DOCUMENT RESUME

ED 194 669 UD 021 007

AUTHOR	Leinhardt, Samuel; Wasserman, Stanley S.
TITLE	Quantitative Methods for Public Management, Package XVI. Introductory Manual; Module II, Revised; Module III, Revised; Module IV, Revised.
INSTITUTION	Carnegie-Mellon Univ., Pittsburgh, Pa. School of Urban and Public Affairs.
SPONS AGENCY	Department of Housing and Urban Development, Washington, D.C. Office of Policy Development and Research.; National Training and Development Service for State and Local Government, Washington, D.C.
PUB DATE	[77]
NOTE	1,157p.; Figures may be illegible due to small, broken print. For related documents see UD 020 926-937 and UD 021 002-009.
EDRS PRICE	MF09/PC47 Plus Postage.
DESCRIPTORS	*Computer Oriented Programs; *Data Analysis; Data Processing; Instructional Materials; Local Government; Postsecondary Education; *Public Administration Education; *Statistical Analysis; Urban Areas

ABSTRACT

This package of modules comprises a portion of the National Training and Development Service Urban Management Curriculum Development Program. The four modules included in the package contain instructional materials covering a broad range of statistical and data analytic procedures chosen on the basis of their probable utility to public managers and administrators. The materials emphasize graphics, robust procedures, model development, and the evaluation and critique of analyses. Each module contains instructor's materials, lecture outlines and supplementary materials; students' materials, reading assignments, exercises, and exams. A computer software system is available for performing data analysis on numerical files. The modules are designed to be used as an entire sequence for one course or as short courses on specific topics. (MK)

Acknowledgements

Assistance in the preparation of this package was provided by Blaine Aikin, Larry Albert, Joseph Chmill, Steve Clark, Marjorie Farinelli, Janice Greene, Gretchen Hemmingsen, Paul W. Holland, J. Michael Hopkins, Gaea Leinhardt, Richard Sandusky, Christine Visminas, Diane Warriner, and Tammar Zeheb.

Table of Contents

Introductory Manual

Need

Policy development and administrative decision making in the public sector often require the collection and analysis of quantitative data and the evaluation of analytic results. Consequently, potential managers and administrators require a solid foundation in statistical reasoning and data analytic procedures if they are to become effective practitioners. Most curricula in public management or administration have long recognized this need and have included an elementary statistics course in their required sequence. However, the ad hoc nature of most policy or administrative analyses, the low quality of most data sources, and the need of policy analysts to communicate results to non-technical audiences suggest that the usual introduction to statistics may not be optimal. In addition, since students of policy management rarely complete more than a two semester introductory sequence in quantitative methods, the course must be comprehensive, covering advanced as well as introductory material. Unfortunately, an additional problem is that most introductory texts in statistics, and therefore most introductory courses, are oriented towards the natural or biological sciences, covering topics and developing examples of relevance chiefly to these disciplines.

Quantitative Methods for Public Management (QMPM) represents a break with the traditional approach. The course contains instructional material covering a broad range of statistical and data analytic procedures chosen on the basis of their probable utility to public managers and administrators. The material emphasizes graphics, robust procedures, model development, and the evaluation and critique of analyses. Besides a specially selected set of topics, the course contains data derived from "real world" policy relevant situations. All examples, exercises, and exam problems derive from actual empirical situations of relevance to public policy managers and administrators. By providing relevant contexts for the development and exercise of abstract methods, the course assures a deeper and more lasting educational experience for the student and enhances the student's likelihood of successfully mastering these methods. Quantitative evaluations of the educational effectiveness of QMPM have shown that it possesses definite advantages over traditional approaches (Leinhardt and Wasserman, forthcoming; Leinhardt, Leinhardt and Wasserman, 1977).

General Overview

The package consists of three elements: (1) a set of detailed lecture outlines and supplemental material for an instructor; (2) a set of reading assignments, exercises, and exams for students; and (3) a computer system for performing data analysis on numerical data files.

Instructors, assumed to be experienced at teaching statistics or quantitative methods, use the lecture outlines as guides in the preparation of each 90 minute lecture. The outlines are extensively detailed and organized in a consistent manner. Learning goals and presentation activities are clearly defined and presentation aids such as overhead projector transparency masters are keyed directly to the lecture outline. Since many topics covered in QMPM do not appear in traditional statistics textbooks, suggested readings are specified to provide instructors with a guide to background material.

Students are expect to have a minimum mathematical preparation of college algebra. Units containing more advanced mathematical material (such as calculus) are preceded by prerequisite inventories intended to detect weaknesses in student preparation and to serve as a remedial resource. Supplementary material for reading by students provides coverage of items whose mastery is prerequisite to mastery of unit material. Reading assignments, exercises, and examinations are keyed to the lecture sequence and are designed to provide students with textual descriptions of methods, relevant examples of empirical applications, and opportunities to exercise newly learned skills on problems whose substance is intellectually interesting and of a contemporary nature. Worked solutions to problem sets are provided so that feedback to the student can be rapid and, therefore, educationally effective.

The reading assignments for students refer to both textbooks in methods and academic journals. Several texts are used since, at the time the course was designed, no single text existed which covered all the topics represented in QMPM. Those texts which are heavily read should be purchased while others can be consulted at the library. Journal articles serve the purpose of exposing students to studies of the type they will likely have to read and digest in performing future professional activities. By and large, the selected articles are reprinted in certain edited volumes and purchase of these is suggested. Other material can be found in university and college libraries.

A computer software system is available to provide students with the opportunity to perform numerous data analyses. One of the most limiting features of traditional approaches to the teaching of data analysis is their reliance on student performance of the arithmetic necessary for the completion of an exercise. While hand calculators have facilitated these operations, many of the procedures covered in QMPM require elaborate arithmetical operations which are arduous to perform even on advanced hand calculators. Additionally, effective learning of data analytic practices requires the student to be ready to try several approaches to the same problem or to repeatedly reanalyze parts of a problem. Such experience provides the student with illustrations of the sensitivity of analytic results to the methods applied with practice at the application of similar techniques in widely differing circumstances.

Although frequent performance of analytic activities contributes to learning they can burden the student with an inordinate amount of reptitious and boring hand work. The computer routine (CMU-DAP), available for use with the OMPM package, obviates this activity by having the computer perform the arithmetical operations. The system is designed so that operations appear natural, i.e., no prior programming experience is necessary. The routines are "called" in a language that is easily understood and employed by novices. While the machine generates graphics and performs computations, the student is free to concentrate on alternative analytic strategies or the evaluation of analytic results. Note that while the computing system enhances the learning experience, it is not an essential feature. In particular, instructional material does not depend on its availability. Other commercially available systems such as SPSS, IBM STATPAK, etc., contain routines for performing many of the procedures covered in OMPM and may be substituted for CMU-DAP. Also, new and planned texts (e.g., McNeil, 1977, and Hoaglin and Velleman, in preparation) provide code for exploratory techniques.

<u>Goals</u>

QMPM is designed to facilitate the education of public managers and administrators in contemporary data analysis, to provide them with skills for understanding and criticizing analyses performed by others, and to provide them with skills for presenting and interpreting technical material to non-technical audiences. The pedagogic structure, topic organization, and instructional material are designed with the objective of providing students with a deep understanding of data analytic methods and assuring that a high proportion of students will acquire mastery of data analytic skills.

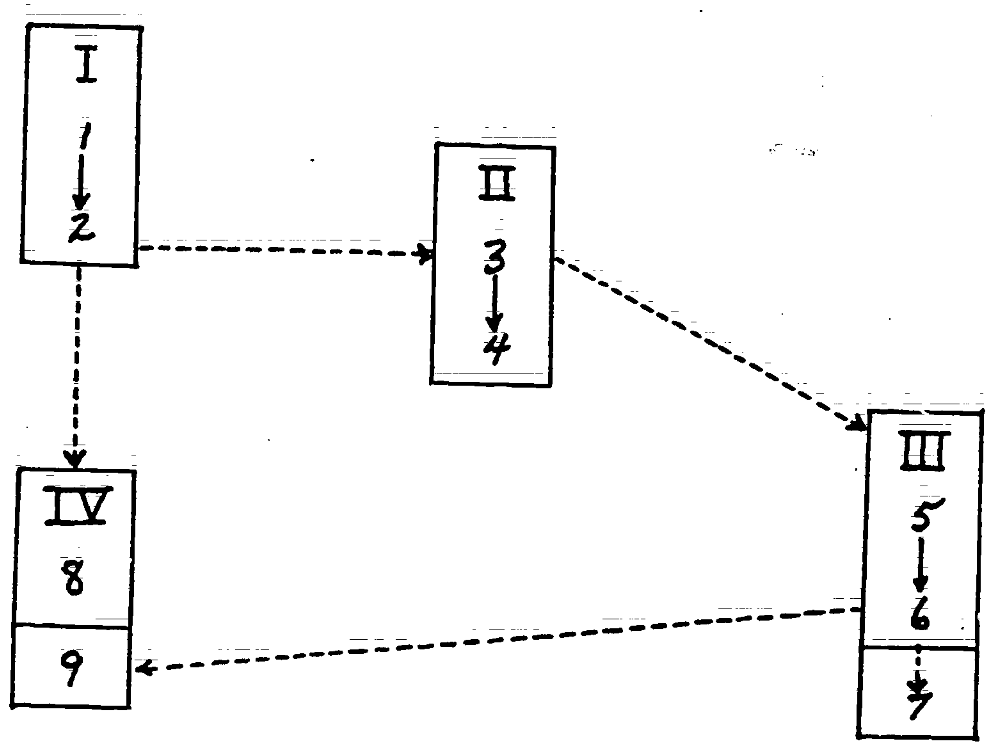
Module and Unit Content

QMPM's curricular material is divided into four independent modules which are further disaggregated into content units consisting of varying numbers of 90 minute lectures.

Suggested Sequencing

A year's length course can be organized by either following the specified sequence of modules and units or modifying this sequence to fit the purposes of the instructor. The designed sequence is based on a hierarchical development of skills for handling increasingly more complicated data sets. Thus, in Module I, single batches of data precede multiple batches, and in Module II regression with one carrier precedes regression with multiple carriers. Note, though, that QMPM's topic organization is non-traditional. The most dramatic deviation from usual sequencing occurs in the presentation of regression as a model fitting procedure <u>before</u> the presentation of probability notions. The assumption here is that probability and inference are not essential to the process of constructing models. Rather, they speak to the issue of selecting best fitting models or estimating parameter values in sampling situations. The logic behind this sequencing is discussed in Leinhardt and Wasserman (1977). An alternative and more traditional approach would place Module III, particularly units 5 and 7, before Module II. Regression could then be covered either directly after Module III or after Module IV. Unit 8 can occur anytime after Module III but Unit 9 should not precede Module III. A diagram of module and unit dependence appears below.

Diagram of Module and Unit Dependence

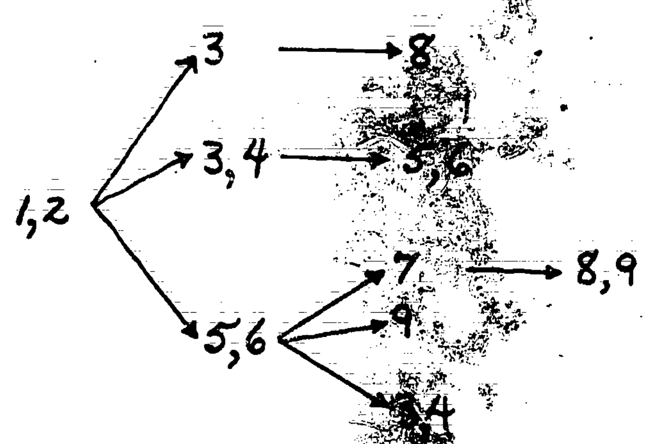


Modules are circled and indicated by Roman numerals; Arabic numerals refer to units. Solid lines indicate design dependence; dashed lines indicate alternatives to the sequence implicit in the unit numbers.

