A Laboratory Model for the Instruction of Survey Data Analysis

by J. Raymond Miyares

Submitted in partial fulfillment
of the requirements
for the Bachelor of Science Degree

Massachusetts Institute of Technology
14 May 1971

Signature of Author: _____
Department of Urban Studies and Planning

Certified by: _____
Thesis Supervisor

Accepted by: _____
Chairman, Departmental Committee
on Theses

for Beverly

Acknowledgements

In the beginning of this project, it must be emphasized, the author was neither aware of the proposed laboratory subject, nor even a student in the Department of Urban Studies and Planning. The real instigators were Professor Ronald A. Walter and Mrs. M. Christine Boyer-Karalis, who introduced me to the project and deserve a major share of the credit for bringing it to fruition. Their contribution to this thesis has been substantial, and is deeply appreciated.

Funds for the experiment were provided by a grant from the Cambridge Project, for which the author is also grateful.

Of course, no subject can be a success without students, and a special thanks is offered to those students who took part in the laboratory for their confidence in an unknown and frankly experimental venture. They are: Charles Cofield, Gerald M. Croan, William Dix, Sophia Driskell, Jeffrey Folinus, Dan Greenbaum, Joseph Hadzima, Roger Jeanty, Christopher Mar, Peter Messeri, Jeffrey D. Morgan, Richard J. Philips, Richard W. Prather, Ellen Reintjes, Eric Shore and Thomas F. Welch.

Table of Contents

A Laboratory Model for the Instruction
of Survey Data Analysis

by J. Raymond Miyares

Abstract

This thesis presents a review of efforts over the past year to develop an urban studies laboratory for the new undergraduate curriculum in Urban Studies at MIT. The laboratory's objectives in teaching survey data analysis are formalized and an evaluation of the initial offering of the laboratory is made. Numerous recommendations for the future of the subject are presented, and the laboratory workbook is thoroughly revised.

Introduction

> We continue to reject over-simplified views of the aims of education that seek to separate the practical from the theoretical, the scientific from the humanistic. Instead, we aim to achieve a balanced education which unites rather than separates the elements of humane learning. The Institute tries to provide students with what Rogers called "truly practical education." Today, as in the past, this can only be an education that unites application with principle, the technical with the perceptive, scientific knowledge with human understanding.
>
> —Commission on MIT Education Report, November 1970

The fact that a Department of Urban Studies and Planning exists at the Massachusetts Institute of Technology implies a special recognition that today's new technical knowledge can be profitably applied to the problems of cities. Facilities, philosophy and geographic location make the Institute uniquely suited for the use of technological research methods in social science. The multitude of public and private organizations engaged in data gathering activities has produced an explosion of information of interest to the student of metropolitan society. The development of computerized data analysis resources within the grasp of the programming novice has made the results of these activities

available for study to all social scientists.

The implementation of an undergraduate curriculum in Urban Studies offered an unusual opportunity for the development of courses which draw upon the facilities and resources in use, challenge the mathematical aptitudes and skills of the MIT undergraduate, and stimulate the use of all of these in an urban context. It was appropriate, then, to consider the laboratory model of a subject -- a model borrowed from the physical sciences -- for instruction in methods of survey data analysis. If scientific reasoning and mathematical experience were to be employed in social research at the undergraduate level, it was believed, the undergraduate laboratory subject such as is offered in physics, chemistry or engineering, would be a valid model to use in developing a structure for data analysis instruction.

The undergraduate laboratory subject has been carefully and specifically defined by the MIT faculty, who offer an explicit pattern which can be followed by anyone who wishes to create a new laboratory subject:

> Typically, a Laboratory Subject consists of an experimental investigation, including an analysis of the problem and evaluation of results. Assumptions based on book knowledge or intuition are experimentally tested against the real world. A student analyzes his results, relating them to his assumptions and methods, and a faculty member participates in the evaluation of the experiment. The Laboratory Subjects call for a major commitment of the student's attention to one or a few experimental problems and emphasize as much as possible work of project type rather than routine experimental exercises. They are designed to stimulate the student's resourcefulness and his own ideas.
>
> --MIT Bulletin 70/71
> General Catalog Issue
> August 1970

With this model in mind, the project which developed into this thesis was begun.

The possibility that a laboratory subject could be created to teach the analysis of urban related, survey research data was first presented to the author in June 1970. Considerable thought had been given to the need for offering a subject to meet the demands of undergraduate students wishing to apply quantitative techniques to social problems, and desiring to relate their laboratory subject requirement to their major interests. Professor Ronald A. Walter had made the decision to attempt to offer such a laboratory, subject number 11.505, in the spring of 1971. Mrs. M. Christine Boyer-Karalis had devoted some time in the spring to outlining a laboratory workbook which could be used in the class. But there were still many problematic wrinkles which had to be ironed out. Professor Walter offered the author the opportunity to attack the problems and test the suitability of the Institute Laboratory Subject model for the instruction of quantitative research methods in urban studies.

While the objectives and focus of the laboratory were not yet formalized, certain of its elements were specified. The laboratory would contain workbook exercises somewhat like those outlined by Mrs. Boyer-Karalis. It would rely on some form of electronic data processing. And, unless an alternative could be defended, it would employ as its basic text, Robert S. Weiss' Statistics in Social Research,[1] chosen because of its readibility and simple approach. The author's participation in the experiment consisted mainly of the tasks

of obtaining and preparing data files, evaluating and arranging for computation tools, developing the workbook from its outline, gathering reading materials, and formalizing the definition of laboratory objectives.

The author, inexperienced in social science programming systems, and having only an elementary knowledge of programming techniques, set out to become informed about the facilities available for use at the Institute. The selection of the Statistical Package for the Social Sciences (SPSS)[2] as the data processing backbone of the course was an easy one. This package, designed for batch processing use, is an integrated system of programs designed to provide the social scientist with a unified and comprehensive capability of performing many different types of data analysis in a simple and convenient manner. SPSS was chosen because it allows a great deal of flexibility in the format of data, provides the user with a comprehensive set of procedures for data transformation and file manipulation, and offers him a large number of statistical routines commonly used in the social sciences. Since the system had been used for some time in the Political Science Department at MIT, the novices in the 11.505 project were able to obtain valuable help in programming with SPSS. This proved to be an important factor because the SPSS Manual had not yet been published, nor had the MIT Information Processing Center begun to support the system's use.

The author found his initial mastery of SPSS to be a slow process, not quickened by his selection of the 1969 Boston Area Study[3] as the first data file with which to work.

Creating the SPSS file BOSAREA, based on this study, proved to be a monumental task, completed only after a substantial number of teeth had been gnashed. Finally, however, the file was successfully stored, and the author could begin to consider its use in the laboratory workbook.

Mrs. Boyer-Karalis' workbook outline was brought out again and reviewed. In her work Mrs. Boyer-Karalis closely relied on the Weiss book, from which a preliminary subject outline had been drawn. However, since the subject material was cummulative in nature, it was necessary to insure that each step be mastered before the next was attempted. Therefore, her outline was reorganized into four sections, each building on the previous one, but each, nevertheless, a single unit. Additional rearrangement was required to accomodate the format of the SPSS system, so that similar programming tasks and similar analytical tasks would be found together in one section. The new outline which resulted, provided the basis from which the 11.505 workbook was written.

Concurrent with the development of a workbook to accompany the laboratory, other efforts to formulate aspects of the subject were being made. The author prepared a demonstration file, based on a study of American small communities, to serve as a sample for beginners' use of the SPSS system. An introduction to SPSS was written, and a sample program was run to facilitate students' learning of the system in the quickest possible time.

In spite of a strong desire to complement SPSS with a second, interactive system, none of those available seemed to

meet all of the requirements of the laboratory. In the end the subject staff agreed to try to use Interdisciplinary Machine Processing for Research and Education in the Social Sciences (IMPRESS)[4] as an adjunct to the batch services offered by SPSS. IMPRESS, which utilizes a local line hook-up with the Dartmouth Time Sharing System consists of approximately twenty-five interrelated programs written in BASIC and time-sharing FORTRAN. These programs, taken together, build a storage, retrieval, reduction and analysis system enabling the user to interact with data, unhampered by a time lag.

After having established the basic format of the urban studies laboratory, it was possible to formulate this structure into the actual subject presentation.

Objectives of the Laboratory

The tools of data analysis are not legitimate ends of a total urban studies curriculum, but rather skills whose utility is defined in the context of other aspects of the curriculum. Thus it is important that an urban studies laboratory in data analysis techniques not exist as an isolate. Instead, it must be a part of the main undergraduate program and serve a purpose auxiliary to it.

In this context, three distinguishable, but overlapping objectives were set up for 11.505. These were: (1) the development of skills and techniques of data analysis, (2) the improvement of creative thinking in the analysis process, and (3) the discovery of interesting information about the populations studied in various surveys used in the laboratory. The laboratory staff hoped that the achievement of these objectives could be accomplished by structuring a series of exercises around the first two. The third objective, then, could be achieved by devoting a substantial amount of class time to discussion of insights drawn from these exercises.

The staff considered certain analytical basics to be primary needs of each student. These basics included understanding how social survey research is initiated, what

purposes it is intended to serve, how these purposes are translated into an actual list of questions, how these questions are answered, by whom, and with what inducement. Comprehension of a survey's philosophy and implementation enables the student to recognize the limitis of its utility. For example, since surveys often require a respondent to give either oral or written responses to a number of questions, these studies are limited by their assumption that the answers given are related to facts. Many authors have questioned this premise. Thomas R. Williams, for example, states flatly that "the assumption that a reply by a respondent to a questions is 'the answer' insofar as his social behavior is concerned is fallacious."[5] Thus, when a respondent says that his home is in "generally sound" physical condition, his answer does not reflect merely the condition of his home, but rather a variety of factors which can color his viewpoint. One must be careful to avoid stating, for example, that "60% of the population live in homes which are in generally sound physical condition." Instead, he must state that "60% of the respondents rated their homes to be in generally sound physical condition." It is clear that the second of these statements is both weaker and more difficult to understand precisely.

