PN-AAM-846
ISN=29321

DRAFT MANUSCRIPT: NOT FOR
QUOTATION OR ATTRIBUTION WITHOUT
EXPRESS PERMISSION

DOING EVALUATION: A HANDBOOK OF PROCEDURES

Howard E. Freeman
Peter H. Rossi
Sonia R. Wright

March, 1978

c Copyright

Final Draft
March 1978

CHAPTER FIVE

EVALUATION AND POLICY MAKING

I. Evaluation in the Larger Context

The previous chapters of this monograph have been concerned primarily with the technical aspects of evaluation, providing guides to persons considering attempting evaluations and to those who may have the responsibility of overseeing evaluation studies. But, as must be abundantly clear to readers, evaluation research is not simply just a technical activity. By virtue of the fact that it is designed to aid in the processes of policy formation and program design, it is also a political activity. By that phrase, we mean that evaluation research is always conducted within a context in which there are many interested parties whose stakes in the outcomes of the research affect both the evaluator's ability to carry out evaluations effectively and the ways in which evaluation research results are employed by decision makers. This chapter attempts to discuss some of the issues arising in the conduct of evaluation research that stem from the fact that such research ordinarily takes place within a more or less politicized environment.

A. The Interests Engaged by Evaluation

To evaluate means to make judgments; to conduct evaluation research means to provide social science based findings that can be used in making judgments. The distinction between making judgments and providing facts upon which judgments can be based is a valid distinction, but one which is often difficult to discern in practice. In short, the results of an evaluation research may be

seen by the parties involved as amounting to judgments, and reacted to accordingly.

What are the parties that are typically involved in the results of an evaluation? Listed below are some of the groups that typically may become parties interested in the evaluation research process:

1.	Policy Makers	Agency (or representatives) ultimately responsible for decisions on whether program is to be instituted, continued, discontinued, enlarged, or decreased.
2.	Evaluation Sponsor	The agency or its representatives who initiates and funds the evaluation research.
3.	Program Sponsor	The agency (or representative) who initiates and funds the program to be evaluated. (Sometimes the evaluation sponsor and the program sponsor may be identical.)
4.	Program Administration	The agency (or representative) responsible for the running of a program. (Sometimes identical to 2 above.)
5.	Program Deliverors	Persons responsible for ultimate delivery of the program (e.g. teachers in an educational program) (Sometimes identical with 3 above)
6.	Evaluator	Agency (or representative) responsible for design and/or of evaluation research.
7.	Clients	Persons, households, or other units who are targets of program that is being evaluated.
8.	Program Competitors	Agencies who compete with program administration for the funds involved.
9.	Context Interests	Agencies, groups, individuals and other units in the immediate environment of a program (e.g. local government officials in a community in which a program is placed.)

Although these nine groups do not completely exhaust all conceivable interests that might be involved in the politics of evaluation, they do constitute groups that typically become involved in one or another ways in an evaluation. Nor is it completely clear how the interests of each of the groups would be engaged by a given evaluation research's outcome. In a particular situation, policy makers may be more pleased by a positive evaluation than a negative one; in other situations their reactions may be quite reversed. Perhaps, the main reliable prediction that one could make is that program sponsors, program administrators and program deliverors are most likely to regard positive evaluations favorably and to have the opposite reaction to negative ones. Such reactions, are, of course, quite understandable since they are the ones with the most stake in the continuance of a program and whose activities appear to be most clearly judged within an evaluation research's report.

All too often the novice evaluator having completed his or her report and proudly announced the results may be devastated by a torrent of criticisms and negative reactions to the findings. While it may not help to know that this is a common occurrence, it is useful to know that evaluation can be controversial and that the evaluator may often be at the center of the controversy involved. Knowing this in advance, it is possible to take steps that can reduce the amount and intensity of such controversies.

First, it is important that the major interests that might be

engaged in an evaluation be brought to understand the terms of the evaluation and to understand something about the method that is to be employed. This is particularly important with respect to the program sponsor, program administration, and program deliverors. The evaluator ought to try to develop some consensus with these groups over the aims of the program that are to be evaluated, over the methods to be employed and the possibility that the evaluation may turn out in ways that could reflect negatively upon them. This process of building consensus (and committment) should begin before evaluation is undertaken.

Secondly, the evaluator is best off when his/her position is as independent as possible from ties with program administration and program deliverors. Some would go so far as to state categorically that evaluation whould not be accomplished within the administration of a program but should always be conducted by an outside agency or person. While we would not go that far in urging separation of evaluation from administration, we would nevertheless urge that such "inside" evaluators work out especially clear understandings of roles and limits with program administrators. It seems to us to particularly appropriate for inside evaluators to conduct the kinds of evaluation activities described in Chapter 2, (i.e. defining target populations, estimating sizes and locations of target populations and monitoring program implementation,) all activities of direct utility to program administration and useful in fine tuning a program to accomplish its objectives. More hazardous is the conduct of

program impact studies (as described in Chapter 3) as an "inside" evaluator, since the evaluator may find conflicts between his/her loyalty to the administration of the program and the demands of the task of evaluation.

Thirdly, one of the major points of conflict ordinarily arises out of whether the terms of evaluation were properly chosen, especially whether the goals of a program were correctly specified. Here it is important to note that it is not necessary that a program have one goal or a small set of goals in order to be evaluated. The criteria we stressed in earlier chapters were only that the goals be clear and that they be measurable. It is possible, therefore, to evaluate a program from several viewpoints, for example, from the viewpoints of goals set respectively by policy makers, program administrators and from the respective viewpoints of clients and even other community members. Again, this process of deciding upon the terms of evaluation and accomodating a range of viewpoints into the evaluation is an act that should be undertaken <u>before</u> the evaluation begins in consultation with the interested parties involved.

Finally, it is important that the more controversial an evaluation may be, the more carefully it should be carried out. Anticipated controversy may be a sure sign that the program is important at least to some of the interested groups: The more important the program the more powerful should be the evaluation research design so that decision errors (False Positives and/or False Negatives, as the case may be) are kept to a minimum.

Furthermore, the more carefully executed the less vulnerable an evaluation may be to carping criticism.

B. Political Time and Research Time

Evaluation research, just as other types of research, takes time, and the tighter the research design, other things being constant, the longer the evaluation will take. For example, the large scale social experiments which gauge the effects of innovative programs ordinarily take from four to eight years to complete and produce reports. The political world moves on a much faster time scale, however. Out of this disparity in time demands grows another problem for evaluators.

The same political atmosphere that makes it possible to do evaluations especially of ongoing social programs is usually the same atmosphere in which decisions are about to be made about such programs. Hence, often there is considerable pressure for evaluation to be done more quickly than the better methods can allow. There is also often much demand for release of preliminary results, often before the researchers are quite certain that their results are completely firm. At times, evaluation researchers have been asked for their "impressions" of effectiveness even after saying that they had not yet produced any firm results.

It is not at all clear what can be done to reduce the degree of pressure arising from the different time schedules of research and decision making. It is clear that an evaluation should not be

undertaken if the information is needed long before the evaluation can be completed. It may be better in such circumstances to suggest that the judgment of experts be relied upon or that some of the more judgmental methods (discussed in Chapter 3) be used for the evaluation in question. In short, it would be better to have some information of the highest technical quality possible, given time constraints, than to have no information at all.

A more strategic approach is to confine the technically more complex evaluations to programs that are not likely to be reviewed in the near future or to programs that might be considered in the future as ongoing programs. Thus randomized controlled experiments may be most appropriate to evaluate the worth of new proposed programs long before such programs appear upon the agendas of decision making bodies. Or, extensive cross sectional analyses may be applied to programs that have a steady support, e.g. pension plans for the aged.

C. Value Conflicts and Evaluations

Although traditional views of the social sciences exhort social scientists to be value-free, it is not obvious that such a stance should be taken to begin with nor that social scientists can be completely value free in practice. After all, each one of us -- whether we are evaluation researchers or ordinary citizens -- lives within political systems and has the ordinary obligations of citizenship and possibly additional obligations of public office or public

employment. It is almost impossible to remain neutral in every aspect of public life and hence to require neutrality of anyone who is concerned with social programs is to demand more than usually can be expected.

As a consequence evaluators are not persons without opinions concerning whether the programs that they are to evaluate are likely to be successfully effective. That such opinions may affect their approaches to evaluation goes without saying. But, the evaluator also has the obligation to consider not only his/her own estimates of how well a program may be doing, but also the contrary positions. Thus an evaluation should always be a test of which of a set of possible outcomes comes closest to being true of the working of a program. In short, the obligation of the evaluator is not to be neutral in thought or opinion, but to be neutral in action. To be neutral in action means that the evaluation design should not be biassed for or against any particular outcome, whatever the evaluator (or other parties to the evaluation) may believe about the worth of the program in question.

Although no actual examples have yet appeared, some writers on evaluation have advocated that the evaluation of a controversial social program ought to be undertaken by several evaluators each representing the major interested parties. Thus, an evaluation of a controversial land resettlement program might be undertaken from the viewpoints, respectively, of the settlers, large landholders whose holdings were being used for resettlement purposes, and of the central government agency sponsoring the program. Each evaluator would be presumably examining

those effects that engaged the interests of their particular evaluation sponsors. Under such circumstances, each evaluator would have no particular obligation to search for those benefits accruing to parties other than the particular sponsor that has employed the evaluator.

There are many reasons why such an approach -- called advocacy evaluation -- does not yet have any examples. For one thing, the costs of doing, say, three evaluations are triple the costs of doing one and evaluation costs are heavy enough under present circumstances. Secondly, each of the sponsors would have to bear the costs of the evaluation taking its point of view, and the resources of the interested parties are rarely equal. Thirdly, under such circumstances evaluations would become simply another form of political struggle and an expensive form at that. Finally, if evaluators take seriously the obligation stated above to entertain seriously the possibility of alternative outcomes and to provide for their testing within the context of a single evaluation, the aims of advocacy evaluation would be fulfilled without becoming another form of political struggle. Although no code of ethics has yet to be adopted by evaluation research societies, it seems likely that the code will not be very different from those that have been adopted by other researchers. First and foremost, the evaluator has an obligation to remain a scientist in the sense that his/her first obligation is to see that the research design and the analyses of the resulting data conform (within

budgetary constraints) to current state of the art standards of quality. Secondly, the evaluation expert has the obligation to take into account in the design the differing perspectives and expectations of major interested parties. Finally, while no evaluator may be expected to be neutral entirely in his/her private opinions, those opinions should not dominate over the first two considerations mentioned above: Neutrality is not to be expected in opinions, but is to be expected in action.

II. The Role of Evaluation in the Political Process

As we stated at the outset of this volume, the purpose of evaluation research is to aid in the making of rational political decisions concerning social programs. All too often, evaluators, having completed their assessments of a social program, have been depressed by the slowness with which the decision making process has reacted to their findings. Even more depressing have been the numerous occasions when evaluation findings have been seemingly ignored in subsequent political processes with programs being either continued or terminated without much explicit regard for assessment outcomes. In short, all too often it has appeared that the valuable and sometimes expensive information obtained through evaluation research has been ignored by the political decision making process.

There can be no doubt about the feelings of the evaluators -- indeed, each of the authors of this volume has on one occasion

or another felt that his or her labors have been expended in vain. But, is this a justified feeling? The answer to this question is that one can only expect that the results of an evaluation will be but one of the inputs into decision making. The parties to a social program, as outlined in the opening section of this chapter, are many. Program administrators, program sponsors, program operators and sometimes the clients of a program often have very strong stakes in the continuation of a program and their representations to that effect to decision makers may count more than the results of an evaluation research. The outcomes of a political process may be viewed as balancing the representations of a variety of interested parties, with the outcome of an evaluation research being simply one more argument on one side or another.

To imagine otherwise would be to provide to evaluators and evaluation researches a special place in the political decision making process that amounts to a veto over such decision making, a move that would strip from the decision makers degrees of freedom in satisfying constituency demands that no one would want to see occur. Indeed, under such circumstances evaluators would become a species of philosopher kings whose pronouncements on particular programs would over-ride the representations of all the other parties involved. In any political system that is sensitive to weighing and assessing the conflicting claims of a number of constituencies, one can expect that evaluation research outcomes will play the role of an expert witness to the effectiveness of a program but that, as with other expert witnesses, such testimony

may be given more weight than shrewd guesses but not determinative weights in deciding an outcome.

What then can one expect to be the role of evaluation research in determining the shape and direction of social policy? There are three important roles that have been distinguished: First, evaluation researches have been useful in deciding technical questions, e.g. whether a program should take one or another form. Secondly, on occasion, evaluation researches have provided definitive information upon which decision makers have acted in line with the outcome of the evaluation, resting their decision on the outcomes in question. Thirdly, evaluation researches have had some very general impacts on political decision making in the form of changing the grounds of political argumentation. These three political roles of evaluation are discussed in more detail below:

A. The Role of Evaluation in Technical Issues

By definition, a technical issue is one which is neutral as far as policy and values are concerned: that is, the various parties involved in a social program have no stake in whether the issue is decided one way or another. For example, it may make little difference to any of the parties in an adult literacy program whether the teaching method employed is one which relies very heavily on teaching persons to recognize whole words or to build reading ability on the individual letters of the alphabet. Under such circumstances, an evaluation research which could prove the clear superiority of one method over the other might have very

little difficulty in getting its results adopted by program administrators. Or, it may make a considerable difference whether one approaches farmers through their local leaders or through the mass media, as far as acceptance of the use of fertilizers is concerned. In the two cases cited above, the parties interested in the program care little which technical means is employed to achieve an end.

To resolve a technical issue in the implementation of a program may be among the most important roles that evaluation researches may perform. Most of the activities described in Chapter 2 address themselves to what are ordinarily regarded as technical issues -- the size, location, and distribution of target populations, the means used to recruit participatns in programs, the precise means employed to persuade target populations to participate, the methods used to teach new techniques, and so on. Clearly, there are many technical issues to be resolved in the design, execution and implementation of social programs and the role that evaluation research can play in deciding such issues is quite considerable and important in assuring that a program is proceeding properly.

B. <u>The Role of Evaluation As Program Validation</u>

The impact assessments described in Chapter 3 are intended to assess whether or not a program is achieving its intended results. The experience in the United States in such impact assessments is that the expected value of estimated effects is zero. There

are many reasons why such is the case, (and can be expected to be the typical outcomes for other countries in which social programs are highly developed or more highly developed). The programs that are tested are ordinarily ones that are slight alterations to or small increments to existing programs. Anticipated effects of such slight alterations are necessarily slight and therefore hard to detect. In addition, program alterations are usually designed to meet needs of target populations that are harder to reach than the program in question has already been reaching. For example, it is easier to reduce illiteracy in a country where the literacy rate is, say, 30% than in a country where the corresponding rate is 90%.

Few social programs have been eliminated because their evaluations have shown them to have essentially zero effects. For example, the very large program to provide pre-school training in reading and the recognition of numbers, Head Start, has been evaluated with results that indicate that young children who have gone through the Head Start experience do not differ substantially (or at all) in their subsequent school performance. While the results of the evaluations have been challenged, neither the partisans of the program nor the program's detractors have ever claimed that at best the program's effects are more than slight. The considerable partisan support for the Head Start Program has managed to generate enough support in the U. S. Congress to keep the program well financed for the past few years.

