be Independents and those in the two income groups are equally likely to be Republicans.

We would then look at the conditional distributions in each partial table. Within both income groups Jews and Catholics are more likely than Protestants and those with other religious preferences to identify with the Democrats. This result parallels that found in the zero order table. Among those earning less than $20,000 both those with an "other" preference and Catholics are more likely than Protestants and Jews to identify with the Independents. Among those with a family income of $20,000 or more only those with an "other" religious preference are more likely than the total group to identify with the Independents. Only the results with this higher income group parallel those found in the zero-order table. Finally, among those with a family income of less than $20,000, only Protestants are more likely than the total group to indicate a Republican preference and no Jews in this income category and only 8% of the Catholics indicate a Republican preference. Among those with a higher family income both Protestants and Catholics indicate a Republican identification more often than the total group. In the zero order table only Protestants were more likely than those in the total group to identify with the Republicans.

Table 4-2
Percentage Crosstabulation of Religious Preference and Party Identification by Annual Family Income

Annual Family Income: Less Than $20,000

Party Identification	Religious Preferences				
	Protestant	Catholic	Jewish	Other	Total
Democratic	50%	60%	80%	40%	53%
Independent	20	32	20	40	25
Republican	30	8	0	20	22
TOTAL	100	100%	100%	100%	100%
N	450	250	10	50	760

Annual Family Income: $20,000 and More

Party Identification	Religious Preferences				
	Protestant	Catholic	Jewish	Other	Total
Democratic	36%	40%	54%	22%	37%
Independent	40	34	33	71	41
Republican	24	26	13	7	22
TOTAL	100	100%	100%	100%	100%
N	548	118	24	58	748

Source: Hypothetical

These results suggest that even when we control for family income, religious preference seems to be related to party identification. However, some differences do appear between the zero order and partial tables, especially with respect to Catholics' identification with Independentsand Republicans. While the zero order table indicates that Catholics identify with the Independents about as often as the total group and with the Republicans less often than the total group, the partial tables indicate that low-income, but not high-income, Catholics are more likely than their total income group to identify with the Independents. High-income, but not low-income Catholics, are more likely than their total income group to identify with the Republicans.

Measures of Association

Measures of association summarize the extent to which one variable depends upon or is related to another. While one can examine the percentage differences found in tables, such as 4-1 and 4-2, to determine the extent to which a relationship exists between two variables, such computations can become tedious and confusing when there are more than two categories within a variable. Thus, researchers have developed single measures which summarize the degree to which two variables are associated with each other.

There are many measures of association, the most useful of which have what is called a proportionate-reduction-of-error (PRE) interpretation. Below we first describe the basic elements of a PRE statistic and then describe two such measures of association, one appropriate for variables measured on a nominal scale and the second appropriate for variables measured on an ordinal scale.

PRE Measures of Association

The designation of PRE measures of association was first proposed by the sociologist Herbert Costner to describe a large variety of measures of association. PRE measures have four common elements:

First, they have a rule for predicting the classification of each subject on the dependent variable, ignoring information about the classification of that member on the independent variable. For this rule we generally look only at the marginal distribution of the dependent variable. The basis for the prediction often involves measures of central tendency such as a mode or a mean.

Second, we need a rule for predicting the classification of each subject on the dependent variable using the information about the classification of that member on the independent variable. In other words, when using the second rule we take information about both the independent and dependent variable into account.

Third, we need a definition of what is meant by a prediction error. This definition varies by the level of measurement of our variables. With nominally measured variables an error is usually a misclassification. With ordinally measured variables, an error involves a wrong prediction of relative order of a pair of variables. With intervally measured variables, an error can involve deviations from a central value such as the mean. The number of errors in classifying the dependent variable when using rule one is termed E_1. The number of errors in classifying the dependent variable when using rule two is termed E_2.

Fourth, a PRE measure of association is defined as

$$PRE = (E_1 - E_2)/E_1 \qquad (4\text{-}1)$$

The term in the numerator $(E_1 - E_2)$ essentially tells us how much knowing the independent variable reduces our error in predicting the dependent variable. The term in the denominator (E_1) is simply our total error in predicting the dependent variable when we don't know the independent variable. Thus the PRE measure tells how much our knowledge of the independent variable has reduced our error in predicting the dependent variable as a proportion of the total error we have if we don't know the independent variable. In other words, a PRE measure tells us how much our error in predicting the dependent variable is reduced once we know the independent variable.

The definition of error used varies from one type of measure to another. The result however, is always a proportion. When a PRE measure equals 0 we could say that there is no association -- we have had no reduction in our error in predicting the dependent variable as a result of knowing the independent variable. When a PRE measure equals - or + 1.00, we would say that there is perfect association -- a total reduction of error in predicting the dependent variable when we know the independent variable.

<u>Lambda: A Measure for Variables on a Nominal Scale</u>

Lambda (λ) is a PRE measure appropriate for variables measured on a nominal scale. Error is defined in this case as a misclassification, not predicting the correct category of the dependent variable.

Rule 1, the rule for predicting categories of the dependent variable when we do not take the independent variable into account, involves simply examining the marginal distribution of the dependent variable. If we know nothing about the joint distribution, we would be most often correct if we predicted that a case fell into the modal category. Our errors by rule one would then be computed by

$$E_1 = N - \max N_{.j} \qquad (4\text{-}2)$$

where $N_{.j}$ are the marginal frequencies of the dependent variable. In other words, E_1 is the total of the frequencies not found in the modal category.

Rule 2 is the rule for predicting categories of the dependent variable when we take the knowledge of the independent variable into account. In this case, we could use as our best prediction of categories of the dependent variable, if we know the joint distribution, the modal category of the dependent variable within each category of the independent variable. The number of errors then corresponds to the total number of cases not found in these modal categories. This may be represented as:

$$E_2 = N - (\text{Max}N_{1j} + \text{Max}N_{2j} + \ldots + \text{Max}N_{kj}) \qquad (4\text{-}2)$$

where N_{ij} = cell frequencies in each category of the independent variable i.

Table 4-3
Joint Distribution of Race/Ethnicity and Unemployment Status for a Hypothetical Sample of Teenagers

	Race/Ethnicity				
Employment Status	White	African-American	Hispanic	Other	Total
Employed	250	50	100	75	475
Unemployed	50	150	100	25	325
Totals	300	200	200	100	800

Consider the data given in Table 4-3 regarding the association between race/ethnicity and unemployment status for a hypothetical group of teenagers. Within this table race/ethnicity must clearly be the independent variable. Then, looking at the distribution of marginals for the dependent variable, employment status, it may be seen that the modal category is "employed," which contains 475 of the 800 cases. Thus, if we were to guess, based on this marginal distribution, that teenagers were employed we would be right in our prediction 475 times and wrong 800 - 475 or 325 times (800 - 475 = 325). In other words, our errors by rule 1 would be 325.

Now, however, if we took into account the distribution of the teenagers across the 4 categories of race-ethnicity, we would have a different picture. For the whites, we would predict that they were employed, and we would be right 250 times. For the African-Americans, we would guess that they would be unemployed, and we would be right in this guess 150 times. For the Hispanics, we could guess that they would be either employed or unemployed and be right 100 times with either guess. Finally, for the "other" group we would guess that they were employed, and be correct 75 times. Thus, if we knew the race-ethnicity of the teenagers we would be right 575 times (250 + 150 + 100 + 75 = 575). Given that there are 800 people all together we would be wrong 225 times. Thus, our errors by rule 2 would be 225.

We can now compute lambda (λ) for this association.

$$\text{lambda} = (E_1 - E_2) \;/\; E_1$$

$$= (325 - 225) / 325 = 100 \;/\; 325 = .31$$

This indicates that we can reduce our error in predicting teenagers' employment status by 31% (almost one-third) if we know their race and ethnicity.

Table 4-4 gives the actual frequencies for the data given in percentage form in Table 4-1 for the relationship of religious preference and party identification. Remember that Religious Preference is the independent variable and Party Identification is the dependent variable. The computations for lambda for these data are shown in Table 4-5.

Table 4-4
Frequency Cross-Tabulation of Religious Preference and Party Identification

Party		Religious Preference			
Identification	Protestant	Catholic	Jewish	Other	Total
Democratic	424	197	21	33	675
Independent	310	120	10	61	501
Republican	264	51	3	14	332
Totals	998	368	34	108	1508

Table 4-5
Computation of Lambda for Data in Table 4-4

$$\begin{aligned} E_1 &= N - \text{Max } N_j \\ &= 1508 - 675 = 833 \end{aligned}$$

$$\begin{aligned} E_2 &= (\text{Max } N_{1j} + \text{Max } N_{2j} + \text{Max } N_{3j} + \text{Max } N_{4j}) \\ &= 1508 - (424 + 197 + 21 + 61) \\ &= 1508 - 703 = 805 \\ &= \frac{E_1 - E_2}{E_1} = \frac{833 - 805}{833} = \frac{28}{833} = .034 \end{aligned}$$

Th lambda value of .034 indicates that we reduce our errors in predicting the respondents' party identification by .034, or 3.4%, once we know their religious preference.

Table 4-6 gives the frequencies corresponding to the percentage distribution given in Table 4-2 for the relationship between religious preference and party identification for those with annual family income under $20,000 and $20,000 and over.

Table 4-6
Crosstabulation of Religious Preference
and Party Identification by Annual Family Income

Annual Family Income: Less Than $20,000

Party Identification		Religious Preferences			
	Protestant	Catholic	Jewish	Other	Total
Democratic	225	150	8	20	403
Independent	90	80	2	20	192
Republican	135	20	0	10	165
TOTAL	450	250	10	50	760

Annual Family Income: $20,000 and More

Party Identification		Religious Preferences			
	Protestant	Catholic	Jewish	Other	Total
Democratic	199	47	13	13	272
Independent	220	40	8	41	309
Republican	129	31	3	4	167
TOTAL	548	118	24	58	748

Source: Hypothetical

Table 4-7

Computation of Lambda for Data in Table 4-6

For those with Family Income < $20,000:

$$E_1 = N - \text{Max } N_j$$

$$= 760 - 403 = 357$$

$$E_2 = N - (\text{Max } N_{1j} + \text{Max } N_{2j} + \text{Max } N_{3j} + \text{Max } N_{4j})$$

$$= 760 - (225 + 150 + 8 + 20)$$

$$= 760 - 403 = 357$$

$$= \frac{E_1 - E_2}{E_1} = \frac{357 - 357}{357} = 0$$

For those with family incomes $20,000 and more:

$$E_1 = N - \text{Max } N_j$$

$$= 748 - 309 = 439$$

$$E_2 = N - (\text{Max } M_{1j} + \text{Max } N_{2j} + \text{Max } N_{3j} + \text{Max } N_{4j})$$

$$= 748 - (220 + 47 + 13 + 41)$$

$$= 748 - 321 = 427$$

$$= \frac{E_1 - E_2}{E_1} = \frac{439 - 427}{439} = \frac{12}{439} = .027$$

The calculations in Table 4-13 indicate that lambda equals zero for those with incomes lower than $20,000. For all of the religious groups, the modal political identification party is Democrat. This would indicate that knowing the respondents' religious preference does not reduce our error at all in predicting their party identification. Remember, however, that our earlier analysis of this partial table indicated that there were patterns of differences among the conditional distributions. This low value of lambda undoubtedly reflects the skewed marginal distribution of the table. Over half of the respondents identify with the Democrats and so, even though the conditional distributions vary from one category of the independent variable to another, the variation is not large enough to counteract these skewed marginals. This is an example of when lambda actually produces misleading results and when we should rely upon examination of percentages instead (as well as inferential statistics we will discuss later in the term).

The lambda value for respondents with a family income of $20,000 or more is .027, indicating that we may reduce our error in predicting the respondents' party identification by about 2.7% if we know their religious preference.

Most computer programs provide lambda. While our discussion has involved a designation of one of the two variables as dependent, there is a form of lambda that does not designate either variable as dependent, but simply tells us the extent to which knowing the categories of one variable helps predict the categories of the other. This is referred to as symmetric lambda.

When computing lambda by hand it is often easier to use computing formulas rather than the definitional formulas given here. The computing formula may be easily derived from the definitional formula used above. Remember that

$$E_1 = N - \mathrm{Max}\ N_{.j}, \qquad (4\text{-}4)$$

where $N_{.j}$ are the marginal frequencies of the dependent variable; and that

$$E_2 = N - (\mathrm{Max}N_{1j} + \mathrm{Max}N_{2j} + \ldots + \mathrm{Max}N_{kj}), \qquad (4\text{-}5)$$

where N_{ij} are the cell frequencies in each category of the independent variable i. Lambda is defined as

$$= (E_1 - E_2) / E_1 \qquad (4\text{-}6)$$

Substituting the values in 4-4 and 4-5 into 4-6 we obtain

$$= \frac{[(N - \mathrm{Max}\ N_{.j}) - (N - (\mathrm{Max}N_{1j} + \mathrm{Max}N_{2j} + \ldots + \mathrm{Max}N_{kj})}{(N - \mathrm{max}\ N_{.j})}$$

The term "N" in the numerator of the above expression cancels out and, taking into account the changing of terms required by multiplying a negative times the second term in the numerator, we are left with

$$= \frac{(\mathrm{Max}N_{1j} + \mathrm{Max}N_{2j} + \ldots + \mathrm{Max}N_{kj}) - \mathrm{Max}\ N_{.j}}{(N - \mathrm{max}\ N_{.j})} \qquad (4\text{-}7)$$

The first term in the numerator of 4-7 is the sum of the modal frequencies of the dependent variable within each category of the dependent variable. The second term in the numerator is simply the modal frequency of the dependent variable.

To briefly summarize, lambda is a useful measure of association when one is examining the relationship between two variables measured on a nominal scale. It varies from 0 to 1. A lambda of 0 indicates that knowing categories of the independent variable does not help in predicting categories of the dependent variable. A score of 1 indicates that knowing the independent variable allows us to perfectly predict the dependent variable (our reduction in error is total). Lambda has a useful proportionate reduction of error interpretation, telling us the proportionate improvement in prediction of the dependent variable that results once we know the independent variable.

Lambda is inappropriate and has misleading results when the dependent variable has a skewed marginal distribution as seen in Table 4-8, illustrating the relationship between the length of couples' marriages and the presence of communication problems. Assuming that communication problems are the dependent variable, we are interested in how length of marriage affects these problems. The computed lambda is equal to zero, but an examination of the table quickly indicates that there is indeed a relationship.

Table 4-8
Length of Marriage by Presence of
Communication Problems

	Length of Marriage in Months				
	0-36	37-47	48-59	60+	Total
Report Communi- cation	40	28	16	9	93
Problem	60	72	84	91	307
Totals	100	100	100	100	400

Table 4-8 (continued)

$$E_1 = N - \text{Max } N_j$$

$$= 400 - 307 = 93$$

$$E_2 = N \ (\text{Max } N_{1j} + \text{Max } N_{2j} + \text{Max } N_{3j} + \text{Max } N_{4j})$$

$$= 400 - (60 + 72 + 84 + 91)$$

$$= 400 - 307 = 93$$

$$= \frac{E_1 - E_2}{E_1} = \frac{93 - 93}{93} = 0$$

Those with longer marriages are much less likely than those with shorter marriages to report communication problems. This is not reflected in the lambda because of the skewed marginals. The modal category is the same, no matter what category of the independent variable is examined. When you have skewed marginals you should always carefully examine the percentage distributions, for there might indeed be substantively important results that are not reflected in the value of lambda. Later we will also examine an inferential statistic that will help us deal with such situations.

Gamma: A PRE Statistic for Ordinally Measured Variables

Gamma is a PRE measure of association designed for use with variables measured on an ordinal scale. It is most useful when the variables have a relatively small number of categories. (While lambda could be used with ordinally measured variables, in doing so we would be wasting information on the order of the subjects.)

