Packet 56

SOC 412

SOCIOLOGICAL RESEARCH METHODS

Professor Stockard

University of Oregon

Winter Term 1992

Copies:	$5.40
Binding	$1.75
Royalties	$0.00
Permission Handling Charges	$0.00
Total cost of packet:	$7.15

TABLE OF CONTENTS

Jean Stockard - Soc 412
Packet # 56

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.

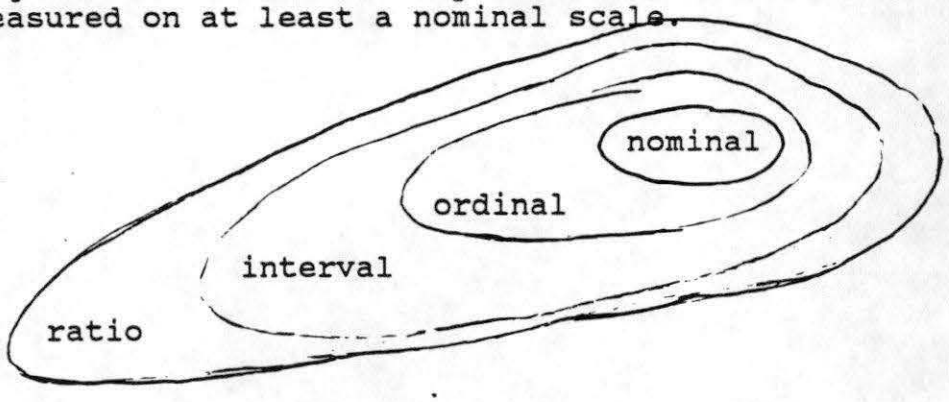


Figure 1-1: Representation of the relative restrictiveness of the four levels of measurement

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

Code	Freq	Adj %	Cum %	Code	Freq	Adj %	Cum %	Code	Freq	Adj %	Cum %
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 to nearest	Rounded to next lower	Rounded to next higher	True-limits of Interval nearest	True-limits of Interval next lower	True-limits of Interval 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	5.0	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	6.020	6.019	6.020	6.0195-6.0205	6.019-6.020	6.019-6.020
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	
Grade	Frequency	Cum. Freq. (less than)	Frequency	Cum. Freq. (less than)
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

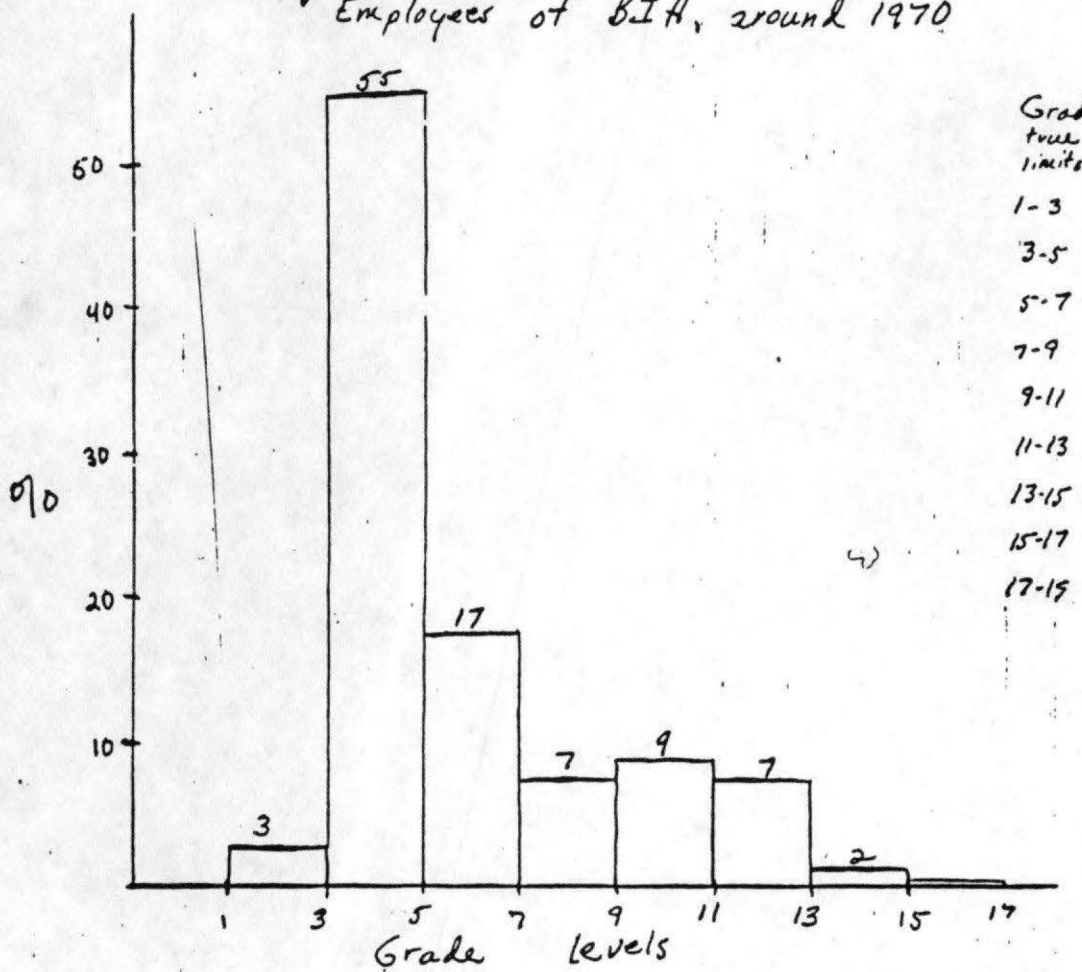
There are three basic graphs that are commonly used to represent quantitative data: histograms, frequency polygons, and ogives (or cumulative frequency graphs). Each of these has a common form in that along the horizontal axis the intervals for the distribution are graphed. These would be the same intervals 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

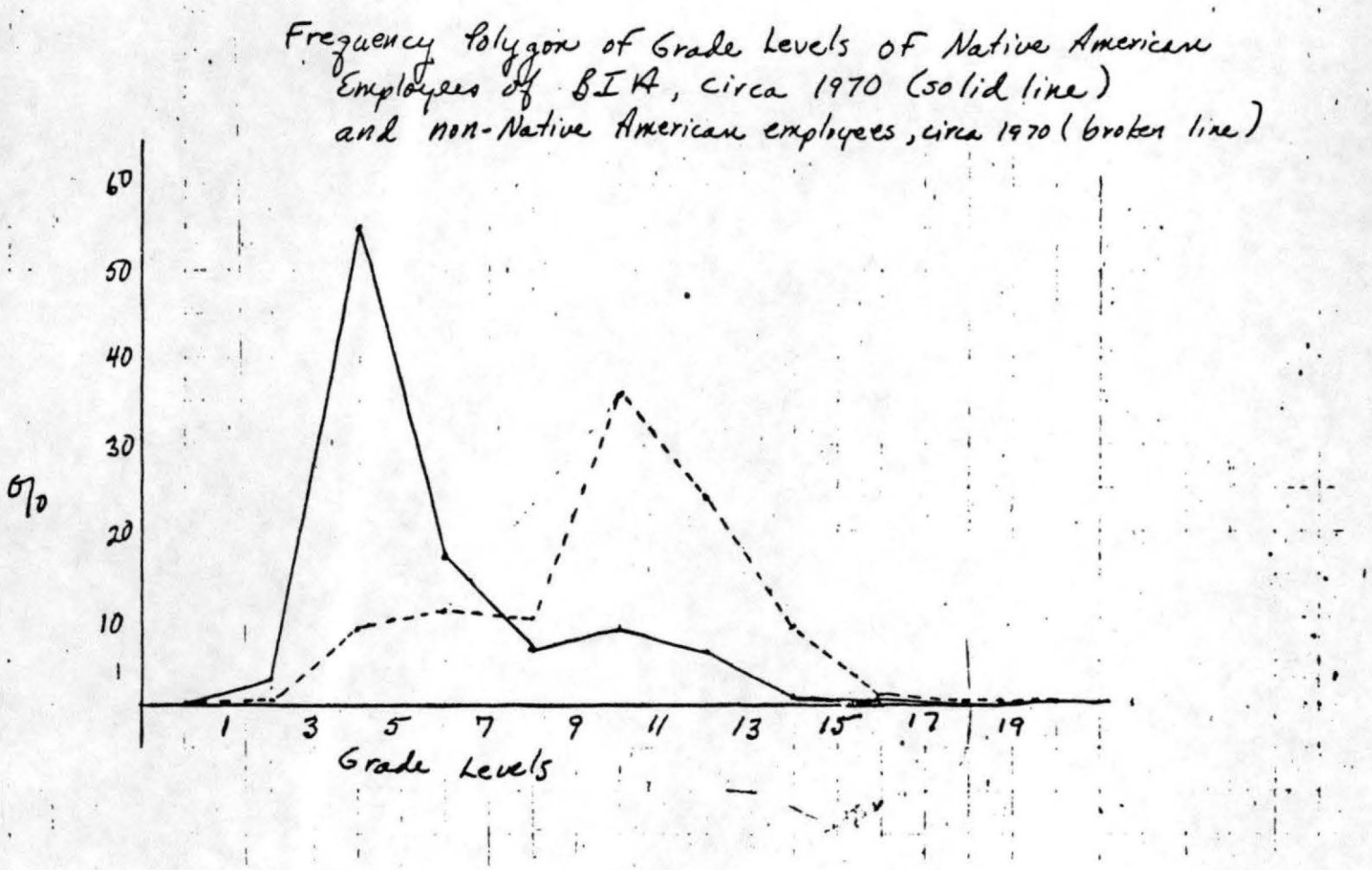
Figure 2-2

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



Grade levels true limits	Grade levels rounded limits	NA	%'s 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	.88
17-19	17-18	0	0.03

Figure 2-3


Frequency Polygon of Grade Levels of Native American Employees of BIA, circa 1970 (solid line) and non-Native American employees, circa 1970 (broken line)
60
50
40
30
%
20
10
1
1
3
5
7
9
11
13
15
17
19
Grade Levels

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 ($\leq$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

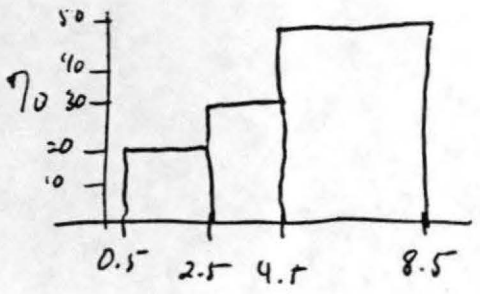
Figure 2-4

Example of Adjusting uneven Intervals in Constructing a Histogram

a)

True Interval Limits	f_0
0.5-2.5	20
2.5-4.5	30
4.5-8.5	50
	100 %

Incorrect Histogram



b)

Adjusted True Interval Limits	f_0
0.5-2.5	20
2.5-4.5	30
4.5-6.5	25
6.5-8.5	25

Correct Histogram

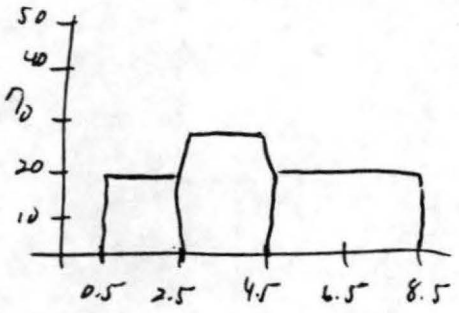
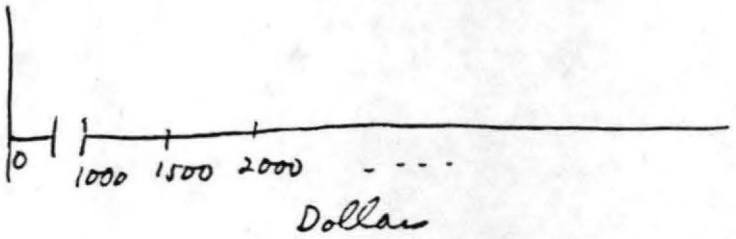


Figure 2-5

Example of Baseline for a Graph w/ Quantitative Data



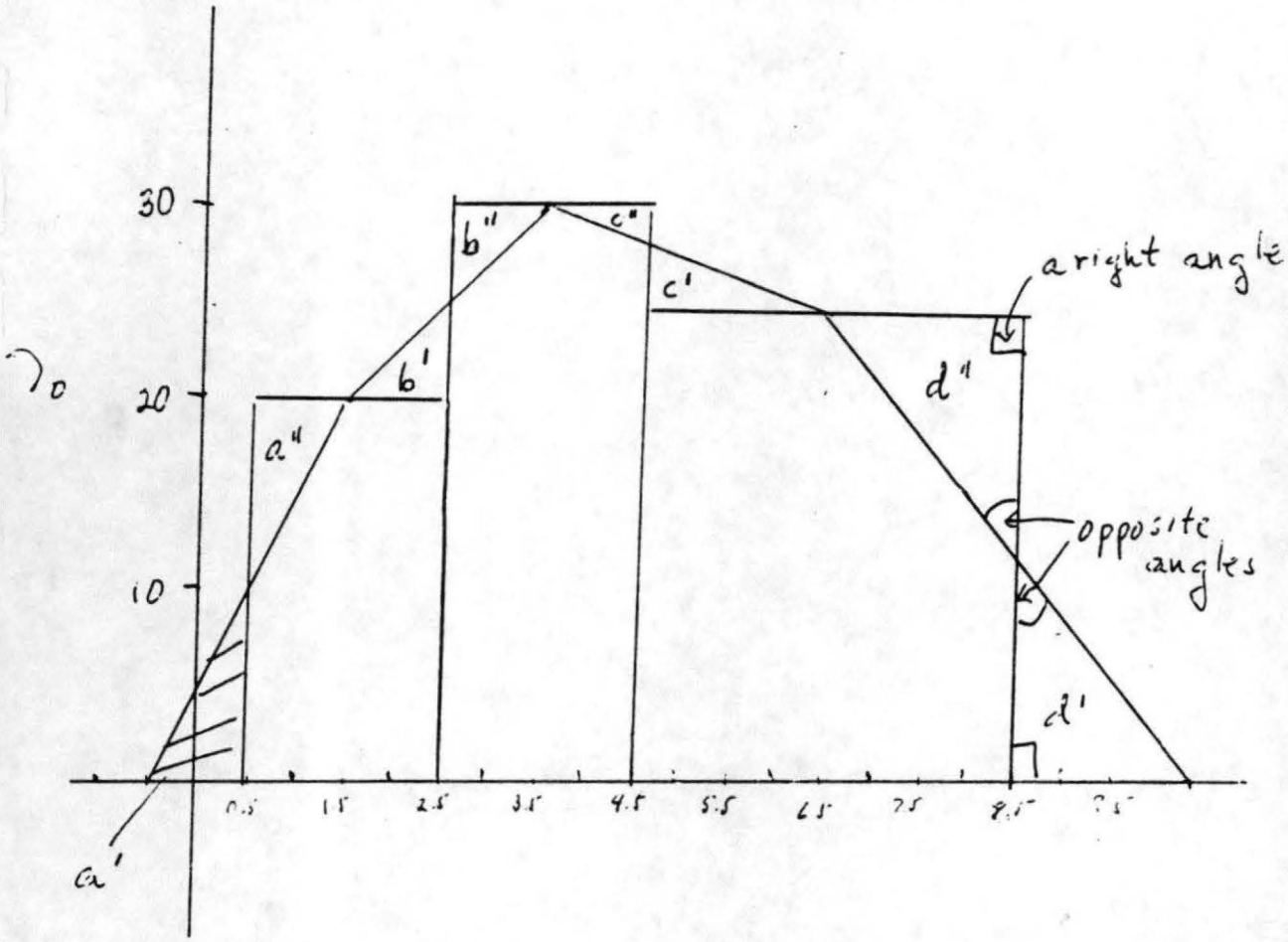
represented within that interval. The data for a given interval within a frequency polygon are actually spread across the area allocated to three intervals.

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.

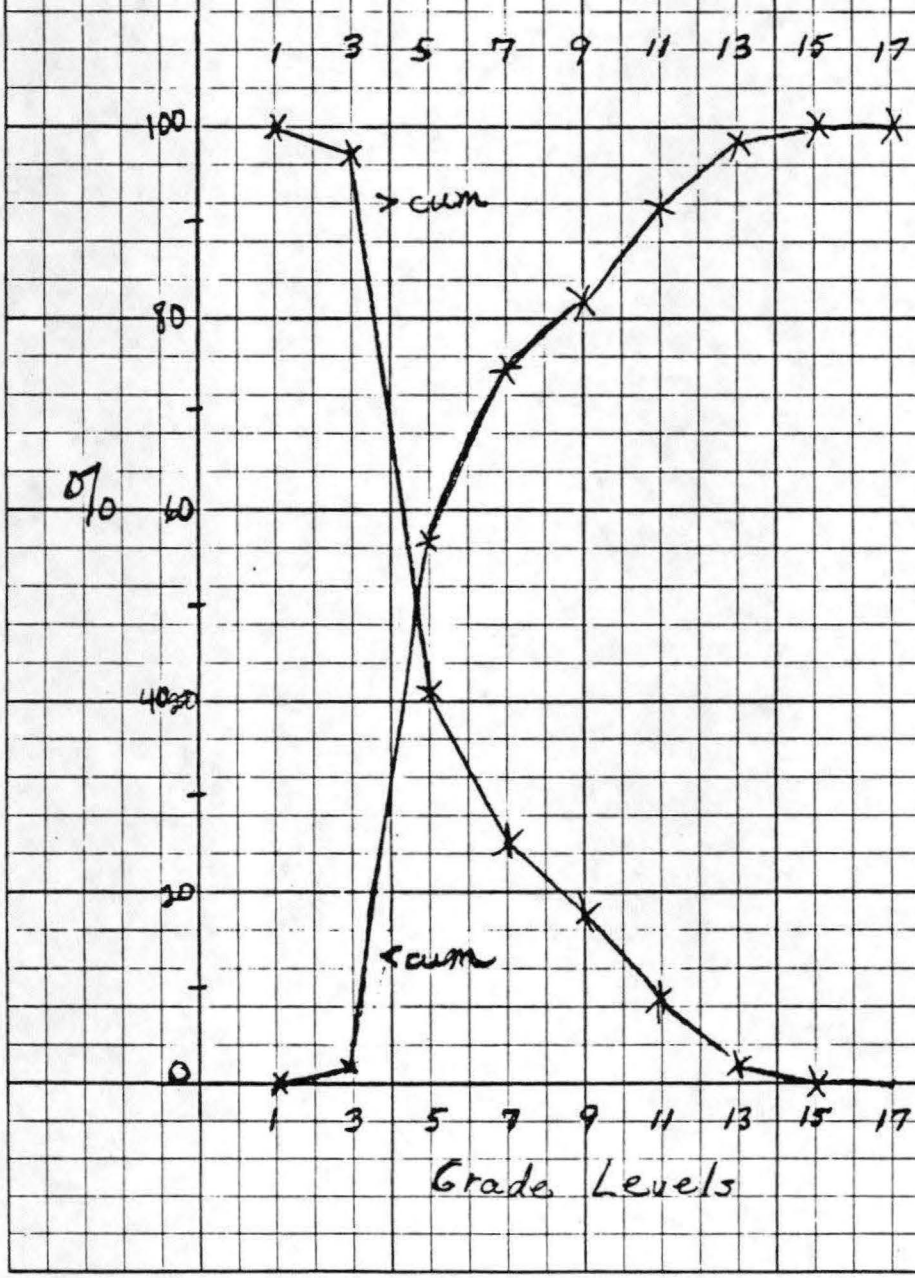
Figure 2-6



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

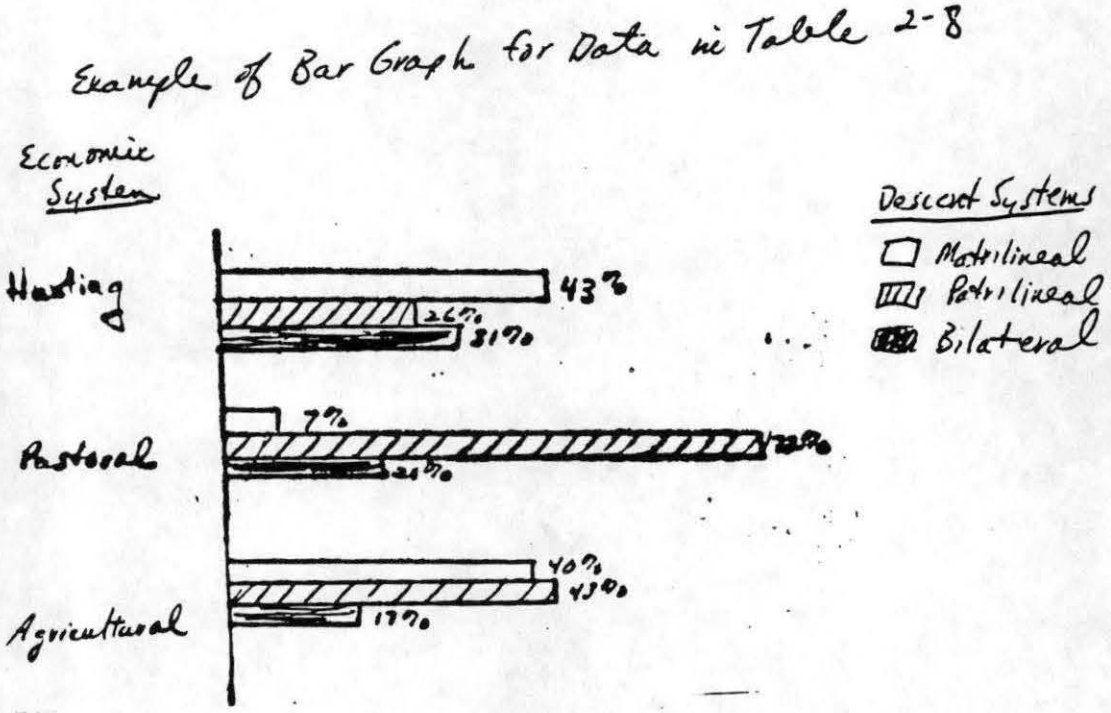
There are a number of graphs that are used with qualitative data. We will focus on bar charts, which are the most common. You may consult various statistics texts for examples of other types. As with the quantitative data, the bar charts are designed to display the data found in the tables in a way that 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
Economic System
Descent Systems
Matrilineal
Patrilineal
Bilateral
Herding
43%
26%
31%
Pastoral
7%
72%
21%
Agricultural
40%
43%
17%

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 sub-graph) 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.

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

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, 7, 8, 9, 11	13, 17, 19, 21, 22 — 5 cases

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

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

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

Ages of People Referred to Clinic

4 cases	6, 7, 8, 9	11 (median value); 13, 17, 19, 21 — 4 cases

Median = 11

Table 2-12 Degree of Support Respondents Report for President

	%
Highly Supportive	20
Mildly Supportive	40 ← Median Category
Neutral	10
Mildly Unsupportive	10 — 40%
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.

	Native Americans	non-Native Americans
1st quartile $Q_1 = L + \left[\frac{N/4 - cf}{f} \times i\right]$ (2-4)	$3.0 + \left[\frac{25-3}{55} \times 2\right]$	$7.0 + \left[\frac{25-20}{10} \times 2\right]$
	$= 3.8$	$= 8.0$
3rd quartile $Q_3 = L + \left[\frac{3N/4 - cf}{f} \times i\right]$ (2-5)	$5.0 + \left[\frac{75-58}{17} \times 2\right]$	$11.0 + \left[\frac{75-66}{24} \times 2\right]$
	$= 7.0$	$= 11.75$
1st Decile $D_1 = L + \left[\frac{N/10 - cf}{f} \times i\right]$	$3.0 + \left[\frac{0-3}{55} \times 2\right] = 3.25$	$5.0 + \left[\frac{10-9}{11} \times 2\right]$
		$= 5.18$

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.

RESPONSE	COLS. 38-39 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

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

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

$\bar{X} = \frac{\Sigma f_i X_i}{n}$

$= \frac{111}{12}$

$= 9.25$

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 Midpoint of Intervals	Frequencies (%) Native Americans	Frequencies (%) non-Native Americans	fx Native Americans	fx non-Native 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

= 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= $\frac{\Sigma f_i X_i}{n_i}$	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.

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

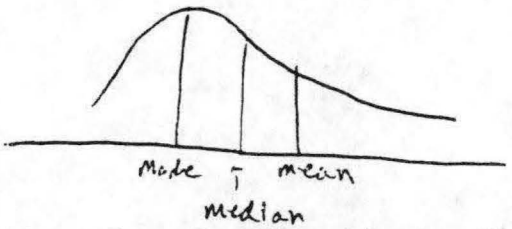


Figure 2-12 Example of a Negatively Skewed Distribution

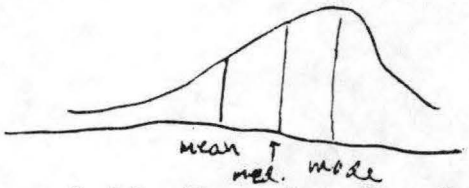


Figure 2-13 Example of a Symmetrical Distribution

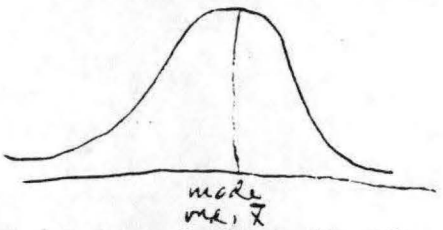
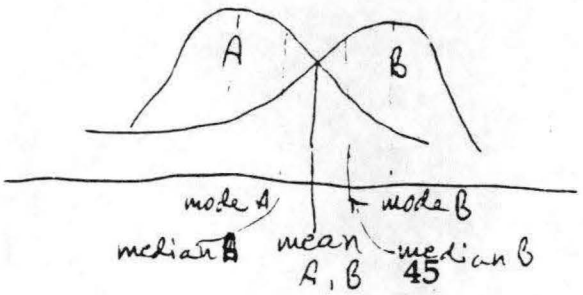


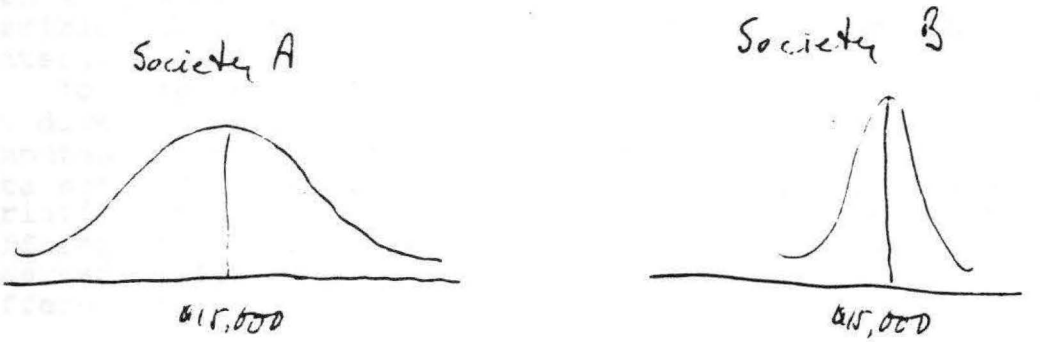
Figure 2-14 Example of Distributions with Equal Means and Unequal Medians and Modes



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, $S_o = (25 \times 25) + (25 \times 25) + (25 \times 25) + (25 \times 25) + (25 \times 25) + (25 \times 25) = 3750$

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

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

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

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

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	$\lvert X-Md \rvert$	$\lvert X-\bar{X} \rvert$
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,

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

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:

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

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

X_i	$\lvert X_i - Md\rvert$ NA	$\lvert X_i - Md\rvert$ non NA	NA freq.	non NA	$f\lvert X - Md\rvert$ 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:

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

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

Xi	Frequencies NA	non NA	$(X_i - \bar{X})$ NA	non NA	$(X_i - \bar{X})^2$ NA	non NA	$f(X_i - \bar{X})^2$ 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 $\hat{s}$ 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.

The Coefficient of Relative Variation -- 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

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

and for the mean

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

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 Inequality (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.

III. Univariate Inferential Statistics

In this section we examine inferential statistics that apply to one variable. Suppose we have information about people in a sample and we want to know how typical this information is of the corresponding total population. We will examine in this section how to make this type of inference.

To do this we must first explore the characteristics of the normal curve. We then go through definitions that are basic to statistical inference and develop the idea of a sampling distribution. Finally, we apply this information to developing confidence intervals around a mean. The confidence interval or band gives us a range of values in which a population parameter is likely to fall. We may specify through various manipulations how likely it is that a given parameter will fall within that band.

In this section all of our inferences will be made regarding the mean, a measure of central tendency. In later sections we will make inferences about other statistics. We will also examine how we can make inferences through hypothesis testing rather than through confidence intervals.

The Normal Distribution and Univariate Inferential Statistics

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. 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.

Normal distributions are most commonly found in natural situations. For instance, shoe size, height, weight, gestation periods, and other biological phenomena are generally normally distributed. 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

Figure 3-1

Comparison of normal curves with the same standard deviations but different means.

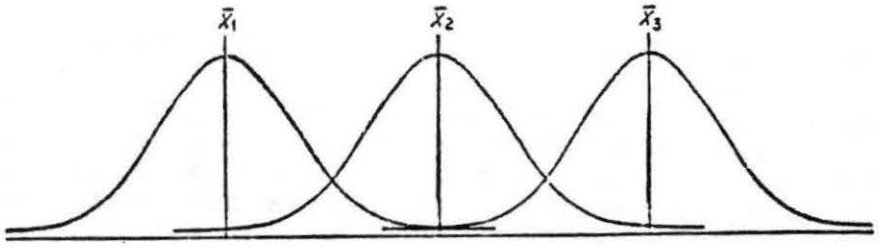
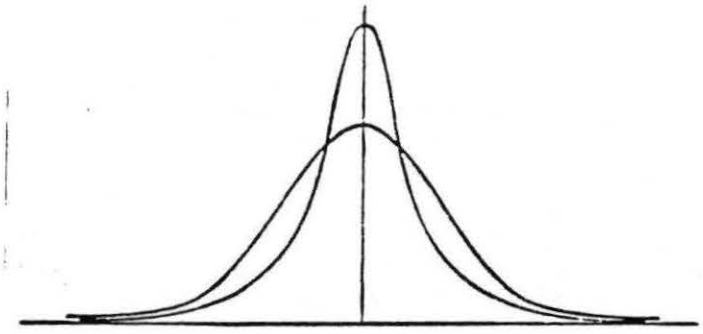


Figure 3-2

Comparison of two normal curves with the same means but different standard deviations.



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, for any normal distribution, the area under the curve and corresponding information (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. 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. Table 3-1 gives the areas under the curve for this standard normal distribution represented as N (0,1). (Some tables give the proportion of area found under the curve beyond a given standard deviation unit from the mean. To convert one table to another you simply would subtract a value for a given standard deviation unit from 0.5000.)

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 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 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 .500 + .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, 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[illegible] 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.

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 3-3, 3-4, and 3-5 and Figure 3-3 illustrate the development of a sampling distribution. Table 3-3 gives data for a total population: the suicide rates for 220 SMSA's in 1970. Table 3-4 gives the results obtained when samples, each sized 30, were taken from this population and the average suicide rate was computed. Table 3-5 gives a tally of these sample means and Figure 3-3 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.

Tables 3-5 and Figure 3-3 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. It is this theoretical sampling distribution that we use in making inferences from samples to populations.

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 3-6 gives the distribution of sampling errors for the group of samples in Table 3-4. 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:

$$\sigma_{\bar{x}} = \sigma / \sqrt{n} \qquad (3\text{-}2)$$

Table 3-3

Table 13.1.1 *Array of Suicide Rates for 229 United States Standard Metropolitan Statistical Areas, 1970*

2.7	7.3	8.7	9.6	10.7	11.6	12.7	14.3	16.9
3.3	7.4	8.8	9.7	10.7	11.7	12.8	14.3	17.0
3.8	7.4	8.8	9.7	10.7	11.7	12.8	14.5	17.2
4.6	7.4	8.8	9.7	10.8	11.7	12.8	14.6	17.5
5.0	7.4	8.9	9.8	10.9	11.8	12.8	14.7	17.8
5.2	7.5	8.9	9.8	10.9	11.8	12.8	14.9	17.9
5.2	7.5	8.9	9.8	11.0	11.8	12.8	15.1	18.3
5.5	7.6	9.0	9.9	11.0	11.9	12.9	15.1	18.4
6.0	7.6	9.0	9.9	11.1	11.9	12.9	15.2	18.6
6.3	7.7	9.1	9.9	11.2	12.0	13.0	15.2	18.7
6.3	7.7	9.1	9.9	11.2	12.0	13.0	15.4	19.0
6.4	7.7	9.2	10.0	11.2	12.0	13.1	15.5	19.4
6.5	7.8	9.2	10.0	11.3	12.0	13.2	15.6	20.0
6.5	7.9	9.2	10.0	11.3	12.1	13.2	16.0	20.1
6.6	7.9	9.3	10.0	11.3	12.2	13.2	16.0	20.6
6.6	8.0	9.3	10.1	11.4	12.2	13.5	16.1	20.9
6.7	8.2	9.4	10.2	11.4	12.2	13.6	16.1	21.0
6.7	8.4	9.4	10.3	11.5	12.3	13.6	16.1	21.8
6.7	8.4	9.4	10.3	11.5	12.3	13.7	16.1	22.0
6.9	8.5	9.4	10.4	11.5	12.4	14.0	16.2	22.1
6.9	8.5	9.4	10.5	11.6	12.4	14.0	16.3	22.5
7.1	8.5	9.5	10.5	11.6	12.5	14.0	16.4	22.5
7.2	8.6	9.5	10.5	11.6	12.7	14.0	16.5	24.8
7.2	8.6	9.5	10.6	11.6	12.7	14.1	16.7	24.9
7.3	8.7	9.6	10.6	11.6	12.7	14.1	16.9	25.0
7.3	8.7	9.6	10.7					

Sources: U.S. Bureau of the Census, *County and City Data Book, 1972*. Table 3. Washington, D.C.: U.S. Government Printing Office. 1973. U.S. Department of Health, Education and Welfare. *Vital Statistics of the United States, 1970*, Vol. II: Mortality, Part B. Washington, D.C.: U.S. Government Printing Office. 1974.

60 g

Table 3-4

Table 13.1.2 Array of 100 Sample Means, n = 30

10.3	10.8	11.5	11.8	12.3
10.4	10.8	11.5	11.8	12.3
10.4	10.9	11.5	11.9	12.4
10.4	10.9	11.5	11.9	12.4
10.4	10.9	11.5	11.9	12.4
10.5	11.0	11.5	11.9	12.4
10.6	11.0	11.6	11.9	12.4
10.6	11.0	11.7	11.9	12.4
10.6	11.1	11.7	11.9	12.5
10.6	11.2	11.7	11.9	12.5
10.6	11.2	11.7	11.9	12.5
10.6	11.2	11.7	12.1	12.9
10.6	11.3	11.7	12.1	12.9
10.8	11.3	11.8	12.1	13.0
10.8	11.3	11.8	12.2	13.0
10.8	11.3	11.8	12.2	13.0
10.8	11.3	11.8	12.2	13.0
10.8	11.3	11.8	12.2	13.0
10.8	11.3	11.8	12.2	13.2
10.8	11.4	11.8	12.2	13.2

$\bar{X} = 11.62$

Table 3-5

Table 13.1.3 Frequency Tally, Empirical Sampling Distribution, 100 Sample Means, n = 30

Class Interval	Tally	Frequency (f)
10.0–10.9	𝍸 𝍸 𝍸 𝍸	25
11.0–11.9	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 I	46
12.0–12.9	𝍸 𝍸 𝍸 𝍸 II	22
13.0–13.9	𝍸 II	7
		100

12.31
12.4

60h

Figure 3-3

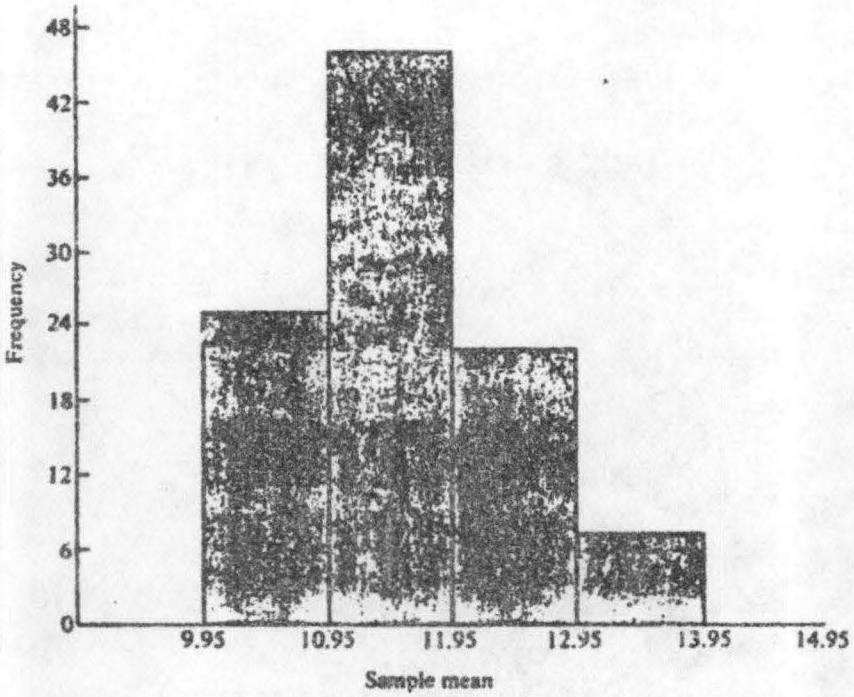


Figure 13.1.1 Histogram of 100 Sample Means, n = 30

Table 3-6

Table 13.1.4 Array of Sampling Errors, 100 Samples, n = 30

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

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

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}} \qquad (3\text{-}3)$$

SPSS routinely computes 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 n is used in equation 3-3 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

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad (3\text{-}4)$$

where

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

It should be stressed that the sampling distribution of the mean is normally distributed even when the frequency distribution for the population is not. 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. (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. Later we will discuss the use of other sampling distributions, but we will always use the same logic we develop below.

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. 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.

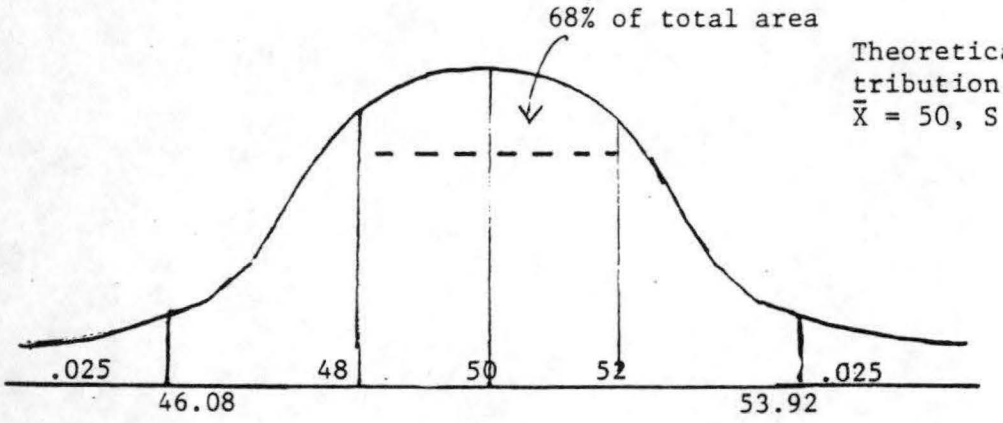
$$\bar{X} = 50, \quad s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} = 26$$

This information may be used to estimate the form of the sampling distribution. As explained above, $\bar{X}$ 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. Here $s_{\bar{X}} = \frac{26}{\sqrt{169}} = \frac{26}{13} = 2.0$

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 3-4. 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 3-4



Theoretical sampling distribution where $\bar{X} = 50$, $S = 26$, $n = 169$ $S_{\bar{x}} = 2$

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 (Table 3-1) 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

$$P\ [48 < \mu < 52] = .6826 \qquad (3\text{-}5)$$

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 Table 3-4

we can see that .025 of the area under the curve is remaining (.475 on one side of $\bar{X}$ 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.

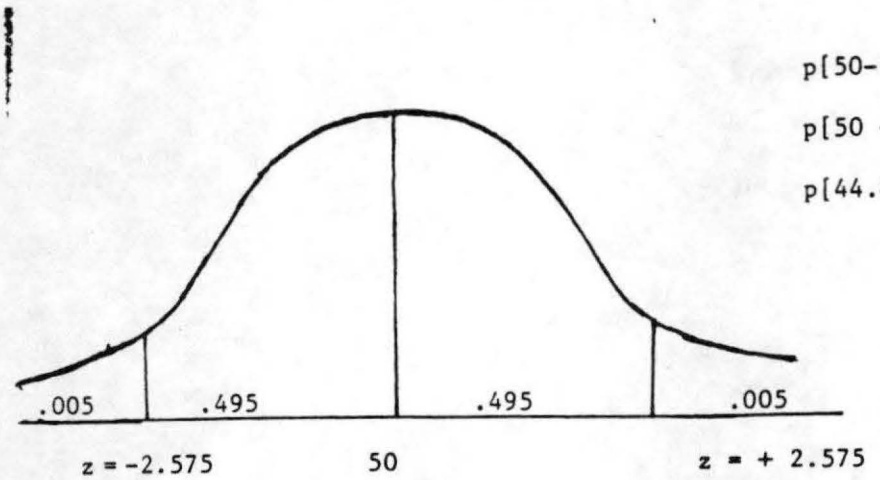
In the present example the estimated mean is 50 and the estimate of the standard error is 2. 1.96 standard errors is equal to 3.92. Thus, we may conclude that 95% of the means in the estimated sampling distribution are included between (50 - 3.92) and (50 + 3.92). This may be written symbolically as

$$P[46.08 < \mu < 53.92] = .95 \qquad (3\text{-}6)$$

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 2.55. The computations below and the figures show how the 99% confidence interval would be computed.

Figure 3-5



$p[50-(2.575(2)) < \mu < 50 + (2.575)(2)] = .99$

$p[50 - 5.15 \mu < 50 + 5.15] = .99$

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

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

$$P[\bar{X} - (1.96)(s_{\bar{X}}) < \mu < \bar{X} + (1.96)(s_{\bar{X}})] = .95 \qquad (3\text{-}7)$$

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

$$P[\bar{X} - (2.58)(s_{\bar{X}}) < \mu < \bar{X} + (2.58)(s_{\bar{X}})] = .99 \qquad (3\text{-}8)$$

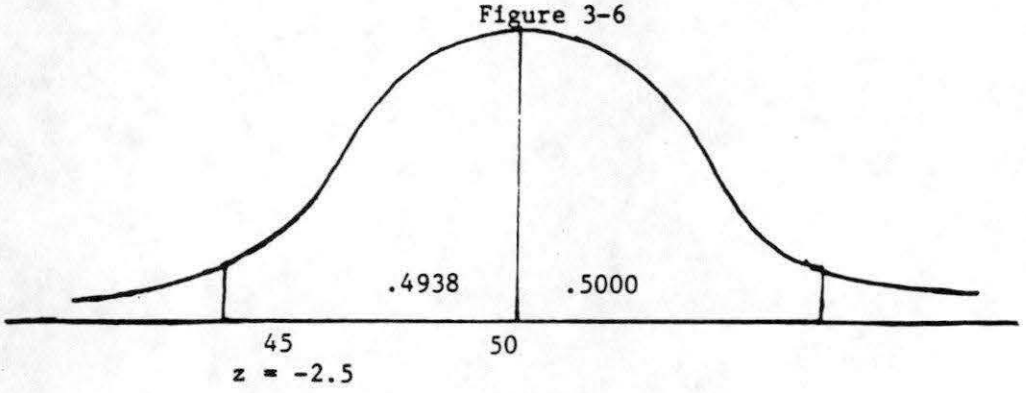
where $\bar{X}$ is the sample mean and $s_{\bar{X}}$ is the estimated standard error.

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 $\bar{X} = 45$ is from the mean of the theoretical sampling distribution. We may do this using standard scores.

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

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

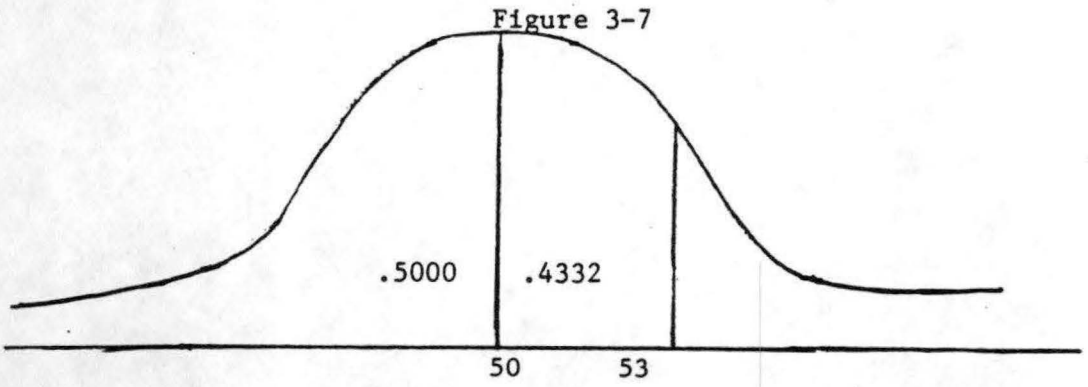
Figure 3-6



Using the table of areas under the normal curve (Table 3-1) 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 $\mu < 53$ we must determine how far away 53 is from the estimated mean of the sampling distribution, 50. $z = (\bar{X} - \mu)/s_{\bar{x}} = (53-50)/\ 2 = 1.5$. This indicates that 53 is 1.5 standard errors above the estimated sampling distribution.

Figure 3-7



Using the table of area under the normal curve (Table 3-1) we can find that

$P [\mu < 50] = .5000$

$P [50 < \mu < 53] = .4332$

and thus $P [\mu < 53] = .5000 + .4332 = .9332$.

There is a 93% probability that the population mean is less than 53. Similarly, $P [\mu > 53] = .5000 - .4332 = .0668$. 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.

The computer output for the subprogram frequencies 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 3-8 below.

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_0.

The alternative or substantive or research hypothesis is the alternative to the null hypothesis. For example, if the null hypothesis, H_0 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: $\mu \neq 100$

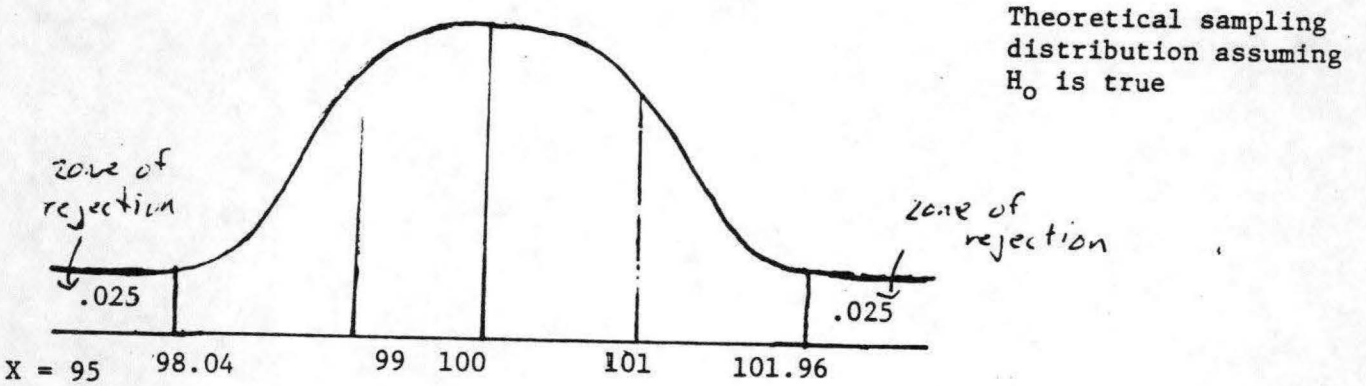
Suppose we draw a random sample from the population involved. In this sample $\bar{X} = 95$, $s = 13$, $n = 169$

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 $S_{\bar{x}}$ 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 3-9 below.

Figure 3-9



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 sampling distribution for the population if H_o 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_o 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

$$z = \frac{\bar{X} - \mu}{s_{\bar{X}}} = (95-100)/1.0 = -5/1 = -5.0$$

Locating this z-value on Table 3-1 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

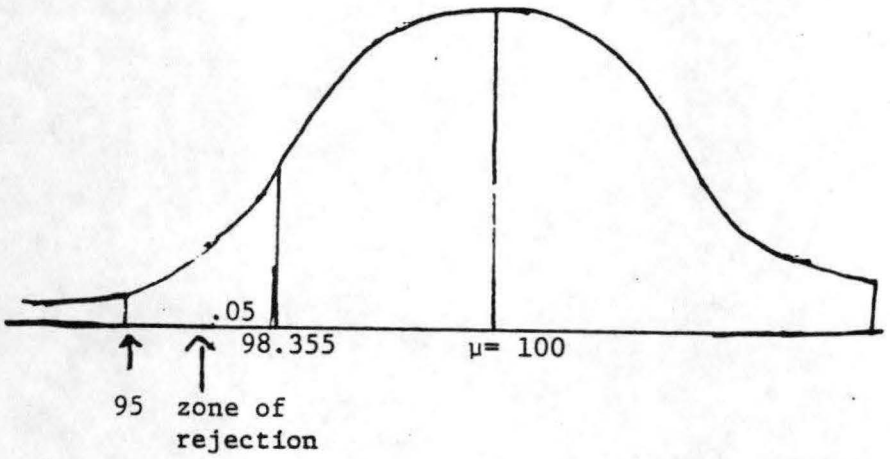
specified above. For instance, suppose in the example above the hypotheses had been

H_o: $\mu = 100$ or ≥ 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 3-10.

Figure 3-10



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_o: $\mu = 100$ in favor of H_1: $\mu < 100$ 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_o is true is developed and the sample values are compared against the "critical values" on that sampling distribution. The critical value 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 theory, 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.

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 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. We will examine the t-distribution in detail in a later section.

Inferences About Proportions

While, technically, inferences about proportions involve the binomial distribution, rather than the normal distribution, they can be seen as simply a special case of inferences about means and use the normal distribution, as long as one has a relatively large sample. The procedure one would use with the binominal distribution is essentially the same as described here. In

addition, the similarity to inferences regarding means holds whether one is interested in confidence intervals or hypothesis tests.

When one has a research problem which involves a proportion, one essentially has a nominally measured variable which is a dichotomy. For instance, if one is interested in the proportion of people who support a ballot measure, one is interested in the people who vote yes and the people who vote no. Suppose that one arbitrarily assigns a score of 1 to those who would vote yes and a score of 0 to those who would vote no. Suppose also that in a sample of 625 people 325 indicated they would vote yes, while 300 indicated they would vote no. This means that 52% of the people supported the measure. The computations for the mean and standard deviation for this sample are shown below, using the assigned scores.

Table 3-11
Example of Computation of Mean and Standard Deviation of Proportions

Score (X)	f	fX	x^2	fX^2	
0	300	0	0	0	
1	325	325	1	325	n=625
Total	625	325		325	

$\bar{X} = \Sigma fX / n = 325/625 = .52$

$s = \sqrt{\frac{\Sigma X^2}{n} - \bar{X}^2} = \sqrt{\frac{(325)}{625} - (.52)^2} = \sqrt{.52 - .27} = \sqrt{.25}$

$= .50$

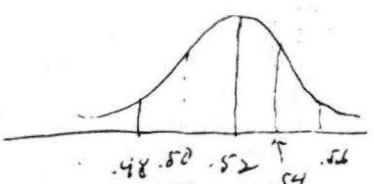
Note that the mean simply equals the proportion who voted yes, which can be signified as p_u. The standard deviation simply equals the square root of the product of the proportion who voted yes (p_u) and the proportion who voted no (q_u). Blalock provides a proof of this relationship: $s = \sqrt{p_u q_u}$

Suppose that one wanted to estimate with 95% confidence the proportion of voters in the population who would support the given ballot measure. One could use the familiar formula for estimating the 95% confidence interval around a mean. The estimated standard error would be

$$s_p = s / \sqrt{n} = .50/25 = .02.$$

The associated sampling distribution is shown below in Figure 3-12.

Figure 3-12



The 95% confidence interval around the estimate of the population proportion of .52 would be calculated as

$$P\,[.52 - (1.96)(.02) < P < .52 + (1.96)(.02)] = .95$$

$$= \quad P\,[.52 - .04 < P < .52 + .04] = .95$$

$$= \quad P\,[.48 < P < .56] = .95.$$

Thus, based on the data from this sample, we could be 95% confident that between 48% and 56% of the people in the population would vote yes on the ballot measure. Note that this confidence interval crosses the 50% mark and thus we cannot be 95% confident that the measure would pass.

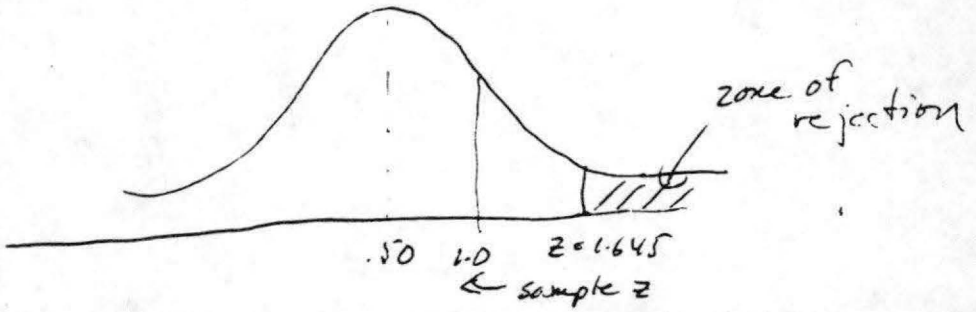
Suppose that we wanted to test a hypothesis regarding the fate of the ballot measure and suppose that we suspected that it probably would be supported. Our research hypothesis then would be

$H_1: P > .50$

and our null hypothesis would be $H_0: P \leq .50$

The associated sampling distribution, assuming Ho is true is shown in Figure 3-13.

Figure 3-13



Suppose that we decided to use a .05 level of significance. With a one-tail test our critical value of z would be 1.645 and we could compute our sample value of z with the familiar formula where we subtract the hypothesized value of the proportion (or mean) from our actual value and divide by the standard error.

$$z = \frac{p - P}{s_p} \qquad = \frac{.52 - .50}{.02} = 1.0$$

Our resulting z value of 1.00 is far from falling in the zone of rejection and we must fail to reject the null hypothesis. In fact, consulting the table of the normal curve (Table 3-1) shows

us that we would actually be wrong 16 times out of 100 if we were to reject the null hypothesis in favor of the alternative.

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.

IV. Statistics for Data Measured on an Ordinal and Nominal Scale: Chi-Square, and Measures of Association

Most of the statistics we will discuss in this class are designed for variables measured on an interval level and for variables that have a normal distribution within the population. These are called parametric statistics. Sometimes, however, we have data that are measured on less than an interval scale and/or which are not normally distributed in the population. For these data it is appropriate to use non-parametric or distribution free statistics. Such statistics have weaker assumptions and requirements than the parametric statistics and do not require interval measurement or normal distributions.

Some cautions should be noted. The so-called distribution free statistics are not always distribution free. For instance, studies have found that the Mann Whitney U (a non-parametric test) is more dependent, in some circumstances, on the shape of a variable's distribution than the t-test (a parmetric test we discuss in the next section). Also, although non-parametric tests have weaker assumptions and requirements than their parametric counterparts they also have less power. That is, when using these tests there is a greater possibility of making a type II error (to be discussed in the next section). They are also often less flexible and thus less useful, especially when dealing with multivariate relationships. Thus, there is a trade-off. In using non-parametric statistics one can relax assumptions, but one can't do as much often in analyzing one's data, and chances of type II errors may be higher.

Measurement is currently a central focus of researchers in methodology, and the problems with non-parametric statistics are one impetus to this research. We want to be better able to measure variables that we believe can be measured eventually on an interval scale. We want to come closer to tapping the true dimensions of a variable. This is especially true when we are measuring something that we believe probably can eventually be measured on an interval scale, but we presently are only measuring it with an ordinal scale. We want to try to get our ordinal scale to more closely approximate the interval scale that best represents the measure so we can use the more powerful statistics designed for intervally measured variables.

Most people in the social sciences then treat ordinally measured variables as if they were intervally measured, trying to approximate the interval scale as much as possible. (One simple way to do this or help do this is to retain as many points in the scale as possible -- e.g. by using a summated Likert scale rather than one item answered in Likert fashion.) The non-parametric techniques I discuss in this section are then most appropriate for variables which can best be measured on a nominal level. If data approach an interval scale I suggest using the more powerful

parametric statistics. I specifically caution against collapsing intervally measured variables into a few categories and proceed to analyze them as though they were nominally measured. This results in using statistics that are not the most powerful or appropriate for the data. There is also a very real danger of having different results depending on where you choose to collapse the data (that is, where you make the cut-points for the collapsing).

Given these cautions I proceed below to discuss two broad areas of statistics appropriate for nominally and ordinally measured data and extensions and elaborations of these statistics. First, I discuss the chi-square distribution and then measures of association for contingency tables. The former is extremely important, not just for its use with nominally measured variables, but also for its use with more advanced and very important statistical techniques.

Chi - Square

The chi-square distribution is commonly used when we want to test hypotheses regarding whether or not observed frequencies differ from those we would expect by chance or by some theoretical model. It can be used with univariate, bivariate, and multivariate contingency tables, and examples of each of these uses are discussed below. Just as with the t and F distributions, which we discuss more later, the chi-square distribution is really a family of distributions, the shape of the curve depending on the number of degrees of freedom. In fact, the mean of the distribution equals the degrees of freedom and the variance equals twice the degrees of freedom. This produces distributions for the lower degrees of freedom (1 and 2) that are J-shaped curves, as shown in the diagram below. As the degrees of freedom increase the curve becomes more skewed to the right and gradually more symetrical.

It is important to note that in the cases we discuss below the degrees of freedom are based on the number of frequencies involved in the study (the number of categories being studied) and not on the sample size. (The chi-square distribution is also used to test hypotheses about a single variance and to put confidence limits around variances and in those cases the degrees of freedom are a function of sample size.)

There are a number of ways to estimate the value of chi-square for a sample. For categorical data (which we will only be concerned with in the following discussion) they involve the difference between the frequencies expected in each category and the actual observed frequencies. The expected frequencies may be those that would be expected by chance or those that would be expected if some kind of particular association or shape of the distribution were true. This format is what makes the chi-square distribution so useful in advanced statistics. One can posit a certain type of model, even a very complex multivariate model, and determine how data should look if that model were true. This would determine your expected frequencies, and you could then compare these expected frequencies to those which you actually have with your data (your observed frequencies). The most common formula used for the computation of chi-square is

$$\chi^2 = \sum \frac{(fo - fe)^2}{fe}$$ where fo = observed frequencies, fe = expected frequencies

Univariate or Single-Sample Tests -- Sometimes a researcher will have information about a particular group of subjects and will be interested in the distribution of certain variables for these subjects. Champion (1981) gives an example from Metz's (1966) study of people who were opposed to the fluoridation of water. Metz had information about their knowledge of the effects of fluoridation and had categorized this knowledge as (1) correct, (2) incorrect, and (3) uncertain. He wanted to know if the people who were unfavorable toward fluoridation differed at all in their knowledge about it. He thus proposed the following hypotheses:

H_o: Persons who are unfavorable toward fluoridation will not differ in their knowledge of the effects of fluoridation.

H_1: Persons who are unfavorable toward fluoridation will differ in their knowledge of the effects of fluoridation.

Let us suppose that he set a .001 level of significance (two-tailed test). Table 4-1 below gives the actual frequencies observed by Metz. Table 4-2 gives the frequencies that would be expected if the subjects did not differ in their knowledge (the results assumed in the null hypothesis). Using the chi-square formula given above we may estimate the chi-square value for this sample. The computations follow the tables.

Table 4-1

Knowledge of fluoridation of subjects who were unfavorable toward fluoridation (observed frequencies)

	Correct Knowledge	Incorrect Knowledge	Uncertain	Total
n	121	40	31	192
%	63	21	16	100

Table 4-2

Expected distribution of knowledge of fluoridation for subject who were unfavorable

	Correct Knowledge	Incorrect Knowledge	Uncertain	Total
n	64	64	64	192
%	33.3	33.3	33.3	100

$$\chi^2 = \sum \frac{(fo-fe)^2}{fe} = \frac{(121-64)^2}{64} + \frac{(40-64)^2}{64} + \frac{(31-64)^2}{64}$$

$$= \frac{57^2}{64} + \frac{24^2}{64} + \frac{33^2}{64} = \frac{3249}{64} + \frac{576}{64} + \frac{1089}{64}$$

$$= 50.766 + 9.000 + 17.016 = 76.782$$

To determine whether or not this sample value of chi-square falls in the zone of rejection on the sampling distribution, we must know which chi-square distribution to use -- that is, we must determine the degrees of freedom. With this single sample test we have used only one degree of freedom. Essentially we could put numbers in all the cells except one, and then we would be required to put a particular value in that last cell to have our correct sample size. Our degrees of freedom in a simple one-sample test such as this are equal then to k - 1, where k = the number of categories. In this case our degrees of freedom equal 3-1= 2.

We can then use a table of chi-square values such as that shown in Table 4-3, as taken from Blalock's appendix. This table gives the proportion of area under the curve of the sampling distribution, or the various possible levels of significance, across the top of the table. The possible degrees of freedom are

given down the left hand side. In the body of the table, at each intersection of probability levels and degrees of freedom are the actual chi-square values (the critical values) that correspond to each probability level and degrees of freedom. For df = 2 and the .001 level of significance it may be seen that the critical value of chi-square is 13.815. Our sample chi-square value of 76.782 is much larger than 13.815 and thus we may reject the null hypothesis with less than a .001 chance of being wrong in doing so.

Note that at this point the researcher must return to the actual data to determine how the observed frequencies differ from those that would be expected. It may be seen that those opposed to fluoridation much more often have correct knowledge and much less often have either incorrect of uncertain knowledge. The researcher would then discuss theoretical explanations for these findings.

In the computations above we used a definitional formula for chi-square. This formula involves numerous subtractions and these repeated subtractions lead to rounding error. It is thus suggested that when doing chi-square tests by hand that the researcher should use a computing formula as follows:

$$\chi^2 = \left[\sum \frac{fo^2}{fe}\right] - N$$

Blalock's text includes a proof that this computing formula is equivalent to the definition. Below we repeat the computations for the example given above using this computing formula.

$$\chi^2 = \left[\sum \frac{fo^2}{fe}\right] - N = \frac{(121)^2}{64} + \frac{(40)^2}{64} + \frac{(31)^2}{64} - 192$$

$$= \frac{14641}{64} + \frac{1600}{64} + \frac{961}{64} - 192$$

$$= 228.76 + 25 + 15.02 - 192 = 76.78$$

Clearly the chi-square test is easy to compute and easy to interpret. It is thus widely used. Its major disadvantage is that it is extremely sensitive to sample sizes. If sample sizes are large the chi-square value can be easily inflated, even though the actual distribution has not changed at all. You can easily demonstrate this to yourself by simply doubling the values in Tables 4-1 and 4-2 and recomputing the chi-square. The chi-square value that you will get will be much larger, even though the actual proportions remain the same. Similarly, if the sample is so small that the expected frequencies in each category are

small (usually five of less for any one cell), then the chi-square values can be easily distorted and much higher than they should be. Again, you may demonstrate this to yourself by trying tests with artificial samples of varying size. In general, you should avoid using chi-square tests with very small samples and take results with large samples with a grain of salt. (You should, in the bivariate case use the measures of association we discuss below in conjuction with the chi-square also.) Sometimes it is possible to collapse various categories to raise the expected cell frequency to exceed five. When this is done you should make sure that the collapsed categories make theoretical sense. If they do not have some theoretical justification then your resulting chi-square will be meaningless.

The most important use of the single-sample or univariate use of chi-square is in testing the hypothesis that a given distribution assumes a particular shape. Oftentimes one would be interested in testing the hypothesis that a sample comes from a population with a normal distribution on the variable being considered, that is, testing the null hypothesis that the variable is normally distributed within the population.