A second limit of survey research is what Milton Rokeach has referred to as its "race-horse philosophy."[6] Since survey research is geared to quantitative analysis, many of its questions are framed in closed alternatives, each one

strictly defined, with one or another of the given alternatives sure to receive the highest number of responses. Since these alternatives are framed by the researcher, they may reflect his values. It is important that these values be understood before any analysis of the results of his research is begun.

No attempt is being made here to define all of the limits of survey research, nor was it considered necessary for the class to spend large amounts of time debating these limits. The staff wished to have students recognize, as well as understand that these limits are restrictions on a survey's usefulness. The implications of these restrictions and the exercise of caution in interpreting survey data were to be stressed.

The subject staff also considered the understanding of methods of preparing large numbers of variables and cases for analysis to be of great importance. While the actual coding of responses into lists of numbers, an important step in the analytical process, was not to be directly covered by the laboratory, it was nevertheless deemed necessary that an understanding of data preparation process be generated. Any mystery associated with the transformation of responses into statistics which spring out of a computer system on demand was to be combatted.

Another foundation the laboratory staff tried to emphasize was the knowledge of the descriptive terminology of statistics. Data analysis has its own, often confusing,

jargon which must be understood by any student who hopes to master the analysis of survey data. These terms define properties and types of data, as well as convenient statistical views which guide the way an analytical procedure will progress.

Another elemental objective of the laboratory was the development of the ability to define and distinguish among measures of statistical properties, and the skill of applying them intelligently. For each statistical property which is commonly studied in social research, there are many measures. These differ only is very subtle ways; yet it was essential that these differences be highlighted for students. There are narrow constraints which restrict the usefulness of some measures. In the hands of a student who comprehends these limits, such measures may become powerful tools. But when they are applied incorrectly, they are nonsense.

Knowledge of approaches to data analysis and logical sequences of analytical procedures was also most crucial to the successful student. It is no simple matter to determine what step should follow a step already completed. Yet the proper determination of a next step is so important to analytical procedure that without it there is a haphazard conglomeration of interesting but unrelated statistical facts which can not even be forced into a meaningful structure for understanding data.

Last among the techniques that 11.505 was to attempt to develop were the skills of clear presentation of analysis

results. The requirement of a laboratory report for each section of the workbook was intended to synthesize the student's skills in using the statistical tools into a unified analysis.

Electronic data processing was to be employed to relieve the burden of endless arithmetic often associated with statistical processes. Mastery of the available computational tools, therefore, was an important objective of the laboratory. However, the use of SPSS and IMPRESS was considered an intermediate step to the analytical end. The laboratory staff postulated the lesser importance of sophisticated programming skills and stressed the primacy of the statistical procedures and social phenomena.

It was deemed that the students should apply these statistical basics to specific problems of analysis, and that this application would provide them with an opportunity to develop a thoughtful dimension to their research. In particular, the subject staff placed a high value on giving students practice in single variable analysis, including frequency distributions, representative values, descriptive statistics and measures of variation, and in multivariate analysis, including measures of association and correlation, significance tests, test variables and causal modelling. These analytical methods were deemed of greatest utility to the social researcher, and constituted the substantive elements of the workbook exercises.

The staff planned to encourage students to develop

hypotheses throughout their work, by observing relationships between and among variables, and to consider how these hypotheses may be supported. The design and execution of tests of hypotheses was to be a primary exercise.

Beyond the development of the ability to use statistical skills in specific analytical problems, the ability to critically evaluate the analysis procedure was a prime objective. Three distinguishable, fundamental dimensions were borrowed from the work of Robert H. Ennis.[7] The first, the logical dimension, "covers judging alleged relationships between meanings of words and statements." In the context of the laboratory subject, the logical dimension was thought to include, in part, grasping the meaning and implications of a hypothesis, judging whether it is a reasonable induction from the data, whether it is really a special case of a more general hypothesis and whether certain variables take values in contradiction of each other, and identifying assumptions.

The critical dimension, understanding "the criteria for judging statements," was to encompass deciding, for example, if a hypothesis is specific enough or if an observation is reliable. The pragmatic dimension "covers the impression of the background purpose on the judgement, and. . .the decision as to whether the statement is good enough for the purpose." The laboratory staff included within this dimension evaluating whether a definition is adequate and whether conclusions are verified sufficiently by hypothesis testing.

The purpose of the development of these critical and

analytical skills was to enable the student to synthesize them into an original study focussing on a topic of interest to him. It was supposed that the ability to apply and evaluate statistical techniques would give the student the confidence to produce an original plan or proposal for analysis, implement it, and evaluate his completed work.

The staff of the urban studies laboratory anticipated that, in pursuit of these objectives, students would uncover interesting information about the populations. A considerable amount of the classroom time was set aside to offer an opportunity to share this information, and thereby to generate enthusiasm for the types of useful insights which the student could carry with him through his further study of urban problems.

Recommendations

In this section, the following recommendations are made for future improvements in the urban studies laboratory:

1. That the roles of the teaching staff be defined in a manner which will give each a responsibility for individual students.
2. That the laboratory divide regularly into small sections of about half a dozen students.
3. That the workbook exercises must be modified to eliminate any doubts in students' minds as to what specific performance is expected of them and to stress the importance of thinking about information learned from data.
4. That each student's assignments should be more closely related to the others' work in the same workbook section.
5. That exercises may be accompanied by readings in the area with which they deal.
6. That the class should have at least two meetings per week which are one hour long.
7. That more time should be allotted to independent student projects.

8. That the option of substituting another interactive system for IMPRESS should be kept open.

9. That close scrutiny and evaluation of the urban studies laboratory should continue.

Having expressed in some detail the objectives set for the urban studies laboratory, it is appropriate to proceed to some comments on how well these objectives were realized. Before this effort is begun, however, one point must be accentuated. That is, this laboratory, as any new subject with an untested format, was an experiment in a rather strict, scientific sense of that word. Those who participated in the experiment were willing to accept, when all the facts had been collected, that the laboratory was a failure and should simply be discontinued and forgotten. Too often, outside the physical sciences, there are pressures to find success in any experimental program no matter how devoid of value it may be. The fact that the subject might have been a total failure is something that can not be stressed often enough. Even as the experiment was drawing to a close, the departmental administration had determined that the laboratory was successful enough to be repeated the following year and to have a second section opened to graduate students. This decision was premature, although

in light of the evaluation which follows, it was not incorrect.

Approximately thirty-five students attended the first meeting of the urban studies laboratory, but this number was too large, both for the type of activity planned and the funds available. A short questionaire was distributed to these students, from which a class of sixteen was to be selected on the basis of the responses given. On the questionaire, students gave details of their programming and statistics background, their interest in the laboratory and some particulars about their degree program and expected graduation date. Two students from the selected group eventually chose not to continue, and two others were added because of their persistent interest in the material.

In the end, twelve urban studies students and four architecture students bacame a part of the class. Fifteen were undergraduates; one was a special student. Fifteen were MIT students; one was from Wellesley College. All fourteen MIT undergraduates stated that they planned to use the laboratory to fulfill the Institute Laboratory Requirement. Many had had at least one programming course, but four students listed no previous experience with computers. Although MIT students are expected to have familiarity with calculus, fully half of the students had had no formal statistics experience. The laboratory was designed for a "typical" student with no programming or

statistical experience and, therefore, presumed none. Still, there is some evidence that those who had additional experience benefited from it.

One way to consider the success of the laboratory subject is to observe the performance of students and note how it demonstrated their grasp of the material covered. Since the number of students was small, it may be most illustrative to consider some individuals whose performance points up particular difficulties found with the laboratory. Of course, it is important to realize that no one student is "typical" of a class as small as this one. The following studies, therefore, are not an attempt to describe such a student. In fact, the students selected are, in many aspects, actually deviant cases from the class norm. However, the cases do illustrate some weaknesses that the laboratory had. Of course, the students' names given in the studies have been changed so that they will not be identifiable.

George

George began the laboratory with somewhat less preparation than the average student, though his apparent preparedness for the subject did not differ radically from the norm. He had, for example, had some FORTRAN IV instruction which had come in a subject offered in his own architecture department. His mathematics training, though it had not gone beyond freshman calculus, was certainly near the norm for the class. However, George's problems began almost from the day he entered the class. He

experienced difficulty mastering the SPSS system, and failed to finish the first exercise of the workbook due to his inability to complete a successful run. And while the first two workbook exercises elicited some confusion among most students in the class, they were catastrophic for George.

George's main problem seemed to be an insecurity in the role of social researcher. While most students were able to note the lack of sufficient direction offered by the early exercises, criticize this failing and work around it, George could not. His work reflected not merely a failure to grasp the statistical material, but a lack of orientation towards applying this material to research. The staff of the laboratory encouraged him to proceed in directions he appeared most natural with, but his work did not seem to reflect any more understanding, nor to be structured any more coherently.

When the laboratory mid-term exam was given, considerable anxiety was expressed by the staff over how George would perform. His performance was, indeed, near the bottom of the class. But the exam provided insight into his difficulty. The exam began with a question which described a complicated sampling and polling procedure with many sources of error, then gave the results of the poll, and asked students to comment on them. Most students found the question especially easy and scored well on it. They responded that the polling procedure introduced many

biases into the results, outlined what some of them were, and then made an estimate of how reliable the poll results might be. George, however, took a different approach. He attempted to apply statistical techniques to the results, and even computed the standard deviation of response frequency. However, he remarked that he could not see how his statistical methods could be valid in light of the rather sloppy sampling process. The fact that it was precisely this sloppiness that he was supposed to document did not occur to him, since it was not explicitly stated in the question. His understanding of the procedure was sufficient for him to realize that there was something drastically wrong with it. But he was not comfortable enough with his understanding to assume to role of its critic.