When social programs have been discontinued after impact assessments have shown them to be ineffective, it has often turned out to be the case that such programs had little partisan support to begin with. For example, the American Experiment in Performance Contract Learning was discontinued after a very hasty evaluation showed it to be largely ineffective.[1] The program was a test whether private contractors who would be paid on a sliding scale depending on how much the students would learn could do a better job of teaching elementary school children than regular school systems. The program met with the opposition of teachers associations, local school systems and other parties interested in the public schools. Faced with the preliminary findings of no effect, testing of the program was discontinued abruptly. Had the contract learning program achieved the support of school personnel and administrators perhaps the decision to terminate the program might not have been made so abruptly and experimental testing of the program might have continued for an additional period of time.

These experiences illustrate some of the points made earlier in this chapter. An impact assessment is only one of the inputs to the decision making process. Depending on whether the program has a great deal of support or opposition, an evaluation may be set

---

[1]Some evaluation experts after reviewing the evaluation have judged that the program was not properly implemented to begin with (e.g. in some of the experimental schools, the contractors had extreme difficulties in getting the appropriate equipment, hiring teachers, and so on) and that there were some signs in schools where it had been fully implemented that the contractors were achieving some slight improvement in learning over the ordinary school experiences in those schools.

aside or be used to bolster a particular decision.

It should also be stressed that these experiences are to some unknown degree conditioned by the nature of the American political system and the stage in which social programs are currently in the United States. It may well be the case that in political systems in which central administrations are stronger and less responsive to a variety of constituencies, impact evaluations may turn out to be more important elements in decision making. It should also turn out to be the case that in countries which are just developing their social programs, impact assessments might be more likely to show strong positive or negative effects for new social programs, a circumstance that might shift the information value of an evaluation research upwards considerably.

C. Evaluation Research As General Information on Social Problems

We have stressed so far in this volume that the main role of evaluation research is to provide information on specific social programs. However, there is another role that in the long run may turn out to be more important in the design and running of social programs. This role is that of providing general information on how social programs operate and on their effectiveness. Thus, while no specific program that is found to be ineffective may be discontinued, once started, it may well be the case that these negative results affect whether new programs using the same philosophy will be started.

For example, one of the major reasons why manpower training programs have received less emphasis in American unemployment policy has been the fact that it has been very difficult to discern that they have had any major impact on either participants or the general unemployment rate. Similarly, the last decade and a half of evaluation researches on educational programs has led the American government to invest more heavily into basic research on learning since the programs all apparently showed that existing innovations at best were marginal improvements over existing educational practices.

One of the major long run effects of evaluation researches, we believe, is an increasing sophistication in our knowledge about how social programs operate. This general rise in knowledge has filtered in the first instance into the administration of social programs, to some degree into legislative bodies and to a possibly lesser degree to the general populace. It has made special inroads into the social science curricula in universities and may well serve as the knowledge base for the decision makers of the future whose training will be undertaken in those institutions.

III. The Future of Evaluation Research

There are many signs that evaluation research is among the fastest growing fields within the social sciences and becoming increasingly important within agencies that have responsibilities

for social programs. There are many reasons for this growth in size and importance.

Perhaps the main reason is an increasing skepticism among decision makers --legislators, public officials, administrators, and so on -- that common sense and conventional wisdom can design social programs so that they achieve their intended goals. Decades of experience with attempts to solve the problems represented by an exploding population growth, the maldistribution of resources within and between societies, with popular discontent, rising crime and delinquency rates, apparent weaknesses in traditional institutions such as the family and kinship, and so on, have led to a realization that these are obstinate and difficult issues to deal with. This skepticism has led decision makers to seek ways in which they can learn more quickly and efficiently from their mistakes and to capitalize more rapidly on effective measures. To fund an evaluation is to express that skepticism or at least to state implicitly that a proposed social program may not be as effective an answer to a problem as hoped.

A second major reason for the growth of evaluation research has been the growth in the knowledge and technical bases of the social sciences. The development of the sample survey has provided the social sciences with a powerful information gathering method. When coupled with the more traditional experimental methods in the

form of field experiments, the social sciences have developed techniques that are peculiarly adapted to the testing of social programs. Advances in measurement, statistical theory, and in substantive branches of the social sciences have also added to the increased ability of the social sciences to take on the special tasks involved in evaluations.

Finally, there are also changes in the political climates of our times, that lead to more concern for social program. Increasingly we have come to believe that social problems are not fixed features of the condition of humankind but rather subject to change and amelioration through the reconstruction of human institutions. We believe more than our ancestors did that societies can be improved, and that the lot of all can be enhanced by the betterment of the disadvantaged and deprived.

All the above trends appear unlikely to subside, much less reverse in the next few decades. This means that we can anticipate increasing interest in evaluation research, more persons with appropriate social science training being turned out by our universities, and more support for evaluative activities in many quarters. Of course, evaluation research as practiced may fall far short of the hopes and aspirations with which we and other social scientists have invested evaluation research. Our best strategy is to practice evaluation research at the highest possible level of the state of art. We hope that this volume will contribute the wide diffusion of current state-of-the-art practices.

Second Draft
March, 1978

T E C H N I C A L A P P E N D I X

I. Sampling

II. Interviewing

III. Measurement

IV. Descriptive Statistics

V. Statistical Inference

VI. Data Reporting Techniques

VII. Computer Software Packages

The purpose of this Appendix is to introduce the logic and concepts for some of the basic methodological and statistical procedures required for evaluation research. For evaluators who have not been trained in social science methods and statistics, these notes might provide a starting point and alert for some of the principles and problems, since topics are covered only sketchily. We strongly urge that additional sources be consulted, both by novices and the more experienced researchers, since ultimately the quality of the research is a function of the extent to which methods are well-executed.

I. Sampling

A. Basic Concepts[1]

For reasons of cost and efficiency, evaluations of large-scale programs often preclude researching the entire target population which received the intervention. Similarly, if the intervention includes a number of activities or treatments, it may not be feasible to study all. Consequently, unless the program is relatively small-scale, evaluation requires that a sample of program recipients (and nonrecipients), or a sample of interventions, be studied.

For example, the evaluation of an agricultural development program operating in a number of regions and intended to impact

---

[1]An excellent reference for sampling is Seymour Sudman's Applied Sampling (New York: Academic Press, 1976). As its title implies, this is a source book with a clear, applied emphasis and invaluable for practitioners. An older "classic" is Leslie Kish, Survey Sampling (New York: John Wiley and Sons, 1965).

several thousand farmers, requires that a sample of the farmers be selected for study, since it would be too costly to study all. In addition, the errors that would arise from gathering, processing and analyzing data for a very large population are avoided. A well-executed sample can therefore be the preferred strategy. Even a pilot program in a small number of communities might require sampling. When implementation and impact can be studied on all the persons involved with the program, a sample of those not involved with the program still has to be selected (depending on the research design as elaborated in Chapter 3), to allow the relative comparisons of output (impact).

Although samples are almost always studied, the purpose of research is to generalize to (make inferences about) the entire population of interest on the basis of sample data, to estimate and test hypotheses about population parameters. Therefore, it is essential to draw samples such that statistically valid inferences can be made about the population. Heuristically what this means is that the sample should be representative or reflect with "fairness" the population. A biased sample will produce results which cannot be validly generalized. Fortunately, a variety of sampling techniques have been developed which, if properly executed, allow valid inferences about the population or universe of interest (the terms are used interchangeably). Which sampling method to use is determined by the nature of the population and the available resources.

The first step is to define the population of interest, by

specifying whether it is a target population of individuals, households, groups, institutions or whatever. These constitute the units of analysis for the study. In evaluation research the specification of units of analysis for the study. In evaluation research the specification of units of analysis should be straightforward: whatever the program or intervention specifies as the target population becomes the unit of analysis. The decision of what units to exclude --that is, establishing the boundaries of the population--can also be derived from program rules. Program eligibility requirements are often stated in reference to geographic, economic and demographic characteristics of the units and thus provide boundary definitions. Examples are regions, villages, agricultural sectors; age categories, employment status, levels of income or education for individuals; size of groups, farming methods, and so on for each of the appropriate units.

The nature and characteristics of the population determine (at least in part) which type of sampling method is feasible. The major difference is whether or not there exist records or lists of the population from which a sample can be drawn, or if such lists are unavailable, how they can be constructed.

For inferential purposes, the most important requirement is that random sampling be employed. Probability or random samples are those samples in which every element (unit) in the population

has a known, nonzero probability of selection. Random does not mean "haphazard" or "capricious" but that some well-known probability method is used to select the sample.

B. Simple Random Sampling

Whenever there exists a list, or records can be used to generate a list, (also referred to as the sampling frame), of the entire population then simple random sampling can be employed. The procedures for simple random sampling are easy and inexpensive once the list has been assembled.

A simple random sample is one in which every element in the population has an equal probability of being included in the sample (hence every sample is also equally likely to be drawn). Furthermore, sampling takes place at one stage, with elements of the sample selected independently of one another.

Suppose, for example, that an evaluation was required of an agricultural training program for adolescents e.g., the Centers for Rural Education established in the Upper Volta during the 1960's by UNESCO. The project consisted of 2,137 centers capable of handling around 30 students each. A monitoring evaluation was to be done, with center as unit of analysis, to ascertain how the program was being implemented, what it consisted of and who was receiving it. Subsequently an impact evaluation was to assess the effect of the training program on the trainees after returning to their villages,

in terms of such outcome measures as farming methods and production increases. Detailed investigation could not be conducted on all the centers nor on the approximately 60,000 students who participated in the three-year training program. Since the lists of centers and students were complete and available, a random sample of centers was first drawn, from which another random sample of students was later selected.

The most convenient and accurate procedure for obtaining a random sample is to use a table of random numbers. Portions of the Rand Corporation tables[2] are reprinted in many statistics and methods texts, are shown in Sudman (1976), and are also available in many computer programs.

The first step is to number the population of elements (units) --for the above example, the 2,137 centers. A table of random numbers is easy to use in drawing the sample. The numbers are random in any direction and from any starting point (they are grouped in the tables only for convenience), and therefore it does not matter where one starts or in what direction one goes. One way is to start at the top (of any page) and move down. All that is required is that one work with sufficient columns to give each element in the list an equal chance of selection. Thus, in sampling from the list of

---

[2]Rand Corporation, A Million Random Digits with 100,000 Normal Deviates, Glencoe, Illinois: The Free Press, 1955.

2,137 centers one needs to use 4 columns of digits. Read down the columns of random numbers and keep in the sample any number from 1 to 2,137 which comes up, discard any number bigger than 2,137. The procedure continues until the required sample size is achieved.

The most common error among novices drawing simple random samples is to reject the sample obtained because somehow it "does not look right" and to start discarding elements, replacing them with new ones until a "more likely looking" sample is obtained. The result is no longer a simple random sample but a judgment sample. An alternative, if indeed the simple random sample exhibits peculiarities (such as the 50 centers selected are all from a very small geographical area), is to discard the entire sample and start from the beginning (from a different random start in the table) and draw another sample.

It is worth stressing in this context that random sampling does not guarantee that any single sample will be "representative" of the population, since random sampling can produce peculiar samples (recall that in simple random sampling any sample of elements could be drawn--e.g., the most effective training centers). For example, in tossing a coin 10 times, it is possible to get 8, or 9, or even 10 "heads." But such "unusual" outcomes are not as <u>likely</u> to occur as, say, 4 or 5 or 6 "heads." This is because if we toss a coin only once or a few times, the outcome is less predictable than if we tossed it 1,000 times and recorded the outcomes.

Thus, the merits of random sampling are understood in terms of

probability processes: in the long run, if random samples were drawn repeatedly from a population, on average all those samples would provide rather accurate estimates of the population. Moreover, for random sampling the extent of variability or error can be estimated, since statistical tests of significance can be applied to address the question of whether or not sampling error could have produced the results.[3]

C. Systematic Sampling

An alternative random sampling method which is particularly useful for sampling from a very large listed population is a systematic sample, where cases are selected at given intervals. For example, if 200 cases are to be selected from a population of 10,000, one can select every fiftieth case. Note that the entire list does not need to be numbered. Two things are required for systematic sampling: the sampling fraction (or its inverse, the sampling interval) and a random start.

The sampling fraction is simply the ratio of the desired sample size, say N, to the number of elements in a population, say M: $f = N/M$. The sampling interval $i = M/N$. In the previous example, for a sample of 50 out of 2137 centers, the sampling interval is $2137/50 = 42.74$. One would round up or down, to 43 or 42 without any noticeable effects on the accuracy of the sample. To obtain a random start, a table of random

[3]These ideas are discussed further in section V.

numbers is used to select a number between 1 and i (the sampling interval). This ensures that every element in the population has an equal chance of selection and avoids the small bias that would be introduced if the first or last element in the population were always selected. In the sample of 50 from the list of 2137 with a sampling interval of 42, suppose one obtains the random start 13. Then the selected elements in the population would be 13, 55, 97, 139 and so on through 2113. If the centers are not numbered, one would simply count and select out every 42nd case, starting with the 13th on the list.

Systematic sampling is commonly used when choosing a sample from directories such as city or telephone. The procedure might amount to taking every fourth page and the tenth entry down. One has to know how many pages the directory contains and how many entries there are per page. Sudman (1976) describes how one can use length, say on a strip of cardboard, to measure off the intervals. Systematic sampling and some of these shortcuts can also be used when the population is not listed but is arranged on cards or file drawers, making sure that the files are all the same thickness.

The most serious problem of systematic sampling, <u>periodicity</u>, occurs when the list is arranged in some order which coincides with the sampling interval and thus results in a nonprobability sample. The most common example is when sampling through time periods since there are often seasonal variations in socioeconomic processes. In sampling from a large number of years, if the sampling interval

happens to be 12, one might end up with the peak unemployment period, or the lowest production month. But periodicity does not only occur as a function of time. For example, in cities or towns it is often the case that corner houses are more expensive. In selecting a systematic sample of houses from block maps (e.g., every tenth house), the sampling interval might coincide with or oversample corner houses. An evaluation of housing quality would thus overestimate quality. Therefore, lists have to be inspected carefully prior to the decision to opt for systematic rather than simple random sampling. In general, alphabetized lists are unlikely to contain periodicities.

D. Cluster Sampling

For many, if not most, evaluation problems, the populations of interest are unlisted and it would be much too costly to have to generate a complete list. In such cases other probability sampling techniques are available which require fewer resources. One is cluster sampling, which takes advantage of the fact that even widely dispersed population units such as individuals and households can be found in geographic or other natural clusters.

If clusters rather than individuals are sampled, then data-gathering costs are reduced. Individuals or units within clusters are obviously much less dispersed than if they are sampled randomly from the entire population. Therefore, listing costs and travel time for interviewing are greatly reduced,

In the simplest cluster design (single-stage), clusters are randomly selected, and then every individual or unit within each cluster is studied. For example, to estimate the adult illiteracy

rate in a large city, we might draw a random sample of city blocks and then interview all the adults in each block. Similarly, in the example of the agricultural centers, if we had to assess impact on trainees, we might draw a random sample of centers and then interview all the trainees from each center.