Each module and each unit is a complete instructional package and can, therefore, be taught independently of other QMPM components. However, each does possess a set of prerequisites which are often covered in preceding components. If mastery of these prerequisites is assured, prior components need not be taught.

Since the package is modularized, components can be used to create short courses focusing on specific topics or as part of in-service training programs that offer a selection of topics. For example, a short course on contemporary exploratory data analysis could be composed of units 1, 2, 3, and 8. A short course on analysis of contingency tables could be based on units 5, 6, 7, and 9. A short course on modern data analytic graphics could be developed using units 1 and 2. A short course in regression could be based on units 3, 4, 5, and 6. A flow chart of these alternatives appears below. Other courses can be conceived and interconnected with these suggested sequences.

Flow Chart of Alternative Unit Sequences
for Short or Specialized Courses



Exploratory data analysis

Short QMPM

No regression; contingency tables

Traditional approach

Module Table of Contents

(There are no solutions to unit 7 homework, since there is no single "correct" answer.)

Package Development

The QMPM package was developed at the School of Urban and Public Affairs (SUPA) of Carnegie-Mellon University (CMU). SUPA offers both doctoral and masters degree programs which emphasize public sector professional activities and research. The school is heavily committed to research and to innovations in teaching. The educational staff at SUPA is quantitatively oriented and recognizes the essential importance of sophisticated quantitative training at all graduate levels.

At SUPA the need to develop skills for performing quantitative studies and for presenting results in an informative and effective fashion has always been recognized. Until 1975, satisfying this need was viewed in the traditional manner of including a year's sequence in introductory statistics either through a course offered in SUPA or through other statistics courses available at CMU.

In 1975 an attempt was made to break with the past. QMPM was put together following the acquisition of information on data analysis problems experienced by a broad range of practicing public managers and administrators. From mailed questionnaires and interviews with practitioners it became evident that both the approach and content of traditional courses were suboptimal as far as the needs of public sector professionals were concerned. As a consequence, a course was developed that emphasized graphics, exploratory procedures and robust analyses. In addition, topics such as survey design, sampling methods and analysis of cross-classified data were added while other, less relevant material, was excised.

In the 1975-1976 academic year an experimental version of QMPM was offered to approximately 20 first year masters students at SUPA. Emphasizing application and based in a relevant empirical context, the course proved to be an outstanding success. When the NTDS/HUD curriculum development project was announced, QMPM seemed to be a natural base for a proposal and it was ultimately funded.

The course development activity took place between May 1976 and August 1977. Simultaneously, during the academic year, an experimental version of the course based on the curricular material developed under the NTDS/HUD subcontract was taught. In addition, a short version of QMPM was taught as part of an in-service training program for personnel in community mental health programs. Feedback in the form of student opinions and outside evaluative observation of student progress were used to revise the instructional material. The final product has been tested under a variety of situations and promises to provide a significant improvement in the educational experience of public administrators and managers in the area of data analysis and statistical methods. Leinhardt, Leinhardt, and Wasserman (1977) reports results of a quantitative evaluation of the experimental 1976-1977 implementation.

References:

Bishop, Y.M. M., S.E. Fienberg, and P. W. Holland (1975), Discrete Multivariate Analysis, Cambridge, Massachusetts: The M.I.T. Press.

Erickson, B.H. and T.A. Nosanchuk (1977) Understanding Data, Toronto: McGraw-Hill Ryerson.

Fairley, W.B. and F. Mosteller (1977) editors, Statistics and Public Policy, Reading, Massachusetts: Addison-Wesley.

Fienberg, S.E. (1977), The Analysis of Cross-Classified, Categorical Data, Cambridge, Massachusetts: The M.I.T. Press, in press.

Hoaglin, D.C. (1976), A First Course in Data Analysis, preliminary edition, Reading, Massachusetts: Addison-Wesley.

Huff, D., How to Lie with Statistics, New York: Norton, 1954

Leinhardt, G., S. Leinhardt, and S.S. Wasserman (1977) "An Experimental Evaluation of Two Approaches to Teaching Applied Statistics and Data Analysis," working paper, Carnegie-Mellon University, School of Urban and Public Affairs.

Leinhardt, S. and S.S. Wasserman (in preparation) "Teaching Regression: An Exploratory Approach."

McNeil, D.R. (1977) Interactive Data Analysis, New York: John Wiley & Sons.

Mosteller, F. and J.W. Tukey (1977), Data Analysis and Regression, Reading, Massachusetts: Addison-Wesley.

Tanur, J., et.al., editors, Statistics: A Guide to the Unknown, San Francisco: Holden-Day, 1972.

Tufte, E.R., Data Analysis for Politics and Policy, Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Tufte, E.R., editor, The Quantitative Analysis of Social Problems, Reading, Massachusetts, 1970.

Tukey, J.W. (1977) Exploratory Data Analysis, Reading, Massachusetts: Addison-Wesley.

Warwick, D.P. and C.A. Lininger, The Sample Survey: Theory and Practice, New York: McGraw-Hill, 1975.

Wonnacott, R.J. and T.H. Wonnacott, Econometrics, New York: Wiley, 1970.

QUANTITATIVE METHODS FOR PUBLIC MANAGEMENT

INSTRUCTOR'S MANUAL

Developed by

SCHOOL OF URBAN AND PUBLIC AFFAIRS
CARNEGIE-MELLON UNIVERSITY

SAMUEL LEINHARDT, PRINCIPAL INVESTIGATOR
and
STANLEY S. WASSERMAN

Under Contract to

THE URBAN MANAGEMENT CURRICULUM DEVELOPMENT PROGRAM
THE NATIONAL TRAINING AND DEVELOPMENT SERVICE
5028 Wisconsin Avenue, N.W.
Washington, D.C. 20016

Funded by

The Office of the Assistant Secretary
for Policy Development and Research
U.S. Department of Housing and Urban Development

Acknowledgements

Assistance in the preparation of this package was provided by Blaine Aikin, Larry Albert, Joseph Chmili, Steve Clark, Marjorie Farinelli, Janice Greene, Gretchen Hemmingsen, Paul W. Holland, J. Michael Hopkins, Gaea Leinhardt, Richard Sandusky, Christine Visminas, Diane Warriner, and Tammar Zeheb.

Table of Contents

Instructor's Manual

Introduction

Quantitative Methods for Public Management contains instructional materials to cover four major areas in data analysis: (I) Exploration of batches of data; (II) Modeling continuous data using regression; (III) Probability, sampling and inference; and (IV) Modeling cross-classified data. Each module consists of lecture outlines, reading assignments, examinations and exercises with solved problems, masters for visuals, prerequisite inventories, and material for distribution to students. The lecture outlines are to be used as presentation guides for instructors. All other material is for student use. Computer routines are also available for student use in performing analyses of empirical data.

The QMPM modules are not self-contained instructional components. Their use depends upon the availability of various commercially distributed texts, readers, and similar resources. These resources are detailed below and in the modules themselves.

All modules are similarly organized. The instructional strategy used follows established educational theory. Prerequisite inventories are employed to determine whether students possess knowledge of various concepts and methods upon which mastery of a unit's substance depends. Handout material and references provide students with detailed and sufficient information on these prerequisites. Broad based understanding of technical areas is avoided at this stage, with focus instead on specific tools or ideas that are used in the units. The units themselves contain introductory material in the form of advanced organizers which sensitize the student to topics and ideas that will be covered in the unit. New material is then presented in an instructional mode with general principles followed by examples of applications or development. Visuals, in the form of overhead projector transparencies, are used extensively. Students should have copies of these slides in hand while the lecture proceeds. All examples are based on empirical data descriptive of or relevant to public policy or administration. Students demonstrate learned skills in three situations: homework, papers and quizzes. Homework is designed to present students with problems to be solved in unpressured time periods. Solution of specific, well-defined problems are at issue here. Papers provide longer study periods and require demonstration of comprehension and interpretation of an unstructured problem. Exams require students to operate under pressure to solve relatively straightforward problems. Text references and readings on empirical studies in which quantitative analytic methods are applied to empirical policy issues provide students with diverse examples of applications. Computer operations permit students to participate personally in numerous analyses. It is recommended that students write one 10 to 12 page paper at the conclusion of each module, the topic should be selected by the student in consultation with the instructor. The paper should contain a quantitative analysis of of a public policy issue and an extensive verbal discussion of the study.

Objectives

The goal of QMPM is to help students of public management and administration master a diverse set of data analytic tools. Closely associated with this goal is that of providing students with a critical sense of what is a good and useful analysis and with skills to present relatively complicated analyses to non-technical audiences in such a manner that results and implications are effectively communicated. Paper assignments, as discussed above, are essential to achieving this goal.

Instructor's Role

Quantitative Methods for Public Management (QMPM) is first and foremost a course in data analysis and statistics. QMPM has been designed under the assumption that students will not continue a course of study in statistics beyond their experience with QMPM. (Although QMPM does provide all essential material for continuation). Thus, the material covered, the presentation process, empirical context and instructional activities have been designed to achieve both a broad introduction to quantitative methods and a deep, lasting learning of analytic skills.

The role of the instructor in accomplishing these objectives is critical. Because courses in quantitative methods are traditionally thought of by students as "hard" courses and even irrelevant to their main concerns, instructors of required quantitative methods courses face a particularly difficult task. When the material covered is as novel as that in QMPM the behavior of the instructor becomes even more central to success.

The instructor must possess self-assurance and be able to demonstrate competence with the methods taught. Instructor familiarity with the substance and procedures of QMPM is, thus, essential. Prior to teaching QMPM the instructor should proceed through all of the instructional material so that the essential features and idiosyncracies of the course are known. To a great extent, QMPM is an attitude--an attitude towards data, their manipulation and analysis. For successful transference of this attitude to students, instructors must be able to demonstrate it in their own behavior, in their willingness to pursue unorthodox analysis and to explore data in attempts to make the data "talk." This attitude is not easily transmitted to students, especially students who may possess only weak mathematical skills or who have been taught that data are sacrosanct. Nonetheless, acquisition of this attitude towards analysis and towards quantitative data may be considered to be the primary behavioral objective of QMPM.

The instructor is expected to perform much in the manner of an instructor in any traditional course. A lecture situation is assumed in which the instructor presents material on a scheduled basis before a group of students. The lecture outlines should be used by the instructor as a guide in the preparation of a lecture. The instructor should promote questioning by students, pursue general problems of understanding in depth but leave for private consultation an individual student's problem when a brief response is unsatisfactory.

The instructor should construct many examples of the application of QMPM procedures. These examples need not be elaborate but should demonstrate how understanding of a policy or administrative issue is improved by use of data analytic tools. The instructor should aim to develop examples based on local situations or topics of current national or international interest. Artificial examples, unless the point they make cannot be covered in any other way, should be avoided.

The instructor should be available and responsive to student inquiries outside of formally scheduled class periods. Students are required to engage in numerous exercises and should be encouraged to try alternative approaches to a given problem rather than seek the one "correct" answer. Since this will inevitably lead some students into situations which they do not have the knowledge to understand, they should know that help is available. Remember, good positive attitudes towards the performance of analysis are essential. Lack of support from instructors sets up a poor role model and turns students off. Since courses in quantitative methods have historically suffered from a poor image, instructors should act to compensate for student insecurity.

Besides being available and supportive, instructors should provide students with rapid feedback on exercises, paper assignments, and exam performance. Because of the diversity of topics covered in QMPM students may be unable to use feedback information regarding a particular behavior or skill if it comes long after demonstration. In addition, feedback is most effective in learning if it follows rapidly on behavioral action. When it does, students can adjust their understanding or modify a behavior while the activity is fresh in their minds and, possibly, demonstrate the correct behavior and have it confirmed in another circumstance.