With gamma, error is defined as occuring when we fail to correctly predict the relative order of two cases. We essentially examine all pairs of cases in the sample and look at their relative order on the two variables that are being studied. Gamma tells us how accurately we can predict the order of a pair of cases on one variable once we know their relative order on the other variable. (Gamma does not differentiate between independent and dependent variables.)

A simple example is given in Table 4-9 below. Here we examine the responses four subjects have given to two questions: 1) attitudes toward the women's movement and 2) attitudes toward voting for a woman president. Suppose that their responses may range along the following continuum: 1) strongly disapprove, 2) disapprove, 3) approve, and 4) strongly approve. Suppose also that we have arbitrarily assigned the codes noted above (1, 2, 3, 4) to each of the categories. These maintain the order of the possible responses.

Table 4-9
Example of Data on Two Ordinally Measured
Variables for Four Subjects

Subject	Attitude Toward Women's Movement	Attitude Toward Woman President
W	Strongly Agree (4)	Strongly Agree (4)
X	Disagree (2)	Agree (3)
Y	Agree (3)	Strongly Disagree (1)
Z	Strongly Disagree (1)	Disagree (2)

Subject Pair	Relative Order on Women's Move.	Relative Order on Woman Pres.	Concordant or Discordant
WX	SA > D	SA > A	C (same)
WY	SA > A	SA > SD	C (same)
WZ	SA > SD	SA > D	C (same)
XY	D < A	A > SD	D (different)
XZ	D > SD	A > D	C (same)
YZ	A > SD	SD < D	D (different)

Also given in Table 4-9 is the joint distribution of responses. This compares the scores of each pair of subjects on each variable, giving their relative order. For instance, when subjects W and X are compared it may be seen that they stand in the same relative order on the woman's movement question as on the woman for president question. In both cases W is more in favor than X. The ordering of this pair on these two variables is the same, or is called concordant. When W is compared to Y we again see that on both variables W has more favorable attitudes than Y. When W and Z are compared it is seen that on both variables W has more favorable attitudes than Z. When X and Y are compared we see that on attitudes toward the women's movement X has less favorable attitudes than Y. On attitudes toward a woman president X has more favorable attitudes than Y. This is a case of different or reverse ordering for the pair of cases on the two variables. This can be called discordant order. In all, of the total of 6 pairs of cases, 4 have the same ordering on both variables and 2 have reverse orders. Four are said to be concordant pairs, while 2 are said to be discordant pairs.

Gamma is defined as

$$\text{gamma } (\gamma) = (C - D) / (C + D) \qquad (4\text{-}8)$$

where C is the number of concordant pairs, the number of pairs of cases with the same relative order on each variable, and D is the number of discordant pairs, the number of pairs of cases with the reverse order on each variable.

Gamma tells us the extent to which same ordered, or concordant, pairs predominate over reverse order, or discordant, pairs as a proportion of the total number of pairs on which order can be determined.

In our example C = 4, D = 2 and gamma = (4-2)/(4+2) = 2/6 = .33. This indicates that our errors in predicting the relative order of pairs of cases on one variable within this sample is reduced by .33 once we know their relative order on the other variable.

Usually we have sample sizes larger than those in the last example, and we display our data in contingency tables or crosstabulations such as those used earlier in this section. When both of the variables in these contingency tables are measured on an ordinal scale we may talk about the relationship between these variables in terms of their relative order. We are essentially interested in how the two variables covary. When subjects have higher scores on one variable do they also have higher scores on the other variable? This situation is described as monotonic increasing and illustrated in part a of Table 4-10. Or when subjects have higher scores on one variable do they have lower scores on the other? This is called a monotonic decreasing relationship and is illustrated in part b of Table 4-10. Or is there no relationship between the two variables? This is shown in part c of Table 4-10.

Gamma describes the extent to which the relationship between two variables is monotonically increasing, monotonically decreasing, or non-existent. It does this by looking at all pairs of cases in a sample in which we can determine order (i.e., in which one is greater than or less than the other on the variables being studied) and examines the extent to which our ability to predict the relative order of any pair of cases is improved by knowing the relative order of that pair on the other variable. In other words, gamma tells us how much our error in predicting order of subjects on one variable is reduced once we know their relative order on the other variable. For part a in Table 4-10, our error would be reduced 100% and our gamma would be +1.00, indicating a perfect monotonic increasing function. For part b in Table 4-10, our error would also be reduced 100%, but gamma would be -1.00, indicating a perfect

Table 4-10
Examples of Possible Relationships Between Variables Measured on Ordinal Scale

a: A Monotonic Increasing Relationship

	Variable A				
Variable B	1	2	3	4	total
1	25	0	0	0	25
2	0	25	0	0	25
3	0	0	25	0	25
4	0	0	0	25	25
total	25	25	25	25	100

gamma = +1.00

b: A Monotonic Decreasing Relationship

	Variable A				
Variable B	1	2	3	4	total
1	0	0	0	25	25
2	0	0	25	0	25
3	0	25	0	0	25
4	25	0	0	0	25
total	25	25	25	25	100

gamma = -1.00

Table 4-10 (continued)

c: No Association

Variable B	Variable A 1	2	3	4	total
1	5	10	10	5	30
2	5	10	10	5	30
3	5	10	10	5	30
4	5	10	10	5	30
total	20	40	40	20	120

gamma = 0.00

monotonic decreasing function. For part c in Table 4-10, our error in predicting order would not be reduced at all, and gamma would be zero.

To expand upon these conclusions we may return to the formula for gamma given in line 4-8. If there are no discordant or reverse order pairs, D = 0, and

$$\begin{aligned} \text{gamma} &= (C - D) / (C + D) \\ &= (C - 0) / (C + 0) \\ &= C / C = 1.00, \end{aligned}$$

and we have a perfect monotonic increasing function.

If there are no same order or concordant pairs, C = 0, and

$$\begin{aligned} \text{gamma} &= (C - D) / (C + D) \\ &= (0 - D) / (0 + D) \\ &= -D / D = -1.00, \end{aligned}$$

and we have a perfect monotonic decreasing function.

Finally, if there are the same number of discordant and concordant pairs, C=D, and

$$\begin{aligned} \text{gamma} &= (C - D) \ / \ (C + D) \\ &= (D - D) \ / \ (D + D) \\ &= 0 \ / \ D = 0.0, \end{aligned}$$

and gamma equals zero.

It is important to note that the formula does not include any cases where the scores or ranks are tied. In computing gamma we simply ignore cases where subjects are tied on either of the two variables.

Table 4-11
Example of a Contingency Table for Computation of Gamma

Attitude Toward Women's Movement	Attitude Toward a Woman President 1-Strongly Dis Approve	2-Disapprove	3-Approve	4-Strongly Approve	Total
1) Strongly Dis Approve	n_{11} 3	n_{12} 1	n_{13} 1	n_{14} 0	5
2) Disapprove	n_{21} 2	n_{22} 2	n_{23} 1	n_{24} 0	5
3) Approve	n_{31} 1	n_{32} 1	n_{33} 2	n_{34} 1	5
4) Strongly Approve	n_{41} 0	n_{42} 1	n_{43} 2	n_{44} 2	5
Total	6	5	6	3	20

Table 4-11 gives another example, with the same variables used in Table 4-9, but with more cases. The data are displayed in a contingency table or cross-tabulation. Suppose that we wished to compute the number of same order pairs involved in this data set. Let us begin first with n_{11}, the cell in the upper left hand corner. There are 3 subjects in this cell. These people hold the same attitude on the women's movement as all the subjects in the first row and the same attitude on women presidents as all subjects in the first column. We say then that they are tied with those subjects and we cannot determine their relative order with these subjects. However, comparing these subjects with the

two cases in cell 22 produces a total of 6 (3 x 2 = 6) pairs of cases with the same relative ordering on both variables. Then comparing subjects in cell 11 with subjects in cell 23 we can see that those in cell 11 have lower scores on both variables. Because there is one subject in cell 23 this produces a total of 3 (3xl = 3) pairs of cases with the same relative order. In general, if we compare the subjects in cell 11 with subjects in all cells below and to the right we will see that those in cell 11 always have lower scores. This produces
n_{11} (n_{22} + n_{23} + n_{24} + n_{32} + n_{33} + n_{34} + n_{42} + n_{43} + n_{44}) =

= 3 (2 + 1 + 0 + 1 + 2 + 1 + 1 + 2 + 2) = 3 (12) = 36
pairs of cases involving n_{11} with the same relative ordering.

We can now move to cell n_{12} and repeat the process finding that there are n_{12} (n_{23} + N_{24} + n_{33} + N_{34} + n_{43} + n_{44}) = 1(1 + 0 + 2 + 1 + 2 + 2) = 1 (8) = 8 pairs of cases including n_{12} with the same relative ordering.

Using cell n_{13} we find that there are n_{13} (n_{24} + n_{34} +n_{44}) = 1 (0 + 1 + 2) = 1 (3) = 3 pairs of cases including n_{13} with the same relative ordering.

If we work through the entire table in this manner, we find that there are a total of 36 + 8 + 3 + n21 (n_{32} +n_{33} +$_{34}$ + n_{42} + n_{43} + n_{44}) + n_{22} (n_{33} + n_{34} + n_{43} + n_{44}) + n_{23} (n_{34} + n_{44}) + n_{31} (n_{42} + n_{43} + n_{44}) + n_{32}(n_{43} + n_{44}) + n_{33} (n_{44}) = 36 + 8 + 3 + 2 (1 + 2 + 1 + 1 + 2 + 2) + 2 (2 + 1 + 2 + 2) + 1 (1 + 2) + 1 (1 + 2 + 2) + 1 (2 + 2) + 2 (2) = 36 + 8 + 3 + 2 (9) + 2 (7) + 1 (3) + 1 (5) + 1 (4) + 2 (2) = 36 + 8 + 3 + 18 + 14 + 3 + 5 + 4 + 4 = 95 pairs of cases with the same relative ordering (concordant pairs).

To find the number of pairs of cases with reverse ordering (discordant pairs) we repeat the same procedure, but begin in the upper right hand corner of the table. Here we will use cell n_{13} since cell n_{14} has no cases. We may note that all subjects in the first row and the third column are tied with subjects in cell n_{13} and may be ignored. If, however, we compare subjects in cell n_{22} with those in cell n_{13} we will find that those in n_{22} have a lower score on attitudes toward women presidents but a higher score on attitudes toward the women's movement. This is a case of reverse ordering. Similarily, comparing n_{13} with n_{21}, we find that subjects in these two cells produce a case of reverse ordering. In general, there are n_{13} (n_{22} + n_{21} + n_{32} + n_{31} + n_{42} + n_{41}) = 1 (2 + 2 + 1 + 1 + 1 + 0) = 1 (7) = 7 pairs of cases with reverse ordering involving cell n_{13}. There are n_{12} (n_{21} + n_{31} + n_{41}) = 1 (2 + 1 + 0) = 1 (3) = 3 pairs of cases with reverse ordering involving cell n_{12}.

In general, to compute the number of reverse order (discordant) pairs for the table we would follow this procedure for all cells.

$+ n_{22} (n_{31} + n_{41})$

$D = 7 + 3 + n_{24} (n_{33} + n_{32} + n_{31} + n_{43} + n_{42} + n_{41}) + n_{23} (n_{32} + n_{31} + n_{42} + n_{41}) + n_{34} (n_{43} + n_{42} + n_{41}) n_{33} (n_{42} + n_{41}) + n_{32} (n_{41}) = 7 + 3 + 0 + 1 (1 + 1 + 1 + 0) + 2(1 + 0) + 1 (2 + 1) + 2 (1) + 1 (0) = 7 + 3 + (3) + 2 + (3) + 2 = 20$

We may now use our general formula to compute gamma:

$$(C-D)/(C+D) = (95 - 20)/ (95 + 20) = 75 / 115 = .65$$

This tells us that there are 75 more concordant or same ordered pairs than discordant or reverse ordered pairs in this table and that there are 115 pairs all together in which order can be determined. The gamma value of .65 tells us that we reduce our error in predicting the order of a pair of cases on one variable by 65% if we know their relative order on the other variable. The fact that the value of gamma is positive indicates that this is an increasing function and that subjects who score highly on one variable tend to score highly on the other.

To review the characteristics of gamma: Gamma is a symmetric measure, that is, no differentiation is made between the independent and dependent variable. We need at least an ordinal level of measurement for our variables in order to compute gamma. Gamma varies from -1 to +1. When it holds a value of zero we say that there is no association between the two variables, in the sense that the number of like-ordered or concordant pairs of cases equals the number of reverse ordered or discordant pairs of cases. A gamma of minus one indicates a perfect monotonic decreasing function; a gamma of positive one indicates a perfect monotonic increasing function. Gamma does not take the relative ranking of the variables into account, but only their relative order. Thus, a pair of cases where one subject strongly agreed and the other strongly disagreed and a pair where one strongly agreed and the other agreed would both be treated in the same way. Finally, gamma has a PRE interpretation. This is not immediately apparent from the formula and the proof is a little too complex for this class. Suffice it to say that a value of gamma may be seen as indicating the percentage of errors in predicting the relative order of a pair of scores on one variable that are eliminated when we know their relative order on a second variable.

It should be noted that there are many measures of association designed for variables measured on an ordinal scale. They all use the difference of the number of concordant and discordant pairs in the numerator, but tend to differ in what is placed in the denominator. This generally involves the number of pairs on which variables are tied, either on the independent or dependent variable or both, plus the total number of pairs on which order can be determined. Because this denominator is larger than that used for gamma, these other measures of association are usually smaller than gamma for the same table.

V. Simple Bivariate Correlation

Quantitative analyses are always linked with research designs. We use statistics to help answer research questions, and we must always adapt our analysis to fit the characteristics of our measures. When we are interested in the degree to which two variables measured on a nominal scale are associated, we use lambda; to examine the association between two ordinally measured variables we use gamma. What, however, if we were interested in the association between two variables, which were both measured on an interval scale? For instance, one could be interested in the association between income and education. What happens to peoples' income as their level of education rises? When studying a question such as this, where both the dependent and the independent variable are measured on an interval scale, we may use regression or correlation techniques.

Below we explore the elements of basic bivariate correlation. We develop the use of the regression line, explain the PRE measure of association, r^2, and also discuss r, the square root of r^2, which is often called the Pearson product moment correlation.

r^2 as a Measure of Association

Consider a case when you have two variables, each measured on an interval scale. What if you thought there were some pattern in the association between the two variables? Suppose they had a positive linear association, as one variable went up, so did the other (as in the example of income and education above) or what if as one variable went up the other went down (say as in an association between educational level and amount of superstitious beliefs), a negative linear association. Figure 5-1 below illustrates the possible association between the income and education of a group of people. On the horizontal axis the amount of education is represented from high to low. On the vertical axis the amount of income is shown. Each dot represents one person. It is apparent that people with low amounts of education tend to have lower incomes, people with higher educations tend to have incomes.