But cluster sampling is even more useful when multistage designs are used. For the illiteracy rate example, after a random sample of blocks has been drawn, a random sample of housing units within blocks could be drawn, and then within households a random sample of adults, where lists would be constructed at every sampling stage. Thus, cluster sampling essentially involves (simple) random sampling in stages. If the clusters are geographic, cluster sampling is also known as "area sampling." It proves to be the most efficient method for obtaining national samples of households or adults. The technique can be used whenever the population units are located in clusters, such as schools, hospitals, firms, villages, clinics, pilot farms, distribution centers and so on.

The characteristics of a cluster sample and its usefulness depend on the <u>heterogeneity</u> of the clusters in terms of the important variables being studied. The aim in cluster sampling is to select clusters that are as heterogeneous or varied as possible. Otherwise, if the clusters are very homogeneous, even very large samples will provide little information. For example, there is less variability within neighborhoods than within villages or towns or cities in

terms of the socio-demographic characteristics of the residents. Similarly, if an intervention is carried out in many different localities and a cluster sample ends up with a very small number of contexts, then the evaluation would tend to underestimate the variation or heterogeneity in program implementation and/or targets. Sudman (1976) provides useful guidelines and examples of cluster sizes and procedures for estimating or measuring homogeneity within clusters.

Cluster sampling does not require that the first sampling units (known as primary sampling units) be of the same size. For some sampling problems, however, clusters can be defined initially so that each cluster contains approximately the same number of elements, in which case the sample is "self-weighting." Where naturally appearing clusters are not of the same size--for example, states, counties, districts, towns--the recommended procedure is to sample proportional to population size. Otherwise, the probability requirements are violated. For instance, in drawing a simple random sample of counties or municipalities and then a simple random sample of counties or households within counties, people living in smaller counties would have a higher probability of appearing in the sample than those living in larger counties.

E. Stratified Sampling

Stratified sampling represents a different approach when the clustering does not occur naturally in the population under study,

but instead, the researcher divides the population into groups or categories (called strata.) Independent samples are then drawn (usually simple random) from each group or stratum. In evaluation problems a common instance is when the strata consist of program participants and nonparticipants. Other stratification criteria might be early versus later stages of program development and implementation, urban versus rural location, or size of center. For other evaluation problems it may be important to compare impact on program recipients according to an important independent or intervening variable, such as ethnicity, employment status, education or geographic location.

Whenever a simple random sample of the population would not include a sufficient number of cases from the relevant categories (strata) which need to be compared, stratified sampling is appropriate. For example if a program has been administered to a small proportion of the population in a given area (e.g., a fertility control program aimed at women under 30) a simple random sample would not be efficient because a very large sample would be required to find enough eligible women. If, on the other hand, a program has been administered to all adults (e.g., a literacy program) and an important research question was whether or not there were any differences in program effectiveness by sex, a simple random sample would be expected to provide approximately equal numbers of men and women.

Two types of stratified samples are possible, proportional or

disproportional. If the sampling fractions for each stratum are equal, that is, strata are sampled proportional to their rate of occurrence in the population, the sample is a proportional (or self-weighting) stratified sample. For example, to estimate the impact of manpower training programs for urban workers, it may be necessary to stratify by size of urban center, especially if the distribution of urban centers by size is highly skewed. A simple random sample of urban places would give the same probability of being chosen in the sample to the very largest places and the very small places. Hence, it would make sense to stratify by size of urban place first, picking places at random within each stratum. Area probability samples of the United States, for example, ordinarily make strata by size of place with the largest places, e.g. New York, Chicago, Los Angeles, etc., being placed in a stratum by themselves in order to insure that such large places are not under-represented in the sample of places eventually chosen.

A more likely rationale for stratifying is in order to obtain enough from a group which otherwise would not appear in sufficient numbers in proportional sampling. In fact, for statistical analysis the optimum sample is one where the subgroups to be compared in the sample are of equal size. Therefore, one has to sample more units from the strata which comprise a smaller proportion of the population. This is referred to as a disproportional stratified sample. When generalizing to the population, estimates must be weighted by the proportion that the strata represent in the population.

Such disproportional stratified sampling is typically applied to circumstances in which there is some special interest in a relatively rare subgroup. For example, in a study of secondary schools, if there is a special interest in vocational schools as compared with academic schools and in the population only 15% are vocational, then it would make sense to make two strata of schools and oversample among the vocational to get an equal number of each. Suppose that there are 1,000 schools and a sample of 100 is required. To get 50 vocational and 50 academic schools, the sampling fractions for each stratum have to be different: 50/150 or 1/3 for vocational schools and 50/850 or 1/17 for academic schools. Of course, to make generalizations about all schools, it is necessary to weight-up the undersampled stratum (academic schools): the weights would be .15 for vocational and .85 for academic schools, where the weights are the proportions of each stratum in the population. More simply, we can simply multiply each stratum by the inverse of the sampling fraction used: 3 for vocational and 17 for academic.

In sum, there are many instances in evaluation research where stratified sampling is necessary. In effect, the notion of disproportionality is important when assigning cases to experimental treatments in order to allow the comparisons desired for policy relevance.[4] In order for stratified sampling to be feasible, a

---

[4]Kish (1965) is a good reference for the issue of disproportionality. Watts and Conlisk (1968) developed an optimal allocation method used in designing the New Jersey Income Maintenance Experiment (David Kershaw and Jerilyn Fair (eds.), The New Jersey Income-Maintenance Experiment, Vol. I, New York: Academic Press, 1976.)

population list is required, as well as additional information on which the stratification is to be based (e.g., age, education, or other characteristics of the units or of the intervention.) The net effect of stratification is to decrease the variation within strata, (at least in terms of the stratifying variable), which requires further statistical adjustments (Sudman, 1976).

F. Quota Sampling

Quota sampling is a non-random technique used in the past by market researchers and pollsters which provides economical, convenient and fast data. The characteristics desired in the elements to be interviewed are specified (e.g., sex, race, length of program participation) and then the selection of elements or respondents is left up to the interviewers. Interviewers thus select and interview a quota of persons who possess the desired characteristics. Such procedures are obviously nonrandom, but a series of constraints can be imposed on the ultimate selection of units and can be combined with prior stages of probability sampling, such as cluster.

In this connection is is worth pointing out that, in general, the field costs of conducting a survey are roughly one-third for actual interviewing, one-half for travel and locating and waiting for respondents, and the rest for clerical costs. Hence the appeal of quota sampling (cluster, as well) to reduce costs. The most serious problems, however, involve taking respondents who are most conveniently available. More modern approaches attempt to rationalize the search procedure by specifying interviewer travel patterns for locating respondents, and requiring that interviewers visit predesignated

households or dwelling units. Sudman (1976) again provides a useful discussion of the costs and potentials of such quota approaches with constraints.

G. Telephone and Mail Samples

Other efforts to reduce field costs involve telephone samples, which unfortunately still have limited application for evaluations in less developed nations. If telephone directories or lists can be used to draw the desired sample, simple random or systematic sampling can be employed. The most common omissions from telephone directories, however, are persons or households of lower socio-economic status who are more likely not to have telephones, and upper income households who might choose not to list their numbers.

The U.S. experience suggests that survey costs can be cut substantially and equally valid and reliable information obtained by telephone as opposed to face-to-face interviews. The newest techniques are known as random digit dialing and use tables of random numbers to select numbers. In this way even unlisted numbers have an equal probability of being selected. If the evaluation problem is amenable to telephone interviewing, a reference such as Sudman (1976) should be consulted. Different equipment can be purchased which selects the random digits.

Mail samples also can reduce data-gathering costs substantially. Whenever addresses for the units being sampled are known or can be obtained, and the research problem is such that a self-administered

questionnaire can provide the necessary data, a mail survey would be feasible. In planning mail samples, however, the researcher should keep in mind that nonresponse can be high and that several follow-up mailings might be necessary for those who do not respond initially. Thus, initial postage and printing costs have to be augmented to include follow-ups, usually as many as 2, and postage so that respondents do not incur the cost of return postage. Additional costs include the time required to receive all the questionnaires.

H. Sample Size

One of the most difficult questions to answer is how big the sample should be. In some cases the sample size may be prespecified by those funding the evaluation. The more likely case is that the total resources, both in terms of funds and time, are set for the evaluation, which essentially amounts to a predetermined sample size. A useful rule of thumb when reserach costs are fixed is to spend about half the total amount for data collection and the other half for data analysis. This determines the sample size once the sample design and data collection procedures are specified.

When costs have not been fixed and the researcher has to make a sample size determination, the precedent of previous research can be useful. In large part, the sample size required depends on the nature of the analysis to be performed, the kind and number of comparisons that will be made, the number of variables that have to be examined simultaneously. For example, if the key analysis

consists of computing averages for experimentals and controls in a program and comparing the differences, then a sample under 100 might be adequate, provided that experimentals and controls can be assumed to be equivalent to each other on all other relevant variables because they were assigned to the experimental or control group by randomization.[5]

However, most often relevant variables have to be controlled statistically (as discussed in Chapter 3) because the groups differ by factors other then chance. If two groups are compared but statistical controls have to be introduced for as few as four variables (e.g., sex, education, occupation, urban-rural residence) and each of the four variables is only dichotomous (i.e., divides the sample in two only), a simple cross-tabulation of treated and nontreated groups controlled for the four dichotomous characteristics requires a table of 32 cells (2^5). Obviously 100 cases would be insufficient for anlaysis. In short, the evaluator must establish in advance the mode of analysis to be employed in order to determine the required sample size for efficient estimation of parameters and for statistical tests.

Finally, more technical considerations suggest that the required

---

[5]Randomization simply assures that the groups differ on the basis of chance alone, which can be tested statistically. "Equivalence," then, is a by-product of randomization, in the same way "representativeness" is a product of random sampling, in terms of probability processes and not single instances.

sample size is a function of the precision of the estimates one wishes to achieve, the variability or variance one expects to find in the population, and the statistical level of confidence one wishes to use. Although the elements of the formula are discussed in Section V below, we can give an example: To estimate mean (average) earnings following a training program, with an accuracy of at least $100 per year, using a 95% confidence interval, and assuming that the standard deviation of earnings in the population is $1,600, the required sample size N is approximately 985.

In general, to estimate a mean:

$$\sqrt{N} = \frac{(1.96)(\sigma)}{\text{precision}}$$

where σ is the population standard deviation for the variable whose mean one is interested in estimating, precision refers to the width of the interval one is willing to tolerate, and 1.96 reflects the confidence level (in this case the value of Z at the .05 level.)

II. INTERVIEWING

Interviewing can be thought of as a continuum of types, where the polar ends are as follows:

Unstructured Interviewing: Interviewing tends to resemble conversation style with no set questions, following a broad outline of topics. Each interview may cover a different set of topics within the broad outline.

Structured Interviewing: The interview follows a strict order of set questions with the respondent picking his/her response from among a series of alternatives offered by the interviewer.

Any given interviewing task, however, is usually somewhere in between the two polar types, containing some set questions and some discretionary questions to be asked by the interviewer when it appears appropriate. As one might suspect there are advantages and disadvantages to each type discussed below.[6]

A. Unstructured Interviewing

Unstructured interviewing requires a considerable amount of skill both with the language involved and with techniques that encourage respondents to talk frankly and at length. In addition, unstructured interviewing requires considerable knowledge on the part of the interviewer concerning the topic under consideration, since it is that knowledge which enables the interviewer to direct skillfully the course of the interview into useful and informative channels.

There are no particular personality features that seem to determine

---

[6] Standard references on interviewing techniques include: Raymond L. Gordon, Interviewing: Strategy, Techniques, and Tactics (Homewood, Ill.: Dorsey Press, 1969); Stephen A. Richardson et al., Interviewing: Its Forms and Functions (New York: Basic Books, Inc., 1965); Claire Selltiz et al., Research Methods in Social Relations (New York: Holt, Rinehart & Winston, 1976).

whether a given person will make a good unstructured interviewer. Clearly, a good interviewer has to have confidence in him or herself, be warm but not ingratiating, and to have the ability to listen attentively and with apparent interest to a respondent.

Given these characteristics, unstructured interviewing is an excellent tool in the early stages of a study where the investigator may need to know a topic from the point of view of target populations. The skills and knowledge needed to conduct good unstructured interviews on a wide scale are simply too scarce to make this a good technique for obtaining information on a widely defined population.

In addition, the data obtained through unstructured interviewing can vary somewhat from respondent to respondent depending on the interests and needs of the respondent and the particular approach taken by the interviewer. Hence, it is difficult to code and use such data in any systematic way: rather the main use of unstructured interviewing information is to provide illustrative materials and to give an investigator a sense of the topic as seen in more subtle ways through a respondent's eyes.

B. Structured Interviewing

A structured interview is only as good as (1) the questions written by the investigator and (2) the abilities of the interviewers hired to ask the questions and to faithfully record the answers given. Question writing is almost an art form, apparently requiring a good sense for language as it is spoken and for the ways in which target populations view the topic in question. Hence, unstructured interviewing is often used first by an investigator to obtain sufficient intimate knowledge of target persons' viewpoints and to tailor the structured questions to the latters' understanding of the topic in question. The art of writing questions has not been systematically explicated. Thus, the best way for a novice to proceed

is to look at questionnaires used by other successful investigators and to copy the style used. Wherever possible, standardized questions should be used.

Structured interviewing requires at a minimum the following skills: first, a good command of both the spoken and written language. Consequently, skillful interviewers have at least some secondary school education. Secondly, structured interviewers need to be impressed with the necessity of asking questions as printed in the schedule or questionnaire in order to provide the same stimulus to all respondents. Thirdly, interviewers need confidence in themselves and in the utility of the study. Finally, they need to be precise and thorough.

The training of interviewing staffs for structured interviewing ordinarily involves the following steps: first, hiring and recruiting interviewers with the characteristics listed above. Secondly, intensive short term training in the purposes of the survey, the intent of questions and of the necessity for asking questions as printed and carefully recording responses. A useful training device is setting up role playing simulations in which prospective interviewers interview each other before a group of recruits, with the instructor commenting on proper procedures and mistakes. Third, it is useful to send interviewers out with a "practice" interview and to thoroughly go over their returned questionnaires noting errors made. Finally, while interviewing is in progress, it is necessary to set up a quality control operation in which the schedules returned by interviewers are checked over carefully by a supervisor and errors and inconsistencies are brought to the attention of the interviewers. Part of a quality control procedure should be interview validation, an independent check with a sub-sample of respondents to ascertain whether interviewers actually conducted

interviews with the right respondent.

Structured interviews lend themselves easily to systematic treatment. Results can be easily transformed into quantitative forms and stored either as punched cards or magnetic computer tapes. Clearly for large scale evaluation research structured interviews are most cost effective.

III. MEASUREMENT

Any research requires that theoretical or conceptual ideas be translated into research operations which provide empirical evidence bearing on those ideas. For example, many development programs are aimed to "improve the standard of living" of the people. Obviously, "standard of living" is not a directly observable characteristic in the world, and thus an evaluation of such development programs requires first that the abstract notions be translated into procedures which will allow measurement to take place. We might measure standard of living, for example, in terms of economic productivity for the community, personal income, education, agricultural productivity, type of housing, and so on. Measurement involves the transition between concepts and operations, and the outcome is data which are manipulated statistically to test hypotheses.