Instructor's Qualifications

The instructor is assumed to have experience teaching quantitative methods or statistics at the graduate level. Experience at performing empirical studies contributes to the instructor's ability to relate abstract notions or methods to real life situations. It is not essential for the instructor to be a statistician or mathematician. Nor is it essential for the instructor to have extensive prior experience with all the topics covered in QMPM. An instructor with knowledge of classical statistics is advised to read carefully all text and reference material cited in the package and, in particular, to read Tukey, J.W., Exploratory Data Analysis, Addison-Wesley, 1977, Mosteller, F.M. and J.W. Tukey, Data Analysis and Regression: A Second Course in Statistics, Addison-Wesley, 1977, Bishop, Y., S. Fienberg and P.W. Holland, Discrete Multivariate Analysis, MIT Press, 1975, S. Fienberg, The Analysis of Cross-Classified Categorical Data, MIT Press, in process, McNeil, D.R., Interactive Data Analysis, Wiley, 1977, and Erickson, B.H. and T.A. Nosanchuk, Understanding Data, McGraw-Hill, 1977.

Staff Support

QMPM can be taught by a single instructor. However, with a sizable class (10 or larger) the need for rapid feedback and availability may infringe upon an instructor's other responsibilities. In such situations it is highly advisable to have teaching assistants available. These individuals should have regular hours in which students can have access to them and should take responsibility for grading homework exercises and quizzes. Since QMPM is supplied with worked problems for exercises and exams, performance of these activities by teaching assistants should pose no difficulties.

If the computer routines supplied with QMPM are employed, then a staff member should take responsibility for interacting with students regarding their usage. The system that is provided has been extensively tested and debugged and, therefore, should need no software work beyond that required for mounting on the local computer. However, students unfamiliar with computer software packages may become unnecessarily frustrated by their own lack of knowledge about the system. This can be relieved by assigning either a teaching assistant or another staff member the responsibility of becoming adept at using the system and relying on that person to act as an inhouse systems consultant. Data acquisition and mounting might be handled by the same person. It is suggested that a large data library should be acquired and left open to student exploration.

Technical Resources

Because of the uniqueness of QMPM no single textbook is fully satisfactory. Indeed, even combinations of texts fail to provide students with complete material for studying some topics covered in the course. For this reason it is advisable to provide students with alternative means for reviewing the contents of a lecture. Instructors might wish to reproduce copies of lecture outlines so students will have a topic outline of covered material. Of particular utility here, however, is the use of video taping equipment. If such resources are available, then lectures should be taped and a tape library of the course constructed which students can exploit at any time. Such devices have been highly regarded by students when employed in experimental implementations of QMPM. A technical requirement for such equipment is the ability to resolve small characters when written on a blackboard.

QMPM is designed to be taught by an instructor in a traditional lecture format. A hall or room which possesses ample blackboard space is required. The instructor should feel free to write examples on the board, draw figures, and otherwise illustrate material as the need arises. If video taping equipment is used, the room should have sufficient lighting to permit high contrast resolution of material written on the blackboard. Since the use of overhead transparencies is assumed, the room should be provided with a projection screen and a location for the projecter. QMPM comes with dense paper masters of transparencies. These should be reproduced onto plastic slides by the instructor through use of appropriate equipment. The instructor should distribute paper copies of these transparencies to students before a lecture and should also assure that reproductions of other hand-out material are available on a timely basis. Since references to contemporary texts and articles occur in both student and instructor material, the availability of a library is advantageous.

If the computer routines are to be used, then the routines should be mounted on a computer before the course commences. While the system has been adequately debugged, there are likely to be local machine idiosyncracies that must be overcome for efficient operation. Some software may have to be written at the implementation site as a consequence. The computer system is designed as an interactive system. A time-shared computer and hard-wired or acoustically coupled printing terminals are suggested. CRT's are not recommended. While the system can be operated in a batch processed mode, its educational utility is maximized when it is operated interactively.

Use of Instructional Materials

The primary curricular components contained in the modules are lecture outlines, one for every 90 minute lecture in the course. These are organized with a zeroth lecture containing advance organizers for students followed by lectures containing substantive presentations.

Instructors are expected to use these components as topic guides and are not expected to adhere to them absolutely. Both the level of preparation of students and the nature of the implementation should condition the actual presentation of material. Similarly, visuals in the form of overhead projection transparencies, homework problems, examples, and test material are provided as guides. While those delivered in the QMPM package can be used as they stand, the instructor should make an effort to construct comparable examples and problems which are relevant to the specific time and place of the implementation. In addition it is the responsibility of the instructor to see that copies of material to be used by students and copies of transparencies be prepared and distributed.

The expected usage is as follows (recalling that each module is organized in units): A prerequisite inventory containing material whose comprehension is required for mastery of QMPM unit topics is distributed. Homework problems on prerequisite material (which can be taken home or done under in-class test conditions) follow. Solutions to these problems are given to students after they have attempted to solve the problems. Difficulties with prerequisite inventory problems should be resolved by the student and confirmed by the instructional staff prior to exposure to new material.

A lecture N-0 (where N indicates the unit) precedes every unit. This lecture (which is discretionary) contains advanced organizers to focus the student's attention on topics that will be covered in the unit. The more complex the material covered in the unit the more important are the advanced organizers. The instructor uses the lecture outline in this and every case as a guide, modifying and adjusting the presentation as style and context dictate.

The material presumes a lecture, i.e., an instructor standing before an audience and making an oral presentation. Each lecture is 90 minutes in duration (which may be organized into one or two class sessions). Presentation aids that are also assumed are a blackboard with space sufficient for copious drawings and writing equations, an overhead projector and screen and duplication facilities for producing handouts prior to a lecture. Suggested presentation sequences for transparencies are keyed by number in the lecture outline (numbers in brackets on righthand side) and summarized in a transparency guide.

Following presentation of the unit's introductory lecture a student reading assignment is distributed. The number of lectures for each unit depends upon the unit's contents. Following the substantive lectures instructors should provide a review lecture, although such lectures are discretionary. Classes of advanced or experienced students may not need review while slower students will find reviews critical to complete mastery.

Homework problems and quiz material with worked solutions follow each unit's lecture. The homework schedule should assure rapid return of graded and corrected problems. Students should require no more than one week to hand in homework and should receive corrected homework in two to three days. Students failing to comply with homework requirements or who consistently hand in erroneous problem sets should be singled out for remedial help. A unit quiz should be conducted in a classroom under examination conditions. Students should be permitted the use of hand-held calculators.

This entire procedure is repeated for each unit. A week of class should normally include three hours of instruction with one to two additional hours available for workshop and review (with instructional staff). Workshops are also used for the presentation of special material (such as "Some Principles of Graphics for Tables and Charts" in Module I). The instructor is responsible for organizing and facilitating the functioning of the course. The instructor is also expected to elaborate on the policy relevant nature of examples, problem sets, and outside readings. These elaborations should be made during normal lecture presentations and through handouts of worked problems derived from local (i.e., the locale where the implementation occurs) situations.

Workshops should include extensive discussions of applications. Students are expected to attend lectures, complete homework problems, and take quizzes and the final examination. In addition, students should be required to produce two papers (10 to 20 pages in length) within a semester's time in which QMPM techniques have been applied to a policy or public management problem of the student's own choosing. These requirements allow the student to perform data analysis in three types of situations: homework provides structured problems with lax time constraints; quizzes provide structured problems with tight time constraints; papers provide unstructured problems with extended time constraints. In all cases grading should be based on the effective solution of the problem and its interpretation by the student. Total student preparation effort should range from two to four hours per lecture.

Use of the computer

QMPM is provided with a computer package containing a set of analytic routines and a data library. These are meant to be used by the student in exercising learned skills. Students are expected to use the system in doing homework problems and writing papers. While not essential to the successful implementation of QMPM, the computer system does provide students with opportunities for performing elaborate studies and for carrying out numerous analyses where one study would otherwise be considered sufficient. Since empirical data are idiosyncratic the effective analyst and critic of analyses should have extensive "hands-on" experience with empirical studies. Such experience often is a consequence of a long career. Students of QMPM, however, by using the computing system, can develop such experience while they participate in the course. Homework problems that require computer assistance are indicated. Alternative software which permits students to perform necessary computations on a machine may be used in place of the routines in the QMPM package.

Audience

QMPM is designed as an entry level masters course of one year duration. Students in such a class are expected to have successfully completed college mathematics courses so that they are proficient in algebra. While some matrix algebra and calculus are used, the prerequisite inventories and handouts supplied with QMPM provide sufficient coverage of these tools. No knowledge of statistics is assumed nor is any experience with computers or programming required.

Because of its unitized-modular structure a variety of courses other than a one year sequence may be generated from the QMPM package. At the graduate school level these may take the form of short courses on specific topics or one semester courses containing selected modules. In-service training programs may also be developed. In each case student prerequisites are the same as for a one year course save that certain advanced units in the package build upon material covered in other units.

QUANTITATIVE METHODS FOR PUBLIC MANAGEMENT

STUDENT'S MANUAL

Developed by

SCHOOL OF URBAN AND PUBLIC AFFAIRS
CARNEGIE-MELLON UNIVERSITY

SAMUEL LEINHARDT, PRINCIPAL INVESTIGATOR
and
STANLEY S. WASSERMAN

Under Contract to

THE URBAN MANAGEMENT CURRICULUM DEVELOPMENT PROGRAM
THE NATIONAL TRAINING AND DEVELOPMENT SERVICE
5028 Wisconsin Avenue, N.W.
Washington, D.C. 20016

Funded by

The Office of the Assistant Secretary
for Policy Development and Research
U.S. Department of Housing and Urban Development

Acknowledgements

Assistance in the preparation of this package was provided by Blaine Aikin, Larry Albert, Joseph Chmill, Steve Clark, Marjorie Farinelli, Janice Greene, Gretchen Hemmingsen, Paul W. Holland, J. Michael Hopkins, Gaea Leinhardt, Richard Sandusky, Christine Visminas, Diane Warriner, and Tammar Zeheb.

Table of Contents

Introduction

Quantitative Methods for Public Management (QMPM) is a course of instruction in data analysis and statistics for students of public management and administration. The course is designed to teach you how to perform and criticize data analyses and how to interpret and present analytic results for effective communication to non-technical audiences. The course was developed at the School of Urban and Public Affairs of Carnegie-Mellon University as part of a curriculum development project funded by the Federal Department of Housing and Urban Development.

QMPM is structured into four modules which cover a diverse set of quantitative analytic methods. Because of its modular structure your instructor has the option of presenting all of the material, in which case a year long course of study is assumed, or selecting components for shorter periods of instruction. The topics covered have been selected specifically for their utility in policy and administrative studies. Contemporary educational theory has been used throughout to assure that you will have the greatest chance of mastering the material and developing a deep understanding of principles. In addition, to improve relevance of the course, all examples, exercises and examinations are based on empirical data that derive from or are relevant to public policy and administrative issues.

Audience

QMPM is designed as an entry level year long masters course. However, its modular structure permits it to be used in shorter course sequences and in in-service training programs. The student is assumed to have successfully mastered a college mathematics course. In some instances mathematical skills beyond this level are required. In these cases a prerequisite inventory test will be administered by your instructor and material will be provided to aid you in acquiring mastery of the necessary concepts and tools.

Instructional Organization

QMPM is designed as a lecture course. Your instructor will prepare presentations based on the instructional material in the package. Each unit of material will be preceded by a presentation in which the instructor will describe the objectives of the unit, the types of skills you will learn and the nature of the problems these skills will enable you to solve. These advance organizers will help focus your expectations about the unit and will enhance your receptivity when new information is provided.