George's difficulty was, in part, the result of a major fault that went uncorrected throughout much of the laboratory. This was the lack of an explicit structure which could unify and blend all of the elements of the subject into a cohesive whole. This lack was reflected in many aspects of the laboratory, and caused each to be less than completely effective. For example, the goals of the workbook exercises were something of a mystery to many students. The correspondence of lectures given in class to the exercises was less than obvious. The focus of the exercises was not placed as strongly as possible on the understanding of data, and students failed to realize the

extent to which programming facilities were an adjunct to, rather than a central activity of, the laboratory.

Much of the rest of the difficulties George experienced was the result of an inability to deal with students on as intimate a basis as should be possible with such a small ratio of students to staff. George's needs required that his difficulties be given extensive attention by a single member of the laboratory staff. But the staff, accustomed to making decisions more or less by consensus, adapted poorly in response to his obvious needs. Each staff member felt equipped to deal with only some aspects of the problem, and none felt that George was his unique responsibility. This situation resulted in George receiving very little continuing, goal-oriented assisstance.

The lack of obvious coordination among various elements of the laboratory was the most serious failure of the experiment. In addition, the staff did not achieve a proper consciousness of this lack until late in the term. This poor coordination had its roots in several aspects of the construction of the laboratory and could have been corrected through small adjustments in its constitution.

The primary source of the problem was the decision of the subject staff to divide labor by function. In the future, it will be the job of Professor Walter to define his own role and those of his teaching assistants in a manner which will give each a responsibility for individual students. Each should relate his own activities to the

totality of the laboratory, rather than limit his participation to his particular area of expertise.

For example, the author dominated the data processing services offered students, and combined with Mrs. Boyer-Karalis to assume the major responsibility for interacting with students on their written work. Professor Walter developed and gave the lectures and led most in-class discussions, and Mrs. Boyer-Karalis aided him by giving a number of lectures on special topics. Mrs. Boyer-Karalis actually completed the semester without mastering SPSS, on of our most basic tools. Yet she did not need SPSS to fulfill the role she assumed, and there was no one to say that such a relationship among members of a teaching staff was ludicrous.

In addition to defining roles across rather than along functional lines, the laboratory should take advantage of its fortunate ratio of students to staff by dividing regularly into small sections of about half a dozen students. Discussions in this small atmosphere, and consequently student involvement in the exercises would improve because students would be forced to overcome their natural reluctance to speak before a larger group. Attempts to bring the class together to examine a single cross-tabulation by means of an opaque projector failed, to some degree, because of the necessity of students' sitting in the dark and speaking in an unnaturally loud voice in order to overcome the noise generated by the projector. Smaller groups require no

projection, and thus discussion would not be inhibited by it.

A further advantage to the sectioned format of the laboratory is a real involvement between each member of the staff and the students in his section. Not only would each staff member become more capable of evaluating and helping his students, but he would be forced to deal with all aspects of the laboratory subject, not just those in which he defines his own expertise. Thus, laboratory sections would insure that a functional division of roles would be prevented.

Substantial confusion was derived from the exercises which were originally designed with something of a planned schizophrenia built into them. They made suggestions of specific exercises the student should perform with specific data, implied that work with these variables was required, and yet asked for original hypotheses and study designs. In class, comments were made to the effect that workbook exercises could be followed as closely or as loosely as one desired. The result was that students often felt that while there may have been deliberate planning on the part of the author, the schizophrenia was largely the students'.

At the end of this section are drafts of the exercises in the laboratory workbook. These drafts attempt to correct the faults which were most offensive to the students in the laboratory. The total effect of the exercises must eliminate any doubts in students' minds as to the degree to which they are to follow the instructions literally, and must encourage students to use more creativity as they progress through the

workbook. The "tooling up" exercise contained in workbook section one has been made as closely defined as possible to eliminate the considerable confusion found in the original. Although the result is a complete overhaul of the entire section, it can be represented by the change from "do something like this" to "do this." Each succeeding exercise allows the student to pursue his own purposes to a wider extent, leading to a term project at the end of the four workbook sections, in which the student may do anything that pleases him. In addition, an exercise preliminary to any dealing with programming has been added to the beginning of the workbook in order to stress the importance of thinking about information learned from data.

Mark

In contrast to George, Mark claimed no experience which would help him in the laboratory. He did not ever seem, however, to fall behind the class in understanding the material. In fact, his occasional remarks in class illustrated a rather more intimate understanding, at least of some points, than the average. However Mark seemed determined that he would not do more than the minimum required to get by in the laboratory. His unwillingness to extend a concerted effort was first noted when he was found to have used another student's work to get through the first workbook exercise. Less dramatically, all of his efforts demonstrated his lack of concern for performing beyond the minimum. His workbook effort was considerably below average,

and repeated attempts to have him put more thought into his work fell on deaf ears. This was almost literally true, because he regularly missed classes where written work was returned, and failed to pick up his work outside of class time. Nor did he seem committed to turning in assignments on time.

A particularly disturbing example of Mark's poor effort is his fourth workbook exercise. The workbook section was a short one, which asked students to observe how the relationship between two variables can be explained or affected by the introduction of a third variable into consideration. Students were expected to identify and explain intervening variables, spurious relationships, and other types of phenomena related to the introduction of test variables. The final paragraph of the assignment stated: "The association you observe in this step will likely lead you to ask about other test variables, for which you should repeat the steps you have already done. Your final product should be another complete discussion of the association between two variables and the effect test variables have on it."

Mark's work consisted of taking the relationship between two variables, income and education, controlling for race, and stating that he was unable to draw conclusions from this procedure. The association he considered did not lead him to ask about other test variables, and so the exercise was, for him, complete. It was obviously an oversimplification of the association between income and

education. Technically, however, it fulfilled the minimum requirement of the exercise.

Part of Mark's indifference stemmed from the special circumstances which induced students to enroll in 11.505. The laboratory was one of the few subjects to fulfill the Institute Laboratory Requirement, the only one from the Department of Urban Studies and Planning to do so, and an easy-sounding way to satisfy the requirement. Students who were interested only in completing requirements may, therefore, have been attracted to 11.505. And their attitudes toward the subject were not always as enthusiastic as they could have been. This effect was blunted, in part, by the inclusion of only students who expressed a specific interest covered by the subject, but even among students who stated such an interest which complemented the material very closely, the laboratory suffered from being, in a sense, required.

Mark, for example, stated in his admissions questionaire that he was working on a project "that will probably involve collection and evaluation of data." His particular interest in 11.505 stemmed from his desire to "see how the use of a computer in social science works and what its possibilities are." Yet, his attention to meeting his own objectives was disappointingly low.

More of an effort, of course, could have been made to maintain a higher level of student interest. The Institute, surely without intending to do so, cultivates a certain level

of indifference among some students toward classes which deal with mathematical or other theoretical topics. In part, the average student finds that he can not depend on lectures for his understanding of material and, thus, does not feel an obligation to attend regularly or to participate in classroom activities. This reticence can be overcome, but efforts to this end were too small to be entirely successful.

Whenever students are not encouraged to be enthusiastic, they will selectively neglect what they can. This fact has been kept in mind in the rather extensive overhaul of the workbook exercises presented in the next section of this thesis. The content of the data files has been changed so that each exercise is fresh with new variables, and so that all of the students' individual projects will be more closely related to others in order that classroom discussions can be held with wider understanding among students. Thus, each exercise focuses on a single subject. Exercise two contains a file with variables related to community identification and involvement in local activities, the third section deals with a study of Boston area housing, and the file used in section four contains variables which illuminate the concept of social class. It has been suggested that these files be accompanied by readings in the area with which they deal. While these readings are not essential to the laboratory, they can serve to guide students' work, and are certainly a valid option to give

the student. A few readings have been suggested at the end of each exercise. It should be emphasized that these exercises do not require that any readings be done, and that the readings merely offer an optional theoretical background for the student who is unsure of himself.

Restructuring class time can help to generate enthusiasm for in-class work. An attempt was made to overcome the poorly thought-out scheduling of the laboratory (two hours each on consecutive days) by having a coffee break midway through the class, by varying the focus of the lecture or discussion from one hour to the next, and by dismissing the class early when it was possible. The lesson learned is a clear one. The class should have at least two meetings per week which are one hour long. The third and fourth hours may still be consecutive if they are reserved for discussion of student findings in less formal, smaller sessions.

It was, perhaps, unfortunate that students were unable to devote as substantial a committment to their final project as would be desired. Time limitations played an important role in determining how thorough an original piece of research they could produce. Two or three weeks is certainly a small span to go from preliminary hypothesis to final conclusions. Yet the project research proved to be one of the most valuable aspects of the entire laboratory, with numerous interesting hypotheses being explored. The students accepted the challenge, and produced quite

respectable results. It was most encouraging to watch both their enthusiasm and their confidence in their role of laboratory researcher, and to judge this aspect of the laboratory experiment. Members of the class found no difficulty in applying their new expertise to questions that they had encountered in other subjects, and several students, with the support of the laboratory staff, decided to combine their project with some other work they were doing in another context. In the future, more time should be allotted to this worthwhile endeavour so that its many benefits can be adequately derived.

It was clear, althouth somewhat surprising, that the SPSS system met with considerably more student approval, and certainly was more extensively utilized, than IMPRESS, despite the fact that IMPRESS offered the student a turn-around time of zero. In part this was due to the students' introduction to IMPRESS after already being thoroughly familiar with SPSS. In addition, students found SPSS a more flexible instrument in many aspects and found the lack of an explicit programming manual a serious barrier to its effective use. Many expressed the hope that an interactive SPSS be developed in the near future, or that IMPRESS be given the data modification and data selection capabilities of SPSS, as well as the facility to handle original data files. Further, more precise evaluation of IMPRESS is impossible because of the limited use that the system received in the laboratory. However, the option of

substituting another interactive system for IMPRESS should be kept open.