A. Level of Measurement[7]

In general, what we do when we "measure" something is to assign numbers in some fashion such that the numbers represent the properties of the phenomenon of interest: we measure income in terms of the local currency,

---

[7]Discussions of level of measurement appear in all methods and statistics texts. The classic reference is S.S. Stevens (ed.) Handbook of Experimental Psychology. New York: John Wiley & Sons, 1951.

education in terms of years of formal schooling attained, agricultural yield in terms of liters of given crops, nutritional intake in terms of calories or grams of protein, literacy in terms of scores on a reading test, and so on. Each of these different ways of operationalizing or measuring phenomena results in numbers which themselves have different properties and can be subjected to different mathematical operations. This is known as the level of measurement and it is essential that researchers be aware of the level of measurement attained in the research operations because the level of measurement determines almost exclusively the nature of the statistical analysis which is appropriate.

Nominal level of measurement involves the simplest of all scientific operations: the categorization of units into classes or categories according to a certain characteristic. Nominal measurement means assigning different names to units. For subsequent analysis, the names are replaced by (arbitrary) numbers: for example, for measuring sex, we can classify all men = 1 and all women = 0. The assigned numbers could have been 1 and 2, or 1 and 5, or 1 and 500. Nothing other than a difference is implied: men and women differ, but one number does not represent more or less than the other. Similarly, villages can be classified according to a numbering system representing region, but the numbers only stand for different names, in a shorthand notation.

Ordinal measurement implies that the categories of a classification system can be ordered or ranked according to some property. Some elements possess more or less than others in terms of some characteristic, but it is not possible to say exactly how much more or less they possess. Thus, individuals vary in terms of socio-economic status; they can be rank-ordered along a continuum such that some are upper, middle and lower status.

At this level of measurement we still do not know the magnitude of differences between assigned numbers. For convenience, we might assign the numbers 1, 2 and 3 to low, medium and high, respectively. But the numbers could also have been 1, 20 and 100. Most attitudinal measurement is ordinal in that people can be classified in terms of, say, level of self-esteem, degree of alienation, level of agreement with certain political issues, or degree of satisfaction with a treatment.

Interval and ratio measurement imply that the units in terms of which the measurement is expressed are constant. Not only is it possible to rank persons or objects with respect to the degree to which they possess a characteristic, but it is now legitimate to indicate the distances between them: for example, years of education, dollars of income, number of persons in a family. The units which separate the scale of the measure are equal at all points on the scale. With interval measures, arithmetic operations of adding and subtracting are now meaningful and legitimate. With ratio measures it is possible to locate an absolute or nonarbitrary zero point, and thus division (taking ratios) and multiplication are also meaningful operations. In practice, researchers usually treat interval measures as if they were ratio, since once a constant interval scale can be established, it is usually possible to assign a nonarbitrary origin.

The different levels of measurement are obviously cumulative in that interval and ratio measurement also implies rank-ordering as well as classification, and rank-ordering implies classification. For research purposes, whenever there is a choice between alternative ways of measuring the same phenomenon, one should choose the highest level of measurement possible because more powerful and sophisticated analytical techniques can be used. For example, in collecting data about age, it is best to

ascertain years of age directly (an interval-ratio measure) rather than classifying respondents into young, middle-aged, or old (an ordinal measure). The first can always be reduced to the second, but not vice versa. Fortunately, in evaluation research, many of the critical outcome variables of an intervention or program can be measured as counts or by enumeration: number of births, income, number of persons vaccinated, agricultural production in liters of crops or hectares under cultivation, and so on.

B. Measurement Error[8]

Measurement error refers to the lack of fit or slippage between the measured value that the researcher has at hand and the "true" value of the observation on the phenomenon in question. The general "theory" of measurement error can be expressed in terms of a simple formula:

$$M = T + e$$

where M is the measured value, T is the true value and e is the measurement error, that is, all the stochastic errors that cause the measured value to differ from the true value. For example, score on an achievement test (M) is a function of ability (T) and measurement error. The obvious problem is that researchers never know what the true value is: T is what we want to know, M is all we have and what we are forced to use to estimate T, and the error is a great unknown, about which many assumptions have to be made. What the researcher can do, is to take steps to guard against committing some well-known avoidable mistakes by taking steps which will reduce the numbers and kinds of errors in measurement.

---

[8] A useful reference is George W. Bohrnstedt, "Reliability and Validity Assessment in Attitude Measurement." In Gene F. Summers (ed.), Attitude Measurement. Chicago, Illinois: Rand McNally, 1970.

Errors can occur at any stage of the research: sampling, measurement, data coding and processing, and so on. The fundamental distinction is between random errors and bias. Random errors tend to cancel each other out in terms of direction and magnitude (technically, the expected value of the error term in the formula is zero). Nonrandom errors are referred to as systematic error or bias. Well known examples are the tendency of people to under-report their age and in some cases income, the tendency to over-report level of education, and all sorts of response sets which lead to biased measurement. When we have a biased measurement, we have invalid measurement (see the discussion of validity in Chapter 3). Put differently, when we expect that the measured value equals the true value, the measure is valid--that is, it measures what it purports to measure.

A large number of measurement biases have been subsumed under the term response effects, which are systematic errors in measurement usually discussed in the context of surveys. These have been decomposed into three major components: interviewer effects, respondent effects and task effects. Interviewer effects are all the errors introduced by the interviewer and as a result of his or her characteristics. A previous section on interviewer training has noted some of the procedures which enhance interviewing quality. In general, the researcher should watch out for any characteristics of the interviewer which might affect responses, if the study is related to those characteristics. It was believed in the past that interviewer characteristics such as sex, race and age would bias responses, but at least in the United States, such concerns have not been substantiated by methodological research (see Sudman and Bradburn, 1974). The extent and nature of the problem in evaluation research in less developed countries is unknown. We simply urge the researcher to be conscious and cautious in avoiding the introduction of interviewer effects in surveys.

Respondent effects are all those systematic errors in measurement produced by non-relevant respondent states, traits or characteristics. Many are response sets, or tendencies to answer questions in certain ways irrespective of their content. One of the best known forms of response set is acquiescence, the tendency to agree to an item regardless of content. It is particularly troublesome in attitude measurement which is based on "agree-disagree" items. Care must be taken in item wording such that a positive or "agree" response denotes the presence of a trait or attitude half the time and absence of the trait the other half. Wherever possible, well-developed scales should be used which have been constructed to guard against the problem.[9]

A second form of response set is called social desirability bias, or the tendency to give what the respondent feels is the desirable or appropriate answer irrespective of his/her own feelings. This may be a problem any time there are certain strong norms in the society which prevail: for example, the UNESCO Experimental World Literacy Program evaluation cites an example where in one country it was simply impolite to answer "no" to any question. There is a scale (in English) which has been developed for the express purpose of measuring the extent of social desirability bias in respondents, and which has been used as a correction factor. In sum, each cultural tradition may produce certain respondent effects which the researcher should be alerted to. Wherever possible, when measuring some characteristic or attitude for which a standardized measure or scale is available (many have been translated into several languages), the researcher should use it since most likely some of the methodological properties will

---

[9]There are several compilations of available measures which should be consulted before deciding to construct new measures, for example, John P. Robinson and Phillip R. Shaver, Measures of Social Psychological Attitudes, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, 1973.

be known and such problems will be avoided.

Finally, task effects refer to the errors introduced by the nature of the task which the respondent is asked to perform. It includes the method of the survey (whether it is face-to-face, telephone, mailed), the length, degree of structure, and also the nature of the information elicited. Two common sources of error (which may or may not be systematic) are due to attempts to retrieve retrospective information. Simply put, people tend to forget about events and experiences, occurring in the past, or they may remember events but report the data inaccurately (for instance when it occurred). The first is sometimes known as memory decay, the second as telescoping. For these reasons, we have urged throughout the discussion of research design in Chapter 3, that adequate evaluation research which requires data at more than one time point obtain those data without resort to recall by respondents. If longitudinal studies are not feasible, then documentary evidence may be preferable to asking respondents to recall past events. This, of course, depends on the nature of the variables under study: obviously major or critical events in a person's life will be remembered and reported more accurately.

One final source of systematic error or bias in measurement is actually a non-response bias, or the general problem of missing data. Data can be missing due to a variety of reasons: respondents refuse to be interviewed, respondents don't answer all questions, participants drop out of programs, controls who are part of an evaluation but are not participating or deriving benefits from a program may drop out, and so on. Whatever the source and magnitude, missing data are very problematic in the data analysis and may lead to biased results. Imagine an intervention program with an experimental design which imparts job training to the

unemployed. Partly through the program, some of those who have the least skills and most difficulty finding jobs simply drop out of the program. An evaluation on the basis of those who remain in the program and complete the training will overestimate the impact (positive benefits) of the program. In different contexts, differential dropping out of the program (known as attrition) can occur among those who benefit the most, and thus results would tend to underestimate impact.

The only entirely satisfactory solution to the problem of missing data is to take all possible steps to avoid it, for instance, by monitoring participation or by offering incentives to those most likely to drop out (such as controls). The problem of data missing intermittently through the file can be dealt with in many different ways, none of which is generally superior to any other. Most computer programs or software packages (e.g., SPSS) designed for analyzing social science data provide some discussion of the problem and offer alternative ways of handling it.

C. Validity and Reliability of Measures and Multiple Item Scaling

Whenever a researcher constructs a measurement instrument, the two criteria applied to assessing methodological quality are validity and reliability. These properties were discussed briefly in Chapter 3, and they ask, essentially: does my instrument measure what it purports to measure (i.e., is it valid) and does my instrument constitute a reliable--i.e., a consistent--measure. There is a very large methodological literature on how to establish the validity of measures and how to estimate the reliability.[10]

---

[10] A very useful summary article in Bohrenstedt's (1970) article, but practically any reference on measurement and scaling would be useful: for example, any psychometric handbook, the Handbook of Social Psychology, most advanced methodology books such as Blalock and Blalock (1968), and a forthcoming Handbook of Survey Research by P. Rossi and J. Wright (Academic Press).

The important thing to know is that there are well known formulas to estimate reliability coefficients as well as guidelines on when and how to use adjustments for unreliability in statistical analyses (the best summary is in Bohrenstedt, 1970). The logic is that reliability can be measured by administering the same instrument at two different points, and assuming that persons have not changed, the same instrument should produce consistent measurement. This is known as test-retest reliability, and it can be computed by a correlation coefficient (see Section IV below). In evaluation contexts, test-retest measurement may have to be done prior to the program, since the whole point of intervention programs is to cause change. Thus, a more likely situation is for evaluation researchers to measure a concept at a single point in time by means of a number of items. Reliability can then be estimated in terms of how consistent the items are: how well correlated are they with each other, do they seem to "hang together" and measure the same thing? An early method involved dividing the items in half and computing the correlation between halves (split-half reliability, usually by means of the Spearman-Brown (1910) formula). The most straightforward is to compute an average inter-item correlation among the items in the instrument, but formulas such as the Kuder-Richardson (1937) and Cronbach's alpha (1951) are also available.

For example, an evaluation of the impact of a housing program on quality of housing would have to measure "quality of housing" if no instruments have already been developed or are inappropriate for a particular context. Suppose the researcher decides to construct 15 items which all together will be combined into a quality of housing indicator. The relevant questions, then, are whether or not the indicator actually measures quality

of housing (is it valid?) and whether it is reliable. An important consideration is how the 15 items will be combined into a single indicator or summary measure, or how the items will be scaled.

Scaling multiple item indicators could be done at the simplest level by adding items or points all the way up to complex scaling procedures involving factor analysis. Again, any standard methodology text will be a useful source for scaling techniques.[11]

---

[11] See A.L. Edwards, Techniques of Attitude Scale Construction, (New York: Appleton-Century-Crofts, 1957) on scaling and R.J. Rummel, Applied Factor Analysis (Evanston, Ill.: Northwestern University Press, 1970) on factor analysis.

IV. DESCRIPTIVE STATISTICS

Data from evaluation research have to be organized and summarized intelligibly and examined systematically to test hypotheses about program effects. Descriptive statistics are used to summarize and describe the characteristics of the program, targets and effects. Inferential statistics provide methods for testing hypotheses about the significance of the observed results and for estimating population values on the basis of sample data. This section gives a very brief introduction to some of the more useful descriptive statistics for evaluation, the next discusses inferential methods.[12]

A. Describing a Single Variable

1 PERCENTS, RATIOS AND RATES

One way to summarize a variable is to report its percentage distribution, converting the frequency distribution into percentages. The frequency distribution (also known as "marginal distribution" is simply the number of cases that fall into each of the categories of the variable.)

Table A-1. Frequency and Percentage Distributions of Employment after Training.

Employment after Training	Frequency	Percentage
Farming	150	60.0 =(150 ÷ 250)x100
Industry	65	26.0
Teaching	11	4.4
Unemployed	24	9.6
Total	250	100.0

The advantage of a percentage over a frequency is that comparisons can be made across distributions based on different numbers of total cases (usually

[12] There are literally dozens of statistics texts which can be consulted. At the introductory level one which is widely available is Hubert M. Blalock, Social Statistics. New York: McGraw-Hill, 1972.

symbolized by the letter N). For example, if we wanted to compare the proportions in Table A-1 with another program that had 300 graduates with 165 (55%) in agriculture, the frequencies 165 and 150 are not directly comparable but 55% and 60% are.

To compute a <u>ratio</u> the frequency in one category is divided by the frequency in another category: e.g., from Table A-1, the ratio of those in farming to those who are unemployed is 6.25 (=150 ÷ 24). A more common example is the sex ratio: if 130 of the trainees are males and 120 females, the sex ratio is 130 ÷ 120=1.08 (sometimes expressed as 108:100, i.e., there are 108 males for every 100 females.)

A <u>rate</u> is like a percentage in that it divides the frequency of observations in a given category by the total number of observations, but then that proportion is expressed as a multiple of 10 (usually larger than 100.) For example, birth rate is usually expressed as the number of births in a population per 1,000: if there are 9,000 births during a year in a population of 200,000, the (crude) birth rate = 9,000 ÷ 200,000 x 1,000 = 45 births per 1,000. Similar rate computations can be carried out for age-specific female cohorts, for example number of births for the population of women ages 15-50. For events which occur very infrequently, such as certain crime rates, rates are expressed in multiples of 10,000 or 100,000. Examples of other rates in evaluation are: yield or revenue divided by hectares cultivated and area cultivated divided by population size. (See Chapter 2 for a discussion of prevalence and incidence.)

A <u>rate of change</u> or <u>rate of gain</u> is useful in comparing the distribution of a variable through time, say before (time 1) and after (time 2) a program:

$$\text{Rate of change} = \frac{\text{Value at time 2} - \text{Value at time 1}}{\text{Value at time 1}}$$

The result can be multiplied by 100 to remove the decimal and to reflect percentage change. For example if the population equals 200,000 in one year and 250,000 5 years later, the rate of change in population is (250,000 - 200,000) ÷ 200,000 =.25 or 25%.