You should prepare for each lecture by reading the text and article assignments <u>before</u> class. These refer to textbook discussions and application examples. Endeavor to become as familiar as possible with each new idea you encounter so that each lecture will be more readily understood. You should expect to spend between two and four hours preparing for each lecture.

Prior to each lecture you will receive a handout of reproductions of any visuals your instructor will use in the presentation.

This will enable you to refer to visuals during the lecture which are not at that moment being projected. The lecture is keyed to these displays, which provide examples and figures illustrating concepts and methods covered in the lecture.

During a lecture you should feel free to raise questions concerning material being presented. It is important that you feel that you understand the procedures taught and can apply them in other contexts.

Following each lecture you should use your notes and copies of the visuals to review the lecture's substance. If the lecture has been video taped you should use this resource to review aspects of the lecture that may seem more difficult than others.

Homework exercises will be distributed each week. These should be attempted as soon as the topics they refer to have been covered in class. It is absolutely critical that you do the homework. Data analysis is a skill which can only be mastered through use. The homework gives you an opportunity to exercise and perfect your newly learned skills by exploring the kind of data that you are likely to encounter in your professional career. You should try to become facile at organizing, analyzing, and interpreting these data. QMPM is designed to maximize the benefit you will derive from learning quantitative methods, but there is no substitute for extensive experience.

In doing your homework and in taking examinations you will find a hand calculator to be invaluable. These devices are relatively inexpensive and one should be purchased at the beginning of the course. You will need a machine that has the four arithmetic functions, logarithms and exponentiation. A memory is advantageous but not necessary.

The QMPM package includes a computer system. If this system is available in your course you should learn to use it early. It will help your learning experience by removing the drudgery of repetitious arithmetical operations from your exercises and permit you to concentrate on analytic strategy and on replicating similar analyses on different data sets. Thus, you will be able to amass more experience with data analysis than if you had to rely on hand calculations. Some exercises supplied with the course are meant to be done on a computer. These will be so indicated.

QMPM is supplied with worked exercises and exam problems. Thus, you can expect rapid feedback from your instructor if you complete your assignments promptly. When feedback on new skills occurs shortly after demonstration of the skill, the learning process is more efficient and effective. This is particularly important in the case of a technical course where new skills build on older ones. There is a cumulative process involved which will be short circuited if you fall behind significantly.

As you progress in QMPM you will discover that other courses that you may be taking will become easier. QMPM covers basic notions and methods in data analysis and statistics. The procedures you will learn in QMPM are used throughout the social and policy sciences, and thus articles or textbooks you may read in other courses can be expected to make use of them. Consequently, successful mastery of QMPM will enhance the successfulness of your entire program of study.

QUANTITATIVE METHODS FOR PUBLIC MANAGEMENT

MODULE I, REVISED

Developed by

SCHOOL OF URBAN AND PUBLIC AFFAIRS
CARNEGIE-MELLON UNIVERSITY

SAMUEL LEINHARDT, PRINCIPAL INVESTIGATOR
and
STANLEY S. WASSERMAN

Under Contract to

THE URBAN MANAGEMENT CURRICULUM DEVELOPMENT PROGRAM
THE NATIONAL TRAINING AND DEVELOPMENT SERVICE
5028 Wisconsin Avenue, N.W.
Washington, D.C. 20016

Funded by

The Office of the Assistant Secretary
for Policy Development and Research
U.S. Department of Housing and Urban Development

Acknowledgements

Assistance in the preparation of this package was provided by Blaine Aikin, Larry Albert, Joseph Chmill, Steve Clark, Marjorie Farinelli, Janice Greene, Gretchen Hemmingsen, Paul W. Holland, J. Michael Hopkins, Gaea Leinhardt, Richard Sandusky, Christine Visminas, Diane Warriner, and Tammar Zeheb.

TABLE OF CONTENTS

Material intended solely for the instructor is denoted by a (I). Material that should also be distributed to the students is denoted by a (S).

General Introduction to Quantitative Methods for Public Management (QMPM)

Many of you are wondering why students of public management are required to study quantitative methods. The reason is this: Public management involves decision making, and making effective decisions requires careful evaluation of information. Today, information of relevance to public managers often comes in quantitative form. Its evaluation requires an operational knowledge of analytic methods. This course is designed to provide this knowledge by teaching you data analytic skills that will enhance your ability to gather quantitative data, operate on them, and use them to make better, more effective decisions. The course's curriculum has been carefully designed to provide you with a variety of tools which will cover most of the data analytic problems you will encounter as practitioners. To help make the course representative of reality it contains an extensive library of real data, the same kind of data that operating public managers use, so that your learning experiences will come as close as possible to the realities of public management. You will be asked to exercise your new skills on these data as you progress. To help you acquire these skills there exists an elaborate support system composed of pedagogic procedures, personnel and audio-visual equipment. The system will be described in this introduction and in a class presentation.

QMPM is a new course. One might even say that it is a revolutionary course. It is revolutionary in that it breaks with traditional approaches to teaching quantitative methods in both its pedagogy and its substance. Pedagogically, it emphasizes mastery learning concepts, the organization of topics into inferred learning hierarchies with clearly specified skill prerequisites. You will need to know the prerequisite material before you can proceed through a topic. These prerequisites include all but the most basic skills that you will need to succeed. There are no hidden assumptions; no special knowledge is required to master a given topic other than what is clearly spelled out in the prerequisite inventories that precede each of the four major sections or modules composing the course. Facilities have been provided and time will be set aside to help you master these prerequisites should they be unfamiliar to you.

Mastery learning also means that you will not be graded on a curve or normed. There is nothing in this course that is too difficult for any of you to master. If you all master the material, you will all pass high. If problems arise because of a lack of comprehension or understanding, numerous resources exist to help you locate the specific difficulty and obtain ultimate mastery of the skill. In general, you should feel assured that every effort will be made to help you master a skill before pushing ahead to new material.

The other new aspect of this course rests in the selection of topics to be covered. QMPM is not a course in statistics. While some topics will be covered that are discussed in traditional statistics courses, they are approached from a pragmatic rather than a theoretical point of view. The theory discussed will be just sufficient to insure comprehension of particular skills and awareness of their limitations. The emphasis will be on doing analysis rather than studying analysis.

Although QMPM is not a statistics course in the traditional sense, it is a course in the analysis of quantitative data. In addition to some traditional methods of statistical analysis, a variety of new tools and analytic methods which were pioneered in the late 1960's by John W. Tukey, a statistician at Princeton University and Bell Laboratories, will be covered. The new methods which Tukey and others have been developing emphasize the exploratory nature of data analysis, the "detective" work that precedes the traditional inferential stage of confirmatory statistics. In exploratory data analysis (EDA), the analyst first organizes the data to understand what kinds of questions can be answered by them and what kinds of operations must precede the application of confirmatory or inferential procedures.

Exploratory data analysis possesses several features that are especially useful to public managers. First, it relies heavily on the use of graphic displays as analytic tools. Traditionally, displays have been used as final summaries presented only after an analysis was completed. In QMPM, however, graphics are used as integral parts of the analytic process, so that they may provide critical information about the data and the process of the analysis. The graphics used in QMPM are relatively simple and easily learned. Indeed, once the graphical methods are introduced you will discover that they have a kind of "face validity"--what they mean is obvious from the way they appear. The face validity of EDA graphics will be a great advantage to you in your professional career. The graphics that you will learn to use can be presented to non-technical audiences and will probably be understood with only minimal explanation. Consequently, rather complicated notions or analytic results can be communicated during presentations to individuals who have a wide variety of backgrounds--a situation you will undoubtedly encounter often in your professional careers.

The second feature of EDA which will enhance its utility to you is the use of resistant or robust procedures. These are procedures which yield results which are relatively unaffected by occasional missing or incorrectly recorded data values or incompletely specified models. Most traditional confirmatory statistical procedures are not resistant in that they are easily influenced by a few widely divergent data values, nor are they robust in that the misspecification of a model can yield invalid results. But public managers often must rely on data of less than highest quality, data collected for other purposes, and models that neglect some variables. EDA procedures are particularly helpful in such situations. The resistant and robust qualities of the procedures covered in QMPM are so important that without them in many situations an investigator would not be able to conduct a thorough study.

Many of the techniques you will learn in this course, both exploratory and confirmatory, are among the newest in the field of data analysis. Learning such up-to-date skills will put you on the "cutting edge" of the field. The newness of these techniques, however, does present some difficulties as far as communicating with others whose training in quantitative methods occurred some years ago. You will be learning procedures which have only recently been made available to the general public. Most texts you will use were published as QMPM was developed. Not many technical data analysis or public management practitioners are familiar with them. Thus, you should expect to find yourself frequently explaining what you have done, even to individuals whom you might normally believe were familiar with data analytic procedures.

Even though many of the techniques we will cover are relatively straightforward, some are complex. Moreover, even simple procedures performed on large data sets can be extremely time consuming when done by hand. Consequently, a computer system (CMU-DAP for Carnegie-Mellon University-Data Analysis Package) exists to facilitate doing analysis. (This is an optional part of the QMPM package.) The system permits data entry, manipulation, and analysis in a simplified format. Doing data analysis on a time-shared computer will facilitate your mastery of analytic skills by allowing you to try many different approaches to the same problem. Thus, you will be able to gain wide experience in applying your skills without fear that you will have to invest an inordinate amount of time on arithmetic operations. By having the "grundge" work of data analysis performed by computer you should be free to concentrate on planning and interpreting your analysis and on exploring alternative approaches. If it is available in your course, you will be introduced to CMU-DAP during the second week of QMPM in a special three hour session, and will be expected to use it for both homework and paper assignments.

The library of real data that has been prepared for your use has already been mentioned. Typically, the data analyzed in traditional statistics courses are fabricated for the purpose of illustrating a particular technique. More often than not, such data are very unrealistic--not the type which you would actually confront in the "real world". Consequently, students, when they leave school and engage in data analysis in the field, often find that their training has not prepared them for the vagaries of reality. In QMPM these situations are avoided. QMPM stresses the analysis of real data, data gathered from practitioners, faculty, students, and published sources. This collection has been organized into a computer based DataBank that can be accessed with CMU-DAP. In addition to choosing data from the DataBank for analysis throughout the course, you will also be expected to gather real data and analyze it.

In summary, QMPM may very well be the most important course that you take in graduate school. You are to participate in a revolutionary approach to quantitative methods--you are co-conspirators in an attempt to make data analysis relevant and useful to public management.

Introduction to Module I

Overview

Module I of the Quantitative Methods for Public Management package contains two units, numbers 1 and 2. Unit 1, Single Batches of Data, introduces the student to the notion of a data batch, a fit, and an effect. It focuses on the organization, condensation, and analysis of simple situations, single batches. The general objective is to familiarize students with data and elementary models and to provide students with a set of basic tools for summarizing, displaying, and working with data. Single batches, essentially a single set of observations on one variable, are considered in depth. The tools introduced include classical procedures such as histograms, sorts, means, and standard deviations. But the emphasis is on tools of exploratory data analysis such as stem-and-leaf displays, order statistics and transformation procedures. The definition and features of a well-behaved or Gaussian batch are also considered, and a special section discusses the features of good graphics and charts.

In unit 2, Multiple Batches of Data-Unordered, the student is introduced to the more complicated situation in which more than one distinct set of observations exist. The tools introduced in unit 1 are used in unit 2 to facilitate comparison of effects among batches. Since differences in spread among the batches can confound thedetermination of differences in level, a procedure for finding a transformation that equalizes spread is introduced. This procedure prepares students for variance stabilizing transformations, introduced in a later unit in the context of multiple regression.