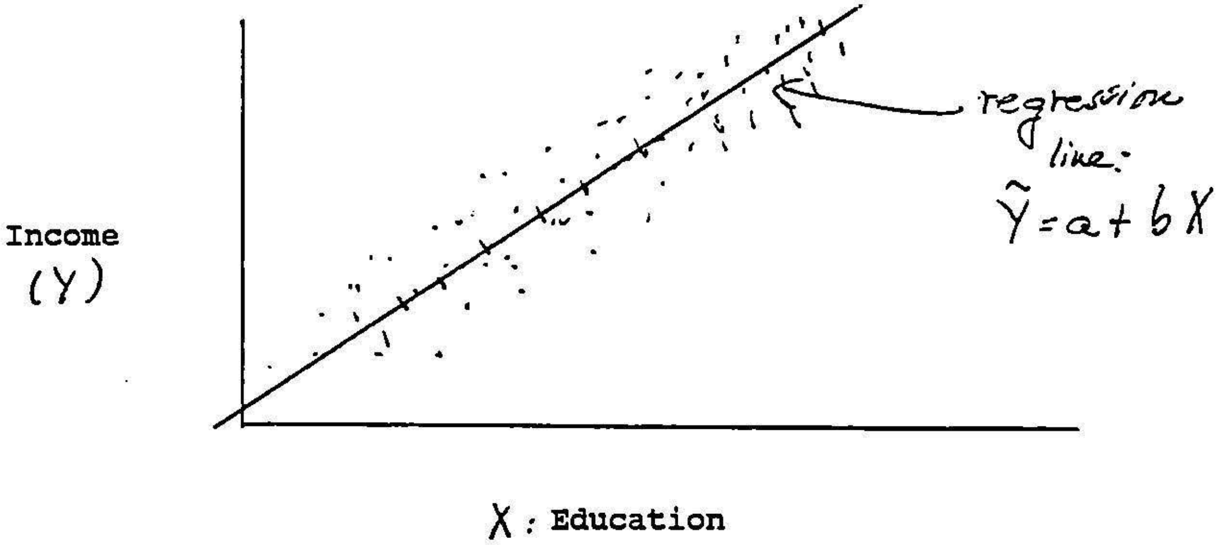
It is possible to draw a straight line through this diagram so that it falls as close to each element of the sample as possible. Such a line is drawn through Figure 5-1.

From elementary algebra you will remember that the equation for a line is

$$Y = a + b X \qquad (5-1)$$

where Y is the dependent variable, the variable on the vertical axis; and X is the independent variable, on the horizontal axis. The value "a" is the Y-intercept, the value of Y when X = 0 or the point where the line crosses the vertical axis. The value "b" is the slope of the line, the amount of changes in Y for each unit change in X.

Figure 5-1
Relationship between Income and Education
for a Hypothetical Group of People



Based on the actual data on two variables for a sample it is possible to construct a line that best predicts the scores of Y, the dependent variable, from the scores of X, the independent variable. This equation is called the <u>regression equation</u> and is written

$$\hat{Y} = a_{yx} + b_{yx} X \qquad (5\text{-}2)$$

where $\hat{Y}$ is the predicted value of Y for any X, a_{yx} is the y - intercept, b_{yx} is the slope of the line, and X is any value of the independent variable. The subscripts, yx, indicate that the coefficients in the equation are predicting the variable Y from the values of X.

Now, because it is possible to construct this line so that it is the best line that predicts values of Y from those of X, we can use these predicted values of Y, $\hat{Y}$, as our best predictors of the dependent variable when we know the values of the independent variable and when we assume the two variables have a linear association. Because $\hat{Y}$ is our best predictor of Y when we assume that the association between X and Y is linear, $\sum(Y - \hat{Y}) = 0$, and $\sum(Y - \hat{Y})^2$ is a minimum for any value of $\hat{Y}$ that can be developed through an equation of the form $a_{yx} + b_{yx}$ X (where X is the value corresponding to that X in the scatter diagram).

Remember that $\sum(Y - \bar{Y})^2$ is our measure of error when all we know is the dependent variable, for the mean is always the best predictor of an intervally measured variable.

Note that we now have all the elements of a PRE measure. We have a rule for classifying subjects on the dependent variable when we only know the dependent variable: We simply would give them the score of the mean, for our deviations around the mean are at a minimum for any value. Our rule for classifying subjects on the dependent variable when we know the independent variable is the regression line, for deviations of scores around the regression line are also at a minimum. Our definition of error can simply be squared deviations of scores around these points (we square to get rid of negative values.)

For the first rule

$$E_1 = \sum(Y-\bar{Y})^2 \qquad (5\text{-}3)$$

or the squared deviations of scores around the mean.

For the second rule

$$E_2 = \sum(Y-\hat{Y})^2 \qquad (5\text{-}4)$$

or the squared deviations of scores around the regression line.

Remember that a PRE measure is $(E_1 - E_2)/E_1$. From the definitions of E_1 and E_2 above we can then construct the following measure of association:

$$\frac{(E_1 - E_2)}{E_1} = \frac{\sum(Y-\bar{Y})^2 - \sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2} = r^2 \qquad (5\text{-}5)$$

In this measure the total variation to be explained, or the error when we only know the dependent variable is $\sum(Y - \bar{Y})^2$. The variation unexplained or left around the regression line, the error when we also know the linear association with the independent variable, is $\sum(Y - \hat{Y})^2$. The difference between the total variation and the unexplained variation is the variation of the dependent variable that is explained by the regression line or by the linea r association between the dependent and independent variable. This ~~measure~~ is r^2. It is the square of the Pearson product moment correlation. It is simply interpreted as the proportion of the variation in the dependent variable (or one variable) that is explained by

(as a proportion of the total variation) its linear association with the independent (or other) variable. It may also be seen as the proportionate reduction of error in predicting values of the dependent variable when we know the linear association between the two variables compared with our error when we only know the dependent variable.

An Example

A simple example can illustrate the meaning of r^2 and its relation to the regression line. Figure 5-2 shows a scatter diagram of data representing the reported monthly church attendance of pairs of mothers and daughters. These data are also summarized in Table 5-1. Note that in family A both mother and daughter attended once in the month; in family B mother attended twice, daughter 3 times; in family C mother attended four times and daughter 3, and so on.

Figure 5-2

Scatter Diagram of Hypothetical Data Regarding the Monthly Church Attendance of a Sample of Mothers and Daughters

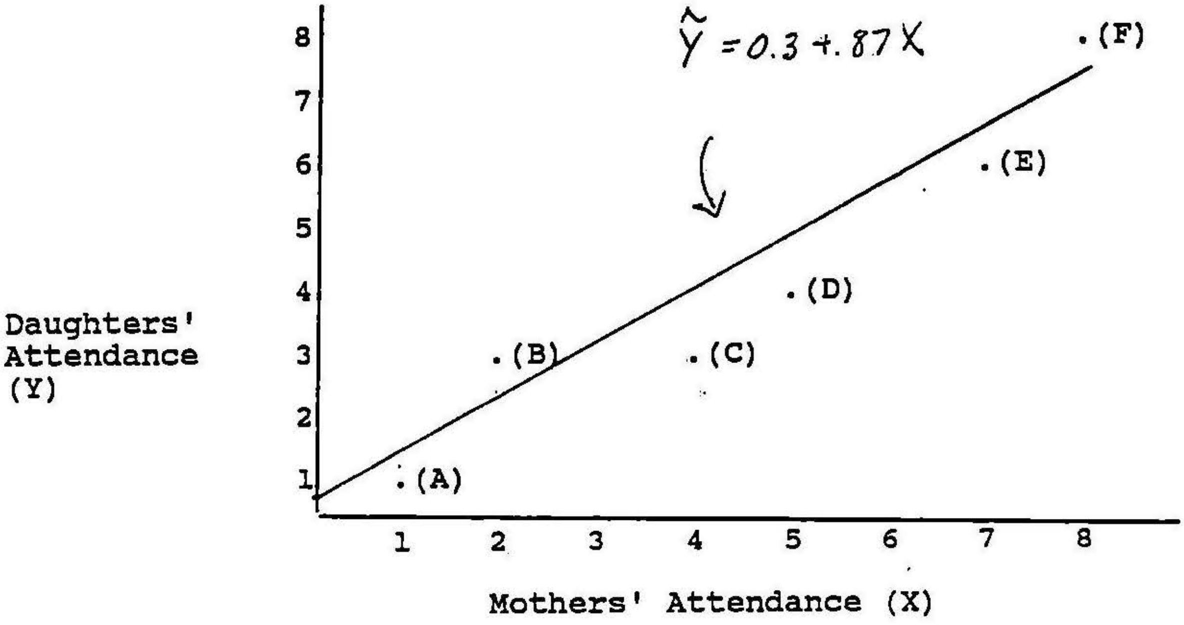


Table 5-1

Computations Needed to Compute r^2 for data in Figure 5-2

Family	Mother (X)	Daughter (Y)	$(X-\bar{X})$	$(X-\bar{X})^2$	$(y-\bar{Y})$	$(X-\bar{Y})$ $(Y-\bar{Y})$
A	1	1	-3.5	12.25	-3.2	11.2
B	2	3	-2.5	6.25	-1.2	+3.0
C	4	3	-0.5	0.25	-1.2	+0.6
D	5	4	- .5	0.25	-0.2	-0.1
E	7	6	2.5	6.25	1.8	+4.5
F	8	8	3.5	12.25	3.8	+13.3
				37.5		
Totals	27	25				32.5

$\bar{X} = \frac{27}{6} = 4.5;$ $\bar{Y} = \frac{25}{6} = 4.2$

A scatter diagram, as in Figure 5-2 is a device used to illustrate the nature of the association between two variables. From the scatter diagram in Figure 5-2 it appears that there is a positive linear association between the daughter's church attendance and the mother's church attendance. As the mother has higher church attendance, so does the daughter.

Now we want to construct a line that can be drawn through this scatter diagram that will best predict values of Y (the daughter's attendance) from our knowledge of the mother's attendance (X). I will not here go through the derivation of the formulas used to get values of b_{yx} and a_{yx}. They involve a knowledge of elementary calculus. Suffice it to say that mathematicians have figured out the equations that will produce these best predictors.

However, an intuitive explanation of the formula for b is possible.

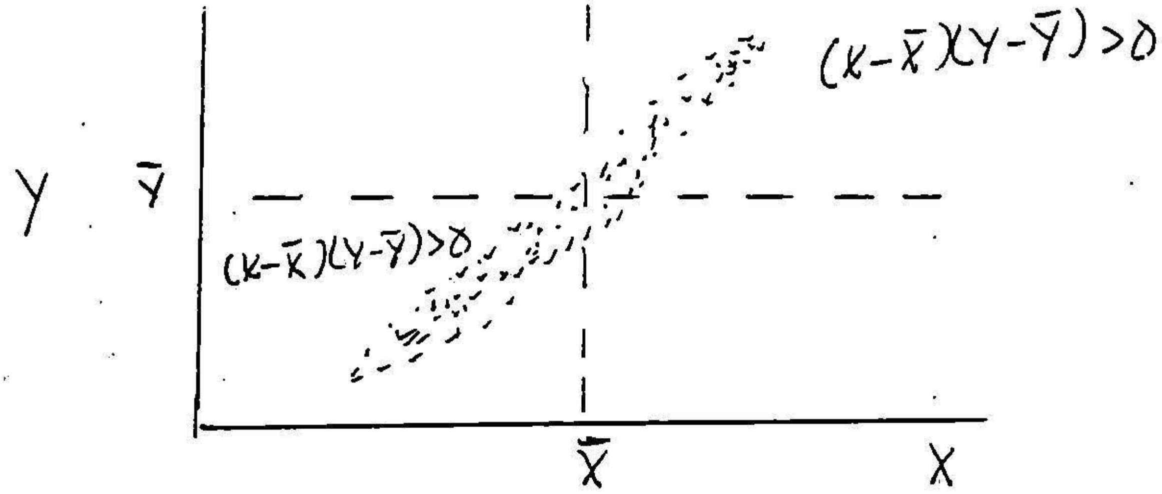
$$b_{yx} = \sum(X - \bar{X})\ (Y - \bar{Y}) / \sum(X - \bar{X})^2 \qquad (5\text{-}4)$$

This is simply the covariation of X and Y $[\sum(X-\bar{X})(Y-\bar{Y})]$ divided by the variation of X $[\sum(X-\bar{X})^2]$, the predictor or independent variable.

Remember that b_{yx} is the slope of the regression line. When it is greater than zero there is a positive association; when it is less than zero there is a negative association (as one variable goes up the other goes down) and when there is no association the slope is approximately equal to zero. The variation of X is always greater than zero (if it were equal to zero there would be no use in conducting an analysis). Thus, to understand how this formula for the slope can result in a positive, negative, or zero value, we need look only at the covariation of X and Y.

We can see that the covariation includes information about the mean of X, which is our best predictor of X when we only know X, and about the mean of Y, our best predictor of Y when we only know about that variable. The covariation takes into account how the actual <u>pairs</u> of scores vary around the best two predictors for each variable. Figures 5-3, 5-4, and 5-5 illustrate situations that will result in different values of b.

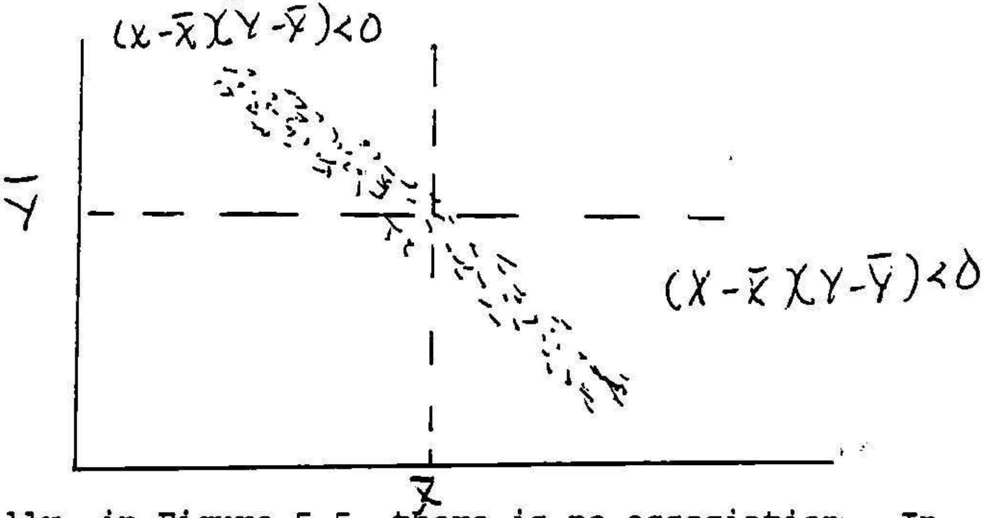
Figure 5-3
A Positive Value of r and b



In Figure 5-3, because the relation is positive most of the pairs of scores fall into the quadrant where both Y-Y and X-X are greater than zero, or in the quadrant where both of these values are less than zero. In both these cases the product of $(Y - \bar{Y})$ $(X - \bar{X})$ would be positive (positive times positive = positive; negative times negative = positive) and thus the covariation would be positive, and b or the slope would be positive.

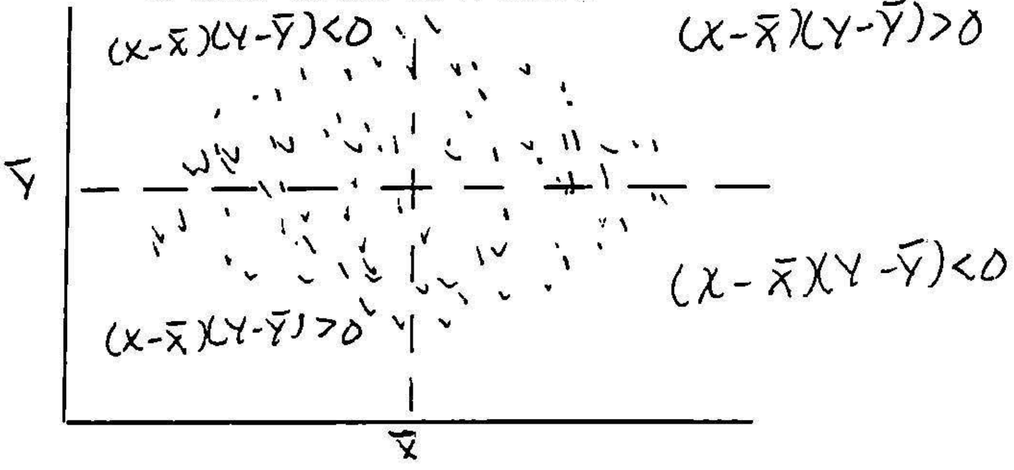
In Figure 5-4 the association is negative. In this case most of the cases fall into the quadrant where $(X - \bar{X})$ is less than zero and $(Y - \bar{Y})$ is greater than zero, or into the quadrant where $(X - \bar{X})$ is greater than zero and $(Y - \bar{Y})$ is less than zero. In this case the product of $(Y - \bar{Y})$ $(X - \bar{X})$ would usually be negative and thus b and the slope would be negative.