2 CENTRAL TENDENCY: MODE, MEDIAN, MEAN

Another way to summarize the distribution of a variable is with one of three measures of central tendency, which convey information about the "average" value of the variable. The mode is the most frequent value in the distribution. For the example, the mode (modal employment) of the distribution shown in Table A-1 is farming. The mode can be reported for any level of measurement, and it requires no computation. It can be used for grouped or ungrouped data.

A second kind of average is the median, or the value which divides the distribution in half--that is, half the cases fall below the median and the other half are above it. The median therefore, reflects the middle position of the distribution, where the values have been ordered. (It is a meaningless statistic for nominal data.) For example, for a hypothetical distribution of size of training centers:

25 30 32 37 38 42 45 50 56

the median is 38 because there are four smaller and four larger centers. If there had been a tenth center of size 70, the median of that distribution would be a value in the middle between 38 and 42: (38 + 42) ÷ 2 = 40. The median is an important measure of central tendency when the distribution is skewed, that is, when there are extreme values in one of its ends. For example, if the first center only was of size 5 instead of 25, the median would still be 38, and likewise if the last one were 150 instead of 56. Thus, the median is a stable measure of central tendency and is not affected

by extreme scores. This is why one often finds the median reported for income data, which tend to be skewed.

The most commonly known measure of central tendency is the arithmetic average or <u>mean</u> ($\overline{X}$), computed by adding up all the values and dividing by the number of cases. For example, for the distribution of center size,

$$\overline{X} = \frac{25 + 30 + 32 + \ldots + 50 + 56}{9} = 39.4$$

A shorthand notation can be introduced to simplify the remainder of this discussion:

X_i = the "raw" or observed values of a variable

N = total number of cases or observations

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N} = \text{mean}$$

Or even simpler,

$$\boxed{\text{mean} = \overline{X} = \sum_{i=1}^{N} \frac{X_i}{N}}$$

The symbol Σ (the Greek capital letter Sigma) simply indicates a sum. It directs us to add all the values of X_i appearing to the right of the symbol. The notation can be made even simpler by omitting the subscript i:

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

For quantitative (numeric) data the mean is the preferred measure of central tendency because it has important arithmetic and other properties used in inferential statistics. It is a good summary measure because its computation utilizes every value of the distribution. However, extreme values do affect the mean (for example, in the above distribution of center size, if the largest center were size 100 and not 56, the mean would be 44.3). Therefore, for skewed distributions, it is best to report both the mean and the median to avoid misleading interpretations.

3 DISPERSION

The average values of a distribution convey only one kind of information about the variable. A second property is the dispersion or variation in the values of the distribution -- that is, how spread out the values are or how close together. One very simple measure of dispersion is the range, or the difference between the highest and lowest values (in the above example, range = 56 - 25 = 31). Sometimes the highest and lowest values are simply given -- for example, program participants range in age from 12 to 16 years. But suppose the largest center were size 100 instead of 56 in the previous example, and the rest of the distribution was unaltered, the range would now be 100 - 25 = 75. Since the range only takes into account the minimum and maximum values of a distribution, it tells us nothing about the distribution of values in the middle.

Therefore, other distance-based measures of dispersion have been developed which are more informative about the middle of a distribution. One such measure is the interquartile range, which is simply the range of values in the middle half (50%) of the distribution. Such a measure is more sensitive to the way values are concentrated around the middle, but it still amounts to the difference between only two values of the distribution.

A more important (in the sense of useful properties and applications in other topics in statistics) characterization of dispersion is the extent to which each of the values of a distribution differs from the mean. The variance of a distribution is defined as the average squared deviation from the mean:

$$\boxed{s^2 = \text{variance} = \frac{\Sigma(X - \overline{X})^2}{N}}$$

Table A-2 shows a distribution of seven cases and the computation of its mean and variance:

Table A-2. Computation of the Mean and Variance for a Simple Distribution.

X	$(X - \bar{X})$	$(X - \bar{X})^2$	Computation:
2	-4	16	$\bar{X} = \frac{42}{7} = 6$
2	-4	16	
4	-2	4	
5	-1	1	
6	0	0	$s^2 = \frac{122}{7} = 17.4$
8	2	4	
15	9	81	
$\Sigma X = 42$	0	122	

The second column shows the deviations from the mean for each of the values of the distribution. Note that the sum of all the deviations from the mean is zero (it always is, for any distribution),[13] which suggests why the next step of squaring the deviations is necessary.

The formula for variance given above is the definitional formula. In practice one does not compute the variance with that formula because simpler computational procedures are possible which are also less likely to introduce rounding error.[14] A computational formula for the variance is:

$$s^2 = \frac{\Sigma X^2 - (\Sigma X)^2 / N}{N}$$

[13] By definition, the mean is that point of the distribution which represents its center of gravity (often the point is illustrated by analogy to a teeter-totter, where the mean is the place where the fulcrum is located).

[14] For example, if (as is almost always the case) the mean of a distribution is not a whole number, each of the deviations will be a decimal number, which when squared will also be a decimal number. So the computations get more tedious and more subject to rounding errors.

Table A-3 shows the simplified computation of the variance for the same distribution in Table A-2. Note that only one additional column of squared terms has to be computed, and the calculation does not involve the mean $\bar{X}$ directly:

Table A-3. Calculating the Variance of a Distribution with the Computational Formula.

X	X^2
2	4
2	4
4	16
5	25
6	36
8	64
15	225
42	374

$$s^2 = \frac{\Sigma X^2 - (\Sigma X)^2 / N}{N}$$

$$s^2 = \frac{374 - (42)^2/7}{7}$$

$$= \frac{374 - 1764/7}{7}$$

$$= \frac{374 - 252}{7}$$

$$= \frac{122}{7}$$

$$= 17.4$$

Note that the first summation term in the formula refers to the sum of the squared values. The second summation term is obtained by first summing up all the "raw" values of the distribution, and then squaring that total. The rest is arithmetic, but the result is the same as in Table A-2 when the definitional formula was used.

While variance is an important statistical concept which is used in different ways, it does not have a direct interpretation for descriptive purposes. For one thing, it is given in units which are squared -- for example, if the above distribution were of the variable "family size," and hence in units of individuals, the variance would be 17.4 (individuals)^2. Another measure of dispersion is therefore the <u>standard deviation</u>, which

is the square root of the variance:

$$\boxed{\text{standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}}$$

Its computational formula is as before:

$$\boxed{s = \sqrt{\frac{\Sigma X^2 - (\Sigma X)^2/N}{N}}}$$

The standard deviation is expressed in the original units of measurement of the distribution: the standard deviation for Table A-3 is $\sqrt{17.4} = 4.2$. Thus, the mean is 6 and the standard deviation 4.2 persons.[15]

4 STANDARDIZED SCORES OR Z-SCORES

The mean and standard deviation can be used to describe individual values or scores of a distribution by means of a formula which is called the Z-transformation:

$$\boxed{Z = \frac{X - \overline{X}}{s}}$$

A Z-score expresses a raw or observed value X in terms of its distance from the mean $(X - \overline{X})$, or the numerator of the formula. Dividing by s converts that distance to "standard deviation units." For example, we might want to know how the observed value of 15 in Table A-3 compares with the rest of the distribution, so we would compute a Z-score which corresponds to the value 15:

$$Z = \frac{15 - 6}{4.2} = 2.14$$

We would say that the value 15 is located 2.14 standard deviations above the mean of the distribution.

Standardizing scores also allows comparing raw scores from two distributions. Suppose we want to compare values (say, test scores) from two

[15] Many statistics texts include tables of square roots and squares of numbers, which can be helpful for the calculations by hand (without computers or calculators).

different distributions (say, before and after an educational program. One student scores 90 before the program and 80 after the program. Without converting those scores to z-scores, comparisons of the two raw values would be misleading (one conclusion might even be that the progam is detrimental). Instead, the appropriate comparison is in terms of standardized scores. Suppose the pre-program distribution of scores had a mean of 75 and standard deviation of 5; post-program scores have a distribution with a mean of 60 and standard deviation of 4. (Note that the second test was harder, and there was less variation around the mean than for the first test. In fact, they might even have been different tests). We can now compute two Z-scores.

$$Z_{before} = \frac{90 - 75}{5} = 3$$

$$Z_{after} = \frac{80 - 60}{4} = 5$$

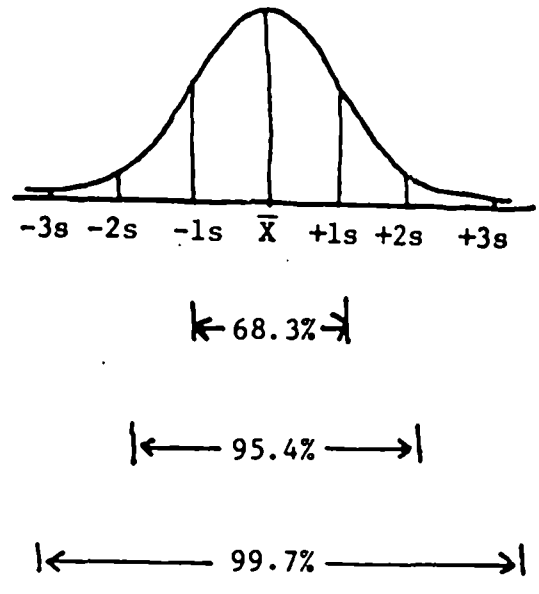
Thus, 80 is a better score than 90, relative to the average score and variation in each of the distributions: 90 is equivalent to 3 standard deviation units higher than the mean, 90 is equivalent to 5 standard deviation units higher than the mean. Note that it is possible to obtain negative Z-scores, which would then indicate that a particular score is below the mean.

5 THE NORMAL DISTRIBUTION

The normal distribution is a mathematical or theoretical distribution of great importance in inferential statistics. We introduce it here to show another useful application of the standard deviation. The normal distribution is defined by a mathematical equation and is based on a theoretically infinite number of cases -- thus it can only be approximated with real data. It is a symmetrical bell-shaped curve. It actually consists

of a whole family of distributions where the mean and standard deviation can both vary from one curve to the next. An important property is that there is always a constant area or proportion of cases between the mean and a given number of standard deviation units:

Figure A-1: Graph of a Normal Distribution and Areas under the Curve.



Thus, in a normal distribution, over two-thirds of the cases (68%) can be found within one standard deviation unit in either direction from the mean. Over 95% of the cases fall within two standard deviation units from the mean, and practically all the cases are within three standard deviation units. What the figure shows is that for a normal distribution, we can determine the proportion of cases that fall between any two scores, between any score and the mean, or above any given score. Statistics texts always include a table of "areas under the normal curve," which is given for a normal curve that has been standardized such that it has a mean of 0 and standard deviation 1 (This allows that only one table be shown which

is applicable for all normal distribution problems when scores are transformed by the Z-transformation.)

A very important word of caution: using the Z-score transformation will not change the shape of a distribution, and will not make a distribution of observed values normal. The reason the normal distribution can be used is that many events do occur in ways that their distribution approximate the normal theoretical distribution. Furthermore, there are guidelines provided by mathematical statistics which tell us in what particular instances we can use the normal approximation (as we'll see in the inferential section, having to do among other things with sample size and sampling method.)

B. Examining the Relationship Between Two or More Variables

1 CROSSTABULATION AND HOW TO PERCENTAGE A TABLE

In order to analyze the effect of a program, two or more variables have to be examined simultaneously. Suppose we have a fertility control program and have gathered data on completed family size (number of children) for an appropriate sample of women who participated and those who did not. A first step is to construct a crosstabulation showing number of children by program participation:

---

[16] An excellent source for the logic and procedures of simple and multivariate crosstabular analysis is James A. Davis, Elementary Survey Analysis, Englewood Cliffs, New Jersey: Prentice-Hall, 1971. In addition, the book has useful and authoritative advice on many practical aspects of research design and analysis.

Table A-4: Frequency Table of Number of Children by Program Participation.

Number of Children	Participated	Did Not Participate
0-3	300	100
4-5	240	275
6 or more	60	125
Total	600	500

Since the cell entries for participants and nonparticipants are not based on the same number of cases, the table has to be percentaged to permit comparisons of cell entries across the two different groups.

Table A-5: Percentage Table of Number of Children by Program Participation.

Number of Children	Participated	Did Not Participate
0-3	50 %	20 %
4-5	40 %	55 %
6 or more	10 %	25 %
Total	100 %	100 %

There is a rule for percentaging a table: always compute the percentages within categories of the independent variable. (In the example, the independent variable is program participation.) Another way to phrase the rule in more mechanical terms is: set up the table so the independent variable is in the columns (the categories of the independent variable are across the top) and the dependent variable or effect is down the rows. Then always percentage down the columns, to make comparisons across the rows.

Table A-5 shows that those women who participated in the program do tend to have fewer children than those who did not participate.[17]

The problem with the results is that family size may be affected by other variables besides program participation, and if the sample does not represent a sample of randomly assigned women to either of the two groups, we must use those other variables as statistical controls (see Chapter 3). For example, it may be that level of education changes the nature of the relationship shown in the preceding table. Adding education as a control variable produces the following:

Table A-6: Number of Children by Program Participation and Education (Cell Entries are Frequencies).

	Primary Level Education		Post-Primary Level Education	
Number of Children	Participant	Nonparticipant	Participant	Nonparticipant
0-3	90	42	210	58
4-5	128	165	112	110
6 or more	32	93	28	32
Total	250	300	350	200

Again the table must be percentaged before it can be interpreted:

[17] In a subsequent section we show how to compute a statistic known as chi-square. This statistic allows an inferential test of the hypothesis that the observed results might have occurred by chance (e.g., due to sampling variability).

Table A-7: Number of Children by Program Participation and Education (Cell Entries are Percentages).

	Primary Level Education		Post-Primary Level Education	
Number of Children	Participant	Nonparticipant	Participant	Nonparticipant
0-3	36%	14%	60%	29%
4-5	51%	55%	32%	55%
6 or more	13%	31%	8%	16%
Total	100%	100%	100%	100%

The effects of the fertility program are quite different for women of different educational levels: the effects are much more pronounced for women with higher levels of education, and less pronounced for women with only a primary education. For instance, 60% of the more educated participants have a small number of children, as compared with 36% of the less educated women. More precisely, the difference between participants and nonparticipants in the percentage who have small families is 22% (36 - 14) for the less educated and 31% (60 - 29) for the more educated.

Two points can be noted: it is clear that one could introduce additional variables as control variables, but only up to the point where there are sufficient cases in the cells. Furthermore, the introduction of several control variables simultaneously in a table can result in a very cumbersome and clumsy analysis. Hence, other statistics can be employed to summarize more succinctly the effects of several variables.

For example, with the ungrouped or raw data on number of children and level of education for participants and nonparticipants, one might

start out by reporting the mean or average number of children for participants and for nonparticipants: 3.5 for participants and 4.6 for nonparticipants, a difference of 1.1 children for the aggregate. Suppose that the means when controlling for educational level are:

Table A-8: Mean Number of Children for Program Participants and Nonparticipants by Educational Level (Number of Cases Shown in Parentheses).