Specific Objectives

Unit 1

Upon successful completion of Unit 1 a student will be able to organize a batch of data using simple sorts, stem-and-leaf displays, and histograms. The student will be able to describe the batch of values using various computed summary numbers and to display the summary by constructing a schematic plot. In addition, the student will be able to determine if a symmetrizing transformation would facilitate contrasting the batch with a well behaved batch and will know how to deter mine a good transformation for this purpose. The student will know how to recognize a well behaved batch and use and evaluate classical summary statistics in their description. The student will also have a critical appreciation for effective graphic and tabular displays and be able to construct uncluttered, informative charts containing quantitative facts.

Unit 2

Upon successful completion of Unit 2 a student will be able to recognize a set of non-ordered multiple batches and use parallel stem-and-leaf displays and parallel schematic plots to compare the batches to one another. To improve the effectiveness of comparison when spreads in the individual batches vary greatly, the student will know how to use median by midspread plots to find a spread stabilizing transformation. The student would then proceed to perform an analysis on the transformed data.

Prerequisite Inventory
Units 1 and 2

Units 1 and 2 of Module I focus on the analysis of single and multiple batches of data. Prior to the presentation of the material in these two units, we shall discuss several elementary concepts. The mastery of these concepts is an essential prerequisite to mastery of the skills taught in Units 1 and 2. Before proceeding to Unit 1, you should assure yourself that you are familiar with these basic concepts.

The inventory is divided into the following five sections:

1. Numbers--Properties and Representation
2. Data Vectors--Observations, Subscripts, Indexing, Summations
3. Data Sets--Variables and Various Transformations
4. Percentages
5. Plots and Graph Paper

Additional references to these topics appear at the end of this inventory. Specific topics in these five areas will be reviewed in class only if the average performance of the class indicates that such discussion is necessary. If areas that you are weak in are not covered in class, you should consult a member of the course's teaching staff to determine how best to achieve mastery.

Section 1. Numbers--Properties and Representation

Throughout this course numbers are used. Consequently, the more important properties of the number system need to be reviewed. These properties are discussed in chapter 1 of Rosenbach, et.al. (see the end of this inventory for full reference).

It is assumed that all numbers worked with in this course belong to the set, or collection, R of <u>real numbers</u>. Real numbers are those which can be represented by terminating or nonterminating decimals. Included in R are those numbers without decimal places, the <u>integers</u>. An integer may be positive, 1,2,3,..., negative, -1,-2,-3,..., or zero, 0.

When writing down single numbers, you should take the time to record all the digits (including all the decimal places) in the number to convey as much information as the number allows. The digits of accuracy required in writing a number are called the <u>significant digits</u> or <u>significant figures</u> of the number. In general, the number of significant digits of a number equals the number of digits of accuracy that the measuring instrument allows. If inches are recorded with a ruler marked with tenths of inches, then the first decimal place of the recorded numbers will always be a significant digit. For example, 10.0 inches has 3, and not 1, significant digits. The population of New England in 1790 is another example. In this year it was 1,009,408 persons. This number has 7 significant figures. If one chooses not to record all the digits of a number, the quantity of significant figures is reduced. If the 1790 population of New England has been approximated by 1,009,000, a number with only 4 significant digits, 3 significant digits would have been lost. It is important to note how "fine" the scale of the measuring instrument is, since this knowledge is essential in determining the total number of significant digits of the recorded numbers.

Occasionally, numbers are expressed in a manner which draws attention to the significant figures of the number. This can be done by writing the number as a product of an integer power of 10, and a number between 1 and 10, that is, a number with one digit to the left of the decimal point. This method of recording numbers is called <u>scientific notation</u>. As an example, the 1970 population of the United States, 203,211,926 persons, can be written as 2.03211926 x 100,000,000 (a hundred million) or 2.03211926×10^8 (9 significant digits). We may wish to approximate this number emphasizing only the number of millions as 203,000,000 or 2.03×10^8 (3 significant digits). Table 1 shows various powers of 10, both positive and negative, that you should be acquainted with.

[Table 1 here.]

Once a number is recorded in scientific notation, the number of significant figures of the number equals one more than the number of decimal places, and the correct power, or <u>exponent</u>, of 10 determines the <u>magnitude</u> of the number. Hence, the 1970 population of the United States has a magnitude of 8.

Occasionally one may wish to record a number with fewer than its usable number of significant figures. This technique is called <u>rounding</u>. It saves time and increases comprehension when more than a few numbers are to be examined. The 1790 New England population may be rounded to 1,009,000 persons (4 significant figures) or even 1,000,000 persons (only 1 significant figure). Digits are always rounded to the

TABLE 1

Powers of 10

Power	Number
-6	10^{-6} = one millionth
-5	10^{-5} = one hundred thousandth
-4	10^{-4} = one ten thousandth
-3	10^{-3} = one thousandth
-2	10^{-2} = one hundredth
-1	10^{-1} = one tenth
0	10^{0} = one
1	10^{1} = ten
2	10^{2} = one hundred
3	10^{3} = one thousand
4	10^{4} = ten thousand
5	10^{5} = one hundred thousand
6	10^{6} = one million
9	10^{9} = one billion
12	10^{12} = one trillion

nearest number, with 0,1,2,3, and 4 rounded down, and 5,6,7,8, and 9 rounded up. (Note the 5-5 split of the digits.) Thus:

$$19.1 \rightarrow 19$$
$$17.7 \rightarrow 18$$
$$16.5 \rightarrow 17.$$

Tukey, in his text Exploratory Data Analysis, suggests rounding numbers whose last digit is 5 to the nearest even number. Thus,

$$16.5 \rightarrow 16$$
$$17.5 \rightarrow 18$$

We recommend the former convention.

Occasionally it is convenient to reduce the number of significant figures by just dropping off the unnecessary digits. This is called cutting and is quicker and easier than rounding. In QMPM cutting is used in certain instances, although when accuracy is desired rounding is generally preferred. When the decimal portion of a number is dropped and only its integer component is recorded, the operation is called truncating. Rounding, cutting and truncating are discussed in chapter 1, pages 3-5, of Tukey (1977).

Section 2. Data Vectors--Observations, Subscripts, Indexing, Summation

A batch of numbers is a set of similar numbers, obtained in some consistent fashion. Simple examples of a batch are: 1) Average family incomes for each of Pittsburgh's 186 census tracts; 2) Population of New York State for each year between 1900 and 1970, inclusive; 3) Distance traveled from home to school by each student in this class.

The expression data vector is used as a synonym for batch. A specific datum, or batch value, is an observation or an element of the data vector. Hence, the batch of family incomes for Pittsburgh census tracts has 186 total observations.

It is convenient to have a mathematical representation for a batch of numbers and the observations in the batch. In QMPM a capital letter, such as X, is used to denote an entire batch of numbers. Each individual observation is identified by attaching a number, written below and on the right of this letter. For example, the first observation in the batch X is denoted X_1, the second observation is X_2, etc. The i^{th} element is denoted X_i. Small numbers attached to X that identify different individual observations are called subscripts. Thus, a batch of 10 numbers, denoted X, can be written $X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8,X_9$, and X_{10}. The subscripts are the integers running sequentially from 1 to 10. A more abbreviated representation of this batch is $X_i, i = 1,2,3,4,5,6,7,8,9,10$, or simply $X_i, i = 1,2,\ldots,10$. In this last form, i is called the index, which in this example runs from 1 to 10. The sequence of periods is an abbreviation of the phrase 'and so forth'. The capital letter 'N' is used to stand for the total number of observations in the batch.

Another special notation is used to denote the sum of a batch of numbers. The notation $\sum_{i=k}^{n} x_i$ indicates that the sum $X_k+X_{k+1}+X_{k+2}+\ldots+X_{n-1}+X_n$ is to be formed; i=k indicates that the summation is to begin with the kth element of the data vector, and n indicates that the summation is to end with the nth element. The symbol Σ, the Greek capital letter sigma, by convention, denotes that a summation is to be performed. The letter i is the index and the summation ranges over the values k to n. Listed below are some rules for summations:

1) $\sum_{i=k}^{n} x_i = x_k + x_{k+1} + \ldots + x_{n-1} + x_n$

2) $\sum_{i=k}^{n} x_i^2 = x_k^2 + x_{k+1}^2 + \ldots + x_{n-1}^2 + x_n^2$

3) $\sum_{i=k}^{n} a = (n - k + 1)\, a$

4) $\sum_{i=k}^{n} a\, x_i = a \sum_{i=k}^{n} x_i$

For example, Table 2 is a batch of numbers corresponding to United States spacecraft launchings per year for the years 1957 to 1964.

[Table 2 here.]

Let X denote this batch. Thus, X_1 corresponds to 1, X_2 to 17,..., and X_8 to 81. Summing the numbers in X,

$$\sum_{i=1}^{8} X_i = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8$$

$$= 1 + 17 + 21 + 31 + 49 + 71 + 71 + 81 = 342$$

Thus, a total of 342 spacecrafts was launched by the U.S. between the years 1957 and 1964, inclusive (i.e., including the two 'end' years, 1957 and 1964). As an exercise, you should verify that $\sum_{i=1}^{n} X_i^2 = 20736$.

Chapter 14 of Rosenbach, et.al. (1963) and Appendix A of Hays (1973) discuss summations in greater detail, with some examples and problems.

TABLE 2

U.S. Spacecraft Launchings

Year	Number of Launchings
1957	1
1958	17
1959	21
1960	31
1961	49
1962	71
1963	71
1964	81

Section 3. Data Sets--Variables and Various Transformations

A single batch, or a collection of related batches that are to be analyzed together, is called a data set. Data analysis is concerned with exploring and understanding data sets. The field of statistics encompasses both data analysis and the study of variables--especially random variables or variables with associated probabilities. A variable is a quantity that may assume any one of a set of values. Population of census tracts, number of homicides in a police precinct, and yearly incomes for professors in a major university are examples of variables. A more 'formal' definition of a batch of data is a set of realizations of a particular variable.

A variable may be classified into one of two types depending on the values it may assume. A discrete variable may take one of a finite or countably infinite set of values. The number of students in this class with yearly incomes in excess of $15,000 is a variable which may assume any member of the finite set $\{0,1,2,...,n\}$, where n is equal to the total enrollment of this class. This variable is discrete. All counts of objects or events are discrete. A continuous variable may take a value from a set of infinite size. Length of a particular r[illegible] surface, daily outdoor temperatures, and the percentage of female students in the U.S. are examples of continuous variables. While a continuous variable may be bounded in its values, as in the case of the variable percentage of female students in the U.S., which cannot be less than zero percent or greater than one hundred percent, there are an infinite number of values within these bounds. For practical purposes, if a variable takes on only integer values, it is discrete;

otherwise, it is continuous. Section 4.1 of Blalock (1972) discusses discrete and continuous data.

Occasionally you will want to <u>reexpress</u> or <u>transform</u> a batch of numbers in the process of performing an analysis. The most common transformations involve raising numbers to various powers, a process called exponentiation, or taking logarithms of numbers. Below are some general rules for exponentiation:

1. $(y^n)(y^m) = y^{n+m}$
2. $(y^n)^m = y^{nm}$
3. $y^{-n} = \frac{1}{y^n}$
4. $y^{\frac{1}{n}} = \sqrt[n]{y}$
5. $y^0 = 1$
6. $y^1 = y$
7. $y^2 = y \cdot y$

A logarithm is an important but easily misunderstood concept. It is closely related to exponentiation. Any number may be represented in scientific notion as $p \cdot 10^k$, where p is a number between 1 and 10, and k is an integer power of 10. It is also possible to represent any positive number, N, as 10^y, where y is any <u>real</u> number. When a number is represented in this fashion, y is called the <u>logarithm of base 10</u> of the number N. Any positive number may be used in a base of a logarithm.