Figure 5-4
A Negative Value of r and b



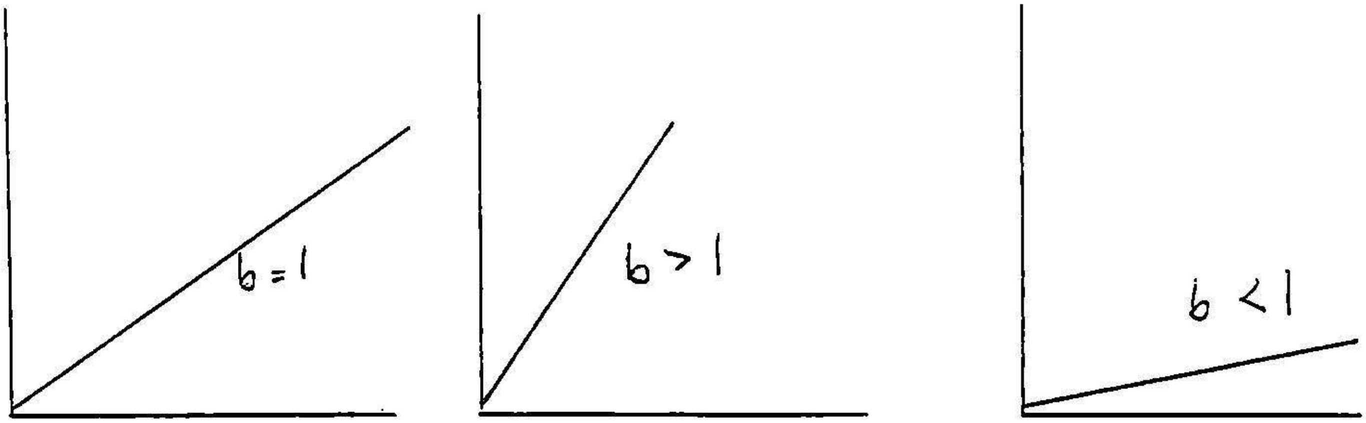
Finally, in Figure 5-5, there is no association. In this case the pairs of cases generally fall equally between the four quadrants. Thus the number of times the product of (X-X) and (Y-Y) is positive should about balance off the number of times the product is negative and thus the overall sum of these products over all cases would be close to zero.

Figure 5-5
A zero value of r and b



If the variation in X ($\Sigma(X-\bar{X})^2$) is about equal to the covariation of X and Y [$\Sigma(X-\bar{X})(Y-\bar{Y})$], then b would be approximately equal to one. This means that the changes in X and Y are about equal, as X moves one unit, Y is predicted to move about one unit. When b is greater than one, the covariation of X and Y is greater than the variation in X, and when X changes one unit Y is predicted to change by more than one unit. Conversely, when b is less than one, the covariation of X and Y is less than the variation in X, and the predicted changes in Y are less than the unit changes in X. Each of these situations is illustrated in Figure 5-6.

Figure 5-6



Given this intuitive feel for the meaning of b_{yx}, let us return to the example involving mothers' and daughters' church attendance. Using the information given in Table 5-1 we can calculate:

$$b_{xy} = \frac{\Sigma(X-\bar{X})\ (Y-\bar{Y})}{(X - \bar{X})^2} = \frac{32.5}{37.5} = .87 \qquad (5\text{-}5)$$

$$a_{yx} = \bar{Y} - b_{yx}\,\bar{X} = 4.2-3.9 = 0.3 \qquad (5\text{-}6)$$

$$= 4.2 - 3.9 = 0.3$$

The regression line that best predicts the values of Y, the daughter's attendance, from the values of X, the mothers' attendance is:

$$\hat{Y} = 0.3 + .87\ X \qquad (5\text{-}7)$$

In Table 5-2 we present the data that can be used to develop the measure of association r^2. This measure tells us how much of the variation in daughters' church attendance can be accounted for by its linear association with mothers' attendnace.

Table 5-2
Data for Calculating r^2 for data in Figure 5-2

Family	X	Y	$\hat{Y}$	$Y - \bar{Y}$	$(Y-\bar{Y})^2$	$(Y-\hat{Y})$	$(Y-\hat{Y})^2$
A	1	1	1.17	-3.2	10.24	-.17	.03
B	2	3	2.04	-1.2	1.44	+.96	.92
C	4	3	3.78	-1.2	1.44	-.78	.61
E	5	4	4.65	-0.2	.04	-.65	.42
D	7	6	6.39	1.8	3.24	-.39	.14
F	8	8	7.26	3.8	14.44	.74	.55
				-5.8	30.84	-1.99	2.67
Totals				+5.6		+1.70	
				-.02		-.29	

Note that the simple sum of deviations of the scores of the dependent variable around the mean are approximately equal to zero. Thus the sum of the squared deviations around the mean are also at a minimum. The predicted values of Y shown in the table are those computed when the given value of X, the mothers' attendance for each family, is substituted in the prediction equation. The simple sum of the scores of the dependent variable around the predicted values of Y from this regression line are approximately equal to zero, and the sum of the squared deviations around the regression line are at a minimum.

We may now use the sum of the squared deviations around these two best predictors to compute r^2. $\Sigma(Y-\bar{Y})^2$ = the variation of scores around the mean, the best predictor when we only know the dependent variable. $\Sigma(Y-\hat{Y})^2$ = the variation of scores around the point on the regression line that is predicted for that family or pair of scores. This is our best predictor of the dependent variable when we know the independent variable and assume that the association between the two variables is linear.

$$r^2 = \frac{\Sigma(Y-\bar{Y})2 - \Sigma(Y - \hat{Y})2}{\Sigma(Y-\bar{Y})^2} = \frac{30.84 - 2.67}{30.84} = \frac{28.17}{30.84} = .91 \qquad (5\text{-}7)$$

Thus, for this sample, when we know the mother's frequency of church attendance we can reduce our error in predicting the daughter's attendance by 91% when we assume that the association between the two variables is linear (can be represented by a straight line). Another way of saying this is that 91% of the total variation in the daughter's church attendance can be explained by its linear association with the mothers' frequency of church attendance.

Note that r^2 is a symmetric measure. In fact, we could work out the equation predicting X from values of Y and compute r^2 that way and come up with the same figure. We could also say then that 91% of the variation in mothers' church attendance is explained by its linear association with daughters' frequency of church attendance.

r^2 is sometimes called the coefficient of determination, representing the extent to which one variable is determined by another. $1 - r^2$ (in this case = .09) is called the coefficient of alienation, the proportion of variation that is not explained by this linear association.

Because our way of computing r^2 above used the definitional formula involved a number of subtractions, ~~and~~ it ~~thus~~ is bound to introduce rounding errors. When you compute r^2 by hand it is preferable to use a computational formula. This is usually written for the value of r itself. To get r^2 we simply square this value. The computational formula for r is simply

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \qquad (5\text{-}8)$$

in our example $r = \dfrac{6(145) - (27)(25)}{[6(159) - (27)^2][6(135) - (25)^2]}$

$$= \frac{195}{[225][185]} = \frac{195}{41{,}625} = \frac{195}{204.02} = .95$$

$$r^2 = (95)(.95) = .90$$

<u>The Pearson Product Moment Correlation, r</u>

While r^2 has an easily understood interpretation in the PRE format, the pearson product moment correlation, r, is more frequently used. While r^2 varies between 0 and 1 (with 0 indicating no association and 1 indicating perfect association), r varies from -1.0 to +1.0. r and r^2 are obviously related in that r^2 is simply the value of r multiplied by itself. Yet, the interpretation of r is somewhat different than the interpretation for r^2. Below we go through four interpretations related to r after exploring more the formula for r itself.

Above we gave the computational formula for r. It is also instructive to examine the definitional formula. The definition of r is

$$r = \frac{\Sigma(X - \bar{X})\ (Y - \bar{Y})}{\sqrt{(X - \bar{X})^2 \quad (Y - \bar{Y})^2}} = \frac{\text{covariation of X and Y}}{\sqrt{\text{(variation of X)(variation of Y)}}} \qquad (5\text{-}9)$$

Note that this is closely related to the definitional formula of the slopes:

$$b_{yx} = \frac{\text{covariation (XY)}}{\text{variation X}} \qquad b_{xy} = \frac{\text{covariation XY}}{\text{variation Y}} \qquad (5\text{-}10)$$

While the slope always has the covariation of X and Y in the numerator, the denominator is either the variation of X or the variation of Y depending on whether X or Y is the predictor variable.

$$r^2 = b_{yx}\ b_{xy} \text{ and thus } r = \sqrt{b_{yx}\ b_{xy}} \qquad (5\text{-}11)$$

The various possible interpretations of r follow these observations. First, by observing the sign associated with the correlation coefficient, we may ascertain whether the association between the two variables is positive or negative. This follows from the logic associated with the sign associated with the slope as explained earlier.

Second, we may simply square the value of r to get r^2, which tells us the proportion of variation in one variable explained by its linear association with the other. This was fully discussed above.

Third, we may remember that r is equal to the square root of the product of the two slopes. This is called a geometric mean, one type of measure ~~of measures~~ of central tendency. The correlation coefficient then is the geometric mean or geometric average of the two different slopes b_{yx} and b_{xy}.

Fourth, r may be interpreted as the slope of the regression line when standard deviation units are used as scores rather than the raw scores. Figure 5-8 illustrates this interpretation for the example used in the previous section. As shown in Table 5-3, each of the scores may be transformed to its corresponding z-score or standard deviation unit score. Based on these scores we may compute b_{yx} and b_{xy}. Note, however, that $b_{z_y z_x} = b_{z_x z_y} = r_{z_x z_y}$.

In other words, r is simply the change in standard deviation units in y for every standard deviation unit change in X.

Also, r = $\sum(z_y z_x) / N$, or the average of the cross-product of the standard errors. This occurs because when standard scores are used the standard deviation of the standard scores is automatically one and the mean is 0. (Remember that the definition of standard scores or z-scores is a distribution where the mean is 0 and the standard deviation is 1). This then means that the sum of the squared deviations of scores from the mean simply equals the sample size, as shown in equations 5-12 and 5-13.

$$s_{zx}^2 = 1 = \frac{\sum(Z_x - \bar{Z}_x)^2}{N} = \frac{\sum(Z_x - 0)^2}{N} \qquad (5\text{-}12)$$

and by multiplying N by each side of the equation:

$$N = \sum(Z_x - \bar{Z}_x)^2 = \sum(Z_x - 0)^2 = \sum Z_x^2 \qquad (5\text{-}13)$$

Thus, the sum of the squares deviations around the mean are simply equal to the sample size. This means that ~~the cross-product = $\sum z_y z_x$ = N^2,~~ and the square root of the product of the variations is equal to the sample size.

$$(\quad \sqrt{\sum(X-\bar{X})^2 \quad \sum(Y-\bar{Y})^2} = \sqrt{(N)(N)} = (N) \qquad (5\text{-}14)$$

Table 5-3
Calculations of r and r^2 for Data
in Table 5-1 Using Standard Scores

Family	X	Y	Zx	Zy	ZxZy
A	1	1	-1.3	-1.33	+1.73
B	2	3	- .92	- .49	+ .45
C	4	3	- .18	- .49	+ .09
D	5	4	+ .18	- .07	- .01
E	7	6	+ .92	+ .77	+ .71
F	8	8	+1.3	+1.61	+2.09
	27	25	0	0	5.1

$\bar{X}$ = 4.5
$\bar{Y}$ = 4.17
S_x = 2.7
S_y = 2.38
n = 6

$$Z_x = \frac{X-\bar{X}}{S_x} \quad Z_y = \frac{Y-\bar{Y}}{S_y} \qquad r = \frac{\sum Z_x Z_y}{N} = \frac{5.1}{6} = .85 \approx .9$$

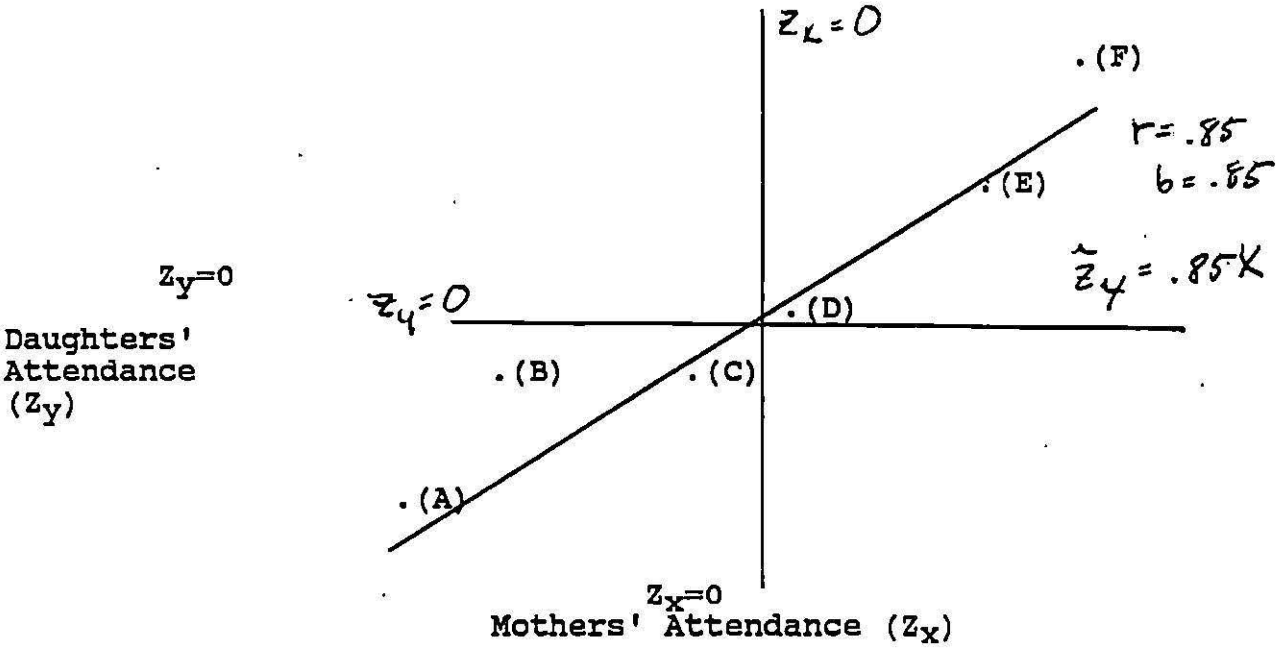
And these results are equal, when rounding errors are taken into account, to those found through other computation methods above

This final interpretation of r is the one that will be the most useful. From it, one can interpret r as being the standard deviation unit change in the other variable. This is analagous to the interpretation of the slopes, but

in one variable produced by one
standard deviation unit change

involves the use of standard scores rather than actual scores. That is, the value of r tells us how many standard deviation units we would expect one variable to change when the other changes one standard deviation unit. The result above says that we would expect daughters to have church attendance patterns that were .85 of a standard deviation higher than the average when mothers' church attendance was one standard deviation above the average. Similarly, if mothers had church attendance patterns that were one standard deviation below the average for mothers, we would expect daughters' church attendance patterns to be .85 of a standard deviation below the average.