In spite of the rather considerable failings in the first formulation of the urban studies laboratory, the basic framework of the laboratory model appears to be a good one. Numerous original and imaginative projects were completed by students with interesting results. Mrs. Boyer-Karalis and the author were successful in providing quick, useful feedback to students on most of their work. Student performance on the mid-term examination was encouraging, and demonstrated that considerable sophistication in data analysis was being developed. Professor Walter, not a lecturer by temperament, put forth a creditable effort, and succeeded in covering the statistical material with a minimum of verbiage, choosing not to cover topics on which the Weiss book was sufficiently explicit. Mrs. Boyer-Karalis' supplementary lectures were particularly well-organized and lucid.

Perhaps more important, and certainly more gratifying to the author, was the appreciation and approval expressed by students in the laboratory class. The subject was fulfilling a genuine need felt by many of the students, and a large fraction of them stated that they would use the skills and information learned inthe laboratory in future contexts. That is the greatest accomplishment of the laboratory.

In sum, then, the laboratory was a limited success.

The important aspect of the total experiment to keep in mind, however, is that few subjects receive the scrutiny and self-evaluation that 11.505 has had. This scrutiny should continue and may lead to other changes than are projected in this thesis. And, in time, the urban studies laboratory may evolve into a most valuable and important part of the undergraduate urban studies curriculum at MIT.

Laboratory Workbook
Introductory Section

Before you begin work in the urban studies laboratory, it is fair to let you know what you should expect from the exercises which make up the laboratory workbook. The four exercises, taken together, are meant to help you to develop skills and techniques of data analysis, creative thinking in the analysis process, and an understanding of the information contained in the data files with which you will be working.

It is important for you to spend some time at the beginning of the laboratory to consider how to structure productive thinking about data, how data can be used to answer research questions, and how data can be misused. These questions will be discussed in an early meeting of the laboratory. In order to be able to participate fully in the discussion, you should do some thinking about the following article which appeared in the Washington Post on May 6, 1971. Articles of this type are found quite often in the popular press and are a common illustration of the influence surveys have on American society.

Poll Splits Evenly on Hoover Quitting
by Louis Harris

The American people are split down the middle over whether or not J. Edgar Hoover, head of the FBI, should resign: 43 per cent think he should quit, while 43 per cent feel he should continue. The remaining 14 per cent are undecided.

The center of public controversy over Hoover is the question of his age and whether or not "he has lost his touch." By 46 to 41 per cent, a plurality feels that the FBI head is beyond his prime and is no longer doing as effective a job as could be done.

By 81 to 13 per cent, Americans agree with the statement that "the FBI has done a first rate job of protecting the security of the United States for many years."

By 71 to 14 per cent, the public agrees that "J. Edgar Hoover has done a good job in catching subversives for many years now."

By 61 to 28 per cent, people think "the FBI has done an effective job in cracking down on organized crime."

A plurality of the country, 48 per cent, regard J. Edgar Hoover as "Mr. Law and Order," although 39 per cent disagree with that description.

In addition, on two controversial issues — alleged FBI spying on public figures and Mr. Hoover's strong statements about young radicals -- pluralities of the public refuse to go along with his critics.

By 48 to 35 per cent, most do not agree with the statement that "the FBI spends too much time spying on college students, politicians, and other civilians and too little time tracking down real criminals." Sen. Edmund Muskie and House Majority Leader Hale Boggs have been particularly critical of Hoover for what they have charged is "indiscriminate FBI spying activities."

By no more than 48 to 41 per cent do people tend to disagree with the claim that "the FBI has lost most of its effectiveness in the past few years."

Finally, a carefully drawn cross section of 1508 households across the country, surveyed between April 12 and 15, was asked:

"All in all, would you rather see J. Edgar Hoover resign as head of the FBI or continue in that job?

	Resign %	Stay %	Unsure %
Nationwide	43	43	14
By Region			
East	49	39	12
Midwest	38	49	13
South	35	49	16
West	53	35	12
By Age			
Under 30	51	36	13
30-40	42	46	12
50 and over	37	48	15
By Race			
Black	46	37	17
White	43	44	13
By Education			
8th grade or less	30	50	20
High School	40	46	14
College	53	36	11

Laboratory Workbook
Section One

0. In this exercise you are asked to perform some very specific and simple analytical procedures using the Statistical Package for the Social Sciences (SPSS). From this, you will acquire experience, both in using SPSS, and, most importantly, dealing with some analytical questions that commonly confront the social scientist. Before attempting the exercise, it is important for you to read Chapters 1-5 and 16 in Weiss, Statistics for Social Research and Pages 1-128 in the SPSS Manual.

1. Read through the list of variables contained in the SPSS file COMSTUDY:

COMCOOl	Community Code
MEDSCH	Median School Years for Population over 25
MEDFINC	Median Family Income
PTGOHS	Percent Total Units Good Housing
PTAGRI	Percent Labor in Agriculture-Forest-Fishing
PTMANU	Percent Labor in Manufacturing
PTTERTRY	Percent Labor in Tertiary Industry
POP60	Total Population in 1960
WHTCOLAR	Percent Civilian Labor in White Collar Occupations
LIFE	Life Magazine Sales per 1000 Population
TIME	Time Magazine Sales per 1000 Population
NEWSWEEK	Newsweek Magazine Sales per 1000 Population
READDIG	Readers Digest Sales per 1000 Population

In this exercise, you are asked to discover something

about the relationships among magazine reading habits, education levels, economic status, population, and industry types of American small communities. In order to do this, you will probably need to make several runs using SPSS. Do not try to cram all of your thinking into a single run!

2. Group the values of variables to aid your understanding. All of the variables in the COMSTUDY file are metric measurements; that is, each has a unit of measurement (years, dollars, etc.). Sometimes, however, it is easier to understand data by grouping metric responses into categories. One way to do this is to use conceptual division boundaries. Using the RECODE feature of SPSS, group the variable MEDSCH according to these conceptual divisions:

1. Less than junior high graduate
2. Less than high school graduate
3. High school graduate

Now select a few more variables in the file. For each one define conceptual divisions, and RECODE the variables into the divisions you establish.

Another way to group responses is according to gaps in frequency distributions. Use SPSS subprogram CODEBOOK to observe the distribution of cases for the variable LIFE. Can you find patterns in the distribution which you can use in a categorization scheme? RECODE this and one or two other variables according to divisions you observed in their distributions.

For many reasons it may be useful to dichotomize

variables. Dividing populations into only two groups allows you to examine hypotheses of the form: "The presence of characteristic A implies the presence of characteristic B" and the converse. When you find a useful division which separates the population into two groups, use this division to simplify your later analysis. What are some criteria for useful cutting points?

Use VALUE LABELS to name each division. Use SPSS subprogram CODEBOOK again to find frequency distributions of each RECODEd variable.

When choosing categories for your values, remember to ask yourself: What are the meanings of the categories chosen? How can you judge the suitability of the categories for statistical analysis of a particular problem? For example, is each category sufficiently broad to contain a meaningful number of cases, or should some categories be combined to form larger ones? How helpful are your new categories? What information is lost by performing your groupings? Remember that how you form your initial categories will greatly influence the validity of your consequent analysis.

What information about the communities you are studying do the frequency distributions you have found yield? What do they leave out?

3. Establish a summary score for some characteristics which are actually made up by combining two or more variables. The simplest way to build a summary score is to add together the values of two variables. For example, an index of socio-economic status can be built by combining the variables

MEDSCH and MEDFINC. First RECODE MEDSCH and MEDFINC in the following manner:

MEDFINC	1. Less than $5000 2. $5000 to $6000 3. More than $6000
MEDSCH	1. Less than junior high 2. Less than high school 3. High school graduate

Then, using the *COMPUTE feature of SPSS, create the new variable SOCIO:

SOCIO = MEDFINC + MEDSCH

In a rather different way, an index which will show if a community is primarily agricultural, industrial or engaged in tertiary industry can be created. Use the IF feature of SPSS to create a dominant industry indicator, DOMINDUS, in the following manner:

DOMINDUS equals 1 if PTAGRI is greater than PTMANU and greater than PTTERTRY,
2 if PTMANU is the greatest of the three,
3 if PTTERTRY is the greatest.

Use CODEBOOK to observe the frequency distribution of DOMINDUS and SOCIO.

Now try to think of a system of indicators that will describe reading habits for the population. You should develop at least two summary scores. The first one should separate those communities with high sales in all magazines from those with low sales. You should also produce an index which will distinguish those communities which have high sales in one magazine from those which have high sales in another. When building these indices, remember to ask: How

do we compare sales of different magazines within a community when the average sales of each magazine for all the communities are different? Do "high" sales of one magazine mean the same as "high" sales of another? Can an index be made to reflect any differences there are in defining "high"? What do the indices you have created measure? Use CODEBOOK to observe the frequency distributions for each.

4. Create table displays to show relationships between two variables. In order to find out how readership patterns for LIFE magazine vary with median education levels, the table of LIFE by MEDSCH can be presented. You have already grouped the values for each of these variables, so it is a simple matter to use subprogram CROSSTABS to make the table. If you place the variable LIFE on the rows and consider MEDSCH as the independent variable, it makes sense to have column percentages printed.

Now, design the following tables:

PTGOHS by MEDFINC
WHTCOLAR by PTTERTRY
SOCIO by DOMINDUS

Also design and create a few tables which will show the relationships between the readership indices you have created and some other variables. When you design the tables, remember to ask yourself: Which dimension of the table should be placed along the rows? Should you present percentages as well as actual numbers? Should percentages be along columns or rows, or should they be percentages of the total number of cases? Should more than one of these

percentages be used?