More formally, a logarithm of a base number b, of a number N, is defined as that power to which b must be raised to obtain N. In mathematics, given any N>0 and b>0, if $b^y=N$, then $\log_b N=y$. 'Log' is an abbreviation of logarithm (which, in Greek, means "reckoning number"). For example, $5=10^{.69897}$; hence, $\log_{10}5=.69897$. Also, $25=10^{1.39794}$ and $100=10^2$; hence $\log_{10}25=1.39794$, and $\log_{10}100=2$. Some general rules for logarithms are given below:

1) Logs "come" in various bases; however, all logs to different bases differ only by a multiplicative constant. Specifically, if a and b are any 2 bases, then $\log_a N=\log_b N \log_a b$ ($\log_a b$ is the multiplicative constant for this particular conversion). Because of this mathematical fact, any base is essentially as good as any other. However, for various reasons of convenience some bases are preferred in certain contexts.
2) In QMPM logarithms to the base 10 are used exclusively. These are written $\log_{10}$, or merely log. This choice is prompted by the decimal number system. Base 10 logs are called common logs.
3) The second most useful base is the irrational number approximated by 2.71828... which is simply denoted by the letter 'e' in honor of the mathematician Euler. The number e plays an important role in calculus, as well as in other areas of mathematics. It occurs frequently in economics. Logs to the base e are written $\log_e$ or ln (for Napierian or Natural logs.)
4) $\log(1)=0$
5) $\log(0)$ is undefined, that is $\log(0) = -\infty$.
6) Log of a product is the sum of the logs;

   $\log (PQR) = \log P + \log Q + \log R$
7) Log of a quotient is the difference of the logs:

   $\log (P/Q) = \log P - \log Q$.

8) Log of a number to a power is found by multiplying the log of the number by the power:

$$\log (P^n) = n\log P$$

9) Log of a root of a number is found by dividing the log by the root:

$$\log (\sqrt[n]{P}) = \frac{1}{n} \log P.$$

All of these rules for the use of logs derive from the basic definition of a logarithm and the rules of exponentiation. Logs and exponentiation are discussed in Chapter 3 of Paul and Haeussler (1973).

Section 4. Percentages

Familiarity with percentages is essential in a policy oriented quantitative methods course. Many data sets contain variables that are originally recorded as percentages, and often analyses are requested in terms of percentage change. A percentage is a portion of a number expressed in hundredths. The following mathematical statement is common: A is B percent of the number C. Since percents are expressed in hundredths, B percent is equivalent to B/100, and the above statement may be written A = (B/100)(C).

There are three common situations encountered when using percentages:

1) A is unknown, B and C are known.

2) B is unknown, A and C are known

3) C is unknown, A and B are known.

Each of these situations is discussed in turn below.

The first problem is generally stated "B percent of C equals what number?" The answer is found by multiplying (B/100)xC. For example, 35 percent of 120 equals (35/100)x120 = .35x120 = 42.

The second problem is stated "A is what percent of C?" The answer is found by dividing A by C and multiplying the result by 100%; i.e., if B is the correct answer, B = (A/C)x100%. For example, to determine what percent 36 is of 144, calculate B = (36/144)x100% = (.25)x100% = 25%.

The last problem occurs when the whole or base number C is unknown. It is usually stated "A is B percent of what number?" The answer is found by dividing A by (B/100), C = A/(B/100). If A is 90 and B is 40%, then C = 90/(40/100) = 90/.14 = 225, that is, 90 is 40% of <u>225</u>.

Chapter 3 of Blalock (1972) is a good reference for percentages, as is chapters 1-3 of Zeisel (1968).

Section 5. <u>Plots and Graph Paper</u>

In Unit 2 plots, or graphs, of pairs of observations are made. So that graphs can be read easily some conventions have been established. The horizontal axis is called the x-axis and the vertical axis is called the y-axis. The x-axis is placed at the bottom of the page, and the y-axis at the left side of the page. A point on the graph is represented as (x,y). Figure 1 illustrates these preliminary steps for a plot on a page of ordinary graph paper.

The graph paper in Figure 1 has linear scales in both the x and y directions. This is the type of graph paper that is used most in this course. Some rules for improving the appearance of a plot are given below, in brief:

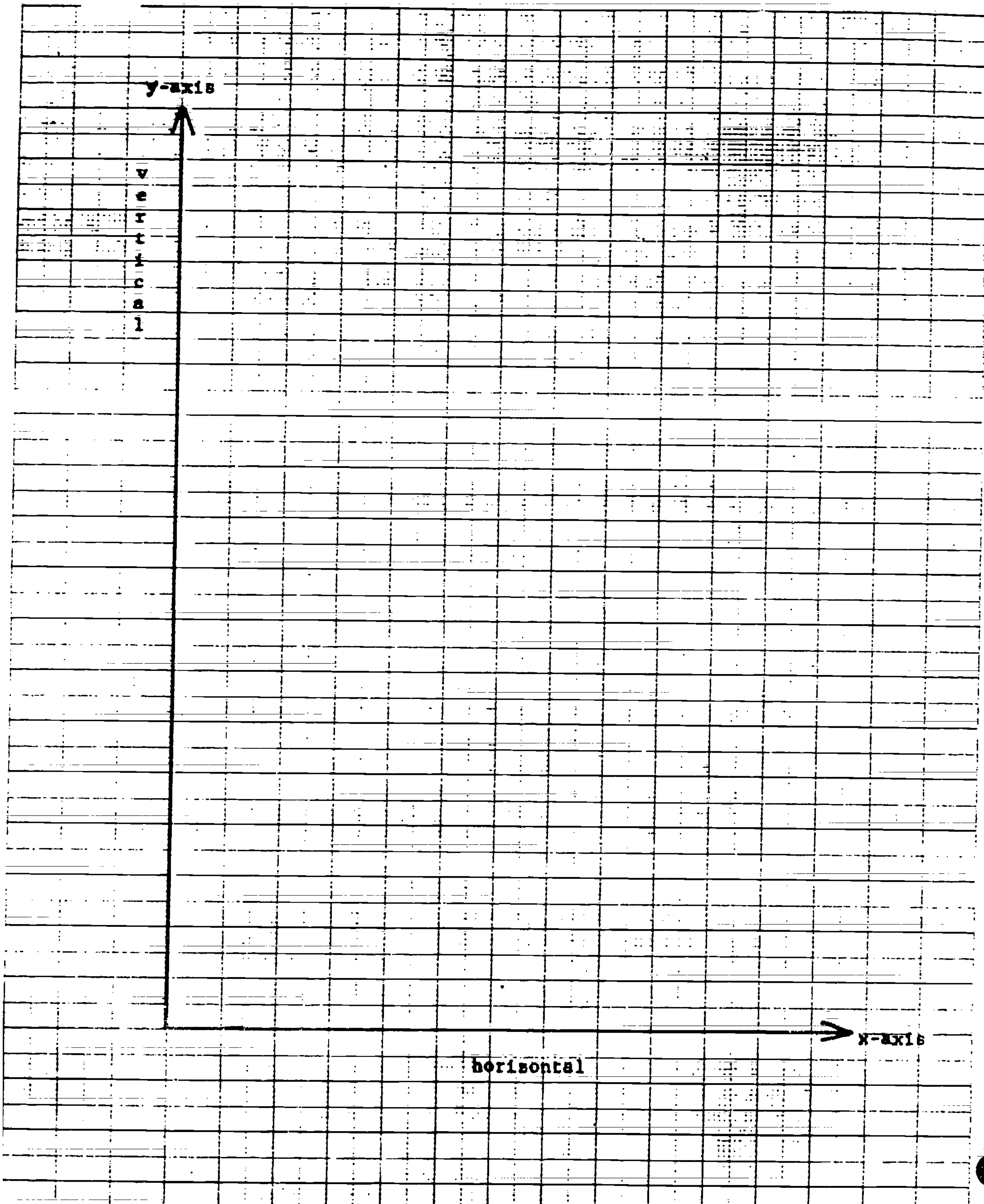
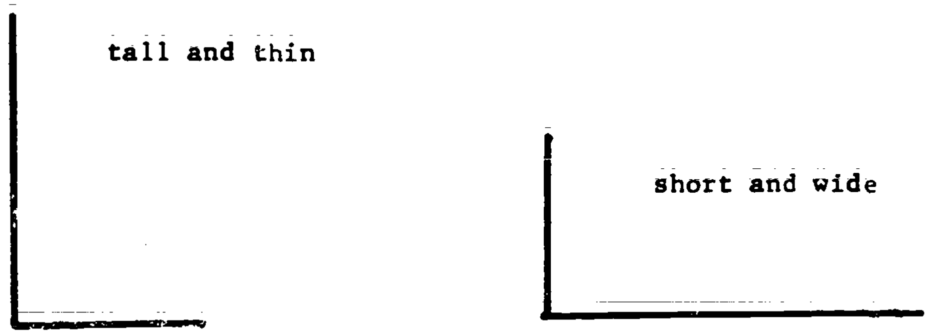


Figure 1.

1) Make plots "tall and thin" (y-axis longer than x-axis) or "short and wide" (y-axis shorter than x-axis)--whichever more effectively conveys your message.



2) Use graph paper with light rulings for the units, heavy rulings every ten units, and intermediate rulings every five units in between. The paper in Figure 1 has these intermediate lines, which make a large difference in speed and accuracy (when plotting).
3) Be clever in assigning numerical values to the basic unit--units other than 1,2, or 5 times a power of 10 are too awkward and tedious.
4) In the finished version of the plot, do not clutter the plot by having too many values marked on the axes.

There are also types of graph paper with non-linear scales. Such paper can save a lot of time when plotting logarithms of the observations. One example is <u>semi-log</u> paper, as seen in Figure 2, with a logarithmic scale for the y-axis. Note that on the log scale the physical distance from 10 to 100 equals the distance from 100 to 1000. This distortion or shrinkage is because log(10) is one unit away from