Figure 5-8
Illustration of r and b when using standard scores with data from Figure 5-2



The term Pearson product moment correlation also comes from the definition of r as $\Sigma z_x z_y$ / N. A moment is an average. The mean is the first moment (the average of the scores). The variance is the second moment (the average of the squared deviations of the scores around the mean). Here we are averaging the products of standard scores, thus, the product moment correlation. Karl Pearson is the mathematician who developed the statistic, and thus the name Pearson.

Computer Work

Various computer programs can provide scatter diagrams and compuations of r and r^2. The output shown below in Figures 5-9 and 5-10 come from data from a western Oregon high school. I requested two scatter diagrams, both looking at the association between scores on a general achievement test taken in the eleventh grade (called VAR11 by the computer) and the students' average grades in the seventh grade (called VAR15). I posited that the grades were dependent upon achievement. These calculations were requested for each social class group. Results for the middle class are given first, results for the working class are given second. Each * on the table represents one person at the intersections of those points. If more than one person falls at a point the computer prints the number of people involved. Note that the cases cluster around the regression line. [In asking for this output I asked the computer to have the plot lines be equal to integer values. This makes it somewhat easier to draw the regression line, but it also results in all of the data being in the lower half of the table (because gpa was measured to two decimal points, but spans only 4 integer values).]

The associated regression line is drawn on both printouts. To draw the line the values for a and b were taken from the printout and the equation for the line developed. Then predicted values of Y were computed for 3 separate values of X and resulting points were plotted. For the middle class students

$$\hat{Y} = 2.322 + .013X. \qquad (5\text{-}15)$$

For the working class students

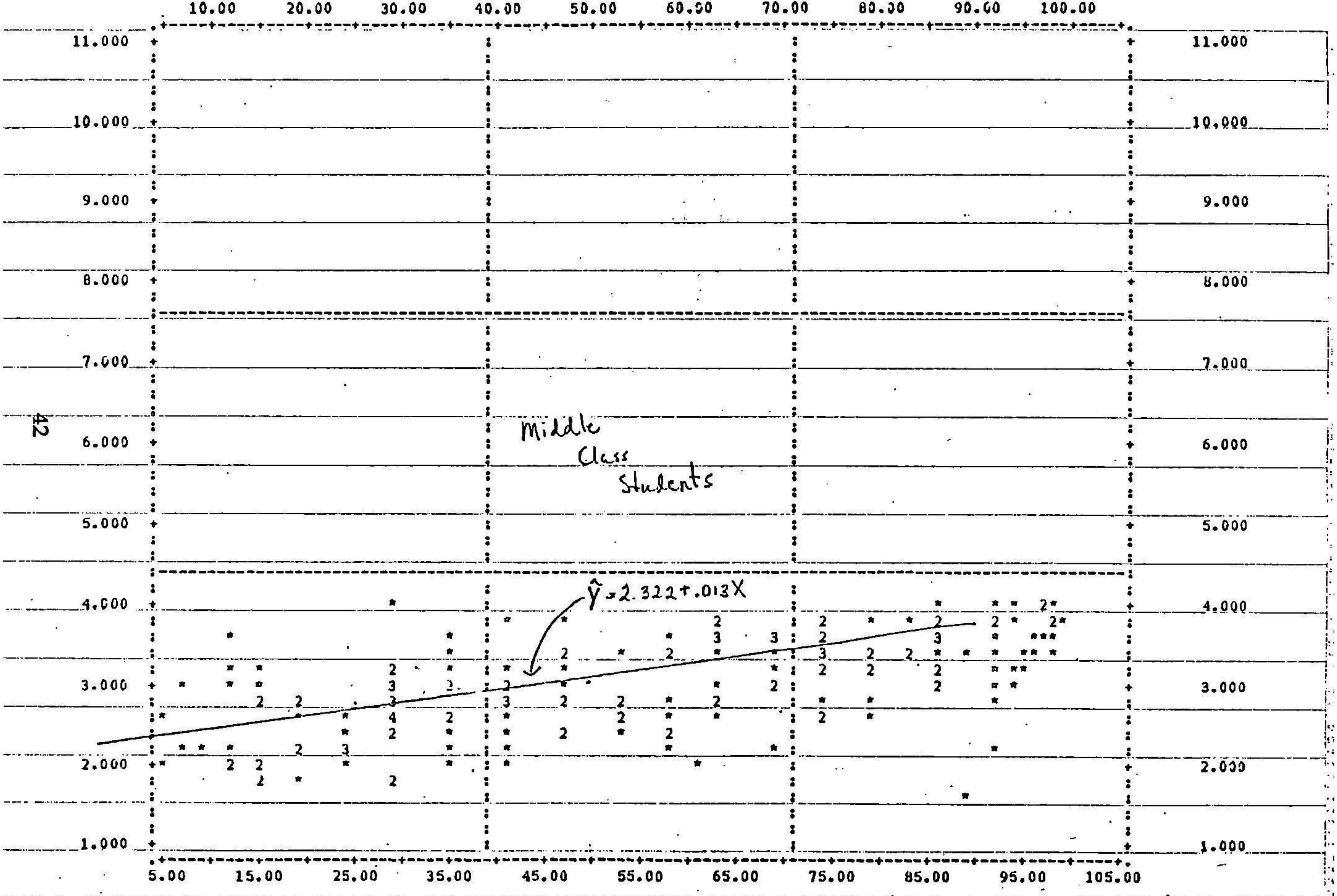
$$\hat{Y} = 2.202 + .012\ X. \qquad (5\text{-}16)$$

Note again that this is the regression line predicting gpa from achievement. GPA is the variable on the vertical axis of the scatter diagram. Note also that both the y-intercept and the slope are lower for the working class students than for the middle class students.

```
EXAMPLE OF CROSSTABS FOR 326                                    13-May-83       Page  2

File    COLEMAN  (Creation date = 01/19/79 )  REPLICATION, SPRINGFIELD 78 DATA
Scattergram of   (down) VAR11    11TH GPA                    (across) VAR15    11TH ITED COMPOSITE
```

Figure 5-9



EXAMPLE OF CROSSTABS FOR 326 Figure 5-10 13-May-83 Page 5

```
File   COLEMAN  (Creation date = 01/19/79 )  REPLICATION, SPRINGFIELD 78 DATA
Scattergram of   (down) VAR11    11TH GPA                (across) VAR15    11TH ITED COMPOSITE
```

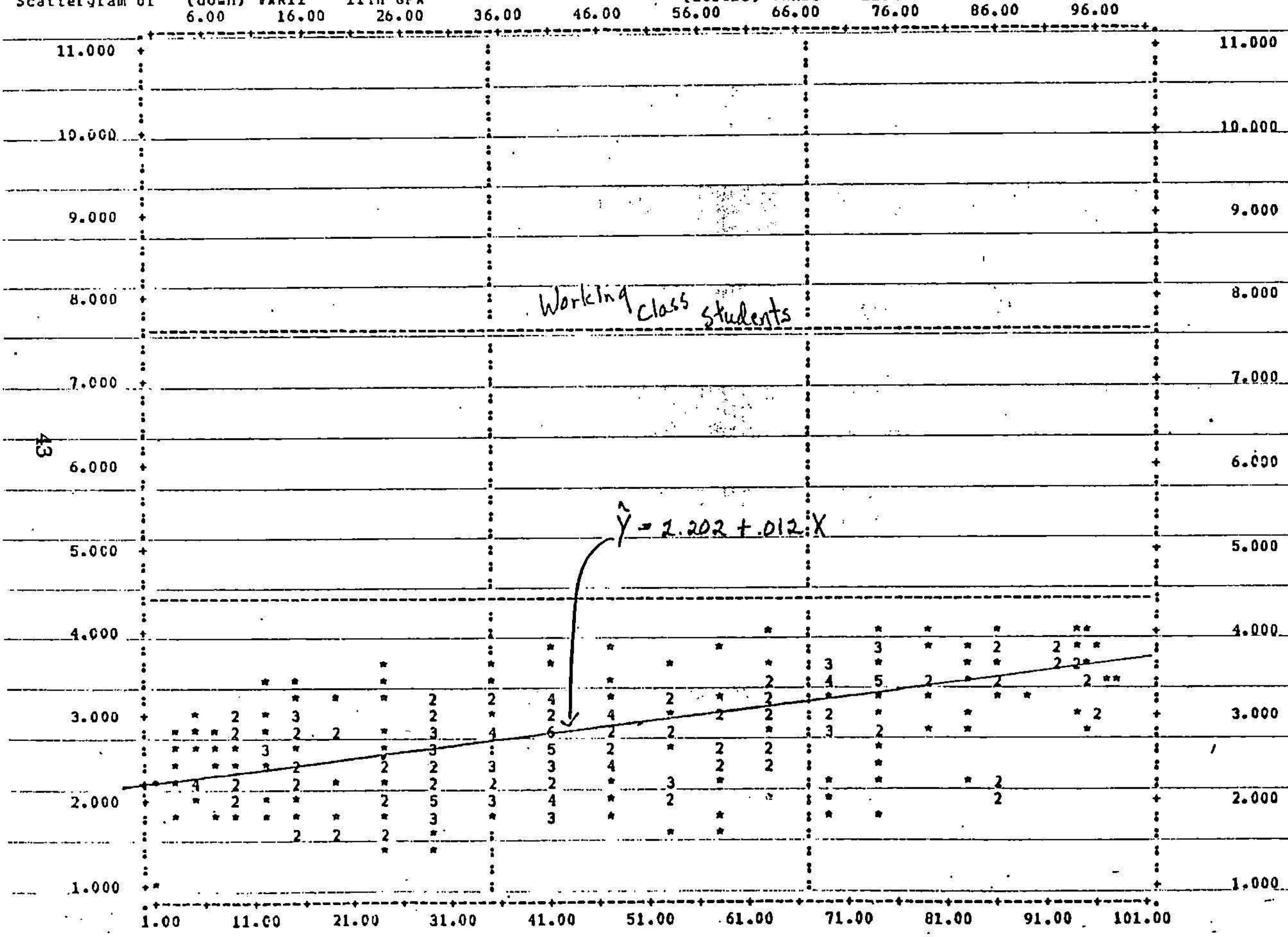


Table 5-4
Calculation of Grades predicted for Middle Class and Working Class Students at Various Levels of Achievement

Achievement	Predicted Grades		
Test Scores	Middle	Working	
(Percentiles)	Class	Class	Difference
X	$\hat{Y}_{mc}$	$\hat{Y}_{wc}$	$\hat{Y}_{mc} - \hat{Y}_{wc}$
0	2.322	2.202	0.102
25	2.647	2.502	0.145
50	2.972	2.802	0.170
75	3.297	3.102	0.195

Table 5-4 shows the results of using the regression equation to compute predicted values of the gpa for working and middle class students using their respective regression equations. It may easily be seen that at all values of achievement middle class students have higher predicted grades than working class students. Because the slope in the regression equation is larger for middle class students than for working class students the gap or difference between the predicted grades becomes larger with higher achievement scores, reaching almost .2 of a grade point for students with achievement test scores at the 75th percentile.

Looking again at the printout results it may be seen that the r^2 between achievement and grades is .36 for middle class students, but .26 for working class students. If we know the linear association of students' achievement scores with their grades we may account for over one-third of the variation in middle class students' grades but only about one-fourth of the variation in working class students' grades. Not only do middle class students receive higher scores than working class students when they have equal achievement, but the variation of scores around the regression line is much smaller for middle class students than for working class students.

Study Guide - Midterm Two
Sociology 326
Winter, 1992

You should understand each of the following:

The characteristics of the normal curve, and how to sketch one given its mean and standard deviation

How to read the table of the normal curve

How to compute z-scores and what they mean

How to determine the proportion and number of cases within a certain range of values in a normal distribution

What a measure of association is

What a PRE measure of association is, and what various values of a PRE measure would indicate regarding a distribution

How to percentage a bivariate table and interpret the results

When to use, how to compute, and how to interpret lambda, gamma, r, and r^2

How to put data into a scatter diagram, how to determine and compute a regression line, how to compute r^2 and r, and how to interpret these values

The exam will be at the regular class time, Friday, Feb. 21. You may use all of your books and notes during the exam. In fact, be sure to bring your text so that you have the table of the normal curve with you. It might also be helpful to bring a calculator.

Packet 149

SOC 326

QUANTITATIVE METHODS

Professor Stockard

University of Oregon

Winter Term 1992

UP-84
c

kinko's

the copy center

860 E. 13th
Eugene • 344-7894

Copies:	$2.68
Binding	$0.00
Royalties	$0.00
Permission Handling Charges	$0.00
Total cost of packet:	$2.68

TABLE OF CONTENTS

Jean Stockard - Packet 149

VI. Probability and Statistical Inference

In this section we examine issues related to probablity and statistical inference. Remember that there are two types of statistics: Descriptive statistics tell us about the nature of a sample or a population. Inferential statistics allow us to see how typical or true a given descriptive statistic from a sample is of a population from which it came. That is, inferential statistics help us generalize from samples to populations. Throughout the first part of this class we have explored various descriptive statistics. Now we are going to look at inferential statistics.

The notions of probability and randomness are basic to this discussion and thus in the first sections below we discuss these concepts. We then move to basic definitions which underlie inferential statistics, and then present examples of the two basic types of inferential statistics: confidence intervals (or point estimation) and hypothesis testing.

Probability

We use the idea of probability all of the time, often without even thinking about it. For instance, we look out the window each morning and wonder how likely it is that it will rain so that we will know whether or not to carry an umbrella. If it is dark and cloudy you will probably take an umbrella; if it is bright and sunny you probably won't take an umbrella. Similarly, in studying how much to study for an exam students decide how probable or likely it is that certain material will be included on the test. If they are certain some material will be on the exam they will study that material much more thoroughly than material they will not be tested on. In both of these cases you are interested in probabilities or likelihoods that certain events will occur.

The term probability simply refers to the possible outcomes of a situation -- how likely or unlikely is it that a given situation will occur. For instance, if you know that it will rain 3 out of the next 10 days, the probability is 3/10 = .30 that it will rain on any one day. Similarly, if a teacher gives you 20 questions that may be on the test, but will choose only 5 of those for the actual exam, the probability that any one question of the 20 will be on the test is 5/20 = .20. In general the probability that an event A will occur is symbolized as p(A) and equals the actual number of times the event occurs divided by the total number of events (see equation 6-1). A probability varies from 0 (which would indicate no chance of occurance) to 1.00 (which would indicate than an event would always occur).

$$p(A) = \frac{\text{number of times A occurs}}{\text{total number of events}} \qquad (6\text{-}1)$$

Consider the data given in Table 6-1 (seen earlier as Table 4-3 in discussion of lambda). Using equation 6-1 it may be seen that in this distribution the probability that a given teenager will be unemployed is

$$p \text{ (unemployed)} = 325/800 = .41 \qquad (6\text{-}2)$$

The probability that a teenager will be employed is

$$p \text{ (employed)} = 475/800 = .59 \qquad (6\text{-}3)$$

The probability that a teenager will be Anglo is

$$p \text{ (Anglo)} = 300/800 = .375 \qquad (6\text{-}4)$$

The probability that a teenager will be African American or Hispanic is

$$p \text{ (African-American)} = p \text{ (Hispanic)} = 200/800 = .25 \qquad (6\text{-}5)$$

The probability that a teenager will be in the "Other" category is

$$p \text{ (Other)} = 100/800 = .125 \qquad (6\text{-}6)$$

Table 6-1
Joint Distribution of Race/Ethnicity and Unemployment Status for a Hypothetical Sample of Teenagers

	Race/Ethnicity				
Employment Status	Anglo	African-American	Hispanic	Other	Total
Employed	250	50	100	75	475
Unemployed	50	150	100	25	325
Totals	300	200	200	100	800

Note that the probability of being in a minority group (i.e. non-Anglo) is equal to

p(minority) = p (African-American or Hispanic or Other)
= .25 + .25 + .125 = .675 (6-7)

This illustrates the addition rule of probability, which states that the probability of being in any of a group of mutually exclusive (non-overlapping) events or situations is simply equal to the sum of the probability of being in each of them.