5. The job control cards needed for this SPSS run are listed here:

```
// 'your name',REGION=250K,CLASS=A
/*MITID         PROB=M8170,PROG=your id number
/*SRI           LOW
/*MAIN          TIME=2,LINES=50
// EXEC SPSS
//FT03F001 DD DSNAME=USERFILE.M8170.8837.COMSTUDY.01,DISP=(OLD,KEEP)
//G.SYSIN DD*
.
.
.          spss cards go here
.
.
/*
```

In order to access the file you need, include a GET FILE card in the SPSS portion of your deck. It should read:

```
GET FILE        COMSTUDY
```

6. Here are some programming tips which will help you to complete this exercise more easily:

A. To run your SPSS program, you must have received your programmer's identification number. Once you have picked up an ID card from Prof. Walter's secretary, you are ready to begin to use the MIT Information Processing Center. The IPC is located in building 39, and all of the facilities you will need to solve most of your problems are located on the second floor. Note especially the dispatching room (39-260), where you take your programs to be processed and pick them up when they have run. Also note the keypunch room (39-233) which is where you punch up the programs you write, and the programming assistance and information (PAI) office (39-219) which is where you go when you don't know

what to do.

The IPC is open 24 hours a day, 7 days a week. However, after 6 pm on weekdays, you must enter through the ground floor entrance in the back of the Center (facing building 13). A campus patrolman will be there to check your student ID card, so be sure to bring that along.

B. For best results, you should always use the model 29 keypunch machines. All but three of the machines in the keypunch room are model 29's. They are the more modern looking of the two models. Keypunch machines look much like typewriters, and if you know how to type, you are about nine-tenths of the way toward learning to keypunch. On of the most difficult aspects of keypunching is learning to turn the machine on. The on-off switch has been carefully hidden away by your right knee as you sit at the machine.

Be sure there are enough cards in the bin at the top of the machine. If there are not, you can get more from the card rack in the keypunch room. Practice using the "feed," "release," and "register" buttons to position the cards for use. For most purposes, you will want to set the "automatic feed" switch to the "on" position. This switch is located above the keyboard in the center.

C. Some points of keypunching etiquette will help you and others to use the keypunch machines most efficiently. To begin with, do not sit at the keypunch machine composing your program or reading your SPSS manual. Others may be waiting to use the machine. Some machines are

marked "Express Keypunch: 5 cards or less." These machines are reserved to enable users with short keypunching jobs to complete them quickly. Other machines are marked "interpret keypunch." These machines are reserved for the special purpose of reading cards generated by a computer run and marking what they say along their top. You will not need to bother with these machines.

D. Once you have punched your desk, it is a good idea to make a listing of the cards using the Data 100 card lister in the room next door to the keypunch room. This machine has instructions for its use on it, and provides you with a handy method of checking to be sure you have not made keypunching errors. Remember, if you have a single word misspelled, it can mean that your job will not run!

E. When you take your deck to the dispatcher, he is supposed to ask you for your programmer's ID. You should have it ready to show him. He will take your job and give you a card which has a job number on it. You can keep track of the progress of your job by watching the queue display in the dispatcher's room, which is posted every once in a while, or by calling the dispatcher at X4121 and giving him your job number. Do not pester the dispatcher! He is a nice fellow, but he has a lot to do, so do not call him unless you have an idea that your job might possibly be done. It is reasonable to assume that your job will be processed within two hours of the time you submit it, which is marked on the card the dispatcher gave you. During the periods of

peak usage, however, your job may take longer to be processed.

7. Present your findings in a lab report. This report should contain an explanation of the procedures you followed in performing the exercises, a rationale for the steps you took, and the results of the analysis you made. Your report should neither contain reams of programming output, nor a listing of your actual SPSS program. Programming techniques are a tool for eliminating endless arithmetic, and while it is important for you to program accurately, you should be most interested in understanding and explaining what your output means.

Laboratory Workbook
Section Two

0. You will want to review pages 1-128 in the SPSS Manual before you begin work on this section of the laboratory workbook. In addition, you should read Chapters 6-8 and 12 in the Weiss book, the Boston Area Study Question Book (yellow), the Boston Area Study Master Code and the BOSAREA CODEBOOK (both on reserve at Rotch Library), and pages 129-142, 196-207 and 272-274 in the SPSS Manual.

In the exercises of this section, you are asked to consider representative values, single values that in some sense can represent the entire distribution of values as its typical, central or average value, and how much variation there is from this representative value. Some of the exercises require you simply to observe the frequency distributions of the variables. This can easily be done by reviewing the BOSAREA CODEBOOK, and there should be no need for you to depend on the SPSS system for these exercises. Remember that SPSS is primarily a tool which eliminates arithmetic for you, and you should apply it where the time needed to do calculations by hand gets to be long. You

should write a laboratory report of your hypotheses, procedures, rationale and results.

1. Select a problem area for study. In this section of the workbook, you will be working with an excerpted, weighted version of the large BOSAREA file which contains about thirty variables related to community identification and involvement in community activities. Also included are background information about respondents, their attitudes, education, occupation and income levels.

The weighting of this file was accomplished using the following cards:

```
IF          (PLACE LT 200) WTFACTOR = HHADULTS
IF          (PLACE GE 200) WTFACTOR = 9*HHADULTS
WEIGHT      WTFACTOR
```

This weighting procedure was performed because the BOSAREA file under-samples the suburbs by a factor of nine, and under-samples all households with more than one adult by a factor equal to the number of adults. For some purposes, you may want to alter the WEIGHTs that have been assigned. This can be done by defining a completely new procedure similar to the one used here, or by modifying the presently assigned WEIGHTs. The WEIGHT assigned each case is stored under the variable name CASWGT. This variable can be modified using any of the data modification features of SPSS.

Here is a complete list of the variables available for this exercise:

```
PLACE     Place of Residence
PER2SEX   Sex of person 2
PER2AGE   Age of person 2
```

LIVBOSAR	1. Years living in Boston area
LIVNEIGH	4. Years in city, town or part of Boston
PER2STAT	Marital Status of person 2
WHYLIVES	8. Why R lives in nborhood in general
NBRPART	11. Does R feel part of neighborhood?
SUMNOASS	15. Number of neighborhood assets
NOASSETS	15. Type of neighborhood assets named
MOVEPROB	18. How likely is R to move?
CIVICLUB	21. Does R belong to civic clubs?
NOCIVIC	22. Number of civic clubs R belongs to
TPCIVIC	22. Type of civic clubs R belongs to
SOCICLUB	23. Does R belong to social clubs?
NOSOCIAL	24. Number of social clubs R belongs to
NOCLUBS	21-24. Number of clubs R belongs to
RELISERV	85. How often R attends rel services
BELONCH	86. Does R belong to church or synagogue
GENERATI	94-99. Summary score of generation
RACE	100. Race
NATIONAL	103. Nationality
NAFRIEND	106. No of friends with same nationality
WIFECHIL	121. Number of wife's children
MINCHILD	122-124. Min no more children expected
MAXCHILD	122-124. Max no more children expected
HHADULTS	Number of adults living in household
REDUCATE	133. R's education
HEDUCATE	138. Head's education
HEADWORK	143. Head's occupation
HEADPATT	146-147. Head's work pattern over year
WELFARE	163. Did R receive welfare, AFDC, unem?
FAMINC	168. Family income

From this list of variables, you should define a problem which you would like to study, and develop one or a few hypothesis about this problem that can be understood from a study of the variables in the file. For example, you may wish to compare levels of involvement in religious activities for various income groups, education levels, etc. You should begin by stating your hypothesis very explicitly. This will help you to build your study thoughtfully and systematically.

2. Find a representative value for the variables you study. Three different measures are commonly used to define a representative value for a distribution. The mode uses a nominal level of measurement, and is the category which contains the highest number of responses. The median requires at least an ordinal level of measurement, and is defined as that score which is larger than or equal to the value of half the scores and smaller than or equal to the other half. The arithmetic mean requires metric measurment, and is the sum of the scores of a variable divided by the total number of valid cases for that variable.

For certain of your variables, you will want to find the mode. Why is the mode an appropriate representative value for these variables? What is the meaning of the mode of each variable? What information does this value tell about the data?

For other variables, you will be able to compute the median and you can compare the median to the mode. Which seems to be a preferred representative measure for your ordinal variables? It is sometimes useful to interpolate the median for grouped ordinal measures. This interpolation is based on the assumption that the cases are distributed evenly within a category. Is this assumption likely to produce a gross error for the variables? Why or why not?

For metric measurements, compute the mean. Using the RECODE feature of SPSS, group the data for these variables and estimate the mean of the grouped responses. How do

these two values of the mean compare for the variables? How do you find the mean for variables with open-ended categories such as "6 months or more"? Some ordinal measures are actually metric measures which have been grouped before inclusion in the BOSAREA file. You can estimate a mean for these variables by using the RECODE feature of SPSS to RECODE the categories to their meanings, choosing a representative value for each category. How accurate a procedure do you think this is? Why?

Create a short presentation using representative values to substantiate or refute your hypotheses. What is the best representative value which summarizes your presentation? Why? Has your analysis verified your hypotheses?

3. Measure variation from the representative values you have found. At times you will be dealing with distributions where the values are about the same for all cases. At other times the distributions fluctuate quite a lot across values. Variation is a measure of this fluctuation.

For example, in the COMSTUDY file, you found this distribution for the variable DOMINDUS:

1. Agricul,For,Fish	7	10.9%
2. Manufacturing	15	23.4%
3. Tertiary	41	64.1%
9. NA	1	1.6%
	64	100.0%

For this nominal data, it is clear that category 3 (Tertiary) is the modal response, and since nearly two thirds of the distribution falls into this category, it is safe to say that it is a typical response. One could say, "about two-thirds

of the communities are dominated by tertiary industries, and the other third are dominated by other industry types."