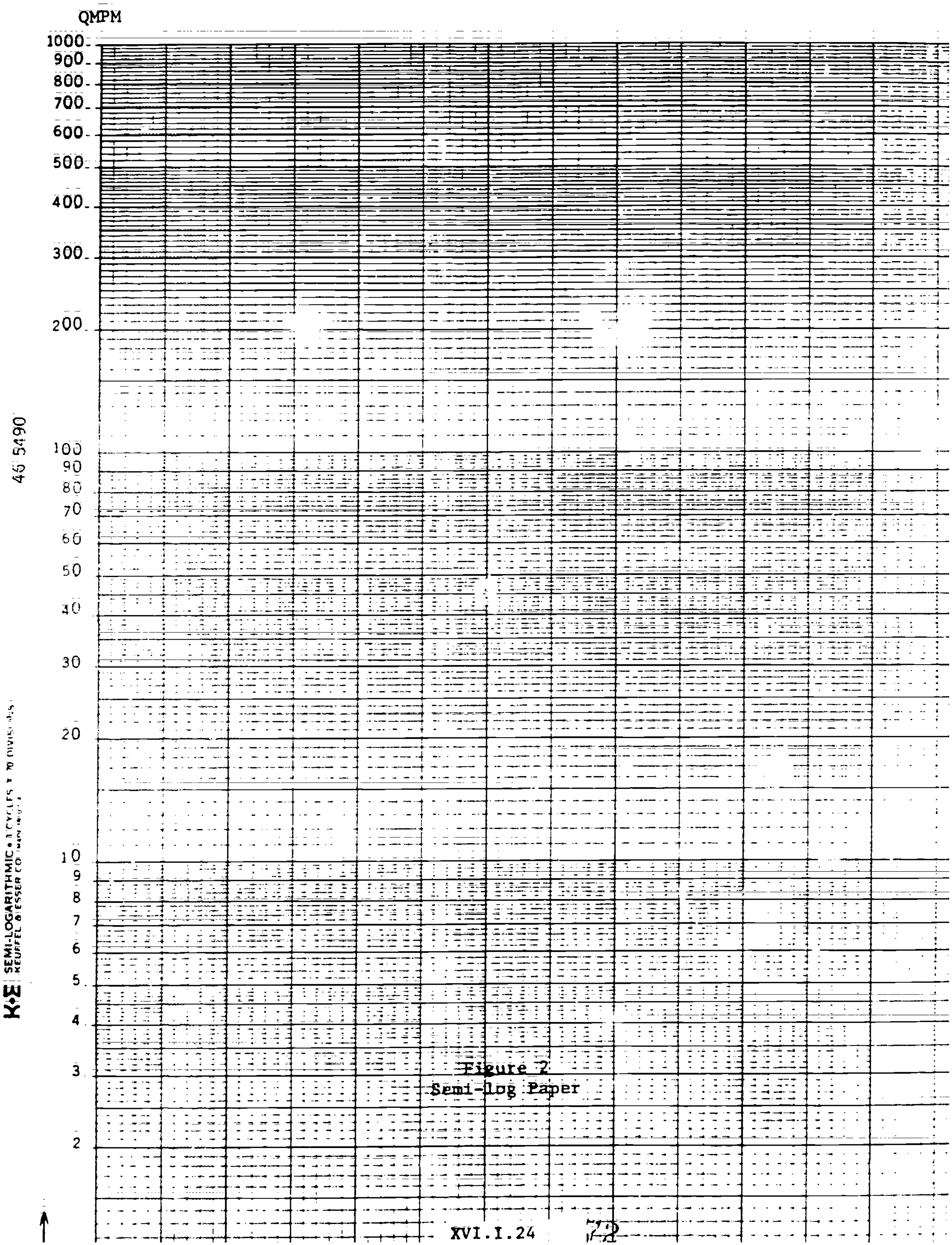


Figure 2
Semi-log Paper

log(100), i.e. log(10)+1=log(100), and log(100) is one unit away from log(1000). Another example of graph paper with non-linear scales is log-log paper which is illustrated in Figure 3. Here, both axes have logarithmic scales. Both semi-log and log-log graph papers will be of use when transforming batches by taking logarithms of the observations because you can go directly from an observed value to its logarithm by simply finding the value on the graph paper's logarithmic scale. This operation makes it unnecessary to first calculate the logarithm using tables, a calculator, or a computer.

Plots are very important in this course. You should reacquaint yourself with the basics of plotting:

1) labelling the axes
2) locating points in the x-y plane using the abscissa, x-coordinate, and the ordinate, y-coordinate, of each point.

Paul and Haeussler (1973) discusses graphing in Chapter 3, Section 3, and, in general, can serve as a useful reference volume.

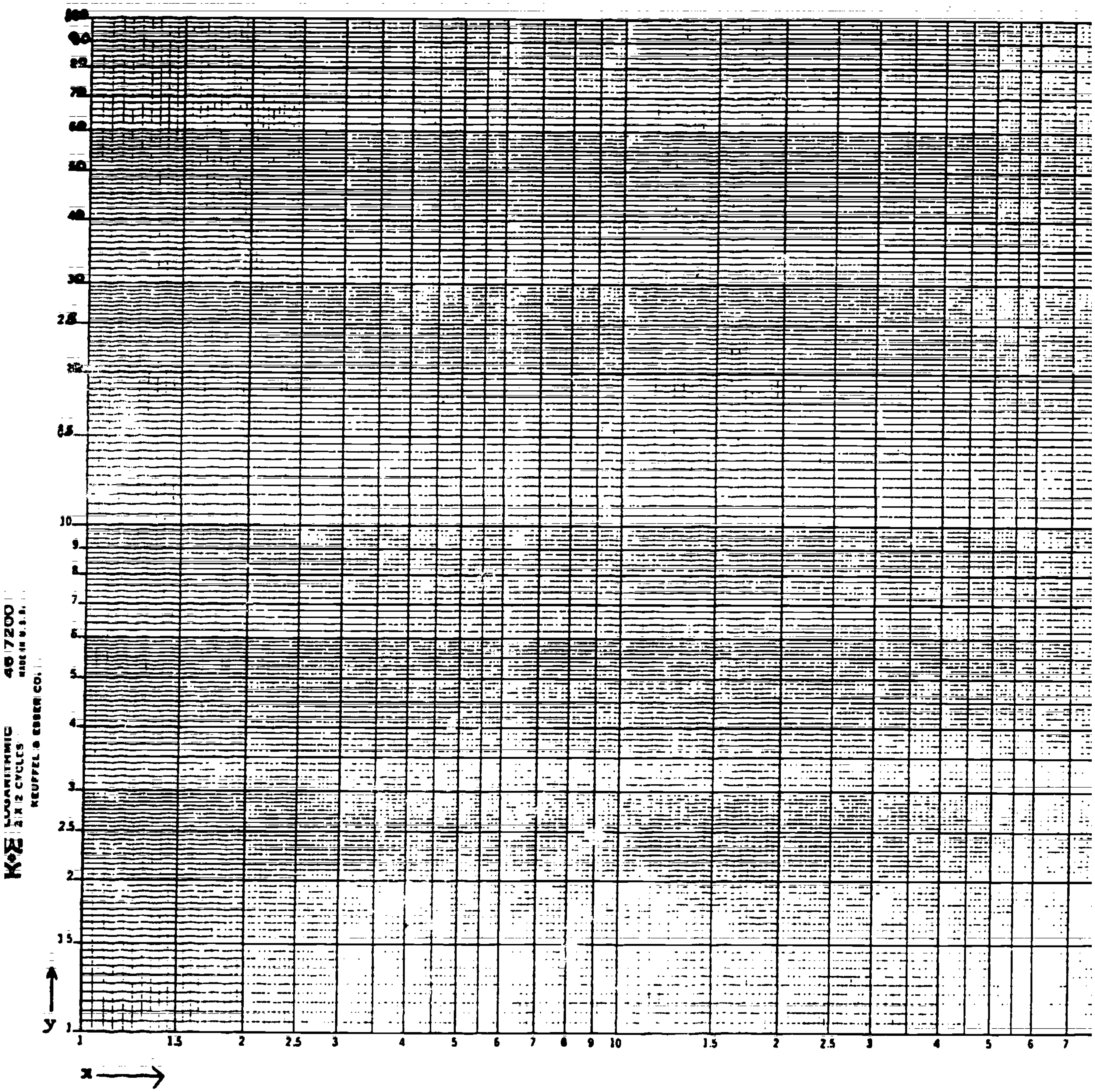


Figure 3

Log-log Paper

References

Blalock, Hubert M. Social Statistics, Second Edition, McGraw-Hill Book Company, 1972.

Hays, William L. Statistics for the Social Sciences, Second Edition, Holt, Rinehart, and Winston, 1973.

Paul, Richard S. and Ernest F. Haeussler, Jr. Introductory Mathematical Analysis: For Students of Business and Economics, Reston Publishing Co., second edition, 1976.

Rosenbach, Joseph B. et.al., College Algebra with Trigonometry, Blaisdell Publishing Co., 1963.

Tukey, John W., Exploratory Data Analysis, Addison-Wesley Publishing Co., 1977.

Zeisel, Hans, Say it with Figures, fifth edition Revised, Harper and Row Publishers, 1968.

All data sets were taken from Statistical Abstract of the United States 1975 U.S. Department of Commerce, Bureau of the Census, Washington, D.C. 1975.

Homework
Prerequisite Inventory, Units 1 and 2

1. $\log_{10} 7.8 + \log_{10} 2.0 - \log_{10} .5 =$

2. $\log_{10} (9.1 + .9) =$

3. $\log_7 1 =$

4. $(5.6 \cdot \log_{10} 100) \cdot (.2 \cdot \log_{10} 1000) =$

Write the solutions to problems 5-8 in scientific notation.

5. $3.524 \cdot 10^2 + .6476 \cdot 10^3 =$

6. $(1.2 \cdot 10^{-4}) \div (.6 \cdot 10^{-6}) =$

7. 47569.532

   a. to five significant digits =

   b. to three significant digits =

8. $(34 \cdot 2 \log_3 27 + 52.9 \cdot \log_8 1) \cdot \log_{12} 144 =$

9. $\dfrac{\log_8 56}{\log_6 56} =$

10. $\log_7 5 \cdot \log_5 49 =$

11. 5/6 of 23 2/3 is

12. 12 is what percent of 300?

13. Change the fraction 7/8 to a percent.

14. Change .1% to a decimal.

15. 38 is 20% of what number?

16. If a man has $1500 in the bank and the annual interest rate is 5%, how much will he have in the bank after one year?

17. Is the square root of 25 a rational or an irrational number?

18. Is $\sqrt[3]{125}$ an integer?

19. Is $(-3)^3$ a positive real number?

20. Can an irrational number ever be an integer?

21. Which of these is $7 \cdot 7 \cdot 7 \cdot 7$?

4^7 $\quad 7^4 \quad 4^3 \quad 14^2$

22. $(5^0)(8^1) =$

23. $(11^3)(11^5) =$

24. $15^5 \div 15^2 =$

25. $\frac{1}{3^{-3}} =$

26. $(-3) - (-7) =$

27. $(-3) - (+3) + (+3) =$

28. $(-3)^7$ equals which one of the following?

3^7 -21 -3^7 -3 3^{-7}

29. Round the following values to integers.

a. 1093.91
b. 0.8
c. $\sqrt{2}$
d. $33.\overline{33}$
e. 0.2
f. -0.956
g. -.001
h. $-1.\overline{77}$

30. The computer has generated calculations on your data that are significant to only 3 digits. Cut the following values to 3 significant digits.

a. $1.0992 \cdot 10^4$
b. $7.7109 \cdot 10^{-3}$
c. $8.0084 \cdot 10^2$

31. If you have negative values in a data batch can you make a logarithmic transformation on the raw data?

32. If you have fractional values in a data batch can you make a square root transformation on the raw data?

33. If a > b and b > c, then which of the following statements is true?

a > c
a - b > c
ab > bc
abc > 0

34. Arrange the following fractions in increasing order: -2/5, -1/2, 1/5.

78

Questions 35-39 pertain to the following data vector:

$$
\begin{aligned}
X_1 &= 1.64\\
&\ 1.72\\
&\ 1.68\\
&\ 1.77\\
&\ 1.56\\
&\ 1.95\\
&\ 1.78\\
&\ 1.91\\
&\ 1.97\\
&\ 1.82\\
&\ 1.85\\
&\ 1.77\\
&\ 1.75\\
&\ 1.93\\
&\ 1.78\\
&\ 1.71\\
&\ 1.63\\
&\ 1.76\\
&\ 1.55\\
&\ 1.66\\
&\ 1.49\\
&\ 1.64\\
&\ 1.70\\
X_{24} &= 1.68
\end{aligned}
$$

This data vector, X, contains the number of inspections in units of a hundred companies conducted by 24 regional federal Occupational Safety and Health Administration offices. Let i denote the office and X_1, X_2 . . . X_{24} denote the number of inspections in hundreds conducted by each office.

35. The actual number of inspections conducted by office X_6 is

36. Office X_4 is in Boston; office X_{23} is in Seattle. How many more inspections did Boston conduct that Seattle?

37. Offices X_1, X_2, . . . , X_6 are in the northeast.
Offices X_7, X_8, . . . , X_{12} are in the southeast.
Offices X_{13}, X_{14}, . . . , X_{18} are in the southwest.
Offices X_{19}, X_{20}, . . . , X_{24} are in the northwest.

What notation would you use to indicate the sum of all the inspections in the southwest?