Similarly note that

p (unemployed) + p (employed) = .41 + .59 = 1.00 (6-8)

and that

p (Anglo) + p (African-American) + p (Hispanic)
+ p (Other) = .375 + .25 + .25 + .125 = 1.00 (6-9)

These two equations illustrate the exhaustive principle of probability, which simply says that the sum of the probabiity of all of the events (all of the employment statuses or all of the race/ethnic groups) must equal 1.0.

The figures given above in equations 6-2 through 6-6 are called marginal probabilities, corresponding to the marginal frequencies of a table. They are also called unconditional probabilities, indicating that they aren't conditional on, or don't depend on, any other event or attribute.

In contrast, the probabilities which correspond to the cells in the interior of Table 6-1 are called the joint probabilities, corresponding to the joint distribution of race/ethnicity and employment status and indicating the probability that an individual falls into any given combination of two categories or events. For instance, the probability that a teenager in the sample is an employed Anglo is

p (Anglo and Employed) = 250/800 = .3125 (6-10)

The probability that a teenager is an unemployed Anglo is

p (Anglo and Unemployed) = 50/800 = .0625 (6-11)

Note, following the addition rule, that equations 6-10 and 6-11 sum to equation 6-4, the probability of being Anglo. That is, the joint probabilities sum to the marginal probabilities.

The data in Table 6-1 suggest that race-ethnicity and employment status are quite likely related to each other. We might then be interested in trying to compute the probability that a teenager of any particular racial-ethnic group would be employed or unemployed. This is called a conditional probability, the probability of employment or unemployment conditional upon one's racial-ethnic heritage.

Consider, for instance, the situation of African-American teenagers. The probability that a teenager is both unemployed and African-American (the joint probability) is

$$p\ (\text{African-American and Unemployed}) = 150/800 = .19 \quad (6\text{-}12)$$

We know from equation 6-5 that the probability that any teenager in the sample is African-American is .25. The probability then that an African-American teenager is unemployed is

$$\frac{p\ (\text{African-American and Unemployed})}{p\ (\text{African-American})} = \frac{.19}{.25} = .76 \quad (6\text{-}13)$$

This is referred to as p (Unemployed|African-American) or the conditional probability of unemployment given that one is African-American. In general

$$p\ (B|A) = \frac{p\ (A \text{ and } B)}{p\ (A)} \qquad (6\text{-}14)$$

The conditional probability of an event B occurring given situation A is equal to the joint probability of A and B occurring divided by the probability of A occurring within the total group.

Note that the conditional probability is simply equal to the proportions one obtains when one computes proportions within the categories of the independent variable. That is, the conditional probabilities are simply ~~the~~ equal to the proportions (or .01 times the percentages) one obtains when one percentages a table within the categories of the independent variable. In Table 6-1, the ~~percentage~~ proportion of all African-American teenagers who are unemployed = 150/200 = .75, which is equal, given rounding error, to the figure obtained in 6-13 above.

This can be proved in general by noting that the joint probability of A and B equals

$$n_{ij} / n..$$

and the general probability of A occurring equals

$n_{.j} / n_{..}$.

Substituting these values into equation 6-14 we can obtain

$(n_{ij} / n_{..}) / (n_{.j} / n_{..}) = n_{ij} / n_{.j}$, which is simply the proportion of cases within a given category of A.

Randomness

Inferential statistics depend heavily upon the notion of randomness and the idea of random samples. Random samples are ones in which all elements of a given population have an equal chance of being selected. Non random samples, also called biased samples, are ones in which all elements of the population do not have an equal chance of being selected. Some stand a greater chance of being included than others. The inferential techniques we discuss here all assume that samples have been randomly selected. They also assume that samples have been independently selected, that is, that choosing one person does not automatically result in other people also being chosen. Each person or case has an equal chance of being selected.

Now, even though events or people we may study have been randomly selected or gathered, the statistics we derive from these samples turn out to be quite predictable. The text by Elifson, et al, gives an example of counting the number of times "heads" appear when a group of people flipping coins. If the people in this group continue to flip coins over a long period of time, the number of people getting "heads" each time will average out to simply half of the group. Moreover, if one plots the number of people getting heads in each of these tries over a long period of time, the frequency polygon will begin to look like a normal distribution. It will be unimodal, symmetrical, and bell-shaped, with a mean at the point indicating half of the people.

The exercise conducted in class illustrated this principle. In fact, the normal curve and its characteristics will be central to all of our later discussions of inferential statistics. Students should make sure that they understand all aspects of the discussion of the normal curve presented earlier before progressing further.

Basic Definitions

The following definitions are basic to the use of inferential statistics. Students should be familiar with all of them.

A population is the entire set or group of scores, people, animals, whatever the elements that are being studied.

A sample is a subset of the population, part of the population.

A random sample is a sample that is selected in such a way that each element of the population has an equal chance of being in the sample.

A representative sample may also be used in making inferences. This is a sample where the researcher knows how the sample was collected and in what way it is representative of the total population. Both random and representative samples, as noted earlier, are probability samples. In this course we will assume, when using inferential statistics, that all our probability samples are simple random samples. (The procedures involved in making inferences are slightly more complex when other types of probability samples are involved.)

A parameter is a specified value of the population, such as the mean or variance. Parameters are generally designated by Greek symbols.

A statistic is a specified value of the sample, such as the mean or variance. Statistics are usually designated by Roman letters.

The sampling error refers to the difference of the true population value and the sample value, the difference between the parameter and statistic. For any given sample taken from a population, a statistic (such as a mean) may differ from the corresponding parameter in the population. The difference between the statistic and parameter is the sampling error, the error introduced by looking at the sample instead of the total population.

The sampling distribution is a distribution of sample statistics obtained by drawing an infinite number of samples from a population. For example, given a large population one would draw one sample from the population, obtain the mean and standard deviation of that sample and plot it. The sample is then replaced and the procedure is repeated an infinite number of times. The eventual result is the sampling distribution.

Tables 6-2, 6-3, amd 6-4 illustrate the development of a sampling distribution. Table 6-2 gives data for a total population: the suicide rates for 220 SMSA's in 1970. Table 6-3 gives the results obtained when samples, each sized 30, were taken from this population and the average suicide rate was computed. Table 6-4 gives a tally of these sample means and Figure 6-1 displays this tally in a histogram. Only 100 samples were drawn in this example, but we could repeat the procedures an infinite number of times. (Data are taken from Muller, et al.)

Sampling theory tells us that when we have an infinite number of samples in our sampling distribution, the mean (average) of the sampling distribution of the means (the mean of the sample means) will equal the population mean. As the samples drawn get larger the distribution assumes the shape of the normal curve.

Table 6-4 and Figure 6-1 illustrate this result. It may be seen that the majority of sample means in the distribution cluster around the true population mean of 11.7. While the distribution of these actual sample means around the population mean of 11.7 is not exactly shaped like a normal distribution (this is called the empirical sampling distribution), if we drew an infinite number of samples, we would expect the sampling distribution around the population mean to be normally distributed. Because we could never draw an infinite number of samples this is referred to as a theoretical sampling distribution. <u>It is this theoretical sampling distribution that we use in making inferences from samples to populations.</u>

The discussion immediately above refers to the most typical value of the means (i.e., the central tendency of the sampling distribution). We are also concerned, however, with how far away from this central tendency most samples are. That is, we know that the values tend to cluster around the population mean, but how much do they vary? What is the sampling error, the difference of the sample mean and the population mean? Table 6-5 gives the distribution of sampling errors for the group of samples in Table 6-3. It is clear that the majority of errors are very small. More extreme errors are relatively less frequent.

It turns out that the standard deviation of the theoretical sampling distribution of means is equal to the standard deviation of the population divided by the square root of the sample size. This standard deviation of the sampling distribution is referred to as the standard error and has the formula:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad (6\text{-}15)$$

Table 6-2
Suicide Rates for 229 United States Standard Metropolitan Statistical Areas, 1970

Rate	Freq.	Rate	Freq.	Rate	Freq.	Rate	Freq
2.7	1	8.8	3	11.7	3	15.6	1
3.3	1	8.9	3	11.8	3	16.0	2
3.8	1	9.0	2	11.9	2	16.1	4
4.6	1	9.1	2	12.0	4	16.2	1
5.0	1	9.2	3	12.1	1	16,3	1
5.2	2	9.3	2	12.2	3	16.4	1
5.5	1	9.4	5	12.3	2	16.5	1
6.0	1	9.5	3	12.4	2	16.7	1
6.3	2	9.6	3	12.5	1	16.9	2
6.4	1	9.7	3	12.7	4	17.0	1
6.5	2	9.8	3	12.8	6	17.2	1
6.6	2	9.9	4	12.9	2	17.5	1
6.7	3	10.0	4	13.0	2	17.8	1
6.9	2	10.1	1	13.1	1	17.9	1
7.1	1	10.2	1	13.2	3	18.3	1
7.2	2	10.3	2	13.5	1	18.4	1
7.3	3	10.4	1	13.6	2	18.6	1
7.4	4	10.5	3	13.7	1	18.7	1
7.5	2	10.6	2	14.0	4	19.0	1
7.6	2	10.7	4	14.1	2	19.4	1
7.7	3	10.8	1	14.3	2	20.0	1
7.8	1	10.9	2	14.5	1	20.1	1
7.9	2	11.0	2	14.6	1	20.6	1
8.0	1	11.1	2	14.7	1	20.9	1
8.2	1	11.2	3	14.9	1	21.0	1
8.4	2	11.3	3	15.1	2	21.8	1
8.5	3	11.4	2	15.2	2	22.0	1
8.6	2	11.5	3	15.4	1	22.1	1
8.7	3	11.6	6	15.5	1	22.5	2
						24.8	1
						24.9	1
						25.0	1

Table 6-3
100 Sample Means, n=30, Taken from Distribution in Table 6-2

Sample Means	Frequency
10.3	1
10.4	4
10.5	1
10.6	7
10.8	9
10.9	3
11.0	3
11.1	1
11.2	3
11.3	7
11.4	1
11.5	6
11.6	1
11.7	6
11.8	9
11.9	9
12.1	3
12.2	6
12.3	2
12.4	6
12.5	3
12.9	2
13.0	5
13.2	2
Total	100

Table 6-4
Frequency Distribution of Sample Means -- The Empirical Sampling Distribution for Data in Table 6-3

Range of Means	Frequency
10.0 - 10.9	25
11.0 - 11.9	46
12.0 - 12.9	22
13.0 - 13.9	7

Figure 6-1
Histogram of the 100 Sample Means, n=30, in Tables 6-3 and 6-4

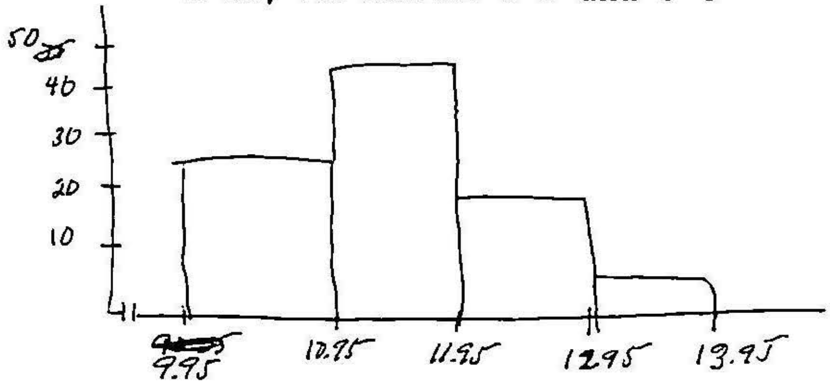


Table 6-5
Sampling Errors, 100 Samples given in Tables 6-3 and 6-4

Error	Frequency	Error	Frequency
-1.4	1	.1	9
-1.3	4	.2	9
-1.2	1	.4	3
-1.1	7	.5	6
-.9	9	.6	2
-.8	3	.7	6
-.7	3	.8	3
-.6	1	1.2	2
-.5	3	1.3	5
-.4	7	1.5	2
-.3	1		
-.2	6		
-.1	1		
0	6		

Note what each part of this formula implies. First, as the population becomes more variable, the samples are less likely to have means like those of the population. Thus samples from more heterogeneous populations will have larger standard errors. Second, as the sample sizes become larger, the standard error decreases and the sample means are likely to be closer to the population mean. This means that if you were to take two samples of different sizes from the same population, the larger sample would have a smaller standard error.

Because one usually does not know the standard deviation of the population we must arrive at some estimate of this standard error. We use the standard deviation of the sample for this estimate, but make sure that the standard deviation is defined as

$$S = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} \qquad (6\text{-}16)$$

Various computer programs routinely compute the standard deviation with this formula, but some statistics books refer to it not as s, but as $\hat{\sigma}$, to denote that it is the best estimate of the population standard deviation. (As explained earlier, the denominator of n-1, rather than n, is used in equation 6-16 because samples tend to vary less than populations and this corrects for this smaller variance.) Using this sample estimate of the population standard deviation, the formula for the standard error becomes

$$S_{\bar{X}} = \frac{S}{\sqrt{n}} = \sqrt{\frac{S^2}{n}} \qquad (6\text{-}17)$$

where

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

It should be stressed that the sampling distribution of the mean is normally distributed even when the frequency distribution for the population is not. <u>No matter what the shape of the population distribution, the sampling distribution of the means will assume the shape of the normal distribution when samples are greater than 100 or so</u>. (We'll discuss the case of smaller samples later. Essentially they have an "almost normal" distribution, called the t distribution.) It is crucial that students understand the difference between a frequency distribution, such as those discussed in the second section, and a sampling distribution, the hypothetical distribution of sample statistics.

The sampling distribution and the standard error are the basis of all inferential statistics. Above, we mainly referred to the sampling distribution of means. However, sampling distributions can be constructed (and have been) for many other statistics. The basic procedure used with all inferential statistics is the same logically, and so in the discussion below we will focus on inferences regarding means.

The important things to remember in the discussion below are the nature of the normal distribution; the fact that with large samples the sampling distribution of the means is normal (with smaller samples it is the t-distribution whose nature is also known and which we will discuss below); and that when we know the mean and standard deviation of the sample we can estimate what the sampling distribution looks like for that population (assuming that the sample is representative of the population). This basic information is used in computing all inferences regarding means.

Confidence Intervals

Confidence intervals are a way of estimating population parameters given knowledge of the related sample statistics. These are also referred to as point estimations. This is done by using knowledge of the sampling distribution. Thus, it is essential that random or representative samples be used. Basically, the statistics from the sample are used as estimates of the population parameters. From these estimates the sampling distribution is reconstructed. Then, using the table giving the area under the normal curve, assuming we have a large sample, the probability of the parameter being within certain ranges may be computed. An example will illustrate this.