Consider the data displayed in Table 4-4. These are the cumulative grade point averages of seniors who graduated from two high schools in a western city in 1978. Suppose we were interested in testing the following hypotheses:

H_o: The variable of cumulative grades has a normal distribution within the population.
H_1: The variable of cumulative grades is not normally distributed within the population.

(Note that the population to which we are inferring here is essentially a hypothetical one and for all practical purposes we are testing the possibility that any deviations of the distribution given from a normal one are only deviations that occur by chance.)

We may use the chi-square distribution to test this hypothesis. Our observed frequencies are those given in Table 4-4, and our expected frequencies are those that would be expected if the distribution were normally distributed. To determine these expected frequencies we need our best estimates of the population mean and standard deviation and these come from our sample data:

$\bar{X} = 2.85$, $s = .59$ (rounded to the nearest hundredth).

Now we may group our data into categories and, using Table 3-1, the table of the normal curve, compute the proportion of area that would be in each category if the data were normally distributed. We can then use these expected proportions to calculate the actual number of cases that would be expected in each category, given that there is a sample size of 569. A chart

Table 4-4

Raw Data, Computation of χ^2, and Histograms of Expected and Observed Frequencies (page 1)

VARIABLE CUM GPA

CODE	FREQ	ADJ PCT	CUM PCT	CODE	FREQ	ADJ PCT	CUM PCT	CODE	FREQ	ADJ PCT	CUM PCT
0.8	1	0	0	2.2	20	4	16	3.2	34	6	74
1.2	1	0	0	2.3	26	5	21	3.3	21	4	78
1.3	3	1	1	2.4	35	6	27	3.4	21	4	81
1.4	1	0	1	2.5	34	6	33	3.5	26	5	86
1.6	3	1	2	2.6	37	7	39	3.6	23	4	90
1.7	3	1	2	2.7	27	5	44	3.7	16	3	93
1.8	6	1	4	2.8	32	6	50	3.8	17	3	96
1.9	14	2	6	2.9	34	6	56	3.9	13	2	98
2.0	21	4	10	3.0	40	7	63	4.0	9	2	100
2.1	17	3	13	3.1	30	5	68				

MISSING DATA

CODE	FREQ	CODE	FREQ	CODE	FREQ
9.9	1				

MEAN	2.848	STD ERR	0.025	MEDIAN	2.854
MODE	3.000	STD DEV	0.591	VARIANCE	0.349
KURTOSIS	-0.450	SKEWNESS	-0.118	RANGE	3.200
MINIMUM	0.800	MAXIMUM	4.000		

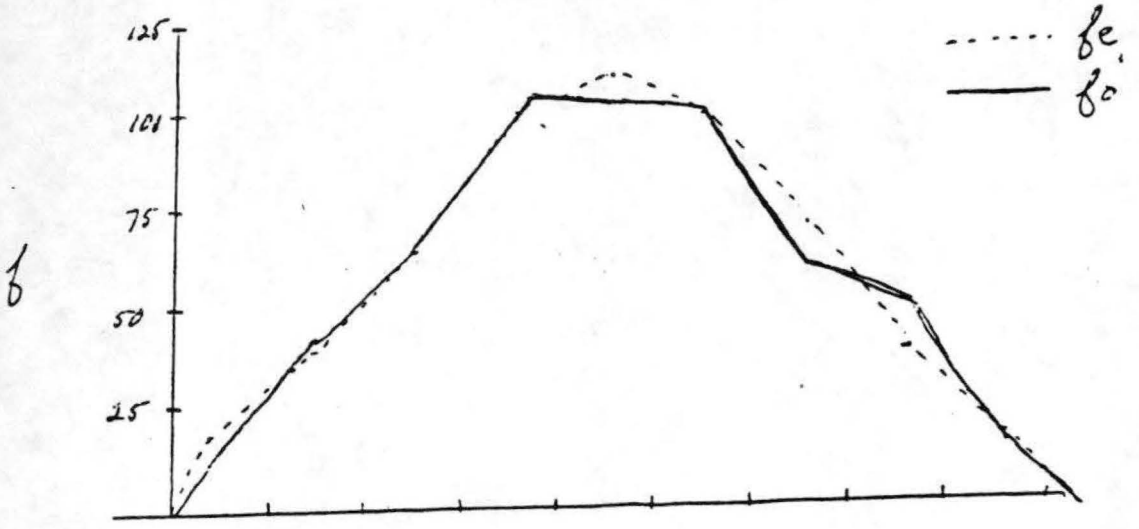
VALID CASES 565 MISSING CASES 1

Table 4-4
(continued)

Range (true values)	Z Corresponding to upper limit	P. of Area under curve	fe	fo	fo^2/fe
<1.75	-1.86	.5000-.4686=.0316	18.0	12	8.0
1.75-2.05	-1.36	.4686-.4131=.0550	31.6	43	58.5
2.05-2.35	-0.85	.4131-.3023=.1108	63.0	63	63
2.35-2.65	-0.34	.3023-.1331=.1699	96.6	106	116.3
2.65-2.95	+0.17	.1331+.0675=.2006	114.0	93	75.9
2.95-3.25	+0.68	.2518-.0675=.1843	104.8	104	103.2
3.25-3.55	+1.19	.3830-.2518=.1312	74.5	68	62.1
3.55-3.85	+1.69	.4585-.3830=.0755	42.9	58	78.4
3.85	—	.5000-.4585=.0415	23.6	22	20.5
		1.00			
Totals			569	569	585.9

$$\chi^2 = \sum \frac{fo^2}{fe} - N = 585.9 - 569 = 16.9 \qquad df = 9-3 = 6$$

$$.001 < p\left[\chi^2 = 16.9 \ df=6\right] < .01$$



$\bar{X} = 2.85$
$s = .59$

that contains all of the relevant data, the computation of the chi-square value, and the histograms of the expected and actual frequencies are included in Table 4-4.

These computations are exactly like those discussed when we studied the normal curve, and thus should be familiar to students. For instance, the first category listed includes all values less than 1.75. Given the mean of 2.85 and standard deviation of .59, we can calculate that a score of 1.75 would fall 1.86 standard deviations below the mean ($z = 1.75-2.85/.59 = -1.86$). The table of the normal curve (Table 3-1) indicates that between the mean and 1.86 standard deviation units below the mean there is .4686 of the area under the curve. Since .50 of the total area is less than the mean, .50 - .4686 = .0316 of the area should be below 1.75. In other words, if this distribution were indeed a normal one, with a mean of 2.85 and a standard deviation of .59, we would expect only .0316 of all the cases to have scores lower than 1.75. Since there are 569 cases all together in the sample, this would translate into 18 cases (.0316 x 569 = 18). In other words, we would expect 18 actual cases in our sample of 569 to be in this interval if the distribution were a normal one. This is our expected frequency if the null hypothesis were true. In actuality, we can see that only 12 cases fell into this interval. That is, our observed frequency was only 12.

Similar procedures are used for each interval. Students should carefully work through the calculations for each interval to ensure that they understand the procedure. To check that computations are correct one should always add up the expected proportion under the curve (it should equal 1.0) and the observed and expected frequencies (they should both equal the sample size).

Once the sample chi-square value is computed we need to determine where this value falls on the sampling distribution that is constructed assuming the null hypothesis is true. We thus need to determine the degrees of freedom for our problem so that we examine the correct sampling distribution. For this problem we have nine categories or nine total frequencies that we are examining. To compute our expected frequencies we need to compute a mean and a standard deviation, thus losing two degrees of freedom. In placing frequencies in categories we may freely choose all frequencies but one, thus losing one more degree of freedom. All together then we lose 3 degrees of freedom from our total of nine and have 9 - 3 = 6 degrees of freedom. We thus use the sampling distribution of chi-square with 6 degrees of freedom and compare our sample value with the critical values in that row of the table for chi-square in Table 4-3. It may be seen that our sample value of 16.9 falls between the critical values for a .01 and .001 level of significance. Thus, we may reject our null hypothesis that the grades are distributed normally in the population and be wrong less than 1 time in 100, but more than 1 time in 1000 in doing so.

Note that in this problem we usually want to fail to reject the null hypothesis, that is, we hope that our distribution is a normal one in the population. Thus, the conservative step here is to use a large level of significance such as .10. With this problem however, our exact probability of being wrong in rejecting H_o is much less, in fact less than .01.

Examining the frequency polygons of the actual and expected frequencies can lend insight into the nature of the deviations of the distribution. In our example, it may be seen that there is no clear pattern in the deviations; it is not one tail or the other that deviates, but a seemingly random pattern of differences between the expected and observed frequencies except for a flatter pattern in the observed frequencies around the mean. One could suspect, then, that at least one source of the high chi-square value in this case is the large sample size.

Bivariate Uses of Chi-Square -- Besides its use in these single sample tests, chi-square is often used in determining if frequencies in contingency tables differ from those that would be expected by chance. In such a case the hypotheses are as follows:

H_o: Variable X and Variable Y are statistically independent of each other

H_1: Variable X and Variable Y are not statistically independent of each other.

Consider the following example of the relationship between the race of workers and the type of job which they have. Each of these variables will have only two categories in this example: Race, white and black; Job, white collar and blue collar. The actual observed frequencies (hypothetical) are shown in Table 4-5.

Table 4-5
Hypothetical Cross-Tabulation of Race and Occupation

		Race: White	Race: Black	Total
Type of Job	White Collar	175	25	150
	Blue Collar	175	75	250
		300	100	400

To compute the frequencies that we would expect by chance (the situation when there is statistical independence) we need to consider the marginal frequencies. Note that 150/400 = 3/8 of all the subjects have white collar jobs. Note also that 300/400 = 3/4 of all the subjects are white. That means that if the data within the contingency tables are arranged as would be expected by chance 3/4 of the blue collar and 3/4 of the white collar workers should be white. Similarly, 3/8 of the whites should have white collar jobs and 3/8 of the blacks should have white collar jobs. An easy way to compute these expected frequencies for each cell is to multiple the two marginal frequencies by each other and divide by the sample size. For instance, the expected frequency for white, white-collar workers is (300)(150)/400 = 112.5. The expected frequencies computed by this logic are shown in Table 4-6.

Table 4-6
Expected frequencies for Hypothetical Cross-Tabulation of Race and Occupation

		Race: White	Race: Black	Total
Type of Job	White Collar	112.5	37.5	150
	Blue Collar	187.5	62.5	250
		300	100	400

The chi-square value may now be computed in the typical fashion by comparing the expected and observed frequencies; and these computations are shown below.

$\chi^2 = \sum \frac{fo^2}{fe} - N$

fo	fe	fo^2	fo^2/fe
125	112.5	15625	138.89
25	37.5	625	16.67
175	187.5	30625	163.33
75	62.5	5625	90.00
400	400		408.89

$= 408.89 - 400$

$= 8.89$

$df = 1$

$.01 < P[\chi^2_{df=1} = 8.89] < .001$

To compare this sample chi-square value to that which would occur if the null hypothesis were true it is necessary to determine the degrees of freedom. Again the degrees of freedom

are determined by looking at the number of free choices of frequencies that are possible given the marginal frequencies. In the bivariate case with two values in each variable it may be seen that there is only one degree of freedom. If a value is placed in the white race/white collar job category the values for all the other cells are automatically determined. Similarly, if a value is placed in the black race/white collar job category the values for all the other cells are automatically determined. Thus, for this example there is only one degree of freedom.

The sample value of chi-square of 8.89 may then be compared to the values in the chi-square table in Table 4-3 for the row with degrees of freedom equal to one. It may be seen that the sample value of 8.89 falls between the critical values for a level of significance of .01 and .001. Thus, we may reject the null hypothesis that race and type of job are independent in favor of the alternative that they are indeed statistically dependent and be wrong in doing so less than 1 time out of 100 but more than 1 time out of 1000.

Chi-square may be used to examine frequencies in contingency tables that are larger than 2x2 cells. In computing chi-square with the larger tables the expected cell frequencies are still computed by using the marginal values and getting frequencies in each row and column that are proportional to the marginal frequencies. As noted above, an easy way to compute these expected frequencies for each cell is to multiply the row total by the column total (that corresponds to the cell) and dividing by the total n. In larger tables the degrees of freedom also increase. In general the degrees of freedom equal (r-1) (c-1), where r is the number of rows and c is the number of columns.

Chi-square retains its advantages and disadvantages in the bivariate case. It becomes unreliable when the expected cell frequencies are small, especially when they are less than 5. It also is inflated when the sample size is large. Most importantly, the chi-square value tells us nothing of the actual shape of the distribution or the nature of the relationship between the two variables that are being studied. Chi-square, however, is easy to compute and is a favorite test of significance when examining bivariate tables.

Multivariate Tables -- In the last 10 to 15 years techniques have been developed to use chi-square in analyzing multivariate contingency tables. Blalock briefly refers to these techniques in his text, although a complete treatment of the area requires a knowledge of topics to be discussed later in this course. These techniques are usually called log-linear models, or also discussed in general terms of categorical models.

The general idea in these models is to develop expected or hypothesized models for relationships among the variables being studied, compute expected frequencies for each of these models, and then compare the actual observed frequencies to those

expected under each model. The researcher is generally interested in finding the simplest model that can best account for the relationships in the data. Thus, as with our work with the test that a distribution is normally distributed, the idea with the log-linear model is to fail to reject the null hypothesis for the model that has the best fit. The researcher looks for the lowest chi-square value in conjuction with the simplest model to describe the data. It is then possible to obtain parameters that describe the extent of various relationships within the model. These parameters are based on "odds-ratios" and the logarithms of these ratios. Many extensions and variations of these methods are possible; they are simply too complex and varied to pursue at this point. Suffice it to say that if one is dealing with categorical data today, the best type of analysis procedure to use is one of the log-linear type models. Most other analysis techniques are not accepted by journals.

Chi-square is also used as a way to test the fit of multivariate models developed through structural equations and with proportional hazard techniques. The former involves an intervally measured dependent variable and interval independent variables and can handle hypotheses regarding two-way causation and panel data. The latter involves over-time data with a number of data points. Again, with both of these techniques elaborate multivariate theoretical models are developed, the frequencies expected with these models are computed, and these expected frequencies are compared with the actual frequencies using a chi-square statistic.

Measures of Association -- Given this pitch for log-linear analysis, it is still important to briefly mention the large number of measures of association that have been developed for non-parametric data. Such measures of association are often easy to compute and handy to use when looking at data from field research or doing exploratory or preliminary analyses. Blalock and other texts give extensive discussions of many of these measures.

A number of measures of association can be derived from the chi-square value for a contingency table. These measures are usually some function of chi-square and the sample size; however, they lack easy interpretations and are not as useful as a family of measures that have a "proportionate-reduction-of-error" interpretation.

Here we will show the computations of only one measure of association, lambda (λ), a measure appropriate for nominally measured variables. There are many such measures, however, and the text may be consulted for details on computations of others.

The general format of a measure with a PRE (proportionate-reduction-of-error) interpretation is as follows:

$$\text{PRE} = \frac{\text{reduction in error with more information}}{\text{original error}}$$

Essentially, the measure compares how much error is reduced once we have more information (usually about an independent variable) from our error that we had with only our original information (usually about the dependent variable alone).

Lambda is an asymetric measure. That means that we need to designate our dependent and independent variable and that the value of lambda varies depending on which variable is called dependent or independent.

In computing lambda we are essentially only concerned with the categories of our variables. We are concerned with predicting the category in which most people fall on our dependent variable and the extent to which knowing about the independent variable helps improve our prediction of the category in which people fall in the dependent variable.

Consider the example in Table 4-7, taken from Loether and McTavish, of the relationship between the marital status of the head of household and the sex and parental status of the household head. Let us say that the marital status is our dependent variable. If we only know the marginal frequencies on this variable we would predict that heads of households are usually married, and we would be wrong in this prediction 64,372 - 45,501 = 18,871 times (the total n minus the number of cases in the modal category). If, however, we know the sex and parental status of the household head (the categories of the independent variable) we can see that our prediction would change. If the head is a male, either with or without children, we would still predict married and we would be right 25,776 + 19,214 = 44,990 times. If the head is female and has children under 18 we would predict divorced; if the head is female and does not have children under 18 we would predict widowed. We would be right here 1,135 + 6,457 = 7,592 times. Altogether, with this knowledge of the independent variable we are right 44,990 + 7,592 = 52,582 times and wrong 64,372 - 52,582 = 11,790 times. (The correct predictions we would make in the dependent variable once we know the categories of the independent variable are underlined in Table 4-7.)

Now, using our PRE format, we can see that we have reduced our error from 18,871 to 11,790 or a reduction of 7,801 cases. This is a reduction of 7,081/18,871 = .375 = lambda. Our reduction in error in predicting categories of the dependent variable once we know the independent variable is .375. This measure is lambda. We have reduced our error in predicting marital status by 38% once we know the type of household involved.

Table 4-7
Relationship of Type of Household and Marital Status of Household Head

Marital Status of Household Head	Type of Household (sex of head and presence of children under 18) Male Head chil.<18	Male Head no chil. <18	Female Head chil.<18	Female no chil. <18	Totals
Married	25,776	19,214	313	198	45,501
Separated	79	502	998	425	2,004
Divorced	74	946	1,135	1,105	3,260
Widowed	181	1,199	942	6,457	8,779
Single	64	2,302	349	2,113	4,828
Totals	26,174	24,163	3,737	10,298	64,372

In general, values of lambda range from 0 to 1.00. A value of 0 indicates that the independent variable does not help at all in gaining knowledge of the dependent variable; a value of 1.0 indicates that perfect prediction of the dependent variable occurs when the independent variable is known. The sampling distribution of lambda is known and computer programs typically give both the value of lambda and its associated level of significance.

It is possible to compute both partial and multiple measures of association such as lambda. The partial measure is simply a weighted average of the measure within each category of the control variable. The multiple measure is that obtained when all the categories of the independent variable are combined, as could be obtained in the above example if a variable such as race were added and there were categories across the table for each combination of the various attributes of the variables of race and type of household.

In general, lambda is easy to compute and easy to interpret. It, however, can be affected by uneven marginals and will go to zero quickly with skewed marginals. There are many measures of association available and these can easily and quickly be used in exploratory work.

Packet 120

SOC 412/512

SOCIOLOGICAL RESEARCH METHODS

Professor Stockard

University of Oregon

Winter Term 1992

UP 756

kinko's

the copy center

860 E. 13th

Eugene • 344-7894

Copies:	$4.54
Binding	$0.00
Royalties	$0.00
Permission Handling Charges	$0.00
Total cost of packet:	$4.54

TABLE OF CONTENTS

Jean Stockard - Packet 120

V. Bivariate Inferential Statistics

In this section we examine statistical procedures that allow us to make inferences about two variables. These involve the t-distribution, mentioned briefly earlier. Thus, we first discuss the nature of the t-distribution and give an example of its use in testing a hypothesis about a single mean. Second, we extend this work to examining differences between "dependent samples." This involves either 1) the case where we want to infer from the differences within a sample between scores on one variable and scores on another variable (for example whether students have higher math SAT or verbal SAT scores) to the population or 2) the case where we want to examine the differences between related groups (say brothers and sisters, husbands and wives, matched pairs) on a particular variable (for example whether wives or husbands contribute more hours per week to household chores) and infer from the sample to the population. Third, we move to examining differences in averages between two independent groups, for instance, looking at whether men or women (unrelated to each other) earn more money on the average when employed in full time teaching positions at the University. We want to test the hypothesis that there is no difference in these averages in the population. Fourth, we briefly describe extensions of this work for developing confidence intervals around differences between means and mean differences. Fifth, we describe the computer procedures that are used in dealing with the second and third areas of inquiry. Then we describe a descriptive statistic that can be used with tests of differences between means, and, finally, we review the nature of inferential statistics and provide cautions about their use.

The t-distribution

We noted briefly above that there are many different sampling distributions. For instance, the chi-square distribution is used in testing hypotheses about frequencies. F-distributions are used to test hypotheses about variances. The t-distribution is used to test hypotheses about means when sample sizes are small. The t-distribution is used in single sample tests such as those discussed above when the sample size is around 125 or less and also in cases where the means from two samples are compared. When the sample size is large (over 150) the t-distribution approaches the shape of the normal distribution and the normal distribution may be used.

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 (as Table 3-1 does for the normal distribution). Instead, the table (Table 5-1) 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.

$$t = (\bar{X} - \mu)/s_{\bar{X}} \quad \text{where } s_{\bar{X}} = \frac{s}{\sqrt{n}} \; ; \; s = \sqrt{\frac{\Sigma (X-\bar{X})^2}{n-1}} \qquad (5\text{-}1)$$

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

$s_{\bar{X}} = 5/\sqrt{25} = 5/5 = 1.0$.

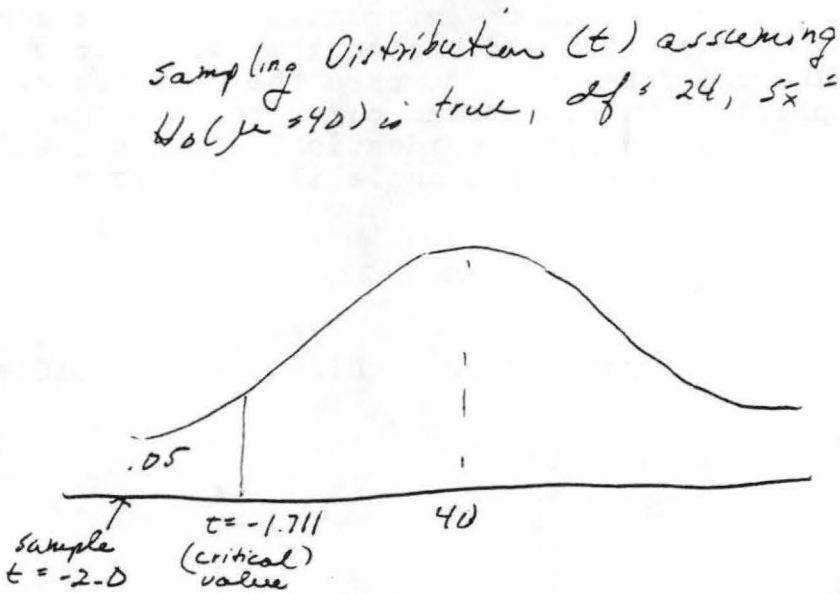
Say we had chosen a significance level of .05.

We turn now to the t-distribution in Table 5-1. 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 Table 5-1, 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. $t = (\bar{X} - \mu)/s_{\bar{X}} = (38-40)/1 = -2.0$. 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 Table 5-1 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 5-1



T-Tests with Matched or Dependent Samples

Sometimes you will want to compare scores on two variables, but the samples will not be independent of each other. Perhaps you have scores on two tests for the same people, say verbal and math scores for boys, and you want to compare their performance on the same subject. Or suppose you had information from brothers and sisters on the same variable and you wanted to compare their responses. In each of these cases the samples are not independent. The brothers and sisters grew up in the same family, they are related to each other. Obviously the same people took the verbal and math tests. In other instances, researchers may specifically try to match respondents on socio-economic or demographic characteristics. All of these are examples of matched or related samples.

When we want to compare the scores on two variables in cases such as this our procedure is a simple extension of the one-sample tests discussed above. This is because, in essence, we are dealing with one sample. Instead, however, of using the actual scores we use <u>difference</u> scores. For instance, suppose we were interested in boys' scores on math and verbal tests. If the tests were standardized in some way so that the scores were directly comparable (say on a base of 100) then we could look at the difference of each boy's math and verbal scores. We would then be interested in the average of these differences (the mean difference).

Our null hypothesis H_0 would be $H_0 : \mu_0 = 0$

and our alternative hypothesis could be $H_1 : \mu_D > 0$ $(X_D = X_m - X_v)$

We chose a directional hypothesis here because some earlier work has suggested that boys have more success with mathematical than with verbal problems. Based on information from the sample we can estimate the sampling distribution that would occur if the null hypothesis were true, then compare the data for our sample with this sampling distribution and carry through the implications. The logic here is identical to that used in the other tests of hypothesis. An example illustrates this.

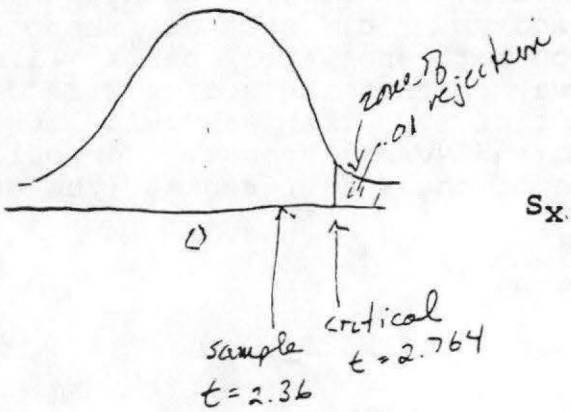
Figure 5-2

Boy	Math Score	Verbal Score	(M-V) Difference	(Difference)^2
A	91	89	2	4
B	85	80	5	25
C	91	95	-4	~~26~~ 16
D	97	98	-1	1
E	73	70	3	9
F	72	71	1	1
G	71	68	3	9
H	76	72	4	16
I	75	74	1	1
J	97	95	2	4
K	84	80	4	16
			20	102

$$\bar{X}_D = \sum(M-V)/N = 20/11 = 1.82$$

$$SD = \frac{\sum (X_D - \bar{X}_D)^2}{n-1} = \sqrt{\frac{\sum X_D^2}{N} - \left(\frac{X_D}{N}\right)^2}\sqrt{\frac{N}{N-1}} = \sqrt{\frac{102}{11} - \left(\frac{20}{11}\right)^2}\sqrt{\frac{11}{10}}$$

$$= \sqrt{9.27 - 3.31}\sqrt{1.10} = \sqrt{(5.96)}\sqrt{1.10} = (2.44)(1.05) = \text{~~2.56~~ } 2.35$$



Theoretical sampling distribution assuming Ho is true

df = n-1 = 11-1 = 10

$$S_{\bar{X}} = \sqrt{\frac{S_D^2}{N}} = \frac{2.56}{\sqrt{11}} = \frac{2.56}{3.32} = \text{~~.77~~ } .71$$

In this case our degrees of freedom = n-1 (there is only one mean to compute so we only lose one degree of freedom) and n-1 = 11-1 = 10. If we want a .01 level of significance, we can see by looking at the t-distribution table (5-1) that the critical value will be a 2.764. Now, given our standard error (.77) and sample mean difference (1.82) we can compute the t-value for our sample:

$$t = \frac{\bar{X} - \mu}{S\bar{x}_D} = \frac{1.82 - 0}{.77} = 2.36 \qquad (5\text{-}2)$$

This value is not high enough to reject the null hypothesis at the .01 level of significance. However, we can see in the table that it is large enough to reject the null hypothesis in favor of the alternative at the .025 level of significance. In other words, we can reject the null hypothesis that boys will do equally well in math and verbal tests in favor of the hypothesis that they will do better in math tests and be wrong less than 2.5 times out of one hundred but wrong more than one time out of one hundred. The researcher would then discuss the substantive and theoretical reasons that would support and explain this conclusion.

Another short example can illustrate this procedure. Suppose that we had asked a group of young women to respond to the following statement: "My father is concerned that I make moral decisions," and the subjects indicated the extent to which they believed that statement was true on a Likert-type scale varying from one to four. The question was also repeated for the same subjects with respect to their mothers. Figure 5-3 shows the results for a set of 12 subjects. Suppose we wanted to test the possibility that the subjects saw their fathers as being more concerned about their behavior than their mothers were. We would then have the following hypothesis:

$H_o = \mu_D = 0$ where $X_D = X_{FA} - X_{mo}$

$H_1 = \mu_D > 0$

The sampling distribution that would occur if H_o were true is also shown in Figure 5-3. The estimate of the standard error is computed from the sample information. In this case df = n-1 = 12-1 = 11. For a .05 level of significance for our directional hypothesis, the critical value of t equals 1.796. The computation of the t-value for this sample is also shown in Figure 5-3. Here t = .89. This does not fall in the zone of rejection and we cannot reject the null hypothesis in favor of the alternative, even at the .10 level of significance. On the basis of this sample we cannot conclude that young women rate their mothers and fathers differently in their concerns about their daughters' behavior.

Figure 5-3

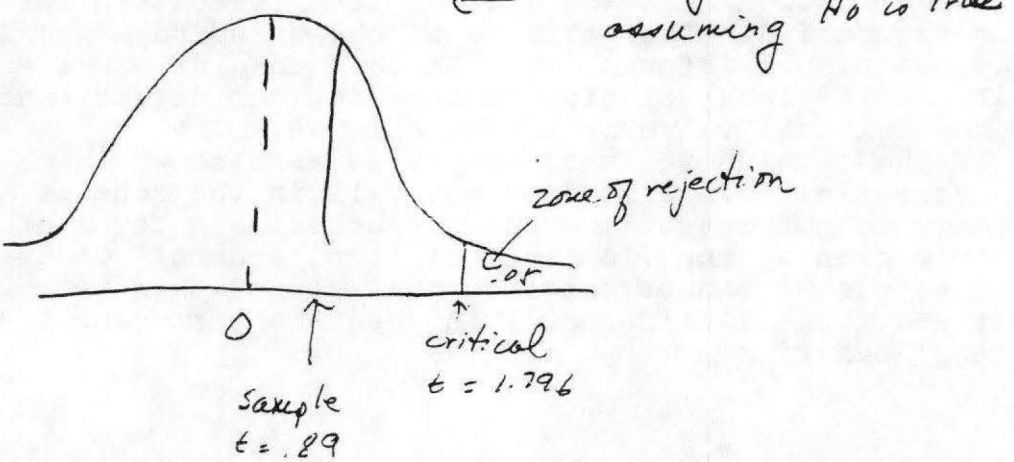
Person	Father Score	Mother Score	X_D Fa-Mo Difference	(Difference)^2
A	3	2	1	1
B	1	2	-1	1
C	2	2	0	0
D	3	2	1	1
E	3	1	2	4
F	4	3	1	1
G	4	2	2	4
H	4	3	1	1
I	2	4	-2	4
J	2	3	-1	1
K	1	3	-2	4
L	4	1	3	9
			5	31

$$\bar{X}_0 = \frac{\Sigma(Fa-Mo)}{N} = \frac{5}{12} = .42$$

$$s_0 = \sqrt{\frac{\Sigma X_0^2}{n} - \left(\frac{X_0}{N}\right)^2}\sqrt{\frac{N}{N-1}} = \sqrt{\frac{31}{12} - \left(\frac{5}{12}\right)^2}\sqrt{\frac{12}{11}} =$$

$$\sqrt{2.58 - .18}\sqrt{1.09} = \sqrt{2.40}\sqrt{1.09} = 1.61$$

$$s_{\bar{X}_0} = \frac{s_0}{\sqrt{n}} = \frac{1.61}{\sqrt{12}} = .47 \qquad t = \frac{\bar{X}_0 - \mu_0}{s_{\bar{X}_0}} = \frac{.42-0}{.47} = .89$$



It is important to carefully watch for situations when you have related samples rather than independent samples. This often occurs when researchers try to match subjects on various variables or when you are concerned with variables regarding related items. Note that when matching is involved problems often occur in deciding to what population the results may be generalized. Computer programs usually automatically take into account if matched samples are involved simply from the way data must be fed in. This, however, does not always occur. In any case the researcher should always think carefully about the nature of the sample and problems involved in planning the analysis.

Two-Sample Tests: Differences of Means

Quite often a researcher will have two independent samples and will want to compare the average scores of people in these two samples on some variable. For instance, you might want to compare the average income of blacks and whites or the average occupational prestige of men and women workers. The t-distribution is used in testing hypotheses in situations such as this.

Let us briefly review the logic involved in the two techniques of inference before moving on. In developing confidence intervals one uses the sample values as the best estimate of the population parameters and then uses this information to develop a range of values in which the population parameters likely fall. How likely it is that they will fall in this range may be specified by the researcher. In hypothesis testing, the null hypothesis is usually the hypothesis of no difference. The researcher assumes that the null hypothesis is true and then follows out the implications of what would happen if that were so by looking at information from the sample. In both techniques the sampling distribution is the basis of the logic and conclusions. The standard error is the standard deviation of the sampling distribution and is estimated from information about the standard deviation and size of the sample. In general, as sample sizes get larger and standard deviations get smaller, the standard error gets smaller and confidence intervals become smaller and the null hypothesis is easier to reject.

Even though it does not make strict sense to use levels of significance when talking about total populations, Blalock notes that they are sometimes used to determine the probability that some event would occur by chance. In other words, we find out what the probability is that the findings are chance ones within some hypothetical larger population.

Suppose you were interested in the math ability of eight year old males and females. Your null hypothesis is that the males and females are equal in ability.

H_o: $\mu_m = \mu_f$ or $\mu_m - \mu_f = 0$

The alternative or research hypothesis (based on your earlier study in the area) was that the males would do better than the females.

H_1: $\mu_m > \mu_f$ or $\mu_m - \mu_f > 0$

To test this hypothesis we need a sampling distribution of the difference between the means, $\mu_m - \mu_f$. This sampling distribution is known and in fact it is the t-distribution. We must assume that both the sample of males and the sample of females have been chosen from their respective populations randomly and that they are not related to each other (e.g., aren't brothers and sisters, or don't sit next to each other in some matched way.) This sampling distribution is the theoretical distribution of all the differences in means that would be computed by drawing successive samples of males and females, computing the differences between the means, and plotting them. The mean of this sampling distribution will be $\mu_m - \mu_F$

and the standard error will be $\sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_F^2}{n_F}} = \sigma_{\bar{X}_m - \bar{X}_F}$.

Note that this standard error for the difference between the means is simply the square root of the sum of the individual standard errors for each mean.

Although with small samples the theory behind the development requires that the populations from which the samples are drawn be normally distributed on the characteristic tested, this requirement lessens as the sample sizes get larger. Also, as the sample sizes get larger, the t-distribution approaches the shape of the normal distribution. It is important that both of the samples be randomly selected, so that the inferences will be correct. It is also important that the samples be independent of each other, so that people in one group are not related in some way to specific individuals in the other group.

Note that the standard error of the sampling distribution of differences between the means $\sigma_{\bar{X}_m - \bar{X}_F}$ is larger than either $\sigma_{\bar{X}_m}$ or $\sigma_{\bar{X}_F}$. This occurs simply because the chances of sampling error are larger when two samples are involved. The standard error reflects this chance of error occurring.

Just as with the single sample tests, we rarely will know the population values of the variance, σ_m or σ_F. Thus, we must get estimates of s_m, s_F and $s_{\bar{x}_m - \bar{x}_F}$. There are two possible ways of getting this estimate.

The two ways of getting the estimate of the standard error of the difference between the means depend on whether or not we can assume that the two groups being compared have equal variances on the dependent variable. In our particular example, we want to know if the variance of math scores is the same for the males as for the females. To examine this, we look at the ratio of one variance to the other, always placing the smaller variance in the denominator. If this ratio is substantially greater than one, we would suspect that the two variances are not equal. If, however, the ratio is close to one we are relatively safe in concluding that the two variances are equal. In fact, the ratio of the variances is a particular statistic called the F-ratio, which has its own sampling distribution (or family of sampling distributions depending on the sizes of samples involved). For this particular level of statistics when doing hand computations it will be sufficient to simply inspect the ratio visually. When using computer output the exact F ratio is given along with a probability level that indicates our chances of error in rejecting a null hypothesis that the two variances are equal.

If we can assume that the variances are equal we may use what is called the "pooled variance estimate" of the standard error of the difference between the means. This is computed by the relatively complex procedure described below.

If we assume that the variance of the two samples are equal:

$$\sigma_m^2 = \sigma_F^2 = \sigma^2$$

In this case $\sigma_{\bar{x}_m - \bar{x}_F} = \sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_F^2}{n_F}} = \sqrt{\frac{\sigma^2}{n_m} + \frac{\sigma^2}{n_F}} = \sigma\sqrt{\frac{1}{n_m} + \frac{1}{n_F}}$

$$= \sigma\sqrt{\frac{n_F + n_m}{n_m n_F}}$$

and we need only estimate σ.

We can do this by what is called a "pooled variance" estimate. This is a weighted average of the sample variances. Defining

$$\tilde{\sigma}^2 = s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}, \quad \tilde{\sigma}^2 = s^2 = \frac{s_m^2(n_m-1) + s_F^2(n_F-1)}{n_m+n_F-2}$$

$$s = \sqrt{\frac{s_m^2(n_m-1) + s_F^2(n_F-1)}{n_m+n_F-2}}$$

$n_1 + n_2 = 2$ is the degrees of freedom associated with $\bar{X}_1 - \bar{X}_2$. In this case we lose two degrees of freedom because two means are involved.

We now substitute this value back into the formula for the standard error. The first term is the pooled estimate of the population's standard deviation. The second term is the factor needed to make the standard error.

$$s_{\bar{X}_m - \bar{X}_F} = \sqrt{\frac{s_m^2(n_m-1) + s_F^2(n_F-1)}{n_m+n_F-2}} \sqrt{\frac{n_m+n_F}{n_m n_F}} \qquad (5-3)$$

If we cannot assume that the two samples come from populations with equal variances we cannot simplify and use the pooled estimate as above. Here it is necessary to estimate separately. The familiar formulas are used

$$s_{\bar{X}_m}^2 = \frac{s_m^2}{n_m}, \quad s_{\bar{X}_F}^2 = \frac{s_F^2}{n_F}, \quad \& \quad s_{\bar{X}_m - \bar{X}_F} = \sqrt{\frac{s_m^2}{n_m} + \frac{s_F^2}{n_F}} \qquad (5-4)$$

This estimate is usually larger than the estimate when we assume that the population variances are equal. It is slightly less efficient and more subject to sampling error.

Now back to an actual example: We hypothesized earlier that with a group of eight year old students a sample of males would score higher on a test of mathematical ability than a sample of females. This was our substantive (or research or alternative) hypothesis. Our null hypothesis was that they would score equally well.

H_0: $\mu_m - \mu_F = 0$
H_1: $\mu_m - \mu_F > 0$

Suppose that we had selected two independent random samples of males and females, gave them a test of mathematical ability and got the following results:

$\bar{X}_m = 76.5$ $\quad$ $\bar{X}_f = 75.0$

$s^2_m = 105.0$ $\quad$ $s^2_2 = 120$ $\quad$ where $s^2 = \sum(X-X)^2/n-1$

$n_m = 100$ $\quad$ $n_f = 100$

From this information we can get an estimate of the standard error of the sampling distribution of the difference between the mean scores of males and females and then draw the estimated sampling distribution when we assume that H_o is true.

To decide which estimate of the standard error to use we must first examine the ratio of the two variances:

$$s^2_2/s^2_1 = 120/105 = 1.14$$

This is enough larger than 1.00 to raise some doubt in our minds as to whether the variances are truly equal, so we will take the more conservative step of using the separate variance estimate of the standard error of the differences of the means.

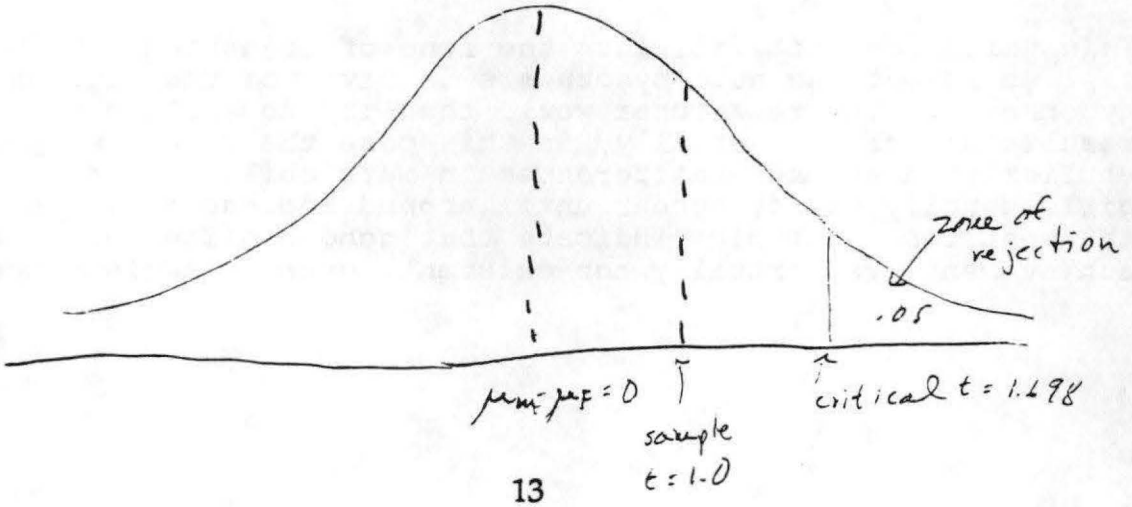
This estimate is

$$s_{\bar{x}_m - \bar{x}_F} = \sqrt{\frac{s^2_m}{n_m} + \frac{s^2_F}{n_F}} = \sqrt{\frac{105}{100} + \frac{120}{100}} = \sqrt{\frac{225}{100}} = \frac{15}{10} = 1.5$$

and the theoretical sampling distribution of the difference between means of the two samples if we assume that H_o is true is shown in Figure 5-4.

Figure 5-4

Theoretical sampling distribution of the difference between the means of the two samples assuming that H_o is true.



Suppose we choose a level of significance of .05 (saying that it is okay with us if we are wrong in rejecting H_o in favor of H_1 5 or less times out of one hundred). We know that we have a one-tailed or directional alternative hypothesis, and we know that our degrees of freedom equal $n_m + n_f - 2 = 200 - 2 = 198$.

We can look at Table 5-1 and see that the zone of rejection begins when the t-score for our samples is greater than 1.645. That is, the critical value of t is 1.645. If the difference in the two means in our sample is more than 1.645 standard t-scores from the hypothesized mean of zero, these results would occur only 5 or less times out of a hundred and we can feel that confident in rejecting the null hypothesis in favor of the alternative that males scored higher than females. Note here how important it is to keep the order of the variables involved straight.

Now, how do we find the t-value for our sample? Remember that t is defined as $(\bar{X} - \mu)/s_{\bar{x}}$ where μ is the mean of the sampling distribution that is being used and $\bar{X}$ is the sample mean value, and $s_{\bar{x}}$ is the estimated standard error of the sampling distribution.

$$t = [(\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)]/ s_{\bar{x}_1 - \bar{x}_2} \qquad (5\text{-}5)$$

where $(\bar{X}_1 - \bar{X}_2)$ is the sample value, $(\mu_1 - \mu_2) = 0$ is the mean of the sampling distribution and $s_{\bar{x}_1 - \bar{x}_2}$ is the estimated standard error.

Therefore

$$t = (\bar{X}_1 - \bar{X}_2)/ s_{\bar{x}1 - \bar{x}2} \qquad (5\text{-}6)$$

when H_o is $\mu_1 - \mu_2 = 0$

In this case $t = \dfrac{(\bar{X}_m - \bar{X}_f) - (\mu_m - \mu_F)}{s_{\bar{x}_1 - \bar{x}_2}} = \dfrac{(76.5 - 75.0)}{1.5} = \dfrac{1.5}{1.5} = 1.0$

This value does not fall into the zone of rejection. We must fail to reject the null hypothesis in favor of the alternative hypothesis. The researcher would then try to explain why these results occurred. Actually, in this case the result supports earlier studies where differences in math ability in boys and girls usually do not appear until around adolescence. In fact, the most recent studies indicate that gender differences in math achievement are virtually non-existent, even at adolescence.

Because our samples are large, the degrees of freedom here equal $n_m + n_f - 2 = (100 + 100) - 2 = 198$. When the degrees of freedoms are so large the t-distribution is the same as the normal distribution. We can then use the table of values on the normal curve to find what the actual probability of being wrong in rejecting H_o in favor of H_1 would be. When we examine this table we see that with $z = 1.00$, the proportion of area under the curve from the mean to that point is .3413. This means that beyond that point there is .1587 of the area. This tells us that if we had rejected H_o in favor of H_1 we would likely be wrong about 16 times out of 100 in doing so.

If our sample sizes had been quite small or if the samples had been quite different in size and we had to reject the hypothesis that the variances were equal our estimate of the degrees of freedom by the formula $n_1 + n_2 - 2 = df$ may be too large.

There is a formula that should be used to estimate the degrees of freedom in this case. The SPSS computer program automatically uses this formula, and it is included in Blalock's text.

Let's do one more example. In this case suppose we had asked young males and females to respond on a survey to the following item which had a Likert-type scale as the possible response: "When I was in high school above all my father wanted me to be happy." We called group one the females and we called the males group 2. Based on our earlier understandings of father-daughter and father-son relationships we expected that daughters would see their fathers as wanting them to be happy more than the sons would. The young men and women were not related to each other so we could assume that the two samples were independent. Our two hypotheses then were

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 > 0$

higher scores means greater wanting of happiness

and our data from our sample were

Group 1, females	Group 2, males
$\bar{x} = 3.2$	$\bar{X} = 2.5$
$s = 1.1$ $s^2 = 1.21$	$s = 1.1$ $s^2 = 1.21$
$n = 65$	$m = 60$

Because the two standard deviations were identical, there was no reason to assume that the population variances were unequal and we can use the pooled variance estimate of the standard error of the sampling distribution. This estimated standard error is computed below.

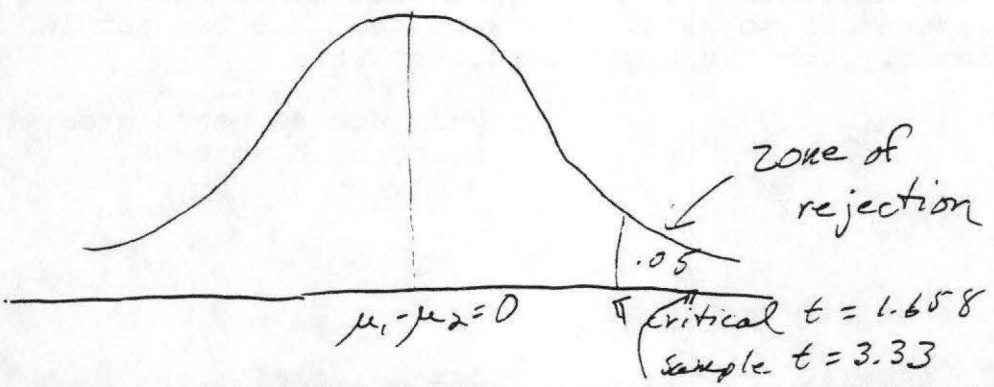
$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}} \sqrt{\frac{n_1+n_2}{n_1 n_2}}$$

$$= \sqrt{\frac{(1.21)(64) + (1.21)(59)}{65 + 60 - 2}} \sqrt{\frac{65+60}{(65 \times 60)}} = \sqrt{\frac{77.44 + 71.39}{123}} \sqrt{\frac{125}{3900}}$$

$$= (1.21 \times .17) = .21$$

In this case our degrees of freedom = 123 = $n_1 + n_2 - 2$. We lose two degrees of freedom here because there are two means involved, not just one. If we choose a .05 level of significance and since we have a one-tailed or directional hypothesis, we can look at Table 5-1 and see that the critical value of t, the value that marks the zone of rejection is 1.658. (We use 120 degrees of freedom since it is the closest smaller degrees of freedom in the table. We generally take the smaller df to have a more conservative decision.) Figure 5-5 shows the sampling distribution that we would have assuming that the null hypothesis is true.

Figure 5-5
Sampling distribution assuming H_o is true, df = 123



Now we must see where the value for our sample lies. We have an estimate of the standard error of the sampling distribution (.21) and we can use this in computing the t-value.

$$t = \left[(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\right] / s_{\bar{x}_1 - \bar{x}_2} = \frac{(3.2-2.5)-0}{.21} = \frac{.7}{.21} = 3.33$$

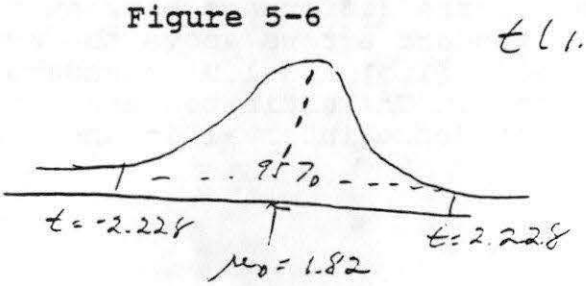
It is obvious that the t-value for our sample falls well within the zone of rejection. In fact, we can compare this value to the t-values for df = 120 in the Table 5-1 and see that the probability of getting a t-value of this size if H_o were really true is less than .005 but not as small as .0005 ($.0005 < p < .005$).

Extensions of the t-test

The elements which are computed in a t-test allow one to also place confidence intervals around the difference between the means or the average difference. The procedure is simply analagous to that used with a single sample estimate. For a confidence interval around the difference between two means, you use the difference between the two sample means as your best estimate of the difference between the population means and the standard error of the difference of the means as your estimate of the standard error. For a confidence interval around the mean difference, you use the mean difference as the estimate of the population mean difference and the standard error of the average difference as the estimate of the standard error. You can then use these elements as the basis of your estimated sampling distribution for the difference between the means. This is shaped like the t-distribution when you have small degrees of freedom and like the normal distribution when the degrees of freedom are sufficiently large (greater than 120 or so).

To illustrate this process I will use data from the examples given above. The first example used in illustrating a test of the hypothesis that the average difference between two related scores equals zero involved boys' scores on math and verbal tests. The calculations given there showed that there was an average difference of 1.82 and a standard error of the mean difference of .77. Given the sample size of 11, the degrees of freedom equal 10. We may use this mean of 1.82 as our best estimate of the population mean and the mean of the sampling distribution of the average differences. The table of the t-distribution indicates that the t-value which corresponds to the .05 level for a two-tail test of significance and 10 degrees of freedom is 2.228. This means that 95% of the area in the t-distribution for 10 degrees of freedom lies between 2.228 standard errors below the mean and 2.228 standard errors above the mean. Thus, to encompass 95% of our estimates of the average difference, we need to go 2.228 standard errors below 1.82 and 2.228 standard errors above 1.82. The estimated sampling distribution is shown below.

Figure 5-6



To compute the 95% confidence interval around the average difference we would need the following simple calculations:

$$P\,[1.82 - (2.228)(.77) < \mu_D < 1.82 + (2.228)(.77)] = .95$$

$$P\,[1.82 - 1.72 < \mu_D < 1.82 + 1.72] = .95$$

$$P\,[.10 < \mu_D < 3.54] = .95$$

We can be 95% confident that in the population the average difference between the boys' math and verbal scores falls between .10 and 3.54.

The t-table indicates that for 10 degrees of freedom 99% of the cases would fall between 3.169 standard errors above the mean and 3.169 standard errors below the mean. With this knowledge one can compute the 99% confidence interval around the average difference as below.

$$P\,[1.82 - (3.169)(.77) < \mu_D < 1.82 + (3.169)(.77)] = .99$$

$$P\,[1.82 - 2.44 < \mu_D < 1.82 + 2.44] = .99$$

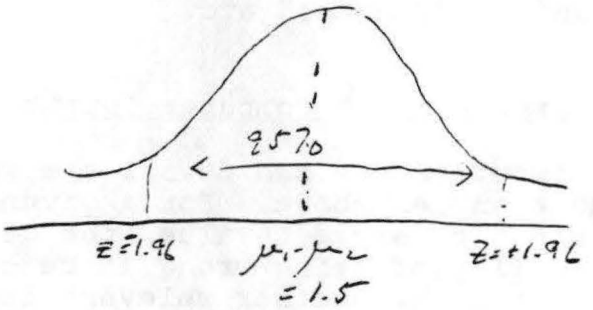
$$P\,[-.62 < \mu_D < 4.26]$$

Thus, we can be 99 percent confident that in the population the average difference between the boys' verbal and math scores mean score of the boys and the mean score of the girls falls between −.62 and 4.26. Note that while the 95% confidence interval does not include zero, the 99% confidence interval does. This corresponds to our decision to reject the null hypothesis at the .05 level, but not at the .01 level with a two-tail test.

This logic can be extended to placing confidence intervals around differences between means. The first example of testing the null hypothesis regarding differences between two means involved examining the differences in math scores of 8 year old boys and girls. The average score for the boys was 76.5 and the mean for the girls was 75.0. The estimate of the standard error of the difference between the means was 1.5, and the degrees of freedom were 198, sufficiently large to use the normal curve table rather than the t-table.

We can use the difference between the sample means of 76.5-75.0=1.5 as our best estimate of the difference between the means of the two populations. Because we can use the normal curve table as our guide, we know that to encompass 95% of the estimated values of the difference between these two means we need to go 1.96 standard errors above the estimated difference of the population means (1.5) and 1.96 standard errors below this estimated difference. The estimated sampling distribution used to compute the confidence interval is shown below.

Figure 5-7



To calculate the 95% confidence interval around the difference between these two means we would have

$$P\,[1.5 - (1.96)(1.5) < \mu_1 - \mu_2 < 1.5 + (1.96)(1.5)] = .95$$

$$P\,[1.5 - 2.9 < \mu_1 - \mu_2 < 1.5 + 2.9] = .95$$

$$P\,[-1.4 < \mu_1 - \mu_2 < 4.4] = .95$$

The 99% confidence interval around the differences between these means would be calculated as below.

$$P\,[1.5 - (2.58)(1.5) < \mu_1 - \mu_2 < 1.5 + (2.58)(1.5)] = .99$$

$$P\,[1.5 - 3.9 < \mu_1 - \mu_2 < 1.5 + 3.9] = .99$$

$$P\,[-2.4 < \mu_1 - \mu_2 < 5.4] = .99$$

Thus, we can be 95 percent confident that in the population the difference between the mean score of the boys and the mean score of the girls falls between -1.4 and +4.4 and 99% confident that it falls between -2.4 and +5.4. Note that this confidence interval includes zero, thus corresponding to our decision to fail to reject the null hypothesis that $\mu_1 - \mu_2 = 0$ at either the .05 or .01 level of significance in favor of a two-tail alternative.

It should also be noted that testing hypotheses about the differences between proportions is a special case of the difference between the means. This situation often occurs when you are dealing with dichotomies, as in investigations of voting preferences.

The computer will automatically treat dichotomies as a special extension of t-tests. When, however, you are doing hand computations you will want to think through the logic and make sure your formulas are correct. You would use the estimate of the standard deviation to get an estimate of the standard error of the sampling distribution and proceed to test the hypotheses as in the case with an intervally measured dependent variable. Confidence intervals can also be computed.

All of this logic can also be extended to more complex cases such as looking at "differences of differences of means," "differences of mean differences," etc.

Examples with Computer Output

Fortunately, computers now can do almost all the work that we have done by hand in the examples above. For a given problem the computer will compute the sample t-value, the degrees of freedom, and the exact probability of being wrong in rejecting the null hypothesis. It will also give other relevant information such as the mean or average difference scores (when dealing with dependent samples) or the mean of each sub-group (when dealing with independent samples).

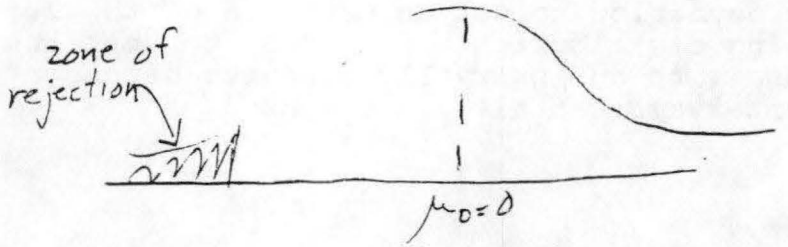
With SPSS you use the T-TEST procedure. The examples below involve data on achievement test scores for high school students in a western Oregon school district. Assume that they are randomly selected. Suppose you were interested in the difference between the composite achievement scores and the reading and language arts scores of the students and suppose also that you wished to examine this difference for male students and for female students respectively. Suppose VAR05 is the variable number for sex, with 1 the code for males and 2 the code for females. VAR15 is the eleventh grade composite achievement test score, VAR16 is the eleventh grade reading achievement score, and VAR17 is the eleventh grade language arts achievement test score.

Further, suppose on the basis of your understanding of the literature that you theorized that the females would have higher reading and language arts achievement test scores than composite test scores. You theorized that the males would probably have reading and language arts test scores that would be about equal to their composite test scores. You would expect then to be able to reject the null hypothesis in favor of the alternative hypotheses for females but to fail to reject the null hypotheses for males. These hypotheses and related theoretical sampling distributions are given below, and a copy of the computer output is on a subsequent page.

For females:

$H_o: \mu_{VAR15} - VAR16 = 0$ $\qquad$ $H_o: \mu_{VAR15} - VAR17 = 0$

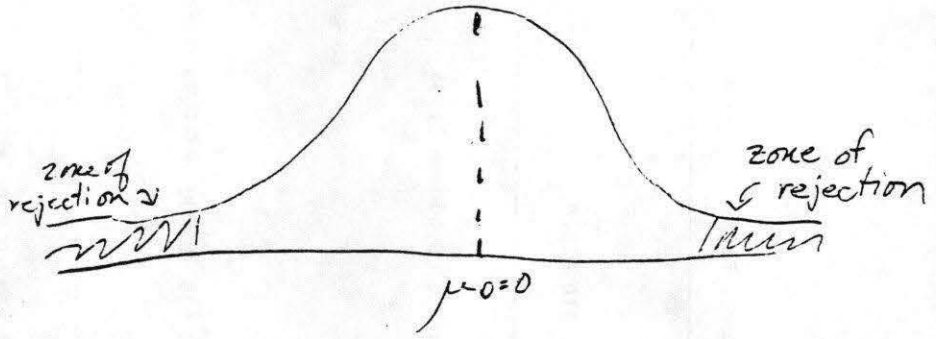
$H_1: \mu_{VAR15} - VAR16 < 0$ $\qquad$ $H_1: \mu_{VAR15} - VAR17 < 0$



For males:

H_0: $\mu_{VAR15} - VAR16 = 0$ H_0: $\mu_{VAR15} - VAR17 = 0$

H_1: $\mu_{VAR15} - VAR16 \neq 0$ H_1: $\mu_{VAR15} - VAR17 \neq 0$



Reading across the output you are first told the variables that are involved in the particular comparison, both by variable name and by label. For the first example they are VAR15 and VAR16, the composite score and the reading score. You are then told the number of subjects or cases involved, 260 for the first example. The mean or average, standard deviation, and standard error for each individual variable are then given. Then, separated by a row of astericks, you are given the mean difference, the average of the differences between the two scores that are being compared. For the first case, the comparison of composite and reading scores for men, this average difference is -2.95, indicating that on the average the men have lower composite scores than reading scores. The standard deviation of this difference is also given, and the standard error (the standard deviation divided by the square root of the sample size). For current purposes ignore the information given in the next two columns and skip over to the last three columns. These give the sample t-value, the degrees of freedom, and the

*SELECT IF (VAR05 EQ 2)

T-TEST PAIRS = VAR 15 WITH VAR16, VAR15 WITH VAR17

The following output was obtained for males

Variable	Number of cases	Mean	Standard Deviation	Standard Error	(Difference) Mean	Standard Deviation	Standard Error	2- Tail Corr.	Prob.	T Value	Degrees of Freedom	2-tail Prob.
VAR15	11th Ited	Composite	29.103	1.805								
		46.3038										
	260	49.2538	27.741	1.720	-2.9500	12.112	0.751	0.910	0.000	-3.93	259	0.000
VAR16	11th Ited	Reading										
VAR15	11th Ited	Composite										
		46.3038	29.103	1.805								
	260				9.8885	14.066	0.872	0.876	0.000	11.34	259	0.000
		36.4154	26.685	1.655								
VAR 17	11th Ited	Langarts										

and for females

Var15	11th Ited	Composite										
		51.7729	25.470	1.608								
	251				-3.1793	10.219	0.645	0.918	0.000	-4.93	250	0.000
		54.9522	24.836	1.568								
VAR16	11th Ited	Reading										
VAR15	11th Ited	Reading										
		51.7729	25.470	1.608								
	251				3.3028	12.595	0.795	0.883	0.000	4.15	250	0.000
		48.4701	26.509	1.673								
VAR17	11th Ited	Langarts										

probability of being wrong in rejecting the null hypothesis. The sample t-value is simply computed with standard formula $t = X - \mu / s_X$ where μ is hypothesized to be zero. Here

$t = -2.95/0.75 = -3.93.$

Our degrees of freedom equal n-1 = 260 -1 = 259, and the probability of being wrong is less than .001 (the computer will usually not print more than three decimal points). Note that this is a two-tailed probability. If we have a one-tail alternative hypothesis (as with the hypotheses for the females) we will need to divide this probability by two). Note that we must reject the null hypothesis that the males have equal reading and composite scores in the population in favor of the alternative hypothesis that they are unequal and be wrong in doing so less than 1 time out of a thousand. This is contrary to what we had theoretically expected.

Students may follow this procedure by examining the results for the comparison of the composite and language arts scores for males and should reach the conclusion that they must reject the null hypothesis in favor of the alternative and be wrong less than one time out of 1000. The average difference here is 9.89, indicating that the males tend to have higher composite scores than language arts scores and that this difference probably occurs within the population.

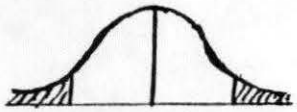
With the females and the difference of composite and reading scores we may reject the null hypothesis in favor of the alternative hypothesis and be wrong less than .0005 times in doing so. With the difference of the composite and language arts scores we may <u>not</u> reject the null hypothesis in favor of the alternative. Students should be able to elaborate on these conclusions and understand why they were reached.

The examples above involved dependent samples. The computer also examines independent samples. Suppose you were interested in comparing the average scores of male and female students and you suspected that the males and females would have equal composite scores, but that in the population the females would have higher language arts achievement scores and the males would have higher math achievement scores. Given the hypotheses listed below (and subtracting the scores of females from those of males) you would expect to fail to reject the null hypothesis regarding composite scores but expect to reject the null hypotheses regarding the language arts and mathematics scores. A copy of the printout is given on the following page.

At the top of the printout it is noted that group one is the code 1, males: group two is the code 2, females. Then the results for each hypothesis are listed: those for the composite scores, then those for the language arts scores, and then those for the math scores. For each variable the number of cases, the average, the standard deviation and the standard error for each

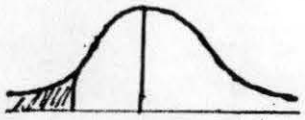
Composite scores $H_o: \mu_m - \mu_F = 0$

$H_1: \mu_m - \mu_F = \neq 0$



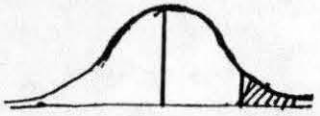
language arts scores $H_o: \mu_m - \mu_F = 0$

$H_o: \mu_m - \mu_F = <0$



Math scores $H_o: \mu_m - \mu_F = 0$

$H_1: \mu_m - \mu_F = >0$



(shaded areas show zones of rejection)

The computer was given the following instructions:

T-TEST GROUPS= VAR05 (1,2)/VARIABLES = VAR15, VAR17, VAR18
and the following results were obtained

GROUP 1 - VAR05 EQ 1.
GROUP 2 - VAR05 EQ 2.