Education	Participants	Nonparticipants	Difference
Primary	3.9 (250)	4.9 (300)	1.0
Post-Primary	3.0 (350)	4.3 (200)	1.3
Total	3.5 (600)	4.6 (500)	1.1

The computation of means allows a much more parsimonious presentation of the data. However, the detail present in the crosstabulations is obviously lost by virtue of the summarization which occurs in computing means. Note finally that each of the means reported in Table A-8 is calculated separately from the raw data with the formula given earlier. Thus, 3.5, the mean for all participants, is computed by adding up all the numbers of children for the 600 participants, then dividing that number by 600.

2 MEASURES OF ASSOCIATION

A large number of measures of association are available which summarize in a single statistic the extent to which two variables vary together or are related to each other. Which one to use depends most importantly on the level of measurement of the variables in question, although there are other considerations such as the sample size and the nature of the distributions. The major distinction about levels of measurement is between the nominal level and the rest, since it is misleading to compute a statistic

such as a correlation coefficient on nominal data (the equivalent is the error of computing an arithmetic average or mean for three categories of a nominal variable such as religion, ethnicity, or region).

If the data are ordinal (categories are just ranks) applied researchers have the choice of computing measures of association developed explicitly for ordinal data, or they may decide to treat the data as if they were interval and use the procedures described in the next section on regression-correlation. The advantage of the latter approach is that those techniques are considered to be more powerful, result in some very useful summary statistics and are capable of handling several variables simultaneously.

Most of the measures of association for nominal and ordinal data have been standardized so that they result in an index which ranges from 0 to 1, and in some cases from -1 to +1 where the sign indicates the direction of association (i.e., a negative sign means higher values of one value are associated with lower values of another, and vice-versa). When the statistic is 0 (or close to 0) we say the variables are uncorrelated; the closer the number is to 1.00, the greater the magnitude of correlation or association.

Since the computation of measures of association can be quite tedious, and each one is based on somewhat different logic, we will not show how to compute such measures. All canned computer programs for social research contain several options. For example, SPSS (Statistical Package for the Social Sciences, see section VII below) computes several chi-square based measures, and others such as lambda, tau and gamma.

3 REGRESSION AND CORRELATION

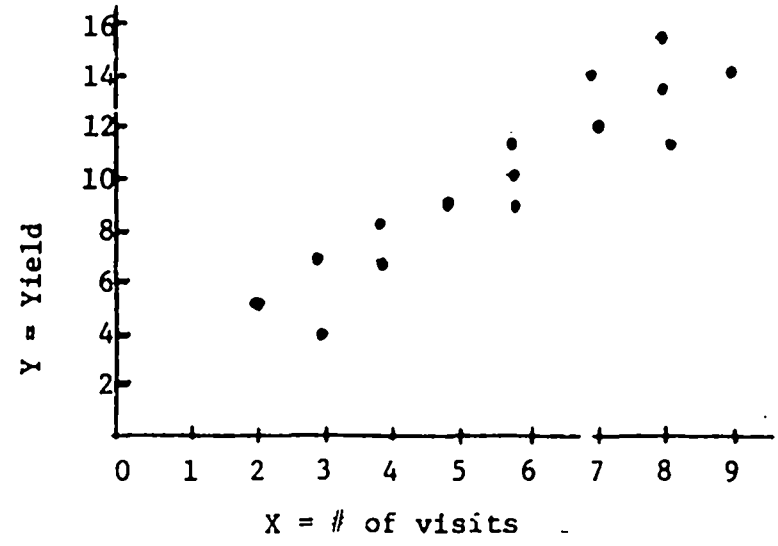
Suppose we want to evaluate the effect of an agricultural training program (which consists of a series of visits by extension staff) on production or yield. The amount of treatment received is simply the number of visits and yield is measured in litres. One way to begin analyzing the data would be to group the data and present them in a cross-tabulation. A more useful analysis entails describing the relationship between the two variables by means of a linear, mathematical equation known as a regression equation. Regression analysis produces estimates of the effect of one variable on another (or of several variables on a dependent variable) as well as a measure of the extent to which the variables are correlated with one another.

In the two-variable case, regression analysis can be easily visualized by first constructing a scattergram or scatterplot, that is, a graph of the data. Let Y represent the dependent variable (yield) and X the treatment (number of visits). Hypothetical data for 15 farms are shown in Table A-9 and graphed in Figure A-2:

Table A-9: Raw Data for Bivariate Regression Example (Effect of Visits on Yield).

Farm	X=# of Visits	Y=Yield (100 liters)
1	2	5
2	3	4
3	3	7
4	4	7
5	4	8
6	5	9
7	6	9
8	6	10
9	6	11
10	7	12
11	7	14
12	8	11
13	8	13
14	8	15
15	9	14

Figure A-2: Scattergram of Data in Table A-9.



The relationship between number of visits and yield can be summarized by a formula or equation describing a straight line fitted through the data points. In order to describe any straight line only two pieces of information are needed: (1) its tilt or slope and (2) its pivot point or the point at which it crosses the X axis, also known as the intercept. Thus, an equation for a straight line takes the form $Y = a + bX$, where a is the intercept and b is the slope.

But the observed data points do not all lie on a straight line; rather, we find different values of yield for any single number of visits (for example, farm #2 has 3 visits and 400 liters, farm #3 has 3 visits and 700 liters). Therefore, fitting a single straight line to the observed data always involves some residual error. No matter where the line is put, there will be points which do not fall on it. More precisely, then, the regression model is:

$$Y = a + bX + e$$

where a is the intercept, b is the slope and e is a stochastic error term (see Chapter 3). The error may be due to sampling, measurement or other factors which produce the observed variability in the data, including variables that should be in the model but are not (for example, rainfall and use of fertilizer).[18] Actually, e is defined as the distance from the observed point to the regression line.

It is easy to see from the scattergram in Figure A-2 that there are many different lines which can be drawn through the data points. Different statistical methods have been developed to solve the problem of how to find the "best-fitting" line through a data set. The most often used is known

---

[18] For inferential tests, certain assumptions have to be made about e, the error term.

is ordinary least squares regression, and it is based on the principle of minimizing the sum of squared errors. This is equivalent to fitting the line that minimizes the square of the distance from each point to the regression line, that is, error as defined above.[19] The least squares solution produces formulas for calculating the intercept a and the slope b from the data:

$$b = \frac{\frac{1}{N}\Sigma(X-\overline{X})(Y-\overline{Y})}{\frac{1}{N}\Sigma(X-\overline{X})^2}$$

$$a = \overline{Y} - b\overline{X}$$

The numerator of b is called the <u>covariance of X and Y</u> (covariance X,Y), the denominator is the <u>variance</u> of X, as we saw above.

Let us now return to the previous example and compute a and b. As in the case of computing variance, there is a simpler computational formula for the regression coefficient b:

Example A-1: Computation of Regression Coefficient and Intercept for Table A-9.

Computational formula for $b = \frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{N(\Sigma X^2) - (\Sigma X)^2}$; $a = \overline{Y} - b\overline{X}$

For the data in Table A-9, all the required terms are:

$\Sigma X = 86$ $\quad\quad$ $\overline{X} = 5.73$

$\Sigma Y = 149$ $\quad\quad$ $\overline{Y} = 9.93$

$\Sigma X^2 = 558$

$\Sigma Y^2 = 1637$

$\Sigma XY = 936$

$N = 15$

$$b = \frac{15(936) - (86)(148)}{15(558) - (86)^2} = 1.35$$

$$a = 9.93 - (1.35)(5.73) = 2.20$$

---

[19] The notion of sum of squared distances is analogous to the variance which calculates the sum of squared deviations from the mean.

We now have an estimated regression line which describes the data. We can summarize our results as follows:

$$\hat{Y} = 2.20 + 1.35X$$

where $\hat{Y}$ is the predicted value of Y (yield), the intercept is 2.20 (interpreted as the predicted value of yield when X = 0, that is, when there are no visits)[20] and the slope is 1.35. Since the measure of yield was in hundreds of liters (we dropped the 0's for computational ease) we should now convert our results and interpret them to mean (1) the predicted yield without a program (0 visits) is 220 liters, and (2) for every additional visit paid by an extension representative, the increase in yield amounts to 135 liters. We now have a precise estimate of the effect of the program on yield (the regression coefficient b) which has direct policy interpretation.

More generally, <u>the regression coefficient is interpreted as the change in Y for a unit change in X</u>. The procedure is referred to as the regression of Y on X. If the coefficient is positive, a unit change in X produces an increase in Y (that is, the line slopes upward). If the coefficient is negative, a unit change in X produces a decrease in Y (the line slopes down--has a negative slope).

We are now in a position to ask: How well does the calculated regression line fit the observed data? In other words, what is the <u>correlation</u> of X and Y? We noted that summarizing data will lead to some error. For example, the predicted value of yield ($\hat{Y}$) for 6 visits (X = 6) can be found easily by substitution: $\hat{Y} = 2.20 + (1.35)(6) = 10.3$. Thus, although we actually

---

[20] Technically, it is incorrect to extend the regression line beyond the range of values of X for which it was estimated.

observed three different values of yield (9, 10 and 11) for 3 different farms that had 6 visits (see Table A-9), we have now summarized the data and the calculated regression line passes through the point [6, 10.3]. For farm #7 which had 6 visits and produced 900 liters (the point [6,9]) our predicted value of 10.3 overestimates production by 1.3 (130 liters). For farm #9 which had 6 visits and produced 1100 liters, our predicted value of 10.3 (1030 liters) underestimates by 70 liters. And so on.

Correlation provides an answer to the question, how good a fit is the regression line? How will the line fit the data depends on the extent to which X and Y are correlated with each other, how much X and Y covary. If all the data points fell precisely on a straight line, the two variables would be perfectly correlated, and the regression line would produce estimates without any error whatsoever. At the other extreme, if the data points were completely haphazard or if they showed no linear pattern whatsoever, the two variables would not be at all correlated. Empirically, results fall somewhere between the two extremes. We can compute another summary measure which indicates the strength or magnitude of the relationship between the two variables.

The Pearson product moment correlation coefficient (more simply known as "r") is defined as the ratio of the covariance to the product of the standard deviations of X and Y:

$$r = \frac{\text{covariance}(X,Y)}{s_x s_y}$$

$$\boxed{r = \frac{\frac{1}{N}\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2/N}\ \sqrt{\Sigma(Y-\bar{Y})^2/N}}}$$

Again, a simpler computational formula is available:

Example A-2: Computation of Correlation Coefficient (r) for Table A-9.

Computational formula for $r = \frac{N(\Sigma XY)-(\Sigma X)(\Sigma Y)}{\sqrt{[N(\Sigma X^2)-(\Sigma X)^2][N(\Sigma Y^2)-(\Sigma Y)^2]}}$

For the data in Table A-9, and using the component terms which were already calculated (Example A-1) to compute the regression coefficient b:

$$r = \frac{15(936)-(86)(148)}{\sqrt{[15(558)-(86)^2][15(1637)-(149)^2}} = \frac{1312}{\sqrt{(974)(2354)}} = \frac{1312}{\sqrt{1514.20}} = .87$$

One way to interpret r is as a measure of the correlation between the observed values of Y and the predicted values $\hat{Y}$. The square of the correlation coefficient, r^2, can be interpreted as the proportion of the variance in Y explained by X. For the example, $r^2 = .76$, and we would say that the linear two-variable model we have used to predict yield from number of visits accounts for or explains 76% of the variance in yield.[21]

The logic of bivariate regression and correlation can be extended to models which include several variables. For example, it would be naive to expect that agricultural yield would be a simple function of the program variable. There are obviously other important variables known to affect yield, such as rainfall, use of fertilizer, size of farm, weeding and others. Although the computations for <u>multivariate regression</u> become quite tedious when done without a computer, the regression model can be stated as:

$$Y = b_o + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k + e$$

where b_o is the intercept (corresponding to a in the two-variable case) and is interpreted as the value of Y when all the other variables are equal to 0, and the b_1 through b_k are the regression coefficients. The coefficient

[21] In practice, it would be unusual to find such a strong correlation between the program and yield.

b_1, for example, is interpreted as the predicted change in Y for every unit change in X_1, holding all the other variables X_2 through X_k constant. If X_1 is the program or treatment variable (as before), then the computed coefficient would be a measure of the effect of the program on yield, controlling for the other variables.[22]

In the multiple (multivariate) regression model several useful statistics can be computed. First, the b_i coefficients are called the partial regression coefficients. It is also possible to compute what are known as partial correlation coefficients (symbolized as $r_{ij \cdot k}$), which measure the degree of relationship between the dependent variable Y and any of the independent variables X_i, controlling for all the other independent variables. (Actually, partial correlation coefficients can be computed for any multivariate problem where one needs to examine the extent to which two variables are correlated when other variable(s) are held constant.) Finally, the measure of correlation between the observed values of Y and the predicted values $\hat{Y}$ in a multiple regression model is symbolized as multiple R; in short, R is a measure of how well the model fits the data. The square of the multiple correlation coefficient, R^2, is a measure of the proportion of the variance in Y explained by all the independent variables X_i in the model; $1 - R^2$ is the variation in Y which is unexplained by the model. The reader is urged to consult a reference on more advanced statistical methods or econometric methods for further details and problems encountered in multivariate regression and correlation applications.

---

[22] For an example of just such a model, see D. Baltin, Evaluation de Dix Ans de Vulgarisation Agricole dans les ORD de Ougadougou, Koudougou et Kaya, Paris, SEAE, 1973. The dependent variable was yield, the independent variables: a single weeding, non-traditional religion, secondary commercial activity, neither a village notable nor rich, saw extension representative regularly, use of fertilizer, recent training session, rainfall of more than 600 mm, use of fertilizer/land available, and at least 5 helpers. The variables were found to explain more than 20% of the variance in yield.

V. INFERENTIAL STATISTICS

> An evaluation researcher has found that a fertility program has an important effect. Subjects had been assigned randomly to treatment and control groups; among the treatment group, the average number of children was 3.5, whereas among the controls, the average was 4.6, a difference of 1.1 children (see Table A-8). In presenting these results to the Fertility Control Council, a policy-maker asks the following question: "How certain are you that the difference of 1.1 children was the result of the treatment? Is it not possible that the difference occurred simply as a result of chance differences between the treatment and control groups?"

In statistical terms, the policy-maker's question involves what is called the null hypothesis, in this case, the hypothesis that the treatment in question really had no effect, that the observed difference was due solely to the operation of chance factors. Hence, a re-run of the same experiment would produce results that tended toward zero differences. The researcher would find it useful to be able to counter that the skeptic's argument that chance was responsible for the observed result was rather implausible. Even better, it would be most useful to be able to state just how probable (likely) the observed result would be if, in fact, the treatment had no true effect but only chance factors (resulting from the initial randomization) were operating.

To make statements of this general sort -- i.e., to state how probable a given observed result would be, given an initial assumption of no true effect -- is the purpose of inferential statistics. The example can be expanded to research involving sampling: whenever evaluation data are gathered for a sample and not the entire population to which one is interested in generalizing, inferential statistics are useful

in addressing the question: "Could the results obtained be due to chance -- that is, due to sampling variability?" If a different sample had been drawn, the results would probably have been somewhat different. Isn't it plausible, then, that an intervention really did not have an effect?