Given a random sample of 169 cases from a very large population.

$$\bar{X} = 50; \quad s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} = 26.$$

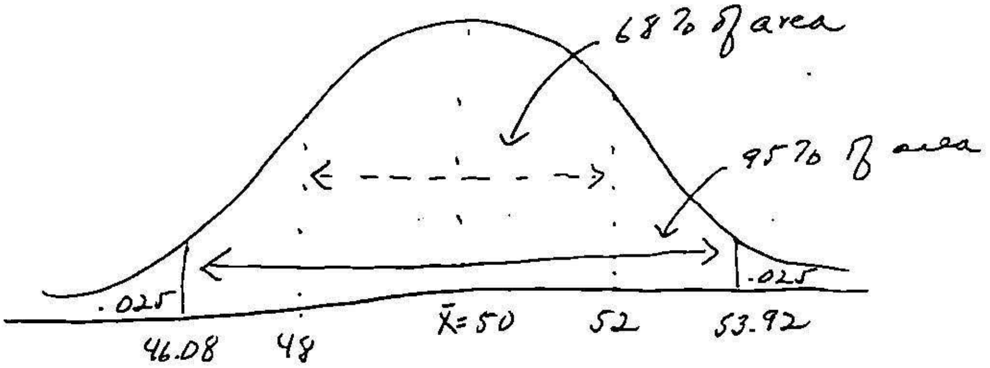
This information may be used to estimate the form of the sampling distribution. As explained above, X is our best estimate of μ, the population mean, $\bar{X}$ = 50.

$s_{\bar{X}} = \frac{s}{\sqrt{n}}$ is our best estimate of $\sigma_{\bar{X}}$, the standard error. $s_{\bar{X}} = \frac{26}{\sqrt{169}} = \frac{26}{13} = 2.0$

Thus, we may estimate the sampling distribution to be normally distributed with a mean of 50 and a standard error of 2,based on our knowledge of the random sample from this population. This sampling distribution is pictured in Figure 6-2. Note that this is a theoretical distribution of means of samples that could be drawn from the population. Because the sample we do have has been randomly drawn, we may assume that it is representative of the population and we use these characteristics to estimate the nature of the sampling distribution.

Figure 6-2

Theoretical sampling distribution where $\bar{X}=50$, $s_{\bar{X}}=2$



We can assume that the sampling distribution is normally distributed because the sample size is relatively large. Using the knowledge of the characteristics of the normal curve we know that between one standard error below the mean and one standard error above the mean there is .6826 of the total area under the curve. In this case the scores in the distribution are sample means and we can say that .6826 of all the sample means in this sampling distribution are between 48 and 52. That is, they are in the area plus or minus one standard error from the mean. If we take these sample means as estimates of the population mean we can say that .6826 of the estimates of the population mean are between 48 and 52. Another, easier, way of saying that is that the probability that the true population mean is between 48 and 52 is equal to .6826.

This can be written symbolically as

$$P\ [48 < \mu < 52] = .6826 \qquad (6\text{-}18)$$

This may be referred to as a 68% confidence interval around the mean. This means that we can be 68% confident that the true population mean lies between 48 and 52.

Note that we switched from talking about the <u>proportion of estimates of the mean of the population</u> that were within a given range to discussing the <u>probability that the population mean was within a given range</u>. This is the essence of statistical inference. We are concerned with the chances of being correct (the probability of being correct) in estimating the value of a population parameter. We use the sampling distribution estimated from the sample values to compute these chances or probabilities.

Confidence intervals or bands equal to 95% or 99% are commonly used. With intervals of this width we are finding the range of values in which 95% (or 99%) of the estimates of the population value fall. For a 95% confidence interval only .025% of the area under the curve would not be included within the interval on each side of the mean. Referring again to the table of the normal curve we can see that .025 of the area under the curve is remaining (.475 on one side of $\overline{X}$ is included) when we are 1.96 standard errors from the mean. Thus, to enclose the area encompassing 95% of the possible means in this theoretical distribution we must go both 1.96 standard errors above the mean and 1.96 standard errors below the mean.

In the present example the estimated mean is 50 and the estimate of the standard error is 2. 1.96 standard errors is equal to 3.92. Thus, we may conclude that 95% of the means in the estimated sampling distribution are included between (50 - 3.92) and (50 + 3.92). This may be written symbolically as

$$P[46.08 < \mu < 53.92] = .95 \qquad (6\text{-}19)$$

This means that we can be 95% confident that the true population mean lies between 46.08 and 53.92 or that the probability that the population mean lies between 46.08 and 53.92 is .95.

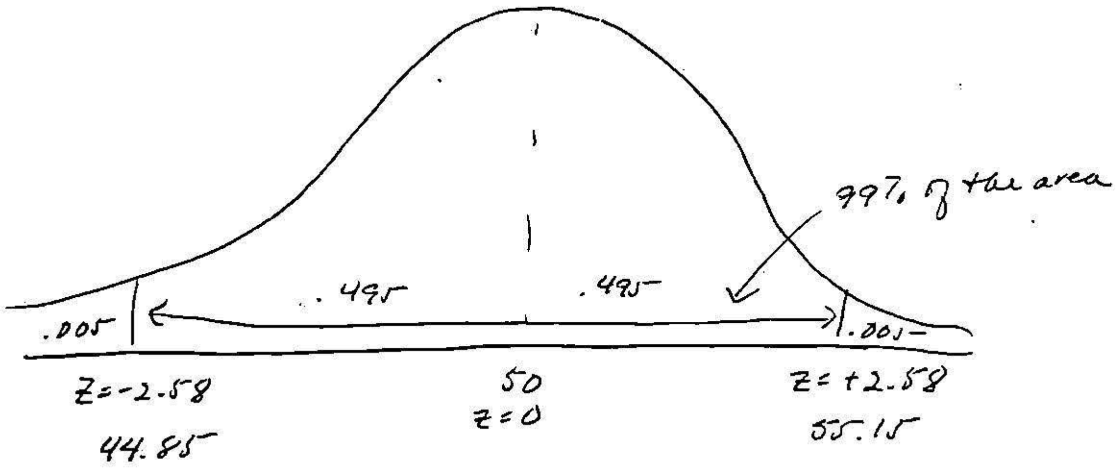
For a ninety-nine percent confidence interval we would need to enclose all but .005 of the area on each side of the mean. This corresponds to an area of .495 between the mean and the given point, which corresponds to a z-score of about 2.58. The computations below and the figures show how the 99% confidence interval would be computed.

$$P[\bar{X} - (2.58)(s_{\bar{x}}) < \mu < \bar{X} + (2.58)(s_{\bar{x}})] = .99$$

$$P[50 - \underbrace{(2.58)(2)}_{5.15} < \mu < 50 + \underbrace{(2.58)(2)}_{5.15}] = .99$$

$$P[44.85 < \mu < 55.15] = .99$$

Figure 6-3



These results indicate that 99% of the means in the estimated sampling distribution fall between 44.85 and 55.15. There is a 99% probability that the population mean falls between 44.85 and 55.15. We can be 99% confident that the population mean lies between 44.85 and 55.15. A general formula for computing confidence intervals is often used. For the 95% confidence interval around the mean, when samples are large, we may use

$$p[\bar{X} - 1.96\, s_{\bar{X}} < \mu < \bar{X} + 1.96\, s_{\bar{X}}] = .95 \qquad (6\text{-}20)$$

and, for the 99% confidence interval, we may use

$$p[\bar{X} - 2.58\, s_{\bar{X}} < \mu < \bar{X} + 2.58\, s_{\bar{X}}] = .99 \qquad (6\text{-}21)$$

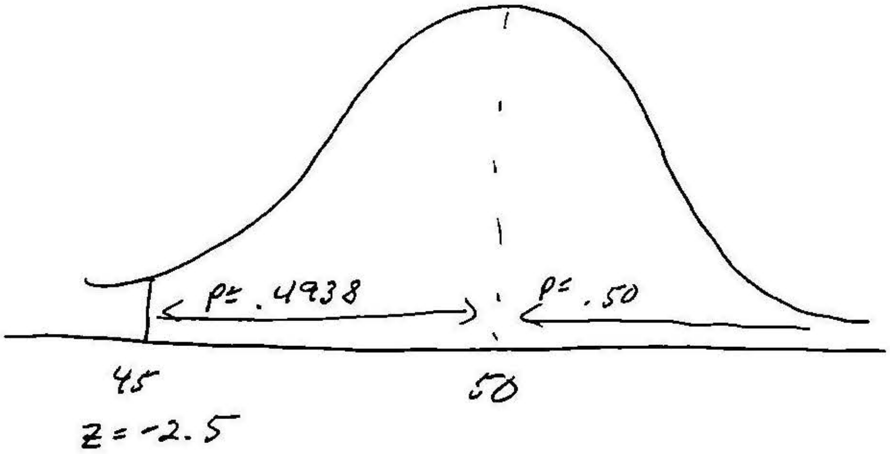
where $\bar{X}$ is the sample mean and $s_{\bar{X}}$ is the estimated standard error.

The logic underlying confidence intervals may also be used in computing the probability that the population parameter is greater than or less than a certain score. For instance, in the example above, we may compute the probability that μ, the population mean, is greater than 45. To do this we must first determine how far this X = 45 is from the mean of the theoretical sampling distribution. We may do this using standard scores.

$$z = \frac{45-50}{2} = \frac{X-\bar{X}}{s_{\bar{X}}} = \frac{-5}{2} = -2.5$$

That is, a score of 45 is 2.5 standard errors below the mean in the sampling distribution.

Figure 6-4

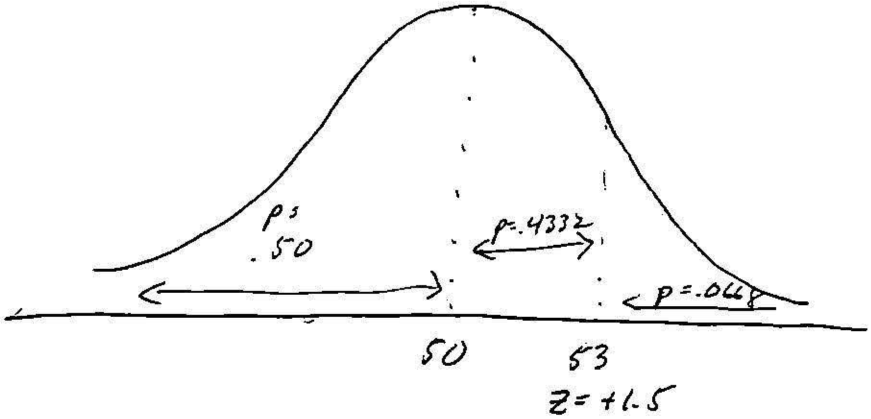


Using the table of areas under the normal curve we can see that the proportion of area under the curve from the mean to X = 45 is .4938. Thus, P [45 < μ < 50] = .4398. We know that P [μ ≥ .50] = .5000 as 50 is the best estimate of the mean of the sampling distribution. Thus, P [μ < . 45] = .4983 + .5000 = .9938.

Similarly, to compute the probability that μ < 53 we must determine how far away 53 is from the estimated mean of the sampling distribution, 50. z = $(X - \bar{X})/s_{\bar{x}}$ = (53-50)/ 2 = 1.5. This indicates that 53 is 1.5 standard errors above the estimated sampling distribution.

mean of

Figure 6-5



Using the table of area under the normal curve we can find that

P [μ < 50] = .5000

P [50 < μ< 53] = .4332

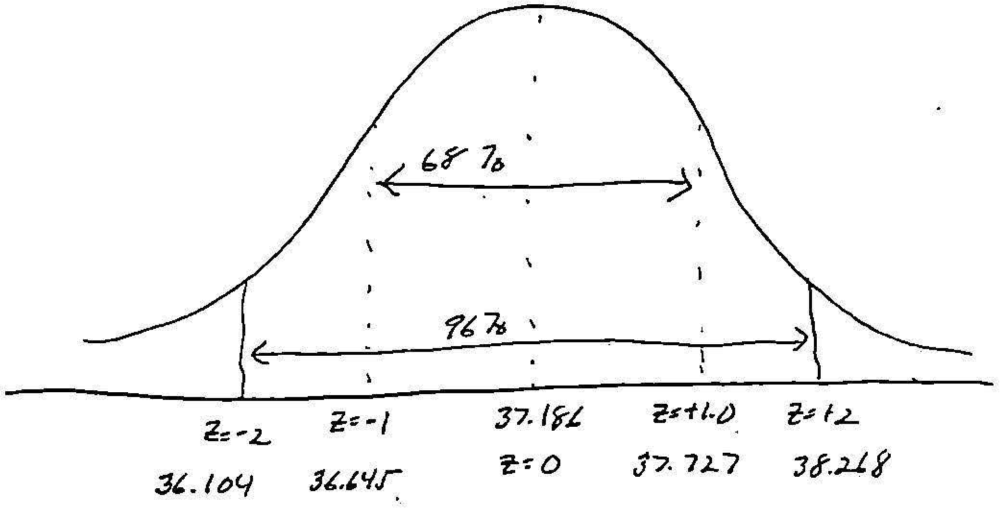
and thus P [μ< 53] = .5000 + .4332 = .9332.

There is a 93% probability that the population mean is less than 53. Similarly, P [μ > 53] = .5000 - .4332 = .0668. Students should work through several more examples of varying types to ensure that they totally understand the logic of confidence intervals.

Note that all of these computations have been based on the theoretical sampling distribution of the mean. If the sample size were different or if the observed mean or standard deviation of the sample were different, the results would have been altered.

Computer output commonly gives the standard error for a distribution. Consider the distribution of ages of the bank employees shown in Table 2-1. Assume the sample has been randomly selected from some larger population of bank employees. The sample mean is given as 37.186, and the standard error is 0.541. The sampling distribution of the means may be estimated as shown in Figure 6-6 below.

Figure 6-6



We may conclude, based on our sample information, that about 96% of the sample means lie between 36.1 and 38.3 years; 68% of the sample means lie between 36.6 and 37.7 years.

Further calculations indicate that

$$\begin{aligned} &P[37.19 - (.54)(1.96) < \mu < 37.19 + (.54)(1.96)] = .95 \\ &= P[37.19 - 1.06 < \mu < 37.19 + 1.06] = .95 \qquad (6\text{-}22) \\ &= P[36.13 < \mu < 38.25] = .95 \end{aligned}$$

$$\begin{aligned} &P[37.19 - (.54)(2.575) < \mu < 37.19 + (.54)(2.575)] = .99 \\ &= P[37.19 - 1.39 < \mu < 37.19 + 1.39] = .99 \\ &= P[35.80 < \mu < 38.58] = .99 \qquad (6\text{-}23) \end{aligned}$$

We can be 95% confident that the average age of bank employees in the total population lies between 36.13 and 38.25 years. We can be 99% confident that the average age of the bank employees in the population lies between 35.80 and 38.58 years.

To find the probability that the average age is less than 39 years we need first to compute the z-score that corresponds to 39: $z = (X - \overline{X})/S_{\overline{X}} = (39-37.19)/.54 = 3.35$. Consulting Table 2-1 and interpolating we find that approximately .49959 of the area under the curve lies between the mean and 3.35 standard errors above the mean. Since .5000 of the area lies below the mean we can reach the following conclusions

$$P[37.19 < \mu < 39] = .49959$$

$$P[\mu < 37.19] = .5000 \qquad (6\text{-}24)$$

$$P[\mu < 39] = .49959 + .5000 = .99959$$

We can be 99.959% confident that the average age of the bank employees in the population is less than 39 years.

Hypothesis Testing

Hypothesis testing, the other major inferential technique, is somewhat more common than confidence intervals. Here, instead of using sample statistics to make inferences about the nature of a parameter, we start with an idea about the population parameters. We then draw out the implications of this idea and test the truth of the implications with the data from the sample.

The null hypothesis is the hypothesis to be tested. It is symbolized as H_o.

The alternative or substantive or research hypothesis is the alternative to the null hypothesis. For example, if the null hypothesis, H_o is that $\mu = 0$; H_1 (the alternative hypothesis) may be $\mu \neq 0$ or $\mu > 0$ or $\mu < 0$.