Variable	Number of cases	Mean	Standard Deviation	Standard Error	F Value	F 2-tail Prob.	Pooled variance estimate: T Value	Pooled: Degrees of Freedom	Pooled: 2-tail Prob.	Separate variance estimate: T Value	Separate: Degrees of Freedom	Separate: 2-tail Prob.
VAR 15 11TH ITED COMPOSITE												
GROUP 1	260	46.3038	29.103	1.805	1.31	.034	-2.26	509	.024	-2.26	504.20	0.024
GROUP 2	251	51.7729	25.470	1.608								
VAR17 11TH ITED LANGARTS												
GROUP 1	261	36.4559	26.641	1.649	1.01	.915	-5.11	511	0.000	-5.11	510.59	0.000
GROUP 2	252	48.4365	26.462	1.667								
VAR18 11TH ITED MATH												
GROUP 1	262	53.7137	29.208	1.804	1.34	.020	0.30	514	.762	0.30	507.38	0.761
GROUP 2	254	52.9843	25.240	1.584								

Table 1. Mean Scores and Standard Deviations, all Variables, Females and Males, Fourth Graders and Adolescents

	Females		Males					
	Mean	(*SD*)	Mean	(*SD*)	*t*	*df*	*p*	*d*[a]
Self-esteem								
Fourth graders (Total group)	2.94	(0.54)	3.07	(0.50)	-3.12	602	<.002	-.25
Fourth graders (Panel study)	3.11	(0.47)	3.32	(0.43)	-1.60	50	.12	-.47
Twelfth graders (Panel study)	3.12	(0.56)	3.26	(0.55)	-.88	50	.38	-.25
Adolescents	2.83	(0.53)	3.05	(0.48)	-6.13	797	<.001	-.43
Self-efficacy								
Fourth graders (Total group)	2.76	(0.45)	3.05	(0.48)	-7.73	598	<.001	-.62
Fourth graders (Panel study)	2.88	(0.38)	3.19	(0.44)	-2.65	50	.01	-.76
Twelfth graders (Panel study)	2.96	(0.50)	3.23	(0.50)	-1.90	50	.06	-.54
Adolescents	3.46	(0.55)	3.75	(0.49)	-7.90	797	<.001	-.56
Relationality								
Fourth graders (Total group)	3.13	(0.54)	2.94	(0.52)	4.20	598	<.001	.36
Fourth graders (Panel study)	3.08	(0.52)	3.07	(0.51)	.11	50	<.91	.02
Twelfth graders (Panel study)	3.31	(0.42)	3.25	(0.66)	.38	50	.70	.11
Adolescents	3.99	(0.47)	3.53	(0.66)	11.17	797	<.001	.80

[a]Cohen's *d* equals the difference between the two means divided by the common standard deviation. Cohen (1977) suggests that effect sizes of .2 could be considered small, those of .5 medium in size, and those of .8 or greater to be large.

Table V. Average Scale Scores of Men and Women in Each Sample[a]

	Sample									
	1972 College		1982 College		1982 High school		1983 Nurses[b]		1984 College	
Scales	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men
Expressiveness										
$\bar{X}$	3.35	3.26	3.45	3.25	3.60	3.33	3.64	3.47	3.49	3.33
s	0.36	0.34	0.36	0.35	0.35	0.37	0.38	0.43	0.41	0.42
N	215	195	152	161	117	125	240	186	250	215
$t(p)$	2.75	(.006)	4.95	(<.001)	5.79	(<.001)	4.30	(<.001)	4.04	(<.001)
Effect size[c]	.26		.56		.75		.42		.38	
Instrumental Industrious										
$\bar{X}$	3.11	3.05	3.16	3.02	3.07	3.03	3.58	3.43	3.16	3.04
s	0.53	0.46	0.48	0.48	0.50	0.44	0.38	0.46	0.55	0.52
N	213	194	154	170	118	126	240	186	249	216
$t(p)$	1.22	(.224)	2.65	(.009)	0.57	(.569)	3.57	(<.001)	2.13	(0.21)
Effect size	.12		.29		.09		.36		.22	
Analytic										
$\bar{X}$	3.04	3.16	3.03	3.17	3.20	3.01	3.32	3.45	3.21	3.24
s	0.55	0.48	0.47	0.45	0.50	0.50	0.50	0.50	0.48	0.49
N	209	192	145	173	113	128	237	186	250	216
$t(p)$	−2.34	(.020)	−2.64	(.009)	2.94	(.003)	−2.51	(.012)	−0.81	(.419)
Effect size	−.23		−.30		.38		−.26		−.06	
Autonomy										
Forceful										
$\bar{X}$	2.84	2.90	2.89	2.93	2.97	3.05	3.02	3.03	2.86	2.89
s	—	—	0.40	0.41	0.44	0.42	0.51	0.50	0.46	0.50
N	196	213	163	167	117	123	237	184	247	208
$t(p)$	—	—	−0.81	(.416)	−1.41	(.161)	−0.09	(.929)	−0.66	(.511)
Effect size			−.09		−.19		−.02		−.06	
Adventurous										
$\bar{X}$	2.90	2.83	3.06	3.05	2.98	3.21	2.88	3.08	3.01	3.04
s	0.70	0.62	0.61	0.58	0.66	0.58	0.86	0.81	0.64	0.68
N	219	195	166	175	121	132	242	189	249	216
$t(p)$	0.24	(.810)	0.02	(.984)	−2.87	(.004)	−2.40	(.017)	−0.45	(.656)
Effect size	.11		.02		−.37		−.24		−.05	

[a]Scores on each scale have been summed and averaged. An average score of 4.0 would mean all respondents had indicated items on the scale were *very true of me*; an average score of 1.0 would mean that all respondents had indicated the items were *very untrue of me*.

[b]In this sample, the expressiveness scale does not include *obliging*; the industrious subscale of instrumental includes *hardworking* instead of *industrious*; the forceful subdimension of the autonomy scale does not include *stern*; and the adventurous subdimension does not include *daring*.

[c]The effect size d was computed with the formula $\bar{X}_1 - \bar{X}_2/S$ where $\bar{X}_1$ is the mean in one sample, $\bar{X}_2$ the mean of the other sample, and S is the within-group standard deviation.

Table I. Mean Scores and Standard Deviations, all Variables, Females and Males, Fourth Graders and Adolescents

	Females		Males					
	Mean	(*SD*)	Mean	(*SD*)	*t*	*df*	*p*	*d*[a]
Self-esteem								
Fourth graders (Total group)	2.94	(0.54)	3.07	(0.50)	-3.12	602	<.002	-.25
Fourth graders (Panel study)	3.11	(0.47)	3.32	(0.43)	-1.60	50	.12	-.47
Twelfth graders (Panel study)	3.12	(0.56)	3.26	(0.55)	-.88	50	.38	-.25
Adolescents	2.83	(0.53)	3.05	(0.48)	-6.13	797	<.001	-.43
Self-efficacy								
Fourth graders (Total group)	2.76	(0.45)	3.05	(0.48)	-7.73	598	<.001	-.62
Fourth graders (Panel study)	2.88	(0.38)	3.19	(0.44)	-2.65	50	.01	-.76
Twelfth graders (Panel study)	2.96	(0.50)	3.23	(0.50)	-1.90	50	.06	-.54
Adolescents	3.46	(0.55)	3.75	(0.49)	-7.90	797	<.001	-.56
Relationality								
Fourth graders (Total group)	3.13	(0.54)	2.94	(0.52)	4.20	598	<.001	.36
Fourth graders (Panel study)	3.08	(0.52)	3.07	(0.51)	.11	50	<.91	.02
Twelfth graders (Panel study)	3.31	(0.42)	3.25	(0.66)	.38	50	.70	.11
Adolescents	3.99	(0.47)	3.53	(0.66)	11.17	797	<.001	.80

[a]Cohen's *d* equals the difference between the two means divided by the common standard deviation. Cohen (1977) suggests that effect sizes of .2 could be considered small, those of .5 medium in size, and those of .8 or greater to be large.

Table V. Average Scale Scores of Men and Women in Each Sample[a]

	Sample									
	1972 College		1982 College		1982 High school		1983 Nurses[b]		1984 College	
Scales	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men
Expressiveness										
$\bar{X}$	3.35	3.26	3.45	3.25	3.60	3.33	3.64	3.47	3.49	3.33
s	0.36	0.34	0.36	0.35	0.35	0.37	0.38	0.43	0.41	0.42
N	215	195	152	161	117	125	240	186	250	215
$t(p)$	2.75	(.006)	4.95	(<.001)	5.79	(<.001)	4.30	(<.001)	4.04	(<.001)
Effect size[c]	.26		.56		.75		.42		.38	
Instrumental										
Industrious										
$\bar{X}$	3.11	3.05	3.16	3.02	3.07	3.03	3.58	3.43	3.16	3.04
s	0.53	0.46	0.48	0.48	0.50	0.44	0.38	0.46	0.55	0.52
N	213	194	154	170	118	126	240	186	249	216
$t(p)$	1.22	(.224)	2.65	(.009)	0.57	(.569)	3.57	(<.001)	2.13	(0.21)
Effect size	.12		.29		.09		.36		.22	
Analytic										
$\bar{X}$	3.04	3.16	3.03	3.17	3.20	3.01	3.32	3.45	3.21	3.24
s	0.55	0.48	0.47	0.45	0.50	0.50	0.50	0.50	0.48	0.49
N	209	192	145	173	113	128	237	186	250	216
t(p)	−2.34	(.020)	−2.64	(.009)	2.94	(.003)	−2.51	(.012)	−0.81	(.419)
Effect size	−.23		−.30		.38		−.26		−.06	
Autonomy										
Forceful										
$\bar{X}$	2.84	2.90	2.89	2.93	2.97	3.05	3.02	3.03	2.86	2.89
s	—	—	0.40	0.41	0.44	0.42	0.51	0.50	0.46	0.50
N	196	213	163	167	117	123	237	184	247	208
$t(p)$	—	—	−0.81	(.416)	−1.41	(.161)	−0.09	(.929)	−0.66	(.511)
Effect size			−.09		−.19		−.02		−.06	
Adventurous										
$\bar{X}$	2.90	2.83	3.06	3.05	2.98	3.21	2.88	3.08	3.01	3.04
s	0.70	0.62	0.61	0.58	0.66	0.58	0.86	0.81	0.64	0.68
N	219	195	166	175	121	132	242	189	249	216
$t(p)$	0.24	(.810)	0.02	(.984)	−2.87	(.004)	−2.40	(.017)	−0.45	(.656)
Effect size	.11		.02		−.37		−.24		−.05	

[a]Scores on each scale have been summed and averaged. An average score of 4.0 would mean all respondents had indicated items on the scale were *very true of me*; an average score of 1.0 would mean that all respondents had indicated the items were *very untrue of me*.

[b]In this sample, the expressiveness scale does not include *obliging*; the industrious subscale of instrumental includes *hardworking* instead of *industrious;* the forceful subdimension of the autonomy scale does not include *stern;* and the adventurous subdimension does not include *daring*.

[c]The effect size d was computed with the formula $\bar{X}_1 - \bar{X}_2/S$ where $\bar{X}_1$ is the mean in one sample, $\bar{X}_2$ the mean of the other sample, and S is the within-group standard deviation.

Table I. Mean Scores and Standard Deviations, all Variables, Females and Males, Fourth Graders and Adolescents

	Females		Males					
	Mean	(*SD*)	Mean	(*SD*)	*t*	*df*	*p*	*d*[a]
Self-esteem								
Fourth graders (Total group)	2.94	(0.54)	3.07	(0.50)	-3.12	602	<.002	-.25
Fourth graders (Panel study)	3.11	(0.47)	3.32	(0.43)	-1.60	50	.12	-.47
Twelfth graders (Panel study)	3.12	(0.56)	3.26	(0.55)	-.88	50	.38	-.25
Adolescents	2.83	(0.53)	3.05	(0.48)	-6.13	797	<.001	-.43
Self-efficacy								
Fourth graders (Total group)	2.76	(0.45)	3.05	(0.48)	-7.73	598	<.001	-.62
Fourth graders (Panel study)	2.88	(0.38)	3.19	(0.44)	-2.65	50	.01	-.76
Twelfth graders (Panel study)	2.96	(0.50)	3.23	(0.50)	-1.90	50	.06	-.54
Adolescents	3.46	(0.55)	3.75	(0.49)	-7.90	797	<.001	-.56
Relationality								
Fourth graders (Total group)	3.13	(0.54)	2.94	(0.52)	4.20	598	<.001	.36
Fourth graders (Panel study)	3.08	(0.52)	3.07	(0.51)	.11	50	<.91	.02
Twelfth graders (Panel study)	3.31	(0.42)	3.25	(0.66)	.38	50	.70	.11
Adolescents	3.99	(0.47)	3.53	(0.66)	11.17	797	<.001	.80

[a]Cohen's *d* equals the difference between the two means divided by the common standard deviation. Cohen (1977) suggests that effect sizes of .2 could be considered small, those of .5 medium in size, and those of .8 or greater to be large.

Table V. Average Scale Scores of Men and Women in Each Sample[a]

	Sample									
	1972 College		1982 College		1982 High school		1983 Nurses[b]		1984 College	
Scales	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men
Expressiveness										
$\bar{X}$	3.35	3.26	3.45	3.25	3.60	3.33	3.64	3.47	3.49	3.33
s	0.36	0.34	0.36	0.35	0.35	0.37	0.38	0.43	0.41	0.42
N	215	195	152	161	117	125	240	186	250	215
$t(p)$	2.75	(.006)	4.95	(<.001)	5.79	(<.001)	4.30	(<.001)	4.04	(<.001)
Effect size[c]	.26		.56		.75		.42		.38	
Instrumental										
Industrious										
$\bar{X}$	3.11	3.05	3.16	3.02	3.07	3.03	3.58	3.43	3.16	3.04
s	0.53	0.46	0.48	0.48	0.50	0.44	0.38	0.46	0.55	0.52
N	213	194	154	170	118	126	240	186	249	216
$t(p)$	1.22	(.224)	2.65	(.009)	0.57	(.569)	3.57	(<.001)	2.13	(0.21)
Effect size	.12		.29		.09		.36		.22	
Analytic										
$\bar{X}$	3.04	3.16	3.03	3.17	3.20	3.01	3.32	3.45	3.21	3.24
s	0.55	0.48	0.47	0.45	0.50	0.50	0.50	0.50	0.48	0.49
N	209	192	145	173	113	128	237	186	250	216
$t(p)$	−2.34	(.020)	−2.64	(.009)	2.94	(.003)	−2.51	(.012)	−0.81	(.419)
Effect size	−.23		−.30		.38		−.26		−.06	
Autonomy										
Forceful										
$\bar{X}$	2.84	2.90	2.89	2.93	2.97	3.05	3.02	3.03	2.86	2.89
s	—	—	0.40	0.41	0.44	0.42	0.51	0.50	0.46	0.50
N	196	213	163	167	117	123	237	184	247	208
$t(p)$	—	—	−0.81	(.416)	−1.41	(.161)	−0.09	(.929)	−0.66	(.511)
Effect size			−.09		−.19		−.02		−.06	
Adventurous										
$\bar{X}$	2.90	2.83	3.06	3.05	2.98	3.21	2.88	3.08	3.01	3.04
s	0.70	0.62	0.61	0.58	0.66	0.58	0.86	0.81	0.64	0.68
N	219	195	166	175	121	132	242	189	249	216
$t(p)$	0.24	(.810)	0.02	(.984)	−2.87	(.004)	−2.40	(.017)	−0.45	(.656)
Effect size	.11		.02		−.37		−.24		−.05	

[a]Scores on each scale have been summed and averaged. An average score of 4.0 would mean all respondents had indicated items on the scale were *very true of me*; an average score of 1.0 would mean that all respondents had indicated the items were *very untrue of me*.

[b]In this sample, the expressiveness scale does not include *obliging*; the industrious subscale of instrumental includes *hardworking* instead of *industrious;* the forceful subdimension of the autonomy scale does not include *stern;* and the adventurous subdimension does not include *daring*.

[c]The effect size d was computed with the formula $\bar{X}_1 - \bar{X}_2/S$ where $\bar{X}_1$ is the mean in one sample, $\bar{X}_2$ the mean of the other sample, and S is the within-group standard deviation.

Assignment Two, Sociology 412
Due February 12, 1992

There are several variables in each of the data sets that divide the sample into two discrete groups. With the GSS and Bank data, examples are sex and race (called minority in the bank data). With the Western Electric data, examples are vital 10 (status at 10 years) and famhxcvr (family history of coronary heart disease). Other variables might be grouped, using the recode command, into two discrete groups. Whenever variables are grouped, however, it is important that a theoretically sound reason be given for the grouping. Students should feel free to consult with the instructor if there are questions about the groups they will use.

A. Choose two dichotomous variables in your data set that you believe are theoretically related to the intervally measured variables you examined in the first assignment. Discuss your theoretical expectations regarding differences between people in each group in each of these variables on both of the intervally measured variables. In other words, treat the intervally measured variables used in the first assignment as dependent variables and the nominally measured variables as independent. What types of differences would you expect to find and why? (For example, what kinds of differences would you expect between men and women?) Be specific and discuss each of the combinations of the dependent and independent variables. Then translate your theoretical expectations into statistical hypotheses. Describe whether your theoretical discussion would lead you to expect to reject or fail to reject each null hypothesis.

B. Use the SPSS program t-test to test these hypotheses. Make sure you have first included any appropriate missing value and recode commands and use instructions like the following.

```
T-TEST  GROUPS = independent variable (a,b)/ VARIABLES =
   dependent variables
```

where "independent variable" refers to the computer name of a given dichotomous variable, "a" and "b" are the two values coded for the dichotomy, and "dependent variables" refers to the intervally measured variables you are studying. You will need a separate t-test command for each independent variable that you use. Please turn in your output with your assignment.

C. Interpret the results which you obtain. Be sure to include both statistical and substantive interpretations, to discuss both inferential and descriptive results (i.e. test and interpret your statistical hypotheses and also compute Cohen's d and interpret it), and relate your discussion to the theoretical expectations developed in Part A. Be specific and detailed, writing in a style appropriate for a research report.

sex group are given. The next two columns are headed F-value and 2-tail probability. These are used in deciding which estimate of the standard error to use. The F is simply the ratio of the two variances. The probability is the chance that we would be wrong in rejecting a null hypothesis that the variances are equal. For our purposes, if this two-tail probability is .10 or less, you should use the separate variance estimate.

The final six columns give estimates of t, the degrees of freedom, and the two-tail probability of being wrong in rejecting the null hypothesis. The first three columns are values based on the pooled variance estimate; the last three columns are based on the separate variance estimate.

With the results for the composite score we should use the separate variance estimate. This is because the F ratio is so large that the probability of being wrong in rejecting H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1.00$ is only .030. We cannot assume the two groups come from populations with similar variances, and so we must use the separate variance estimate of the standard error.

The sample t with this estimate (based on $t = (\bar{X}_1 - \bar{X}_2)/S_{\bar{x}1 - \bar{x}2}$) is -2.26, the degrees of freedom are 504.2, and the probability of being wrong in rejecting the null hypothesis is .024 (assuming we have a two-tailed alternative hypothesis). Thus, we may reject the null hypothesis that males and females have equal composite scores in favor of the alternative that they are unequal and be wrong in doing so only about 2 times out of one-hundred. We can note that the direction of the difference indicates that females have higher scores than males. This result was unexpected.

Students should be able to verify that with the results regarding language arts we would use the pooled variance estimate and that we would reject the null hypothesis in favor of the alternative and be wrong in doing so less than .0005 times out of a hundred. (Because we have a one-tail test we must divide the given probability by two). With the results with the math scores we would use the separate variance estimate, but fail to reject the null hypothesis.

Cohen's D: A Descriptive Statistic

The t-test for the difference between two means is an inferential statistic. Sometimes we are also interested in having a descriptive statistic that can describe the difference between two means. In the paragraphs below I explain why it is important to use both descriptive and inferential statistics; show how Cohen's d, a descriptive statistic appropriate for use with the t-test is computed; and then give examples of its use and briefly discuss its utility in meta-analyses.

Statistical vs. Substantive Significance

Whenever you use an inferential statistic you need to be careful that results that provide statistical significance also produce substantive significance. For instance, a given result may not be a chance occurrence, it may not occur by random; yet, it may be so small that it really doesn't mean much for people's daily lives. The results may be statistically significant, but they may not be substantively or practically significant. This situation can often arise when you have large samples. Large samples are important for good multivariate analyses, yet a large sample helps produce a much smaller standard error and thus a much greater chance that any given result will be statistically significant.

For a while, in the late 1950s through the 1960s and early 1970s, some researchers suggested that we should avoid using tests of significance and rely on descriptive statistics instead. Others argued that inferential tests can provide benchmarks or criteria to use in analyzing data and that they should not be discarded.

It is the latter view that has prevailed, helped along by the development of descriptive statistics that are used in tandem with the inferential results. The descriptive statistics can tell us the magnitude of a result, can describe how extensive it is; the inferential statistic can tell us the probability that a result would occur by chance. Descriptive statistics can be small, even with highly significant inferential results, as can occur when a sample is quite large. On the other hand, inferential results can yield no significance, even though the descriptive statistics are quite large, as can happen when a sample is small. Thus, both types of statistics should be used.

Cohen's d

Cohen's d is a very simple statistic to compute that describes the magnitude of the difference between two means relative to their standard deviation. It is simply computed as

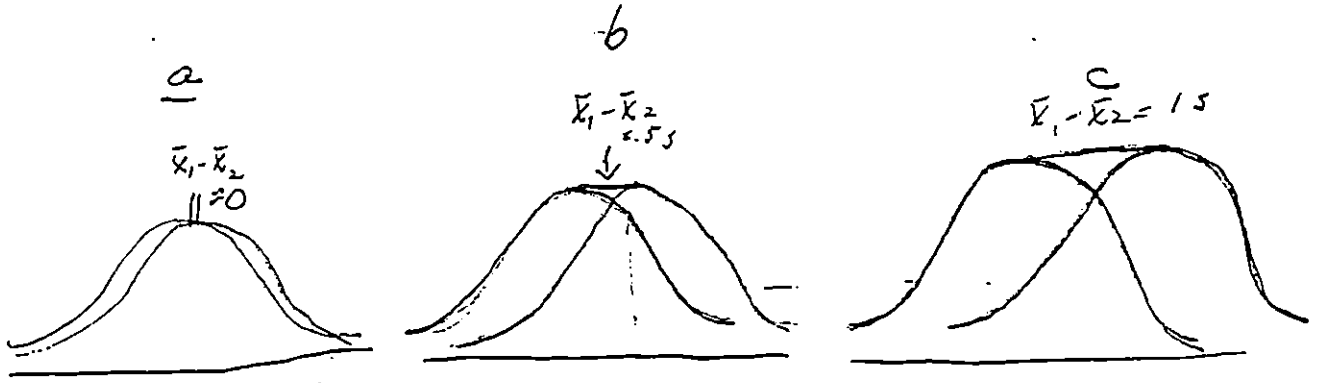
$$d = \frac{\bar{X}_1 - \bar{X}_2}{s} \qquad (5\text{-}7)$$

where $\bar{X}_1$ and $\bar{X}_2$ are the means of the two samples being compared and s is their common standard deviation. (For practical purposes the weighted average of their standard deviations, or the standard deviation of the two combined groups can be used.)

Cohen's d has a very simple interpretation. If it is near zero it simply means that there is no difference between the means (illustrated in part a of Figure 5-8 below). If it equals

one it means that the two averages differ by an entire standard deviation (as shown in part c of Figure 5-8). If it equals .5, the two means differ by one-half of a standard deviation (part b of Figure 5-8). The measure simply provides a handy description of the extent to which the two groups differ, totally independent of their sample size. As a rule of thumb, Cohen has suggested that effect sizes of .2 could be considered small, those of .5 medium in size, and those of .8 or greater to be large.

Figure 5-8
Illustration of Cohen's d Values



For the results given on the computer printout discussed above, the following values of Cohen's d could be computed. Values of t and the associated probability levels are given for comparative purposes.

VAR15 - 11th Grade ITED Composite score

	Mean	Standard Deviation	Cohen's d	t (p)
Males	46.3	29.1	46.3-51.8/27.3 = -5.5/27.3 = -.20	-2.26 (.02)
Females	51.8	25.5		

VAR16 - 11th Grade ITED Language Arts Score

Males	36.5	26.6	36.5-48.4/26.6 = -11.9/26.6 = -.45	-5.11(<.001)
Females	48.4	26.5		

VAR17 - 11th Grade ITED Math Score

Males	53.7	29.2	53.7-53.0/27.2 = 0.7/27.2 = .03	0.30 (.76)
Females	53.0	25.2		

Other Examples

Cohen's d has proved especially useful in comparing results from several samples and with different groups. The following two pages include examples of the use of Cohen's d from articles written by the instructor. Both tables give results of t-tests between the sex groups and the corresponding Cohen's d for different measures and different samples.

The first example (Table V) gives results for 5 different scales (all listed down the left hand side of the table) and 5 different samples (listed across the top of the table). For each scale and sample the mean score, standard deviation, and sample size are given for both men and women (reading down within each sample and within each scale). This is followed by the t and the associated probability level (the probability of being wrong in rejecting H_0: $\mu_1 - \mu_2 = 0$). The last line in each section is the effect size, or Cohen's d. By comparing across the columns within each row one may assess the extent to which similar results are found across the various samples. By comparing between the rows (down the columns) one may assess the extent to which results differ across the different scales. The positive or negative value attached to the d indicates the direction of the difference. (A positive value indicates that women had a higher average score; a negative value indicates that men had a higher average score.)

The second example (Table 1) gives results for four samples and three scales, all listed down the left hand side of the table. Within each row the mean and standard deviation are given for females and males, followed by the t value, the associated degrees of freedom and probability level, and Cohen's d. The results in this table can help illustrate the difference between substantive and statistical significance. Consider the results for self-esteem with the fourth graders (panel study) and twelfth graders (panel study). They are not statistically significant, yet the d for the fourth graders is the highest of all given for that scale and the d for the twelfth graders equals that for the fourth graders (total group), which had a t-value significant at the .002 level. The panel study group was much smaller than the other groups (n=52 vs. n=604 and n=799), thus making it much easier to obtain statistical significance in the latter groups.

In recent years meta-analyses have become much more common, especially in the social psychology literature. Meta-analyses involve the review of a large number of replications of the same study in order to determine general patterns and trends. When these studies involve the comparison of two groups Cohen's d is typically used as a descriptive statistic. Cohen's d is computed for each study in the meta-analysis and these are then averaged to determine the average effect size over a group of studies. Sometimes a researcher is interested in variations in the studies and will compare Cohen's d across various groups.

For instance, a large number of studies of psychological sex differences have been conducted and meta-analyses of these results have become fairly common in recent years. A researcher might compute an overall average Cohen's d for a group of studies (say on gender differences in aggression), and then examine the average d for studies of people in various age groups, with different measures of the dependent variable, with different types of study designs, etc. As you should remember from your methodology coursework, replications are central to the scientific enterprise, and Cohen's d has become a standard element in summarizing and analyzing patterns found in replications.

A Few Cautions and Other Comments

All of the results regarding t-tests are based on assumptions that statisticians used in developing these tests. We assume that the samples are randomly selected from a population that we are interested in. We assume also that the subjects have been selected independently of each other. We also assume that the variables in the populations are normally distributed and that the dependent variable is measured on an interval scale. If you have a representative sample (i.e., a probability sample) that is not a simple random sample (e.g., a cluster sample or a stratified random sample) the estimates of the standard errors will be slightly different. Advanced statistics and sampling texts can help you with these problems.

Researchers have investigated what happens if you do t-tests with samples from populations that are not normally distributed. They conclude that if you have large samples (over 25 for each group is sufficient) you generally will not have problems with faulty inferences if the populations are not normally distributed.

Others have examined what happens if you use variables measured on an ordinal rather than an interval scale, that is, treating ordinally measured variables as though they were intervally measured. They concluded that there are not extreme problems. Inaccuracies are most apt to occur with one-tail tests, unequal n's in the two groups and badly skewed populations.

VI. Analysis of Variance

With t-tests we are able to compare the means of two groups. What, however, if one wants to compare the averages of more than two groups? Then one must use analysis of variance. This technique has the same kinds of assumptions that underlie the formation of the t-test (interval measurement of the dependent variable, normal distributions of the populations, random and independent sampling, and equal variances of the populations--an assumption that was relaxed through using a different estimate of the standard error in t-tests). With analysis of variance our null hypothesis is that the means of the k categories or groups are equal:

$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

and the alternative hypothesis is that they are unequal:

$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \ldots \neq \mu_k$

While in fact the t-test is a subset of analysis of variance, the logic of analysis of variance and the nature in which the work is done differ considerably. Below we discuss both one-way analysis of variance (where only one independent variable is involved), two-way analysis of variance (using two independent variables), and other variations only briefly. We describe the logic behind the arguments, give examples of the computations, and also show a descriptive statistic that assesses the amount of association between the independent and dependent variables. Note that this means that analysis of variance can yield both an inferential test or statistic and a descriptive statistic.

One-Way Analysis of Variance

Below we first describe the logic behind analysis of variance by trying to give an intuitive understanding of the manipulations or logic involved. Then two examples are given after the various computations are explained, and finally a related measure of association is explained.

The Logic and Required Computations

In analysis of variance we are interested in understanding how an independent variable, which need be only categoric or nominally measured, can help explain the variation in an intervally measured dependent measure. For instance, the dependent variable might be school performance as measured on some kind of percentage scale. The independent variable could be the type of instruction method used such as lecture classes, reading classes, and discussion classes. We would be interested in whether the nature of the classes affected the students' academic performance, in other words we would be asking if experience in different kinds of classrooms was related to

different levels of performance. Then our null hypothesis would be that students in each different class setting would do equally well.

H_o: $\mu_l = \mu_r = \mu_d$

Our alternative hypothesis would be that the students in the different classes would score differently.

H_1: $\mu_l \neq \mu_r \neq \mu_d$

(Note that we did not predict the direction in the difference between each of the groups because more than two groups were involved. If desired a later step in the analysis would involve testing hypotheses regarding the differences between each pair of means. Caution needs to be used however if a large series of means is compared to assure that significant results do not occur by chance alone. Special techniques are available for these comparisons.)

If we knew nothing about the classes to which the students were assigned, our best single predictor of the students' achievement would be the overall average of their achievement scores. That is because variations of the scores about the mean are always at a minimum: $\Sigma(X-\bar{X}) = 0$ and $\Sigma(X-\bar{X})^2$ is a minimum. In the discussion below we call $\Sigma(X-\bar{X})^2$ the variation. (Note that this contrasts to the variance, which is simply the average variation, $\sigma^2 = \frac{\Sigma(X-\bar{X})^2}{n}$ for the population and $\hat{\sigma}^2 = s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}$ as the best estimate of the population value for a sample.

It is possible to see the variation as a measure of error, a measure of how far off the individual scores are from the mean. The higher the variation, the farther removed the scores are from that best guess, the mean. The smaller the variation is, the closer are the scores to the mean and the better the mean is as a predictor of the actual scores. Thus, as variation gets smaller, there is less error, and our prediction of the scores is better.

In analysis of variance it is assumed that the population variance, σ^2, is the same in each category of the independent variable. In this case we assume that $\sigma_l^2 = \sigma_r^2 = \sigma_d^2 = \sigma^2$. This common value σ^2 is called the common variance. The variation (and the variance) in each category of the independent variable indicates the amount of error in predicting the dependent variable in each category. We are interested in understanding how this variation or variance in the dependent variable can be explained or accounted for, and, specifically, if the various categories of the independent variable can help explain it.

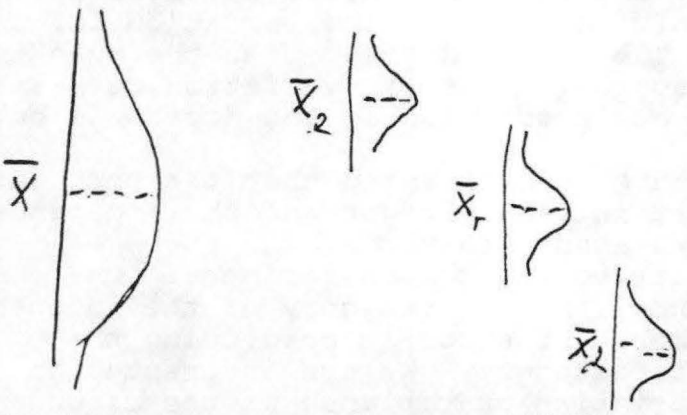
The technique of analysis of variance involves getting two estimates of this common variance. One of these estimates is unbiased. It always gives an accurate estimate of the true common variance, σ^2. This estimate is called the <u>within categories estimate</u> of the common variance and is based on the variation of the scores within each category of the dependent variables.

The second estimate of the common variance is unbiased only if the null hypothesis is true. This estimate comes from looking at the variation of the category means around the mean of the total sample and is called the between categories estimate of the common variance.

Now, because these two estimates of the common variance are equal only when the null hypothesis is true, we can use a ratio of their values (between categories estimate / within categories estimate) to test the null hypothesis. This ratio is an F-ratio or F-statistic. If this ratio is close to unity, then the null hypothesis is likely true. If the ratio is quite a bit larger than unity, then we may cast doubt on the null hypothesis. As we noted in the last section, the F statistic is the ratio of 2 variances and its sampling distribution is known for samples of various sizes. It is possible then to find out if the F statistic from this ratio falls into a zone of rejection, That is, we can determine the probability that such an F ratio would occur if the null hypothesis were true.

Now we will go into each step of the above description in more detail. First, what are the two estimates of the common variance? Remember that the common variance is the variance of the scores in each category. This is illustrated in Figure 6-1 below:

Figure 6-1



For each category we can find the variation or error around the category mean ($\Sigma(X-\bar{X}_1)^2$; $\Sigma(X-\bar{X}_r)^2$; $\Sigma(X-\bar{X}_d)^2$. These estimates of the variations are also called the sums of squares within categories, SS_w. For each category we can compute the sums of squares, the variation of individual scores around the

category mean. We can then add up this variation in each category. $SS_w = \sum_j \sum (X-\bar{X}_{.j})^2 = \sum (X-\bar{X}_l)^2 + \sum (X-\bar{X}_r)^2 + \sum (X-\bar{X}_d)^2$. This is the variation of the scores around the means within the categories. It is also called the <u>within categories sum of squares</u> or WSS It is also called the unexplained sums of squares, referring to the fact that it is the variation that is unexplained, or not accounted for, by the categories of the independent variable.

To get an estimate of the common variance we must divide this sum of squares, or variation, by the appropriate degrees of freedom. As with computations of degrees of freedom in other cases, this involves the amount of freedom we would have in assigning values to subjects after certain parameters are fixed. For the within categories sums of squares the degrees of freedom are always the number of cases involved minus the number of categories (N-k). This is because there are N subjects involved and k means. Another way of remembering this is realizing that instead of dividing by N-1 to get the best estimate of the population, we are really trying to get the best estimate of the variance for several (k) populations (even though this estimate is assumed to be the same in each case). Thus we cannot divide the sums of squares by N-1, but must divide by N-k, to get the best estimate of the population variances. <u>This estimate of the common variance that is obtained from the within category sums of squares is the within categories estimate of the common variance</u>. It is also called <u>the mean sum of squares within categories</u> (MSS_w) referring to the process of averaging involved in getting the value for the variance. <u>This value is always an unbiased or correct estimate of the population variance.</u>

The second way of estimating the population variance, σ^2, is unbiased only if the null hypothesis is true, if the category means are equal. This estimate involves looking at the variation of the means in each category around the grand mean or overall mean of the sample. In general this may be written as

$$\sum_{j=1}^{k} (\bar{X}_{.j} - \bar{X}_{..})^2$$

In this example, the estimate would equal

$$(\bar{X}_{.l} - \bar{X}_{..})^2 + (\bar{X}_{.r} - \bar{X}_{..})^2 + (\bar{X}_{.d} - \bar{X}_{..})^2$$

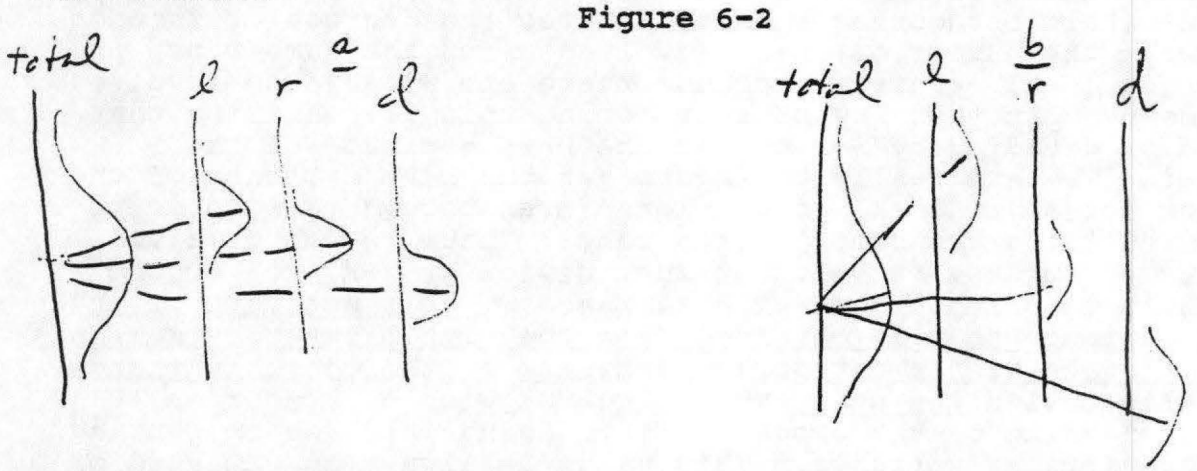
This is called the <u>sum of squares between categories or BSS</u>. It is also called the explained sums of squares, referring to the fact that it involves the amount of variation that can be accounted for by the means of the categories of the independent variable.

We can divide this value by the appropriate degrees of freedom to obtain the between categories estimate of the common variance. In this case the appropriate degrees of freedom are the number of categories involved minus one (k-1). This is simply because k different values (category means) are used in

calculating the variation. To have a good estimate of the population value, we must have the number of cases less one in the denominator rather than the number of cases itself (as we would in the case of the total population). This estimate of the common variance is called the between categories estimate of variance of the Mean sum of squares between, MSS_b.

If the null hypothesis is true and the categories do have equal averages, this estimate (MSS_b) will be an unbiased estimate of the true common variance and $MSS_b = MSS_w$. But, if the null hypothesis is not true, there will tend to be more variation of the category means around the grand mean and MSS_b will be greater than MSS_w. Figure 6-2 gives an intuitive illustration of why this occurs.

Figure 6-2



In part a of Figure 6-2 each of the category means is quite close to the overall mean of the total population. In part b, however, each of the category means is farther away from the overall mean of the total group. Clearly, $\Sigma(\bar{X}_j-\bar{X})^2$, the between sums of squares, will be larger in part b than in part a, even though the variation in the three categories (the within sums of squares) has not changed. Thus, in part b the MSS_b, the between categories estimate of the common variance will be much larger than the MSS_w, the within categories estimate.

To tell if this difference in the estimates of the variance is greater than would occur simply by chance, we use the F-ratio. This is a ratio of variances and it is simply

$$F=MSS_b/MSS_w.$$

The larger estimate of variance (which will be MSS_b if H_o is to be rejected) will always go on top.

If the null hypothesis is true, these two estimates will be approximately equal and the ratio will be close to one. If H_o is not true, the ratio will be greater than one. As we mentioned in the previous section, the sampling distribution for F is known and is summarized in tables in Table J in the appendix of Blalock. This table gives the critical values of F needed to

reject the null hypothesis for the .05, .01, and .001 levels of significance. If our computed F-value for a sample is larger than a critical value, then we may reject the null hypothesis in favor of the alternative at that level of significance.

The degrees of freedom correspond to denominators used for each estimate of the common variance. This is k-1 for the MSS_b and N-k for the MSS_w, seen as degrees of freedom one and two respectively in the tables.

The notation used in analysis of variance is actually quite simple. As shown in Table 6-1 each individual score can be assigned a subscript. The first number in the subscript refers to its position within the category. The second subscript refers to the category to which it belongs. Thus, X_{33} is the third case in the third category, and X_{41} is the fourth case in the first category. $\bar{X}_{.1}$ is the average of the values in the first category. $\bar{X}_{.k}$ is the average of the values in the kth category. $\bar{X}_{..}$ is the average of all the values.

$\sum_{i=1}^{n_1} X_{i1}$ means the summation of all values in the first category;

$\sum_{i=1}^{n_k} X_{ik}$ means the summation of all values in the kth category.

$\sum_i \sum_j X_{ij}$ means the summation of all values in the table, over all the cases i and over all categories j.

Table 6-1

i	2	3	- - -	k
X_{11}	X_{12}	X_{13}	- - -	X_{1k}
X_{21}	X_{22}	X_{23}	- - -	X_{2k}
.	.	.		
.	.	.		
.	.	.		
X_{n1}	X_{n2}	X_{n3}	- - -	X_{nk}
sums $\sum_{i=1}^{n_1} X_{i1}$	$\sum_{i=1}^{n_2} X_{i2}$	$\sum_{i=1}^{n_3} X_{i3}$	- - -	$\sum_{i=1}^{n_k} X_{ik}$
averages $\bar{X}_{.1}$	$\bar{X}_{.2}$	$\bar{X}_{.3}$	- - -	$\bar{X}_{.k}$
sample size $n_{.1}$	$n_{.2}$	$n_{.3}$	- - -	$n_{.3}$

grand total $\sum\sum X_{ij}$ | $\bar{X}_{..}$ | $n_{..}$

The computation of the values of MSS_b and MSS_w for a sample all involve the notion of variation. $\Sigma(Xij-X..)$ is the variation of a single score around the grand or total mean. This single variation or deviation from the grand mean may be broken into two parts as shown in equation 6-1 below.

$$(X_{ij} - \bar{X}..) = (X_{ij} - \bar{X}.j) + (\bar{X}.j - \bar{X}..) \qquad (6\text{-}1)$$

The first element is the deviation of an individual score from the mean in each category and the second element is the deviation of the category mean from the grand mean. Note that if the parentheses were removed the two terms of the category mean would cancel each other out and the terms on both sides of the equation would be the same.

Now, in just a mathematical twist, we can square both sides of the equation, and we get as in 6-2 below:

$$(X_{ij}-\bar{X}..)^2 = (X_{ij}-\bar{X}.j)^2 + 2(X_{ij}-\bar{X}.j)(\bar{X}.j-\bar{X}..) + (\bar{X}.j-\bar{X}..)^2 \qquad (6\text{-}2)$$

We could then add up all the deviations of each score from the mean. In other words we could sum across all cases and in each category, and we would have the results shown in 6-3.

$$\sum_i\sum_j (X_{ij}-\bar{X}..)^2 = \sum_j\sum_i (X_{ij}-\bar{X}.j)^2 + 2\sum_i\sum_j (X_{ij}-\bar{X}.j)(\bar{X}.j-\bar{X}..) + \sum_i\sum_j (\bar{X}.j-\bar{X}..)^2 \qquad (6\text{-}3)$$

Look now at only the middle term in equation 6-3. The calculations in 6-4 below show that this term actually equals zero. That is because in any column the category mean $\bar{X}.j$ is always a constant. Thus the summation across the categories can be moved outside, and we need only look at the summations of the deviations of the scores around the category mean. Yet, by definition the deviations of the scores around the category mean equals zero and so the whole term reduces to zero.

$$\begin{aligned} &\sum_j\sum_i (X_{ij}-\bar{X}.j)(X_{ij}-\bar{X}..) = \\ &\quad [2(\bar{X}.j-\bar{X}..)\sum_i (X_{ij}-\bar{X}.j)] \\ &\quad = \sum_j [2(\bar{X}.j-\bar{X}.\cdot)(0)] = 0 \end{aligned} \qquad (6\text{-}4)$$

Finally, then, the sum of the squared deviations around the mean can be written as

$$\sum_i\sum_j (X_{ij}-\bar{X}..)^2 = \sum_i\sum_j (X_{ij}-\bar{X}.j)^2 + \sum_i\sum_j (\bar{X}.j-\bar{X}..)^2 \qquad (6\text{-}5)$$

$$SS_t = SS_w + SS_b$$

The first term is the total sum of squares, the variation of scores around the grand mean. The first term on the right hand side is the within categories variation, the sum of squared deviations around the categories means. The second term on the right hand side is the between category sum of squares, the between categories variation, or the sum of squared deviations of the category means around the grand means. In other words, the

total variation of the sample (which can be used to estimate the total variance of the population) can be broken into the variation within the sample categories and the variation between the sample categories. (Note that this refers to variation, not to variance.) The SS_w, the within categories variation, is sometimes called the unexplained or error sums of squares. This is variation that remains after the division of the sample into categories of the independent variable and cannot be accounted for by the independent variable. The SS_b, the between categories sum of squares, is called the explained sum of squares, that part of the total variation that can be accounted for or explained by the categories of the independent variables.

Note that the within categories sum of squares (SS_w) can be calculated if the variance of the dependent variable in each category is known.

$$SS_w = \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 = \sum_{i=1}^{n_1} X_{i1} - \bar{X}_{.1})^2 + \sum_{i=1}^{n_2} (X_{i2} - \bar{X}_{.2})^2$$

$$+ \ldots + \sum_{i=1}^{n_k} (X_{ik} - \bar{X}_{.k})^2 = (n_1 - 1)(s_1^2) + (n_2 - 1)(s_2^2) + \ldots \quad (6\text{-}6)$$

$$+ (n_k - 1)(s_k^2) \quad \text{where } s_k^2 = \frac{\sum (X_{ik} - \bar{X}_{.k})^2}{n_k - 1} = \tilde{\sigma}_k^2$$

Remember that $s_k^2 = \frac{\sum (X_{ik} - \bar{X}_{.k})^2}{n_k - 1}$, also called $\tilde{\sigma}_k^2$ by Blalock. Because $(n-1)s_k^2 = \sum_{i=1}^{n_k} (X_{ik} - \bar{X}_{.k})^2$, SS_w may be written as in the second line of (6-6).

To get estimates of the variance in each category, the common variance, we must divide the sums of squares by the appropriate degrees of freedom as described earlier. For the total sample remember that to get the best estimate of the population variance, we must divide the total sum of squares by n-1. (We lose one degree of freedom here by computing just one mean.) For SS_b we have k-1 degrees of freedom, losing one degree of freedom for the overall or grand mean. For SS_w, we have N-k degrees of freedom, losing one degree of freedom for each category mean.

Note that

$$N - 1 = (N - k) + (k - 1) = N - 1 \quad (6\text{-}7)$$

df for total = df for SS_w + df for SS_b

That is, the total degrees of freedom are equal to the degrees of freedom for the sums of squares within plus the degrees of freedom for the sums of squares between, just as the total sums of squares equals the sums of squares within plus the sums of squares between.

The two estimates of the common variance are

$$MSS_w = SS_w / (N - k) = \sum_i \sum_j (X_{ij} - \bar{X}_{.j})^2 / (N - k) \qquad (6\text{-}8)$$

$$MSS_b = SS_b / (k - 1) = \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 / (k - 1) \qquad (6\text{-}9)$$

And the F ratio = MSS_b / MSS_w with the degrees of freedom = (k-1, N-k). If the null hypothesis, that the means within the categories are equal, is true then this ratio should be close to one. If it is not true, then it will be enough larger than one that the F-value falls into the zone of rejection for the F-distribution corresponding to the appropriate degrees of freedom.

Examples

Below a hypothetical example of the achievement of students in three different classrooms is shown. The dependent variable is the student's achievement, measured by the number of correct answers given on an exam. The independent variable is the classroom to which the students were assigned, either a lecture, discussion, or reading section. The analysis of variance model assumes that the students were randomly assigned to the classroom (they weren't matched in any way), that the population of achievement scores in each category of the independent variable is normally distributed, and that the population variance in each category is equal. The null hypothesis is that the population means in each category are equal, and the alternative hypothesis is that these means are not equal. Table 6-2 shows the actual scores of students in each class. For computation purposes the squares of each of these scores are also given.

Ho: $\mu_l = \mu_r = \mu_d$

Hi: $\mu_l \neq \mu_r \neq \mu_d$

Table 6-2

The Scores of Students in Each Group

Lectures		Reading		Discussion	
X	X^2	X	X^2	X	X^2
50	2500	50	2500	55	3025
55	3025	60	3600	65	4225
60	3600	60	3600	65	4225
90	8100	70	4500	75	5625
95	9025	75	5625	75	5625
70	4900	80	6400	80	6400
75	5625	85	7225	80	6400
80	6400	90	8100	85	7225
80	6400	95	9025	90	8100
90	8100	95	9025	95	9025

$\sum_{i=1,..}^{n_1} X_{i1} = 745$ $\quad \sum_{i=1}^{n_2} X_{i2} = 760$ $\quad \sum_{i=1}^{n_3} X_{i3} = 765$

$\sum_{i=1,..}^{n_1} X_{i1}^2 = 57,675$ $\quad \sum_{i=1}^{n_2} X_{i2}^2 = 60,000$ $\quad \sum_{i=1}^{n_3} X_{i3}^2 = 59,875$

$N_1 = 10$ $\quad N_2 = 10$ $\quad N_3 = 10$

$\bar{X}_{.1} = 74.5$ $\quad \bar{X}_{.2} = 76$ $\quad \bar{X}_{.3} = 76.5$

$\bar{X}_{..} = \frac{2270}{30} = 75.66$ $\quad \sum_i \sum_j X_{ij} = 745 + 760 + 765 = 2270$

$N_{..} = 30$ $\quad \sum_i \sum_j X_{ij}^2 = 57,675 + 60,000 + 59,875 = 177,550$

TSS (total sums of squares) $= \sum_{j=1}^{3} \sum_{i=n}^{n_i} X_{ij}^2 - \frac{\left(\sum_j \sum_i X_{ij}\right)^2}{N}$

$= 177,550 - \frac{(2270)^2}{30} = 177,550 - 171,763.33 = 5786.67$

BSS (between sum of square) $= \left[\frac{\left(\sum_{i=1}^{n_1} X_{i1}\right)^2}{n_1} + \frac{\left(\sum_{i=1}^{n_2} X_{i2}\right)^2}{n_2} + \frac{\left(\sum_{i=1}^{n_3} X_{i3}\right)^3}{n_3}\right] - \frac{\left(\sum_i \sum_j X_{ij}\right)^2}{N}$

$= \left[\frac{(745)^2}{10} + \frac{(760)^2}{10} + \frac{(765)^2}{10}\right] - 171,763.33 = 171,785 - 171,763.33 = 21.67$

WSS (within sums of square s) = TSS - BSS = 5786.67 - 21.67 = 5765

$E^2 = \frac{BSS}{TSS} = \frac{21.67}{5786.67} = .004$

Table 6-3
Testing Ho

Source of Variation	SS	df	MS = $\frac{SS}{df}$	F
Total	5786.67	n-1 = 29	199.54	
Between	21.67	k-1 = 2	10.835	.0507
Within	5765	n-k = 27	213.52	

To reject Ho at .05 level of significance with df = 2,27
We need F > 3.35. Obviously we cannot reject the null hypotheses of no difference between the category means.

In our discussion above we gave the definitional formulas for the total sum of squares and the sum of squares within and between categories. Anytime, however, that you repeatedly do subtractions, rounding error may enter into the calculations, and it is more accurate to use computing formulas. These computing formulas for the sum of squares are shown in Table 6-2 above. Note that the total sum of squares is computed, the between sum of squares is computed, and the within sum of squares is computed by simple subtraction from these two values. (We will explain in the next section what E^2 is, it is simply a measure of association.)

Table 6-3 summarizes the information needed to test the hypotheses. This is the standard form of a table used in analysis of variance. It shows the source of the variation that is being measured (1st column), gives the values of the sum of squares (2nd column), the degrees of freedom or the value in the denominator when obtaining the best estimate of the population variance (3rd column), the mean sum of squares, or the estimate of the population variance (4th column), and in the final column the F ratio, obtained by comparing the MSSb to the MSSw in a ratio. In this case our degrees of freedom are 2 and 27. From Table J in the appendix of Blalock it is apparent that to reject the null hypothesis with these degrees of freedom we need an F-value greater than or equal to the critical value of 3.35. Obviously, the F-value for our sample is much less than this (and in fact less even than one) and so we must fail to reject the null hypothesis of no difference between the category means. Apparently, from this study we would conclude that the three teaching methods do not result in different amounts of student achievement. The researcher would then speculate on possible reasons why this might be so.

(at $p = .05$)

A second example comes from the field of physical education (like psychology they very often use analysis of variance models, largely because they use experimental designs). Suppose that students were randomly assigned to one of four different activity programs: yoga, calisthenics, football, and nothing. We were interested in how these programs affected their overall

flexibility. The dependent measure is the difference in flexibility in inches measured from the beginning of the term to the end of the term. The null hypothesis is

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (with the numbers representing each activity)

and the alternative hypothesis is

H_1: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

The analysis of variance model assumes that the dependent variable, difference in flexibility, is intervally measured (here it is actually measured on a ratio scale), that the subjects have been independently and randomly selected, that the scores are normally distributed in the populations, and the variances in the populations are equal. (If you do suspect that the variances are not equal, you could check this hypothesis, although simulations ~~report~~ suggest that this need not be a large worry when the populations are not skewed or the examples markedly different in size).

The data for the example are given in Table 6-4 and the summary information is given in Table 6-5.

Table 6-4
Differences in Flexibility

Calisthenics		Yoga		Football		No Activity	
X	X^2	X	X^2	X	X^2	X	X^2
0	0	1	1	0	0	-2	4
1	1	2	4	1	1	-1	4
1	1	3	9	2	4	0	0
0	0	0	0	0	0	0	0
2	4	2	4	-1	1	1	1
3	9	3	9	-2	4	-4	16
1	1	4	16	-3	9	-3	9
2	4	4	16	-3	9	2	4
1	1	5	25	1	1	1	1
0	0	1	1	-4	16	0	0
Σ 11	21	25	65	-2	36	-2	36
$\bar{X}_{\cdot j}$ 1.1		2.5		.2		-.2	
n_j 10		10		10		10	

$\sum_i \sum_j X_{ij} = 36$ $\sum_i \sum_j X^2_{ij} = 158$

Each score measures change in flexibility from beginning to end of term. Larger scores indicate a greater change.

Table 6-4 (continued)

$$TSS = SS_t = \sum_i \sum_j X_{ij}^2 - \frac{(\sum\sum X_{ij})^2}{N} = 158 - \frac{(36)^2}{40} = 125.6$$

$$SS_b = \sum_j \left[\frac{\sum X_{ij}}{n_{\cdot j}} \right] - \frac{(\sum\sum X_{ij})^2}{N} = \left[\frac{(11)^2}{10} + \frac{(25)^2}{10} + \frac{(2)^2}{10} + \frac{(2)^2}{10} \right] - \frac{(36)^2}{40} = 43.0$$

$$SS_w = SS_t - SS_b = 125.6 - 43.0 = 82.6$$

$$E^2 = SS_b/SS_t = BSS/TSS = 43.0 / 125.6 = .34$$

Table 6-5
Analysis of Variance Summary Table

Source of Variation	SS	Degrees of f.	MSS	F
Total	125.6	N-1 = 39		
Between	43.0	k-1 = 3	14.33	6.25
Within	82.6	n-k = 36	2.29	

In this case, with degrees of freedom equal to 3, 36, Table J in the appendix to Blalock shows that at the .05 level of significance we need an F-ratio of 2.92 to reject the null hypothesis that the means in the categories are equal. Obviously, 6.25 is greater than 2.92, and we can reject the null hypothesis at this level. Further inspection of Table J shows that this F-ratio for our sample is large enough to reject the null hypothesis at the .01 level of significance, but not at the .001 level. Thus, we can reject our null hypothesis that the activities produce the same amount of change in flexibility in favor of the alternative that they differ in the amounts of change induced and be wrong in doing so less than one time out of one hundred, but more than one time out of a thousand.

We can then inspect the category means and see that yoga produces the most change, calisthenics the next greatest amount, and football and no activity almost none. One could then, if desired, compare the differences between these separate activities, noting that repeated tests do carry a problem of getting significant results simply by chance. (A Sheffe test can be used for such comparisons and takes into account the problem of repeated tests of significance. Advanced texts in psychology

discuss this procedure.) One would also want to speculate about why yoga would be a better activity at producing flexibility changes than the others.

The best computer programs to use on SPSS-PC for a one analysis of variance for most sociology purposes are MEANS or ONEWAY. MEANS gives a table with the category means, the related F-ratio and the measure, E^2, the measure of association discussed below. ONEWAY includes tests for comparing multiple means with each other.

E^2, a Measure of Association for Analysis of Variance Problems

Sometimes, a researcher is not as much interested in whether or not the category means are equal as she or he is interested in how much the independent variable helps us understand the variation in the dependent variable. How much does the independent variable help reduce our error in predicting values of the dependent variable? The answer to this comes immediately from the summary tables used in computing analysis of variance, by looking at the total sum of squares and the breakdown of this total variation of the dependent variable into that which can be accounted for by the independent variable (SS_b) and that which cannot be accounted for by the independent variable (SS_w). The measure E^2 (called eta^2 by SPSS, E^2 by Blalock) summarizes this information.

$E^2 = SS_b / SS_t =$
variation explained by independent variable/total variation.
Note that this also equals
(total variation - unexplained variation)/ total variation =
$(SS_t - SS_w) / SS_t$. This measure can simply be interpreted as telling us the proportion of the total variation or error in the dependent variable that can be explained or accounted for by the categories of the independent variable.

A measure of association should always be used, when possible, in conjunction with an inferential test. This can guard against unwarranted substantive conclusions from the inferential results based on large or small sample sizes. For instance, if the sample is small, but the E^2 measure indicates that a substantial proportion of the variation is explained by the independent variable even though the F statistic only approaches significance, then one should be sure to report that there is good reason to believe that association exists and that replication seems in order.

It is also important to note briefly some confusion in the literature in notation. What Blalock calls E^2, SSPS calls eta^2. What Blalock calls eta^2 is an analogue to E^2 but uses variances rather than variation. Blalock's eta^2 then has no easy intuitive interpretation. We will see in the next unit on correlation that E^2 is analogous to the square of the correlation coefficient r^2. What Blalock calls E^2 and SPSS calls eta^2 is then the most useful

measure. It is the one that I will be using throughout the term. However, you should always check carefully for the formula used when people deal with this measure to make sure that they are talking about proportion of variation (or sums of squares) that is explained.

Two-Way Analysis of Variance

In this section we review the logic of two-way analysis of variance, the situation with two independent variables and one intervally measured dependent variable. We review the logic, go through an example, discuss more complicated instances only briefly, and review the associated computer procedures.

The Logic of Two-Way Analysis of Variance

With one-way analysis of variance we compare the total distribution of a group of scores (say income of a group of people) with the distribution of the scores within a number of categories in a nominally measured variable (say religious groups). We test the null hypothesis that the average incomes of the people in each religious group are equal. Here income is the dependent variable and religion is the independent variable. We are essentially trying to see if breaking the scores or people into religious groups helps us to know more about their income. We use variation of the incomes around the mean ($\sum\sum (X_{ij} - \bar{X}..)^2$ as our basic measure. The variation of all the incomes around the average of the total group may be broken into two parts: that unexplained by religious groups (the variation within each of the religious categories or the within categories sum of squares -- SS_w) and that explained by the religious breakdown (the between categories sum of squares or the variation between the means of each category and the total mean -- SS_b). The within category sum of squares is defined as $\sum\sum(X_{ij} - \bar{X}.j)^2$ and the between category sum of squares is defined as $\sum\sum_{i\ j} (\bar{X}.j - \bar{X}..)^2$.

In two-way analysis of variance we have not one but two independent variables, each measured on a nominal scale. For instance, we may be interested in the effect of both race and religion on income. (The formulas given below and in class relate only to the case where we have an equal number of scores in each subcategory, i.e. there are the same number of people in each combination of race and religious group. I will return to this later when discussing the computer calculations.) With two-way analysis of variance we may again break the total variation of the dependent variable into parts. The unexplained variation is that which has not been explained by the two independent variables. This is the variation within each of the subcategories of race and religion such as the variation of incomes of white Protestants and the variation of incomes of white Catholics. This may be called the within subclass sum of squares or variation.

The explained variation is the between-subclass sum of squares or the variation between the means of each subcategory and the total mean. Note the direct analogy to the between category sum of squares in simple one-way analysis of variance. This explained variation can then be broken into three component parts: 1) that explained by one independent variable (say religion; 2) that explained by the other independent variable (say race); and 3) that explained by the interaction of these two independent variables or the special joint effect of race and religion on income (for instance the special discrimination given to Black Jews over and above that which they get separately as Blacks or as Jews).

These components of the explained variation are called respectively the between columns sum of squares (if religion is placed in the columns in the table); the between-rows sum of squares (if race is placed in the rows in the table); and the interaction sum of squares. They are defined as follows:

The within subclass sum of squares: $\sum_i \sum_j \sum_k (X_{ijk} - \bar{X}_{.jk})^2$

The total explained variation or between subclass sum of squares : $\sum_i \sum_j \sum_k (\bar{X}_{.jk} - \bar{X}_{...})^2$

This total explained variation includes:

The between columns sum of squares: $\sum_i \sum_j \sum_k (\bar{X}_{.j.} - \bar{X}_{...})^2$

The between rows sum of squares: $\sum_i \sum_j \sum_k (\bar{X}_{..k} - \bar{X}_{...})^2$

The interaction sum of squares: between subclass SS - (between columns SS and between rows SS)

Note that the interaction sum of squares is simply that part of the total explained variation that cannot be accounted for by the column variable alone or by the row variable alone.

Also note that since $SS_t = SS_{\text{within subclass}} + SS_{\text{between subclass}}$, we may compute $SS_{\text{within subclass}}$ by simply $SS_t - SS_{\text{between subclass}}$ in a direct analogy to the procedure with one way analysis of variance.

The notation used in two-way analysis of variance is very similar to that with one-way anova with the simple addition of one more subscript to refer to the row variable. This is illustrated in Table 6-7.

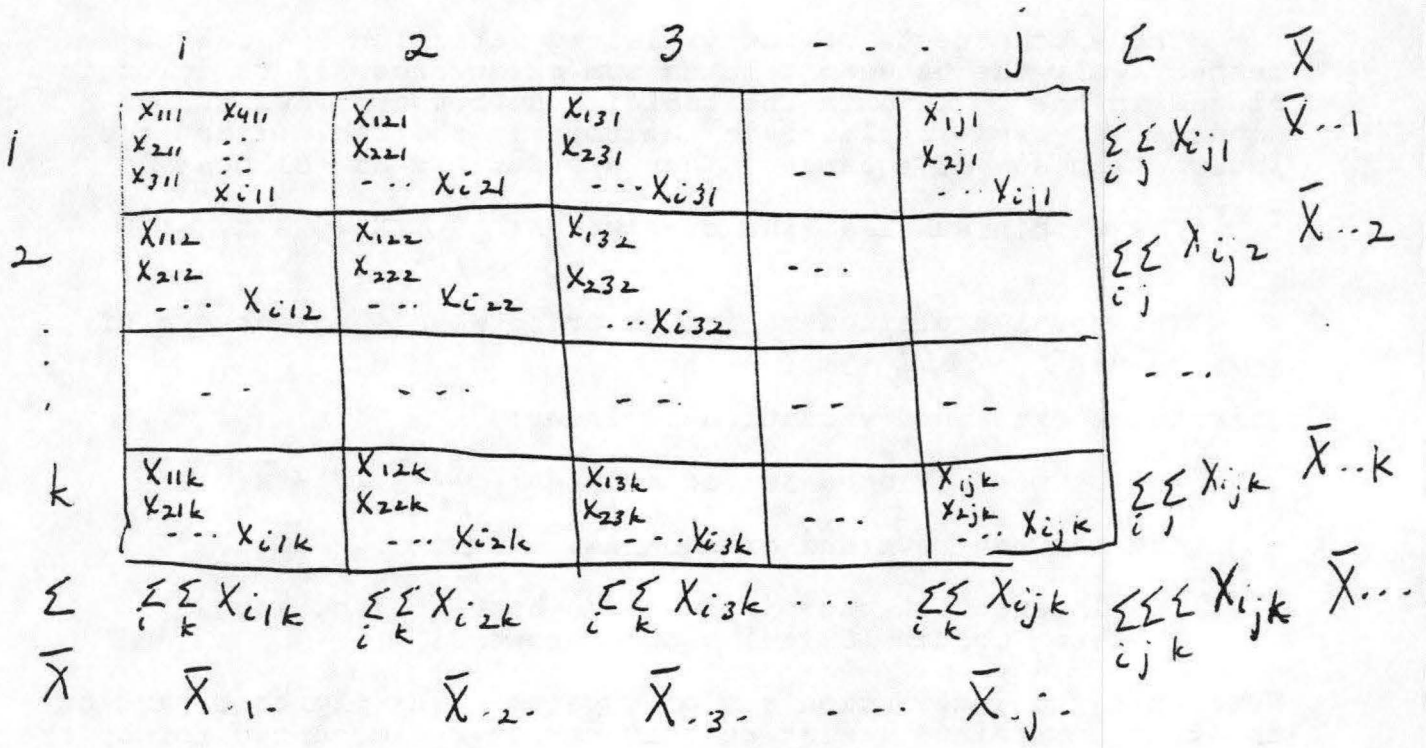
Table 6-7
Notation in Two-way anova

X_{ijk} -- each score: i refers to the score, j refers to the column, k refers to the row

$\bar{X}...$ -- the mean of all scores

$\bar{X}..k$ -- the mean in each row k

$\bar{X}.j.$ -- the mean in each column j



Interaction Effects

Before giving an example, we will try to give more of an intuitive idea of what interaction is. Interaction is simply additional influence of the independent variables that comes from their joint action. Some examples can help. Suppose that we had an experiment where we were interested in levels of aspiration. We randomly assigned college men to three treatment groups. They were all given a game that apparently involved their own skill in winning. In actual fact the outcome of the game was controlled and everyone would get the same score. Yet, the subjects were told that they had done above average, average, or below average. They were then asked to predict how well they would do on the next try. The average predicted score on the next try is the score given in the tables below. Suppose that we did this with a random sample of athletes and a sample of regular college

students. If there were no interaction the results shown in Table 6-8 would occur.