In short, the issue arises because in every experiment or sample survey the argument can be put forth that the observed results arise from sampling or randomization variability and not from the intervention.[23] However, if we are able to rule out "chance" (sampling or randomization variability) as likely explanation of our findings, the alternative hypothesis that the intervention did produce the results becomes more plausible.

Unfortunately, we never can be totally certain that the intervention actually did produce the observed differences. Put differently, the logic of science is such that we are never able to prove that one thing caused another; we can only state that some explanations are implausible and can therefore be rejected. For this reason, the logic of inference often seems counterintuitive as researchers first encounter it. Thus, inference does not allow us to prove that hypotheses are correct or true, but it does provide tools for rejecting hypotheses which sample data render implausible, such as the Fertility Control Council critic's null hypothesis. Finally, in addition to

---

[23] Note that this does not refer to other plausible, substantive variables known or presumed to affect the outcome variables of interest, and which need to be included as statistical controls whenever randomization or random sampling have not been carried out. For this reason, we have stressed throughout this volume the desirability of using random procedures in the evaluation design. Then even the possibility that chance fluctuations are responsible for the observed outcomes can be tested with inferential statistics. The only plausible conclusion that cannot be rejected is that the intervention caused the observed outcomes.

tests of hypotheses, inference also provides methods for estimation of population parameters based on sample data.

A. The Logic of Inference and Basic Concepts

1. SAMPLING DISTRIBUTION

In order to test the hypothesis that our observed results are simply due to chance, that is, random fluctuations in sampling or randomization, we must first know how chance operates. Such chance fluctuations are known as random error.[24] A sampling distribution can be thought of as the distribution of random error, and it is an essential element of inferential statistics because it serves as a standard of comparison against which empirical results are checked to determine whether or not random error could have produced them. Sampling distributions are derived mathematically with the use of probability theory, but the logic can be understood intuitively by first considering a sampling problem and then a randomization problem.

The basic notion is: "What would happen if we drew random samples repeatedly from a (fixed) population?" Imagine the following scenario: there is a very large population list, we draw a random sample of size N, compute the mean ($\bar{X}$) age for that sample, and throw the sample back.

---

[24] In the vocabulary of statistics, "error" does not imply "mistake," but means the extent to which a given measurement departs from the "true" measurement. "Random errors" are those fluctuations that arise archetypically from random sampling or from randomization. Their characteristic is that in probability terms, their expected or long-run value is zero--that is, some errors will be positive, some will be negative, and so on, but in the long run they will all average out to zero.

We draw another random sample of size N, compute their mean age, and so on. We do this over and over again, in fact, an almost infinite number of times. Now we have a set of numbers which are average ages of different samples from a single population. That set of numbers has a distribution, we can compute its mean (the average of all those averages), and we can compute its standard deviation (known as the standard error). It is a sampling distribution of means which shows how sampling error operates (we have done nothing to affect the mean age of all those samples, but each time we draw a different sample, the mean is likely to vary, just because different people make up each particular sample).

Consider the fertility example again as a context to illustrate the notion of sampling distribution in reference to randomization error. Imagine a population of females whose true fertility averages 4.0 children (the population parameter). Imagine further that this population was randomly divided into two halves and the mean fertility of each half was found to be 4.0. Then a second random division is done, and the fertility of one group is measured as 3.9 children while the fertility of the second group is measured as 4.1 children. Yet another division results in scores of 3.8 and 4.2; a fourth again produces means of 4.0 for each group, and so on for a large number of divisions. If this were done over and over again, and each time we calculated the difference between the two means, we would end up with a sampling distribution of differences between means. What we would find is that the random division of the population into groups will produce a wide variety of outcomes: in some cases the fertility

averages for each group are exactly the same (the difference between them is zero); in other cases the averages differ by a substantial amount; but in all cases, the observed difference between one group and the other results only from the randomization itself.[25]

The policy-maker on the Fertility Control Council who asked whether the observed difference between means could have occurred just by chance was saying: "We know that random division of a population into two groups will in some cases produce fairly large differences between the two groups, just because of chance factors alone. Is it possible that the researcher's reported difference is just one of those cases?"

Obviously, researchers would never carry out such exercises to construct sampling distributions; they usually draw only one sample or perform one randomized experiment from which they wish to generalize. Thus, we rely on mathematical derivations based on probability theory for the important sampling distributions needed for inferential applications. There are also theorems which serve as guides in knowing which sampling distributions are appropriate. It turns out, for example, that those mean ages for all the different samples will tend to cluster around the mean age for the entire population (that is, the average of all those sample averages will tend to equal the population average). The shape of the distribution is normal when the sample size is large and regardless of the shape of the entire population--that is, the distribution of means (also of differences between means

---

[25] Actually, the smaller the sample size, the more likely that the differences will be large--differences due to chance alone.

as in the randomization example) is a normal distribution. Furthermore, the amount of variation around the mean of the sampling distribution (i.e., the standard error of the sampling distribution)also depends on the sample size: the smaller the size of the samples being drawn repeatedly, the more likely that sample means will tend to be farther away from the population mean.[26] For example, in drawing small samples, a sample comprised of mostly old people is relatively more likely to appear, as well as one of mostly young people, as opposed to the case where large samples are being drawn, since unusual ages would be averaged in with all the rest. Similarly, differences between two sample means based on small samples can also be quite large as a result of sampling or randomization error alone.

These characteristics of a sampling distribution imply that we now have a standard of comparison with known properties, with known relationships to population, distributions, and which can be used to test hypotheses based on data from samples. When a null hypothesis is stated, sampling distribution properties are used. Put differently, we can take observed data from a single sample, or from a single randomized experiment, and because of the known properties of sampling distributions we can assess the likelihood that random error or sampling error could have produced the observed results. Inferences about a

---

[26]These three characteristics of a sampling distribution are known from the Central Limit Theorem--perhaps the most powerful theorem of inferential statistics--which states that the sampling distribution of means will be normally distributed as the sample size gets large, will have a mean equal to the mean in the population, and a standard deviation equal to the population standard deviation divided by the square root of the sample size.

population based on samples, in other words, are possible because sampling distributions show how sampling or randomization error operates, and because the characteristics of sampling distributions are known and stated in terms of the (unknown) but hypothesized population parameters.

Finally let us consider the relationship between sample size and standard error (sampling or randomization error) and its relevance to a sampling issue raised in a previous section: How large should a sample be? Even an intuitive understanding of what a sampling distribution is helps us understand that sampling error increases with small samples (the instance where a small sample could be made up of all older people, therefore have a large mean age). Conversely, as the sample size increases, sampling error decreases. This relationship holds regardless of the size of the entire population: interestingly, the relative proportion of the population represented by the sample is not crucial. Rather, larger samples will always have smaller sampling errors, regardless of population sizes. Hence, it is best to take large samples whenever feasible.

2. DECISION ERRORS AND STATISTICAL SIGNIFICANCE

Statistical tests are based on probability theory and make use of sampling distributions, and are subject to what are known as <u>decision errors</u>. There are two types of errors which can occur when a statistical decision is made on the basis of sample data to either reject or accept a null hypothesis.[27] Let us return again to the fertility program with

---

[27]Technically, we never "accept" an hypothesis--we "fail to reject" it. The language, however, becomes so confusing that we will use the terms accept and reject for simplicity.

the difference of 1.1 children. We want to test statistically whether or not this difference could be the result of randomization. The null hypothesis is that the two groups have the same number of children on average, that the true difference between the two groups is zero. The null is therefore a hypothesis of no difference: what we are asking is whether 1.1 children as an observed difference is so large that it is unlikely to have been produced by random error. A sampling distribution of random error will in fact have some extreme values (we might have randomly assigned more women with small families to the treatment group), but extreme or unusual values are not as likely to occur--that is, they occur infrequently.

We ask: what is the probability of finding a difference this large or larger if the null hypothesis is true? Or: is this observed difference of 1.1 a big enough difference that is is very unlikely that chance or random error produced it? "Big enough" is in reference to the appropriate sampling distribution.[28] But we must make an arbitrary decision as to what constitutes "big enough"--in more general terms, what criterion to use in deciding what constitutes an unusual outcome. That criterion is known as the level of significance. In order to be useful as a general criterion, the level is set not in terms of an absolute magnitude (e.g., a difference of 5 children is big) but in terms of a probability level. For example, we consider a difference or program effect to be big enough to represent an unusual outcome if such a difference (or an even larger difference) would only occur 5 times out of 100 due to random error (probability is .05) if the null

---

[28]In this case, the sampling distribution of differences between means.

hypothesis of no difference is true. The level of significance of a statistical test is usually set at .05 or .01, and sometimes at .10.

There are two types of decision errors, shown in the diagram below:

Figure A-3: Decision Errors Involved in a Statistical Test.

Decision	Null Hypothesis Is: True	Null Hypothesis Is: False	
Reject Null	TYPE I error	correct decision	(Data show a significant difference)
Accept Null	correct decision	TYPE II error	(Data show no significant difference)
	(Program has no effect)	(Program has an effect)	

In the rows we see that a statistical test of a null hypothesis results in one of two outcomes: either we reject or we accept the null hypothesis. In the columns we see that in the real world, the null hypothesis is either true or false. If the null hypothesis is false (there is a real difference) and we reject it (because our data show a large difference), our decision is a correct one. If the null is true (there is no difference between our groups) and we accept the null (because our sample data show small differences) our decision is also correct.

Type I error occurs when we reject the null hypothesis (because we find a large difference between the two groups), but acutally the program had no effect. The observed difference is due to random error. And large differences, that is, unusual events, do occur as a result

of random error. If a difference as large as the one we observe (or one even larger) can occur by chance only 5 times out of 100, we conclude that our effect is statistically significant--i.e., we reject the null hypothesis of no difference at the .05 level--i.e., we are wrong one in twenty times. But if chance produced the difference, we have made an erroneous decision, unknowingly, to be sure, since we never know what the real effect is. Thus, a Type I error consists of concluding that a program produced an effect, when in fact it did not; our study showed a statistically significant difference, but we have capitalized on random error. For this reason Type I error is also called a "false positive" error or conclusion. The probability of making a Type I error is equal to the level of significance which was chosen.

The second type of decision error, Type II error, occurs because it is possible that a program has a real effect, but we fail to detect it. In this case, the null hypothesis is false because there is a real difference, but we fail to reject the null (we accept it). In short, we have a "false negative." Unfortunately, the probability of making a Type II error is never known precisely, although we do know what affects that probability. For example, if the level of significance is set very low (e.g., .001), we are more likely to make a Type II error, because our criterion for what constitutes a real effect is so strict (one that occurs only 1 time in 1,000 by chance) that real but small differences will simply not meet the criterion. In short, real differences will go undetected. Type II error is affected by other factors, such

as the sample size, the amount of variability there is in the population and the magnitude of the true program effect in the population.

Equipped now with some of the basic concepts of inference, we turn to some examples of statistical tests, show their applications and how to compute them.

B. Confidence Intervals

Estimation is the problem, quite literally, of estimating or ascertaining the value of a parameter (say the mean) for a population on the basis of data from a single sample. Suppose we want to estimate the effect of a program in a population, but we only have the resources to study a sample. The sample provides us with only one "best guess" about the effect: we might compute a mean and wish to infer that it represents the true of real effect in the population. But we just saw that the average effect obtained from a sample reflects the population effect plus a random error component associated with sampling variation. Thus, we might construct a confidence interval around the estimate, and state that the true effect in the population falls within an interval constructed around the sample mean. By using the appropriate sampling distribution we can construct such an interval, and we can attach a probability value which reflects the proportion of times (say 95 out of 100) that we will capture within the range of the interval the real effect in the population. Put differently, a 95% confidence interval is interpreted to mean that if we repeatedly drew samples of a given size, in 95% of the samples the real or true program effect would fall within the range indicated by the confidence interval.

The formula for computing a 95% confidence interval for the mean ($\bar{X}$) is:

$$\boxed{C.I._{95\%} = \bar{X} \pm 2 \; s/\sqrt{N-1}}$$

where s is the standard deviation (of the variable whose mean has been computed) and N the sample size. This is an approximation and not a precise formula,[29] appropriate for a sample size (N) from about 20 to 100. For a sample larger than 500, the appropriate formula is $\bar{X} \pm 1.96 \; s/\sqrt{N}$. For small samples, the term "2" or "1.96" in the formula actually varies between "2" and upwards to about "2.5" as the sample size gets smaller. The appropriate value should therefore be obtained by consulting a table of the "t-distribution," which can be found in any statistics text. An example of the computation of a confidence interval for a mean is shown below:

Example A-3: Computation of a Confidence Interval for a Mean.

A researcher needs to estimate the nutritional level of a population in order to implement a supplement program. A random sample of 122 cases was drawn from the population, and was found to have a mean calorie intake of 1,500 calories per day, with a standard deviation of 500 calories. A 95% confidence interval to estimate the caloric intake of the population is:

$$\begin{aligned} C.I._{95\%} &= \bar{X} \pm 2 \; s/\sqrt{N-1} \\ &= 1{,}500 \pm 2 \; (500/\sqrt{122-1}) \\ &= 1{,}500 \pm 2 \; (500/11) \\ &\doteq 1{,}500 \pm 90.9 \end{aligned}$$

The confidence interval is therefore (1409.1, 1590.9)

Based on the sample data, we estimate that the mean caloric intake for the population falls in the interval. If we drew repeated samples of size 122 from the population, in 95% of the samples the real population mean caloric intake would fall within the range indicated.

---

[29] The exact formula is: $\bar{X} \pm t \; \sigma_{\bar{x}}$, where t is the value of the t-distribuli above which 5% of the cases fall, i.e. the .05 level of significance, and $\sigma_{\bar{x}}$ is the standard deviation of the sampling distribution of means, approximated by $s/\sqrt{N-1}$ or $s/\sqrt{N}$ is N is large.

Confidence intervals can also be computed for proportions, by using the following formula:

$$\boxed{C.I._{95\%} = p \pm 1.96\sqrt{p\,q\,/\,N}}$$

where p is the observed proportion around which an interval is being set, q = 1 - p, and N is the sample size. An example is given below, using the data shown in Table A-1:

Example A-4: Computation of a Confidence Interval for a Proportion.

The data shown in Table A-1 indicate that 9.6% of a random sample of 250 persons who received a training program are still unemployed after the training. The 95% confidence interval around that estimate (p = .096) is:

$$C.I._{95\%} = p \pm 1.96\sqrt{p\,q\,/\,N}$$

$$p = .096$$

$$q = 1 - .096 = .904$$

$$\sqrt{p\,q\,/\,N} = \sqrt{(.096)(.904)/250} = \sqrt{.087/250} = \sqrt{.000347} = .019$$

$$1.96\sqrt{p\,q\,/\,N} = (1.96)(.019) = .037$$

$$C.I._{95\%} = .096 \pm .037$$

The 95% confidence interval is therefore (.059, .133) or (5.9%, 13.3%).

Based on this sample of program trainees, we estimate that the proportion of unemployed program graduates among all those persons who received the training falls within the interval--i.e., the percent unemployed is between 5.9 and 13.3.