For your nominal measurements, how much of the distribution falls into a modal category? What is the typical response, and how much variation is there from the typical? Perhaps the typical category is the one with the highest number of responses, but what if no category has a significantly higher number than the others? What grouping of categories reveals a typical response and what percentage accounts for "all others"?

Now consider another variable from the COMSTUDY file, XMEDFINC. The combination

```
COMPUTE          XMEDFINC = MEDFINC
RECODE           XMEDFINC (1 THRU 4000=1) (4000 THRU 5000=2)
                 (5000 THRU 6000=3) (6000 THRU 7000=4)
                 (7000 THRU HIGHEST=5)
MISSING VALUES XMEDFINC (0)
VALUE LABELS     XMEDFINC (1) $4000 OR LESS (2) $4000 TO $5000
                 (3) $5000 TO $6000 (4) $6000 TO $7000
                 (5) $7000 OR MORE (0) MISSING
CODEBOOK         XMEDFINC
```

yields the following frequency distributions:

1.	$4000 or less	9	14.1%
2.	$4000 to $5000	16	25.0%
3.	$5000 to $6000	17	26.6%
4.	$6000 to $7000	12	18.8%
5.	$7000 or more	9	14.1%
0.	Missing	1	1.6%
		64	100.0%

A quick glance at the distribution shows that the median falls in category 3. By interpolation, a median of $5412 can be estimated. One can say that "among the communities the middle 70% have median family incomes between $4000 and $7000." This is a statement describing the variation from

the median response. That is, about 15% are above and 15% are below this middle range.

For your ordinal measurements, between what values in the range of data does a given percent of the data fall? What is the best range of data to consider? Should the complete range be considered, or the middle 80%, the middle 60% or the middle 50% (the interquartile range)? What are the extreme values of each of these ranges? Interpolation is valid here; use the same method of interpolation that you used to compute the median.

Finally, consider the variable POP60. Its mean is 13450.453. To measure the variation from the mean, the standard deviation is a measure of variation based on the distance each case is from the mean. Its formula is:

$$\text{standard deviation} = \sqrt{\frac{\Sigma(x - \bar{x})^2}{N}}$$

The standard deviation of POP60 is 14235.031. You can get a feel for the large amount of variation in POP60 by noting that the standard deviation is larger than the mean!

For your metric measurements, measure variation by computing the standard deviation. Then define an intermediate range and find what values fall within it. Compare these two methods of measuring variation. Which measure expresses variation better? Why?

Create a short presentation which expands on the conclusions you drew in your analysis using representative values. Compare the variation in variables for the various populations you are studying. You may want to use SPSS

subprogram CODEBOOK, FASTABS or BREAKDOWN in your analysis. What do your measures of variation mean for each variable? What does this method of analysis say about variation among various subpopulation groups? Which groups have the greatest variations, and how are these variations best expressed? Why do you think so?

4. In order to complete this section, you will need to access the excerpted weighted BOSAREA file with the identification variables in it. This file can be gotten by using the following JCL card:

```
//FT03F001 DD DSNAME=USERFILE.M8170.8837.IDENT.1969,DISP=(OLD,KEEP)
```

The first card in the SPSS part of your deck should be:

```
GET FILE        BOSAREA
```

5. A few readings which may help you to form your hypotheses are listed below. These readings are optional, and are merely a representative sample of the available literature. However, other literature would be equally appropriate:

Litwak, Eugene; "Voluntary Associations and Neighborhood Cohesion"; American Sociological Review, Vol. 25 (Feb. 1960) pp. 258-271

Stein, Maurice R.; The Eclipse of Community: An Introduction of American Studies; Princeton: Princeton University Press 1960

Seeman, Melvin; "On the Meaning of Alienation,"; American Sociological Review, Vol. 24, No. 6 (Dec. 1959)

Hillery, George A.; Communal Organizations; Chicago: University of Chicago Press 1968

Laboratory Workbook
Section Three

0. Some readings will help you to complete the exercises in this section of the workbook more easily. You definitely should read Chapters 9-11, 13 and 14 of the Weiss book, and pages 143-146 of the SPSS Manual.

1. Measuring association and correlation between pairs of variables is the major focus of this exercise. You are asked to consider the relationships among variables contained in a file drawn from the BOSAREA file. These variables were selected because they relate to Boston area housing. In addition to measuring association, you are asked to make a judgement about whether the associations you measure are significant. The measurements you make should enable you to state with a known degree of certainty some conclusions about housing in metropolitan Boston. Your work should be presented in a laboratory which contains: (1) a statement of the hypotheses you chose to study, (2) some justification of each, (3) a description of methods to test each, (4) the results of performing these tests, and (5) some explanatory comments on the results.

Remember that you have two sets of goals: (1) learning about housing phenomena, and (2) learning about association, correlation and significance tests. Each goal should affect the way you proceed through this exercise and guide you in selecting the steps you follow in it.

2. The file with which you will be working has not been weighted because so many of its variables deal with households, rather than individuals. You may want to WEIGHT the suburbs by a factor of 9, however, if this seems appropriate to you. The variables contained in the file deal with various aspects of housing: reasons for living in the neighborhood, or in the particular house, the amount of space in the home, what its cost is, rating of local services, and opinions related to open housing for blacks, as well as some background information about the respondents. The variables are:

NOSCOUNT	Number of people living in household
HHMINORS	Number of minor children in household
HEADSEX	Sex of Head
HEADSTAT	Marital Status of Head
ISRHEAD?	Is R Head?
PER2AGE	Age of person 2
LIVNEIGH	4. Years in city, town or part of Boston
WHYLIVES	8. Why R lives in nborhood in general
RACECOMP	14. Racial composition of neighborhood
CLOSWORK	15. Is neighborhood close to work?
CLOSCHOL	15. Is neighborhood close to schools?
CLOSSHOP	15. Is neighborhood close to shopping?
CLOSTRAN	15. Is nborhood close to transportation?
CLOSRELA	15. Is neighborhood close to relatives?
CLOSFRND	15. Is neighborhood close to friends?
HOMOGEN	15. Do people like R live in the area?
NBRFREIN	15. Are R's neighbors friendly?
GOODHOUS	15. Has neighborhood good housing?
INEXHOUS	15. Has nborhood inexpensive housing?

MOVEPROB	18. How likely is R to move?
WHERLOOK	19. Where R would look for another home
WHYCHOSG	25. Why R chose his home in general
HOUSECON	26. Condition of R's home
NATUPROB	27. Nature of condition prob in R's home
CENTHEAT	28. Does R's home have central heat?
NOHOUSRM	29. Number of rooms in R's home
RMPERPER	29. Rooms per person is R's home
NOBDROOM	30. Number of bedrooms in R's home
BDRPRPER	30. Bedrooms per person in R's home
HOMESAT	31. Satisfaction with home
OWNHOME	32. Does R own his home?
TOTORENT	33. Rent plus utilities for R's home
SALPRICE	34. Sale price of R's home
TOTOCOST	34. Total cost per month of R's home
HOUSCOST	33-34. Cost of R's h me as % of income
HOUSMUCH	35. Does R pay too much for housing
BUYHOUSE	39. Would R like to buy house
NOBUYWHY	40. Why R would not like to buy house
TRASHCOL	44. Rating of trash collection
PLAYGRND	45. Rating of parks and playgrounds
POLICE	46. Rating of police
PUBSCHOL	57. Rating of public schools
RACE	100. Race
BLACKOUT	112. Opinion: keeping blacks out
BLACPUSH	113. Opinion: blacks should not push
MORALE	Morale
STRUCTYP	Type of structure where R lives
PUBLICHO	Is R in public housing?
QUALITHO	Quality of R's housing

This file can be accessed by inserting the following card into your control deck:

```
//FT03F001 DD DSNAME=USERFILE.M8170.8837.HOUSING.1969,DISP=(OLD,KEEP)
```

You should use the same GET FILE card you used in section two.

3. For nominal data, some common measures of association are the percentage difference, lambda, and measures based on the chi-square statistic. Useful significance tests often applied are the Fisher Exact Test, and a test based on the chi-square statistic. These measures are outlined and

illustrated here, and you should select from them according to your particular needs. It is important to consider why you use a particular measure or test.

Two qualities are associated when the distribution of values of the one differs for different values of the other. A very simple asymetric measure of association for the two-by-two case is the comparison of the percentages on different columns of the same row category, or in different rows of the same column category.

Choosing an example from the COMSTUDY file, a table of WHTCOLAR by PTTERTRY can be made using the following statements:

```
RECODE          PTTERTRY (LOWEST THRU 40=1)(40 THRU HIGHEST=2)
RECODE          WHTCOLAR (LOWEST THRU 40=1)(40 THRU HIGHEST=2)
VALUE LABELS    PTTERTRY, WHTCOLAR (1) LOW (2) HIGH
FASTABS         VARIABLES PTTERTRY, WHTCOLAR (1,2)/
                TABLES WHTCOLAR BY PTTERTRY
OPTIONS         5
```

The printed output of such a table would look like this:

			PTTERTRY		
		COUNT ROW PCT COL PCT	LOW 1.	HIGH 2.	ROW TOTAL
WHTCOLAR	LOW	1.	17 85.0% 53.1%	3 15.0% 9.4%	20 31.3%
	HIGH	2.	15 34.1% 46.9%	29 65.9% 90.6%	44 68.7%
		COLUMN TOTAL	32 50.0%	32 50.0%	64

Now it is a very simple procedure to compare the row precentages for LOW and HIGH values of WHTCOLAR. The

difference is:

$$85.0 - 34.1 = 65.9 - 15.0 = 50.9\%$$

The column percentages for LOW and HIGH values PTTERTRY can be compared just as easily:

$$53.1 - 9.4 = 90.6 - 46.9 = 43.7\%$$

Now consider the variables in the BOSAREA housing file. To compare the percentage differences, use FASTABS to create tables of variables which you have dichotemized in the manner of the example. Percentage the tables along their rows and columns, and find the precentage differences along each. What do these values explain about the direction of association between the two variables?