Table 6-8
No Interaction, Only a Column Effect

	Above Av.	Average	Below Average
College students	28	33	35
Athletes	28	33	35

In Table 6-8 there is no interaction, and only the experiment itself had an effect. Whether or not the students were athletes had no influence. Note that for both cases the difference from being told one was above average to average was 5 points and the difference from average to below average was 3 points.

In Table 6-9 again there is no interaction. This time, however, the experimental variable has no effect and the only variable with an influence is the status of the students. Consistently, the college students predicted a four point better performance than the athletes, regardless of the experimental condition.

Table 6-9
No Interaction, Only a Row Effect

	Above av.	Average	Below Average
College Students	34	34	34
Athletes	30	30	30

In Table 6-10 there is again no interaction, but this time both the experimental variable and the student type affect the scores. The effect of moving from above average to average is 5 points, the effect of moving from average to below average is 2 points -- for both the college students and the athletes. Consistently also the college students differ from the athletes by four points.

Table 6-10
No Interaction, Both Row and Column Effects

	Above Av.	Average	Below Average
College students	30	35	37
Athletes	26	31	33

Finally, in Table 6-11 we may see interaction effects. In this situation the athletes show a different pattern between the three experimental groups than the college students do. Conversely we may say that the college students and athletes differ from each other in different ways in each experimental

situation. For the above average situation the college students and athletes are the same. They differ widely, with the college students 16 points higher in the average situation. And they differ less strikingly, by only 4 points but in the other direction, in the below average situation.

Table 6-11
Interaction Effects

	Above av.	Average	Below Average
College Students	28	41	33
Athletes	28	25	37

Interaction effects are very important theoretically, and generally cannot be detected from simply inspecting the marginals. Table 6-12 illustrates this possibility. In this case suppose that two methods of instruction (#1 and #2) were used in teaching students, both boys and girls. The average achievement scores for each group are shown with the marginals, or average scores for all the students, on the outside columns and rows. From comparing the averages for all boys and girls we would conclude that they did not differ in their achievement. From comparing the overall averages for the two methods we would conclude that method one was superior to method 2. However, by looking inside the table it appears that method 2 is better for girls, while method one is better for boys. Interaction has occurred. The methods do not have the same effects on boys as they do on girls.

Table 6-12
Example of Interaction Effect and How Cannot be Detected From the Marginals

	Method One	Method Two	Total Scores
Girls	55	65	60
Boys	75	45	60
Total Scores	65	55	60

Hypotheses

Because there are three components to the explained sums of squares in two-way analysis of variance, three hypotheses must be tested. These are:

H_{01}: the population means of the categories in the row variables are equal $\mu_{\cdot\cdot 1} = \mu_{\cdot\cdot 2} = \ldots = \mu_{\cdot\cdot k}$

H_{11}: the population means of the categories in the row variables are unequal $\mu_{\cdot\cdot 1} \neq \mu_{\cdot\cdot 2} \neq \ldots \neq \mu_{\cdot\cdot k}$

H_{02}: the population means of the categories in the column variable are equal $\mu_{\cdot 1 \cdot} = \mu_{\cdot 2 \cdot} = \ldots = \mu_{\cdot j \cdot}$

H_{12}: the population means of the categories in the column variable are unequal $\mu_{\cdot 1 \cdot} \neq \mu_{\cdot 2 \cdot} \neq \ldots \neq \mu_{\cdot j \cdot}$

H_{03}: there is no interaction (essentially the sum of squares accounting for interaction and the associated estimate of common variance is zero)

H_{13}: there is interaction

If it turns out that we must reject H_{03}, in other words if there is interaction, then it makes no sense to go ahead and examine the first two hypotheses. This is simply because if interaction exists then different things are happening between the rows depending on which column you look at and vice versa. Thus, separate generalizations about the rows and columns make no sense. Thus, the first hypothesis to be tested is that regarding interaction. Then the other two hypotheses are also tested.

An Example

On the following pages an example is given that is an extension of one of the examples used with one-way analysis of variance. In this example we are interested in examining the influence on a person's change in flexibility over a term of assignment to three different activities: none, sports and yoga as well as the kind of body (loose versus stiff) that a person has. As before this change in flexibility is measured in inches.

Three hypotheses are listed: that there is no interaction (actually testing one of the assumptions needed to complete the analysis), that the column means are equal -- that each activity contributes equally well to changes in flexibility, and that the row means are equal -- that people of both body types have equal amounts of change.

Note that all the equations used apply only to cases where there are equal number of cases in each subcategory. This is because this automatically makes the activity type and the body type not associated with each other. If this did not occur the various sources of explained variance would not sum to the total explained variance as they should. Computer programs can easily handle the more complex equations needed when there are unequal n's across the cells. Below the definitional formulas for each source of variation are again given. Then the computing formulas and the computations for this example are given. Note again that one should always use the computing formulas rather than the definitional formulas in computations to minimize rounding errors and also to prevent simple arithmetic mistakes.

First the hypothesis that there is no interaction is tested (Table 6-14). This is done by commputing the sum of squares for interaction and for error (or within each sub class) along with the associated degrees of freedom and then computing the estimates of the common variance. The estimates of the common variance are shown in the fourth column of the table. All of these procedures are just as in one-way analysis of variance except that here we have more sources of explanation of the variation. To obtain the F-ratio to test the hypothesis regarding interaction, we look at the ratio of the estimate of variance from the sums of squares due to interaction and that due to error. This F ratio = .067 which is far below unity and we may fail to reject the null hypothesis of no interaction. Thus, we may assume that additivity exists. This means that the column variable and the row variables act independently on the dependent variable and we may analyze their separate effect.

Blalock suggests that at this point one should add the sum of squares or variation due to interaction back with the error sum of squares, the variation within each subclass. His justification for doing this is that if we can fail to reject the null hypothesis that interaction exists, then any variation explained by interaction must be simply due to error and so we can treat it as such. This is the procedure we have used above. However, other sources do not do this. For instance, Hays, the premier text in psychology does not add the interaction sum of squares into the error term, but retains it as a separate element. Similarly, SPSS retains it separately. If you do add it to the error term note that the degrees of freedom are also added in.

Table 6-13
Example of Two-Way Analysis of Variance

We will test each of the following null hypotheses:

H_{01}: no interaction among the cells (between the 2 variables) (Note this is testing one of the assumptions on which 2-way anova is based and is necessary to do before proceeding to the next 2 hypotheses.)

H_{11}: There is interaction among the cells (between the 2 variables).

H_{02}: $\mu_1 = \mu_2 = \mu_3$ (column means are equal or means within each category of activity are equal)

H_{12}: $\mu_1 \neq \mu_2 \neq \mu_3$ (column means are unequal or means within each category of activity are unequal)

H_{03}: $\mu_1 = \mu_2$ (row means are equal or category means within each body type are equal)

H_{03}: $\mu_1 \neq \mu_2$ (row means are not equal or category means within each body type are unequal)

Note: The method illustrated here is derived for cases with equal n's in each subcategory.

Body Type	ACTIVITY: None	Sports	Yoga	Total
Loose	-1 1 0 2 0 1 0 0 -2 1 $\Sigma X = 2$ $\Sigma X^2 = 12$	-1 1 -1 3 0 2 1 1 -1 2 $\Sigma X = 7$ $\Sigma X^2 = 23$	0 2 1 2 0 3 1 1 2 2 $\Sigma X = 14$ $\Sigma X^2 = 28$	$\Sigma X = 23$ $\Sigma X^2 = 63$
Stiff	-2 -1 0 2 -1 1 0 0 -1 1 $\Sigma X = -1$ $\Sigma X^2 = 13$	-3 3 0 2 -1 0 0 1 -1 1 $\Sigma X = 3$ $\Sigma X^2 = 27$	-1 1 3 2 0 2 1 -1 0 1 $\Sigma X = 8$ $\Sigma X^2 = 22$	$\Sigma X = 10$ $\Sigma X^2 = 62$
	$\Sigma X = 1$ $\Sigma X^2 = 25$	$\Sigma X = 10$ $\Sigma X^2 = 50$	$\Sigma X = 22$ $\Sigma X^2 = 50$	$\Sigma X = 33$ $\Sigma X^2 = 125$ $N = 60$

In two way anova our unexplained variation is within each subclass: $\Sigma\Sigma\Sigma (X_{ijk} - \bar{X}_{.jk})^2$.

The explained variation comes from 3 sources:

1) that explained by the variable in the columns
2) that explained by the variable in the rows
3) that explained by the interaction or joint effect of these two variables.

Table 6-13 (continued)

The total explained variation is called the between subclass variation and = $\sum_j \sum_k \sum_i (\bar{X}_{ijk} - \bar{X}...)_2$

The between columns SS = $\sum_k \sum_j \sum_i (\bar{X}.j. - \bar{X}...)_2$

The between rows SS = $\sum_k \sum_j \sum_i (\bar{X}..k - \bar{X}....)^2$

The interaction SS = between subclass SS - (between rows SS and between columns SS).

The error or within subcategory SS = TSS - between subclass SS.

If the contribution of the interaction SS to the total variation is not significant, Blalock suggests that we add this interaction SS back into the error SS and use this value for estimating the within category variance for the F ratios in testing the second and third hypotheses.

Computing formulas and computations:

$$TSS = \sum_k \sum_j \sum_i X^2_{ijk} - \frac{\left(\sum_j \sum_k \sum_i X_{ijk}\right)^2}{N} = 121 - \frac{(33)^2}{60} = 106.85$$

$$\text{Between columns SS} = \left(\sum_j \frac{\left(\sum_k \sum_i X_{ijk}\right)^2}{n_j}\right) - \frac{\left(\sum_j \sum_k \sum_i X_{ijk}\right)^2}{N}$$

$$= \left(\frac{1^2}{20} + \frac{10^2}{20} + \frac{22^2}{20}\right) - \frac{(33)^2}{60} = 11.1$$

$$\text{Between rows SS} = \left(\sum_k \frac{\left(\sum_i \sum_j X_{ijk}\right)^2}{n_k}\right) - \frac{\left(\sum_i \sum_j \sum_k X_{ijk}\right)^2}{N}$$

$$= \left(\frac{23^2}{30} + \frac{10^2}{30}\right) - \frac{33^2}{60} = 17.63 + 3.33 - 18.15 = 2.82$$

$$\text{Between subclass SS} = \left[\sum_j \sum_k \frac{\left(\sum_i X_{ijk}\right)^2}{n_{jk}}\right] - \frac{\left(\sum_j \sum_k \sum_i X_{ijk}\right)^2}{N}$$

$$= \frac{2^2}{10} + \frac{7^2}{10} + \frac{14^2}{10} + \frac{1^2}{10} + \frac{3^2}{10} + \frac{8^2}{10} - 18.15 = 14.15$$

Error SS = Total SS - between subclass SS = 106.85 - 14.15 = 92.7

Interaction SS = Between subclass SS - (Between Col. SS + between rows SS)

= 14.15 - (11.1 + 2.82) = 0.23

Table 6-14
Test for Interaction

Source of Variation	SS	df	Estimated Variance	F
Total	106.85	n-1 = 60	1.811	
Between subclass	14.15	kl-1 = 5	2.83	
Between columns	11.1	k-1 = 2	5.55	
Between rows	2.82	l-1 = 1	2.82	
Interaction	0.23	(k-1)(l-1)=2	.115	
Error	92.7	n-kl = 54	1.717	.067

$F = \frac{.115}{1.717} = .067$

And since the F ratio is far below unity we can fail to reject our null hypothesis of no interaction - i.e. there is additivity and we may proceed with the analysis.

Table 6-15
Testing for Column and Row Effects
(Interaction SS - added to error term)

Source of Variation	SS	df	Estimated Variance	F
Total	106.85	59		
Between columns	11.1	2	5.55	3.3433
Between rows	2.82	1	2.82	1.697
Error	92.93	56	1.66	

We may reject the null hypothesis of no difference between the means of the column categories at the .05 level of significance, but we must fail to reject the null hypothsis of no difference between the means of the row categories. That is, the type of activity, but not the body type, appears to influence an increase in flexibility.

Table 6-16
Calculation of E^2

E^2 by activity = $\frac{\text{Between columns SS}}{\text{TSS}} = \frac{11.1}{106.85} = .1039$

E^2 by body type = $\frac{\text{Between rows SS}}{\text{TSS}} = \frac{2.817}{106.85} = .026$

E^2 act, body & interaction = $\frac{\text{Between subcl. SS}}{\text{TSS}} = \frac{14.15}{106.85} = .1324$

Table 6-15 gives the summary table that shows the test of the two hypotheses regarding the influence of the row and column variables. For the influence of the column variable of activity type, the degrees of freedom are 2 and 56. From Table J in the appendix we see that an F ratio of 3.23 is needed to reject the null hypothesis at the .05 level of significance and an F ratio of at least 5.18 is needed for rejection at the .01 level. In this case the F ratio is 3.34, large enough to reject the null hypothesis at the .05 level and to tell us that the activities do indeed differ in their influence on flexibility, independent of the influence of body type.

For the impact of body type the degrees of freedom are 1 and 56 (df is less here because there were only two types of bodies, but three types of activities). For rejecting the null hypothesis at the .05 level of significance, Table J tells us that the F-ratio would have to be at least as large as 4.08. Our F-ratio is not that large and we must fail to reject the null hypothesis that body type influences the change in flexibility over the term. The researcher would then proceed to discuss why the activities influence the changes and which activities are most influential and why differences in body type apparently do not.

Table 6-16 shows the second important part of the analysis of variance procedure. These are the measures of E^2, the proportion of the total variation in the dependent variable that can be explained by each of the independent variables. Note, that as we would expect from the F-ratios, the most variation is explained by activity, over 10%. Body type alone explains almost 3% of the total variation. Together these two variables (including interaction) explain over 13% of the total variation. In trying to understand how important these variables are, think that if one had 6 more variables as good as these, one could explain all the variation. With some substantive areas that will be terrific. In others, you may not impress many people. With just the activity variable alone you would only need ten other variables just as effective and you would explain all the variation in flexibility. Obviously, it will sometimes be easier to find explanations of variables that are physical in nature such as flexibility than those that are more cognitive and social in nature.

It might be helpful to remember that E^2 is what Costner has called a PRE, a proportionate reduction of error, measure. For a PRE measure you must have a definition of error. In this case this is the variation of scores around the mean: $\Sigma\Sigma\Sigma X_{ijk} - \bar{X}...)^2$ Then we have two rules for finding this error. Rule one is when you know only the dependent variable. This is simply the total sum of squares and the squared deviations around the grand mean $SS_t = \Sigma\Sigma\Sigma(X_{ijk} - \bar{X}...)^2$. Rule two gives the error if we also know the independent variable. This would be the variation left in each category, or the within category or within subclass sum

of squares: $SS_w = \Sigma\Sigma\Sigma(X_{ijk} - \bar{X}_{.jk})^2$. A PRE measure is simply then (error by rule 1 - error by rule 2)/ error by rule 1. In this case it is $SS_t - SS_w / SS_t = SS_b / SS_t = E^2$. This is the proportion of the total variation or error in the dependent variable that can be explained by knowing the categories of the independent variable(s).

Extensions

There are many types of analysis of variance designs. These are closely associated with various experimental designs. You will hear of block designs, nested designs, etc. If you are to be involved at all with experimental work it might be a good idea to take a statistics course in the psychology department that seriously deals with these variations. Three-way analysis of variance would involve the addition of a third independent variable and the computations and logic used here would be extended.

Because most sociologists don't use experimental designs we will not often use the formats the psychologists use. After studying multiple regression, I will show you how analysis of variance is simply a subset of multiple regression and how we may use those techniques when we want to do analysis of variance.

To prepare students for this step Blalock has a discussion on pp. 356-357 that describes how an individual score may be broken into component parts. Unfortunately, the notation in those paragraphs is somewhat inconsistent. Below I have tried to translate his work with more consistent notation starting with the second line from the bottom of page 356.

"Letting the score of the ith individual in the kth row and the jth column by represented by X_{ijk}, we may conceive of this score as being composed of the following components: (1) a component "due to" the overall population mean $\mu_{...}$; (2) a second due to the effects of being in a particular row k, which we shall label the row effect α_k; (3) a similar effect β_j owing to being in column j; (4) an interaction effect γ_{jk} due to the peculiar combination of the kth row and jth column; and (5) a unique effect, or error term $\varepsilon_{.jk}$ produced by factors not explicitly considered in the equation. The equation then becomes

$$X_{ijk} = \mu + \alpha_k + \beta_j + \gamma_{jk} + \varepsilon_{ijk}$$

which of course refers to population parameters that must be estimated from the sample data. It turns out that if all of the required assumptions for two-way analysis of variance are met, we may obtain unbiased estimators of the parameters in the above equations as follows:

$$\hat{\mu} = \bar{X}... \qquad \hat{\gamma}_{jk} = \bar{X}_{.jk} - \bar{X}_{..k} - \bar{X}_{.j.} + \bar{X}...$$

$$\hat{\alpha}_k = \bar{X}_{..k} - \bar{X}... \qquad = \bar{X}_{.jk} - (\hat{\alpha}_k + \hat{\beta}_j + \hat{\mu})$$

$$\hat{\beta}_j = \bar{X}_{.j.} - \bar{X}... \qquad \hat{\varepsilon}_{ijk} = X_{ijk} - \bar{X}_{.jk}$$

"Each of these estimates makes intuitive sense except, perhaps, for the estimate of the interaction effect $\hat{\gamma}_{jk}$. We use the sample grand mean $\bar{X}...$ to estimate μ and the deviations of the row and column sample means from $\bar{X}...$ to estimate the row and column effects, α_k and β_j, respectively. The deviation of X_{ijk} from the subcategory sample mean $\bar{X}_{.jk}$ of course represents unexplained variation in the sample that estimates the comparable residual term ε_{ijk}. The estimate of the interaction component can then be obtained by subtraction. In effect, then, we have expressed each individual X_{ijk} in terms of the following components: grand mean, row effect, column effect, interaction effect

$$X_{ijk} = \bar{X}... + (\bar{X}_{..k} - \bar{X}...) + (\bar{X}_{.j.} - \bar{X}...) + (X_{ijk} - \bar{X}_{..k} - \bar{X}_{.j.} + \bar{X}...) + (X_{ijk} - \bar{X}_{.jk})."$$

error

One important exception to the case of translating analysis of variance to multiple regression is the case of repeated measures. The case of repeated measures is similar to that of matched cases that we found in t-tests. Sometimes you will have measures on the same person or on related people for several different categories and you want to compare them. I ran into this once with a study comparing lesbian feminists, heterosexual feminists, and heterosexual traditionals on their reports of behavior of their mothers and fathers. The analysis format is illustrated below.

	LF	HF	HT
Mother score	--	--	--
Father score	--	--	--

Note that here while the groups of three women are independently selected, the mothers and fathers within each group are obviously related to each other. In any situation like this (while you could use difference scores of mother and father this is not really recommended because you may mask some interaction effects) you would use what is called repeated measures analysis of variance. This can test for different patterns occurring for the mother and father scores within the three groups (the interaction effect), and differences between the mother and father scores (the row effects) and the differences between the three groups (the column effects). The calculations, however, are different than those described here and you should consult a psychology textbook for further guidance.

While the MEANS program or ONEWAY can be used for one-way analysis of variance, the ANOVA program should be used for two-way analysis of variance. It can also be used for one-way analysis of variance. It is quite versatile and extensive.

Assumptions Underlying Analysis of Variance

The assumptions underlying analysis of variance are essentially identical to those that are associated with the t-test. In fact, the t-test is simply a special case of the F-test. It is the square root of the F-ratio when the first degrees of freedom is one (categories of the independent variable = 2). This may be easily seen by comparing the square roots of the values in the F-table with the corresponding values in the T-table for a two-tail test. Boneau's 1960 article and Lindquists's Design and Analysis of Experiments in Psychology and Education deal with this issue. Lindquist cites an unpublished Ph.D. dissertation done in 1952 that examines the impact of altering variances and population snapes on the F values. Boneau adds the impact of changing sample sizes.

Essentially, these people conclude that if we have both unequal n's and unequal ~~in the~~ population variances we may have problems with inaccurate results. Also, if the populations are skewed, especially if skewed in different ways, and there are unequal n's we may have problems. If, however, the sample size is increased these problems diminish. If the samples are larger than around 25 or 30 we can virtually ignore the assumptions.

I have not yet found an article that specifically deals with violating the assumption of interval measurement of the dependent variable with F ratios. This likely needs to be specifically tested. However, since t is a special case of F, mathematically the results obtained with the t statistic should extend to F. Boneau (1960) cites articles that show that other results regarding t extend to results regarding F.

Packet 160

SOC 412

SOCIOLOGICAL RESEARCH METHODS

Professor Stockard

University of Oregon

Winter Term 1992

Copies:	$1.98
Binding	$0.00
Royalties	$0.00
Permission Handling Charges	$0.00
Total cost of packet:	$1.98

TABLE OF CONTENTS

Jean Stockard - Packet 160

VII.

VII. Correlation and Regression

Quantitative analyses are always linked with research designs. We use statistics to help answer research questions. Sometimes we are interested in comparing the magnitude or size of results in two groups, in which case we would use a t-test, or in three or more groups, in which case we would use analysis of variance. As we saw in the previous section, analysis of variance may be extended to the case where two or more independent variables are used to explain or account for variation in the dependent variable. In this case the independent variables need be measured only on a nominal scale.

What, however, if we were not interested in comparing magnitudes or central tendency, but were interested in the association between two variables, especially in the association between two variables when both the independent and dependent variable were measured on an interval scale. Here we are interested in association or correlation, what happens to the dependent variable when the independent variable changes. For instance, one could be interested in the association between income and education, what happens to peoples' incomes as their levels of education rise. When both the dependent variable 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, the nature of the PRE measure of association, r^2, interpret the correlation coefficient r, give examples of computing both of these values, test hypotheses that the association between two variables is linear and that the correlation coefficient is equal to zero, set confidence intervals around the population conterpart of r, and briefly discuss computer programs used in doing these processes.

test the hypothesis that 2 correlations are equal,

r^2 as a Measure of Association

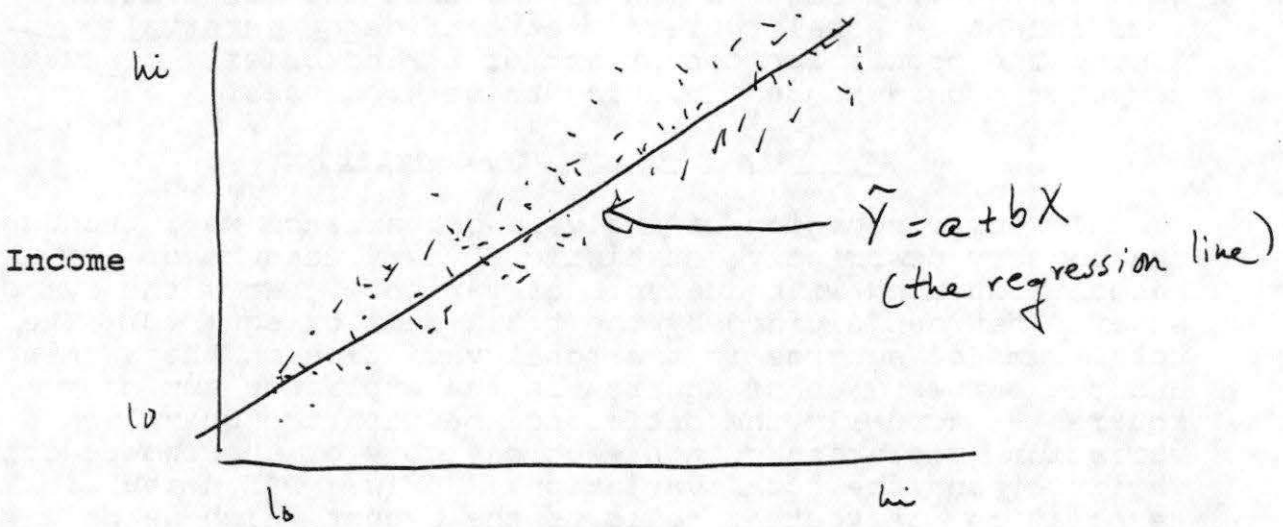
In our discussion of analysis of variance we introduced the summary descriptive statistic E^2 as a measure of association used with analysis of variance. E^2 = the sum of squares between divided by the total sums of squares. The total sums of squares is the total variation in the sample, and the between sum of squares is the explained sum of squares -- actually the difference between the unexplained variation (variation within each category of the independent variable) and the total variation. E^2 is a PRE measure of association, in that it tells us the amount which we have reduced our error in predicting the dependent variable when we knew something about the independent variable. In this case the information we use about the independent variable is its categories, and our best predictor of the dependent

variable when we know this information is the category means. Thus, E^2 tells us how much of the variation of the dependent variable is explained when we know the categories of the independent variable.

E^2 is an extremely useful measure. However, it tells us nothing about the particular nature of the association between the independent and dependent variable, only that when the value is greater than zero the two variables are associated in some way or another. E^2 requires that the independent variable only be measured on a nominal scale. What, however, if you thought there were some pattern in the association between the independent and dependent variable? What if, say, you thought that they had a positive linear association -- as one varible went up, so did the other (as in the example of income and education above). Or, what if there were a negative linear relationship -- as one variable went up the other went down (as in an association between educational level and amount of superstitious beliefs).

Figure 7-1 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. higher

Figure 7-1
Relationship between Income and Education
for a Hypothetical Group of People



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 7-1.

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

$$Y = a + b\,X \qquad (7\text{-}1)$$

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.

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

$$\hat{Y} = a_{yx} + b_{yx}\,X \qquad (7\text{-}2)$$

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, $\Sigma(Y - \hat{Y}) = 0$, and $\Sigma(Y - \hat{Y})^2$ 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 $\Sigma(Y - \bar{Y})^2$ 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

$$E_1 = \sum (Y-\bar{Y})^2 \qquad (7\text{-}3)$$

or the squared deviations of scores around the mean.

For the second rule

$$E_2 = \sum (Y-\hat{Y})^2 \qquad (7\text{-}4)$$

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:

$$\frac{(E_1 - E_2)}{E_1} = \frac{\sum (Y-\bar{Y})^2 - \sum (Y-\hat{Y})^2}{\sum (Y-\bar{Y})^2} = r^2 \qquad (7\text{-}5)$$

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 lineaar association between the dependent and independent variable. This 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 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.

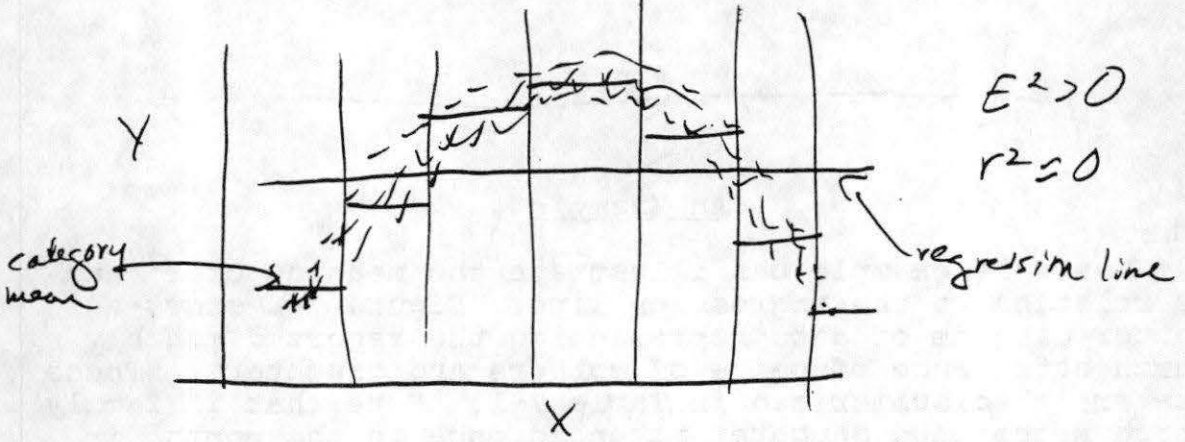
as a proportion of the total variation

Note how this measure is analagous to E^2. Both measures use the same definition of total variation or total error. They differ in how they use information from the independent variable. E^2 simply uses the average variation within each category of the dependent variable. r^2 uses the total pattern of variation or association between the two

variables to develop a straight line that best illustrates this association. Because there is almost always some variation from a straight line association between two variables, E^2 is always bigger than or equal in size to r^2.

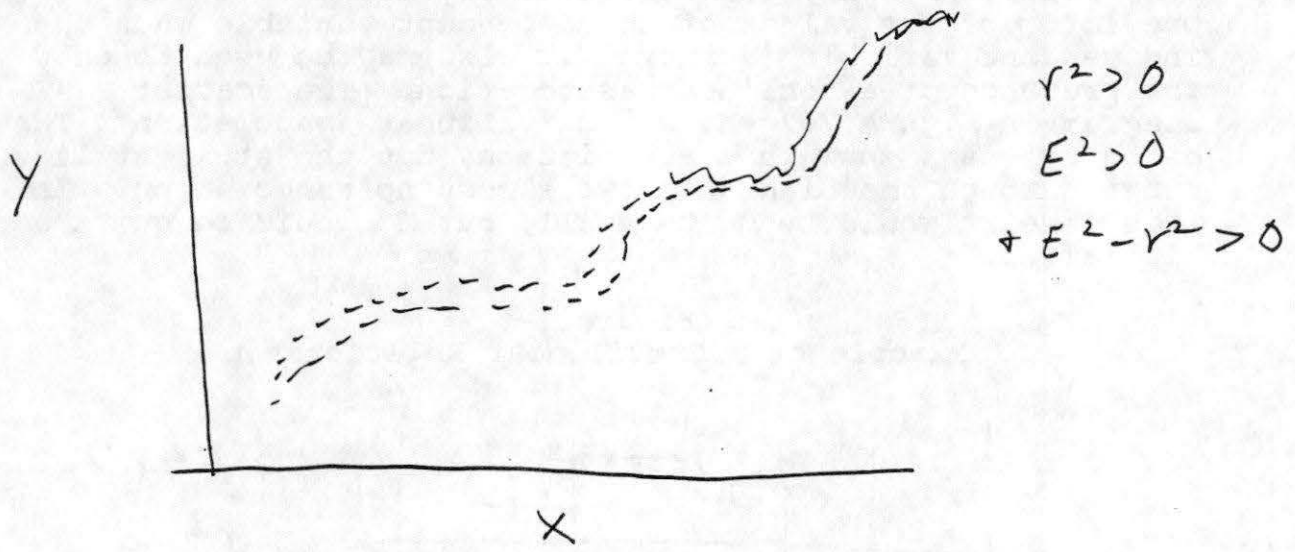
Because E^2 only uses the category means as a best predictor of the values of the dependent variable when the independent variable is known, it also may be used to show the presence of a nonlinear association. The scatter diagram in Figure 7-2 shows a curvilinear association. The category means show this association, but the straight line drawn through the diagram shows almost no association. In this case r^2 would be quite small, but E^2 would be quite a bit larger.

Figure 7-2
Example of a Curvilinear Relationship



Because E^2 and r^2 use the same basic measure of total variation [$\Sigma(Y - \bar{Y})^2$], it is legitimate to compare their values. In fact, $E^2 - r^2$ can indicate the degree to which an association between two variables is not linear. If $E^2 - r^2$ is much greater than zero , then it indicates that the two variables are associated, but that their association is not primarily a linear one. (Later we will test the null hypothesis that the difference between these values in the population is equal to zero.) Note that $E^2 - r^2$ may be greater than zero even if r^2 is also significantly greater than zero. (See Figure 7-3.) In this case, however, the fact that the nonlinear relationship is much larger than the linear one would indicate that it would be the one to pay most attention to.

Figure 7-3
Example of a Situation Where Both E^2 and r^2 are Greater than Zero



An Example

A simple example can illustrate the meaning of r^2 and its relation to the regression line. Figure 7-4 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 7-1. 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.

A scatter diagram, as in Figure 7-2 is a device used to illustrate the nature of the association between two variables. From the scatter diagram in Figure 7-2 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.

Figure 7-4

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

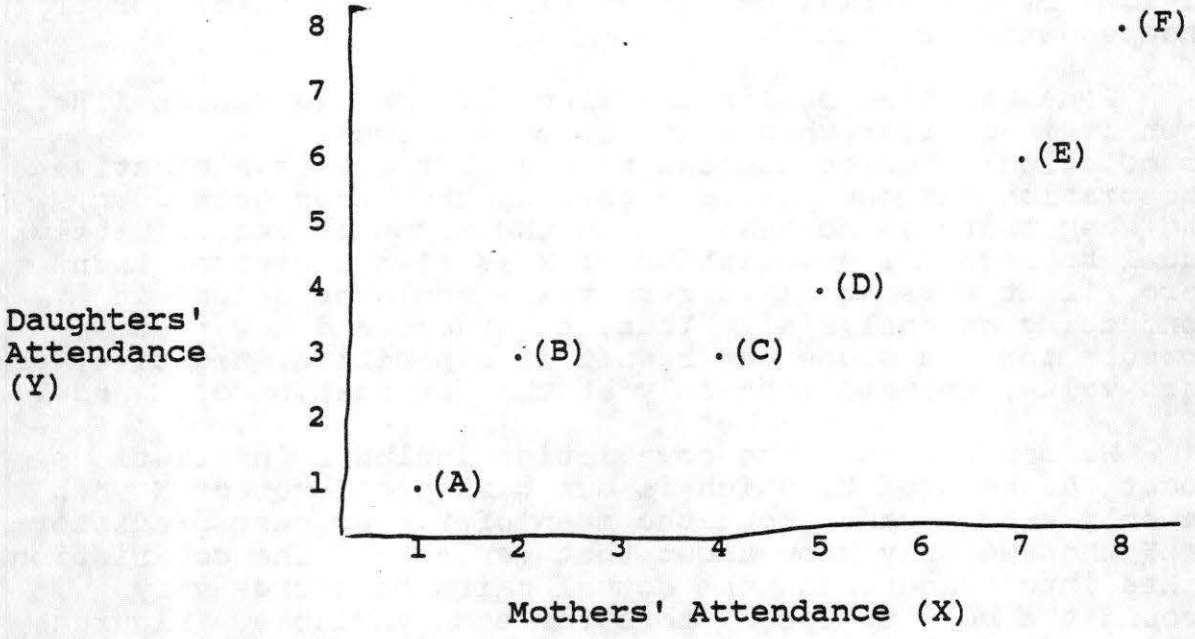


Table 7-1

Computations Needed to Compute r^2 for data in Figure 7-2

Family	Mother (X)	Daughter (Y)	$(X-\bar{X})$	$(X-\bar{X})^2$	$(Y-\bar{Y})$	$(X-\bar{X})$	$(Y-\bar{Y})$
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	
Totals	27	25				32.5	

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

An intuitive explanation of the formula for b is possible.

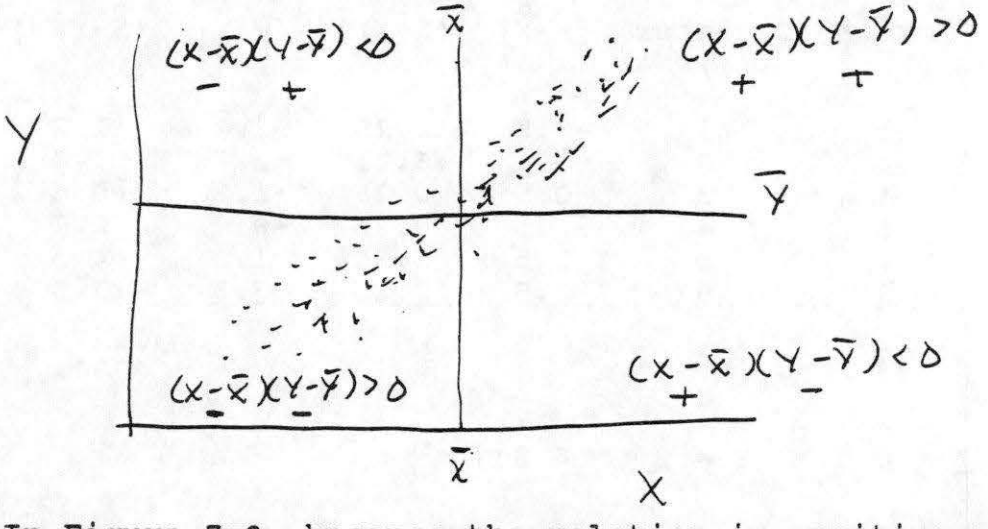
$$b_{yx} = \sum (X - \bar{X})\ (Y - \bar{Y}) / \sum (X - \bar{X})^2 \qquad (7\text{-}4)$$

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 7-3, 7-4, and 7-5 illustrate situations that will result in different values of b.

Figure 7-3
A Positive Value of r and b

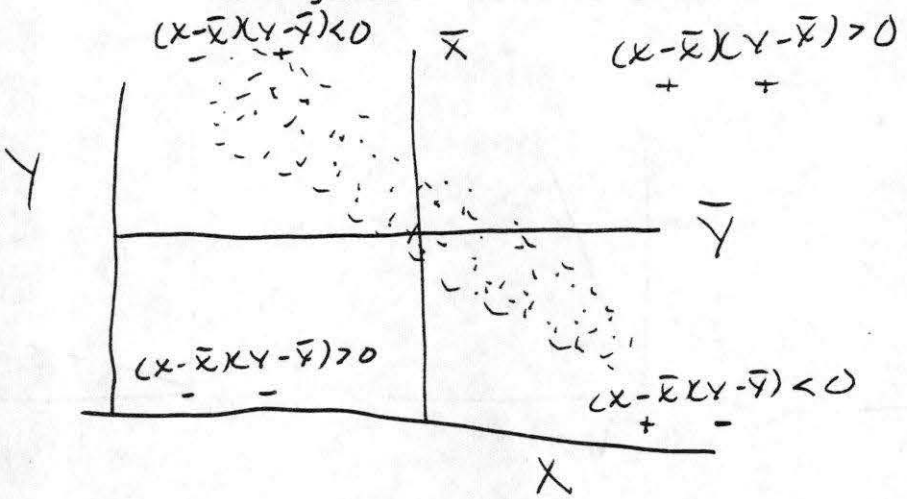


In Figure 7-3, because the relation is positive most of the pairs of scores fall into the quadrant where both Y-Y and X-X are greater than zero, or in the quadrant where both of these values are less than zero. In both these cases the product of $(Y - \bar{Y})$ $(X - \bar{X})$ would be positive (positive times positive = positive; negative times negative = positive) and

thus the covariation would be positive and b or the slope would be positive.

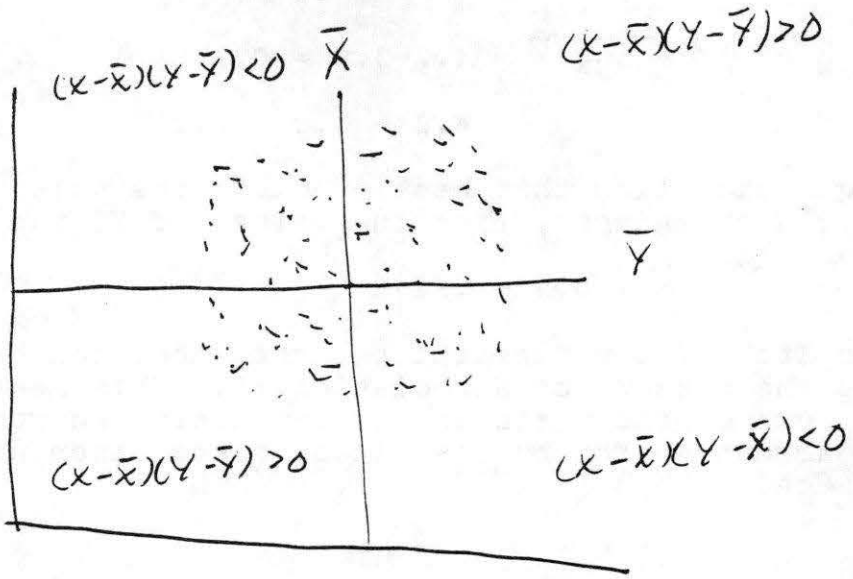
In Figure 7-4 the association is negative. In this case most of the cases fall into the quadrant where $(X - \bar{X})$ is less than zero and $(Y - \bar{Y})$ is greater than zero, or into the quadrant where $(X - \bar{X})$ is greater than zero and $(Y - \bar{Y})$ is less than zero. In this case the product of $(Y - \bar{Y})(X - \bar{X})$ would usually be negative and thus b and the slope would be negative.

Figure 7-4
A Negative Value of r and b



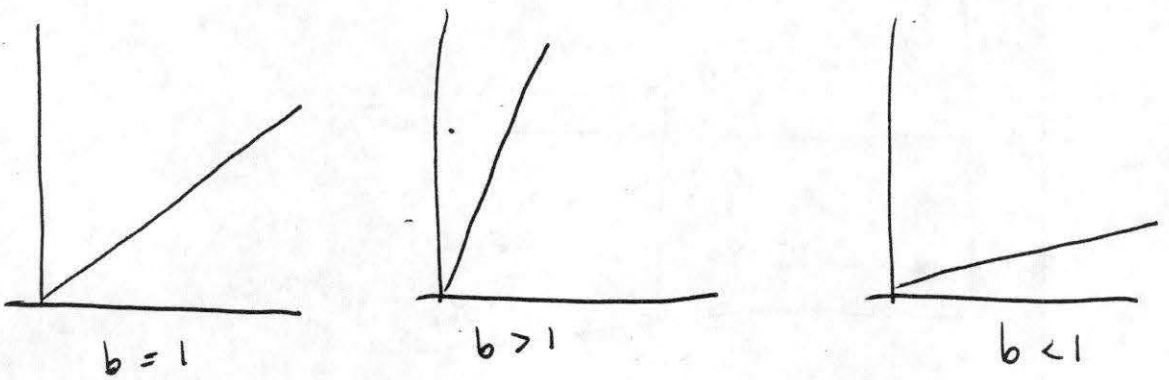
Finally, in Figure 7-5, there is no association. In this case the pairs of cases generally fall equally between the four quadrants. Thus the number of times the product of $(X-\bar{X})$ and $(Y-\bar{Y})$ is positive should about balance off the number of times the product is negative and thus the overall sum of these products over all cases would be close to zero.

Figure 7-5
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 7-6.

Figure 7-6



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 7-1 we can calculate:

$$b_{xy} = \frac{\Sigma(X-\bar{X})\ (Y-\bar{Y})}{\Sigma(X-\bar{X})^2} = \frac{32.5}{37.5} = .87 \qquad (7\text{-}5)$$

$$a_{yx} = \bar{Y} - b_{yx}\,\bar{X} = 4.2-3.9 = 0.3 \qquad (7\text{-}6)$$

$$= 4.2 - 3.9 = 0.3$$

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

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

In Table 7-2 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' attendance.

Table 7-2
Data for Calculating r^2 for data in Figure 7-2

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.

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

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 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

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \qquad (7\text{-}8)$$

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

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

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

The Pearson Product Moment Correlation, r

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

$$r = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma (X - \bar{X})^2 \; \Sigma (Y - \bar{Y})^2}} = \frac{\text{covariation of X and Y}}{\sqrt{(\text{variation of X})(\text{variation of Y})}} \qquad (7\text{-}9)$$

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

$$b_{yx} = \frac{\text{covariation (XY)}}{\text{variation X}} \qquad b_{xy} = \frac{\text{covariation XY}}{\text{variation Y}} \qquad (7\text{-}10)$$

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.

$$r^2 = b_{yx}\, b_{xy} \text{ and thus } r = \sqrt{b_{yx}\, b_{xy}} \qquad (7\text{-}11)$$

The various possible interpretations of r follow from 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 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 7-8 illustrates this interpretation for the example used in the previous section. As shown in Table 7-3, 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, $r = \Sigma(z_y z_x) / N$, 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 7-12 and 7-13.

$$s_{zx}^2 = 1 = \frac{\Sigma(Z_x - \bar{Z}_x)^2}{N} = \frac{\Sigma(Z_x - 0)^2}{N} \qquad (7\text{-}12)$$

and by multiplying N by each side of the equation:

$$N = \Sigma(Z_x - \bar{Z}_x)^2 = \Sigma(Z_x - 0)^2 = \Sigma Z_x^2 \qquad (7\text{-}13)$$

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

$$\left(\sqrt{\Sigma(X-\bar{X})^2 \; \Sigma(Y-Y)^2} = \sqrt{(N)(N)} = (N)\right. \qquad (7\text{-}14)$$

Table 7-3
Calculations of r and r^2 for Data
in Table 7-1 Using Standard Scores

Family	X	Y	Zx	Zy	ZxZy	
A	1	1	-1.3	-1.33	+1.73	$\bar{X} = 4.5$
B	2	3	- .92	- .49	+ .45	$\bar{Y} = 4.17$
C	4	3	- .18	- .49	+ .09	$S_x = 2.7$
D	5	4	+ .18	- .07	- .01	$S_y = 2.38$
E	7	6	+ .92	+ .77	+ .71	$n = 6$
F	8	8	+1.3	+1.61	+2.09	
	27	25	0	0	5.1	

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

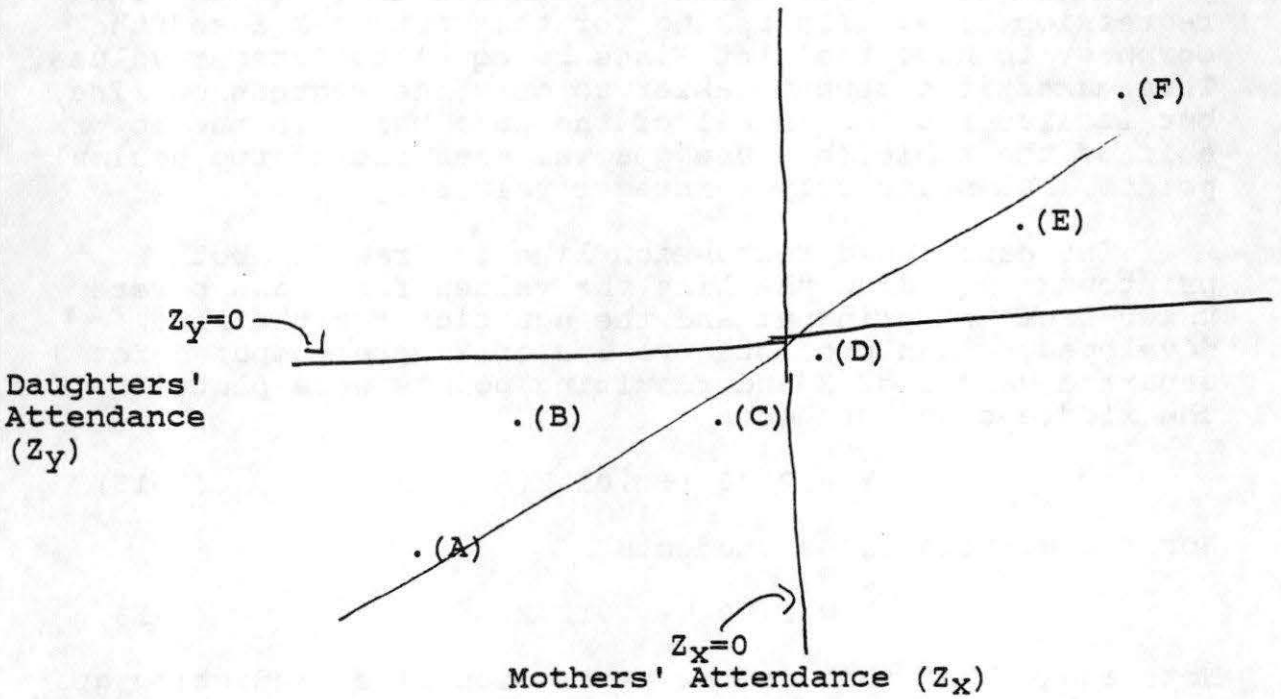
And these results are equal, when rounding errors are taken into account, to those found through other computation methods above

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 7-8
Illustration of r and b when using standard scores with data from Figure 7-2



The term Pearson product moment correlation also comes from the definition of r as $\sum z_x z_y / N$. 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.

Example With

Computer Work

Various computer programs can provide scatter diagrams and compuations of r and r^2. The output shown below in Figures 7-9 and 7-10 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

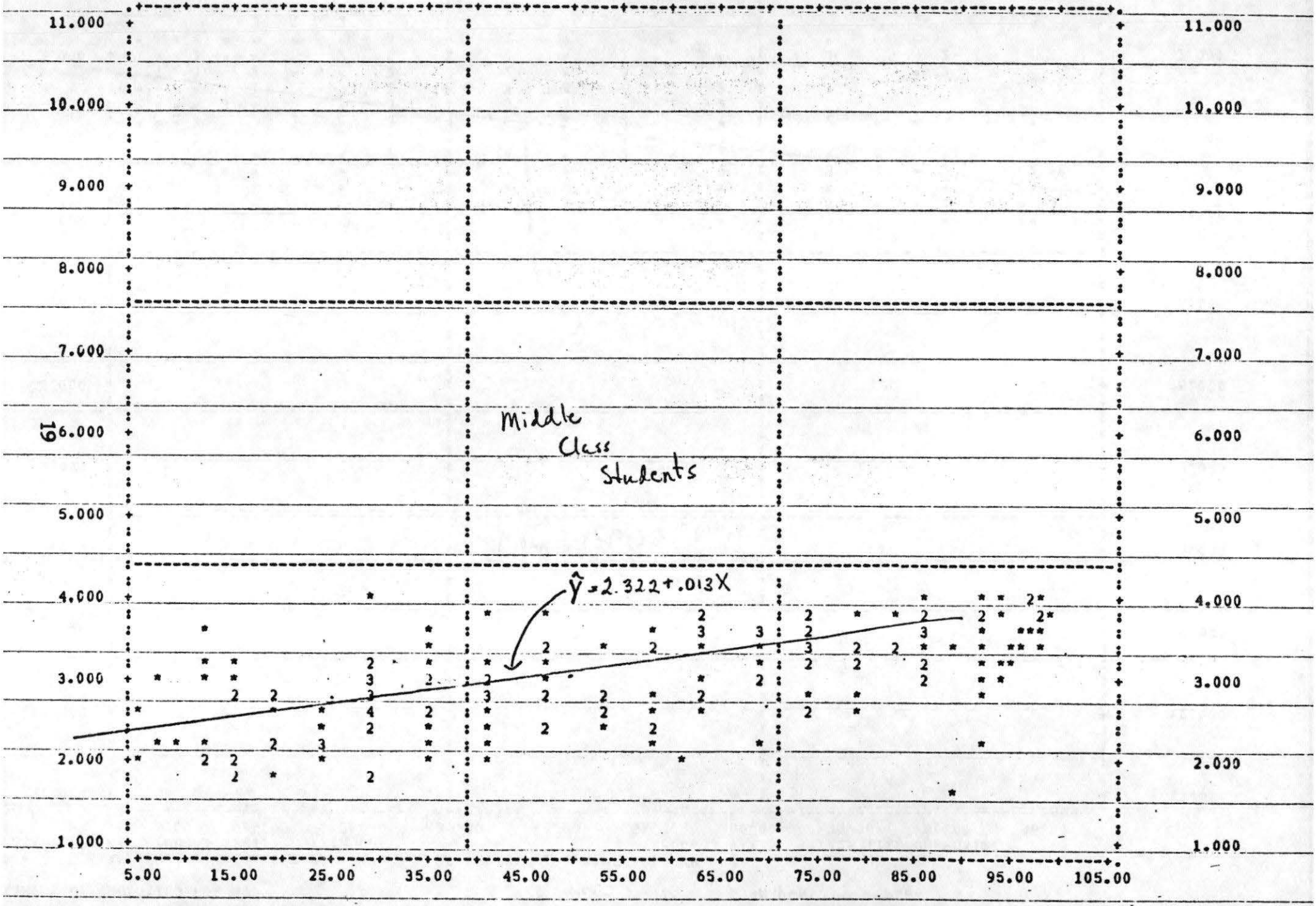
$$\hat{Y} = 2.32 + .013X. \qquad (7\text{-}15)$$

For the working class students

$$\hat{Y} = 2.20 + .012\ X. \qquad (7\text{-}16)$$

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.

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



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

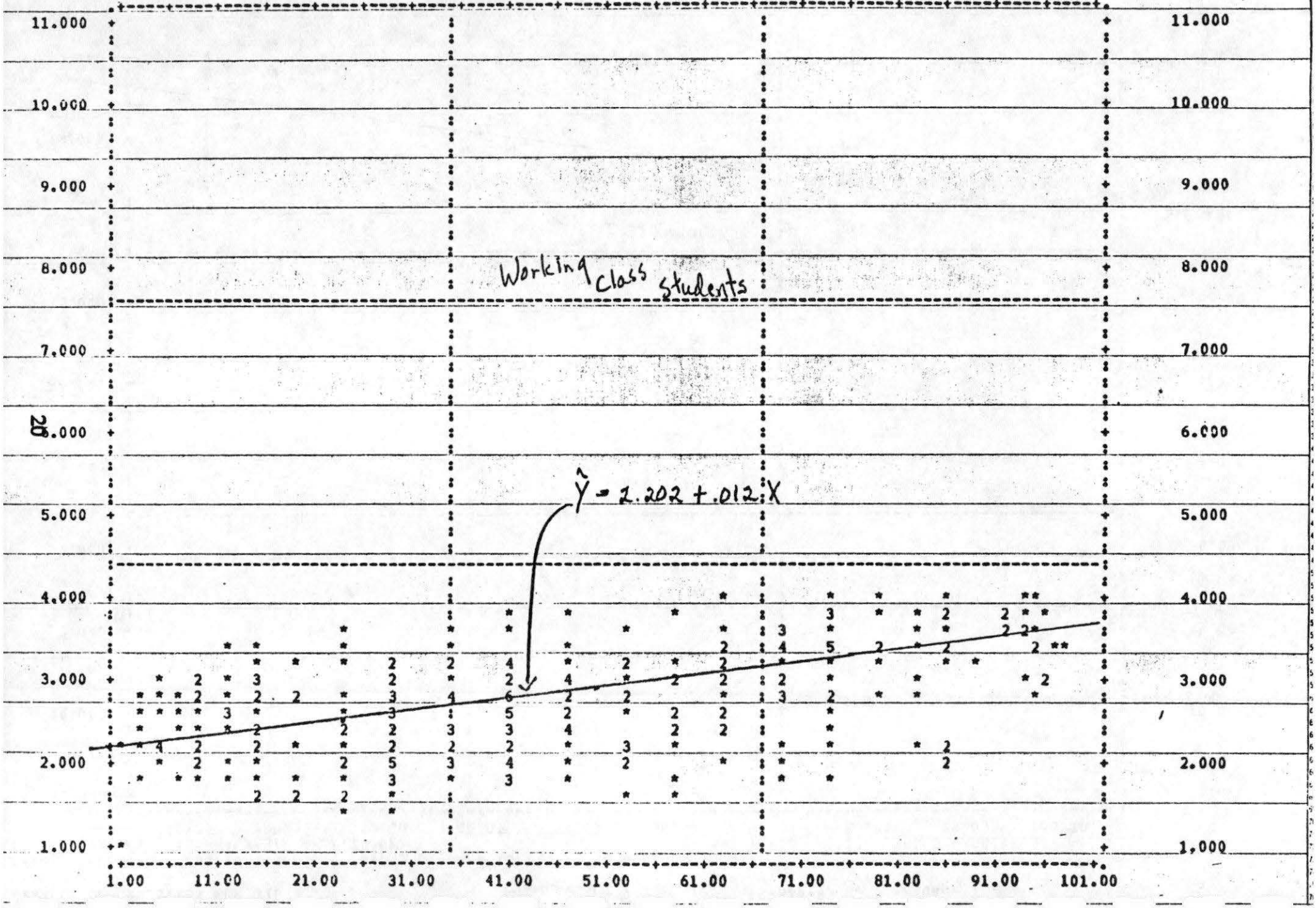


Table 7-4
Calculation of Grades predicted for Middle Class and Working Class Students at Various Levels of Achievement

Achievement Test Scores (Percentiles)	Predicted Grades		
	Middle Class	Working 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 7-4 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.

When researchers report results regarding the correlations between a number of variables they typically use what are called correlation matrices. Examples of correlation matrices were handed out in class. They are an efficient way of showing the relationship between a large number of variables.

Inferential Tests
(and the regression equation)

The r and r^2 values, are ~~both~~ descriptive statistics. r tells us the standard deviation unit change produced in one variable by one standard deviation unit change in the other. The other measure (r^2) is a PRE statistic that tells us what proportion of the variation in one variable is accounted for by its linear association with another. Yet, we also might want to generalize these results to the population from which the sample was drawn. As with any inferential test, we must assume then that the sample used is representative of some larger population. Below we show how to test the null hypothesis that the correlation coefficient in the population is equal to zero, how to test the null hypothesis that the association between the two variables is linear rather than non-linear, how to put confidence intervals around the value of the correlation coefficient, and how to test the null hypothesis that two correlations are equal within the population.

Testing the Null Hypothesis that Rho Equals Zero

To test the hypothesis that within the population the correlation coefficient is equal to zero, that the correlation occurs only by chance, we would need the following hypotheses.

$$H_0: \rho^2 = 0$$

$$H_1: \rho^2 \neq 0$$

Rho (ρ) is the population counterpart to the sample term of r. To complete this test we must assume that the sample was independently and randomly (or representatively) drawn from the given population. We must also assume that X and Y have a bivariate normal distribution. This means that each variable is distributed normally about the other. Blalock has a good three-dimensional representation of this in his text. He describes it as a fireman's hat. One can also think of it as a slightly melted ice-cream sundae. In any case, what happens in a bivariate normal distribution is that if one sliced into it at any point, one would find a normal distribution of one variable about the values of the other.

Implied by the existence of the bivariate normal distribution is the condition of homoscedasticity. This is simply the assumption that for each value of y the values of x in the population have equal variances. Similarly, if we would look at the population at each value of x, the values of y would have the same variance.

Note that all of these assumptions are analagous to those needed for the development of analysis of variance. There we assumed that within each category of the population the values of the dependent variable were normally distributed. Here because both variables are intervally measured (and the measure is symetric) we extend this to normality of both variables. Also, we assumed in analysis of variance that within each category the values of the dependent variable had equal variances. This is like the homoscedastic assumption. Figure 7-4 below illustrates grossly what these assumptions imply with both X and Y given as the dependent variable. Figure 7-5 gives two examples of when homoscedasticity does <u>not</u> exist.

Figure 7-4
An Example of Homoscedasticity

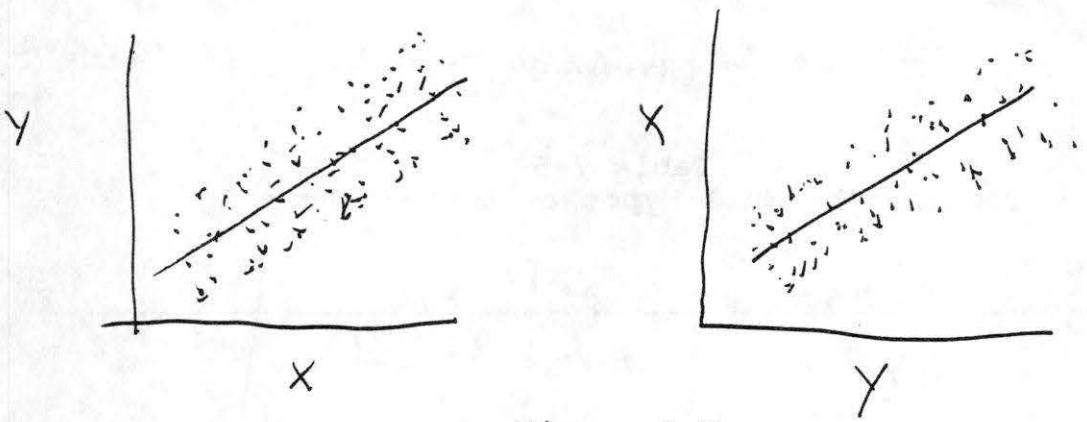


Figure 7-5
An Example of non-homoscedasticity

The test of the hypothesis given above then follows the format used with analysis of variance. Remembering our interpretation of r^2, you will recall that

$$r^2 = \frac{\text{explained SS}}{\text{total SS}}$$

The explained sums of squares refers to the variation that can be explained by the linear association between the two variables. The total sums of squares is written as $\Sigma(Y-\bar{Y})^2$. Blalock calls this Σy^2. From the above then we can see that the explained SS = r^2 $(\Sigma (Y-\bar{Y})^2) = r^2 \Sigma y^2$. These results are summarized in Table 7-5 below. The unexplained sum of squares equals the proportion of variation that is unexplained $(1-r^2)$ times the total sum of squares. This follows directly from

$r^2 = \frac{\text{total SS} - \text{unexplained SS}}{\text{total SS}} = \frac{\text{total SS}}{\text{total SS}} - \frac{\text{unexpl SS}}{\text{total SS}} = 1 - \frac{\text{unexpl SS}}{\text{total SS}}$

$\rightarrow r^2 - 1 = -\frac{\text{unexpl SS}}{\text{total SS}} \rightarrow (r^2-1)(\text{total SS}) = -\text{unexpl SS} \rightarrow (1-r^2)(\text{total SS}) = \text{unexpl SS}$

Table 7-5
Testing the Null Hypothesis that rho = 0

Source of Variation	Sum of Squares	df	MSS	F
Total	$\Sigma(Y-\bar{Y})^2$	$n-1$	$\frac{\Sigma(Y-\bar{Y})^2}{n-1}$	
Explained	$r^2\Sigma(Y-\bar{Y})^2$	1	$\frac{r^2\Sigma(Y-\bar{Y})^2}{1}$	$\frac{r^2\Sigma(Y-\bar{Y})^2(n-2)}{(1-r^2)\Sigma(Y-\bar{Y})^2}$
Unexplained	$(1-r^2)\Sigma(Y-\bar{Y})^2$	$n-2$	$\frac{(1-r^2)\Sigma(Y-\bar{Y})^2}{n-2}$	$= \frac{r^2(n-2)}{1-r^2}$

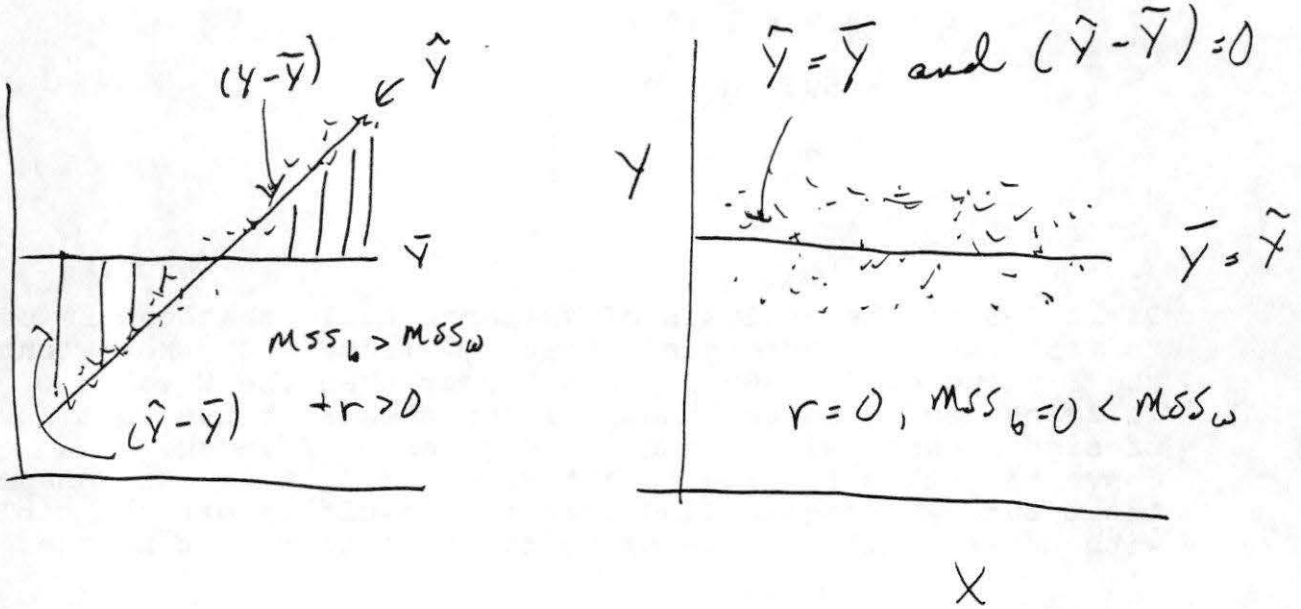
What are the degrees of freedom associated with these values? As with simple analysis of variance, we only lose one degree of freedom with the total variation. This comes from the computation of the mean. Thus, the degrees of freedom associated with the total variation is N-1. Recalling the explanation of r^2, the unexplained variation = $\Sigma(Y-\hat{Y})^2$, the squared deviations of scores around the regression line. To make the regression line we need to know two

points. (Two points are needed to draw any line; we also need to calculate two values: a and b.) Thus, for this sum of squares we lose two degrees of freedom. Finally, since the total degrees of freedom must equal that for the explained variation plus that for the unexplained variation, the degrees of freedom for the explained variation must equal $df_{total} - df_{unexplained} = (N - 1) - (N - 2) = -1 + 2 = +1$

To get the mean sum of squares, the sums of squares are simply divided by the degrees of freedom. These give us the two estimates of the common variance. As with analysis of variance, only the estimate from the unexplained sum of squares is always unbiased. The estimate based on the between sums of squares is valid only if the null hypothesis (that the correlation coefficient in the population equals zero) is true.