Similarly, confidence intervals can also be computed for differences between means or for differences between proportions (i.e., percentage differences). The logic is the same, the calculations somewhat more complicated because more terms are involved in the arithmetic. A statistics text should be consulted for the formulas, as different ones may be appropriate depending on the nature of the estimation problem and the characteristics of the samples.

The 95% confidence interval for a difference between two means from large samples is:

$$\boxed{C.I._{95\%} = (\bar{X}_1 - \bar{X}_2) \pm 1.96\ \hat{\sigma}_{diff}}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the two sample means, $\hat{\sigma}_{diff}$ is the estimated standard error of the difference between means and equals $\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$, where s_1^2 and s_2^2 are the variances of the two samples.

An example is shown below:

Example A-5: Computation of a Confidence Interval for a Difference Between Means for Large Samples.

Suppose that the fertility program mentioned throughout this section had provided the following information:

Group 1 (Nonparticipants):	Group 2 (Participants):
$\bar{X}_1 = 4.6$ children	$\bar{X}_2 = 3.5$ children
$s_1 = 1.2$ children	$s_2 = 1.0$ children
$N_1 = 500$	$N_2 = 600$

The 95% confidence interval for the difference in means is:

$$C.I._{95\%} = (\bar{X}_1 - \bar{X}_2) \pm 1.96\ \hat{\sigma}_{diff}$$

$$(\bar{X}_1 - \bar{X}_2) = 4.6 - 3.5 = 1.1$$

$$\hat{\sigma}_{diff} = \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}} = \sqrt{\frac{(1.2)^2}{500} + \frac{(1.0)^2}{600}} = \sqrt{\frac{1.44}{500} + \frac{1}{600}}$$

$$= \sqrt{.003 + .002} = \sqrt{.005}$$

$$= .071$$

$$1.96\ \hat{\sigma}_{diff} = (1.96)(.071) = .139$$

$$C.I._{95\%} = 1.1 \pm .139$$

The confidence interval is therefore: (.961, 1.239).

The fertility program has an estimated effect of 1.1 children less for the participants, and our estimate is that the real effect is a difference which can range from .961 to 1.239 children. We are confident that these calculated bounds capture the real program effect, since in 95% of all replications of the experiment, the real effect would fall within the indicated range of values.

Note that the Fertility Council skeptic's hypothesized value of zero difference between the two groups does not fall within the estimated confidence interval, since the lower boundary of the interval is .961. We can therefore show the skeptic that even when we take into account randomization variability which could occur in 95% of all possible randomizations, the estimated program effect for the population falls well above the value of zero.

What if another critic had said: "Previous research has shown that such a program does have an effect, but never as large as one child. The largest effect reported previously was .8 children." A second null hypothesis that can be tested with the sample data, that the difference or effect equals .8, can also be rejected on the basis of the observed data because the value .8 falls outside the calculated confidence interval. Phrased differently, the confidence interval amounts to all the plausible null hypotheses which cannot be rejected on the basis of data from this sample (experiment), at a 95% level of confidence.

Finally, note that the width of a confidence interval is affected by the sample size: larger samples yield interval estimates, another reason why large samples are desirable.[30] The particular level of significance chosen also affects the size of the interval. More precise estimates (narrower intervals) can be constructed if the confidence level is reduced (e.g., to 90%); conversely, increasing the confidence level (e.g., to 99%) will result in less precise estimates (wider intervals). In short, confidence level and precision are <u>inversely</u> related.

---

[30] For Example A-4 which calculated a confidence interval (5.9%, 13.3%) around an observed unemployment of 9.6%, had the sample size been 1,000 instead of 250, the confidence interval would have been (7.8%, 11.4%).

C. Tests for Differences Between Means and Proportions

Here we ask the question: does my observed difference between two sample means, i.e., my observed program effect, represent a statistically significant difference? Is it a large enough difference that it is very unlikely that randomization error produced it? What is the probability that error produced the effect we found? To answer the question we test the null hypothesis that the real difference is actually zero (i.e., there is no program effect), hoping to reject the null hypothesis in favor of an alternative (substantive) hypothesis that the program did have an effect. We use the same data as shown in the previous page when we calculated the confidence interval around the difference. But now the problem is set up in a slightly different way. We must compute a test-statistic, which allows us to, in a sense, translate our sample data to sampling distribution units, in order to find out the likelihood that the observed difference represents random error. The test-statistic for a significant test for a difference between two means is:

$$\boxed{t = \frac{\overline{X}_1 - \overline{X}_2}{\hat{\sigma}_{diff}}} \quad .$$

Where $\overline{X}_1$ and $\overline{X}_2$ are the observed means and $\hat{\sigma}_{diff}$ is the standard error of the sampling distribution of the difference between means, estimated for large samples as before:

$$\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}} \quad .$$

We determine in advance of the test the level of significance we are willing to use as a criterion -- for example a .05 probability level. This implies that a computed t-value greater than or equal to 1.96 will in fact represent a "statistically significant difference," tested against the null hypothesis that the difference is zero.

All the terms we need were computed for Example A-5:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{diff}}$$

$$= \frac{1.1}{.071}$$

$$= 15.49$$

The t-value is obviously greater than the value required (1.96) to reject the null. We conclude, therefore, that the observed difference of 1.1 children is a statistically significant difference, contrary to the skeptic's argument. A difference as large as this or larger is extremely unlikely to have been produced by random error (we can verify this statement by examining a table of probability values of the t-distribution.)

A <u>significance test for the difference between two proportions</u> requires the same logic. The test-statistic is

$$\boxed{Z = \frac{p_1 - p_2}{\hat{\sigma}_{diff}}}$$

where p_1 and p_2 are the proportions (from random samples) being compared and as before, $\hat{\sigma}_{diff}$ = the estimated standard error of the sampling distribution of differences between proportions, calculated by $\sqrt{\frac{p_1 q_1}{N_1} + \frac{p_2 q_2}{N_2}}$,

where $q_1 = 1-p_1$ and $q_2 = 1-p_2$. The null hypothesis being tested is that there is no difference in the population.

Example A-6: Computation of a Test of Differences Between Sample Proportions.

The agricultural training program evaluation mentioned previously seems to indicate that the impact of the centers varies by region. Thus, in one region there is a smaller percentage of trainees who are engaged in agriculture after the program. The data are based on random samples of training graduates from the two regions and show:

Region 1:	Region 2:
$p_1 = .55$	$p_2 = .48$ (p = proportion in agriculture)
$N_1 = 400$	$N_2 = 300$
$q_1 = 1 - p_1 = .45$	$q_2 = 1 - p_2 = .52$

Set the level of significance at .05; the required Z value must be 1.96 or greater to reject the null hypothesis that the impact of the centers does <u>not</u> vary by region.

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1 q_1}{N_1} + \frac{p_2 q_2}{N_2}}}$$

$$= \frac{.55 - .48}{\sqrt{\frac{(.55)(.45)}{400} + \frac{(.48)(.52)}{300}}}$$

$$= \frac{.07}{\sqrt{\frac{.248}{400} + \frac{.250}{300}}} = \frac{.07}{\sqrt{.001 + .001}} = \frac{.07}{\sqrt{.002}}$$

$$= \frac{.07}{.045}$$

$$Z = 1.56$$

We conclude that we <u>cannot reject</u> the null hypothesis of no difference between regions, since the Z value is less than 1.96. Therefore, we continue to hold as tenable the hypothesis that program impact does not differ by region.

D. The Chi-square Test for Independence Between Two Variables

Whenever we have a crosstabulation (contingency table) of two variables, the question we most want to ask is: are the two variables related to each other, or are they independent? In other words, does one variable affect another? Once again the statistical model for inference requires that a null hypothesis be tested; in this case, the null hypothesis is that the two variables are statistically independent, that sampling (randomization) variability could have produced the observed results.

Consider a crosstabulation from an agricultural training program evaluation. Table A-10 shows what appears to be an effect on employment for program participants:

Table A-10: Frequency Table of Employment After an Agricultural Training Program by Program Participation.

	Program Participation		
Employment	Participant	Nonparticipant	Total
Agriculture	200	145	365
Other	140	100	240
Unemployed	40	55	95
Total	400	300	700

The computation of chi-square (symbolized as χ^2) requires a frequency (not a percentage) table. Essentially a chi-square test involves comparing the observed frequency table with what is called the expected frequency table, or the table we would expect to find if the two variables were independent, i.e., unrelated to one another.

The computation of χ^2 is carried out with the following formulas:

$$\chi^2 = \Sigma\frac{(O-E)^2}{E}$$

$$E = \frac{(R)(C)}{T}$$

$$d.f. = (r-1)(c-1)$$

Where O = the observed frequency for a given cell

E = the "expected" frequency for a given cell

R = the row total frequency

C = the column total frequency

T = total number of cases in the table

d.f. = degrees of freedom (needed to consult a χ^2 distribution table)

r = number of rows in the table

c = number of columns in the table

For simplicity in the computation, let us name each of the four cells in Table A-10 as follows:

a	d
b	e
c	f

The computation of the chi-square statistic is shown below:

Example A-7: Computation of Chi-square for Table A-10.

cell	0	E	0 - E	$(0 - E)^2$	$\frac{(0 - E)^2}{E}$
a	220	(365)(400)/700=208.57	11.43	130.64	.63
b	140	(240)(400)/700=137.14	2.86	8.18	.06
c	40	(95)(400)/700= 54.29	-14.29	204.20	3.76
d	145	(365)(300)/700=156.43	-11.43	130.64	.84
e	100	(240)(300)/700=102.86	- 2.86	8.18	.08
f	55	(95)(300)/700= 40.71	14.29	204.20	$\frac{5.02}{10.39} = X^2$

Degrees of freedom (d.f.) = (r - 1)(c - 1) = (3 - 1)(2 - 1) = 2, since the table has 3 rows and 2 columns.

All that remains is that we consult a table of the probability values associated with the chi-square distribution, to determine whether or not the obtained value of 10.39 for d.f. = 2 is too large to have been produced by chance if the program and employment are independent of one another. A portion of the chi-square distribution table is reproduced in Table A-11:

Table A-11: Values of Chi-square Corresponding to Given Probabilities.

d.f.	p = .05	p = .01
1	3.84	6.64
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09

For the example, we can therefore reject the null hypothesis of independence or no relationship between program participation and employment, since the required value of chi-square at the .05 level, for a 3 by 2 table or 2 degrees of freedom, is 5.99. In fact, we can reject the null even at the .01 level of significance, since the calculated value of chi-square is 10.39 and therefore exceeds the required 9.21. We conclude, therefore, that there does appear to be a relationship between program participation and employment, because the probability is less than 1 in 100 that chance fluctuations in the data would have produced Table A-10 if the two variables were unrelated.

In sum, the logic of a chi-square test involves two-steps: (1) determining what the table would have looked like if there were no relationship between the two variables, i.e., if the two variables were statistically independent. The expected frequency table is calculated from the marginal frequency totals only and represents a "theoretical" frequency table showing no relationship between the variables. (2) Once we determine what the table would have looked like, had there been no relationship between the variables, we simply compute a chi-square measure which gives the amount of discrepancy between the observed and expected frequencies for each cell in the table.

E. Other Tests of Significance

Numerous other tests of significance can be performed other than the ones we have discussed. The logic is always the same, but the calculations involve different test-statistics and sometimes different sampling distributions. Tests of significance are available, for example to compare the differences among a set of group means and variances (Analysis of Variance), to test the significance of a correlation coefficient r (against the null hypothesis that r = 0), to test the significance of a regression coefficient (also against the null that the coefficient = 0), and so on. The computations involved can be quite complicated and tedious, and wherever feasible, we recommend that computer packages be employed (see Section VII below.)

Most of the tests of significance which we have reviewed technically require one or more assumptions about the population from which the sample has been drawn. As we saw previously, random sampling or random assignment to experimental and control groups was also a requirement for inferential tests. Perhaps the most frequently encountered assumption for such tests (for example, t-tests for means) is that the population distribution is normal. Several tests have been developed, referred to as nonparametric tests (or distribution-free tests), which do not require the assumption of a normal population. Although other assumptions are still required, they tend to be less restrictive than those required in parametric tests. In practice, the normality assumption obviously cannot be met when the data represent only nominal or ordinal levels of measurement. Hence, if nonparametric tests are appropriate, other sources should be consulted for further details.[31]

[31] At minimum, Blalock (1972) can be consulted. Other references are: Sidney Siegel, Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill, 1956, and J.V. Bradley, Distribution-free Statistical Tests. Englewood Cliffs, New Jersey: Prentice-Hall, 1968.

VI. Data Reporting Techniques

In this section we briefly comment on the important features to consider when the statistical results of an evaluation are reported. Whenever tables are reported, it is necessary that they be presented clearly, with a precise title and labels so that anyone reading the table would be able to understand what the table is about. The title can be as long as is necessary, so long as it is absolutely clear what the table reports. All of the tables in this appendix have been labeled with sufficient detail to clarify what the cell entries are.

Besides the title and labels, every table should show what the relevant sample size is: even if a percentage table is reported, at least the marginal frequencies should also be reported. When reporting tests of significance, the table should contain the probability level chosen (e.g., $p = .05$), where appropriate, the number of degrees of freedom, and so on. In short, the rule of thumb is that a table should stand by itself: it ought to be possible to read and understand an entire table without having to refer to the text that surrounds it. If this rule is followed, each table will be reported with all the necessary detail.

The same rule applies to reporting any other data which are summarized. For example, it may be useful to show results graphically (e.g., as a histogram, line graph, or whatever). Graphic presentation of the data may show differences more dramatically. But the researcher must take care not to distort the findings when showing them graphically. And, again, any graph should contain sufficient labels and title information so that it can be understood without reference to the text.

VII. Computer Software Packages

A number of computer software packages are available which have been designed to process and analyze social science data. Many of these computer packages (programs) are widely disseminated and have been written in several computer languages and for practically all existing computer systems (e.g., IBM, CDC, Univac, Honeywell, Burroughs, etc.). The addresses where additional information can be obtained for three commonly used packages are given below:

SPSS (Statistical Package for the Social Sciences)

SBSS "Incorporated"
Suite 1234
111 East Wacker Drive
Chicago, Illinois 60601
USA

BMD (Biomedical Computer Programs)

BMD Program Librarian
Health Sciences Computing Facility
AV-111 Health Sciences Center
University of California
Los Angeles, California 90024
USA

OSIRIS III (Organized Set of Integrated Routines for Investigation with Statistics)

Inter-University Consortium for Political and Social Research
University of Michigan
Box 1248
Ann Arbor, Michigan 48106

Non-IBM OSIRIS Distributors:

UNIVAC OSIRIS: CHI Corporation
11000 Cedar Avenue
Cleveland, Ohio 44106
USA

CDC OSIRIS: Universitaet zu Koeln
Rechenzentrum
OSIRIS DISTRIBUTION
5 Koeln 41
Robert--Koch--Strasse 10
West Germany

Siemens OSIRIS: Ms. Carol A. Cassidy
Zentrum fuer Umfragen, Methoden und
Analysen (ZUMA)
68 Mannheim 1
B2, 1
West Germany