Now consider the sample table again. If one were to guess whether a community in the COMSTUDY were included in the HIGH category for WHTCOLAR, and he had no other information about the particular community, he would be smart to guess that the community was included. The table shows that 68.7% of the communities do have HIGH values of WHTCOLAR. But suppose that one other bit of information was known, namely that the community has a LOW value for PTTERTRY. Then the smart guesser would give a different answer, because 53.1% of the communities with LOW values of PTTERTRY also have LOW values of WHTCOLAR.

In the first case, where there was no additional information, the smart guesser would have been wrong 20/44 (31.3%) of the time. But, if he were told whether the value for a particular community of PTTERTRY was known to be either

HIGH or LOW, the smart guesser would have improved his probability of being correct. He would have been wrong 15/32 (46.9%) of the cases with LOW values of 3/32 (9.4%) of the cases with HIGH values. That is, the smart guesser would have been wrong in 18 of the 64 cases. Lambda asymmetric (which is referred to in the Weiss book as "g") measures the improvement in guessing that would result from knowing this extra fact. Its formula is:

$$\text{lambda asymmetric (row/column)} = \frac{\sum_{\text{over all columns}} \left(\text{largest cell frequency in column} - \text{largest row total}\right)}{N - \text{largest row total}}$$

In a similar manner, the formula which will measure the improvement in guessing PTTERTRY that is derived from knowing a value for WHTCOLAR, can be written:

$$\text{lambda asymmetric (column/row)} = \frac{\sum_{\text{over all columns}} \left(\text{largest cell frequency in row} - \text{largest column total}\right)}{N - \text{largest column total}}$$

For the table WHTCOLAR by PTTERTRY, both of these can be computed:

$$\text{lambda asymmetric (row/column)} = \frac{17 + 29 - 44}{64 - 44} = \frac{2}{20} = .100$$

$$\text{lambda asymmetric (column/row)} = \frac{17 + 29 - 32}{64 - 32} = \frac{14}{32} = .438$$

It is clear that knowing something about the value of WHTCOLAR for a community is a great aid in guessing PTTERTRY. The association is not as strong in the reverse direction.

You may use the lambda measure of association for tables of any dimension, not just the two-by-two case required for use of percentage differences. For some of the tables for which you use the percentage difference, you may want to compare lambda to it. Remember to ask yourself which of these measures presents a better picture of the association which you measure.

Now, returning to the sample table of WHTCOLAR by PTTERTRY, try to imagine what the frequency count in each cell would have been if there had been no association between the variables at all. These expected frequencies would have made up a table which looks like this:

			PTTERTRY		
		COUNT ROW PCT COL PCT	LOW 1.	HIGH 2.	ROW TOTAL
WHTCOLAR	LOW	1.	10 50.0% 31.3%	10 50.0% 31.3%	20 31.3%
	HIGH	2.	15 50.0% 68.7%	29 50.0% 68.7%	44 68.7%
		COLUMN TOTAL	32 50.0%	32 50.0%	64

The chi-square statistic is a measure which compares the frequency count of each cell in a table with the frequency count of each cell in a table of expected values such as this one. It is not by itself a measure of association. However, several measures of association including phi, Cramer's V and the contingency coefficient, C, are based on chi-square. Its formula is:

$$\text{chi-square} = \sum \frac{(o - e)^2}{e}$$

where o = observed frequency for each cell
e = expected frequency for that cell

From this information, it is easy to find:

$$\text{chi-square} = \frac{(17-10)^2}{10} + \frac{(3-10)^2}{10} + \frac{(15-22)^2}{22} + \frac{(29-22)^2}{22}$$

$$= 14.255$$

The following definitions show the relationship between some measures of association and the chi-square statistic:

$$\text{phi} = \sqrt{\frac{\text{chi-square}}{n}}$$

Phi is a symmetric measure of the extent to which a two-by-two table displays mutual association. It will be zero when the observed table is identical to that expected on the assumption of independence, and one when chi-square reaches its maximum value, which is n. You will want to compare phi to the percentage difference. Both are measures used only for two-by-two tables, but phi is symmetric, while the percentage difference is directional. As you work, remember to ask yourself which of these two types of measure best meets your requirements. What might you consider in deciding the answer to this question?

$$\text{Cramer's V} = \sqrt{\frac{\text{chi-square}}{n(\text{MIN:}(c-1)(r-1))}}$$

where MIN:(c-1)(r-1) is the number of columns or the number of rows in the table (whichever is less) minus 1.

This formula adjusts phi for the number of rows or the number of columns, depending on which is smaller. For tables

larger than two-by-two, phi has no upper limit, and therefore should not be used. V, however will always range from 0 to 1.

$$C = \sqrt{\frac{\text{chi-square}}{\text{chi-square} + N}}$$

The contingency coefficient should only be used to compare tables which have the same dimensions, and is only useful as a measure compared to other C's. This is because its upper limit is a function of the dimensions of the table. It can be interesting to compare two table using a number of different measures and to note how these measures emphasize or mask the differences that exist. Remember to consider how each measure you use is defined, and to note how the difference in each one's definition is reflected in the comparison you make.

In addition to being a basis for several measures of association, the chi-square statistic can also be used as a basis for a test of the significance of a distribution. A significance test measures the probability that evidence as strong or stronger than that observed to support the alternative hypothesis could have occurred by chance.

It is often useful, when measuring association to test a hypothesis, to make explicit two hypotheses which between them represent whatever the data could possibly show. The first of these is the hypothesis of interest, or alternative hypothesis which expresses the relationship that is being tested. The other is the tested hypothesis, or null hypothesis,

which expresses the lack of the relationship expressed in the hypothesis of interest. For example, one hypothesis of interest which could have been drawn from the COMSTUDY file is that communities with a high percentage of tertiary industry also have high percentages of workers in white collar jobs The tested hypothesis that goes with this is that communities with a high percentage of tertiary industry do not have an especially high percentage of workers in white collar jobs.

The level of significance of a relationship is the probability of rejecting the tested hypothesis in favor of the hypothesis of interest when the tested hypothesis is, in fact, the correct one.

In order to compute the probability of a frequency distribution yielding a chi-square as high or higher than the one observed assuming the tested hypothesis to be true, one must first determine the number of degrees of freedom associated with a table. This number is equal to the number of cells in the table where observed counts are not determined be frequency counts in other cells. For example, if the following two-by-two table is given:

a	b	r1
c	d	r2
c1	c2	n

it is possible to find b, c, and d if a, r1, r2, c1 and c2 are known. The second column is determined once the first is known since it must be equal to the marginal minus the first column count for each row. In a similar manner, the second row is

determined once the first is known since it is equal to the marginal minus the first row for each column. Therefore, all the cell frequencies can be deduced from the cell frequency of one cell, and the two-by-two table is said to have one degree of freedom. In general, the number of degrees of freedom is given by the formula:

$$\text{degrees of freedom} = (r-1)(c-1)$$

where r = the number of rows in the table
c = the number of columns in the table

The distribution of chi-square is a function of the number of degrees of freedom. It is different with one degree thatn with two, and so on. For this reason, it is necessary to consult a table of significance levels associated with the chi-square statistic, such as the one printed in the back of the Weiss book, to determine the probability of obtaining a chi-square statistic as good as the one observed assuming the tested hypothesis to be true.

For your nominal data, you will want to consider computing the probability of achieving a particular distribution, or a stronger one when the tested hypothesis is true. For the two-by-two case, this procedure is not very difficult. The probability of a given distribution is given by the formula:

$$p = \frac{(r1!)(r2!)(c1!)(c2!)}{(n!)(a!)(b!)(c!)(d!)}$$

It is often easy to write down all of the two-way distributions which support the alternative as much or more than the given one. For the WHTCOLAR by PTTERTRY example, for instance, the distributions would be:

17	3	20	18	2	20	19	1	20	20	0	20
15	29	44	14	30	44	13	31	44	12	32	44
32	32	64	32	32	64	32	32	64	32	32	64

The Fisher Exact Test simply computes the probability of each of these distributions, assuming the tested hypothesis to be true, and then sums these probabilities. The result is the probability of evidence as inconsistent or more with the tested hypothesis, assuming it is true.

For your nominal data, you will want to choose from among these measures of association, and from the various significance tests. It is important for you to be conscious of the reasons you choose to measure association or test significance in a particular way. Be ready to defend the criteria you use. What do the values of the association measures explain about the relationship between the variables? With what level of significance in your hypothesis confirmed? Is this level sufficient to give you confidence in rejecting the tested hypothesis?

4. <u>For ordinal data</u>, it is often most convenient to speak of correlation rather than association. This term describes two variables increasing or decreasing together, rather than merely ocurring together. It is also possible to speak of negative correlation, which occurs when one variable increases while the other decreases.

Several useful measures of correlation involve the concept of positive and negative pairs. This concept is best understood by considering an example, such as the COMSTUDY variables WHTCOLAR and PTTERTRY. Consider the following

three cases:

COMCOOL	PTTERTRY	WHTCOLAR
72	47.1	40.8
82	35.0	32.7
123	45.1	25.9

This table says, for example, that Community #72 has 47.1% tertiary industry and 40.8% of its labor force in white collar jobs. Similar information is given for the other two communities. The three communities can be ranked according to their values for each variable. Community #72 is first in both variables, but Community #82 is second in WHTCOLAR and third in PTTERTRY:

COMCOOL	PTTERTRY	WHTCOLAR
72	1	1
82	3	2
123	2	3

Now it is possible to compare each pair of communities, and to classify each pair as positive or negative. A positive pair is one in which the values of each variable for one of the pair are greater than the values of each variable for the other. In other words, if the cases were ranked according to the values of each variable, a positive pair would be one in which one of the cases ranked higher than the other on both lists. Pairs which have the same value of one or both variables are considered ties. All other pairs are negative.