Remember that the unexplained sums of squares is the variation of scores around the regression line $[(\Sigma Y - \hat{Y})^2]$. The explained sums of squares is the variation of scores between the regression line and the overall mean, or the squared deviations of the predicted values of Y from the mean of Y $[(\Sigma \hat{Y} - \bar{Y})^2]$. The estimate of variance from the unexplained sums of squares is always a good estimate of the common variance (the variance around the regression line), analagous to the mean sums of squares within in analysis of variance. The estimate of variance from the explained sums of squares is only a good estimate when the correlation coefficient equals zero, or when the regression line essentially lies on the mean and the two variables are unassociated with each other. The sketches in Figure 7-6 below illustrate these two possibilities.

Figure 7-6



The two estimates of the common variance are compared in the F-ratio, and the resulting value may be compared to the critical values in the F-table in the end of the book. Note that in the F-ratio as shown in Table 7-5, the sums of squares cancel each other out so that the F ratio = $r^2(N-2) / (1-r^2)$. Thus, the F-ratio is simply a function of the size of the correlation coefficient and the sample size. It is simple to see that as the sample gets larger and as the correlation gets larger, the F value will also rise. This should make intuitive sense. We would be more likely to expect a correlation different from zero in the population if we had a larger sample value and if we had more cases in our sample.

Note also that since the degrees of freedom equal (1, N-2) that we may use the square root of this value as equal to the t-statistic. Some text books use the t-statistic in testing the hypothesis that the correlation coefficient in the population is equal to zero. The results from the two methods are equal. The advantage to using the t-distribution is that you can have a one-tail test rather than a two-tail test. In doing such a test, t simply equals the square root of the F-ratio defined above.

$$t = (r\sqrt{n-2}) / \sqrt{1-r^2}$$

Note again, that as the value of r becomes larger and the sample size becomes larger it is easier to reject the null hypothesis. The degrees of freedom equal n-2.

As an example of these tests, consider the data displayed in the scatter diagram in Figure 7-7. These data describe the relationship between delinquency rates and average monthly rental costs in a variety of housing areas. Calculations of the regression line, r, r^2, and E^2 produced the following results:

$Y = 68.9 - 1.19 X$

$r = -.59, \quad n=140$

$r^2 = .35$

$E^2 = .37$

Table 7-6 is the analysis of variance table that can be used to test the null hypothesis that ρ^2 (rho)$^2 = 0$. Note that the F-value of 73.6889 is much larger than the F value of 11.38 needed to reject the null hypothesis at the .001 level of significance with 1 and 138 degrees of freedom. Converting this F-value to a t-value of 8.58 we can consult the t-table and again find that the result is off the table with a very small chance of being wrong in rejecting the

null hypothesis that there is no association between delinquency rates and average monthly rentals.

Figure 7-7
Scatterplot of Data Showing the Relationship between Delinquency Rates (Y) and Monthly Rental Price (X)

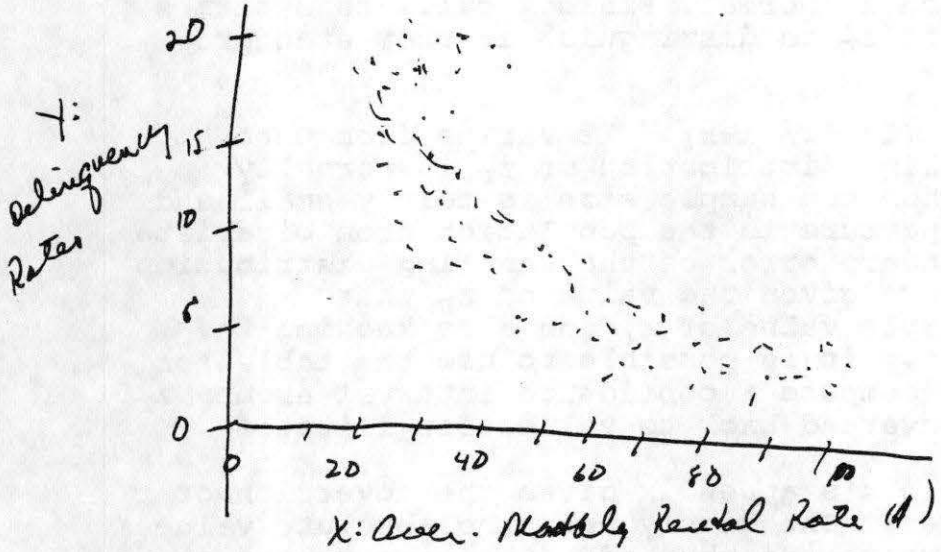


Table 7-6

Source of Variation	Sums of Squares	df	Mean Sums of Squares	F
Total	$\Sigma(Y-\bar{Y})^2$	$n-1$		
Explained	$r^2\Sigma(Y-\bar{Y})^2$	1	$\frac{r^2\Sigma(Y-\bar{Y})^2}{1}$	$\frac{r^2(N-2)}{1-r^2}$ $= \frac{.35(138)}{.65}$ $= 73.69$
Unexplained	$(1-r^2)\Sigma(Y-\bar{Y})^2$	$n-2$	$\frac{(1-r^2)\Sigma(Y-\bar{Y})^2}{n-2}$	

Note also that a test of the hypothesis that $\rho^2 = 0$ (and $\rho = 0$) is equivalent to the test that in the population the slope is equal to zero, for both of these terms have the same numerator. $(r+b)$

Confidence Intervals Around Rho

You may also want to set confidence limits around the value of rho. In other words, you might want to find what would be a reasonable expectation of the size of the correlation coefficient in the population. The sampling distribution of correlation coefficients is not normally distributed, nor symetrical, so we cannot use the r directly, but we can use a transformation of r whose sampling distribution is normal. Blalock calls this term z. I will try to call it z_r to distinguish it from standard scores.

z_r = 1.151 log (1 + r/ 1-r). It varies from 0 to infinity. The sampling distribution of z_r is normally distributed, even when the sample size is fairly small and if there is some departure in the population from bivariate normality. The standard error of the sampling distribution of $z_r = 1/\sqrt{N-3}$. Now, given the value of z_r that corresponds to a sample value of r, and also knowing the standard error for z_r, it is possible to use the table for the normal curve to compute a confidence interval around z_r. This can then be converted back to values for r itself.

Table K in Blalock's appendix gives the coversion of r values into z_r values. One simply uses the absolute value of r to get the z_r value and then adds the sign to the z_r value.

For our example r = -.59, z_r = -.6777

The standard error = $1 / \sqrt{N-3} = 1 / \sqrt{137} = 1 / 11.705 =$ .0854.

Then, following the standard format we used earlier for computing a 95% confidence interval:

$$P[z_r - (1.96)(S_{zr}) < Z_r < Z_r + (1.96)(S_{zr})] = .95$$

$$P[-.6777 - (1.96)(.08541) < Z_R < -.6777 + (1.96)(.0854)] = .95$$

$$P[-.8452 < Z_r < -.5103] = .95$$

We may now use the Table K to convert these z_r values back into actual correlation coefficients.

$$P[-.69 < \rho < -.47] = .95$$

We can be 95% confident that in the population the correlation between delinquency rate and rentals is between -.69 and -.47.

Testing the Null Hypothesis that Two Correlations are Equal

If one has correlations between two variables from two independent samples one may be interested in testing the hypothesis that these two correlations are equal. This is analogous to the situation where we tested the null hypothesis that the means from two samples were equal, and the procedure to test this hypothesis is very similar to that situation. In this case

$H_0: \rho_1 - \rho_2 = 0$

$H_1: \rho_1 - \rho_2 \neq 0, < 0, > 0$

To conduct a test of this hypothesis one simply transforms the two r's or correlations into z_r's using Table K, as shown above. Then one uses a formula for the standard error for the difference between the two z_r's, which is simply

$$s_{z_{r_1} - z_{r_2}} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

Note that this is equal to the square root of the sum of the standard errors for each r alone and is analagous to the formula for the standard error of the difference between two means.

One can then simply use a familiar z-score to test the null hypothesis, with

$$z = \frac{(z_{r_1} - z_{r_2}) - 0}{s_{z_{r_1} - z_{r_2}}}$$

Zero is in the second part of the formula to represent the value of the difference between the correlations in the null hypothesis. Note how this is exactly analagous to our earlier use of the z and t scores in testing hypotheses about means, with the difference between the sample value and the hypothesized value of the mean in the numerator and the standard error in the denominator.

As an example, suppose that one went to a second community than the one which provided the data in Figure 7-5 and found that the correlation between delinquency rates and monthly rents was only -.35. The sample size in the second

city was 150. Suppose that we expected the correlation in the population from which the sample for the first city was taken would actually be a greater negative value than the correlation for the population from which the data for the second city was taken. In this case, we would have the following null hypothesis of no difference between the two correlations and a one-tail or directional alternative hypothesis. Note that the direction of the one-tail alternative indicates that the first correlation will be smaller (more below zero) than the second.

H_0: $\rho_1 - \rho_2 = 0$

H_1: $\rho_1 - \rho_2 < 0$

Using the formulas given above, we can calculate the standard error as

$$s_{z_{r_1}-z_{r_2}} = \sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}} = \sqrt{\frac{1}{137} + \frac{1}{147}} = \sqrt{.0073 + .0068} = \sqrt{.0141} = .12$$

Consulting Table K in the appendix to Blalock, we may find that the z_r corresponding to r = -.35 is -.3654. Remember that the z_r corresponding to the r for the first sample of -.59 is -.6777.

We can then substitute the values into the formula for the z- ratio as follows

$$z = \frac{z_{r_1} - z_{r_2}}{s_{z_{r_1}-z_{r_2}}} = \frac{-.6777 - .3654}{.12} = \frac{-.3123}{.12} = -2.60$$

Consulting the table for the normal curve we can see that the resulting z value of -2.60 allows us to reject the null hypothesis that the two correlations are different, in favor of the alternative that the one for the first community is more negative, with the chance of being wrong only 47 times out of 10,000. Our theoretical views were supported. (Note that different procedures are involved if you are testing hypotheses about correlations from dependent samples.)

Testing the Null Hypothesis that the Association is Linear

Finally, we may want to test the hypothesis that the association itself is linear. In other words, it may be possible that two variables are associated, but that the pattern of association is not best represented by a straight line. We noted earlier the comparability of E^2 and r^2 and that the difference between these two values can be

used to represent the extent to which the association departs from a linear one. In essence, $E^2 - r^2$ represents the proportion of variation in the dependent variable that may be explained by the independent variable, but cannot be explained by a linear association. Table 7-7 summarizes this test.

H_o: $E^2 - \rho^2 = 0$
H_1: $E^2 - \rho^2 \neq 0$

Table 7-7

Source of Variation	Sums of Squares	df	Mean Sums of Squares	F
Total	$\Sigma(Y-\bar{Y})^2 = \Sigma y^2$	$n-1$		
Explained by linear relation	$r^2 \Sigma y^2$	1		$\frac{(E^2-r^2)\Sigma y^2 (N-k)}{(k-2)(1-E^2)\Sigma y^2}$
Additional explained by nonlinear relation	$(E^2-r^2)\Sigma y^2$	$k-2$	$\frac{(E^2-r^2)\Sigma y^2}{k-2}$	$=\frac{(E^2-r^2)(N-k)}{(1-E^2)(k-2)}$
Unexplained	$(1-E^2)\Sigma y^2$	$N-k$	$\frac{(1-E^2)\Sigma y^2}{N-k}$	

Again note that the F-ratio includes only the measures of association. The sums of squares cancel out. In our example the F-ratio equals:

and degrees of freedom

$$\frac{(E^2-r^2)(n-k)}{(1-E^2)(k-2)} = \frac{(.37-.35)(40-9)}{(9-2)(1-.37)} = .43$$

check

$$= \frac{.62}{4.55} = .14$$

Obviously, this ratio is less than unity and we may fail to reject the null hypothesis that the relation is not linear. In other words, the best representation of the association between delinquency rates and monthly rents is the regression line, and our earlier test of the hypothesis that rho equals zero assured us that this relationship was non-zero, with only a small chance of being wrong in this conclusion.

If you did reject the null hypothesis that the association was linear, you could then test the hypothesis that E^2 in the population equaled zero. This is equivalent to the analysis of variance test.

Note, also, that as with all tests the F-ratios used here are very much affected by sample sizes and that you should look at the descriptive statistics of the amount of variation explained as well as the tests of hypotheses about the nature of the results in the population. If you have a large sample, a very small r will be significantly different from zero. Conversely, with a small sample, a large r will sometimes not be significant.

Computer Work

Several programs may be used with SPSS studentware to obtain correlation coefficients. The CORRELATION program is probably the best to use if you are interested in getting a correlation matrix. It, however, does not provide a test for linearity nor the regression equation. The MEANS program should be used when one wants a test of linearity, for it provides r^2, E^2, and the analysis of variance results for the difference between the category means and test for linearity. The PLOT program provides the scatter diagram, the regression equation, and the correlation coefficient. Note that with both the Means and Plot programs you must be very careful in how you designate your dependent and independent variables. You must make sure these designations match your hypotheses. While r is a symmetric measure, E^2, "b," and "a" generally have different values depending on which variable is designated as dependent.

I know of no way to have the SPSS program do the calculations for the confidence interval around r or to test hypotheses about differences between correlation coefficients. You can easily do this by hand.

Packet 37

Soc 413/513

SOCIOLOGICAL RESEARCH METHODS

Professor Stockard

University of Oregon

Spring Term 1992

UP
75 L

kinko's

the copy center

860 E. 13th

Eugene • 344-7894

Copies:	$9.34
Binding	$0.00
Royalties	$0.00
Permission Handling Charges	$0.00
Total cost of packet:	$9.34

TABLE OF CONTENTS

PACKET 37 Jean Stockard
Sociology 413/513

VIII: MUltiple Regression

In this section we discuss multiple regression including tecnniques of partialling or partial correlation. In the first part we discuss the logical reasons that multiple regression tecnniques are used, referring back to a discussion early in the term about the nature of experimental designs and causal inferences in sociological research. In the second section we develop an intuitive notion of the various measures. In the third section we work through an example, computing the descriptive statistics and the associated inferential statistics. In the final section we discuss the computer techniques that are used to get these measures. In the next and last section for this term (Ch. IX) we discuss the general linear model as an extension of multiple regression and show how it can incorporate not only the techniques in this and the previous unit but also analysis of variance.

Causal Inferences in Sociology

Early in this term we discussed the nature of experimental designs and how we rarely approach or use these designs in sociology, for both practical and ethical reasons. In one sense this makes conclusions about the nature of social phenomena difficult, for a controlled experimental situation (with replications) is the only way in which we can ever conclusively show the existence of a causal relationship. This is because, only that situation can ensure the presence of a known time order, the nature of the covariation, and rule out other causal variables. (See earlier discussion for details here.) With demographic, survey, observational data, infact almost all data used in the social sciences except for those from experiments, we cannot easily assess the nature of time order. Usually we must infer this, either logically or on the basis of observations. In most cases, this is not too large a problem, and in many substantive areas the logical explanation of time sequence is not highly arguable. In other instances, this may be more difficult.

In almost all cases, however, the problem of ruling out other causal variables is more difficult. We can observe that there is an association between two variables (that covariation exists) and we may be even able to say which variable preceded the other, but we cannot assess with any accuracy (outside of the controlled experimental situation) if there are other variables that might be influencing this assocation.

While the techniques associated with multiple regression cannot solve this problem, they can, with the aid of the sociologists' imagination, help to rule out the possibility of other possible causal variables. This can be done through techniques using multiple regression equations (simply an extension of the bivariate regression equation used in VII -- with more than one predictor variable) or with partial correlation coefficients (an extension of the simple correlation coefficient - but one that takes into account and removes the influence of one or more other variables). Below we will show how partial correlation coefficients can be used to help rule out the impact of other causal variables when examining causal relationships. The example may also be used to illustrate the importance of one's theoretical position in establishing the nature of causal relations. Later we will illustrate how multiple regression equations can serve the same purpose.

Suppose one were interested in the association between marijauana smoking (called variable X) and heroin use (variable Y). Let's say also that one theorized that although a computation of r_{xy} showed that it had a value that was greater than zero, this correlation was simply due to an association of both marijuana use (X) and herein use (Y) with the amount of exposure people had to drugs (Z). This situation is illustrated in Figure 8-1 below.

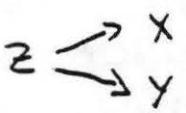


Figure 8-1

From this model we would predict that the association between x and y was only due to the association of both these variables with the prior variable Z. Thus $r_{xy.z} = 0$, or the partial correlation coefficient between X and Y, controlling for Z equals zero. This means that the correlation between X and Y when the impact of Z is removed is non existent. Note that this model also assumes that the correlation between z a nd x and between z and y are unequal to zero ($r_{zx} \neq 0$ and $r_{zy} \neq 0$).

To illustrate the importance of theory in establishing the nature of the causal relationships note that the situations in both figures 8-2 and 8-3 would have the same prediction and assumption equations. However, in figure 8-2 it is suggested that smoking marijuana induces people to have a need for drugs which then leads them to use herein. Figure 8-3 suggests that smoking heroin leads to a need for drugs which then influences the use of marijuana. While figure 8-3 may not often be posited, figures 8-2 and 8-1 are and it might be that in an area with conflicts such as this survey or doemgraphic data may not be what one wants to use.

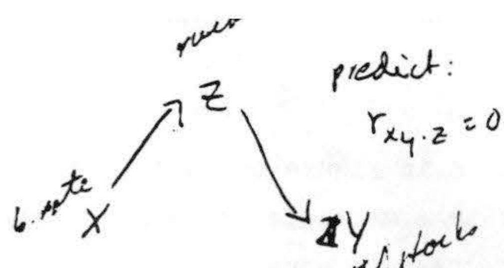


Figure 8-2

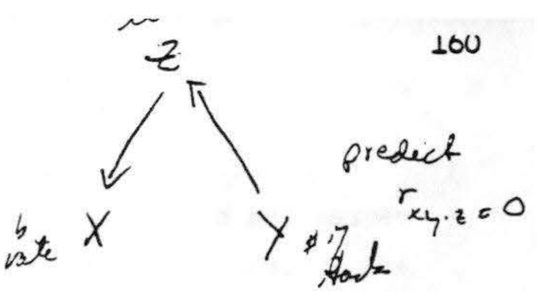


Figure 8-3

In many areas the arguments ~~mf~~ over time priority may not be as ~~mixxxxix~~ hähsh, or it may be so difficult to gain experimental data that one must resort to logical conclusions about time order and to techniques such as those briefly discussed above to rule out other causal variables. In figure 8-4 below an association between four different variables is shown. X and Y are hypothesized as influencing Z and only Z is seen as influencing W. The usual practice is to have the time order go from left to right. Note that in this situation X and Y are seen as occuring simultaneously, Z occurs next, and is followed by W.

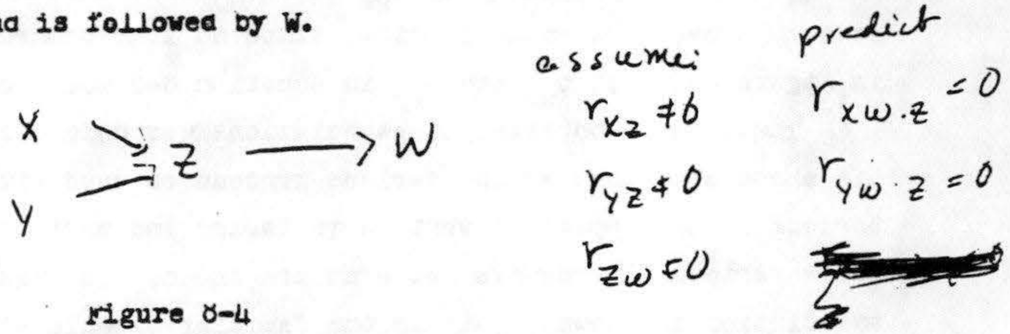


Figure 8-4

From this model the assumption equations and the prediction equations listed above can be determined. Assumption~~x~~ equations are those that represent the associations theorized to be unequal to zero, in this case between X andY and Z and between Z and W. The prediction equations are those that are predicted to be zero. In this case this involves the partial correlation coefficients ~~between X and Y and W when the impact of Z is ruled out and also~~ between ~~Zx~~ X and W when Z and Y are removed and between Y and W when Z and X are ~~m~~ removed. In general, in models such as this, when there is no direct link between two variables, the corresponding partial ~~mpmx~~ coefficient is predicted to equal zero. When there is a direct link it is assumed to be unequal to zero. Both prediction and assumption equations will have to be tested with empirical ~~mx~~ data, however, to see if they are accurate. With a model with more than three variables, one should be careful not to control for dependent variables. for this makes no sense in terms of the causal hypothesis

One could also analyze the situation in Figure 8-4 with a multiple regression equation. The coefficients in a multiple regression equation are analgous to those in a bivariate regression equation discussed in VII except that because more than one causal variable is used the predicted values will not form a line (but a plane or hyperplane) and the slopes which indicate the expected change in the dependent variable with each unit change in the independent variables are computed with the impact of the other variables in the equation removed (we will discuss this all in more detail later). Thus we could have the following two equations for the model in Figure 8-4

$$Z = a_z + b_{zx \cdot y} X + b_{zy \cdot x} Y \qquad (8\text{-}1)$$

$$W = a_w + b_{wz \cdot xy} Z + b_{wx \cdot zy} X + b_{wy \cdot zx} Y \qquad (8\text{-}2)$$

From Figure 8-4 we would predict that b_{zx} and b_{zy} in equation 8-1 would be unequal to zero and that b_{wz} in equation 8-2 would also be unequal to zero. However, we would predict, since no line connects X and Y to W in Figure 8-4 that b_{wx} and b_{wy} in equation 8-2 would both equal zero.

The term elaboration of associations was coined by Lazarfeld and his associates to describe various procedures used with contingency tables to look at the impact of various variables and various patterns of causation. Below various alternative patterns are shown. In Figure 8-5 a spurious association is shown. This is the familiar example of the association between the dollar loss at a fire (Y) and the number of firemen there (X) both of which are actually caused by the prior variable Z, the size of the fire. In this case if the prediction equation $r_{xy \cdot z} = 0$ holds true, then we would say that the association between dollars lost and # of firemen is spurious.

predict $r_{xy \cdot z} = 0$

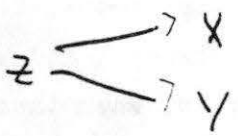


Figure 8-5

Three different patterns of independent effects are also possible. These are shown in figures 6-6, 6-7 and 6-8. In these figures X represents SES, T represent interest in politics, and Y is likelihood to vote. This example is taken from work by Lazarfeld. In figure 6-6, T is an intervening variable to X and Y with the idea that people with higher SES are more likely to be interested in politics and therefore more likely to vote. It is predicted that $r_{xy.t} = 0$. An alternative explantion is in figure 6-7, which predicts that both X and T influence Y. Here none of the partials are predicted to equal zero and both X and T influence Y although X is also seen as influencing T. Finally, in fig ure 6-8 both X and T are seen as independently influencing Y, but as doing so independently of eachother with X not necessarily influencing T. If one still saw X as preceding T in time, one would predict that $r_{xt.y} = 0$. If one first positied the situation in 6-6 and found that it didn't hold then one would generally turn to one of the two models in figures 6-7 or 6-8.

assume: $r_{XY} \neq 0$

$r_{TY} \neq 0$

predict: $r_{XY.T} = 0$

X -> T -> Y

Figure 6-6

assume: $r_{XY} \neq 0$

$r_{TY} \neq 0$

$r_{XY.T} \neq 0$

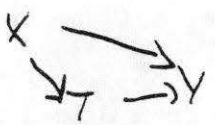


Figure 6-7

assume: $r_{XY} \neq 0$

$r_{TY} \neq 0$

if X+T occur at same time need not

Figure 6-8

Finally, there may be suppressor variables. In other words, when an association equals zero, this may not be the whole story and if the influence of a third variable is introduced the association will rise. This is a situation that is sometimes hard to predict and a place where one's theory is especially important. Sometimes even the association may go in one direction with some values of the control variable and in other directions with other values. This is very hard to detect without contingency tables, and if such a situation is suspected, great care should be taken.

Measures of Association and other Measures in Multiple Regression

The work in section VII involved only zero order correlations or bivariate measures of association. When there is a linear association (or when we think there may be one) between two variables we learned to compute a regression equation for a line that would best represent the association between the two variables. This line took the form

$$\hat{Y} = a_{yx} + b_{yx} X \qquad (8\text{-}3)$$

where we predict Y from X. (The equation could also be formulated predicting X from Y, but the coefficients a and b would be different. Convention usually holds that Y is designated the dependent variable.) By definition this line and the values $\hat{Y}$ are closer to each of the actual values of Y than any other straight line could be. $\Sigma(Y - \hat{Y}) = 0$ and $\Sigma(Y - \hat{Y})^2 = \text{min.}$ Based on this knowledge, the measure r^2 as a measure of association is defined as

$$r^2 = \frac{\Sigma(Y-\bar{Y})^2 - \Sigma(Y-\hat{Y})^2}{\Sigma(Y-\bar{Y})^2} = \frac{SS_{tot} - SS_{res}}{SS_{tot}} \qquad (8\text{-}4)$$

This may be interpreted as the proportion of total variation in one variable that may be explained by its linear association with the other variable. (Note that r^2 is symmetric, its value is the same whether X or Y is designated as the dependent variable. The formula in (8-4) lists Y as the dependent variable, but the measure could also be written with X as dependent.)

The measure r^2 varies from 0 to 1 and has a PRE interpretation. The measure r is also used. It varies from -1 to +1 and may be interpreted as the slope of the regression equation when standard scores of X and Y are plotted. In other words, r gives the expected change in standard deviation units of one variable when the other variable changes exactly one standard deviation unit.

If the measure r and the measure r^2 are to be really accurate, we must assume that other variables that might be having an effect on the dependent variable in equation (8-3) are uncorrelated with the dependent variable X. For instance, if we were using the number of the firemen at a fire as variable X to predict the dollars lost at a fire (Y), we would have to assume that if we were to accept the resulting correlation coefficient as valid that there were no other variables that had an effect on Y that were also related to X. Obviously, however, that isn't so. Another variable, the size of the fire, is related to both the dollars lost and to the number

of firemen at the fire. The correlation that would result from simply looking at the relation between X, dollars lost, and Y, number of firemen, would be inaccurate because this other variable was actually associated with both variables in the equation.

In more exact, or statistical, language, the above paragraph deals with the assumption that within the population each score Y_i of the dependent variable may be represented as

$$Y_i = \alpha_i + \beta_{YX} X_i + \varepsilon_i \qquad (8\text{-}5)$$

where α is the population counterpart of a, the Y-intercept of the regression line; β_{YX} is the slope of the regression line in the population, and ε_i is a measure of how far away each actual score Y is from the score that would be predicted from the regression equation. This error has two parts. One component comes simply from measurement error -- how far off our measurement of Y is from the actual value of Y. Generally, we hope that these measurement errors balance out to be zero, that half the time we measure above the true value and half the time below the true value, although it is really hoped that the variance of the measurement error will be as small as possible. It is difficult to detect this measurement error, and it is here that techniques such as using multiple indicators of a variable are most useful.

The other aspect of error in equation 8-5 above comes from other possible causes of the dependent variable Y. For instance with the example directly above if we had used number of firemen to predict dollars lost, part of the error component would be other variables such as size of the fire. Now, if the values a, b, and r are to be accurate, this aspect of the error term (as well as the total error term) must be uncorrelated with the predictor variable X. This is often very hard to do empirically, as in our example, so we try to take these other possible causal variables into account. Hopefully, as other possible causal variables are taken into account our measure of actual influences on the dependent variable will become more accurate. Note, however, that when this is done we are no longer dealing with a symmetrical measure, but definitely have a dependent variable chosen.

A controlled experimental design takes the possible influence of other variables into account by matching on these variables between the groups and/or by randomization. In most work done by sociologists, this is very difficult to do. As noted in the first section here this makes the sociologists' theoretical task extremely important. It is up to the researcher to think of other possible confounding variables, other variables that might be influencing the association noted, and to pull these into the study. Also note that one can never be sure that one has taken into account all possible variables. But, with continued trying and good theorizing it is possible to consider many of those that might be important. Note that this problem is faced in all research designs or methodologies with the exception of experiments (which have their own problems of generalizability, ethics, etc.). Anytime a researcher discusses associations or relationships, unless there have been explicit design measures to control for other causal variables as through randomization techniques in experiments, the possibility of other variables actually causing the relationship is there. This happens in field research, historical research, survey, in all but the most controlled of experiments. (although there it is not the experimental variable - but something about the experiment that is sometimes later found to be causing an association.)

To take these other variables into account, more than one variable is used in the regression equation. When there are two causal variables, the resulting equation produces a plane in three-space, rather than a line in two-space as in the case of zero-order correlation. When there are more than two causal variables (more than three variables in all) the resulting formation is in more than three space and is hard for all but science-fiction freaks to envision intuitively. Thus here we will simply use a case with three variables -- two causal variables -- for illustration. Suppose as in the above, the dollars lost at the fire was the dependent variable Y. The number of firemen at the fire was the causal or independent variable X_1 and the size of the fire was the independent variable X_2. Then the regression equation

$$\hat{Y} = a_y + b_1 X_1 + b_2 X_2 \qquad (8\text{-}6)$$

could be computed. The resulting format of this equation would be a plane in three-space. By definition, as with the two-variable case,

this plane is closer to all the actual cases of Y then any other plane would be. This means then that

$$\Sigma(Y - \hat{Y}) = 0 \qquad \text{and that} \qquad \Sigma(Y - \hat{Y})^2 = \text{minimum.} \quad (8\text{-}7)$$

It is then possible to define a measure of association

$$R^2 = \frac{\Sigma(Y-\bar{Y})^2 - \Sigma(Y-\hat{Y})^2}{\Sigma(Y-\bar{Y})^2} \quad (8\text{-}8)$$

This measure is analagous to r^2 in the zero-order case. It tells us the proportion of variation in Y, the dependent variable (dollars lost at the fire), that can be explained by the two independent variables X_1 (number of firemen) and X_2 (size of the fire), when the two independent variables are seen as being linearly related to Y. This last clause of the above sentence is important, because we are noting through the regression equation and the R^2 value only the linear association between the variables. This is because the plane itself is flat and because each of the slopes represents a straight line within a plane. About half the cases of Y are above the predicted plane given by the equation in (8-8) and about half the cases of Y are below it, and all together this plane is closer to these points than any other plane would be. The measure R^2 tells us how much better knowing this plane helps us predict the values of Y than if we simply used the mean of Y as the predictor variable. R^2 varies between 0, meaning no linear association, to 1, a perfect linear association. The measure R itself has no real meaning and is not analagous to r. When R is given on a computer printout, I would suggest simply converting it to R^2. The measure $(1 - R^2)$ is sometimes used as the coefficient of alienation just as in the zero order case.

Now, besides knowing how well two or more variables can predict a dependent variable or the extent of their linear association, a researcher would also be interested in the exact nature of the association. In this case the researcher would look at the regression equation itself (eq. 8-8) or partial correlation coefficients (see below). Within the regression equation each of the slopes b_1 and b_2 tells the amount of change that would be predicted to occur in the dependent variable Y when X_1 and X_2 respectively changed one unit and the other variable (X_2 or X_1) remained constant.

Again, both parts of this first sentence are important. If the actual values in the regression equation (8-6) were

$$Y = 100 - 50X_1 + 700X_2 \qquad (8\text{-}9)$$

the values in the equation would tell us the following: The Y-intercept 100 would tell us that we would expect to lose $100 if no firemen came and if the fire registered zero on the scale of size (perhaps in amount of area burning). Obviously, this by itself seems rather ridiculous, but may occur in an actual situation where the plane has been drawn to best represent all points. The real case of zero in both independent variables would probably never occur in reality. More important are the values b_1 and b_2. $b_1 = -50$ tells us that when the number of firemen at the fire increases by one, and the size of the fire does not change (remains constant), we would expect the dollars lost to diminish by $50. $b_2 = 700$ tells us that when the size of the fire increases by one unit, and the number of firemen does not change, we would expect the dollars lost to <u>increase</u> by $700. Obviously, this equation tells us that the size of the fire has a positive influence on the dollars lost, while the number of firemen there, actually acts to decrease the dollars lost. Note that both coefficients tells us what the <u>independent</u> or <u>separate</u> impact of each independent variable is. Note also that this separate impact may not be at all noticed in the zero-order regression, especially when (and generally only when) an independent variable is associated with both the dependent variable and with the other independent variables.

The defining equation for each of these partial slopes helps illustrate how they are related to the zero order equations and also how the impact reported is of the separate influence of each variable. To get the coefficient b_1 above that predicts the influence of X_1 on Y when the impact of X_2 is controlled, the following equation would be used.

$$b_{yx_1 \cdot x_2} = \frac{b_{yx_1} - (b_{yx_2})(b_{x_1x_2})}{1 - (b_{x_2x_1})(b_{x_1x_2})} \qquad (8\text{-}10)$$

(note f ... w: 55 - ie eq. 34 - similar [illegible])

The subscripts to the coefficients in this equation are important. $b_{yx_1 \cdot x_2}$ is the partial regression coefficient that is being computed. The coefficients on the right hand side of the equation all come from different zero order regression equations. The first subscript tells us what the dependent variable is in the equation and the second subscript identifies the predictor variable in that equation.

Consider the numerator first. The first element is the zero order regression coefficient that corresponds to the partial that is being computed. The second element is a product of two coefficients. The first coefficient is the prediction of the dependent variable with the other independent variable in the equation. The second coefficient in the product is the slope from the equation predicting the predictor variable from this second variable. In essence then, this product is the product of the two slopes from the equations predicting the values of the dependent and predictor variable from the control variable or other variable in the equation. This may be seen as removing that part of the influence of the independent variable on the dependent variable that comes from this control variable. If the control variable does a good job predicting both the dependent and independent variable, then this product will be greater. If the control variable is uncorrelated with either of these variables ($by_{x2} = 0$ or $b_{x_1x_2} = 0$) then it will have no impact on the partial regression coefficient. But if it is correlated with both of these variables it will lower the regression coefficient considerably (if the zero order coefficient is positive). Note that in some cases the effect of taking out this information -- if the influence of the control variable on the independent and dependent variable is quite high -- will actually change the direction or sign of the slope. This occurs when the product of b_{yx_2} and $b_{x_1x_2}$ is greater than the zero order coefficient b_{yx_1}.

While the numerator of this equation is important in determining the magnitude and direction of the influence of the partial regression coefficient, the denominator may be simply seen as a correction factor. By subtracting the product of $b_{x_2x_1}$ and $b_{x_1x_2}$ from one, the amount of influence that the two independent variables have on each other is removed and taken out of any influence on the partial slope. Note that if the two independent variables are generally unrelated to each other then this denominator is close to one and has no influence. If they are highly positively

associated with each other, then the denominator becomes less than one and the partial slope is thus accounted for (or supplemented for) this intercorrela- by an increase. If they are negatively associated with each other the correction goes in the other way. Essentially, this denominator may be seen simply as a correction factor, ~~removing~~ accounting for the joint influence of the two independent variables on each other. The numerator is important in that it takes out the influence the control variable has on both the causal and the dependent variable.

Just as in zero order correlation where the Y-intercept is a function of the means of the dependent and independent variables and the slopes, the Y-intercept when more than one independent variable is involved may simply be x written as a function of the averages of each variable and the partial slopes.

$$a_{y \cdot x_1 x_2} = \bar{Y} - b_{yx_1 \cdot x_2} \bar{X}_1 - b_{yx_2 \cdot x_1} \bar{X}_2 \qquad (6\text{-}11)$$

These formulas may be easily extended to computations when there are three independent variables. In this case

$$a_{y \cdot x_1 x_2 x_3} = \bar{Y} - b_{yx_1 \cdot x_2 x_3} \bar{X}_1 - b_{yx_2 \cdot x_1 x_3} \bar{X}_2 - b_{yx_3 \cdot x_1 x_2} \bar{X}_3$$

$$b_{yx_1 \cdot x_2 x_3} = \frac{b_{yx_1 \cdot x_2} - (b_{yx_3 \cdot x_2})(b_{x_3 x_1 \cdot x_2})}{1 - (b_{x_1 x_3 \cdot x_2})(b_{x_3 x_1 \cdot x_2})}$$

When there are more than three independent variables, hand computations get more cumbersome, although they are still possible. It is usually more accurate at this point to simply use the computer to obtain the calculations. Note that the interpretations of the partial slopes and the Y-intercept remain analagous to the interpretations with two independent variables when more than two independent variables are used.

You may have noticed in the example above that it is hard to tell from simply comparing b_1 and b_2 which variable has more impact. We can guess from looking at the units involved, but sometimes we will want a more standardized measure to use in comparing the direct influence of several independent variables on a dependent variable. A measure analagous to r that gives the predicted change in standard score form would be desirable. And, of course, there is such a measure. If the dependent and independent

variables are changed to standard score form and the regression equation is again computed, the resulting regression coefficients will tell us the expected standard deviation unit change in the dependent variable, when each independent variable changes one unit and the other variables are held constant (do not change). Because each variable is in standard score form (with a mean of 0 and standard deviation of 1) then each of the coefficients is comparable. When one standard score coefficient is larger than another that means that the impact of that variable is greater than the impact of the other. (This may not necessarily be the case with the measures with the regular scores if they are measured in widely varying units where one variable is very compact with a small change in actual numbers producing a large impact and another variable is quite variable with a small change producing not so large an impact. Because the one variable is so compact the magnitude of its impact would not be apparent unless standard scores were used and we could see how large its impact was in comparison to another.) Note also that because standard scores are used, the Y-intercept equals zero, the predicted regression line simply goes through the intersects of the means of the variables at zero.

These standardized regression scores are called beta weights. They are equivalent to path coefficients that are used in simple causal models (see discussion below). The standardized regression equation for the equation originally given in (8-8) may be written

$$Y = \beta_{Yx_1 \cdot x_2} X_1 + \beta_{Yx_2 \cdot x_1} X_2 \qquad (8\text{-}14)$$

$\beta_{yx_1 \cdot x_2}$ tells what the expected change in the standard score of the dependent variable Y would be with a standard deviation score change of in X_1 when X_2 did not change. $\beta_{yx_2 \cdot x_1}$ tells the expected change in the standard score of Y with a one standard deviation score change of X_2 when X_1 does not change.

The standardized regression coefficients may be computed directly from the unstandardized coefficients. They are simply a function of the relative standard deviations of the dependent and independent variables involv... For instance

$$\beta_{YX_1 \cdot X_2} = b_{YX_1 \cdot X_2} \frac{S_{X_1}}{S_Y} \qquad (8\text{-}15)$$

and

$$\beta_{YX_1 \cdot X_2X_3} = b_{YX_1 \cdot X_2X_3} \frac{S_{X_1}}{S_Y} \qquad (8\text{-}16)$$

Note that in both cases the standard deviation of the independent variable is in the numerator and the standard deviation of the dependent variable is in the denominator. Thus, if the independent variable is more variable than the dependent variable the standardized regression weight is smaller than the unstandardized coefficient. If the independent variable is more compact or less variable than the independent variable then the ratio is greater than one and the standardized regression weight is larger than the unstandardized measure.

Note also that in the zero order case the correlation coefficient r is equivalent to the beta coefficient. This comes directly from the interpretation of r as the expected change in standard deviation units. Note however that while r is a symetrical measure, the beta weights are asymetrical as soon as there is more than one independent variable.

(moved from p. 172)

In our discussion of causal relationships we used the partial correlation coefficients. Partial correlation coefficients are related to the beta weights and also have an interpretation of their own. Because a partial correlation coefficient is symetric, ~~in that it tells us the expected amount of standard deviation unit change in one variable with one standard deviation unit change in the other when the impact of a third (or more variables) is controlled~~, while the beta coefficients are asymteric, they are not directly comparable. What happens is that the square of a partial correlation coefficient is equal to the product of its two associated beta weights.

$$r^2_{YX_1 \cdot X_2} = (\beta_{YX_1 \cdot X_2} \times \beta_{X_1 Y \cdot X_2}) \qquad (8\text{-}17)$$

The beta weights may also be directly computed from the correlation coefficients. Note how analagous this equation is (eq. 8-18) to the equation for computing the unstandardized slope. In the numerator the products of the correlations of the independent and dependent variables with the control variable is subtracted from the zero-order counterpart of the partial standardized coefficient. In the denominator, the square of the correlation of the independent variable with the control variable is subtracted from one to use as a correction factor.

$$\beta_{YX_1 \cdot X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{1 - r^2_{X_1X_2}} \qquad (8-18)$$

Because a partial correlation coefficient is symetric, it is important to remember that its interpretation is different from that of the standardized slope. You should use only the beta weights and the unstandardized regression coefficients in talking about expected changes in the dependent variable from alterations in the independent variables. The squares of the partial correlation coefficients can be used to tell the amount of variabion explained in one variable by its linear association with another, when the impact of a third variable is removed.

The definition of the partial correlation coefficient is similar to that for the partial slope

$$r_{YX_1 \cdot X_2} = \frac{r_{YX_1} - (r_{YX_2})(r_{X_1X_2})}{\sqrt{1 - r^2_{YX_2}}\ \sqrt{1 - r^2_{X_1X_2}}} \qquad (8-19)$$

Note that in the numerator the product of the correlations between the two variables of concern and the control variable is subtracted from the zero-order counterpart of the desired partial. What this does essentially is to take out of the zero order correlation the part of that correlation or explained variation that is held in common with the control variable. The denominator is a correcting factor, removing the part of the variation that is held in common with the control variable so that it does not throw off the size of the correlation.

There are two possible ways to understand what the partial correlation is. The first involves seeing the partial correlation between a variable Y and X_1, while controlling for X_2, as the correlation between the residuals of the regressions of Y on X_2 and X_1 on X_2. To understand this, we must return to the zero-order regression equations. With the two equations $\hat{Y} = a + b_{yx_2} X_2$ and $\hat{X}_1 = a + b_{yx_2} X_2$ we are predicting values of Y and X_1 from their linear association with X_2. The resulting regression lines are the best fitting of any possible line that can be drawn through the set of variables. Thus the deviations of the actual scores of Y and X_1 from their predicted scores is zero and the squares are a minimum $\Sigma(Y - \hat{Y})^2$ = minimum and $\Sigma (X_1 - \hat{X}_1)^2$ = a minimum. Each of these differences of an actual score of Y or X_1 from the predicted score in the regression equation using X_2 is termed the residual value. It represents how far off the score is from its predicted value. Sometimes the actual score will be greater than and sometimes it will be less than the predicted value. These residuals represent the error component in the representation of each score Y_i as $a + b X_2 + e_i$ and each score X_{1_i} as $a + b X_2 + e_i$.

We noted earlier that part of this error component was made up of measurement error, but that part of the error term could be due to other variables that were correlated with both the predictor and the dependent variable. In partial correlations we are essentially trying to understand more about this error component. This is done by taking the amount each predicted score $\hat{Y}$ is off from its corresponding actual score Y and pairing it with the associated residual of X_1 from $\hat{X}_1$. We can then look at the association or correlation between these two residuals. Since X_2 has explained all it can in both Y and X_1 before these residuals are calculated, the correlation between these residuals represents the correlation between Y and X_1 when the influence of X_2 has been removed. A partial correlation coefficient then can be seen as the correlation between errors with respect to the control variable or the correlation between the amount not accounted for by the control variables.

The partial correlation coefficient can also be seen as a weighted average of the individual correlations between the two variables in each category of the control variable. In other words, if we were interested in $r_{yx_1 \cdot x_2}$ we could see this as a weighted average (weighted by the number of cases in each category) of the correlation between y and x_1 in each small category of X_2.

As an actual example, suppose we were interesed in the association between occupational prestige (x_1) and income (Y) when the variable of education (X_2) was controlled. With the first explanation above we could see $r_{yx_1 \cdot x_2}$ as the correlation that would occur if we predicted occupational prestige from education and income from education and then computed the residuals or difference of each actual income and occupational prestige score from the predicted scores. The partial correlation would then be the correlation between what was left (the parts of occupational prestige and income that were left after education was taken into account) Alternatively, we could see the partial correlation as a weighted average of the correlation between occupational prestige and income in each category of education.

An Example

On the next pages an example using data on self-reported delinquent behavior of junior high school females is given. The matrix on page 175 is of the zero order correlations between six variables: the girls' reprots of their delinquent behavior, the amount of delinquent behavior they expect in the future, the expectations they perceive that their peers have for their own (respondents') behavior, the perceived expectations of the parents, the expectations of the teachers, and the actual delinquency levels of the peers (defined as best fried) as reported by the respondents. All of the correlations are positive, ranging from r = .293 to r = .745.

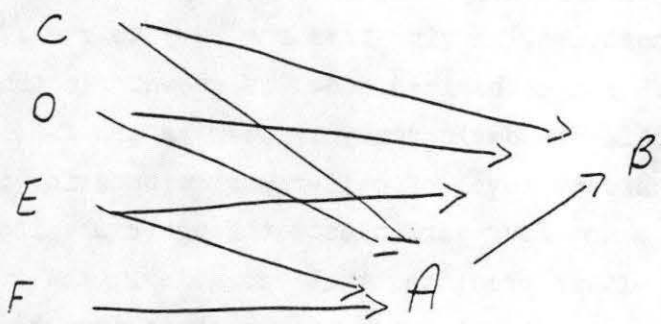
At the bottom of page 175 a hypothesized model is shown. In this model B, the measure of anticipated future delinquency is seen as the final dependent variable. A, the current level of self-reported behavior is an intervening variable, and the four variables C through F are independent variables seen as causing A. It is predicted that variable F, the actual delinquency of peers has no direct influence on the future delinquency level, except through its influence on the young people's current behavior. The assumption equations and prediction equations for this model are given by the model on page 175.

Lecture II · an example w. data from female junior high school students - self reported delinquency.

The Original Data Matrix $n = 162$

	A.	B.	C.	D	E	F
A Self-reported delinquency	1.00					
B. Anticipated future del.	.680	1.00				
C. Peer Expectations	.648	.551	1.000			
D. Parental Expectations	.348	.329	.428	1.00		
E. Teacher Expectations	.285	.327	.357	.341	1.00	
F. Peers' Actual Delinquency	.717	.543	.745	.351	.293	1.00

The predicted model



assume:

$r_{CA} \neq 0$ $r_{EB} \neq 0$ $r_{BC|A} \neq 0$

$r_{DA} \neq 0$ $r_{AB} \neq 0$ $r_{EC|A} \neq 0$

$r_{EA} \neq 0$ $r_{BE|A} \neq 0$

$r_{FA} \neq 0$

$r_{CB} \neq 0$

$r_{DB} \neq 0$

Predict:

$r_{FB|A} = 0$

We may see from the correlation matrix that the assumption equations are confirmed.

(For the smallest coefficient in the table ($r_{AE} = .285$) the test of $H_0: \rho_{AE} = 0$ yields an F of $\frac{r^2(N-2)}{1-r^2} = \frac{(.081)(160)}{.919}$ = 14.102 with df = 1, 160 + we may reject H_0 in favor of $H_1: \rho \neq 0$ at beyond the .001 level of significance.)

To test the prediction equation $r_{FB|A} = 0$ we must compute

$$r_{FB|A} = \frac{r_{FB} - (r_{FA})(r_{BA})}{\sqrt{(1 - r_{FA}^2)(1 - r_{BA}^2)}} = \frac{.543 - (.717)(.680)}{\sqrt{(1-.717^2)(1-.680^2)}}$$

$$= \frac{.055}{\sqrt{(.486)(.538)}} = \frac{.055}{.511} = .108$$

$r^2_{FB|A} = .012$. This indicates that about 1% of the variation in future anticipated delinquency is explained by the young women's reports of peers' (or friends') actual delinquent behavior when the influence of current delinquent activities is removed.

We may also test the hypothesis that $\rho_{FB|A} = 0$ in the usual analysis of variance fashion.

testing assumption equation

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - (r_{yx_2})(r_{x_1x_2})}{\sqrt{1-r^2_{yx_2}}\sqrt{1-r^2_{x_1x_2}}}$$

$$r_{BC|A} = \frac{r_{BC} - r_{BA}r_{CA}}{\sqrt{1-r^2_{BA}}\sqrt{1-r^2_{CA}}} = \frac{.551 - (.680)(.648)}{\sqrt{1-.680^2}\sqrt{1-.648^2}} = \frac{.551-.441}{(.73)(.76)} = \frac{.11}{.55}$$

$$= 0.20 \qquad F = \frac{r^2_{BC|A}(N-3)}{1-r^2_{BC\cdot A}} = \frac{(.20)^2(159)}{1-.20^2} = \frac{6.36}{.96} = 6.625$$

$$r_{BO|A} = \frac{r_{BO} - (r_{BA})(r_{OA})}{\sqrt{1-r^2_{BA}}\sqrt{1-r^2_{OA}}} = \frac{.329 - (.68)(.348)}{\sqrt{1-.68^2}\sqrt{1-.348^2}} = \frac{.329-.237}{(.73)(.94)} = \frac{.092}{.69}$$

$$= .13 \qquad F = \frac{(.13^2)(159)}{1-.13^2} = \frac{2.687}{.98} = 2.74$$

$$r_{BE|A} = \frac{r_{BE} - r_{BA}r_{EA}}{\sqrt{1-r^2_{BA}}\sqrt{1-r^2_{EA}}} = \frac{.327 - (.68)(.285)}{\sqrt{1-.68^2}\sqrt{1-.285^2}} = \frac{.327-.194}{(.73)(.96)} = \frac{.133}{.70}$$

$$= .19 \qquad F = \frac{(.19)^2(159)}{1-.19^2} = \frac{5.74}{.96} = 5.979$$

$H_0: \rho_{FB|A} = 0$ $H_1: \rho_{FB|A} \neq 0$ assume: multivariate normal distributions, independent random sampling

Table 8-1

Source of variation in X_B	Sums of Squares	degrees of freedom	Estimate of variance	F
Total	$\Sigma(X_B - \bar{X}_B)^2$	$N-1 = 161$		
Explained by X_A	$r_{BA}^2 [\Sigma(X_B - \bar{X}_B)^2]$ let $\Sigma(X_B - \bar{X}_B)^2 = \Sigma z_B^2$	1		
Unexplained by X_A	$(1 - r_{BA}^2)\Sigma z_B^2$	$N-2$		
Explained by X_F (that part of the variation unexplained by X_A that is explained by X_F)	$r_{BF.A}^2 (1-r_{BA}^2)\Sigma z_B^2$	1	$r_{BF.A}^2 (1-r_{BA}^2)\Sigma z_B^2$	$\dfrac{r_{BF.A}^2 (N-3)}{1 - r_{BF.A}^2}$
Unexplained by X_F	$(1 - r_{BF.A}^2)(1-r_{BA}^2)\Sigma z_B^2$	$N-3$	$\dfrac{(1-r_{BF.A}^2)(1-r_{BA}^2)\Sigma z_B^2}{N-3}$	

so here $F_{(1,159)} = \dfrac{(.012)(159)}{.988} = 1.931$

$t = \sqrt{F} = 1.390$ $df = 159$ and by consulting the t table we see that it would be possible to reject H_0 in favor of H_1 at the .20 level but not at the .10 level. Given that $r_{FB.A}^2 = .012$ and the smallest other r^2 assumed in the model = .081, we leave the model as is and conclude that it is correct for this sample ~~but~~ (& within the sampling limitations for the population).

The test of the hypothesis H_o: $\rho_{YB/A} = 0$ on page 177 is done in the same manner that inferential tests were done in section seven in association with zero-order correlation. In doing inferential tests associated with regression it is necessary to assume that the sample has been randomly drawn from or is representative of a population and that in the case of testing hypotheses about partial correlations that there is a multivariate normal distribution. This is the multivariate extension of a bivariate normal distribution, with each variable distributed normally about the others in the population. This also implies that homoscedasticity exists. Note that these assumptions are not needed to compute the values of the partial correlation coefficients, but they are needed if we are going to make inferences about these values to the population.

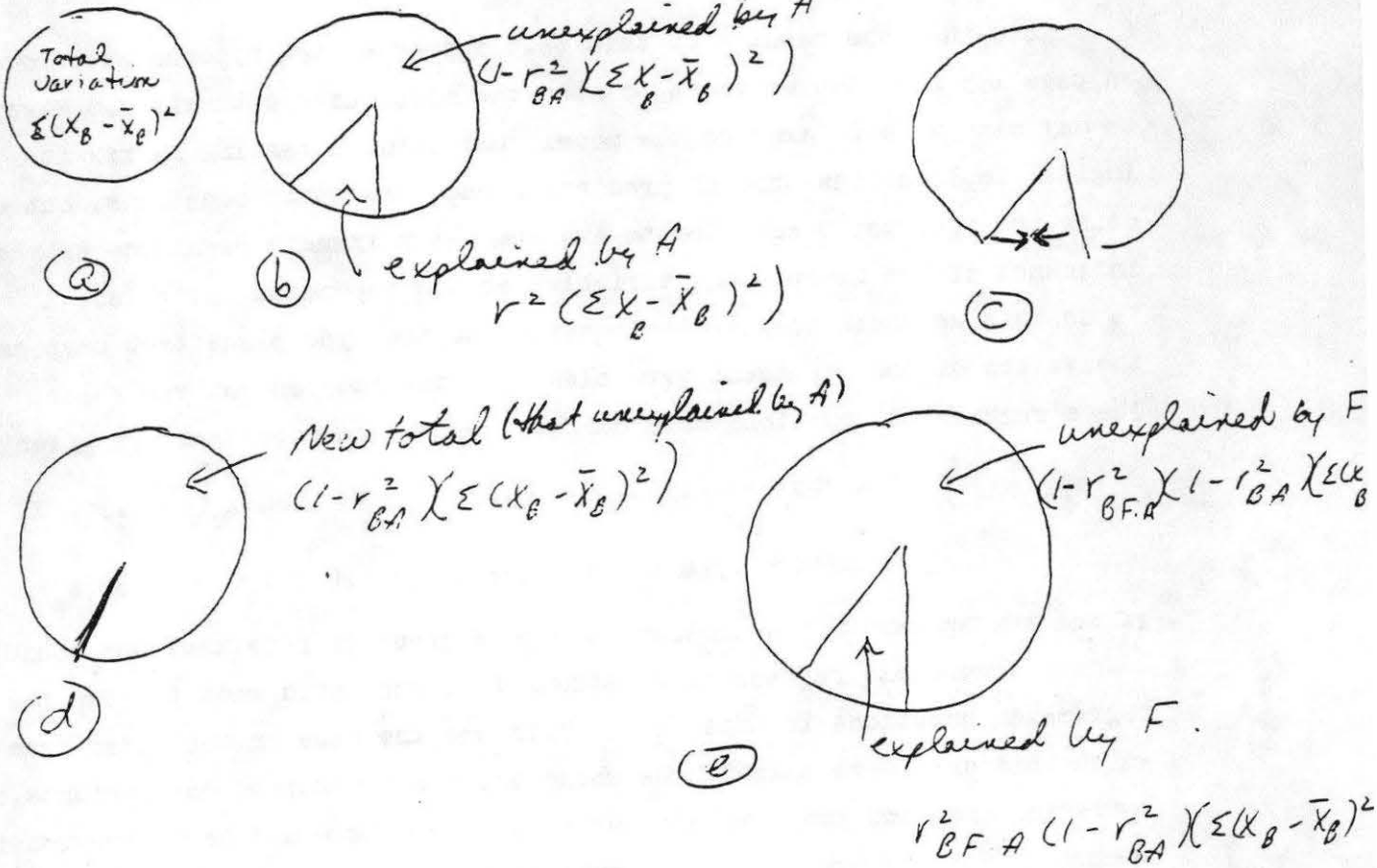
In table 8-1 it may be seen that as usual the total variation is defined as the variation around the mean. The degrees of freedom associated with the total variation or sum of squares is N-1, losing one degree of freedom for the overall mean. The unexplained and explained variation has two aspects. First the variation explained by the control variable (A) and unexplained by the control variable is considered. As with the zero order case, the explained variation is $r^2_{ba}(\Sigma(X_b - \bar{X}_b)^2)$ and the associated degrees of freedom is one. The unexplained variation is $(1 - r^2_{ba})(\Sigma(X - \bar{X})^2)$ and the associated degrees of freedom are N - 2, losing two degrees of freedom when computing the regression line. This part of the argument is identical to that used with zero-order correlation. We then may look at the proportion of the variation that is explained by the other independent variable once the influence of the control variable has been taken into account. This is $r^2_{b(1-a)}(1 - r^2_{ba})(\Sigma(X_b - \bar{X}_b)^2)$. Note that this is simply taking that which is unexplained by X_a as the new definition of the total variation and then looking at what part of this new total variation is accounted for by the other variable. The unexplained variation is then $(1 - r^2_{bA \cdot a})(1 - r^2_{bA})(\Sigma(X_B - \bar{X}_B)^2)$

(in table 8-1 $\Sigma(X_B - \bar{X}_B)^2$ is written sometimes as Σx_B^2)

The degrees of freedom explained by X_1 in table 8-1 are one, and the degrees of freedom for the variation unexplained by the variable F when the influence of A is removed is N - 3, losing one more degree of freedom with the addition of another variable. The estimates of the common variance in the population are computed in the usual way by dividing the sums of squares or variations by the degrees of freedom, and the F-ratio comes from the ratio of the estimates of variance from the unexplained or within variation and that from the estimate from the explained variation.

The "squishy pie" example can help explain how the unexplained and explained variation come about.

Figure 8-1



As with zero-order correlation we start out with the total variation as in part a of figure 8-1. This then can be broken into that which is explained by linear association between b and a and that which is unexplained by that association as shown in part b of the figure. Then, if we take out this explained part, as in part c of the figure, we just have the part of the total variation in b that is unexplained by A. We can then squish this together (as one might squish together the pieces of a pie so that noone would know you had taken a piece) and then use this new configuration as the total variation. This may then be broken into that which is explained by f and that which is unexplained by f as in part e. The important point is, that now we are dealing with the proportion of variation explained by f when that part explained by a is removed.

Note that the results of this test supported the hypothesized model on page 175 and that we chose to keep the model as hypothesized. However, we may also wish to examine the model, not through testing its logical implications through prediction and assumption equations, but by using the regression equation to examine the magnitude of the influence of the independent variables on the dependent variables. To do this we could compute the equation of the hyperplane that best estimates the values of the dependent variables from the independent variables. These regression equations were computed by the computer and are given below.

$$X_A = 4.965 + .192\,X_c + .115\,X_d + .044\,X_e + .362\,X_f \qquad (8\text{-}20)$$

$$X_b = .6211 + .514\,X_a + .117\,X_c + .077\,X_d + .135\,X_e \qquad (8\text{-}21)$$

If one had similar data on a group of boys that one could compute comparable regression equations for, one would want to keep the regression euqations in this form. This was the case in the paper from which this data were taken. One would use unstandardized coefficients in comparing from one group to another because one would be interested in seeing in which group a certain variable had more or less of an impact and comparing the magnitude of that impact in each group.

Because they are more flexible and better suited to testing of theory, regression equations are more often used than partial corelations.

However, if one were interested only in this sample, one would probably want to convert the unstandardized regression coefficients into standardized ~~ones~~ coefficients or beta weights. Because the standard deviations of each of the variables in the model are not equal, the above equations don't tell us about the relative importance of the effect of the independent variables on the dependent variables. We cannot easily compare the impact say of variable C on variable A and variable D on variable A. The unstandardized slopes in the equations above tell us the predicted actual change in the dependent variable with an actual change of one value in the independent variables when the other independent variables do not change. However, because each variable may vary in its variance or compactness, it is difficult to know for sure from simply looking at the unstandardized slopes which variables have the most impact.

To standardize the slopes we compute beta weights. This is done by adjusting the slopes, essentially changing all the scores to standard deviation units (z-scores) and recomputing the regression equation. As noted earlier this is equivalent to simply multiplying the unstandardized slopes by a ratio of the standard deviations of the independent and dependent variables. In this case

$$\beta_{AC.ID} = b_{AC.ID} \frac{s_c}{s_a} \qquad (8\text{-}22)$$

Note also that we could compute the beta weights directly from the correlation coefficients as discussed earlier, although with the number of variables in this model the computations quickly become tedious.

With standardized scores the regression equations become

$$X_a = .2225 X_C + .0592 X_D + .0328 X_E + .5213 X_F \qquad R^2 = .5482 \qquad (8\text{-}2$$
$$(6.934)* \quad (0.943) \quad (0.310) \quad (41.862)**$$

$$X_B = .5428 X_A + .1428 X_C + .0418 X_D + .1068 X_E \qquad R^2 = .4966 \qquad (8.24)$$
$$(52.570)** \quad (3.293) \quad (0.419) \quad (2.944)$$

Note that even though the interpretation should rightly see the influences as representing changes in standard scores, the convention is generally to simply write the equation using the regular notation for the variables.

Note that the Y-intercept here equals zero because the regression plane goes through the intercept of all the means which is zero. Also note that we can see from these equations that variable F, the actual delinquency of peers is the most important influence on the current level of reported behaviors and that c, the expectations of peers is next most important. The most important influence on anticipated future delinquency is the current level of self-reported behavior and next most important are the expectations of peers.

It is also possible to test the hypothesis that the regression coefficients are equal to zero. Again this is done in the familiar analysis of variance format. As with the partial correlation coefficients the associated degrees of freedom are one and N - k, where k represents the numberx of variables (both independent and dependent) in the equation. The F-ratios associated with each of the coefficients in equations (8-23) and (8-24) are given in parentheses below each coefficient. In equation (8-23) by comparing the F-ratios to Table J in Blalock we may see that the influence of X_c on X_a would occur by chance less than 1 time out of one hundred and

that the influence of X_f on X_a would occur by chance less than one time in a thousand and probably much more rarely (the F-table isn't exact enough to tell how rarely.) Similarly, with equation (0-24) we may reject the null hypothesis that the regression coefficient is zero in favor of the alternative that it is not equal to zero with the influence of only X_a on X_D. The influence of X_c approaches significance at the .05 level, but because of the crudity of the table we cannot tell what the exact probability of being wrong in rejecting the null hypothesis would be in this case. Substantively, these conclusions tell us that parents' and teachers' expectations have probably no real effect on young women's current reported behavior and that current behavior is the major influence on future anticipated delinquency. The influence of peers' expectations and also somewhat of teachers' expectations approaches the traditional .05 level of significance.