The null and alternative hypotheses are phrased so that we can reject the null hypothesis with certain probabilities of being wrong and that by rejecting the null hypothesis we can put corresponding confidence in the truth of the alternative hypothesis. The null hypothesis is always phrased in the format of the population parameter equaling some constant (either zero or some other number). The alternative hypothesis is phrased so that the population parameter is either unequal to that constant or greater or less than that constant.

Note that we can never prove the truth of the null or alternative hypotheses. We fail to reject or we reject the null hypothesis with a certain degree of confidence that our decision is correct. We do this by assuming that the null hypothesis is true and then drawing implications from this assumption. Using the sampling distribution we determine the probability of certain sample values appearing. This is the logic of falsification that is basic to work in the social sciences.

The level of significance refers to the decision of how rare a sample outcome must be if it is to cast doubt on the null hypothesis. Usually researchers use levels such as

.05, .01, or .001. However, these are arbitrary levels,and I recommend always noting the actual probability that a given result would occur. This is especially important when we consider what would happen if we consistently received results that were in the same direction, but only marginally significant. For instance, suppose we found that we could reject the null hypothesis in favor of the alternative with a .20 probability of being wrong. Normally, we would fail to reject the null hypothesis. But, suppose we repeated the study and found identical results with a second sample. The chance of finding this same result two times in a row is (.20)(.20) = .04. This is a result that would be acceptable at standard levels of significance, but if we simply reported n.s. (not significant) in our write-up, no one would know how important the results really were.

The zone of rejection is the sample values which lie in the area where their probability of occurrence equals or is lower than the level of significance. Another way of seeing this is as the sample values whose occurrence is so rare that they would occur (given the truth of the null hypothesis) only as frequently as the level of significance.

An example may help to make this clearer. Suppose we had the following null and alternative hypotheses

H_o: $\mu = 100$
H_1: $\mu \neq 100$

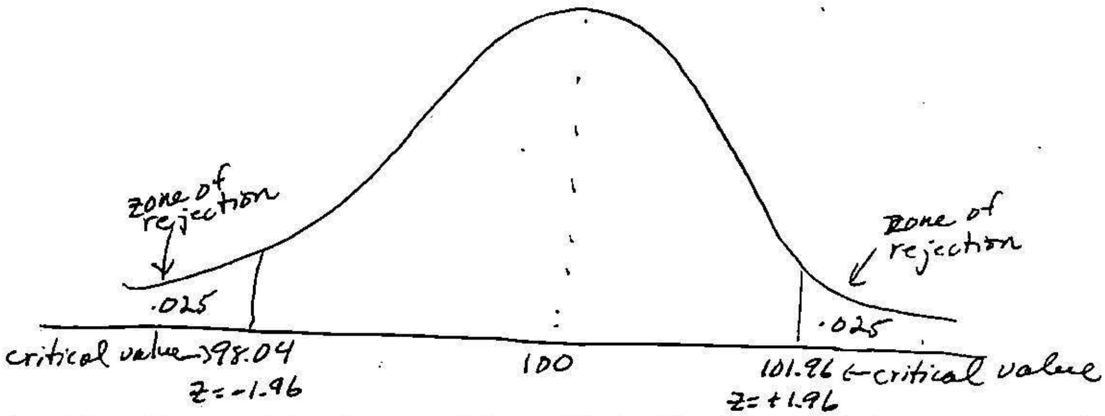
Suppose we draw a random sample from the population involved. In this sample $\bar{X} = 95$, $s = 13$, $n = 169$

Now we shall suppose that H_o is actually true, that the population mean really equals 100. Then we shall use 100 as the mean of the sampling distribution of the means. Given that the sample is a random one of the population, we may use S_x as the estimate of the standard error.

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{13}{\sqrt{169}} = \frac{13}{13} = 1.0$$

Because the n is large, the sampling distribution is normally distributed. The theoretical sampling distribution that would occur given that H_o is true and with the standard error estimated by the sample value of the standard deviation, is drawn in Figure 6-7 below.

Figure 6-7



This is the theoretical sampling distribution with a mean of 100, standard error of 1.0. ~~and~~ It is normally distributed. This distribution would be the true estimated sampling distribution for the population if H_0 were true.

Suppose we choose a level of significance of .05. That is, we decide that to reject the null hypothesis we must have a sample value that would occur only 5 times out of one hundred.

Our alternative hypothesis is that $\mu \neq 100$. We have not hypothesized that μ is less than or greater than 100. Thus, our zone of rejection may be on either side of 100. Because our level of significance is equal to .05 the combined probability of scores in the zone of rejection must equal .05. Thus, the probability of scores in the zone of rejection on both sides of the mean must equal .025 + .025 = .050.

Referring again to the table of area under the normal curve we can find that the score or z value marking off this zone of rejection will be 1.96 standard errors away from the mean on either side. Thus, if a sample value falls either 1.96 standard errors above the mean or 1.96 standard errors below the mean, given that H_0 is true, it will fall in the zone of rejection. That is, if the sample value falls into the zone of rejection the probability of that actually occurring if the null hypothesis were true is less than the level of significance, less than .05.

In this example, the standard error is equal to 1.0. Thus, the zone of rejection equals all values below (100) - (1.96) (1.0) = 98.04 and all values above 100 + (1.96) (1.0) = 101.96. All scores less than 98.04 or greater than 101.96 fall into the zone of rejection.

We return now to the sample chosen. In this sample the mean was 95. This value clearly falls into the zone of rejection. The probability of this value occurring when H_o is true is less than .05. In other words, we may reject the null hypothesis that the population mean does not equal 100 with less than 5 chances out of one hundred of being wrong.

Note that quite likely the probability of being able to reject H_o in favor of H_1 is much lower than .05. In actual practice it is much more useful to give the actual level of the probability of occurrence. As noted above, this is most useful for replication. The computer generally prints the exact probability. We can easily calculate the exact probability of an event occurring simply by finding the z-value that corresponds to the actual sample value on the sampling distribution that assumes that H_o is true. In this case

$$z = \frac{(\mu - \bar{x})}{s_{\bar{x}}} = \frac{(95-100)}{1} = -5.0$$

Locating this z-value on the table of the normal curve we see that the actual probability of this value occurring is < 2 (.0001) = < .0002. We had to multiply the proportion times 2 because our hypothesis did not specify a zone of rejection on just one side of the mean, but on both sides.

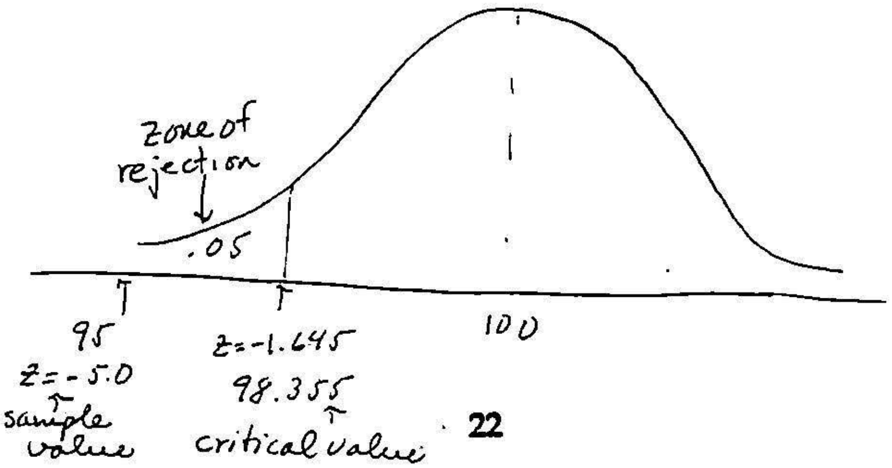
Sometimes a researcher may have reason to suspect that the true population mean fell above or below a certain level. In this case the researcher would use what is called a directional alternative hypothesis instead of the non-directional hypothesis specified above. For instance, suppose in the example above the hypotheses had been

H_o: $\mu = 100$ or $\mu \geqq 100$

H_1: $\mu < 100$

Again assume that a random sample was drawn, with $\bar{X} = 95$, s = 13, n = 169. The theoretical sampling distribution assuming that H_o is true is given below in Figure 6-8.

Figure 6-8



The zone of rejection in this case would fall only below the mean. That is, we are only concerned with samples in this sampling distribution with means less than 100. With a .05 level of probability, this means that all means less than 1.645 standard errors below the hypothesized mean would fall into the zone of rejection. In this distribution this corresponds to all sample scores less than or equal to (100) - (1.645) (1.0) = 98.355. Thus, if a sample mean were be 98.355, or less, we could reject H_0: $\mu \geq$ 100 in favor of H_1: μ < 100 at the .05 level of significance. Note, however, that the exact probability of getting the sample value of 95 when the null hypothesis is true and the alternative hypothesis is true and the alternative hypothesis is directional is <.0001.

This basic logic of testing hypotheses can be extended in many ways. Always the format of the null and alternative or research hypothesis is used. Also, the sampling distribution, assuming that H_0 is true is developed and the sample values are compared against the "critical values" on that sampling distribution. The <u>critical value</u> is the value on the sampling distribution that denotes the start of the zone of rejection. It is important to note that the nature of the alternative hypothesis depends on the <u>theory</u>, what you as a researcher are interested in. For instance, someone interested in the IQ scores of college students would likely have as the research hypothesis that $\mu > 100$. Note that the null hypothesis includes all values of 100 and below.

The theory of inferential statistics has been developed with the assumption that the populations involved are infinitely large. Sampling is usually done with replacement (that is once a sample has been drawn it is replaced). In real sociological research we will sometimes have samples that are relatively large in relation to the population. As your sample approaches the size of the population your sampling error and also your standard error tend to go down. If you are involved in having to make inferences in cases where the sample approaches the population size you should consult a textbook for the rather simple calculations involved in correcting the size of the standard error. In essence, these calculations make it even easier to reject the null hypothesis.

In general, all tests of hypotheses involve the basic steps we have followed here. First one determines a null hypothesis, then one determines an alternative or research hypothesis. Third, one sketches the sampling distribution one would have if the null hypothesis were true. Fourth, one determines the probability level at which one wishes to reject the null hypothesis and the associated critical value and zone of rejection. Fifth, one computes the test statistic, here the z-value, that corresponds to the sample value. Sixth, one compares the sample test statistic with the critical value and decides whether one should reject or fail to reject the null hypothesis. Finally, one computes the actual probability of being wrong if one were to reject the null hypothesis.

Inferences About Means with Small Samples

In the discussion above it has been stressed that the sampling distribution of the means is normally distributed when samples are large, generally over 100 or so. What about smaller samples?

It is possible to make inferences about means when you have samples smaller than 100 using the same procedure as that outlined above. The only difference is in the shape of the sampling distribution. It assumes the shape of the t-distribution. The t-distribution is similar to the normal distribution in that it is symmetrical, unimodal, and infinite. It, however, varies depending on what is called "degrees of freedom." These correspond to the size of the samples being studied. With very small samples the t-distribution, is much broader and shorter than the normal distribution, but as the degrees of freedom (or sample size) become larger (n's over about 150) the shape of the t-distribution becomes more like the normal distribution until with large samples they are identical.

When making inferences about means with small samples you calculate the degrees of freedom by subtracting one from the sample size (n-1). You can then look up the critical values for the sampling distribution on the table summarizing these for the t-distribution and use these critical values in your analysis.

There is not just one t-distribution, but a whole family of distributions. The t-distribution is essentially flatter and wider than the normal distribution, and as the n gets larger it approaches the normal shape more and more. Because there are so many different t-distributions, the table describing the t-distribution does not give all the values. Instead, the table gives the critical values (the values of t found at the edge of the zone of rejection) for a number of levels of significance. This is given for both the case when the alternative hypothesis is two-tailed (no direction given) and when it is one-tailed (directional). These values then can also be used for confidence intervals.

The formula used to calculate the t-value for any value along the distribution is directly analogous to the computation of the z-score.

$$t = \frac{X - \bar{X}}{s_{\bar{X}}}$$

What this formula does is to locate the sample value along the sampling distribution. It tells us how far the sample value is from the estimated or hypothesized mean of the sampling distribution, which is shaped like a t-distribution.

A simple example can illustrate this. Say we had the following hypotheses:

H_0: $\mu = 40$; H_1: $\mu < 40$; $s = \sqrt{\Sigma(X-X)^2/n-1} = 5$; $n = 25$; $X = 38$;

$s_{\bar{X}} = 5/\ 25 = 5/5 = 1.0$.
Say we had chosen a significance level of .05.

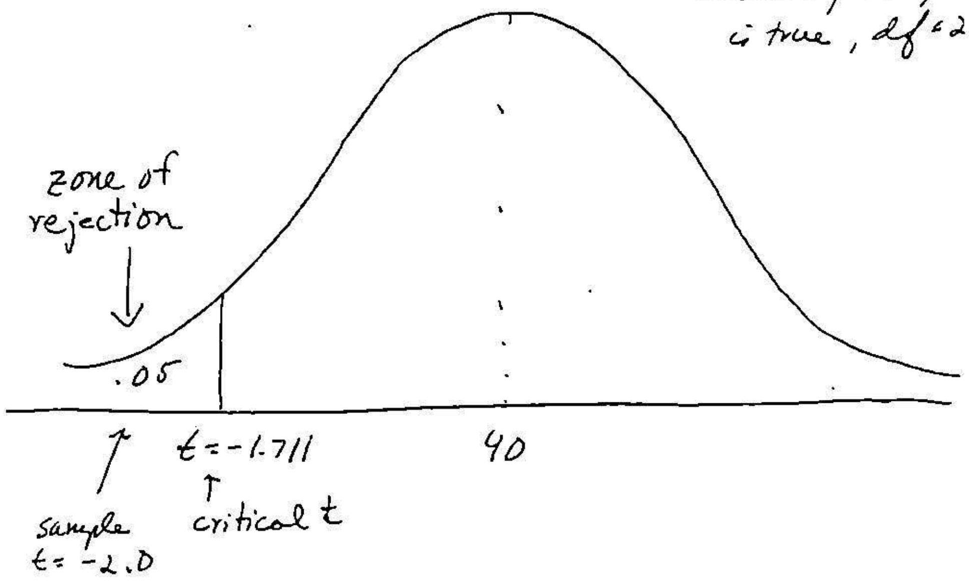
We turn now to the t-distribution. To read this table you need to understand the nature of degrees of freedom. Degrees of freedom are related to the size of the sample. In one sample tests, such as this, the degrees of freedom simply equal n-1. Degrees of freedom come from the number of free guesses one has in determining the value being examined. In this case that value is the mean. In choosing sample values for a particular mean we can choose values randomly for all the cases except one. To make the mean correct, or equal to a particular value, we must set one score equal to some specific number. Thus, we lose one degree of freedom. Here, our df = n-1 = 25-1 = 24.

Now, reading the table for a one-tailed test (from our directional hypothesis), for 24 degrees of freedom, for a .05 level of significance, the critical value is 1.711 for us to reject the null hypothesis in favor of the alternative that μ is less than 40. Because the alternative hypothesis is worded so that the expected population mean is less than that in the null hypothesis, our t-value will also need to be negative.

To compute t we simply substitute in the formula that is so similar to the formula for z-scores. $t = \bar{X} - \mu / s_{\bar{x}} = \frac{38-40}{1} = -2.0$ In other words, along the t-distribution, as shown below, our sample value falls at two standard errors below the hypothesized mean in the null hypothesis. This is indeed in the zone of rejection and we can reject the null hypothesis in favor of the alternative at the .05 level of significance. By examining the t-table we see that this t-value is not large enough to reject the null hypothesis at the .025 level of significance. Therefore, the probability of being wrong in rejecting H_o is less than .05, but greater than .025.

Figure 6-9

The sampling distribution (t) assuming H_0 ($\mu = 40$) is true, $df = 24$, $s_{\bar{X}} = 1.0$


zone of rejection
t = -1.711
critical t
sample t = -2.0