For the three communities listed, it is clear that Communities #72 and #82 form a positive pair because 47.1 is greater than 35.0 and 40.8 is greater than 32.7. Communities #82 and #123 form a negative pair because 45.1 is greater than

35.0, but 25.9 is less than 32.7. If P is the number of positive pairs and Q is the number of negative pairs, then

$$S = P - Q$$

For the example given, P is 2 and Q is 1. Therefore S is 1 for this three case sample. The procedure outlined can be applied to all measurements which are at least ordinal, and that, while the counting procedures may be complicated, it is valid regardless of the number of categories or cases being processed.

Of itself, S is not a measure of correlation. To be sure, it is near zero for low correlation, but how large it gets is closely related to the number of cases. However, some simple measures can be constructed from S. The simplest is the Goodman-Kruskal gamma which divides S by the total number of positive and negative pairs:

$$\text{gamma} = \frac{S}{P + Q}$$

A second measure of correlation involving S is Kendall's Tau_b. Tau_b compares S to an estimate of what the total number of positive or negative pairs would be if there were the same marginal values both in rows and columns and also complete mutual association between the attributes. Its formula is:

$$\text{Tau}_b \quad \frac{S}{\frac{1}{2}(n^2 - \sum c^2)^{\frac{1}{2}}(n^2 - \sum r^2)^{\frac{1}{2}}}$$

where $\sum c^2$ and $\sum r^2$ equal the sum of the squares of the column and row totals, respectively.

You will want to use either gamma or Tau_b to measure

association for your ordinal variables. Which seems to be a better measure of correlation? How would you make this judgement?

The test of significance associated with gamma and Taub, involves making an estimate of the probability of achieving a value of S at least as large as the value observed, assuming the tested hypothesis to be true. Simple observation will show that the distribution of S when the tested hypothesis is true will have a mean of zero, and will be normal about this mean. Thus, if the standard deviation of S is known, a measure of the probability of obtaining at least a certain S can be had by comparing this value of S to the standard deviation. Fortunately, the SPSS system does all of this, and prints out the results of the significance test along with Taub. It should be noted that the same level of significance is associated with gamma.

5. For metric data, one of the most common methods used to measure correlation is to compute the Pearson product-moment correlation coefficient:

$$r \quad \frac{(x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \quad (y - \bar{y})^2}}$$

If your analysis includes metric variables, use SPSS subprogram PEARSON CORR to compute r for them.

6. Three suggested readings which might help you to form your hypotheses are:

President's Committee on Urban Housing; A Decent Home; Dec. 1968 (Government Printing Office: 1969 0-313-937)

National Commission on Urban Problems; Building the American City; 1968 (House Document No. 91-34)

Rossi, P.H.; Why Families Move; Glencoe: The Free Press; 1955

Laboratory Workbook
Section Four

0. Readings that will be useful to you in this exercise are Lazarsfeld, "Interpretations of Statistical Relations as a Research Operation" (distributed in class) and Chapters 6 and 7 of Hyman, Survey Design and Analysis. You may also want to skim Rosenberg, The Logic of Survey Analysis. The latter two are on reserve in the Rotch Library.

1. The file with which you will be working in this exercise has been drawn, as before, from the large SPSS file BOSAREA. It has been left unweighted so that each student can handle this problem for himself. The variables were selected because they may help to illuminate a variety of aspects of the concept of social class. The variables in the file are:

PLACE	Place of residence
NOSCOUNT	Number of people living in household
HEADSEX	Sex of head
HEADSTAT	Marital status of head
PER2AGE	Age of person 2
WHERHDSP	Where Head's spouse is
LIVBOSAR	1. Years living in Boston area
LIVNEIGH	4. Years in City, Town or part of Boston
RACECOMP	14. Racial Composition of neighborhood
GENERATI	94-99. Summary of Generation
RACE	100. Race

NATIONAL	103. Nationality
ETHSCORE	105-108. Ethnicity score
MARIWIFE	120. Year wife first married
WIFECHIL	121. Number of wife's children
RTOHEAD	143. Relationship of R to Head
WHYLIVES	8. Why R lives in nborhood in general
MOVEPROB	18. How likely is R to move?
NOCLUBS	21-24. Number of clubs R belongs to
HOUSECON	26. Condition of R's home
HOMESAT	31. Satisfaction with home
BUYHOUSE	39. Would R like to buy house?
OCCUPAPA	42. Occupation of Head's father
HOLSCOPA	42. Hollingshead Occupation Score of father
REDUCATE	133. R's education
HEDUCATE	138. Head's education
HEADWORK	143. Head's occupation
HEADHOLL	143-151. Hollingshead Score of Head
HEADTOHO	143-151. Head's weighted Hollingshead score
HEADHOSO	143-151. Head's Hollingshead Class
OCCUPATE	143. Occupation
HOLLINGS	143-151. Hollingshead occupation score
TOTAHOLL	143-151. Total Hollingshead Score
HOLLISOC	143-151. Hollingshead Social Class
FAMINC	168. Family income
WHOBLAME	55. Who is to blame for city problems
CANHELP	56. Who can help cities most
CITSERV2	44-46,57. Rating of city services
RELIMPOR	89. Importance of Religion
CITHELP1	109. Whom city government helps most
CITHELP3	109. Whom city government helps least
PROINTEG	73,112,113,114. Pro-integration score
JOBSATIS	162. How satisfied Head is with job
SPENJOY	174. Opinion: Spend today and enjoy
ACASCOME	175. Opinion: Accept things as they come
SLFORGOV	176. Opinion: Rely on Self or Government
LOTOFMAN	177. Opinion: Lot of Man is worse
TYPROBLE	179. Most important problem type
DOESPLAN	180. Does R plan?
STRIVONE	174-176,180. Striving Index Number 1
CHANLIFE	181. How R would change his life
SATISCAL	17,31,162,169. Satisfaction
MORALE	Morale
BESTHING	178. Best thing about R's life

2. In this exercise, you are to use skills that you have already developed to look at the way one variable affects the relationship between two others. There are a number of ways that this can happen:

First, an observed association may disappear when values of a test variable are specified. Such a relationship may be considered spurious, such as the example offered by Lazarsfeld; the correlation between stork population and birth rate is found to be spurious when a level of urbanization is specified. Or the relationship may be said to be explained by an intervening variable. An example of such a relationship is that between the comin of May and the increase in usage of MIT's Information Processing Center. This relationship is explained by controlling for the number of assignment due dates per day.

Second, an observed relationship can be specified by introducing a test variable as a condition. One might, for example, observe a correlation between income and attendance at classical theater productions. If one were to control for education, however, one might find that this relationship only exists for those with college degrees. The test variable specifies under what conditions the correlation is observed.

Third, an expected relationship which was not observed may be found to exist when a test variable is introduced. For example, one might be very distressed to find no association between practicing family planning and earning a certain income. The relationship may not be noticed until

education is introduced into the analysis. Rich plumbers are then separated from college professors, and the association is observed.

3. Pick and independent and a dependent variable between which you expect to observe an association. Think of test variables that you might expect to affect the association. Observe and measure the association between the test variables and the dependent and independent variables you chose and between the original pair, controlling for the test variable. You will want to apply significance tests to determine if the association you observe could have occurred merely by chance.

The association you observe in this step will likely lead you to ask about other tests variables, for which you should repeat the steps you have already completed. Your laboratory report should be a rather complete discussion of the association between two variables and the effect test variables have on it, and should reflect an understanding of the Lazarsfeld and Hyman materials.

4. To access the file for this exercise, the following JCL card must be used:

```
//FT03F001 DD DSNAME USERFILE.M8170.8837.CLASS.1969,DISP (OLD,KEEP)
```

As in previous exercises, your GET FILE card should contain the name BOSAREA.

5. The literature on social class is quite extensive, but is you are unsure about how to find some material from which to build a hypothesis, here are a few especially interesting sources:

Banfield, Edward C.; The Unheavenly City; Boston: Little, Brown; 1970

Suttles, Gerald; The Social Order of the Slums; Chicago: University of Chicago Press; 1968

Cohen, Albert K. and Hodges, Harold, Jr; "Characteristics of the Lower-Blue-Collar-Class"; Social Problems, Vol. 10, No. 4 (1963) pp. 303-334

Hollingshead, A.B.; Elmtown's Youth; New York: John Wiley; 1967

Gordon, Milton M.; Social Class in American Sociology; New York: McGraw-Hill; 1963

Notes

1. New York: John Wiley & Sons, Inc.; 1968

2. Nie, Norman H., Bent, Dale H., and Hull, C. Hadlai; Statistical Package for the Social Sciences; New York: McGraw-Hill; 1970

3. Joint Center for Urban Studies; How the People See Their City: Boston, 1969; Cambridge: Joint Center for Urban Studies, 1970

4. Meyers, Edmund D., Jr.; "Project IMPRESS: Time-Sharing in the Social Sciences" in AFIPS Conference Proceedings, Vol. 34 (1969) Spring Joint Computer Conference, Boston, Mass.; Montvale, New Jersey: AFIPS Press. See also: Davis, James A. and Sternick, Joanna H.; "The IMPRESS Primer," unpublished paper, February, 1971

5. Williams, Thomas R.; "A Critique of Some Assumptions of Social Survey Research"; Public Opinion Quarterly; Vol. 23 (Spring '59) page 57

6. Rokeach, Milton; "The Role of Values in Public Opinion Research"; Public Opinion Quarterly: Vol. 32 (Winter '68-'69) page 548

7. Ennis, Robert H.; "A Concept of Critical Thinking: A Proposed Basis for Research in the Teaching and Evaluation of Critical Thinking Ability" in B.P. Komisar and C.B.J. MacMillan (eds.); Psychological Concepts in Education; Chicago: Rand McNally & Co.; 1967; pages 117-8