We may also measure how much of the total variation of the dependent variable is explained by the independent variables. This value is given by R^2, the multiple correlation coefficient squared as shown by each of the equations 0-23 and 0-24. About 55% of the variation in current reported behavior is explained by the expectations of peers, parents, teachers and the actual behavior of peers. About 50% or half of the variation of future anticipated behavior is explained by the variables in equation 0-24. Again using the analysis of variance format we may test the hypothesis that the multiple correlation coefficient squared in the population is equal to zero. The degrees of freedom associated with the unexplained variation would be $N - K - 1$, where K represents the number of independent variables. Essentially one degree of freedom is lost for each variable involved in the computation (adding together independent and dependent variables). The explained degrees of freedom are K, the number of independent variables. As with zero order correlation, the associated F-ratio is

$$\frac{R^2}{1-R^2} \quad \frac{N-k-1}{k} = F \qquad (0\text{-}25)$$

For our two examples the F ratios are

$$F = \frac{.55}{(1-.55)} \frac{(162-4-1)}{4} = (1.22)\left(\frac{157}{4}\right) = 47.885 \quad (6.2$$

$df = 4, 157$

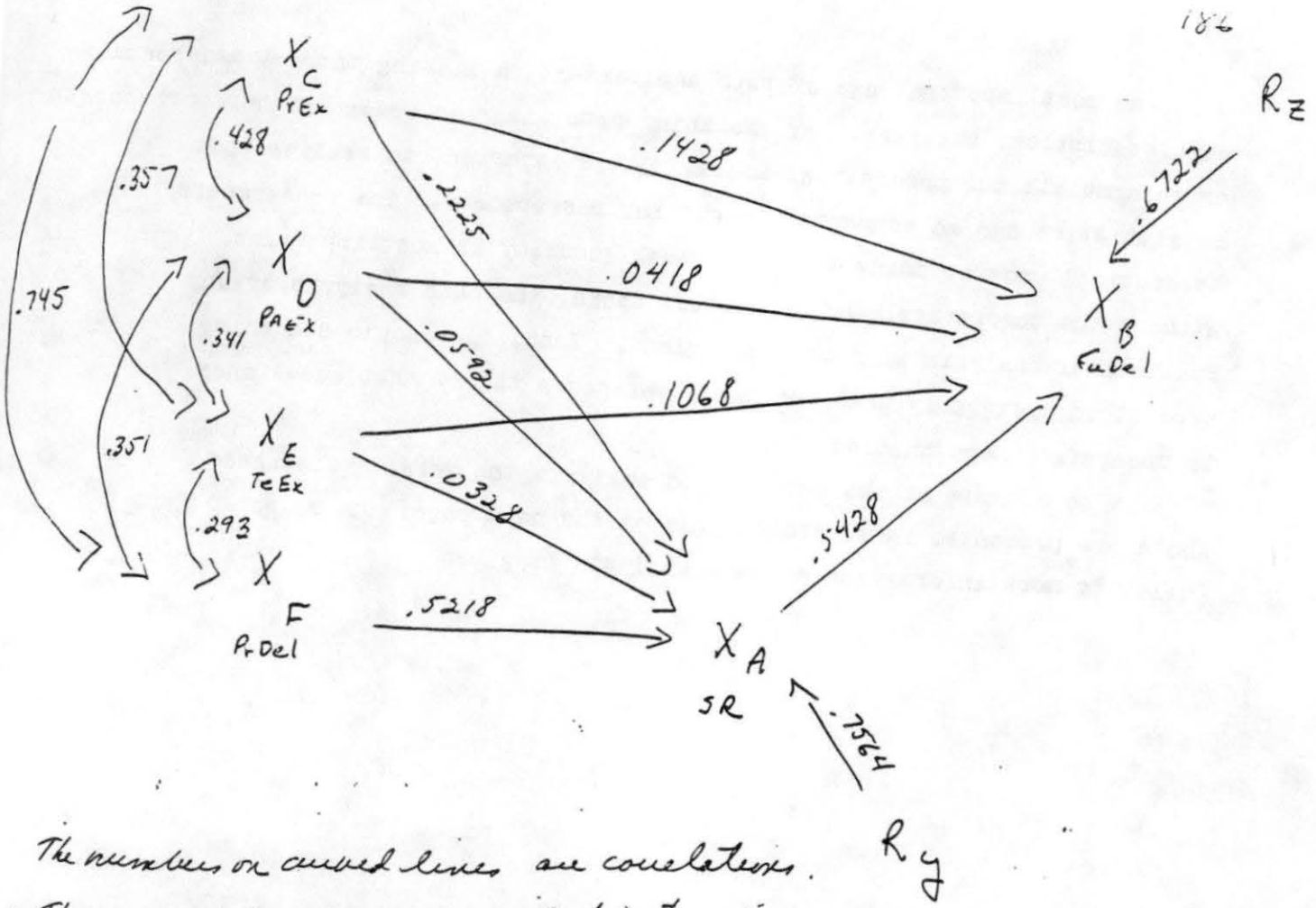
$$F = \left(\frac{.50}{.50}\right)\left(\frac{157}{4}\right) = 39.25 \quad df = 4, 157 \qquad (6.27)$$

and by referring to Table J we can see that in both cases we can easily reject the null hypothesis that in the population $R^2 = 0$ in favor of the alternative that $R^2 \neq 0$ with less than one chance in one-thousand of being wrong in doing so.

In recent years standardized ~~xxx~~ regression coefficients have been used by sociologists in path diagrams. Path analysis is simply the use of multiple regression analysis in analyzing formal models. In path analysis we pictorially represent a model of interrelationships. Standardized ~~xxxx~~ regression coefficients or beta weights or path coefficients (they are the all the same thing) represent the direct influence of an independent (or exogenous) variable on a dependent (or endogenous) variable. By examining the path diagram we can see how the exogenous variable also indirectly influences the endogenous variable. The influence of other variables not included in the model is represented by a "residual variable." The percentage of variation explained by these "residual variables" is simply the persentage of variation that has been unexplained by the other variabl es of $1 - R^2$. Then the path coefficient from the residual to the dependent variable is $\sqrt{1-R^2}$. Obviously this is equivalent to the correlation between the residual variable and the dependent variable. This is then equivalent to ~~R~~ or comparable to the other beta weights, because in the zero order case a correlation coefficient is the same as the beta weight and if we are to assume that the path coefficients or beta weights are accurate, we must assume that the residual is not correlated with the other variables in the model. (Just as with any correlation we must assume that the error term is uncorrelated with the variables involved.)

The most important use of path analysis is in linking theoretical results with statistical analyses. By examining path diagrams we can either corroborate or reformulate our theoretical model. It is important to realize that no statistics can do conceptual magic for researchers. The researchers' theoretical understanding and conceptual accuracy is most important. Without the theoretical and conceptual basis, the path analysis or any statistical analysis will be meaningless. Thus, one should use this type of analysis only when one has developed a theory that one wishes to understand more thoroughly.

As an example of the use of path analysis the results discussed above are presented in a path diagram on the next page. On pages 107 and following more information on path analysis is given.



The numbers on curved lines are correlations. They are used when no causal direction is posited. The numbers on the arrows are standardized regression weights or path coefficients.

The standard notation is $\beta_{AC/DEF} = P_{AC}$, the path coefficient that represents the direct influence of variable X_C on X_A.

Mathematically $r_{ac} = \sum_{c=c}^{b} p_{ai} r_{ic}$ or in general it means through all variables with paths going to A

Then $r_{ac} = p_{ac} + p_{AO} r_{Oc} + p_{AE} r_{Ec} + p_{AF} r_{Fc}$

p_{ac}: direct infl. of C on A

$p_{AO} r_{Oc}$ + $p_{AE} r_{Ec}$ + $p_{AF} r_{Fc}$: indirect infl. of C on A (three variables O, E, F)

Additional information on interpreting path diagrams:

Path coefficients are standardized regression coefficients. Thus we may see $p_{bc} = \beta_{bc}$ as telling us that when X_c changes one standard deviation unit, X_b changes .1428 standard deviation units. In constrast, $p_{bd} = \beta_{bd} = .0418$ when X_d changes one standard deviation unit + X_c, X_e + X_a are constant, X_b changes .0418 of a standard deviation unit. Since we are dealing now with standard deviation units (the same thing as using normal scores with s=1) we can directly compare these beta coefficients or path coefficients. We can see then that variable X_c has much larger influence (when the impact of variables D, and E are removed) on X_b, than does X_d (when the impact of variables C, and E are removed). Similar comparisons can be made with p_{be}. In general, beta weights or path coefficients may be interpreted as telling us the standard deviation unit change in the dependent variable when the influence of the other independent variables (or other variables directly causing the dependent variable) is controlled. They are perhaps most useful when we want to compare the influence of several independent variables within a given sample (or population).

In our example we can see that the expectations of peers and the actual delinquency of peers have the most important influence on the students' current reported delinquency. The students' current delinquency has the most important influence on their projected future delinquency.

Land (1969) quotes Wright (1934) as saying that "the squared path coefficient measures the proportion of the variance of the dependent variable for which the determining variable is directly responsible (Land, p. 10, emphasis in original)." This comes from seeing the path coefficient as measuring "the fraction of the standard deviation of the endogenous variable ... for which the designated variable (the exogenous variable) is directly responsible" (pp.8-9, Land, 1969)
Its important to note that this is percentage of variance, not variation, which is the basis of the interpretations of r^2, E^2 and R^2.

This leads to another important interpretation from path diagrams. By displaying the associations in a path diagram we may examine both the indirect and direct effects of the exogenous or independent variable on the endogenous or dependent variable. Without going into why this happens, we may represent the correlation between each independent and dependent variable as a function of the path coefficient between that independent variable and the dependent variable plus combinations of the path coefficients from other independent variables and the correlations of our independent variable (the one in the first correlation) and all the others. This association is shown at the bottom of page 9.

Here $r_{ac} = p_{ac} + p_{ad}r_{dc} + p_{ae}r_{ec} + p_{af}r_{fc}$ or (1)

$$.648 = .2225 + (.0592)(.428) + (.0328)(.357) + (.5218)(.745) \quad (2)$$

$$.648 = .2225 + .0253 + .0117 + .3887 = .2225 + .4257 \quad (3)$$

$p_{ac} = .2225$ represents the direct influence of X_c on X_a; but $r_{ac}=.648$. $r_{ac}-p_{ac} = .4257$ which represents the influence of X_c on X_a that is indirect.
From examining lines 2 and 3 directly above we can see that the largest part of this indirect effect is through variable X_f, the actual delinquency of peers.

To understand the patterns of direct and indirect effects you can use either the general formula given at the bottom of page 9 or follow the diagram. To do this with r_{ac} you may start at X_c follow the direct influence along p_{ac} to X_a, from X_a go back to X_d and then through r_{dc} back to X_c; then along r_{ce} to X_e and along p_{ae} to X_a; back along p_{af} to X_f and back along r_{cf} to X_c.

A third important interpretation from path diagrams involves the residual. With regression equations we were interested in R^2, the total percentage of variation of the dependent variable that can be explained by the independent variables. By definition R^2 = (explained variation in dependent variable)/(total variation of the dependent variable). This = $\Sigma(Y' - \bar{Y})^2 / \Sigma(Y - \bar{Y})^2$. This is exactly analagous to the measures of association, r^2 and E^2. Again, this definitional formula is not used in computations. Say X_1 is the dependent variable and X_2 and X_3 are independent variables. If $r_{23} = 0$, $R^2_{1(23)} = r^2_{12} + r^2_{13}$. But usually r_{23} is not equal to zero. We have to then take into account the overlapping explanations of X_1 provided by X_2 and X_3. This can be done by multiplying each of the correlations of the independent variable with the dependent variable by its related standardized beta coefficient. In our example.

$$R^2_{a(cdef)} = r_{ac}\beta_{ac} + r_{ad}\beta_{ad} + r_{ae}\beta_{ae} + r_{af}\beta_{af}$$

(note that really above $\beta_{ac} = \beta_{ac \cdot def}$, and so on)

or $$.5482 = (.648)(.2225) + (.348)(.0592) + (.285)(.0328) + (.717)(.5218)$$

$$.5482 = .1442 + .0206 + .0093 + .3741 = .5482$$

You may also compute R^2 directly from the zero-order r's as shown ~~on page 7~~ earlier.

Now since R^2 is the percentage of the total variation of the dependent variable that may be explained by the independent variables, $1-R^2$ gives the percentage of the total variation that is unexplained. In path diagrams variables that are not in the model are represented by one residual variable for each endogenous variable. All the variables not included in the model that could explain the variation that is unexplained are seen as being in this residual variable. Then $1-R^2$ tells how much variation this residual variable explains. $\sqrt{1-R^2}$ is the path coefficient between the dependent variable and the residual variable, and, since the residual variable is assumed to be uncorrelated with the independent variables (more on this below), this is also equal to the correlation between the residual variable and the dependent variable. The size of the residual path is commonly seen as indicating how "good" the model is at explaining the dependent variable. Theoretically this path should approach zero as we get better and better at understanding the social world.

It is important to note that as the number of predictor variables approaches the sample size, R^2 is artificially inflated. In fact, if N=k (k=the number of independent variables) R^2=1.00. This is because the hyperplane best fitting the points can go exactly through each of the points. There is then a correction factor that corrects for this inflation effect.

$$R^2_c = 1 - \left(\frac{N-1}{N-k}\right)(1-R^2)$$ where N is the sample size, R^2 the uncorrected multiple correlation coefficient and R^2_c the corrected coefficient, and k is the number of independent variables.

One final comment on path analysis. If one variable intervenes between an antecedent and a dependent variable, this association may be examined through path analysis, also, and the nature of the association may be examined for fit by recreating the zero-order associations between the antecedent and dependent variable as illustrated above for the independent and dependent variables.

A few comments on SPSS and multiple regression:
If you are using prediction and assumption equations to test the logical implications of your model you will want to use the subprogram PARTIAL CORR. This a fairly straightforward program. You may call for the zero order correlations, the means and standard deviations, and partial correlations with up to five control variables.

If you want to compute the multiple regression equation (with either unstandardized b's or the standardized beta coefficients) and the multiple correlation coefficient you will use the subprogram REGRESSION. The output for this program gives the coefficients for the multiple regression equation (for the unstandardized equation you are given the B's for each of the independent variable and the constant value or a; for the standardized regression equation you are given the beta weights, the standardized regression coefficients). For each of the B's (the unstandardized coefficients) you are given the standard error. This value can be used in testing the hypothesis that the population counterpart of the B equals zero and/or for placing confidence limits around the population coefficient. For each of the B's an F-value is given. This F value may be used in testing the hypothesis that the slope coefficient in the population is equal to zero.
In this program each of the independent variables may be added to the regression equation by itself. It is possible then to test if adding a given variable explains a siggificant additional amount of variation in the dependent variable. For each step in this addition process you are given both the F value testing the hypothesis that the B value of the variable(s) in the equation equal zero in the population and the F-values that would apply to the remaining independent variables if they were the next variable to enter the equation. The standardized regression coefficient of these remaining variables (if they were to enter alone on the next step) is given; we are also given the partial correlation coefficient between the remaining variables and the dependent variable if the other independent variables are controlled; nad we are given the "tolerance" of the independent variables not yet added to the equation. The tolerance refers to the proportion of the variance of the independent variable (not yet in the equation) that can not be explained by the independent variables already in the equation. So, if a variable has a tolerance equal to zero, that means that it is very highly related to variables already in the equation while if the tolerance is close to 1.0 it is fairly independent of the variables already predicting the dependent variable. Finally, the printout gives the multiple correlation coefficient, the square of this multiple coefficient, and the adjusted coefficient. The standard error of R is also given for computing cofidence limits.

To compute standardized beta coefficients if you want to do it by hand requires some knowledge of algebra. The easiest way to do it is with matrix algebra. The example below shows how you would compute the beta weights in the equation predicting X_a, either by using four equations with four unknowns or in the matrix format.

$$r_{ac} = \beta_{ac} + \beta_{aD} r_{Dc} + \beta_{aE} r_{Ec} + \beta_{AF} r_{Fc}$$

$$r_{aD} = \beta_{ac} r_{cD} + \beta_{AD} + \beta_{aE} r_{ED} + \beta_{AF} r_{FD}$$

$$r_{aE} = \beta_{ac} r_{cE} + \beta_{AD} r_{DE} + \beta_{AE} + \beta_{AF} r_{FE}$$

$$r_{aF} = \beta_{ac} r_{cF} + \beta_{AD} r_{DF} + \beta_{AE} r_{EF} + \beta_{AF}$$

or

$$\begin{bmatrix} r_{ac} \\ r_{aD} \\ r_{aE} \\ r_{aF} \end{bmatrix} = \begin{bmatrix} r_{cc} & r_{cD} & r_{cE} & r_{cF} \\ r_{cD} & r_{DD} & r_{DE} & r_{DF} \\ r_{cE} & r_{DE} & r_{EE} & r_{EF} \\ r_{cF} & r_{DF} & r_{EF} & r_{FF} \end{bmatrix} \begin{bmatrix} \beta_{ac} \\ \beta_{aD} \\ \beta_{aE} \\ \beta_{aF} \end{bmatrix}$$

↑ known ↑ known

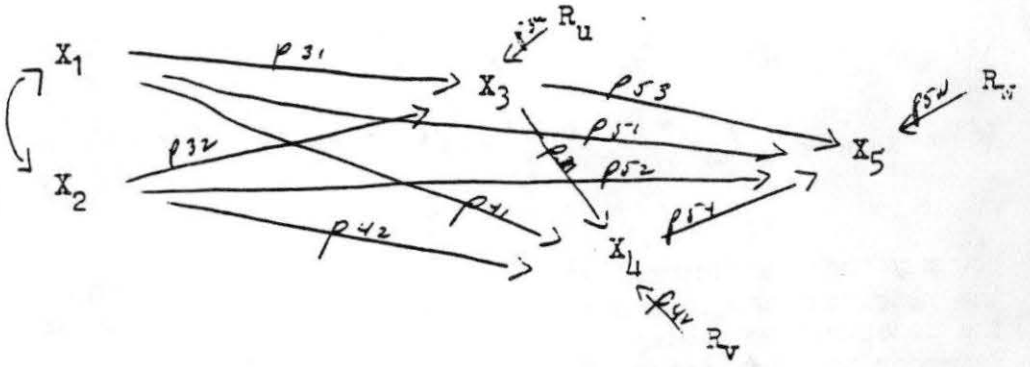
appropriate pages from the SPSS manual follow

Path Analysis

In this section we discuss the basic elements of path analysis in more detail than we did last term. First we review the basic theorems and understandings, then discuss the possibility of using ordinally measured variables in path diagrams, then look at the use of multiple indicators in path analysis, and finally discuss more complex techniques by only briefly describing the use of simultaneous equation models. Obviously this area is complex and vast. It all stems from understanding of the general linear model, but there is no way to thoroughly cover all its aspects in only one term.

In quant II we briefly discussed path analysis in the context of regression. The standardized regression coefficients are equal to path coefficients. This term we will extend this discussion going into the details of the general theorem of path analysis and interpretations of path diagrams; the use of ordinal variables in path analyses; and using multiple indicators of a variable in path analysis. These last two techniques should be helpful in developing more precise measures of sociological concepts.

The causal model below is identical to that discussed in the first part of Duncan's article. This is what we call a "fully identified model." Each of the prior variables in time is seen as causing a later variable. In the model the variables are spaced from left to right as they occur in time.



Suppose that the X_i's in the model are standard or z scores (if the real measure of the first variable is V_1, $X_1 = (V_1 - \bar{V}_1)/\sigma_1$).
Then we could represent each of the dependent variables in the model above with the following prediction equations. These are the standardized regression equations we studied last term.
The β_{ij}'s represent the standard deviation unit change in the dependent variable i with a change of one standard deviation in the independent variable j. So in the equations where the X_i's are all in standard deviation units, the β_{ij}'s represent the amount of standard deviation unit change in the dependent variable X_i with a change of one unit in X_j when the impact of the variables in the equation is removed.

(1) $\hat{X}_3 = \beta_{31}X_1 + \beta_{32}X_2$

(2) $\hat{X}_4 = \beta_{41}X_1 + \beta_{42}X_2 + \beta_{43}X_3$

(3) $\hat{X}_5 = \beta_{51}X_1 + \beta_{52}X_2 + \beta_{53}X_3 + \beta_{54}X_4$

For each of these prediction equations we may obtain an R^2, a multiple correlation coefficient, which indicates how much of the variation in the dependent variable is explained by the independent variables.
Now suppose that there was some other variable for each of the equations above that could explain all of this remaining variation.
We could call these variables (again in standard score form) R_u, R_v, and R_w respectively for each of the three equations above.

We could then rewrite the prediction equations 1,2, and 3 above including these new variables. How, however, all the variation in the dependent variables is explained by variables in the equations. Part of these variables will be measured (the X_i's) and one in each equation is unmeasured (the R_i's).
These equations written below (equations 4,5, and 6) are called structural equations. p_{ij}'s replace the β_{ij}'s. These p's are called path coefficients. These equations may be seen as representing the "structure" of the model, the nature of influences on the variables in the model.

(4) $X_3 = p_{31} X_1 + p_{32} X_2 + p_{3w} R_w$

(5) $X_4 = p_{41} X_1 + p_{42} X_2 + p_{43} X_3 + p_{4v} R_v$

(6) $X_5 = p_{51} X_1 + p_{52} X_2 + p_{53} X_3 + p_{54} X_4 + p_{5w} R_w$

How may these structural equations be used to help us interpret the model and the nature of effects(both direct and indirect)on the dependent variables?
Remember the definition of the correlation coefficient as
covariation (XY)/$\sqrt{(\text{variation X})(\text{variation Y})} = r_{xy}$.
When we are using standard scores variation X = variation Y = 1, by definition. The covariation XY = ΣXY. (X and Y are standard scores.)
Suppose then we look at the correlation between X_3 and X_5. Since both of these variables are written in standard scores we know that $r_{35} = \Sigma X_3 X_5 / N$. (see back)
We can now expand this formula by substituting in the structural equation for X_5.

$$r_{35} = \frac{\Sigma X_3 X_5}{N} = \frac{\Sigma X_3}{N}\left(p_{51} X_1 + p_{52} X_2 + p_{53} X_3 + p_{54} X_4 + p_{5w} R_w\right)$$

$$= p_{51} \frac{\Sigma X_3 X_1}{N} + p_{52} \frac{\Sigma X_2 X_3}{N} + p_{53} \frac{\Sigma X_3 X_3}{N} + p_{54} \frac{\Sigma X_4 X_3}{N} + p_{53} \frac{\Sigma X_3 R_w}{N}$$

by definition of r $\quad p_{51} r_{13} + p_{52} r_{23} + p_{53} r_{33} + p_{54} r_{43} + p_{53} r_{3w}$

In general $\quad r_{35} = \sum_{i=1}^{w} p_{5i} r_{i3} \qquad i = 1, 2, 3, 4, w$

$r_{33} = 1$ by definition

$r_{3w} = 0$ again by definition - we must assume this if the values of ρ_{31}, ρ_{32} are to be valid.

Substituting these values we get

(7) $r_{35} = \rho_{51} r_{31} + \rho_{52} r_{23} + \rho_{53} + \rho_{54} r_{43}$

This process can be repeated for all combinations of the dependent and independent variables. This yields the following formulas.

(8) $r_{34} = \rho_{43} + \rho_{42} r_{23} + \rho_{41} r_{13}$

(9) $r_{32} = \rho_{32} + \rho_{31} r_{12}$ (10) $r_{31} = \rho_{31} + \rho_{32} r_{21}$

In general (11) $r_{ij} = \sum \rho_{iq} r_{jq}$ where i and j are two variables in the system and q runs over all variables with paths leading directly to X_i.

If we keep expanding these formulas by substituting in we can get them to the point where we can see any correlation between an independent and dependent variable as a function of paths between independent and dependent variables and correlations between variables with no causal relation. For example,

(12) $r_{35} = \rho_{51} r_{31} + \rho_{52} r_{23} + \rho_{53} + \rho_{54} r_{43}$

$= \rho_{51} (\rho_{31} + \rho_{32} r_{23}) + \rho_{52} (\rho_{32} + \rho_{31} r_{12}) + \rho_{53} +$

$\rho_{54} [\rho_{43} + \rho_{42} (\rho_{32} + \rho_{31} r_{12}) + \rho_{41} (\rho_{31} + \rho_{32} r_{12})]$

~~$= \rho_{51} r_{31} + \rho_{52} r_{23} + \rho_{53} + \rho_{54}$~~

$= \rho_{51} \rho_{31} + \rho_{51} \rho_{32} r_{21} + \rho_{52} \rho_{32} + \rho_{31} \rho_{52} r_{12} + \rho_{53}$

$+ \rho_{54} \rho_{43} + \rho_{54} \rho_{42} \rho_{32} + \rho_{54} \rho_{42} \rho_{31} r_{12} + \rho_{54} \rho_{41} \rho_{31}$

$+ \rho_{54} \rho_{32} \rho_{41} r_{12}$

Remember that path coefficients are the same as standardized regression coefficients; we know that a path coefficient from variable j to variable i represents the standard deviation unit change in i we would expect with one standard deviation unit change in variable j when the influence of other variables affecting i is controlled. Thus, p_{53} represents the change in variable 5 caused by variable 3 when the influence of other variables causing 5 (variables 1, 2, and 4) is removed. p_{53} is then the direct influence of 3 on 5. r_{53} is the total relationship between variables 5 and 3, or the standard deviation unit change we would expect in 5 with one standard deviation unit change in 3 when no other variables were involved. But, in our instance, other variables do affect variable 5 besides variable 3. Thus $p_{53} < r_{53}$. and $r_{53} - p_{53}$ must equal the influence of variable 3 on variable 5 that is not direct. $r_{53} - p_{53}$ represents the indirect influence of variable 3 on variable 5. By using the expansion of r_{53} in equation 12 we can examine the nature of this indirect effect.

The final expansion of r_{53} expresses this correlation only in terms of path coefficients and the correlation between 1 and 2. No causal order is posited between variables 1 and 2 so there can be no path between them. This expansion then can let us see how variable three directly influences variable 5 (thru p_{53}) and then how the indirect influence ischanneled through associations of variable 3 with other variables and the way these other variables cause variable 5. For instance:

(13) r_{53} (total influence of variable 3 on variable 5) =

p_{53} (direct influence of X_3 on X_5) +

$p_{51} p_{13}$ (indirect influence of X_3 on X_5 through X_1) +

$p_{52} p_{23}$ (indirect influence of X_3 on X_5 through X_2) +

$p_{54} p_{43}$ (indirect influence of X_3 on X_5 through X_4) +

$p_{51}\, p_{32}\, r_{21} + p_{31} p_{52} r_{12}$ (indirect influence of X_3 on X_5 through X_1 and X_2) +

$p_{54}\, p_{42}\, p_{32}$ (indirect influence of X_3 on X_5 through X_4 and X_2) +

$p_{54}\, p_{41}\, p_{31}$ (indirect influence of X_3 on X_5 through X_4 and X_1) +

$p_{54} p_{42} p_{31} r_{12} + p_{54} p_{32} p_{41} r_{12}$ (indirect influence of X_3 on X_5 through X_1, X_2 and X_4

Similar expansions can be made for all associations between dependent and independent variables. This ability to dissect the direct and indirect influences of the independent variable on the dependent variable is the primary asset of path analysis in interpreting relationships.

Equation 11 is called the general theorem of path analysis. It is at the base of this ability to dissect the direct and indirect influences on the dependent variable.

Also note that it is possible to trace the direct and indirect influences of 3 on 5 right on the model given on the first page. Taking each of the terms in equation 13 you may trace the relation between variables X_3 and X_5. The notations in parentheses can help guide you on this tracing. (Within any one term of coefficients multiplied by each other the order given can be changed.)

In your own work you will want to compute the values of each indirect and direct influence. Comparing these values is a fantastic aid to interpretations.

To get the path coefficients on the computer use the subprogram REGRESSION. As mentioned above the path coefficients are simply the standardized regression coefficients. You will need one regression statement for each structural equation in your model.

What if your model is not fully identified? That is, what if one of your exogenous (independent) variables is not seen theoretically as directly influencing your dependent variable.
Your structural equations would then be different and your analysis of direct and indirect influences on the dependent variables would be different. For instance, if you hypothesized that variable X_1 did not directly influence X_5 your structural equation predicting X_5 would be as in eq. 14.

$$(14)\quad X_5 = p_{52} X_2 + p_{53} X_3 + p_{54} X_4 + p_{5w} R_w \ .$$

X_1 has been ommitted and p_{51} is then zero. You could now compute new formulas estimating r_{53}, r_{54}, r_{52}, and r_{51} without the direct influence of X_1 on X_5. If indeed X_1 does not directly influence X_5 then these estimates should, within sampling error, reproduce the correlations.

This technique of reproducing the correlations is a tedious process and there is a much easier way computationally to deal with the issue of variables with theoretically only indirect effects, which produces the same results theoretically.

What we are saying in equation 14 is that X_2, X_3, and X_4 explain all the variation in X_5 that can be explained by measured variables in our model. If then, variable X_1 were added as a predictor it should add no additional explanation of the variabion in X_5. That is, the R^2 obtained when predicting X_5 from ~~XX~~ X_2, X_3, and X_4 should be essentially the same as the R^2 obtained when X_1 is added as a predictor. Correspondingly, $\beta_{51 \cdot 234}$ should be equal to zero. It is possible to examine both of these possibilities when using a regression program on the computer. R^2's are given as part of the standard output as are F ratios testing the hypothesis that beta weights (path coefficients) are equal to zero. To ~~xxxxxx~~ make X_1 enter the regression equation last, however, you will need to request a hierarchical ordering of the variables in the equation.

What about the residual variable? You remember that the residual variable (represented by R_1) represents all other variables that might be causing the dependent variable. For instance, R_w is all other influences on X_5. It is necessary to assume that R_w is uncorrelated with the measured variables in the model that are prior to X_5. If we didn't assume this we would be unable to get values for the path coefficients representing influences on X_5. This is actually a basic assumption of all analyses. If we take our results as vaild we are assuming that there aren't other variables around that could destroy this relationship (that there aren't suppressor, distorter, etc. variables to use Lazarfeld's terms). Obviously this is a very difficult assumption. The important point to remember is that we always have this problem. You cannot change your analysis techniques or methods and escape it. (Unless you do a controlled experiment.)

Last term we simply stated that the path coefficient from the residual equaled $\sqrt{1 - R^2}$. You remember that when there is only one variable influencing a dependent measure the path coefficient equals the correlation coefficient. Thus $p_{5w} = r_{5w} = \sqrt{1 - R^2_{5\cdot1234}}$

We can support this contention both through logic and through using the basic theorem of path analysis.

Via logic: Remember that $R^2_{5\cdot1234}$ = the proportion of variation in X_5 that can be explained by all variables acting on it in the model.

$1 - R^2_{5\cdot1234}$ = the proportion of variation in X_5 that is not explained by these variables.

If we remember that R_w represents all variables outside the model that explain X_5 then $r^2_{5w} = 1 - R^2_{5.1234}$

Then $\sqrt{r^2_{5w}} = r_{5w} = p_{5w}$.

Via the general theorem:

Consider $r_{55} = 1 = \sum_q p_{5q} r_{q5}$ $q = 1, 2, 3, 4, w$ (from the general theorem eq. 11)

So:

$$1 = p_{51} r_{15} + p_{52} r_{25} + p_{53} r_{35} + p_{54} r_{45} + p_{5w} r_{w5}$$

p_{5w}^2

$$+ p_{5w}^2 = 1 - (p_{51} r_{15} - p_{52} r_{25} - p_{53} r_{35} - p_{54} r_{45})$$

the is the total effect of 1, 2, 3 + 4 on 5 from expansion of the general theorem

$$= p_{5w}^2 = 1 - R^2_{5\cdot1234}$$

This result is more obvious w. ρ_{3u}

$$\rho^2_{3u} = 1 - \rho_{31} r_{13} + \rho_{32} r_{23}$$

$$= 1 - (\beta_{31.2} r_{13} + \beta_{32.1} r_{23})$$

which is defined as $R^2_{3.12}$

One final comment on path analysis: Any type of causal analysis (apart from full experimental designs) cannot prove causality. You must substantiate the causal nature and order of your model through previous work or theoretical arguments. The theoretical basis is most important and your path analysis will mean nothing without it.

IX: The General Linear Model

In this final section we pull together the various topics that have been discussed this term by showing how analysis of variance may be seen as simply a subset of multiple regression. In the first short section on page 210 the assumptions underlying tests of hypotheses with multiple regression are briefly discussed. Note how these parallel the assumptions used with analysis of variance directly. In the second section (p. 210 to 214 an example of using multiple regression to do one-way analysis of variance is shown. These analyses were done by hand. In the next section this procedure is extended to two-way analysis of variance. This time the procedure is more complex and is done by computer. All computations are done with the REGRESSION program and the notes on that program given in the previous section may be consulted. Actual examples from a run are included in these notes.

Assumptions:

Because we are dealing with the regression or linear model, we still have the assumptions that we have dealt with previously. The model is based on the assumption that the variables are measured on an interval scale; that the X values are normally distributed around the predicted values of the dependent variable with equal variances around the predicted values (actually this is necessary for the testing of hypotheses, but not strictly for computing R^2, partials or the B's or Beta's); and that the error term is uncorrelated with the independent variables. In addition, in testing hypotheses we assume that the sample is independently and randomly selected from the population. We also assume that the influence of the independent variables is additive. This means in our example that the influence of expectations of peers adds to the influence of expectations of parents which adds to the influence of teacher's expectations and so on. There are ways to deal with cases where this last assumption is problematic; there are also ways of dealing with variables measured on less than an interval scale; and we may test normality assumptions. The really tricky assumption involves the error term. To put it simply, this involves the constant problem in social science of making sure we have ruled out other causal factors. If there are other variables not included in the regression equation that could be suppressing the correlation or making it spuriously high, then our results are not valid. To assume that our results are valid we must assume that there are no such other variables; that is, that any other variables influencing our dependent variable are uncorrelated with those we are considering and would not alter our results. This is the basis of the assumption given in path analysis that the residual variable (the error, essentially) is uncorrelated with the independent variables. Obviously this is a difficult assumption, but it is not one that disappears with switching analysis techniques.

A Review and a synthesis:

This term we have reviewed t-tests and studied analysis of variance and regression. The t-test may be seen as a special case of one-way analysis of variance (with df=1,n-1 and $t = \sqrt{F}$); and each of the other techniques we have studied are cases of the general linear model. In each of the models of analysis we have studied, it was assumed that the dependent variable was measured on an interval scale. We can then use the mean as the best predictor of the dependent variable when we only have knowledge of that variable, and use deviations from the mean as a measure of error. The squared deviations of scores from the mean of the dependent variable is called the variation of the dependent variable. We used the various techniques of analysis to try to explain or account for this variation. With analysis of variance the independent variables were measured on only a nominal scale, while with regression we assume that the variables are measured on an interval scale. Techniques with one and with more than one independent variable were used. We computed both summary measures of association (E^2, r^2, R^2, and partial r^2) that indicate the extent to which the total variation has been explained or accounted for by the independent variables and tests of hypotheses. The tests of hypotheses are used to test whether the results obtained could be expected to occur in the population, usually that the measure of association would be zero in the population. There is a good deal of controversy over the use of such tests of significance (see earlier handout with references), primarily because they are so easily affected by sample size and the differences between statistical and substantive significance, and thus we suggested that both measures of association and tests of significance be used.

analysis of variance, regression, and analysis of covariance all derive from the general linear model. Through using this model we can see that the major difference between these techniques is in the different levels of measurement of the explanatory variables. With analysis of variance they are nominally measured; with regression they are interval measured; and with analysis of covariance they may be measured on both levels.

With linear regression the application or fit with the general linear model is obvious. (in fact by definition). Any Y (dependent variable) in the data set or sample may be represented as

$$Y_i = \alpha_i + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon_i$$

where we are predicting Y from the k independent variables X_i, i =1,...,k. We were able to test the hypothesis that for the entire the β =0 and to compute a measure of association, R^2, that indicates the amount of spread of the data around this predicted plane.

We can see how analysis of variance fits this model by simply altering the way we view the independent variables. Say we had data as given below with scores of 15 subjects who had been divided among three different treatment groups. We were interested in the way the treatment groups affected their scores, a typical analysis of variance problem.

Group	Y (score)	X_1	X_2
	4	1	0
A_1	5	1	0
	6	1	0
	7	1	0
	8	1	0
	7	0	1
	8	0	1
A_2	9	0	1
	10	0	1
	11	0	1
	1	0	0
A_3	2	0	0
	3	0	0
	4	0	0
	5	0	0
Σ	90	5	5
$\bar{X}$	6	.333	.333
s	2.9277	.488	.488
$\Sigma (Y)^2$	640	5	5

We could compute a standard analysis of variance within these three categories and that analysis is given at the top of the next page.
But we can also reconceptualize these three groups. What if we made two new variables: one of them X_1 will have a value 1 if the person is a member of A_1 and 0 otherwise; the second, X_2, will have a value of 1 if the person was in A_2 and 0 otherwise. Then each person's group membership is represented by their score on the variables X_1 and X_2. This technique is called dummy variables.

So we can also compute a regression analysis that would predict Y from the person's score on X_1 and X_2:

$$Y_i = a + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i \quad \text{for each Y}$$

These computations are also shown below.

Analysis of variance (traditional)

$$\text{Total sum of squares} = \Sigma [X_{ij}]^2 - \frac{(\Sigma\Sigma X_{ij})^2}{N} = 660 - \frac{(90)^2}{15}$$

$$= 120$$

$$\text{Between SS} = \left[\frac{(4+5+6+7+8)^2}{5} + \frac{(7+8+9+10+11)^2}{5} + \frac{(1+2+3+4+5)^2}{5}\right] - \frac{(90)^2}{15}$$

$$= 180 + 405 + 45 - 540 = 90$$

	Sum of Sq	df	est. of variance	F
Total	120	N-1 = 14	8.571	
Between	90	k-1 = 2	45	18.0
W. thin	30	N-k = 12	2.5	

$$E^2 = \frac{\text{Between SS}}{\text{Total SS}} = \frac{90}{120} = .75$$

Regression

$$r_{y1} = \frac{15\Sigma Y X_1 - (\Sigma X)(\Sigma Y)}{\sqrt{[15\Sigma X^2 - (\Sigma X)^2][15\Sigma Y^2 - (\Sigma Y)^2]}} = \frac{15(30) - (75)(90)}{\sqrt{[15(55) - 25][15(660) - (90)^2]}} = \frac{0}{\sqrt{(50)(1800)}}$$

$$= 0$$

$$r_{12} = \frac{15(0) - (75)(75)}{\sqrt{[15(55) - (75)^2][15(55) - (75)^2]}} = \frac{-25}{50} = -.5000$$

$$r_{y2} = \frac{15(45) - (90)(75)}{300} = \frac{225}{300} = .75$$

$r_{y1} = 0$; $r_{12} = -.5000$, $r_{y2} = .75$

$s_y = 2.928$ $s_1 = .488$ $s_2 = .488$

$b_{y1} = r_{y1} \frac{s_y}{s_1} = 0$; $b_{y2} = (.75)\left(\frac{2.928}{.488}\right) = 4.5$

$b_{12} = (-.5000)\left(\frac{.488}{.488}\right) = -.5000$ $b_{21} = -.5000$

$$b_{y1.2} = \frac{b_{y1} - (b_{y2})(b_{21})}{1 - (b_{12})(b_{21})} = \frac{0 - (4.5)(-.50)}{1 - .5^2} = \frac{+2.25}{.75} = 3.00$$

$$b_{y2.1} = \frac{b_{y2} - (b_{y1})(b_{12})}{1 - (b_{21})(b_{12})} = \frac{4.5 - (0)(-.5)}{.75} = \underline{6.0}$$

$a_{y.12} = \bar{Y} - b_{y1.2}\bar{X}_1 - b_{y2.1}\bar{X}_2$ $\bar{Y}_{.1} = 6$

$= 6 - (3.0)(.333) - (6.0)(.333)$ 4.2^5

$= 3.003$

so our best prediction is

$$Y = 3.003 + 3.00 X_1 + 6.00 X_2$$

$$R^2_{y12} = \frac{r_{y1}^2 + r_{y2}^2 - 2 r_{y1} r_{y2} r_{12}}{1 - r_{12}^2} = \frac{0 + (.75)^2 - 2(0)(-.5)(.75)}{1 - (-.5)^2}$$

$= \frac{.5625}{.75} = .75 = E^2$

~~Quant II -- More on using regression to do analysis of variance.~~

Regression techniques become especially important when we need to do two way analysis of variance and when we have unequal cell sizes and/or when there is a possibility of interaction effects. below is an example using the second assignment done on the computer where VAR16, liberties granted, is the dependent variable and the recoded variable, VAR19, political views, and VAR05, gender, are the independent variables. VAR05 has only two categories, and for the regression I recoded them so that code 0 = female and code 1 = male.
Because VAR19 had three categories, 1=liberal, 2=moderate, and 3 = conservative I had to recode it to two dummy variables.
The IF statements to make these dummy variables are shown on the next page. P1 is the first dummy variable with a code 1 indicating being liberal and 0=either conservative or moderate. P2 has a code 1 = moderate and 0 =either liberal or conservative. Therefore someone with a 0 on both P1 and P2 would be a conservative. Missing values were assigned a 9 on both variables.

Because we also have the possibility of interaction we need to deal with this with the dummy variables. These are called I1 and I2.
As you can see from the next page, I1 equals the product of the score on P1 and the score on gender. I1 would equal one if P1 equaled one (a liberal) and if var05 equaled 1 (a male). Therefore all people who were not male liberals would be 0 on I1. I2 equals 1 if P2 and VAR 05 equal one. Therefore all people who are not moderate males would be 0 on I2. Later below it will become apparent how all possible cases are taken care of by these two interaction variables. (Essentially knowing these two enables us to tell the interaction effect with all other combinations--it is analagous to degrees of freedom in chi-square: the number of categories in one variable minus one times the number of categories in the other variable minus one (r-1)(c-1))

To be most concise and accurate I should also have listed missing values for I1 and I2, but the default option on regression that calls for listwise deletion, the removal of any case if it is missing on one variable saved me. If a case was missing on VAR05 or P1 or P2 it would automatically be excluded. If a case was missing on VAR05 or P1 or P2 it would automatically also be missing on I1 and I2.

As with the one-way analysis of variance on the other handout, we are interested here in explaining the dependent variable with the two independent variables. We are interested in accounting for the variation in VAR16 through the influence of VAR05, VAR19 (now seen as P1 and P2) and any joint effect these variables may have (the interaction terms I1 and I2).
To see this then we ask for a regression equation predicting VAR16 from VAR05, then predicting VAR16 from VAR05 and P1 and P2 (VAR19) and finally predicting VAR16 from VAR05, P1 and P2, and I1 and I2.
We should also have asked for a regression equation predicting VAR16 from P1 and P2 alone (use the statement REGRESSION = VAR16 WITH P1,P2 (1)) but I forgot to do this and computed the R^2 that would result from this by hand. (R^2 = .046)

```
RECODE          VAR05 (2=0)
IF              (VAR18 EQ 1) P1=1
IF              (VAR18 EQ 2 OR VAR18 EQ 3) P1=0
IF              (VAR18 EQ 0) P1=9
IF              (VAR18 EQ 2) P2=1
IF              (VAR18 EQ 1 OR VAR18 EQ 3) P2=0
IF              (VAR18 EQ 0) P2=9
COMPUTE         I1 = P1 * VAR05
COMPUTE         I2 = P2 * VAR05
MISSING VALUES P1, P2 (9)
REGRESSION      VARIABLES= VAR16, VAR05, P1, P2, I1, I2/
                REGRESSION= VAR16 WITH VAR05 (8) P1, P2 (6) I1, I2 (4)
STATISTICS      1,2
```

VARIABLE	MEAN	STANDARD DEV	CASES
VAR16	2.8093	1.2759	1327
VAR05	0.4635	0.4989	1327
P1	0.2962	0.4567	1327
P2	0.3956	0.4892	1327
I1	0.1590	0.3658	1327
I2	0.1530	0.3601	1327

Correlation matrix

	VAR16	VAR05	P1	P2	I1	I2
VAR16	1.00000	0.02281	0.21214	-0.13763	0.17648	-0.10389
VAR05	0.02281	1.00000	0.09554	-0.12458	0.46786	0.45726
P1	0.21214	0.09554	1.00000	-0.52483	0.67033	-0.27567
P2	-0.13763	-0.12458	-0.52483	1.00000	-0.35180	0.52526
I1	0.17648	0.46786	0.67033	-0.35180	1.00000	-0.18479
I2	-0.10389	0.45726	-0.27567	0.52526	-0.18479	1.00000

DEPENDENT VARIABLE.. VAR16 OPINION ON HOMOSEXUALITY

VARIABLE(S) ENTERED ON STEP NUMBER 1.. VAR05 GENDER

MULTIPLE R	0.02281	ANALYSIS OF VARIANCE	DF
R SQUARE	0.00052	REGRESSION	1.
ADJUSTED R SQUARE	-0.00023	RESIDUAL	1325.
STANDARD ERROR	1.27609		

---------- VARIABLES IN THE EQUATION ----------

VARIABLE	B	BETA	STD ERROR B	F
VAR05	0.05835	0.02281	0.07025	0.690
(CONSTANT)	2.78230			

* *

VARIABLE(S) ENTERED ON STEP NUMBER 2.. P1
P2

MULTIPLE R	0.21437	ANALYSIS OF VARIANCE	DF
R SQUARE	0.04596	REGRESSION	3.
ADJUSTED R SQUARE	0.04379	RESIDUAL	1323.
STANDARD ERROR	1.24769		

---------- VARIABLES IN THE EQUATION ----------

VARIABLE	B	BETA	STD ERROR B	F
VAR05	-0.00041	-0.00016	0.06927	0.000
P1	0.53944	0.19310	0.08819	37.416
P2	-0.09470	-0.03631	0.08261	1.314
(CONSTANT)	2.68724			

DEPENDENT VARIABLE.. VAR16 OPINION ON HOMOSEXUALITY

VARIABLE(S) ENTERED ON STEP NUMBER 3.. I2
I1

MULTIPLE R	0.22246	ANALYSIS OF VARIANCE	DF
R SQUARE	0.04949	REGRESSION	5
ADJUSTED R SQUARE	0.04589	RESIDUAL	13 1
STANDARD ERROR	1.24632		

---------- VARIABLES IN THE EQUATION ----------

VARIABLE	B	BETA	STD ERROR B	F
VAR05	-0.02050	-0.00801	0.12327	0.028
P1	0.41277	0.14776	0.12650	10.647
P2	-0.04494	-0.01723	0.11087	0.164
I2	-0.13414	-0.03786	0.16635	0.650
I1	0.23762	0.06813	0.17633	1.816
(CONSTANT)	2.69712			

ALL VARIABLES ARE IN THE EQUATION

The above information is the resulting printout with the mean, standard deviation and number of cases for each variable involved; the zero order correlation matrix, and the results of the stepwise regression. In step one of this regression we are given the equation that predicts the liberties granted from gender. In step two we are given the equation that predicts liberties granted from gender, and the two dummy variables for political affiliation. In step three, gender, political affiliation and the interation of these two variables are all entered as predictors of liberties granted. In each of these steps we get a measure of R^2, the proportion of the total variation in liberties granted that may be explained by the independent variables. You will remember that because R^2 = explained variation/ total variation; explained variation = R^2 (total variation). We will use this to develop the tests of analysis of variance as shown below.

With two-way analysis of variance we have three hypotheses:
1) H_0: the row means are equal or the means of the categories of one variable (say gender) are equal and H_1: the means of the categories of gender are not equal.
2) H_0: The mean liberties granted in the categories of political affiliations are equal and H_1: The mean liberties granted in the categories of political affiliations are not equal.
3) H_0: There is no interation and H_1: There is interaction.
We must test the third hypothesis first. In the standard way of analysis if we reject the null hypothesis we cannot complete the analysis. With regression it still makes no sense to complete the analysis, but we can make furtner interpretations and analysis. The following table gives the analysis of variance for this problem.

~~p.10~~ 220

Below are the F tests for each of the hypotheses given on page four. You may compare these to the results from the anova program to see that they are identical. (p. 9) 223

1) to test the hypothesis 3, of interaction

$$F = \text{MSS(interaction)} / \text{MSS (unexpalined)} = \left[.003 \Sigma y^2 / 2\right] / \left[.951 \Sigma y^2 / 1321\right]$$

$$= \left(\frac{.003}{2}\right)\left(\frac{1321}{.951}\right) = \frac{3.96}{1.90} = 2.08$$

2) To test hypothesis 1 — of no difference between the means in the two gender groups

$$F = \text{MSS(gender)} / \text{MSS(unexplained)} = \left(\frac{0}{1}\right)\left(\frac{1391}{.951}\right) = 0$$

3) To test hypothesis 2 — of no difference between the means in the three political groups

$$F = \text{MSS (political groups)} / \text{MSS (unexplained)} = \left(\frac{.045}{2}\right)\left(\frac{1321}{.951}\right) = 31.254$$

Note that we did not add the SS from interaction back into the error term. The ANOVA program does not do that unless you specifically ask it to do so. You could do it with the regression work if you wanted .

As you can see from either the ANOVA program or from looking up the cricitcal F values in a table we may fail to reject the hypotheiss of no interaction (although this is quite close) and fail to reject the hypothesis of differences between the gender groups; but we can reject the hypothesis of differences between the political groups at the .001 level of significance.

Using the regression equaitions given on pp. 2-4 we may also examine the differences between the liberty scores for each of the groups we are concerned with.
If we assume there is no interaction (i.e. we reject the null hypothesis three) we may use the regression eqution given in step two to analyse the influence of gender and political affiliation on liberties granted. From page 2 we get 216

$$\text{VAR16} = 2.687 + (-.0004)\,G + (.539)\,P1 + (-.095)\,P2$$

51 should be 52

Analysis of variance -- political liberties-- VAR16 (Y)

Source of Variation	Sums of Squares from regression	degrees of freedom	Mean Sum of Squares
1) Total variation	$\Sigma y^2 = \Sigma(Y-\bar{Y})^2$	N - 1	$\Sigma y^2 / n - 1$
2) Total explained -- between sub-class SS explained by gender, political views and interaction, $P1 \hat{+} P2 \quad I1 \hat{+} I2$	$R^2_{G,P1,P2,I1,I2} \Sigma y^2$ $= (.049)\Sigma y^2$	5	$(.049 \Sigma y^2) / 5$
3) Explained by additive model	$R^2_{G,P1,P2} \Sigma y^2$ $= (.046) \Sigma y^2$	3	$(.046 \Sigma y^2) / 3$
a) explained by gender adjusted for political views	$(R^2_{G,P1,P2} - R^2_{P1,P2})\Sigma y^2$ $= (.046 - .046)\Sigma y^2$ $= 0$	1	0
b) explained by political views adjusted for gender	$(R^2_{G,\ P1,P2} - R^2_{\ G})\Sigma y^2$ $= (.046 - .001)\Sigma y^2$ $= .045 \Sigma y^2$	2	$(.045 \Sigma y^2) / 2$
4) explained by interaction	$(R^2_{G,P1,P2,I1,I2} - R^2_{G,P1,P2})\Sigma y^2$ $= (.049 - .046)\Sigma y^2$ $= .003 \Sigma y^2$	2	$(.003 \Sigma y^2) / 2$
5) Unexplained	$(1 - R^2_{G,P1,P2,I1,I2})\Sigma y^2$ $= (1 - .049)\Sigma y^2$ $= .951 \Sigma y^2$	N - 6 = 1321	$(.951 \Sigma y^2) / 1321$

By looking at the calculations under step three above you can see why the results in the regular ANOVA program we did with this data produced sums of squares that did not add up in the way they do when you have equal cell sizes. Because the number of cases in each subcell are not equal (or proportional to the marginals) the ~~xxxxxx~~ effect of gender and political views on liberties are not independent of each other (as we could force them to be in a controlled experiment) and so some of the total explained variation in the additive model comes from this overlapping explanation which cannot be attributed solely to either independent variable.

should be 51

Substituting the values of VAR05 and P1 and P2 into the ~~same~~ equation for each of the groups under consideration we get the predicted values shown below. Again comparing these results to those obtained on the ANOVA printout, we see that these correspond to the category means predicted when we assume there is no interaction.

$$VAR16 = 2.687 - .0004(6) + .539P1 - .095P2$$

male liberals : $VAR16 = 2.687 - .0004(1) + .539(1) - .095(0) = 3.2256$

female liberals : $VAR16 = 2.687 - .0004(0) + .539(1) - .095(0) = 3.226$

male moderates : $VAR16 = 2.687 - .0004(1) + .539(0) - .095(1) = 2.5916$

female moderates : $VAR16 = 2.687 - .0004(0) + .539(0) - .095(1) = 2.592$

male conservatives : $VAR16 = 2.687 - .0004(1) + .539(0) - .095(0) = 2.6866$

female conservatives $VAR16 = 2.687 - .0004(0) + .539(0) - .095(0) = 2.687$

	liberals	moderate	conservative	
male	3.2256	2.5916	2.6866	> .0004
female	3.226	2.592	2.687	
	.634	−.095		

From examining these predicted means we see that the largest jump is between liberals + other political groups. When we examine the F ratios corresponding to the regression coefficients in this prediction equation we see indeed that the most important influence does come from being liberal. ~~It is the~~ The coefficient for P1 is the only one that has an F-ratio large enough to allow us to reject the null hypotheses.

You will remember that the F ratio testing the hypothesis of no interaction was high enough to reject the null hypothesis at the .05 level of significance. We can examine the prediction equation including the interaction terms (step 30 on p. 4) to see the effect of interaction.

prediction equation assuming there is interaction. ~~p.~~ 22

$VAR16 = 2.70 - .026 + .41 P1 - .05 P2 + .24 I1 - .13 I2$

predicted / actual values = subclass means

male liberals: $VAR16 = 2.70 - .02(1) + .41(1) - .05(0) + .24(1) - .13(0) = 3.33$

female liberals: $VAR16 = 2.70 - .02(0) + .41(1) - .05(0) + .24(0) - .13(0) = 3.11$

male moderate: $VAR16 = 2.70 - .02(1) + .41(0) - .05(1) + .24(0) - .13(1) = 2.52$

female mod: $VAR16 = 2.70 - .02(0) + .41(0) - .05(1) + .24(0) - .13(0) = 2.65$

male conservative: $VAR16 = 2.70 - .02(1) + .41(0) - .05(0) + .24(0) - .13(0) = 2.68$

female cons.: $VAR16 = 2.70 - .02(0) + .41(0) - .05(0) + .24(0) - .13(0) = 2.70$

	liberals	moderates	conservatives
males	3.33	2.52	2.68
females	3.11	2.65	2.70

Here we can see that there is a slight interaction effect with liberal males being more supportive than liberal females, but moderate + conservative males being less supportive than their female political counterparts. By examining the F ratios corresponding to the regression coefficients we see again that it is being liberal and also being liberal + male that produces the most influence.

On the next and final page is the printout from the Anova program on this same problem. (the result from the computer assn. # 2)

ANOVA
STATISTICS VAR16 BY VAR05, VAR18 (1,3)/
1,2

* * * * * * * * * * A N A L Y S I S O F V A R I A N C E * * * * * * * * * *

VAR16 OPINION ON HOMOSEXUALITY
BY VAR05 GENDER
VAR18 RECODED POLITICAL VIEWS

* *

SCURCE OF VARIATICN	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	99.208	3	33.069	21.291	0.001
VAR05	0.000	1	0.000	0.000	0.999
VAR18	98.085	2	49.043	31.575	0.001
2-WAY INTERACTIONS	7.627	2	3.813	2.455	0.084
VAR05 VAR18	7.627	2	3.813	2.455	0.084
EXPLAINED	106.835	5	21.367	13.757	0.001
RESIDUAL	2051.784	1321	1.553		
TCTAL	2158.619	1326	1.628		

1499 CASES WERE PROCESSED.
172 CASES (11.5 PCT) WERE MISSING.

* * MULTIPLE CLASSIFICATION ANALYSIS * * *

VAR16 OPINION ON HOMOSEXUALITY
BY VAR05 GENDER
VAR18 RECODED POLITICAL VIEWS

* *

;RAND MEAN = 2.81

/ARIABLE + CATEGORY	N	UNADJUSTED DEV'N	ETA	ADJUSTED FOR INDEPENDENTS DEV'N	BETA	ADJUSTED FOR INDEPENDENTS + COVARIATES DEV'N	BETA
/AR05							
1	615	0.03		-0.00			
2	712	-0.03		0.00			
			0.02		0.00		
/AR18							
1	393	0.42		0.42			
2	525	-0.22		-0.22			
3	409	-0.12		-0.12			
			0.21		0.21		

ULTIPLE R SCUARED 0.046
-ULTIPLE R 0.214

Types of Coding for "Dummy Variable" Regression

Kerlinger and Pedhazur discuss three types of coding that may be used in dummy variable regression. All of these produce the same results in terms of variation explained and the results of the various hypotheses tests. They differ in the form of the coding used and the way in which once may interpret the regression equations that result. In all cases the number of variables (what Kerlinger and Pedhazur call vectors) that are used to represent a given nominal variable is equal to the number of categories in that variable minus one.

The type of coding used in the examples earlier in these notes is called dummy coding by Kerlinger and Pedhazur. In this kind of coding each dummy variable or each vector is made up of zero's and one's. When a subject is in a group it is assigned a one, when it is not it is assigned a zero. One category then has codes of zeros in all the variables. This group or category may be seen as a "control group" or the one to which comparisons are made. (The example of dummy coding with three groups and a one-way analysis of variance can be consulted for an example of this method. Here the third group, A_3, may be seen as the "control group".) This method is then especially useful in experimental or other situations where one wants to compare all but one of the categories to that one group. The regression equation in this case yields an intercept that equals the mean of this control group. Each slope coefficient (b) equals the difference between the associated mean of the group designate independent variable and the mean of the control group.

A second method of coding is called effect coding. While in dummy coding one group is given zero's in all vectors or "dummy variables," in effect coding this group would be given the score of -1 in each of these vectors. The advantage of using this form of coding is that it gives results that are easily transferable to the general linear model (especially when there are equal n's in the subgroup). The R^2 and F-tests for testing hypotheses are exactly like those obtained with the other methods. There are however different regression equations. When there are equal sub-group

n's, the intercept equals the grand mean of the dependent variable. Each slope coefficient or b equals the "treatment effect," the difference of the mean of the group with which the vector is associated (where it has a +1) from the overall or grand mean. When there are unequal subgroup n's the intercept equals the unweighted mean of the subgroups (this means that the larger subgroups don't contributed as much to this overall mean). Each b then equals the difference of the subgroup mean from this unweighted mean. The data that were used in a dummy variable code analysis for one-way anova are shown below with effect coding. By the way, this coding is called effect coding because each b tells the effect of the treatment in terms of deviations from the overall mean.

Group	Subjects	Y	X_1	X_2
A_1	1	4	1	0
	2	5	1	0
	3	6	1	0
	4	7	1	0
	5	8	1	0
A_2	6	7	0	1
	7	8	0	1
	8	9	0	1
	9	10	0	1
	10	11	0	1
A_3	11	1	-1	-1
	12	2	-1	-1
	13	3	-1	-1
	14	4	-1	-1
	15	5	-1	-1
Σ		90	0	0
Σ^2		660	10	10
Mean		6	0	0
	5	2.9277	.84515	.84515

$R^2_{Y \cdot X_1 X_2} = .75$

$Y' = 6 + 0X_1 + 3X_2$

The third type of coding is ~~xxx~~ orthogonal coding. It is used when the researcher wants to make specific comparisons between the means of some of the groups involved. The type of coding used is determined by the kind of comparisons that one wants to ~~xx~~make. One can make as many orthogonal comparisons ~~xx~~ as there are vectors. Thus, if as in ~~x~~our example there are three categories, one can make two orthogonal comparisons. Two comparisons are called orthogonal when the sum of the products of tneir ooefficients for tneir elements is zero. In the first comparison below, as shown, the comparisons are orthogonal; in the second, however, they are not.

① $D_1 = (1)(\bar{Y}_1) + (-1)(\bar{Y}_2) + (0)(\bar{Y}_3)$

$D_2 = (-\frac{1}{2})(\bar{Y}_1) + (-\frac{1}{2})(\bar{Y}_2) + (1)(\bar{Y}_3)$

$(1)(-\frac{1}{2}) + (-1)(-\frac{1}{2}) + (0)(1) = 0$

② $D_3 = (1)(\bar{Y}_1) + (-1)(\bar{Y}_2) + (0)(\bar{Y}_3)$

$D_4 = (-1)(\bar{Y}_1) + (0)(\bar{Y}_2) + (1)(\bar{Y}_3)$

$(1)(-1) + (-1)(0) + (0)(1) = -1$

When one uses orthogonal coding in comparions one simply uses the coefficients in the hypothesized contrasts as the codes. One the next page the same example used earlier is repeated with orthogonal coding. The coefficients in the first comparison above that was shown to be orthogonal are used (those ~~xxx~~ in the second vector were simply multipled by two to simplify computations). Note that because these two vectors are orthogonal, their intercorrelation is zero (orthogonal means uncorrelated). Note that the same value is obtained for R^2, but that the regression equation is somewhat different. Here tne intercept equals the grand mean (for both cases with equal and unequal subgroup ~~x~~ sizes). The slope coefficients relate to the specific comparison involved. The F-ratio (or t-ratio) associated with each slope actually tests the hypothesis that the two means in tne comparison are equal. In this example $b_{yx_1 \cdot x_2}$ is associated with the comparison between A_1 and A_2 (because in X_1) their sums add to zero. $b_{yx_2 \cdot x_1}$ is associated with the comparison between A_3 and the average of the means of A_1 and A_2.

Example of Orthogonal Coding

Group	Y	X_1	X_2
A_1	4	1	-1
	5	1	-1
	6	1	-1
	7	1	-1
	8	1	-1
A_2	7	-1	-1
	8	-1	-1
	9	-1	-1
	10	-1	-1
	11	-1	-1
A_3	1	0	2
	2	0	2
	3	0	2
	4	0	2
	5	0	2
	6	0	2
Σ	90	0	0
mean	6	0	0
SS	120	10	30
s	2.92770	.84515	1.46385

$R^2 = .75$

$Y' = 6.0 - 1.5X_1 - 1.5X_2$

$r_{YX_1} = -.43301$

$r_{YX_2} = -.75$

$r_{X_1X_2} = .000$

There are a few differences in the notation that Kerlinger and Pedhazur use from that we have used previously. They refer to the regression sum of squares. This is simply the unexplained sum of squares that we have used in testing hypotheses related to regression. (The term regression sum of squares is the same term that SPSS uses in the printout and book.)

In table 10.3, page 236, the variable labeled 4 is simply the interaction term. What Kerlinger and Pedhazur try to do in the first part of this chapter is actually to present the logic of analysis of covariance and only introduce the term itself in the last chapter. This is probably to avoid the problems associated with just seeing analysis of covariance as associated with experimental designs.

On page 238, the last paragraph, the term multiple comparisons refers to comparisons between means to see which are significantly different from each other.

Pp. 245 ff, note here the distinction between ordinal and disordinal interactions and calculating the point of intersection with interaction effect. This was not mentioned in the notes or in Blalock, but is a useful way of describing the nature of the interaction.

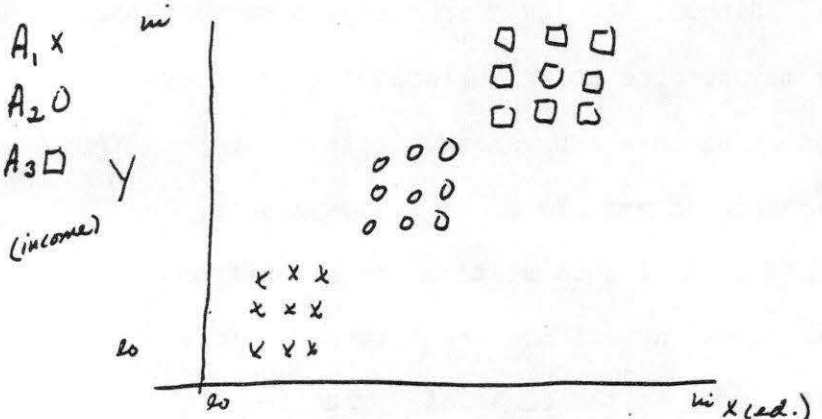
In pp. 260-265 they discuss non-linear relations and the use of these in regression. The technique they use is the common one of using a squared or cubed term (or even higher order) and simply adding this to the regression equation such as in $Y' = a + b X + b X^2$.

On page 274 MSR refers to the mean square of the residuals (mean square residuals) this is the mean square or estimate of variance for the sum of squares (variation) associated with the error term or within estimate.

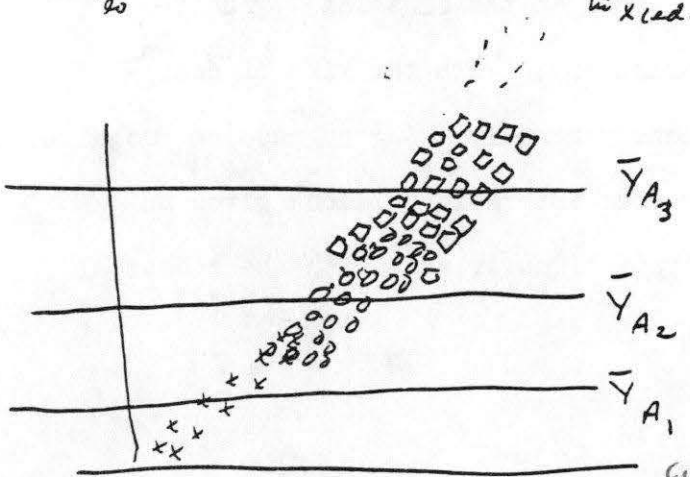
X. Analysis of Covariance

In this section we explore analysis of covariance. This is a technique that combines analysis of variance and regression. Continuing from the work of last term we will discuss the logic in analysis of variance terms and also show how the work may be done using regression techniques. In the first part of this section we discuss the general logic of the technique. In the second part we show an example of an analysis using regression techniques. In the last part of this section the pages from the SPSS manual that discuss analysis of covariance are attached. This includes the analysis of covariance runs in the classical tradition (not using regression and dummy variables). With the kind of data sociologists usually have I recommend generally using regression techniques. The other material is appended for your interest. Blalock gives an example of working through an analysis of covariance with the classical techniques by hand.

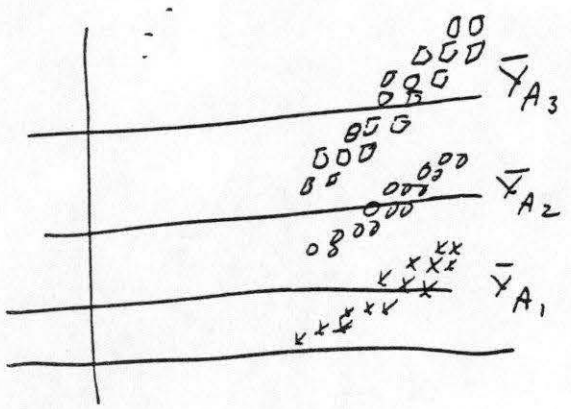
In analysis of covariance we have a dependent variable that is intervally measured and two or more independent variables, one of which is nominally measured and the other measured on an interval scale. (The additional independent variables may be either intervally or nominally measured.) The diagrams below illustrate possible cases where you might want to employ analysis of covariance. You might want to examine the influence of the intervally measured independent variable (called X) on the dependent variable (called Y) when the influence of the nominally measured independent variable (A) is controlled or look at the influence of A on Y when X is controlled.



In this hypothetical situation you can see that when we control for A the relation between X and Y disappears. Substantively, if Y were income and X were education, and A were racial groups this configuration would suggest that race was the determining factor in income rather than education (because within each racial group the association between X and Y disappears, yet the differences in Y between each group of A remain



In this hypothetical situation we see that education retains an effect on income



In this hypothetical situation education does not appear to affect Y because all groups of A have the same range of values in X, but very different Y values. In fact, if this situation appeared in your scatter diagram it would likely be almost a waste of time to do an analysis of covaraiance (especially if you were doing it by hand).

Figure 1-1

An intuitive explanation of analysis of covariance:

If you remember how analysis of variance and especially two-way analysis of variance worked, you will easily understand the logic underlying analysis of covariance. In analysis of variance we were concerned with accounting for the variation and we broke up the sums of squares, $\Sigma\Sigma(Y-\bar{Y})^2$, into portions that could be explained by the independent variables and partions that were unexplained. Analysis of covariance is analagous to this except that here we are concerned with covariance instead of variance or actually with covariation instead of variation. ($\Sigma\Sigma(X-\bar{X})(Y-\bar{Y})$ instead of $\Sigma\Sigma(Y-\bar{Y})(Y-\bar{Y}) = \Sigma\Sigma(Y-\bar{Y})^2$)

As with analysis of variance, the total covariation $\Sigma\Sigma(X_{ij}-\bar{X}..)(Y_{ij}-\bar{Y}..)$ can be broken into the unexplained covariation, the covariation within each of the $\Sigma\Sigma(X_{ij}-\bar{X}_{\cdot j})(Y_{ij}-\bar{Y}_{\cdot j})$ categories of the nominally measured variable (i.e. that which is unexplained by either the nominally measured variable or the intervally measured independent variable), and the explained covariation - analagous to the between variation, the covariation of the X and Y means in each category around the grand means of the two intervally measured variables. $\Sigma\Sigma(\bar{X}_{\cdot j}-\bar{X}..)(\bar{Y}_{\cdot j}-\bar{Y}..)$

As with two-way analysis of variance this total explained variation can be broken into even more parts: that explained by the nominally measured variable A, that explained by the intervally measured variable X, and that explained by interaction between these two independent variables. Also, as with analysis of variance, we divide the covariation by the appropriate degrees of freedom to get estimates of the covariance, examine the ratios of the explained and unexplained estimates of the covariance, and use the F-distribution to check if the results could be obtainable by chance.

To complete the analysis of covariance it is necessary that there be no interaction. And as with analysis of variance we can examine the data to see if there is indeed interaction. Here, however, we don't compare the magnitude of means in each category to have additivity but the slopes of the regression line (b_{yx}) in each category of A. The figures below illustrate the case of no interaction and interaction. The tables below the figures show (in gross ways) what the same data would look like if X were divided into three categories on a nominally measured scale. This illustrates how the concept of interaction is basically the same in analysis of variance and analysis of covariance.

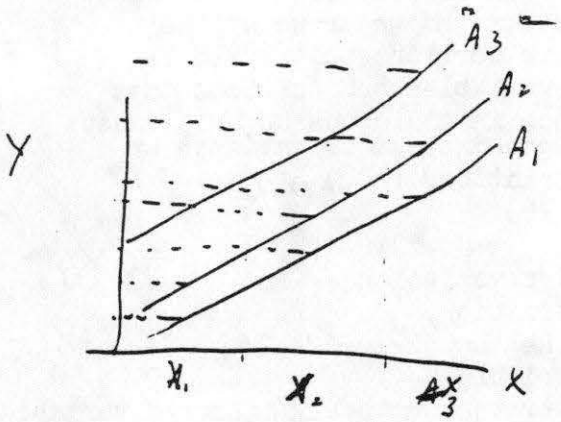
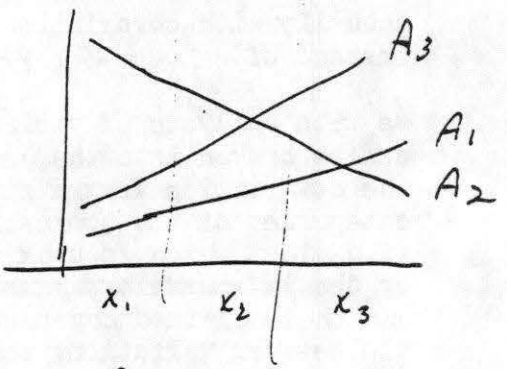


Figure 1-2



	A_1	A_2	A_3
X_1	3	5	10
X_2	8	10	15
X_3	12	14	19

	A_1	A_2	A_3
X_1	3	17	5
X_2	6	13	11
X_3	9	9	17

In two-way analysis of variance the between subclass sums of squares was the total explained variation (including interaction plus that explained by the two independent variables). Here the analagous term is the total explained covariation given above, that explained by X and A together.

The variation in Y that is unexplained when we assume there is no interaction between X and A will always be greater than or equal to the variation that is unexplained when we assume that there is interaction. That is, if we let interaction explain some of the variation it can do nothing but explain more of the variation; it cannot add to the unexplained variation. This is equivalent to the situation with analysis of variance. What we do then to look at how much variation is actually explained by interaction between X and A is use the total that is unexplained when we assume there is no interaction (i.e. assuming that X and A each separately influence Y) and compare this to how much is unexplained when we assume that there is interaction (i.e. assume that X and A may act together in their influence on Y). The difference between these two figures is then how much variation is actually explained by interaction.

When there is no interaction the slopes (b_{yx}) in each category of A are equal. Thus, when we assume there is no interaction we use a common slope (b) for all the categories of A. This is simply a weighted average of the slope within each of the separate categories of A. If this common slope is a good estimate of the actual slopes in the categories then the variation in Y explained by using this estimate should be about equal to that when the individual slopes in each category are used. Any difference between these two figures may be attributed to variation explained by interaction. This amount is treated as the explained or ~~within~~ between variation. The amount of variation unexplained when we assume interaction is the error, ~~or~~ within, or unexplained. Using the appropriate degrees of freedom we may get estimates of the variances, have ratios of the explained to unexplained variances and test the hypothesis that there is no interaction.
If we ~~find~~ find that there is no interaction, Blalock suggests putting the variation explained by interaction back into the error term and then directly examining the variation of Y explained by X and that explained by A (when controlling for the other variable).

Looking at variation explained by X (the intervally measured variable) when we control for A:
Note: we can go to this step only if there is no interaction (or if the interaction was shown to be a non-significant contribution to the explanation of the variation of Y)
Because the regression line in each of the categories of A has a common slope, we can compute an average correlation coefficient for the association between X and Y in all the categories of A. This is called the average within ~~cat~~ class coffelation coefficient, $r_{xy \cdot a}$, and is analagous to the partial

within
w. type

correlation coefficient. This average within class correlation coefficient is simply a weighted average of the correlation between X and Y within each category of A. It only makes sense if there is no interaction, that is, if the slopes of the regression lines are the same in each category of A. $r^2_{xy \cdot a}$ may be interpreted as simply the proportion of the variation of Y that is explained by X when A is controlled (when the influence of A is removed). The hypothesis that $\rho_{xy \cdot a} = 0$ may be tested using the familiar analysis of variance procedures.

Note how this within class correlation coefficient has an advantage over the partial correlation coefficient. With partials, we assume that there are linear associations; with the within-class coefficient we can actually test this by looking at the possibility of interaction.

Looking at the variation in Y explained by A (the nominally measured variable) when we control for X (the intervally measured variable):
Here we are testing the hypothesis that the means of Y in each category of A are equal when we remove the influence of X on A. We can't really "control" for X here, but we can adjust for the influence of X. The term control implies a "holding constant," looking at the influence of A and Y in "categories" of X. But X is here operating as an intervally measured variable. As X changes, Y changes. This rate of change is measured and is equal to the slope, b_{yx}, the change in y for each unit change in x.

$$b_{yx} = \Delta y / \Delta x .$$

When we adjust for X we essentially want to adjust for or control the influence of X on Y. This we can do by holding X constant, adjusting the Y values to this constant value of X and then looking at the influence of A on Y by comparing the adjusted Y means in each category of A.

Again, this step only makes sense when we can assume that there is no interaction between X and A in their influence on Y. Because we can assume there is no interaction, we have an estimate of the common slope of the regression line of X predicting Y in the categories of A. This is the average within class b_{yx}, which is equal to the predicted change in y for each change in x.

To adjust $\bar{Y}$ in each category of A for the influence of X we use this common slope. The average value of X varies within categories of A. To adjust for the influence of X we adjust each of these values of X to a common value, and then see what influence this has on the Y values. We know that $b_{yx} = \Delta y / \Delta x$. Then, using simple algebra,

$\Delta y = (\Delta x) \times (b_{yx})$. We know the value of b_{yx}.

If we let $\Delta x = (\bar{X}.. - \bar{X}._j)$, the difference of the grand mean of X and the mean of X in each category of A, and $\Delta y = (\bar{Y}'.. - \bar{Y}'._j)$, the difference betwee

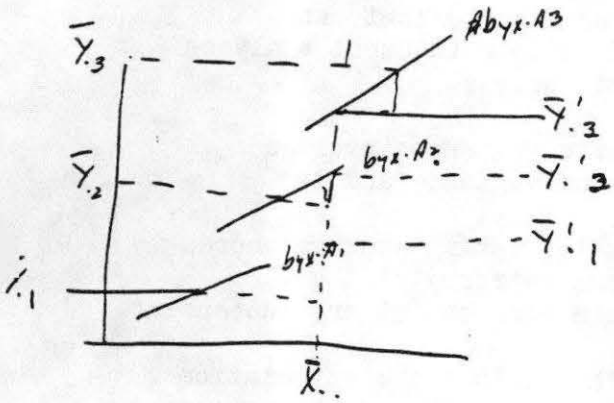
then $\Delta y = b_{yx} (\bar{X}.. - \bar{X}.j)$. Δy is simply $(\bar{Y}'_{.j} - \bar{Y}.j)$, the difference of the actual mean of Y in the category, j, of A and the predicted mean value of Y when we adjust for the change in X.

Then $\bar{Y}'_{.j} = \bar{Y}.j - b_{yx} (\bar{X}.j - \bar{X}..)$. (Remember that b_{yx} refers to the average within class correlation coefficient.)

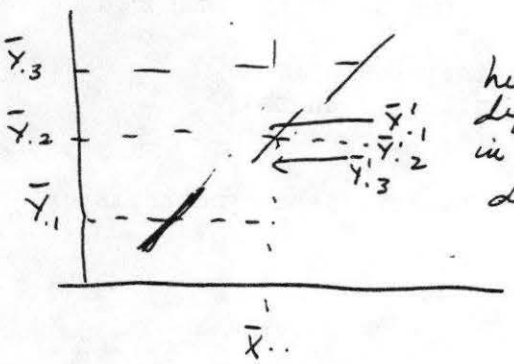
(you can also see the above equation as $\bar{Y}'j = \bar{Y}.j + b (\bar{X}.. - \bar{X}._j)$.)

If you then do this adjustment process in each category of A you get adjusted means of Y, $\bar{Y}'_j$, for each category of A. You can then use these adjusted means of Y in testing the hypothesis that the means of Y in the categories of A are equal, knowing that the influence of X on Y is removed. The regular analysis of variance format is used in testing this hypothesis.

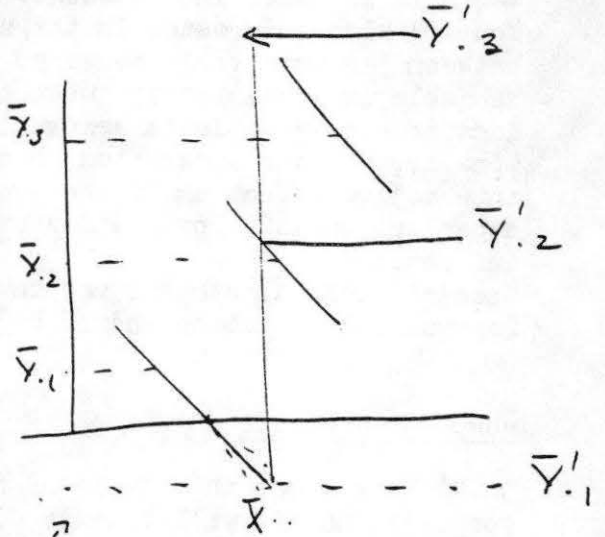
The diagrams below illustrate possible adjustments that may be made and the results.



here the differences between the $\bar{Y}.j$'s remain after adjusting for X, but are slightly smaller



here the difference in $\bar{Y}.j$ virtually disappears



here the difference between $\bar{Y}.j$ become larger after controlling for X.

Analysis of covariance may be extended to incorporate more than one nominally measured variable and/or more than one intervally measured independent variable. However, hand computations become extremely tedious with these additions, especially if the n's in the subcells are unequal. Because of the tedious nature of the hand computations, I suggest using dummy variables in regression analysis in dealing with analysis of covariance. If the interadtion term provesto be significant, using regression with dummy variables will allow you to continue the analysis.

You may continue the analysis if interaction occurs by looking at the pattern of interaction, seeing where it occurs. You may also then (should really) try to see if it follows some kind of pattern, if perhaps by ordering the categories of the nominal variable from those with the smallest to largest slope there is some underlying variable (maybe interval in measurement) that can explain this apptern. This can then be added to subsequent analyses to try to understand more about the relationship.

It is important to realize how analysis of covariance involves both types of hypotheses used in analysis of variance and in regression analysis.
You are interested in the hypothesis that the average scores of the dependent variable are equal in each category of the nominally measured independent variable once the other independent variable is taken into account.
You are also interested in the null hyp thesis that the association between the intervally measured independent variable and the dependent variable is equal to zero when the influence of the nominally measured independent variable is removed.
Finally, you are interested in the hypothesis that there is no interactive effect of the two independent variables on the dependent variable over and above their separate, independent influences.
Each of these hypotheses may be tested using F-ratios. The hypothesis regarding interaction should be tested first.

Computer Analysis

Blalock gives an example of doing analyis of covariance with hand computations. I will briefly discuss here results of an analysis with regression. teachniques. The same format used with analysis of variance and discussed in last term's notes is used.

This analysis was done with the 1972 (I think) NORC data. The dependent variable is male respondents' occupational prestige. The two independent variables were the respondents' self-indication of class status (either lower, working, or middle -- there were so few calling themselves upper class that they were omitted) and the respondents' educational level. The class status variable in the notes below is var09. The education variable is var26. Occupational prestige is var04.
Dummy variables X1 and X2 computed below represent the class level. The interaction terms ED1 and ED2 represent the interaction between education and class status.
The following computer instructions were needed.

```
GET FILE          WHATEVER

IF                (VAR09 EQ 3) X1=1
IF                (VAR09 EQ 2) X2=1

COMPUTE           EDX1 = X1 * VAR26
COMPUTE           EDX2 = X2 * VAR26

ASSIGN MISSING    X1,X2,EDX1,EDX2 (99)

REGRESSIOn        VARIABLES=VAR09,VAR26,X1,X2,EDX1,EDX2/
                  REGRESSION= VAR09 WITH  X1,X2 (8) VAR26 (6) EDX1, EDX2 (4)/
                  REGRESSION= VAR09 WITH VAR26
```

The resulting R^2 and the analysis of covariance table written in the same manner as analysis of variance tables and tables testing hypotheses regarding regression is given below.
Note that with an F= .9517 we may fail to reject the null hypothesis that there is no interaction.
The SS due to interaction was then added back into the unexplained SS assuming that this SS was simply due to chance.
The F=239.35 with df=1,687 tells us that we can reject the null hypothesis that there is no association between education and occupational prestige when we remove the influence of subjective class placement and be wrong in rejecting this null hypothesis less than once out of 1000 times.
The F=6.462 with df= 2,687 is very close to significance at the .05 level. There is some slight indication then that we could perhaps reject the null hypothesis that there are equal means of occupational prestige in each subjective class category when we remove the influence of educational level attained. Note however how only about one percent of the variation in occupational prestige is explained by this variable, compared to about 24% by education apart from class status. (See SS column below.) Together both of these variables do have a significant impact (F=104.96, df=3,687)

; ; time we get the following R^2 's w/ $n = 691$

$R^2_{X_1X_2} = .0754$ (R^2 when use both X_1 & X_2 to predict occ. prestige)

$R^2_{ED} = .3014$ (R^2 when predict Y, occ. prestige, from ed-level)

$R^2_{X_1X_2,ED} = .3143$ (R^2 when use X_1X_2 + ed to predict Y)

$R^2_{X_1,X_2,ED,EDX1,EDX2} = .3162$ (R^2 when add interaction terms)

and this yields the following analysis of variance table.

Source of variation	SS	df	F
1) Total explained	(.3162) SS_y	$k_1+k_2+k_1k_2=K$ $=5$	~~$\frac{(.3162)}{5}\frac{(687)}{(.6857)}$~~ =
2) Total explained wo interaction	(.3143) SS_y	$k_1+k_2=3$	$\frac{(.3143)}{3}\frac{(687)}{(.6857)} = 104.96$
a) by class alone	(.3143 − .3014) SS_y = (.0129) SS_y	$k_1=2$	$\frac{(.0129)}{2}\frac{(687)}{(.6857)} = 6.462$
b) by ed-alone	(.3143 − .0754) SS_y = (.2389) SS_y	$k_2=1$	$\frac{(.2389)}{1}\frac{(687)}{(.6857)} = 239.35$
3) Interaction	(.3162 − .3143) SS_y = .0019 SS_y	$k_1k_2=2$	$\frac{(.0019)}{(2)}\frac{(685)}{(.6838)}$ = .9517
4) Error	(1 − .3162) SS_y = (.6838) SS_y	$N-1-K=$ $691-6=685$	

Because the SS explained by interaction does not add an amount significantly greater than zero to the explanation of Y we can add this SS explained by interaction back into the error SS (within SS) and get error SS = .6857 SS_y, df = 687.

The F-ratios ~~may~~ may then[70] be computed.

With the prediction equation we may predict values of occ. prestige in each category of class status:

~~mid~~ prediction eq. (wo. interaction): $\tilde{Y} = 14.008 + 2.620 X_1 - .677 X_2 + 2.073 E$

middle class: $\tilde{Y} = 14.008 + 2.62 + 2.073 Ed = 16.628 + 2.073 Ed$

working class: $\tilde{Y} = 14.008 - .677 + 2.073 Ed = 13.331 + 2.073 Ed$

lower class: $\tilde{Y} = 14.008 + 2.073 Ed$

With interaction

$$\tilde{Y} = 17.191 - 2.412 X_1 - 2.354 X_2 + 1.649 Ed + .574 Ed X_1 + .285 Ed X_2$$

for middle class: $\tilde{Y} = 17.191 - 2.412 + 1.649 Ed + .574 Ed$

$= 14.779 + 2.223 Ed$

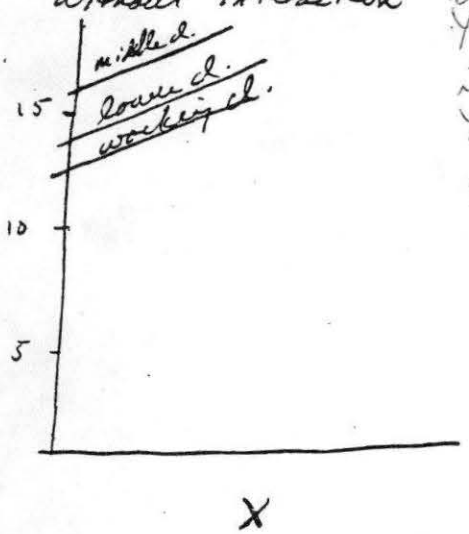
working class

$\tilde{Y} = 17.191 - 2.354 + 1.649 Ed + .285 Ed$

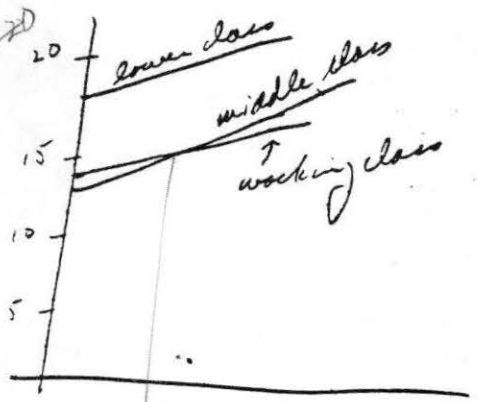
$= 14.837 + 1.934 Ed$

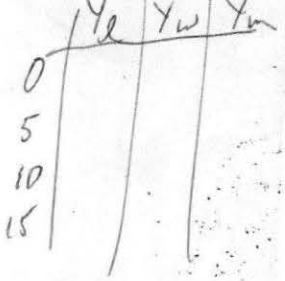
lower class $\tilde{Y} = 17.191 + 1.649 Ed$

without interaction



with interaction





Even though the diagrams on page 11 showed some interaction, the F-ratios showed that this was not significant. The largest difference in slopes is from the lower class to the others. This group had the smallest number of subjective choices and may have affected the results. Note, however, that when interaction terms are included the predicted slopes as well as the Y-intercepts will change. If interaction terms had been signifcant one would want to examine the nature of each of these terms and their associated F-ratios.

Besides doing analysis of covariance via regression it can be done with the ANOVA program used in the analysis of variance work last term. As with the ANOVA printout, however, this program gives fairly abbreviated results, especially if there is interaction. The program with analysis of covariance is exactly as with analysis of variance except that the covariates (the independent variables that are nominally measured) are added after WITH on the anova card.

Using Categoric Variables in Regression -

→ Dependent Variable must be measured on an interval scale

→ Code dichotomous independent variables 0,1
[Other coding schemes are also possible, but interpretations vary - See Pedhazur.]

→ If independent variables are categoric and have more than 2 attributes - (say k attributes) use k-1 dummy variables - coded as below

		Dummy Variables		
Original Independent Variable codes		X_1	X_2	X_3
	1	1	0	0
	2	0	1	0
	3	0	0	1
	4	0	0	0

The actual schema may vary. The important point is that k-1 dummy variables are used.
All the dummy variables are coded 0,1 with 1 indicating presence of an attribute, 0 its absence
One category is designated as the omitted category — coded 0 in all the dummy variables

→ To analyze the effect of the categoric variables on the dependent variable, ~~[illegible]~~ multiple regression is used.

Possible cases:

A - One dichotomous independent variable, X_1, coded 0 - 1
result:

$Y = a + b_{YX} X_1$

R^2 indicates the percentage of variance in Y explained by X_1.
The F associated with R^2 ~~[illegible]~~ tests the significance of this effect and is equivalent to t^2 in the [illegible] test of the [illegible]

A (continued)

a is the predicted value of Y when $X=0$ (i.e. $\bar{Y}$ when $X=0$)

b is the expected change in Y when X changes from 0 to 1 (i.e. $\bar{Y}_{X=0} - \bar{Y}_{X=1}$)

$a+b = \bar{Y}_{X=1}$ the expected value of Y when $X=1$

the t associated with "a" tests

$H_0: \mu_{Y(X=0)} = 0$

the t associated with "b" tests

$H_0: \mu_{Y_{X=0}} - \mu_{Y_{X=1}} = 0$

B: One independent variable, with k categories - transformed into k-1 dummy variables

$X_1, X_2, \ldots, X_{k-1}$ all coded 0, 1

$$Y = a + b_{YX_1} X_1 + b_{YX_2} X_2 + \ldots + b_{YX_{k-1}} X_{k-1}$$

R^2 indicates the percentage of variance in Y that is explained by the various categories of the independent variable (Note if this variable is categorized differently, R^2 will probably be different.)

The F associated with R^2 tests the significance of the effect and is equivalent to the F test obtained in the one-way anova testing

$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

a in the regression equation is the predicted value of Y when $X_1 = X_2 = X_3 = \ldots = X_{k-1} = 0$ (i.e. for the omitted category in the dummy variables).

B (continued)

b_{YX_i} is the expected change in Y when X changes from the omitted category to the category represented by X_i (remember all other categories are controlled)

$a + b_{YX_i}$ is the expected value of Y when $X_i = 1$ ($\bar{Y}_{X_i=1}$)
in general,
$a + b_{YX_i}$ is the expected value of Y when $X_i = 1$ ($\bar{Y}_{X_i=1}$)

C: Two independent variables - each dichotomous, both coded 0, 1
X_1 and X_2

Here Y may vary as a function of X_1, X_2 and any special additional effect of the combination of categories of $X_1 \times X_2$ (interaction)
This interaction is represented by a dummy variable equal to the product of $X_1 X_2$ + is also coded 0, 1

here $Y = a + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2$

(1) First one must test of $X_1 X_2$ - the interaction effect - adds anything to the explanatory power of $X_1 + X_2$ separately
This is done by testing $H_0: R^2_{Y \cdot 123} - R^2_{Y \cdot 12} = 0$

If this hypothesis is rejected, the separate effects of $X_1 + X_2$ cannot be examined + one must examine the full equation above

Here R^2 indicates the percentage of variance in Y explained by $X_1 + X_2$ + interaction
+ the F associated with this test the hypothesis that this =0

c (continued):

a = the expected value of Y when $X_1 + X_2 = 0$
b_1 = the expected change in Y when $X_1 = 1$
b_2 = the expected change in Y when $X_2 = 1$
b_3 = the expected change in Y when both $X_1 + X_2 = 1$

$a + b_1$ = expected value of Y when $X_1 = 1 + X_2 = 0$
$a + b_2$ = expected value of Y when $X_1 = 0 + X_2 = 1$
$a + b_1 + b_2 + b_3$ = expected value of Y when $X_1 = 1 + X_2 = 1$

(2) If one fail to reject $H_0: R^2_{Y.123} - R^2_{Y.12} = 0$ one examines

$$Y = a + b_1 X_1 + b_2 X_2$$

here a = expected value of Y when $X_1 = X_2 = 0$
b_1 = expected change in Y when $X_1 = 1$ $(\bar{Y}_{X_1=1} - \bar{Y}_{X_1=0})$
b_2 = expected change in Y when $X_2 = 1$ $(\bar{Y}_{X_2=1} - \bar{Y}_{X_2=0})$

$a + b_1$ = expected value of Y when $X_1 = 1, X_2 = 0$
$a + b_2$ = expected value of Y when $X_1 = 0, X_2 = 1$
$a + b_1 + b_2$ = expected value of Y when $X_1 = 1 = X_2$

t associated with a test $H_0: \mu_{Y(X_1 = X_2 = 0)} = 0$

t associated with b_1 test $H_0: \mu_{Y_{X_1=0}} = \mu_{Y_{X_1=1}}$ (when X_2 con

is equivalent to $\sqrt{F}$ in the 2 way anova w/ $X_1 + X_2$ as independent variables

t associated with b_2 test H_0 $\mu_{Y_{X_2=0}} = \mu_{Y_{X_2=1}}$ (when X_1 is conste

and is equivalent to $\sqrt{F}$ in the 2-way anova associated w/ X_2
(with $X_1 + X_2$ as independent variables)

O: Two independent variables - both categoric - but at least one with more than 2 attributes assume

1st categoric variable represented by dummy variables
$A_1, A_2, \ldots, A_{k-1}$, all coded 0,1

2nd categoric variable represented by dummy variables
$B_1, B_2, \ldots, B_{k-1}$

the interaction effect of A & B is represented by
$A_1B_1, A_2B_2, \ldots A_1B_{k-1}, A_2B_1, A_2B_2, \ldots, A_2B_{k-1}, \ldots$
$A_{k-1}B_1 \ldots A_{k-1}B_{k-1}$ (all coded 0,1)

the full model (main effects + interaction) would be represented by

~~[illegible]~~

$$Y = a + b_{YA_1}A_1 + \ldots + b_{YA_{k-1}}A_{k-1} + b_{YB_1}B_1 + \ldots + b_{YB_{k-1}}B_{k-1}$$
$$+ b_{YA_1B_1}A_1B_1 + \ldots + b_{YA_{k-1}B_{k-1}}A_{k-1}B_{k-1}$$

the model with only main effects would be represented by

$$Y = a + b_{YA_1}A_1 + \ldots + b_{YA_{k-1}}A_{k-1} + b_{YB_1}B_1 + \ldots + b_{YB_{k-1}}B_{k-1}$$

the effect of interaction is tested via

$$H_0: R^2_{Y \cdot A_1 \ldots A_{k-1}B_1 \ldots B_{k-1}} - R^2_{Y \cdot A_1 \ldots B_{k-1}} = 0$$

if the hypothesis is rejected the regression equation associated with the full model is examined - as described under C.1.

if this hypothesis is not rejected the regression equation with only the main effects is examined

0 (continued)

In the regression equation with only main effects

a = the expected value of Y when $A_1 = A_2 = \dots A_{k-1} = B_1 = B_2 = \dots B_{k-1} = 0$

b_{YA_1} = the expected change of Y when $A_1 = 1$ + all other values are 0

b_{YA_2} = the expected change in Y when $A_2 = 1$ + all other dummy variables = 0

$\vdots$

$b_{YB_{k-1}}$ = the expected change in Y when $B_{k-1} = 1$ + all other dummy variables = 0

$a + b_{YA_1}$ = the expected value of Y when $A_1 = 1$ + $A_2 = A_3 = \dots = A_{k-1} = B_1 = \dots = B_{k-1} = 0$

$a + b_{YA_2}$ = the expected value of Y when $A_2 = 1$ + all other dummy variables = 0

$\vdots$

$a + b_{YA_{k-1}} + b_{YB_1}$ = the expected value of Y when $A_{k-1} = 1 = B_1$ + all others are = to zero

+ so on

t - values associated w/ a ~~[illegible]~~ test to $H_0: \mu_{Y_{A_1 = A_2 = \dots A_{k-1} = B_1 = \dots B_{k-1} = 0}} = 0$

(i.e. the hypothesis that the mean of Y in the omitted category of both variables = 0)

t - values associated w/ each b test the H_0 $b = 0$

(i.e. that the difference between the mean of the omitted category + that represented by the mean = 0)

to test $H_0: \mu_{A_1} = \mu_{A_2} = \dots = \mu_{A_{k-1}}$ (controlling for B) = (what is tested in 2-way anova & - in main effects)

you need to test $H_0: R^2_{Y \cdot A_1 \dots A_{k-1} B_1 \dots B_{k-1}} - R^2_{Y \cdot B_1 \dots B_{k-1}} = 0$

+ to test

$H_0: \mu_{B_1} = \mu_{B_2} = \dots = \mu_{B_{k-1}}$

you need to test

$H_0: R^2_{Y A_1 \dots A_{k-1} B_1 \dots B_{k-1}} - R^2_{Y \cdot A_1 \dots A_{k-1}}$

XI. Factor Analysis

Basic Ideas

Factor analysis is based on the general linear model. Variance (note not variation) is used as the measure of total error in prediction and we look at ways that we can explain this variance. In standard regression analysis we try to explain the variation in one variable (say Y the dependent variable) by looking at the influence of specific other variables ($X_1, X_2, \ldots$) In factor analysis we try to explain the variance in one variable in a set of variables by what it hodls in common with other variables in that set. This is not a causal analysis, but a reduction of the variance of each variable in a set into factors held in common with other variables in the set and a portion that is unique to that variable. The solutions obtained then in a factor analysis vary with the variables included in the analysis.

If all the variables are seen as standard scores (mean = 0, standard deviation and variance =1), then the parts of the variance explained by other variables and unique to the variable may be seen as proportions (part of one)

The variance of a variable = 1 and may be broken into

1 =	what is held in common with other variables in the group	+	what is unique to that variable	
=	communality	+	uniqueness	
=	h^2	+	$(b^2 + c^2)$	.

The communality (represented commonly as h^2) gives the proportion of variance that is held in common with the other variables. The uniqueness ($b^2 + c^2$) gives the proportion that is unique to that variable. The two parts of the uniqueness are

b^2, representing specificity, what is specific to that variable, and

c^2, error variance --what occurs by error in measurement.

This error variance, c^2, is also called the unreliability of a variable.

Together $h^2 + b^2 = 1 - c^2$ is termed the reliability of the variable, that proportion which is not attributable to error.

Sometimes researchers use an index of completeness of factorization. This equals

$$H_j = 100\ (h_j^2 / (h_j^2 + b_j^2)) = 100\ (\text{communality/reliability}).$$

This index ranges from zero to one and gives the percentage of the reliable variance of a variable that can be accounted for by the common factors. (Note the analogue to a PRE measure.)

When you have a set of n variables you may represent each variable as a linear function of the common (called factor) and unique components. This pattern of linear relations is called a factor pattern. Note how similar it is to the regression equations and to structural equations used in path analysis. In the factor pattern shown below there are n variables, represented by z_i. Each F_i represents a factor or some part of the variables that are held in common. In this case there are m different factors that represent the variance the variables hold in common. The a_{ij}'s are the coefficients (called loadings) that tell us how much influence each variable has from a factor. These coefficients range from -1.00 to +1.00 just as beta weights do. As we see below they may be interpreted in the same way as beta weights may. The U_i's in the equations represent the part of the variance of each variable that is unique to it. (Note that in these coefficients the subscript i refers to the variables and j refers to the factors. We will try to keep this convention throughout.)

Table 2-1
A Factor Pattern

$$z_1 = a_{11} F_1 + a_{12} F_2 + \cdots + a_{1m} F_m + a_1 U_1$$

$$z_2 = a_{21} F_1 + a_{22} F_2 + \cdots + a_{2m} F_m + a_2 U_2$$

$$\vdots$$

$$z_n = a_{n1} F_1 + a_{n2} F_2 + \cdots + a_{nm} F_m + a_n U_n$$

The number of common factors for a set of variables is referred to as the complexity.

Each loading or influence of a factor on a variable is represented above by the a_{ij}'s. If the factors are not correlated with each other, then each of the a_{ij}'s is the correlation between the variable and the factor. In other words, a_{11} above gives the correlation between factor one and variable z_1, a_{n2} gives the correlation between Factor two and variable z_n, if the factors are not correlated with each other.

The proof of this statement is given below.

Proof - that $a_{ij} = r_{ij}$ when factors are uncorrelated

1) $z_1 = a_{11} F_1 + a_{12} F_2 + \ldots + a_{1m} F_m + a_1 U_1$ (given - see Table 2-1)

2) $F_1 z_1 = a_1 F_1 F_1 + a_{12} F_2 F_1 + \ldots + a_{1m} F_m F_1 + a_1 U_1 F_1$ (by multiplication)

3) $\frac{\Sigma F_1 z_1}{N} = a_{11} \frac{\Sigma F_1 F_1}{N} + a_{12} \frac{\Sigma F_2 F_1}{N} + \ldots + a_{1m} \frac{\Sigma F_m F_1}{N} + a_1 \frac{\Sigma U_1 F_1}{N}$

(sum over all values + divide by N -)
an arithmetic manipulation

4) Thus then equals

$$r_{F_1 z_1} = a_{11} r_{F_1 F_1} + a_{12} r_{F_2 F_1} + \ldots + a_{1m} r_{F_m F_1} + a_1 r_{U_1 F_1}$$

(from definition of r when variables are standard scores as r = average of the sum of the cross-products)

5) & this equals.

$$r_{F_1 z_1} = a_{11}(1) + 0 = a_{11} \qquad (2E0)$$

(correlation between $F_1 + F_1 = 1$ by definition
+ correlations between factors = 0 by definition
+ correlations between uniqueness + factors = 0 by definition)

If the factors are correlated and we know the correlations between the factors (which can be determined) then we can still use the equation in step four of the proof to determine the correlation between each variable and each factor. The correlations coefficients of the factors with the variables make up the structure of the factor analysis. Note that elements of the factor pattern and factor structure are equal when the factors are uncorrelated with each other.

In general, the factor pattern (the set of equations in Table 2-1) is simply a classic regression equation where $\hat{z}_i$ = the predicted score of the variable; $\beta_{ij} = a_{ij}$; and all variables are in standard score form.

$$\hat{z}_j = \beta_{j1} F_1 + \beta_{j2} F_2 + \cdots + \beta_{jm} F_m + \beta_j U_j$$

When the factors are uncorrelated, $\beta_{ij} = r_{ij}$

and when the factors are correlated, the betas represent the independent influence of each factor on the variables. This is directly analagous to basic theorems used in path analysis. If the factors are uncorrelated, the influence of a factor on a variable is just shown by its loading and is all direct influence. If the factors are correlated, some of this influence must be indirect, through other factors. This is directly analagous to interpretations of direct and indirect influences in path analysis.

The basic idea in computations in path analysis is to get the predicted scores of the variables, $\hat{z}_i$'s, as close as possible to the real scores, z_i's. We want to choose the betas or factor loadings so that the difference between these two values is minimized. h^2, the communality, measures how well the factors predict the variables. h^2 is then analagous to R^2 in multiple regression. The proportion of variance not explained is the uniqueness and equals $1 - h^2$ (and is analaogus to the square of the residual path in path analysis).

IN path analysis it is possible to reproduce correlations between any two variables by looking at associations in the path model (the regression coefficients). We can do the same thing in factor analysis and represent any correlation between two variables as a function of the factor loadings (the analguges to the regression coefficients). In other words, we can trace the association between two variables through their loadings on common factors. The proof of this theorem is given below and is similar to the proof of the similar result in path analysis. It is important to remember that by definition when using standard scores

$$r_{z_1 z_2} = \frac{\Sigma z_1 z_2}{N}$$

Proof of basic theorem of factor analysis
(depicting $r_{z_1 z_2}$ as function of factor loadings)

(1) $r_{\hat{z}_1 \hat{z}_2} = \frac{\Sigma \hat{z}_1 \hat{z}_2}{N} = \frac{1}{N} \Sigma (a_{11} F_1 + a_{12} F_2 + \ldots + a_{1m} F_m + a_1 U_1)(a_{21} F_1 + a_{22} F_2 + \ldots + a_{2m} F_m + a$

(by definition + substitution from factor pattern)

(2)
$$= a_{11} a_{21} \frac{\Sigma F_1 F_1}{N} + a_{11} a_{22} \frac{\Sigma F_1 F_2}{N} + \ldots + a_{11} a_{2m} \frac{\Sigma F_1 F_m}{N} + a_{11} a_2 \frac{\Sigma F_1 U_2}{N}$$

$$+ a_{12} a_{21} \frac{\Sigma F_2 F_1}{N} + a_{12} a_{22} \frac{\Sigma F_2 F_2}{N} + \ldots + a_{12} a_{2m} \frac{\Sigma F_2 F_m}{N}$$

$$+ a_{12} a_2 \frac{\Sigma F_2 F_a}{N} + \ldots + a_1 a_2 \frac{\Sigma U_1 U_2}{N} \quad \left(\begin{array}{l}\text{by multiplying} \\ \text{out step one}\end{array}\right)$$

By definition, $\frac{\Sigma F_i F_j}{N} = r_{F_i F_j}$ and also by definition U_i is uncorrelated with all other unique components + all the factors.

(2) Then reduces to

(3) $r_{\hat{z}_1 \hat{z}_2} = a_{11} a_{21} \frac{1}{\Sigma_1} + a_{11} a_{22} r_{F_1 F_2} + \ldots + a_{11} a_{2m} r_{F_1 F_m} + \ldots + a_{1m} a_{2,m-1} r_{m,m-1}$

$a_{1m} a_{2m}$

and if the factors are uncorrelated with each other

(4) $r_{\hat{z}_1 \hat{z}_2} = a_{11} a_{21} + a_{12} a_{22} + \ldots + a_{1m} a_{2m}$

Equation 4 above simply says that we may reproduce the correlation between any two variables through the factor loadings (and knowing the correlations between the factors when they are correlated). In the case where the factors are not correlated with each other the correlation simply equals the sum of the cross-products of the loadings of each variable on each factor. For instance, the correlation between variables one and two equals the sum of the product of the loadings of factor one on the two variables plus the sum of the product of the loadings of factor two on the two variables and so on.

In general, $r'_{jk} = \sum_i a_{ji} a_{ki}$, where i runs over all factors common to the two variables j and k.

Almost always $r'_{jk} \neq r_{jk}$. Some of this discrepancy will come from sampling error, but some will come because the communality, h^2, is less than one. That is, not all the variance of each variable will be held in common with the other variables. As more common factors are added the difference between the reproduced and actual correlation will be less. When the number of factors equals the number of variables than the match will be exact, but then what use is it to have a factor analysis because the factor pattern will be as complex as what you started with.

Once a factor pattern is found, we can test its adequacy by reproducing the correlations and comparing these reproduced correlations with the true correlations. We use $\overline{r_{jk}} = r_{jk} - r'_{jk}$

$\overline{r_{jk}}$, the difference, is called the residual correlation.

How large the residual correlation may be depends at least partly on what you want to do with the results. Obviously, you would like to have the difference be within sampling error of zero. But, you also don't want so many factors that it is so complex that it is as hard to analyse the factors as to analyze the variables. Generally, you can get the residual correlation to within sampling error of zero without having the factor pattern too complex. However, sometimes you may want to be satisfied with less fit so you will have more simplicity with the factors.

It is important to note that the communality, h^2, will be a unique result for a given analysis. However, the loadings of the factors on the variable is not unique. This may be affected by what is called rotation, a technique of trying to get the best and most understandable fit of the factors to the data. This is not really cheating or anything underhanded. The communality and the number of factors tell us how much variance is held in common by the variables involved and how many different factors are needed to represent this. This does not change. However, <u>how</u> these factors can best fit the data is a descriptive process and we may want to try several different fits to find one that makes the most sense.

Computations

Below we first discuss some basic concepts from matrix algebra and geometry and then discuss two techniques of computing factor analyses.

Basic concepts in matrix algebra and geometry -- A matrix is simply a square or rectangular arrangement of numbers in a table. A correlation matrix with the correlations of all variables with each other is a familiar example. If the number of rows and columns in a matrix is equal, then the matrix is a square matrix.

A matrix may be transposed. A transpose of a matrix is obtained by simply switching the rows and columns. Below the transpose of the matrix Z is Z' (this is the common notation.

$$Z = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \qquad Z' = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

A determinant of a matrix is a multiplicative function obtainable in a square matrix. Although it is a farily complicated procedure for a matrix larger than two by two, in the two by two case shown below the determinant is ad - bc, the difference of the two crossproducts.

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix}$$

The two lines along the sides, rather than the brackets used above designate that a determinant is to be taken.

The minor of a determinant is the determinant of a matrix when one or more columns and rows have been deleted. A first order minor is when one row and column have been deleted, a second order minor is when two rows and columns have been taken out, and so on.

A singular matrix is one whose determinant goes to zero. When a determinant of a matrix equals zero, it means that at least one row or column is a function of the others. For instance, in a correlation matrix, if one variable is totally a function of the other variables, then its determinant will be zero and it will be a singular matrix. (You may now be getting an idea of how this relates to factor analysis. In a correlation matrix if the variables are highly related to each other, some may be functions of the others and thus the determinant of the matrix will be zero. If, however, we can reduce the matrix to minors whose determinants will not go to zero, we will know how many factors underlie the matrix of correlations.) The rank of a matrix is the number of rows (or columns) in its largest non-vanishing determinant.

It is possible to add, subtract, and multiply matrices. (We will use this below with an example.) There are also identity matrixes (by which when multiplied a matrix stays the same) and inverse matrices which will change a matrix to its inverse by multiplication.

The fundamental theorem of factor analysis discussed above that showed how to reproduce the correlations between variables from the factor pattern may also be written in matrix form. This format shows the fundamental theorem for all the variables involved.

$R = A\Phi A'$

where A is the factor pattern
A' is the transpose of the factor pattern
Φ is the matrix of correlations between factors; if the factors are uncorrelated this is the identity matrix with ones on the diagonal
R is the correlation matrix with h^2 (communalities) in the diagonal

If we assume that the factors are uncorrelated we may write this as

$$\begin{bmatrix} h_1^2 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & h_2^2 & r_{23} & & \\ \vdots & & & & \\ r_{n1} & r_{n2} & r_{n3} & \cdots & h_n^2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2m} \\ & & \vdots & & \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ a_{13} & a_{23} & & a_{n3} \\ \vdots & & & \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{bmatrix}$$

(Φ, the identity matrix, may be omitted above because it won't affect the multiplication.)

Matrix multiplication is done element by element. For example

$$r_{21} = a_{21}a_{11} + a_{22}a_{12} + a_{23}a_{13} + \cdots + a_{2m}a_{1m}$$

Here it is clear that r_{21} equals the cross product of elements in the second row and first column of A and A', the row and column corresponding to the place of r_{21} in R.

Note that $h_1^2 = a_{11}^2 + a_{12}^2 + a_{13}^2 + \cdots + a_{1m}^2$

the sum of the aquare of variable one's loading on each factor. This is analagous to the computation of R^2 when variables in the prediction equation (analagous here to factors) are uncorrelated.

Note that the matrix format allows us to reproduce the basic theorem of factor analysis in a much more succinct form than previously. Also note that if we did not have h^2, the communality, in the diagonal, we could not get correct reproductions of the correlations.

Now to geometry. It is possible to represent correlations geometrically. We may see a correlation between two variables, if they are both represented by vectors (arrows) of the same length, as equal to the cosine of the angle between them. $r_{ij} = \cos \phi$, where ϕ is the angle between i and j.

Figure 2-1
The Cosine Function

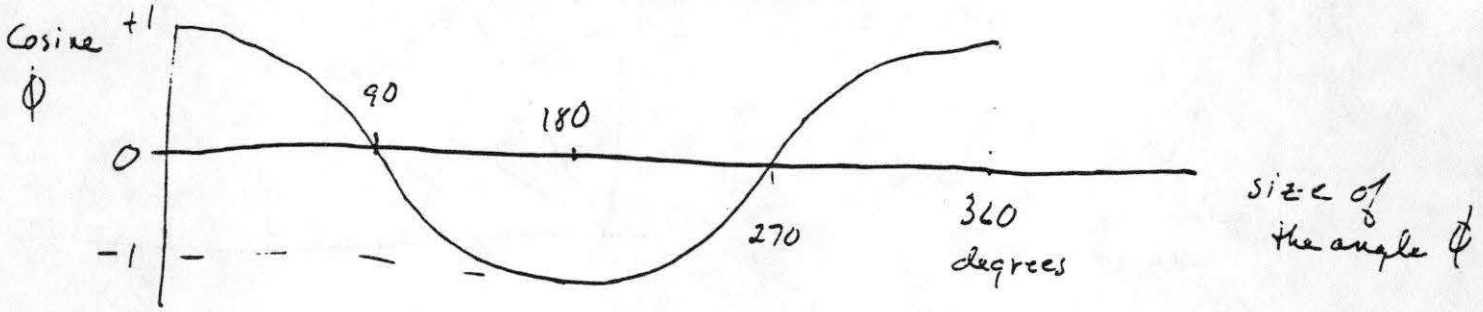


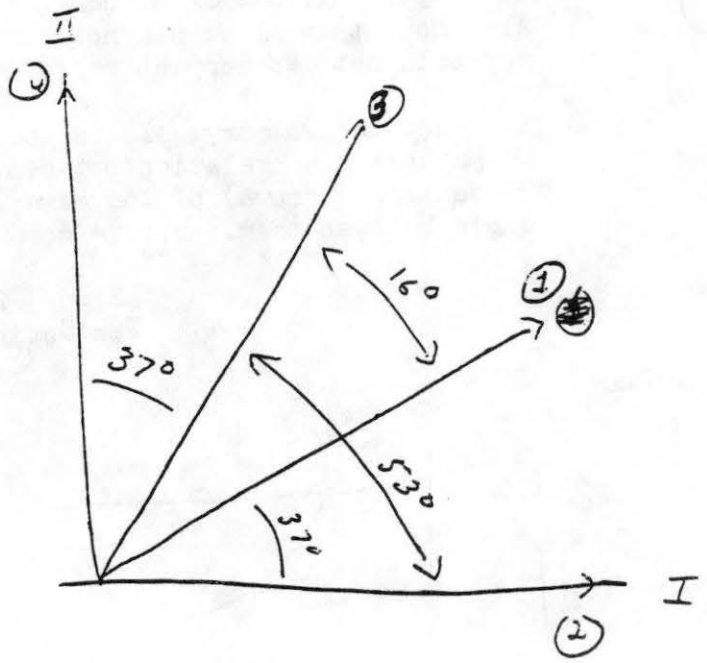
Figure 2-1 shows the relationship between the size of an angle and its cosine. You can see that when an angle is 90 degrees, its cosine is 0, when an angle is 0 degrees, its cosine is one. Thus, if two variables are perfectly correlated ($r = 1.00$) then they would fall on the same line, the angle between them would be zero. On the other hand, if two variables were not correlated ($r=0$) they would be at a 90 degree angle to each other. If two variables were perfectly negatively correlated with each other ($r = -1.00$) they would fall at opposite ends of the same line, with an angle of 180 degrees between them.

Figure 2-2 shows how we can represent loadings of four variables, all with communalities equal to one which are represented by two factors, uncorrelated with each other. The two factors, I and II, are at right angles or 90 degrees to each other because they are uncorrelated. Because variable 2 is uncorrelated with factor II and is perfectly correlated with Factor I it lies at a 90 degree angle with Factor II and on the line of Factor I. Variable 1 has a correlation of .8 with factor I. The angle whoe cosine is .8 is 37 degrees and so variable 1 is at a 37 degree angle from factor I. It is correlated .6 with factor II. The angle whose cosine is .6 is 53 degrees and so it is at a 53 degree angle from factor II. These results are simple because the communalities are exactly equal to one, but they serve to illustrate the geometric principles involved.

Variable	Factor I	Factor II	h^2
1	.8	.6	1.0
2	1.0	0	1.0
3	.6	.8	1.0
4	0	1.0	1.0

factor loadings (a_{ij}'s)

Figure 2-2



Computations -- There are a number of different methods of computing results for a factor analysis. Some methods are mainly of historical interest because they were developed for use before we had computers. One of the simplest hand computation methods will be described. Then I will review a much more accurate method that you can sanely do only with the computer. It is far too tedious to attempt to do by hand. A number of other methods are also available. Fruchter and Harman both describe these methods in sections of their books that were assigned.

The Diagonal method is an easy technique especially if you are not dealing with many variables. It is necessary to have fairly accurate estimates of the communalities for the technique to produce accurate results. You will need to have the correlation matrix as shown below with the communalities for each variable placed on the main diagonal. You will also need the factor matrix as shown. This method obtains orthogonal factors, or factors that are uncorrelated with each other. Note that the form of the factor matrix assumes that the first variable only has loadings on the first factor, the second variable has loadings on only the first and second factor; the third variable is loaded only on the first, second and third factor and so on. This is a basic definition to this method. You know that you cannot have more factors than you have variables, so this assumption is perfectly logical. Theoretically your analysis and computations could extend until you have computed loadings on as many factors as you have variables, but you likely wouldn't want to do this.

$R = A \phi A'$ ← assumes $= I$ (orthogonal factors)

$$\begin{bmatrix} h_1^2 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & h_2^2 & r_{23} & \cdots & r_{2n} \\ r_{31} & & & & \\ \vdots & & & & \\ r_{n1} & r_{n2} & \cdots & & h_n^2 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 & \cdots & & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ a_{31} & a_{32} & a_{33} & 0 & \cdots 0 \\ \vdots & & & & \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} & a_{31} & \cdots & a_{n1} \\ 0 & a_{22} & a_{32} & & a_{n2} \\ \cdot & 0 & a_{33} & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & & a_{nn} \end{bmatrix}$$

The matrices above show the fundamental theorem of factor analysis. From this theorem we know that for any individual correlation r_{jk}, (w/ orthogonal factors)

$r_{jk} = \sum_F a_{jF_i} a_{kF_i}$ and here

$r_{12} = a_{11} a_{21} + a_{12} a_{22} + \cdots + a_{1n} a_{2n}$

Here, however, $a_{12} = a_{13} = \ldots = a_{1n} = 0$

and $r_{12} = a_{11} a_{21} \longrightarrow a_{21} = \dfrac{r_{12}}{a_{11}}$ + $a_{11} = h_1^2 \rightarrow a_{21} = \dfrac{r_{21}}{h_1^2}$ (known)

Similarly,

$r_{31} = r_{13} = a_{11} a_{31} \longrightarrow a_{31} = \dfrac{r_{13}}{a_{11}}$

$a_{11} = h_1^2$, as ~~both~~ r_{12}, r_{13} + a_{11} are known + we may easily compute

a_{21} + a_{31}.

You may consider this pattern to get all the loadings on Factor I.

One of the byproducts of this process is the eigenvalue. Eigenvalues tell us the amount of the total variance that is explained by a factor. Because each variable in standard score form has a variance of one, the sum of the variances of all the variables = N, the number of variables. Eigenvalues can vary in size from zero to N. If they are zero, then the factor explains none of the variance of the total group. If an eigenvalue equals N, then it explains it all. If an eigenvalue of a factor is less than one than it explains less than one variable accounts for. The computer program will generally print out all the eigenvalues for all the factors (up to the number of variables involved). However, in the final calculations only those with eigenvalues greater than one are included. You can also specify other eigenvalues as cutoffs for what factors are included.

The net result of all this is a series of factors that best represent the common nature of a group of variables. Note that as an estimate of the communality a common practice is to use R^2 predicting one variable from all the others in a set. The principal factors method gives results that are similar to the diagonal method, but are more exact and also generally involve less factors.

Both of these methods of calculation adequately reflect the number of underlying factors. They may however yield loadings that do not give the best fit of variables to these factors. For instance, the picture below (Figure 2-3) illustrates the same set of variables that are best represented by two orthogonal factors, but in part b the factors are rotated approximately 30 degrees from their position in part a and the fit of the factors to the variables is much closer. This principal

Figure

of ~~matrix~~ moving the factors so they better fit the data is called rotation. While Figure ~~3-3~~ 3-3 involved only two factors, the same technique can work in larger space (with more factors).

There are two techniques of rotation. Orthogonal rotation means that the factors are uncorrelated with each other. In rotating, the factors are kept perpendicular to each other. In oblique rotation the factors may correlate. Note that rotation can yield an infinity of possible solutions or fit to the data. The main aim is to make the data set more understandable. The communalities and the number of factors remain constant, it is the loadings or fit of the factors to the variables that change.

Three orthogonal rotations are possible with SPSS. ~~The~~ Quartimax makes the complexity of a variable a minimum. With this type of rotation the factors are moved so that a variable loads on as few factors as possible. This kind of rotation would be used when you wanted simple interpretations about each variable. Varimax rotation simplifies the loadings for each factor. It makes the loadings on a given factor as close to zero or one as possible. This is best for easy interpretations of each factor. Equimax is the third technique ~~is~~ and is essentially a compromise between the other two.

Oblique rotations have a similar aim as the orthogonal rotations in the sense of making easier interpretations. However, they don't require that the factors be uncorrelated. You can specify how correlated the factors may be. This is somewhat more difficult to interpret than the ~~oblique~~ orthogonal rotations simply because you must also deal with relations between the factors.

On the following pages the pages from the SPSS manual describing the computer work are attached. (pp. 54-64.) There are two ways of entering data for a factor analysis, using the raw data and using a correlation matrix. Both procedures use the same procedure card. An example of this is shown below. First you must list the variables involved, then list the type of factor analysis program you desire, and then a number of ~~no~~optional additions in case you want to alter the diagonal elements in the matrix, specify the number of factors to be extracted, minimize the eigenvalue,

or specify the number of iterations. In all these cases there are very sensible default procedures used by the computer. The computer will authomatically use a principal axes method of extracting the factors that estimates the communalities. This is the most accurate procedure for using the principal axes method. The default rotation is varimax. There are several statistics that can be used.

With raw data input you simply use the FACTOR card. With matrix input, you should use the following set of cards.

```
RUN NAME
FILE NAME
VARIABLE LIST        use list of variables in the matrix
INPUT MEDIUM         CARD (if matrix is on cards)
N OF CASES           use the number of cases (estimate if necessary)
FACTOR               VARIABLES = .... (as before)
OPTIONS
STATISTICS
READ MATRIX
 enter matrix cards here
FINISH
/*
```

If you are going to type your own data for the matrix input it must be in the format of 8 F 10.7, that is with 8 correlations on a card and and seven values to the right of the decimal point. The decimal point need not be typed on the cards.

From the computer printout you will get information about the percentage of total explained variance that is accounted for by each factor. This is given with the information about the eigenvalues. The communalities tell what percentage of variance of each variable is held in common with the other variables. The percentage of the total variance that is explained by the factors may be computed by adding up the eigenvalues of the factors involved and dividing by N. Interpreting what the factors mean is the most involved and trickiest task. Here you must look at the nature of the loadings, looking at what variables are highly loaded and in what direction and trying to figure out what kind of theoretical meaning this can have. Consulting the results in the Expressiveness Reevaluated X article may help here. Note that for best results you should generally interpret the rotated factor matrix.

Finally, you may want to use factor scores. You remember from the earlier discussion that each variable can be represented as a function of the factors. But, if the factors make some kind of theoretical sense, you may want to do furtner analyses using the factors. You can build factor scores, in other words construct new variables representing each factor as a fuction of the variables, and use these in analyses. In general

$$F_1 = a_{11} z_1 + a_{21} z_2 + \cdots + a_{n1} z_n$$

$$F_2 = a_{12} z_1 + a_{22} z_2 + \cdots + a_{n2} z_n$$

$$\vdots$$

$$F_m = a_{1m} z_1 + a_{2m} z_2 + \cdots + a_{nm} z_n$$

where z_i is the standard scores of the variables.'
The machine can in fact output these factors scores and can even punch them onto cards or print them onto tape. These factor scores can then be used in further analyses.

XII. Discriminant Analysis

Discriminant analysis is used to represent the distinction between two or more groups as a linear function of 1 or more intervally measured variables. It is used for explanatory purposes, primarily in sociology and psychology, and for prediction, primarily in areas such as business and education. The groups under study may be seen as dependent variables or independent variables, depending upon the theoretical notion under investigation.

Suppose we were interested in the extent to which socioeconomic variables could discriminate between people who lived in the country or on farms at age 15 and those who lived in or near large cities. In other words, we have a nominally measured variable, place of residence at age 15, with two attributes. Suppose also that we were interested in how three intervally measured SES variables: education (X_1), occupational prestige (X_2), and income (X_3), were related to place of residence. Using discriminant analysis we can derive a function that best describes the difference between these two groups on these three variables.

$$D = d_1 Z_{X1} + d_2 Z_{X2} + d_3 Z_{X3} \qquad (12\text{-}1)$$

where D represents the discriminating function, d_i (i=1,2,3) represents the relative weight of each variable in discriminating between the two groups, and Z_{Xi} are the standardized scores (z-scores) of each variable.

Note how the function resembles the standardized multiple regression equation. In fact, in the two-group case the results of a multiple regression are proportional to the results of a discriminant analysis of the same data. The values of d_i may be compared to see which variables contribute the most to the discrimination, just as one compares the beta weights (the standardized regression coefficients) in a multiple regression equation. Essentially, higher discriminant scores mean that a variable is a more effective predictor of the difference between two groups.

Computer programs for discriminant analysis also commonly provide unstandardized discriminant coefficients. These are analogous to the unstandardized regression coefficients and are used with the raw data rather than the standardized scores.

One may use the discriminant function scores to predict the group placement of each member of the sample. That is, one can multiply the standard score of each member of the sample on education, occupational prestige, and income by the associated standardized discriminant coefficient, sum these values, and obtain that case's predicted value on the

discriminant function. Similarly, one could obtain a case's predicted value on the discriminant function by multiplying the unstandardized discriminant coefficients by the actual values of a case on the three independent variables and summing the results (plus a constant value analogous to the intercept in regression). Note again how this procedure is like what one would use in obtaining the predicted values for a given case with a regression equation with either standardized (betas) or unstandardized (b's) regression coefficients. The difference is that instead of the predicted values being a straight regression line or a plane that best fits the pattern of relationship between two intervally measured variables, the predicted values fall on a line that best represents the difference between two points in space (the two attributes of the nominally measured variable). The average overall score on this discriminant function is zero, and cases have discriminant scores that are either positive or negative.

Based on these scores on the discriminant function cases can then be sorted into two groups: those with positive scores and those with negative scores. Those with positive scores would be predicted to belong to one group, those with negative scores would be predicted to belong to another group. One can then compare the actual group classification of each case with its predicted group classification. If the chosen variables do a good job of discriminating between the groups, a large number of the cases should be predicted to fall within their actual groups. If the variables really aren't related to group membership at all then the predicted membership should be essentially unrelated to the actual membership. The computer output reports the percentage of cases that have been correctly classified. 50% would be expected to be correctly classified by chance and so one would hope that a considerably larger percentage would be correctly classified by the function if one's theory were to receive support.

After computing the value for the discriminant function for each case, one can compute the average value of these discriminant function scores for all the members in each of the two groups. These are referred to as the " group centroids." By comparing the average discriminant score for each group on each function the relative placement of each group can be compared. One can also see then how (or in what direction) each of the predictor variables explains the placement of cases on the discriminant function. A plot of discriminant scores is often useful to examine, for this gives the actual distribution of discriminant scores calculated for each case.

Note that it is also possible to apply the discriminant function to cases whose group membership is unknown. This

may be done to help classify them or to predict their group membership.

Suppose the following results had been obtained for a standardized discriminant function differentiating those from rural and urban origins.

D = 1.05 (education) + (-.34)(income)

centroids: rural origin: -.34
urban origin: +.30

To produce a discriminant score typical of people from rural backgrounds (negative), one would have lower education and lightly higher income than other subjects. To produce a score typical of those of urban origin (positive), one would have higher education and lightly lower income. Note that education contributes substantially more discriminatory influence than income.

Wilk's lambda is associated with each variable entered into a discriminant function and with each discrimination function that is computed. It is an inverse measure of how well the variables discriminate. If variables discriminate well, they will have a low lambda. Lambda is associated with the chi-square sampling distribution and this is used to test its significance.

There are apparently a number of ways to compute lambda. The one that is most intuitively appealing is to see lambda as unexplained variation/ total variation. That is, lambda represents the proportion of variation that is unexplained (the coefficient of alienation). With discriminant analysis the term variation refers to the variation or differences between the categories or groups being studied. It is analogous to, but not identical to, variation with intervally measured variables.

Canonical correlation is used when there are multiple independent and multiple dependent variables. A canonical correlation represents the association between these independent and dependent variables. With discriminant analysis a canonical correlation is associated with each discriminant function. This correlation represents the degree of association between variables in the function and the groups being discriminated. The square of the correlation may be seen as the proportion of variation in these dependent variables that is explained by the discriminating variables. (This is analogous to R and R^2 in multiple regression.) Note that $1 - R_c^2$ is approximately

equal to lambda. The chi-square associated with lambda then also tests the significance of the canonical correlation.

Discriminant analysis can easily be extended to differentiating between more than two groups. The procedures and results described above also apply in this case. The only difference is that the number of possible discriminating functions increases. The number of possible discriminating functions is always one less than the number of groups. This stems from the simple geometric fact that if one has two points (or group centroids) one represents their difference with a line (a one-dimensional space). If one has 3 points they can be represented by a plane (a 2-dimensional space).

XIII. Multivariate Analysis of Contingency Tables

Throughout this class and the previous class we have focussed on parametric statistics. These are statistics that have an intervally measured dependent variable. Yet, very often sociologists will have only nominally measured variables. Although it is possible to have nominally measured variables ~~as~~ independent ~~ones~~ within the general linear model, anything beyond a dichotomy cannot be used as a dependent variable. In this section we explore a technique that has become popular only rather recently (and was in fact developed only within the last 10 years or so) to analyze models with more than two variables measured on a nominal scale. This technique has been mainly developed by Leo Goodman of the University of Chicago and has been generally called hierarchical models or the log linear technique. (Note the term log linear -- that is a clue that in fact the form that this technique takes does have some ~~a~~ similarity to the general linear model that we have been studying.) Below I first briefly discuss the nature of contingency tables, ~~and~~ as they have been analyzed previously, ~~xxxxxxxxxx~~ I then discuss basic definitions and ways of seeing contingency tables and introduce the concept of cross-products ~~and~~ ratios, ~~and~~ the ~~idea of~~ use of logarithms in relation to these. Finally I show how these relate to analyzing contingency tables and work through an example.

Analysis of Contingency Tables

Sociologists have long used contingency tables -- both for analytic purposes (witness the common use of four-fold tables in theories) and in research. Our analysis techniques with these tables have, however, been relatively simple. Percentages are often used, chi-square statistics can be used to see if the distribution of cases differs from what would be expected by chance, and various measures of association (e.g. lambda, gamma, tau) are used to describe the nature of the association.

With multivariate questions the issue is more difficult. Lazarsfeld and his associates talked of elaboration of tables -- trying to specify relations. In the same way we will use below he suggested shoring the association between two variables within each category of a third (or combinations of the third & maybe a fourth). Lazarsfeld and his associates rarely used ~~xxxxxxx~~ tests of significance (or in some cases even measures of association) but preferred percentage analyses. James Davis in his work in the 1960's did some of the most sophisticated work building off of these techniques. (See also Hyman's work and Rosenberg's work)

In the 1960's Goodman and Kruskal introduced their measures of association for contingency tables; the lambda, tau, and gamma that are now widely used. They also introduced the multiple and partial measures of association associated with these. The multiple measure ~~xxx~~ is simply a measure with the dependent variable remaining the same and the categories of the independent variables all combined. The partial measures are usually weighted averages of the measures within each category of the control variable or combinations of categories of the control variables. All of these techniques had problems. They involved the often tedious task of searching for associations and the partial measures could not be used if the patterns of association were different in the various categories of the control measure. Also the Lazarsfeld techniques had no way of seeing if the results were due to chance, + sampling distributions associated with Goodman + Kruskal measures were not always available or used.

The new technique proposed by Goodman that we discuss here is superior to these in that it requires the researcher to posit ahead of time a model or theory of what kind of association s/he thinks will be in the data.

Essentially then the researcher uses this model as the expected frequencies, the actual data as the observed frequencies and constructs a chi-square statistic. This chi-square statistic can then tell the researcher if the differences between the observed frequencies and expected were due to chance. In contrast to the traditional use of chi-square where the researcher usually hoped to reject the null hypothesis that the difference was zero, in this case the researcher wants to <u>fail to reject</u> the null hypothesis, because s/he hopes that the data actually fit the proposed model.

Basic Concepts

The two articles assigned for this section are the simplest ones I have found that explain Goodman's model. Probably as it is more widely used other writings will become available. Both of these articles list the basic references in Goodman's own writing and the work of others. Those of you who feel relatively confident with your mathematical backgrounds may want to read these original works. Davis' article, published in Sociological Methodology 1973-74, deals only with the logic of hierarchical models. Reynolds' work also covers this and discusses the nature of the log linear model itself. The two articles are indeed complementary and do support each other.

The first basic concept used by Goodman is <u>odds ratio</u>. This is simply the ratio of the frequencies for two categories of some variable. It can be applied to the marginal frequencies, to the interior of one table or for comparisons across several tables. It can be used with dichotomies or with polytomies. Obviously, if the two categories are of equal size

the odds ratio will be ~~xxxxx~~ equal to one. ~~[illegible] one~~

In table 4-1 below, if we look at the marginals, we see that the odds ratio in variable ~~B~~ is 1.00. The odds ratio for variable ~~E~~ A is 40:60 = .67. In each case the number of cases in one category was divided by the number of cases in the other.

Table 4-1

B \ A	hi	lo	
hi	10	40	50
lo	30	20	50
	40	60	

AB: $\frac{40}{60} = .67$

B: $\frac{50}{50} = 1.0$

Davis uses the term conditional odds ratio to refer to the odds ratio for two categories within one category of another variable. For instance, in table one above, we could have the following conditional odds ratios:

$\left[\left(\begin{smallmatrix} hi \\ lo \end{smallmatrix}\right) / A = hi\right] = 10:30 = .33$ $\qquad$ $\left[\left(\begin{smallmatrix} hi \\ lo \end{smallmatrix}\right) / A = lo\right] = 40:20 = 2.0$

$\left[\left(\begin{smallmatrix} hi \\ lo \end{smallmatrix}\right) / B = hi\right] = 10:40 = .25$ $\qquad$ $\left[\left(\begin{smallmatrix} hi \\ lo \end{smallmatrix}\right) / B = lo\right] = 30\ 20 = 1.5$

Note that we could see a conditional odds ratio for the categories of A in each category of B and for the categories of B in each category of A.

Obviously, these conditional odds ratios are not equal. They are less than one for categories of B when A is hi and categories of A when B is high, and greater than one in the other categories. If we just compare the conditional odds ratios for B in categories of A we have the ratio .33/ 2.0 = .167 Or, if we compare the conditional odds ratios for A in categories of B we have .25/ 1.5 = .167. In other words, these ratios show that there are different relations between the categories in B depending on whether one is looking at the high or low category of A (and vice versa). This comparison of the conditional odds ratios is called the relative odds ratio by Davis. It is also simply the cross-product odds ratio, the familiar ratio of the product of the diagonal elements in a four-fold table. Below in Table 4-2 the cross products odds ratio equals ad/bc. For table one this equals (10)(20)/(40)(30)=.167 the same value we obtained above. By various manipulations of the elements Davis shows that the cross-products odds ratio and the ratio of the relative odds ratios are identical.

Just as with the other odds ratios, the cross products odds ratio equals one when there is no association, that is when the values in the cells are proportional to those in the marginals. The value of the cross-product odds ratio remains the same when the values in the cell are multiplied by a constant. If the values in the cells are rearranged (compare table 3 to table one) the cross-product ratio becomes the inverse of the early ratio. The values of the ratio may then vary from zero to infinity.

Table 4-2

a	b
c	d

cross product odds ratio = ad/bc

Table 4-3

30	10
20	40

$$\frac{(30 \times 40)}{(10 \times 20)} = \frac{1200}{200} = 6.0$$

($.167 = \frac{1}{6}$ from table 1; $6.0 = \frac{6}{1}$)

With tables larger than the two-by two ones discussed here we can still compute these various ratios, but usually do so by comparing each cell to one specific cell, often the one in the lower right hand corner of the table.

This is also the procuedure that is used when a third variable is added to the analysis. For instance, in Table 4 below we could examine the various ratios of each cell value to the criterion cell in the bottom right hand corner of the second table.

a	b	c
d	e	f

g	h	i
j	k	l

~~cross product~~ odds ratios: e.g. $\frac{a}{e}$; $\frac{b}{l}$; $\frac{c}{l}$; etc.

essentially are looking at the odds of being in one cell rather than another

Logarithms -- Because the cross-product ratios vary from zero to infinity, with a value of one when there is equality or no association, The cross-product ratio is not symetric. This means that if there were two tables with the columns reversed (see table 5 below) the cross product ratios would not be equidistant from one. Instead, however, they are reciprocal values of each other (see note above about changes in table one with cells moved, also).

Table 5

$\frac{(5)(8)}{(6)(7)} = \frac{40}{42} = \frac{20}{21} = .952$

5	6
7	8

6	5
8	7

$\frac{(6)(7)}{(5)(8)} = \frac{42}{40} = \frac{21}{20} = 1.050$

It is handy to have a measure of association that is symmetric (for instance a positive r has then same interpretation as a negative r of the same magnitude except for the direction). Because the cross-product ratio is not symetric, we want to transform it to a quantity that is.

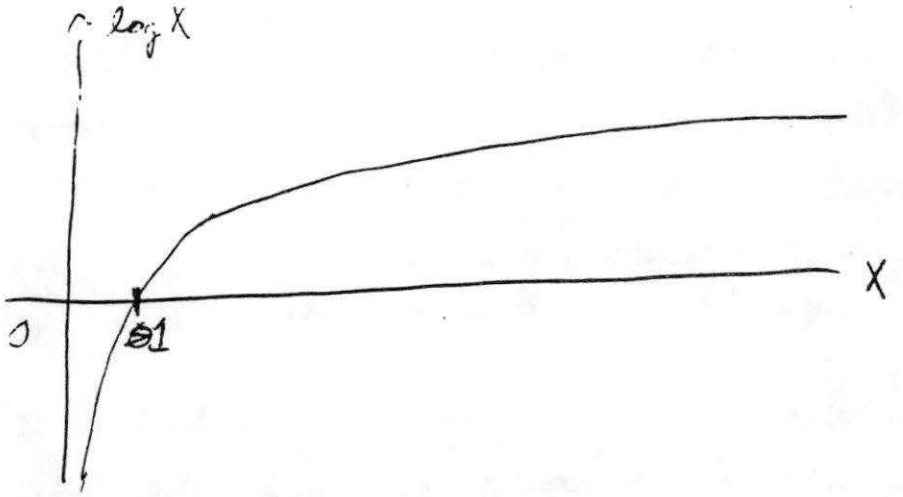
It turns out that the natural logarithm of the cross-product ratio is symetric around zero, varying from miums infimity to infinity. The natural logarithm is the exponent which e (a number aproximately equal to 2.718) must have to equal the cross-product. For instance, if the cross-product ratio equals 1.0, the log equals zero. This is because any number to the zeroth power is one. The graph below (Figure 4-1) shows the relationship between x (the value of the cross-product ratio) and the log of x.

natl. logarithm (ln)

$e^{(k)} =$ cross-product

eg. $e^{x} = 1.0$ $x = 0$ (nonelectric

Figure 4-1



The properties of exponents (and thus logarithms) are also extremely useful. These are theorems which can be proved (if you want to see the proofs, any calculus text should include them), but I will just state them here, both in terms of logarithms and in the exponent form.

$(10^4)(10^3) = 10^7 = 10^{4+3}$

$\log_b xy = \log_b x + \log_b y$ $\qquad e^x e^y = e^{x+y}$

$\log_b \frac{x}{y} = \log_b x - \log_b y$ $\qquad \frac{e^x}{e^y} = e^{x-y}$ $\left(\frac{10^4}{10^3} = 10^{4-3} = 10^1\right)$

$\frac{(10)(10)(10)(10)}{(10)(10)(10)}$

$\log_b x^r = r \log_b x$ $\qquad (e^x)^r = e^{xr}$

$\log_b x = \log_b y$ if and only if $x = y$ $\qquad e^x = e^y$ if and only if $x = y$

$\log_b x < \log_b y$ if and only if $x < y$ $\qquad e^x > e^y$ if and only if $x > y$

Given these properties, we can transform the cross-product odds ratios into their natural logarithms and they will be symetric around zero. (They will also make an additive function, a property which will be extremely useful in interpretations as we see below.)

For this transformation we take the

$$\log((ad)/bc) = \log a + \log d - \log b - \log c$$

from the table

a	b
c	d

and the properties of logarithms. Reynolds calls this value, $\log \alpha$, the log odds. (Davis does not discuss the logarithms, but proceeds to simply discuss the logic of effects and models.)

Effects and Models -- In other units we have discussed various causal models and used multiple regression and path analysis to analyze these models. With contingency tables we can also propose various theoretical models. These models are comprised of several effects, which are composed in a hierarchical manner from the simplest to the most complex. In contrast to the regression analyses where the dependent variable ranges over a wide range of values and we are interested in explaining this variation, in the log linear model our dependent variable may be seen as a cell probability, or the probability that a given member of the population falls into a given combination of the categories of each of the variables involved. Davis goes through the logic of this using cross-product ratios; in the discussion below I will use logarithms as Reynolds does.

The simplest model that is possible is one with no effects. This simply means that absolutely nothing is happening. In the example below we have three variables. One ~~xxxxxxxxx~~x (x) measures attitudes toward having a woman be an elementary principal. It has three categories of strongly approve (SA), approve (A), and disapprove (DA). Two other variables, region (Y) of residence (categories of metro (M), cities (C) and rural (R) and education level (Z) attained (high meaning some college and above (H) and lo (less than some college) are hypothesized as possible having some influence. The hypothetical array of the data in table 6 below shows how the data would be arranged if there were no effects.

Table 6

Z: Ed = hi Ed = lo

Y: Region

X: attitude	Metro	Cities	Rural	M	C	R	totals
SA	100	100	100	100	100	100	600
A	100	100	100	100	100	100	600
O	100	100	100	100	100	100	600
totals	300	300	300	300	300	300	1800

It is clear that all the categories have exactly the same number of cases and that each of the rows has the same number of cases and each column ~~xx~~ has the same number of cases. In other words, nothing seems to be happening.

107

We could express this no effectsmodel in symbols (following Reynolds) as

$F_{ijk} = \mu' = 100$ (where i,j, and k represent the categories in each of the three variables and F represents the cell frequency)

or we could transform this to a natural logarithm

$L_{ijk} = \mu = \log(100) = 4.61$

In the no effects model every cell has the same frequency or the same natural log.

Obviously the no effects model ~~probably~~ may not be likely to occur with this example (and probably not with any others). One possible [redacted]n effect is a row effect (or single effect). In other words, as shown in Table ~~6~~ 7 below, ~~[illegible]~~ the cells in different rows are ~~of~~ different sizes (have different frequencies). There are no column effects so that within each row all the columns have equal cell frequencies and the column marginals are all equal.

Table 7

	Z: Ed = Hi			Ed = Lo			
	Y: Region			Region			
X: attitudes	M	C	R	M	C	R	totals
SA	80	80	80	80	80	80	480
A	100	100	100	100	100	100	600
OA	120	120	120	120	120	120	720
total	300	300	300	300	30	300	1800

Here symbolically and mathematically we can represent our model as

$$F_{ijk} = \mu' \mu'^{X}_{i}$$

where μ'^{X}_{i} differs for each value of i

$\mu'^{X}_{1} = .8$; $\mu'^{X}_{2} = 1$; $\mu'^{X}_{3} = 1.2$

and in logarithmic form

$$L_{ijk} = \mu + \mu^{X}_{i}$$

e.g. $F_{1jk} = (100)(.8) = 80$

$F_{2jk} = (100)(1) = 100$

$F_{3jk} = (100)(1.2) = 120$

$\ln(80) = 4.38 = L_{1jk}$, $4.61 + \ln .8 = 4.61 - .22 = 4.39$

$\ln(100) = 4.61 = L_{2jk} = 4.61 + \ln(1) = 4.61 + 0$

$\ln(120) = 4.79 = L_{3jk} = 4.61 + \ln(1.2) = 4.61 + .18 = 4.79$

We can continue this process by adding column effects and ~~effect~~ single effects of the third variable. In Table 8 we show hypothetical cell frequencies that could appear if there were row effects, column effects, and effects of the third variable (which we can call the control variable for lack of a better term--may also call this the specifier variable).

Table 8

	Z: Ed = hi¹				Ed = lo²				
X: Att. / Y: Region	M¹	C²	R³	tot.	M¹	C²	R³	tot	grand tot.
SA¹	43	(53)	64	(160)	86	106	128	(320)	480
A²	53	67	80	(200)	(106)	134	160	(400)	600
O³	64	80	(96)	(240)	128	160	192	(480)	~~[illegible]~~ 720
tot.	160	200	240	(600)	320	400	480	(1200)	1800

101

We can express this model of single effects in symbols as

$$F_{ijk} = \mu' \mu'_i{}^{X} \mu'_j{}^{Y} \mu'_k{}^{Z}$$

where $\mu'_1{}^{X} = .8$; $\mu'_2{}^{X} = 1.0$; $\mu'_3{}^{X} = 1.2$

$\mu'_1{}^{Y} = .8$; $\mu'_2{}^{Y} = 1.0$; $\mu'_3{}^{Y} = 1.2$

$\mu'_1{}^{Z} = .6667$; $\mu'_2{}^{Z} = 1.333\ldots$

& $\mu' = 100$

& for example

$F_{121} = (100)(.8)(1.0)(.6667) = 53$

$F_{212} = (100)(1.0)(.8)(1.333) = 106$

$F_{331} = (100)(1.2)(1.2)(.6667) = 96$

We can also transform this to logarithms so that

$$L_{ijk} = \mu + \mu_i{}^{X} + \mu_j{}^{Y} + \mu_k{}^{Z}$$

for example $L_{121} \log_f(53) = \log_f(100) + \log_f(.8) + \log_f(1.0) + \log_f(.6667)$

$3.97 = 4.60 + (-.22) + 0 + (-.40)$

$= 4.60 - .66 = 3.94$ ≈

Notice that this model has consistent changes from one column to the next, from one row to the next, and from one level of education to the next. The magnitude of these changes are given by the size of the above terms. This model is called a single effects model because of the consistent changes.

It is also said that this single effects model does not have any interaction effects. That is, region does not seem to be related to attitudes, education does not seem to be related to attitudes, and education is not related to region. True enough, more people disapprove than approve; more people approve than strongly approve. But this patterns occurs in the same magnitude in each different region and in each level of education. Similarly, more people live in rural areas than

in cities and more people live in cities than in metro areas, but again this pattern appears consistently in all categories of attitudes and in each education level. It can be said then that there is no interaction in Table 8. Note how this pattern of consistent changes in the proportions of cell frequencies is similar to the case of no interactions in analysis of variance. Note also how the expression of the model in logarithmic form resembles the equations of effects used in the general linear model. This resemblance is why this model is called the log linear model.

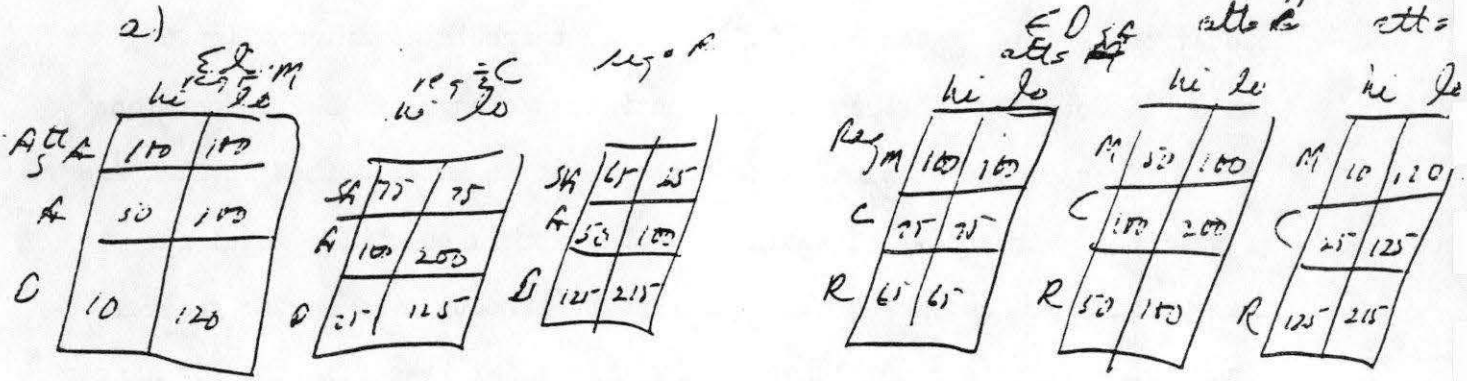
Now, of course it may occur that there really is some interaction between people's educational level and their attitudes and also their region of residence and their attitude, as well as some relationship between their educational level and region of residence. In table 9 below we show a hypothetical example of what could occur if there were an interaction (association) between region and attitudes and education and attitudes, but no association between education and region.

Table 9

		Region M	C	R	Σ	M	C	R	Σ	
Ed	SA	100	75	65	240	100	75	65	240	480
	A	50	100	50	200	100	200	100	400	600
	D	10	25	125	160	120	125	315	560	720
	Σ	160	200	240	600	320	400	480	1200	1800

Table 10

a)



From Table 9 and the reconstructions of the tables in 10 it is clear that rural people in each level of education are more likely to disapprove (an interaction between attitude and region, independent of education); that more highly educated people are more likely to approve in each region (an interaction between education and attitude, independent of region); but that there is no association between education and region (that is people in rural areas are not more or less likely to be highly educated) in each category of attitude.

These associations may be represented symbolically as before in the effect on frequencies and also by transforming these to logarithms.

$$F_{ijk} = \mu \mu_i^{X} \mu_j^{Y} \mu_k^{Z} \mu_{ij}^{XY} \mu_{ik}^{XZ}$$

$$L_{ijk} = \mu + \mu_i^{X} + \mu_j^{Y} + \mu_k^{Z} + \mu_{ij}^{XY} + \mu_{ik}^{XZ}$$

Note again the analogue of the logarithmic form to the form of the general linear model. μ may be seen as representing the grand mean, the size of the log of the category when there are no effects.

$\mu_i^X, \mu_j^Y, \mu_k^Z$ may each be seen as representing respectively the single effects of variables X (attitudes), Y (region), and Z (education) on the marginal categories. In other words they tell howmuch each category log of the marginal differs from the expected category frequency when there are no effects. The two terms $\mu_{ij}^{XY} + \mu_{ik}^{XZ}$ represent the special (interaction) effect that influences each combination of the categories of X and Y and X and Z. The fact that each of these terms has the subscripts i, j, and/or k indicates that for each category and combination of categories a different effect is possible. This is different from the general linear model where these parameters would usually have a constant effect across the model.

The model for table 9 assumed that there was no interaction of Z and Y independent of X and also that the patterns of interaction between X and Y held in categories of Z and the interaction of X and Z held in categories of Y. In fact, Table 9 was constructed so that situation did occur. It would be possible of course that Y and Z could be associated, and that in addition to the effects given in Table 9 we would also need to add to the model the interaction of Y and Z independent of X. (for instance, rural people could have less education and this result would occur in each category of attitudes.) The model for this situation could be written as

$$F_{ijk} = \mu' \mu_i^{X} \mu_j^{Y} \mu_k^{Z} \mu_{ij}^{XY} \mu_{ik}^{XZ} \mu_{jk}^{YZ}$$

take logarithms

$$L_{ijk} = \mu + \mu_i^X + \mu_j^Y + \mu_k^Z + \mu_{ij}^{XY} + \mu_{ik}^{XZ} + \mu_{jk}^{YZ}$$

Finally, it may be that attitudes are associated with people's education, but that this varies in region (maybe with rural people being even more likely to disapprove when they have low educations than are urban people with low educations). Or maybe the association between region and attitudes varies among categories of education, being much stronger among the lower educated group. This would be termed a three-way interaction and could be represented symbolically as

$$F_{ijk} = \mu' \mu'^{X}_{i} \mu'^{Y}_{j} \mu'^{Z}_{k} \mu'^{XY}_{ij} \mu'^{XZ}_{ik} \mu'^{YZ}_{jk} \mu'^{XYZ}_{ijk}$$

+ in logarithms

$$L_{ijk} = \mu + \mu^{X}_{i} + \mu^{Y}_{j} + \mu^{Z}_{k} + \mu^{XY}_{ij} + \mu^{XZ}_{ik} + \mu^{YZ}_{jk} + \mu^{XYZ}_{ijk}$$

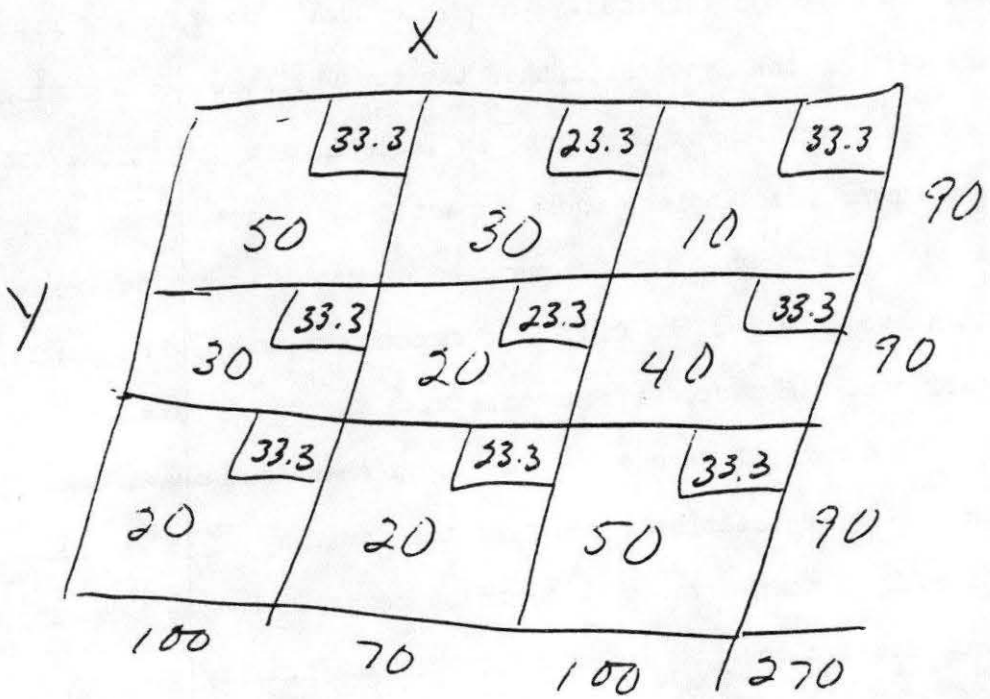
The various parameters have the same interpretation as before. The one addition is the three way interaction terms μ'^{XYZ}_{ijk} or μ^{XYZ}_{ijk}. These represent the effect noted above that the association between any two variables may differ depending on the category of the third variable.

In this model every expected cell frequency or its associated logarithm has a unique combination of terms in the model equation. The model is then called saturated. It is analogous to the complete equation with interaction when multi-way analysis of variance is expressed in regression terms.

Testing the Models

Once you have decided theoretically what kind of effects model your problem should show, you will want to compare your actual data with this model that you have hypothesized. This is actually done in the usual chi-square test of the hypothesis that the two variables in a contingency table are independent. For instance, in Table 11 a hypothetical table with the observed frequencies and those expected by chance (when there is no association) is shown. The frequencies expected by chance are those that we would expect to find in the model with single variable effects. (Expected frequencies with the no effects model would imply that the marginals were all equal too.) To get the chi-square statistic one uses the formula, $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$ This is actually a sample value that can be compared to the sampling distribution of chi-square that would occur when the null hypothesis of no association (that the model exists) is true. Just as with F-distributions and the others that we have studied, if the chi-square statistic falls into a zone of rejection we may cast doubt on this null hypothesis. Just as with the t and F distribution, there are a number of chi-square distributions, each one varying by degrees of freedom. For instance, with table 11 there are 4 degrees of freedom, 4 chances to freely choose cell frequencies based on the size of the margianls. In general, the degrees of freedom represents the number of unspecified parameters, the number of free choices, the numberof cell frequencies that are not specired by the model.

Table 11



$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(16.7)^2}{33.3} + \frac{(-3.3)^2}{33.3} + \frac{(-13.3)^2}{33.3} + \frac{(6.7)^2}{23.3}$$

$$+ \frac{(-3.3)^2}{23.3} + \frac{(-3.3)^2}{23.3} + \frac{(-23.3)^2}{33.3} + \frac{(6.7)^2}{33.3} + \frac{(16.7)^2}{33.3}$$

$$= 42.902$$

$$\text{expected freq. } f_{ij} = \frac{f_{\cdot j}\, f_{i\cdot}}{f_{\cdot\cdot}}$$

$P[\chi^2 = 42.9, df = 4] < .005$

$df = (3-1)(3-1) = 4$

In Table 11 and the calculations of the chi-square statistic shown we have essentially posited that there is no association or interaction between the two variables. Although we do posit that there are single effects or that the marginals may differ and base our expected frequencies on this, we posit no interactions. The expected frequencies for each cell equal the product of the two associated marginals divided by the grand total of cases. The model implicitly posited in the usual chi-square test is $F_{ij} = \mu' \mu_i^{X} \mu_j^{Y}$

or in logarithms $L_{ij} = \mu + \mu_i^{X} + \mu_j^{Y}$

To compute the chi-square statistic we compare the expected cell frequencies under this model with those we actually have or observe in the data. With the data in table 11 we obtained a chi-square value of 43. Comparing this to the chi-square values in the sampling distribution assuming the null hypothesis (that the model holds) is true for 4 degrees of freedom we found that these results would rarely occur by chance alone. In other words the model of only single variable effects cannot hold and the model with interaction effects is the one that must fit the data. With two variables this is the saturated model.

$$F_{ij} = \mu' \mu_i^{X} \mu_j^{Y} \mu_{ij}^{XY}$$

$$\text{or } L_{ij} = \mu + \mu_i^{X} + \mu_j^{Y} + \mu_{ij}^{XY}$$

This kind of logic is the type used whenever one tests the various models that are possible. Obviously the two variable case is a simple one and we can readily see the nature of the association. With more variables the situation becomes more complex, both in ascertaining the nature of

the association simply by inspection and also in estimating the frequencies that would be expected if the various models held. The results of testing the various models will guide the inspection of the data. Getting the expected frequencies for each possible model can be done through various computations that can get tedious when done by hand. Davis describes one method of doing this by hand. You may also consult Goodman's work. The easiest way, however, is to use a computer program. These programs (and the hand methods) essentially go through several iterations or tries at getting expected cell frequencies that fit the model and sum to the marginals expected by the single effects model. Each iteration approaches the sum of the marginals more closely. Because the programs that do the iterations are becoming more and more available I will not describe how to do the iterations.

Once one has the frequencies expected for a data set when a given model holds, one may compare these expected frequencies with those actually observed. One may then compute the usual chi-square statistic and compare that value to the sampling distribution. If one is using log ratios the chi-square statistic must be modified slightly to the likelihood ratio chi-square. Some techniquesalso use a statistic called the information statistic (Ku and Kullback). All of these statistics have sampling distributions equal to or very similar to the chi-square distribution.

An example

We can work through an example using actual survey data from a sample of residents of Oregon who were representatively selected in January, 1977.

Each of the respondents was asked how they would feel about a woman being hired as an elementary school principal in their local schools. They could respond on a ten point scale ranging from strongly disapprove to strongly approve. Because the responses tended to be skewed toward the strongly approve end of the scale the responses were trichotomized into three categories: strongly approve, approve, and disapprove. Because the residents of the state live in fairly distinct regions we also noted the area of residence and divided these into three categories: metropolitan Portland area; the Willamette Valley area; and the rural areas of the state including the Coastal region, Southern Oregon, and Eastern Oregon. These divisions were made to separate out the metropolitan residents who generally have different interests and concerns than other residents; the Willamette Valley people who tend to be more liberal; and the rural residents who are generally seen as more conservative. Finally, the measure of the respondents' education was used, divided at the college/non-college level, with everyone with some college education included in the first category. Based on previous studies of attitudes toward women school administrators we expected that more conservative attitudes would be displayed in rural areas and by those with less education.

The program CONTAB was used to analyze the data. The cards used to access the program (they may not be applicable after spring, 1978) and to define the runs wanted are shown below.

```
//jobname JOB (#    ), name
//JOBLIB ␣DD␣ UNIT=2314,VOL=SER=WORK22,DSN=USER
.OA.WJ282586,DISP=OLD
```

(all one card)

```
//␣␣␣␣ EXEC ␣␣␣ PGM=CONTAB
//PRINT ␣␣ DD ␣␣ SYSOUT=A,DCB=(RECFM=FBA,
LRECL=133,BLKSIZE=1330)
```

(one card)

```
//SYSPRINT ␣␣ DD ␣␣ SYSOUT=A,DCB=(RECFM=FBA,
LRECL=133,BLKSIZE=1330)
```

(one card)

```
//SYSIN ␣␣ DD ␣␣ *

FACTORS = 3    FL(1) = 'ATTELM'  FL(2) = 'EO'
FL(3) = 'REGION'

TITLE = 'ATT ELEMENTARY PRIN'  LIST = 'ORMEO';

3  2  3

3  4  7  8  17  13  13  49  30  23

57  42  87  115  49  86  147  67

TERMS = 3;  1  1  1  1  2  3

TERMS = 3;  2  2  2  1  2  13  23

TERMS = 2;  2  2  13  2  3

TERMS = 2;  2  2  1  2  2  3

TERMS = 1;  3  1  2  3

/*
```


12

Table 12

	Education hi				lo			
Region	Wm Val.	Metro	Rural	C	Wm Valley	M	R	C
D	3	4	7	14	8	17	13	38
A	13	49	30	92	23	57	42	122
SA	87	115	49	251	86	147	67	300
C	103	168	86	357	117	221	122	460

ΣΣΣ = 817

The data for this problem are shown above. The first six cards on the previous page are job control langauge cards. The first is the familiar job card. The second accesses the program from disc. The third and fourth instruct the computer to output the results on the printer. The fifth jcl card tells the computer to go ahead. The remaining cards instruct the contab program as to what kind of data is being used and what hypotheses to test. The factors equal 3 tells it that 3 variables are included. The FL terms give labels to these three variables. The title term gives a name to the run. The list paremeter asks for various kinds of output. These are described in more detail in the description of the program attached to these notes. After the list the first three numbers (3 2 3)

tell the computer that the first variable (attitudes) has three categories; the second (education) has two categories; and the third variable has three categories. After this the data itself (earlier gotten by using an SPSS crosstabs run) is given. The cell frequencies are typed in the order of cell frequencies 111, 112, 113, 121, 122, 123, and so on with the final digit varying most quickly and the first digit varying least quickly. The final cards present the various hypothese or models that are to be tested.

The first terms statement hypothesizes that all the terms are independent. In other words, it suggests that only single effects and no interactions are present. $L_{ijk} = \mu + \mu_i^X + \mu_j^Y + \mu_k^Z$

The second term hypothesizes that there are second order interactions but no third order interactions. It hypothesizes interactions between each pair of variables, but that these interactions are the same in each category of the variable not included. In symbols this may be seen as $L_{ijk} = \mu + \mu_i^X + \mu_j^Y + \mu_k^Z + \mu_{ij}^{XY} + \mu_{ik}^{XZ} + \mu_{jk}^{YZ}$

The third term hypothesizes that there is only an interaction between the first term and the third term and between the second term and the third term, and that this interaction is the same in each category of the variable not included in the interaction. $L_{ijk} = \mu + \mu_i^X + \mu_j^Y + \mu_k^Z + \mu_{ik}^{XZ} + \mu_{jk}^{YZ}$

The fourth term hypothesizes that education is associated with attitudes and that education is associated with (interacts with) region and this pattern is similar in all categories of the variable not included. in each term.

$L_{ijk} = \mu + \mu_i^X + \mu_j^Y + \mu_k^Z + \mu_{ij}^{XY} + \mu_{jk}^{YZ}$

Finally, the fifth term is the saturated model, hypothesizing all possible interactions both of the second and third degree.

$L_{ijk} = \mu + \mu_i^X + \mu_j^Y + \mu_k^Z + \mu_{ij}^{XY} + \mu_{ik}^{XZ} + \mu_{jk}^{YZ} + \mu_{ijk}^{XYZ}$

124

In a summary table the contab program prints a summary of the results of testing each of these models against the actual data. This information is summarized below.

hypothesized model #		Is	degrees of f.	prob. of being wrong in rejecting the model	decision
(X)(Y)(Z)	1.	39.262	12	0.0001	(reject)
(XY)(XZ)(YZ)	2.	4.007	4	0.4051	fail to reject
(XZ)(YZ)	3.	10.608	6	0.1013	marginal reject
(XY)(YZ)	4.	30.815	8	0.0002	reject
(XYZ)	5.	0.000	0	-1.000	

From this summary we can conclude that the model that best fits the data (aside from the saturated model which we would like to avoid if a simpler one will suffice) is the second one ~~xx~~ presented, that of interaction effects of all variables on the second order, but no third order effects.

Now to analyze the specific nature of the effects of this model we turn to the table of data and to the output given regarding the test of this model. This output is shown on the next several pages. The first part of the output, labeled marginals essentially gives the tables of frequencies in each cross-tabulation of variables shown. You can compare these tables to Table 12 to see that they match. Percentages have been added to these tables. It is clear that lower education is associated with disapproving as is residence in rural areas. Also, there is a tendency for rural residents to have less education than metro residents and metro residents to have less education than willamette valley residents. We knwo from the results of testing models 3 and 4 that these results are all independent of each other.

88

HYPOTHESIS 2

MARGINALS: ATTELM * ED.

	1 (%)		2 (%)	
1 O	14.0000	3.9	38.0000	8.3
2 A	92.0000	25.8	122.0000	26.5
3 SA	251.0000	70.3	300.0000	65.2
	357	100	460	100

MARGINALS: ATTELM * REGION.

	1		2		3	
	Wm Val	(%)	Metro	(%)	Rural	(%)
1 O	11.0000	5	21.0000	5.4	20.0000	9.6
2 A	36.0000	16.4	106.0000	27.2	72.0000	34.6
3 SA	173.0000	78.6	262.0000	67.4	116.0000	55.8
	220	100	389	100.0	208	100

MARGINALS: ED * REGION.

	1 Wm. Valley		2 Metro		3 Rural	
		%		%		%
1 hi	103.0000	46.8	168.0000	43.2	86.0000	41.3
2 lo	117.0000	53.2	221.0000	56.8	122.0000	58.7
	220		389		208	

RESIDUALS: ATTELM * ED * REGION. FIRST 1 LEVELS: 1

			1	2	3
1	OBSERVED	1	3.000000	4.000000	7.000000
1	PREDICTED	2	3.231145	5.592946	5.176024
1	RESIDUAL	3	-0.231145	-1.592946	1.823976
1	OUTLIER	4	0.015706	0.506466	0.582304
1	LOG RATIO	5	-3.005071	-2.457002	-2.534470
2	OBSERVED	6	8.000000	17.000000	13.000000
2	PREDICTED	7	7.768652	15.407074	14.824302
2	RESIDUAL	8	0.231348	1.592926	-1.824302
2	OUTLIER	9	0.007113	0.162703	0.237453
2	LOG RATIO	10	-2.128410	-1.443680	-1.48223

RESIDUALS: ATTELM * ED * REGION. FIRST 1 LEVELS: 2

			1	2	3
1	OBSERVED	1	13.000000	49.000000	30.000000
1	PREDICTED	2	16.596924	45.309418	30.092560
1	RESIDUAL	3	-3.596924	3.690582	-0.092560
1	OUTLIER	4	0.858728	0.310617	-0.000212
1	LOG RATIO	5	-1.369289	-0.364992	-0.774229
2	OBSERVED	6	23.000000	57.000000	42.000000
2	PREDICTED	7	19.403030	60.690552	41.907333
2	RESIDUAL	8	3.596970	-3.690552	0.092667
2	OUTLIER	9	0.645275	0.246046	0.000424
2	LOG RATIO	10	-1.213078	-0.072719	-0.443046

RESIDUALS: ATTELM * ED * REGION. FIRST 1 LEVELS: 3

			1	2	3
1	OBSERVED	1	87.000000	115.000000	49.000000
1	PREDICTED	2	83.171875	117.097610	50.731400
1	RESIDUAL	3	3.828125	-2.097610	-1.731400
1	OUTLIER	4	0.193468	0.043287	0.063274
1	LOG RATIO	5	0.242402	0.584500	-0.251962
2	OBSERVED	6	86.000000	147.000000	67.000000
2	PREDICTED	7	89.828308	144.902237	65.268311
2	RESIDUAL	8	-3.828308	2.097763	1.731689
2	OUTLIER	9	0.185097	0.036875	0.049364
2	LOG RATIO	10	0.319392	0.797552	-0.000001

NONZERO EFFECTS: GENERAL MEAN = -3.393629

NONZERO EFFECTS: ATTELM.

			1	2	3
1	EFFECT	1	-1.308747	0.160270	1.148477
1	STAN. DEV.	2	0.111822	0.075480	0.066771
1	STANDARDIZE	3	-11.703875	2.123339	17.200195

NONZERO EFFECTS: ED.

			1	2
1	EFFECT	1	-0.236916	0.236916
1	STAN. DEV.	2	0.061454	0.061454
1	STANDARDIZE	3	-3.855161	3.855161

NONZERO EFFECTS: REGION.

			1	2	3
1	EFFECT	1	-0.325946	0.373773	-0.047827
1	STAN. DEV.	2	0.095308	0.080827	0.083923
1	STANDARDIZE	3	-3.419924	4.624360	-0.569893

NONZERO EFFECTS: ATTELM * ED.

			1	2
1	EFFECT	1	-0.253553	0.253553
1	STAN. DEV.	2	0.111822	0.111822
1	STANDARDIZE	3	-2.267473	2.267473
2	EFFECT	4	0.106971	-0.106971
2	STAN. DEV.	5	0.075480	0.075480
2	STANDARDIZE	6	1.417200	-1.417200
3	EFFECT	7	0.146581	-0.146581
3	STAN. DEV.	8	0.066771	0.066771
3	STANDARDIZE	9	2.195277	-2.195277

NONZERO EFFECTS: ATTELM * REGION.

			1	2	3
1	EFFECT	1	-0.065848	-0.148869	0.214718
1	STAN. DEV.	2	0.173158	0.148425	0.151694
1	STANDARDIZE	3	-0.380279	-1.002990	1.415468

2 EFFECT	4	-0.259012	0.113596	0.145416
2 STAN. DEV.	5	0.118784	0.097593	0.102712
2 STANDARDIZE	6	-2.180533	1.163979	1.415788
3 EFFECT	7	0.324861	0.035273	-0.360134
3 STAN. DEV.	8	0.102022	0.087428	0.093263
3 STANDARDIZE	9	3.184210	0.403451	-3.861484

NONZERO EFFECTS: ED * REGION.

		1	2	3
1 EFFECT	1	0.051839	-0.016192	-0.035647
1 STAN. DEV.	2	0.095308	0.080827	0.083923
1 STANDARDIZE	3	0.543912	-0.200327	-0.424755
2 EFFECT	4	-0.051839	0.016192	0.035647
2 STAN. DEV.	5	0.095308	0.080827	0.083923
2 STANDARDIZE	6	-0.543912	0.200327	0.424755

INFORMATION STATISTIC	4.006716
PROBABILITY OF A GREATER VALUE	0.405098
LN(REFERENCE/(N/NUMBER OF CELLS))	0.389425
LN(REFERENCE*/(N/NUMBER OF CELLS))	0.363240
SMOOTH	0.000000
ZERO	0.000001
SAMPLE SIZE	817.000000
MAXERROR	0.001000
OUTLIER BOUND	7.000000
EFFECTS - INTERACTION LEVEL PRINTED	2
MARGINALS	
RESIDUALS	
DEGREES OF FREEDOM	4
FACTORS	3
NUMBER OF CELLS	18
ZERO CELLS	0
MARGINAL ZERO TOTALS	0
TERMS	3
NUMBER OF OUTLIERS	0

Following the data for the two by two tables the data for each cell inputted is given. This data is given for the association between attitudes and education in each categ ry of region. The first set gives the results for the three levels of attitude and the two categories of education for the willamette valley region. The observed value is that which has been given as input data. The predicted value is the value which was predicted to occur (estimated) if the model hypothesized were tyre. The residual term is the difference between the observed and predicted value. The outlier term is an indicator of how well the cell fits the model. It is a function of the residual value and the sample size. A small outlier value indicates that the cell closely fits the model. A larger value indicates that it "outlies" the model more. The log ratio is the logarithm (base e) of the ratio of the predicted value for that cell to the predicted value for the reference cell (the last cell in the table). These ratios will vary around zero and will equal zero when the predicted cell values are equal and there is no effect. Each of these values is given for each cell.

The next section of the printout tells about the effects in the model and can be used to reproduce the model. These are given in logarithm form and can be used to fill in the model equation

$$L_{ijk} = \mu + \mu_i^{X} + \mu_j^{Y} + \mu_k^{Z} + \mu_{ij}^{XY} + \mu_{ik}^{XZ} + \mu_{jk}^{YZ}$$

The effect term gives the actual logarithm for each particular effect. The standard deviation is the standard deviation for this effect, which

can be used in testing hypothese about effects or putting confidence bands around them (see Reynolds' discussion). The term standardize refers to the ratio of the effect to the standard deviation. This row can be helpful if interactions are important by helping you find where they are. Based on the summary of effects we can provide the following parameters for the model.

$L_{ijk} =$ $L_{ijk} =$

$\mu \quad -3.394$

$+ \mu_i^X \quad \mu_1^X = -1.309 \quad \mu_2^X = .160 \quad \mu_3^X = 1.148$

$+ \mu_j^Y \quad \mu_1^Y = -.237 \quad \mu_2^Y = .237$

$+ \mu_k^Z \quad \mu_1^Z = -.326 \quad \mu_2^Z = .374 \quad \mu_3^Z = -.048$

$+ \mu_{ij}^{XY} \quad \mu_{11}^{XY} = -.25 \quad \mu_{12}^{XY} = .25 \quad \mu_{21}^{XY} = .11 \quad \mu_{22}^{XY} = -.11$

$+ \mu_{ik}^{XZ} \quad \mu_{31}^{XY} = .15 \quad \mu_{32}^{XY} = -.15 \quad \mu_{11}^{XZ} = -.06$

$+ \mu^{YZ} \quad \mu_{12}^{XZ} = -.15 \quad \mu_{13}^{XZ} = .21 \quad \mu_{21}^{XZ} = -.26 \quad \mu_{22}^{XZ} = .11$

$\mu_{23}^{XZ} = .14 \quad \mu_{31}^{XZ} = .32 \quad \mu_{32}^{XZ} = .04 \quad \mu_{33}^{XZ} = -.36$

$\mu_{11}^{YZ} = .05 \quad \mu_{12}^{YZ} = -.02 \quad \mu_{13}^{YZ} = -.04$

$\mu_{21}^{YZ} = -.05 \quad \mu_{22}^{YZ} = .02 \quad \mu_{23}^{YZ} = .04$

Finally, the information testing the hypothesis is summarized. The information statistic and the probability of being wrong in rejecting the model being tested is given. The two log reference numbers are ratios of the reference cell (last one in the table) to first the average observed cell count and second to the average predicted cell count. You may want to compare this to the various log ratios obtained for each cell earlier. The smooth and zero values are those you can input to modify your data if necessary. The outlier bound used was a default value and no cells exceeded this bound.

In completing a substantive interpretation of the model one would examine the parameters for the model in conjunction with the contingency tables and make interpretations. In this case the important information would be the independent effects of both region and education on attitudes and the tendency for rural residents and less well educated residents to disapprove (and the maintenance of this effect despite the tendency for rural residents to have less education).

Sociology 413, Spring, 1988
The General Linear Model

Most of the statistical procedures you will use in your professional life as a social scientist are based on the general linear model. All of these involve the notion of explaining variation (or an analogue to this) in a dependent variable(s) through its association with one or more independent variables. The outline below is meant to be only a sketch of the various techniques that may be used. It is not meant to be exhaustive. Remember that each of these techniques incorporate both <u>inferential</u> and <u>descriptive</u> statistics and that each is appropriate only for certain kinds of variables and research questions. The key to success in using statistics is understanding not just the mechanics of applying the techniques but in knowing when each type of statistics should or should not be used and in accurately interpreting the results. With the basic knowledge of the general linear model gained in this class you should be able to pursue advanced course work or reading in these areas with relatively little difficulty.

1. Simple bivariate regression

$$Y = a_{Y\cdot X} + b_{Y\cdot X} X$$

descriptive statistics:
- a - the y-intercept
- b_{YX} - the slope, predicting y from x
- r - the Pearson product moment correlation
- r^2 - the square of r (the proportion of variation in Y that is explained by its linear association with X)

inferential statistics:
- test of $H_o: r_{YX} = 0$ (equivalent to test of $H_o: r^2_{YX} = 0$ and $H_o: b = 0$)
- confidence intervals can also be placed both r (using the z transformation) and b (using the standard error of b)

2. Multiple regression

$$Y = a_{Yx} + b_{Yx_1}X_1 + b_{Yx_2}X_2 + \ldots + b_{Yx_k}X_k$$

$$Y = \beta_{Yx_1}X_1 + \beta_{Yx_2}X_2 + \ldots + \beta_{Yx_k}X_k$$

descriptive statistics: a - the y-intercept
b - unstandardized regression coefficients
β - standardized regression coefficients
$R^2_{Y \cdot 12 \ldots k}$ - multiple correlation coefficient squared (the proportion of variation in Y that is explained by its linear association with $X_1 \ldots X_k$)
partial correlation coefficients
part correlation coefficients

inferential statistics: test of $H_0 : R^2 = 0$
test of $H_0 : b_{Yx_1} = 0$ (equivalent to testing $H_0 : \beta_{Yx} = 0$) (can also test $H_0 : a = 0$)
Confidence intervals can also be placed around $R_{Y \cdot 1 \ldots k}$ and around the y-intercept and each unstandardized regression coefficient using the standard errors.)

3. Analysis of Variance

$$Y = a + b X_1 + b X_2 + \ldots + b X_K$$

where the X_i are dummy variables (or variations of dummy variables which represent the categories in the nominally measured variables and any possible interaction terms)

descriptive statistics: R^2 = the proportion of variation in Y that can be explained by the categories of the nominally measured variable(s). It is equivalent to E^2 used in analyses of variance.

a and b : interpretation of these values depends upon the way in which the dummy variables were coded, but in general they can be used to describe differences in average values of the dependent variable in each category of the independent variable

inferential statistics: test of $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$
(null hypothesis that the average [mean] of the dependent variable is equal in each category of the independent variable) When more than one independent variable is involved, a test would be conducted for each independent variable and for each possible interaction term.

4. Analysis of Covariance

$$Y = a + b X_1 + b X_2 + \ldots + b X_K$$

where some of the X_i's are dummy variables (or variations of dummy variables which represent the categories in the nominally measured variables), some of the X_i are intervally measured variables, and the rest represent interactions between these.

descriptive statistics: R^2 = the proportion of variation in Y that can be explained by the categories of the nominally measured variables, the intervally measured variables, and any interaction between them.

a and b = interpretation of these values depends upon the way in which the dummy variables were coded, but in general they can be used to describe differences in average values of the dependent variable in each category of the nominally measured independent variables and the relationship between the intervally measured independent variable(s) and the dependent variable.

inferential statistics: test of $H_o : \mu_1 = \mu_2 = \ldots = \mu_K$
(null hypothesis that the average [mean] of the dependent variable is equal in each category of the nominally measured independent variable once they are equal on the other variables)

test of $H_o : \rho_{YX} = 0$
(null hypothesis that the association between the dependent variable and the intervally measured independent variable equals zero, once they are equal on the other variables)

test of H_o : there is no interaction
(null hypothesis that the association between the intervally measured independent variable and the dependent variable is the same in each category of the nominally measured independent variable)

5. A <u>Test of Curvilinearity</u> is used when you are concerned that the association between a dependent variable and an intervally measured variable may not be linear in nature. You may then convert the intervally measured variable into dummy variables, conduct an analysis of variance (as shown in 3 above), compute E^2 and compare this value to the R^2 (or r^2) obtained in the simple regression which treated the variable as intervally measured. You may test $H_o : \eta^2 - \rho^2 = 0$ ($E^2 - r^2 = 0$). If you reject the null hypothesis it is not appropriate to use the simple linear model. However, you can examine the scatter diagram and use various transformations of the dependent variable to try to approximate the relationship as shown in the example below.

$$\tilde{Y} = a + b X^{\frac{1}{2}}$$

[Note: In techniques 6-9 described below nominally measured variables may be used as intervally measured independent variables if converted appropriately to dummy variables or some varient thereof. Dummy variables (or their varients) should not be used as dependent variables in these techniques.]

6. Multivariate Analysis of Variance should be used when you are conducting several analyses of variance all with the same independent variables, but different dependent variables, all of which are measured with equivalent scales (e.g. examining the influence of social class, race, and sex on achievement test scores in a variety of subjects). The techniques of multivariate analysis of variance (manova) control for the possibility that you will get significant results from a series of hypothesis tests simply by chance. (e.g. if you do 100 hypothesis tests just by chance you will get 5 results that are significant at the .05 level). Canonical correlation provides similar results when the independent variables are interval, rather than nominal. SPSSx does canonical correlations as a sub-program within manova.

7. Factor Analysis is used when you believe that there are certain types of dimensions which underlie a set of intervally measured variables all of which are measured on the same scale (e.g. a set of variables which measure attitudes of subjects toward Communism). The "factors" are hypothetical variables which represent these variables and which may be represented as linear combinations of each of the variables. A hypothetical "factor structure" with three factors "underlying" 15 variables is shown below.

$$F_1 = f_{11}X_1 + f_{12}X_2 + f_{13}X_3 + \ldots + f_{1,15}X_{15}$$
$$F_2 = f_{21}X_1 + f_{22}X_2 + f_{23}X_3 + \ldots + f_{2,15}X_{15}$$
$$F_3 = f_{31}X_1 + f_{32}X_2 + f_{33}X_3 + \ldots + f_{3,15}X_{15}$$

The "factor loadings" represent the association (actually a correlation) between each variable and the hypothetical factors. Each variable has associated with it a "communality," analagous to an r^2, which represents the proportion of variation which the variable holds in common with the other variables in the factor analysis. Each factor has an "eigenvalue" which is used to indicate the proportion of the common variance which that factor can account for. Factor analysis was traditionally used as an exploratory technique to determine "underlying structures" to sets of variables. In recent years we have realized the problems of such atheoretical searching for results and the use of confirmatory factor analysis has become more popular. This may be done as part of the LISREL package of structural equation models.

8. Structural Equation Models incoporate elements of both path analysis and factor analysis, allowing researchers to develop causal models that use multiple measures of concepts and feedback loops. The term "two stage least squares" is also sometimes used to describe parts of these techniques. The techniques continue to build on the general linear model but incorporate both a "measurement model," using aspects of factor analysis (specifically confirmatory factor analysis) in describing how multiple measures contribute to a concept and a "structural model" that represents the associations between the actual variables of interest. This type of analysis is preferable when you have panel data, multiple measures, and/or feedback loops.

9. Discriminant Analysis is used when you wish to use several intervally measured variables to distinguish between cases in two or more groups. The result is an equation of a "discriminant function" as illustrated below.

$$D = b X_1 + b X_2 + \cdots + b X_k$$

The equation represents a spatial dimension along which the cases are clustered. If there are only two groups, one dimension represents the differences among them; with three groups, two dimensions may be used; in general, with k groups, up to k-1 dimensions may be used to describe their differences. The dimensions are structured so that they maximize the differences between the "centroids" of the groups which are being distinguished. Each of the coefficients in the discriminant function represents the contribution of a given variable to the function and is analagous to a standardized regression coefficient.

10. Cronbach's alpha is a convenient statistic used to summarize the association among variables which may be combined together into an additive scale (e.g. a Likert scale). It is commonly used in reliability programs to assess the reliability of a scale. It essentially represents the degree of association among variables combined together into an additive scale or the extent to which they all appear to be measuring the same thing.

11. Log-linear analysis is a whole family of techniques which are appropriate for multivariate analyses of nominally measured (or ordinally measured) variables. They essentially look at the "odds" of a case falling into a certain category and then represent it as the product of various other category memberships. As a very simple example, and assuming there are no other variables involved, the odds of a person in the U.S. being a democrat, rather than in another party, might be 3/2. The odds of being a democrat are increased if one is non-white, rather than in another racial-ethnic group (say 2/1), and Catholic, rather than non-Catholic (3/2), but decreased if one is high-income, rather than low-income (by say 1/2). Together, however, one can combine the information on race/ethnicity, religion, and income to find the odds of being democrat.

(2)

$$3/2 = (2/1)(3/2)(1/2) = (6/4) = (3/2)$$

odds (dem) = (odds race)(odds Cath)(odds income)

Mathematically, this multiplicative function is difficult to deal with, but one can simply take a logarithm of the function and turn it into an additive function. (Logarithms are exponents and you may remember that when we multiple exponents, we simply add them, as in

$$(10^2)(10^3) = 10^{2+3} = 10^5$$

By taking the odds of the equation above we get a function similar to that below

Ln (democrat) = a + b ln (Race) + b ln (Cath) + b ln (Inc)

As you can see the resulting equation is in the standard format of the linear model. There are standard errors associated with each of the coefficients. Thus, with this technique, it is possible to examine, with relatively succinct models, multivariate relationships among <u>qualitative</u> variables using an interesting and useful variant of the linear model.

Sociology 413/513
Spring, 1992
Final Exam

Please answer each of the questions below as thoroughly as you can, but try to limit your answer to each question to less than two typed pages. You may consult any of your books or notes in developing your answers, but please work independently of your classmates. Do not put your name on the front of the paper. Instead, place your social security number only on the front page. If you would like your grade posted by social security number on my door at the end of the term, please leave me a separate note telling me this. If you want me to mail your exam to you, you may leave me a stamped, self-addressed envelope or an envelope for campus mail. All exams and the fifth assignment are due by noon on Friday, June 12 in my mailbox in room 736 PLC.

1. Suppose a researcher had data from a random sample of the residents of Eugene. She knew how many times the people attended church in the previous month, their age (in years), income (in dollars), and education (in years). She was interested in how these latter three variables could help predict variations in church attendance. What is (are) the dependent variable(s) in this analysis and on what level is it (they) measured? What is (are) the independent variable(s), and on what level is (are) it (they) measured? What type of analysis procedure(s) would you suggest that this researcher use? Why? What descriptive statistic(s) could she use, and what would they tell her? What inferential tests could she use and what would they tell her?

2. Suppose this researcher was interested in how both gender and age affected church attendance and suspected that the influence of age on church attendance might be different for men and women. What would be the dependent variable(s) in this analysis and on what level is (are) it (they) measured? What is (are) the independent variable(s), and on what level is (are) it (they) measured? What type of analysis procedure(s) would you suggest that this researcher use? Why? What descriptive statistic(s) could she use, and what would they tell her? What inferential tests should she use and what would they tell her?

3. Suppose another researcher had information from a random sample of people in Lane County on their political affiliation, specifically whether they were registered as Democrats, Republicans, or Independents. (Those with other affiliations were ignored.) He was interested in how respondents' gender and race influenced their political affiliation. What would be the dependent variable(s) in this analysis and on what level is (are) it (they) measured? What is (are) the independent variable(s), and on what level

is (are) it (they) measured? What type of analysis procedure(s) would you suggest that this researcher use? Why? What descriptive statistic(s) could he use, and what would they tell him? What inferential tests should he use and what would they tell him?

4. Suppose this researcher was also interested in how age (measured in years), income (measured in dollars earned), and education (measured in years of education attained) could help predict respondents' political affiliations. What type of analysis procedure(s) would you suggest that this researcher use? Why? What descriptive statistic(s) could he use, and what would they tell him? What inferential tests should he use and what would they tell him?

The General Linear Model
Addendum to Spring, 1988 Notes

12. Time series is a way to examine changes in variables over time, especially when one has data over a long period, often involving large groups and aggregated measures. In these analyses time becomes an independent variable and we can assess how variations in a dependent variable (Y) are a function of time. Additional variables can be added as predictors so that one can see how these indpendent variables predict changes in Y, either as a constant influence or as covarying with time. For instance, the equation below would represent a situation where some variable changes relatively constantly over time (b indicates the rate of change) and has a relatively constant influence of the variable X_1.

$$Y = a + b\ (Time) + b\ (X_1)$$

13. Proportional hazard models are also used to analyze changes over time. These techniques are also called survivor functions, hazard models, and event history analysis. Unlike most uses of time series, proportional hazard models (or hazard models in general) are ways in which we can predict the probability (or hazard) that individuals will attain a certain condition or have a given event happen to them, such as dying, having a disease, terminating a job, dropping out of school, etc. Thus, the dependent variable is dichotomous and is measured over a large number of time periods. The analysis focuses on predicting the probability that individuals will "survive," or not succumb to the condition. The predictive, or independent variables include time as well as conditions related to the individual and his or her environment. These additional independent variables may be constant over time (a proportional hazard model) or may vary over time.

14. Hierarchical linear models are recently developed techniques that are extremely well suited to multi-level data and analyses, as when a researcher has a dependent variable measured on the individual level of analysis and independent variables measured on both the individual level of analysis and aggregated or grouped levels. If one were to subject these data to a simple regression analysis, one would have to assume that the influence of the individual level independent variables was the same in all the groups (e.g. individual ses influences students achievement in the same way in schools with all levels of funding). A hierarchical linear model, on the other hand, allows one to observe whether or not the influence of individual ses differs among different types of schools (which analysis of covariance does - assuming that the school level funding is

only a nominally measured variable), but also allows one to estimate how the grouped level variable of school funding influences these different influences of the individual variable. Thus, you essentially have regression equations within regression equations. You can see how the individual level independent variables affect the dependent variable and how the group level independent variables affect the way in which the individual level variables affect the dependent variable. These techniques, to date, are not included in standard statistical packages such as SPSS.

15. Meta-analyses are a way of summarizing results from other studies. They have become increasingly important as we begin to build up results in fields and have been used a great deal in recent years in psychology and related areas. In these studies the results of a whole set of studies become the dependent variable. The results of these earlier studies are often summarized through descriptive statistics, such as Cohen's d. Thus the sample would include all studies of a certain phenomenon, such as gender differences in mathematics achievement, and the dependent variable could be the size of Cohen's d obtained in each of these studies. Then characteristics of the studies, such as the nature of the sample (e.g. the age of the subjects or the educational level or the date the data were gathered), are used to help explain variations in these descriptive statistics.

The Sage series of books which you were exposed to this term has monographs on each of these techniques. Some, as you know, are more technical than others, but the more technical ones generally have references to more easily understood materials. In addition, review journals and monographs, such as the Annual Editions series and Sociology Methodology, often have nice review articles of new techniques.