38. What is the total number of inspections that took place in the southeast?

39. Are these data values discrete or continuous?

40. $\sum_{i=10}^{19} 3 =$

41. $\sum_{i=5}^{24} 4 X_i =$

42. $X_i = 1, 3, 5 \ldots\ldots, 11$

$\sum_{i=2}^{4} X_i^2 =$

For problems 43-46 state whether the variable described is discrete or continuous.

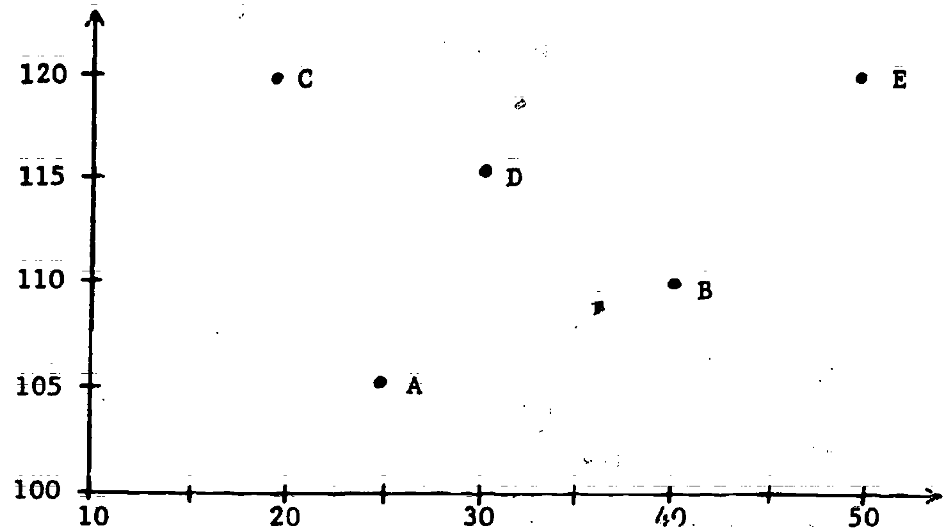
43. The proportion of blacks in each census tract in Pittsburgh.

44. The number of persons living in Pittsburgh that are black.

45. The number of traffic fatalities in the U.S. in 1975.

46. The percentage of vehicle defect caused traffic fatalities in the U.S.

Questions 47-50 refer to the following graph.



47. List the coordinates for points A-E.

48. Order points A-E by increasing values of the abscissa.

49. What is the horizontal distance between points A and E?

50. What is the vertical distance between points C and E?

Homework Solutions
Prerequisite Inventory Units 1 and 2

1. $\log_{10} 31.2$
2. 1
3. 0
4. 6.72
5. $1.000 \cdot 10^3$
6. $2.0 \cdot 10^2$
7. a. $4.7569 \cdot 10^4$
   b. $4.75 \cdot 10^4$
8. $4.08 \cdot 10^2$
9. $\log_8 6$
10. $\log_7 49$ or
11. 19 13/18 or 355/18
12. 4%
13. 87.5%
14. .001
15. 190
16. $1575
17. rational
18. yes
19. no
20. no
21. 7^4
22. 8
23. 11^8
24. 15^3
25. 27
26. 4
27. −3
28. -3^7
29. a. 1094
    b. 1
    c. 1
    d. 33
    e. 0
    f. −1
    g. 0
    h. −2
30. a. $1.09 \cdot 10^4$
    b. $7.71 \cdot 10^{-3}$
    c. $8.00 \cdot 10^2$
31. no
32. yes
33. a > c
34. −1/2, −2/5, 1/5
35. 195
36. 7

37. $\sum_{i=13}^{18} X_i$

38. 1110

39. discrete

40. 30

41. $4 \sum_{i=5}^{24} X_i$

42. 83

43. Continuous

44. Discrete

45. Discrete

46. Continuous

47. A (25,105)
B (40,110)
C (20,120)
D (30,115)
E (50,120)

48. C A D B E

49. 25 units

50. 0

Units 1 and 2
Reading Assignments

Readings should be completed before the indicated lecture.

Lecture	Reading
1-0	"Introduction to QMPM" Tufte, Chapter 1 Tukey & Wilk
Workshop	Prerequisite Inventory Tukey, pp. 1-7
1-1	Tukey, pp. 7-26 McNeil, pp. 1-6 Wallis & Roberts, pp. 177-182
1-2	Tukey, Chapter 2 McNeil, pp. 6-9, 16-17
Workshop	"Some Principles for Graphics of Tables and Charts" Huff, pp. 60-142
1-3	Tukey, Chapter 3 McNeil, pp. 9-16
1-4	Huff, pp. 7-36 Wallis & Roberts, pp. 182-193 Tanur, pp. 229-236
2-1	Tukey, pp. 97-101 McNeil, pp. 27-31 Profiles in School Support, pp. 15-69
2-2	Tukey, pp. 102-115 NcNeil, pp. 32-39 Fairley & Mosteller, pp. 87-109

Texts:

Fairley, William B. and Frederick Mosteller, Statistics and Public Policy, Reading, Mass.: Addison-Wesley Publishing Co., 1977.

Huff, Darrell, How to Lie with Statistics, New York: W.W. Norton & Co., 1954.

McNeil, Donald R., Interactive Data Analysis, New York: John Wiley & Sons, 1977.

"Profiles in School Support," prepared for the National Center for Education Statistics by Eugene P. McLoone, Washington, D.C.: U.S. Government Printing Office, 1974.

Tanur, Judith, et al. Statistics: A Guide to the Unknown, San Francisco: Holden-Day, Inc., 1972.

Tufte, Edward R., Data Analysis for Politics and Policy, Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Tukey, John W., Exploratory Data Analysis, Reading, Mass.: Addison-Wesley, 1977.

Tukey, John W. and M. B. Wilk, Data Analysis and Statistics: An Expository Overview, AFIPS Conference Proceedings, volume 29, 1966, pp. 695-709.

Wallis, W. A. & H. V. Roberts, Statistics: A New Approach, Glencoe, ILL.: Free Press, 1956.

Lecture 1-0. Introduction to QMPM and Unit 1

General Introduction to Quantitative Methods for Public Management and Specific Introduction to Unit 1, Analysis of Single Batches of Data

Lecture Content:

1. Discuss purpose and organization of course, and the nature of data analysis
2. Introduction to the objectives, problem, and notation of Unit 1

Main Topics:

1. Introduction to QMPM--Detailed structure
2. Introduction to Unit 1

Topic 1. Introduction to QMPM--Detailed structure

I. Nature of Data Analysis: Numerical detective work
What does a data analyst do?

1. Take apart data to find structure: Data = Signal (Fit) + Noise (Residual)

2. Familiarity with various forms of data sets and ways of "handling" them

3. Analytical methods to take data apart

4. Exploration preceding confirmation--Detective work (Investigative vs. Judicial evaluative process)

5. Iterative and Interactive process--uses data as guide to procedure

II. Structure of QMPM

1. Rationale: Analysis for decision making
Problems facing data analysts

a. Public managers need to make effective decisions

b. Must be able to process data, present results to nonquantitative audiences

c. Relevant data usually quantitative and "messy" or "dirty" (measurement error, NA values

d. Analyses are usually unplanned, post hoc, second hand

e. Implication--need operational analytic skills that can handle data problems and change data into information

2. Objective: Provide these necessary skills
Students learn to:

a. Gather and prepare data

b. Analyze data to uncover structure and evaluate the analysis

c. Present data and interpret analytic results for improved communication

3. Philosophy: Preparation for practice
   QMPM's emphasis:

   a. Quantitative--numeric data, relating to quantities or measures

   b. Real data--empirical, based on real-life observation or experiment

   c. Policy relevant data--data used in making decisions

   d. Graphics--visual displays or pictures

   e. Resistant and robust techniques--unaffected by deviant values or erroneous assumptions

   f. Data analytic--not just statistical; models for data

   g. Computer orientation--exploits special user-oriented computing system

III. Op[illegible]alization and organization of course: Elaborate structure, requires cooperation between students and instructional staff

1. Instructors--Identify, offices, office hours

2. Module/Unit design

3. Introductory lectures--precede substantive learning units

4. Prerequisite inventories--establish base upon which new skills are built

5. References and texts--describe

6. Computer system--location, staff--if used

7. Video tapes--location, staff--if used

8. DataBank--documentation, staff--if used

9. Calculators--promote purchase

10. Homework--frequency, schedule

11. Workshops--1 per week, flexible role in course

12. Calendar

Topic 2. Introduction to Unit 1, Analysis of Single Batches of Data

I. Introduction to the objectives of Unit 1

1. Questions to be answered in Unit 1

a. What is a batch of numbers?
A set of similar values obtained in some consistent fashion (from Prerequisite Inventory, Module I) (1

b. What analyses can be done on a single batch?
What can we say about a batch?

2. Skills to be mastered in Unit 1 (2)

a. Perceiving and recognizing a batch

b. Organizing a batch to facilitate presentation, comprehension, and analysis

c. Condensing a batch to facilitate summarization

d. Transformations to promote symmetry

e. Definition and recognition of well-behaved batches

II. Introduction to the problems of Unit 1

1. What is a batch? Look at an example.

a. Similar numbers--counts of persons in census tract scale
Example: 1970 populations of the 186 census tracts in Pittsburgh

b. Consistent feature--data collected in 1970 census enumeration of city of Pittsburgh (3)
(Note trailing zeros and tract sequence arrangement; discuss notion of "census tract" and source of census data)
Example: 1970 populations, in thousands of persons (comment on units) (4)

What can we say about a batch? What are its features? How can we summarize it? (5)

a. Minimum value--how small is the smallest value of the batch?

b. Maximum value--how large is the largest value of the batch?

c. Typical value--what is the "average" value of the batch?

d. Variability--how spread out is the batch?

e. Uniformity--how clustered is the batch?

f. Shape--how symmetric is the batch?

3. Example: Population data again (6)

a. Minimum value--334 persons, tract 185

b. Maximum value--7910 persons, tract 95

c. Cannot answer remaining questions

4 example: Number of blacks in each of the census (7)
f Pittsburgh in 1970

Minimum value--0 Blacks, tracts 105, 129, 146, 171, 176, 177, 185

b. Maximum value--4611 Blacks, tract 74

c. Cannot answer remaining questions, but note large numbe of small values

5. Conclusion

a. Need methods to organize data

b. Need tools to summarize important features

c. Methods should be easily performed

d. Summaries should be readily comprehended

III. Introduction to Notation of Unit 1
(See Prerequisite Inventory, Module I, for reference text)

1. Conventions (8)

a. Capital letter ("X") denotes entire single batch of values

b. Individual values identified by single subscripts

2. Example: Pittsburgh population

   a. Let X = Population of Pittsburgh census tracts in 1970

   b. Let X_1 = Population of tract 1 = 972 persons

      Let X_2 = Population of tract 2 = 4082 persons

      $\vdots \qquad\qquad\qquad\qquad \vdots$

      Let X_{185} = Population of tract 185 = 334 persons

      (Note arbitrariness of assignment of tract numbers)

   c. In general, there are n tracts (in this case, n = 185)

   d. Thus $X_n = X_{185}$ = 334 persons

QMPM

Lecture 1-0
Transparency Presentation Guide

Lecture Outline Location	Transparency Number*	Transparency Description
Topic 2		
Section I		
1.a	1 **	Definition of Batch of Data
2	2	Topics for Unit 1
Section II		
1.b	3	1970 Populations of Pittsburgh Census Tracts
1.b	4	1970 Populations, in Thousands of Persons
2	5	Questions to be Answered for Single Batches
3	6	1970 Populations, Maximum & Minimum Indicated
4	7	1970 Black Populations of Pittsburgh Census Tracts, Maximum & Minimum Indicated
Section III		
1-2	8	Conventions & Example of Notation

*Refers to numbers in parentheses on righthand side of lecture outline.

**Bracketed transparencies are [illegible] the same page.

BATCH

A batch of data is a set of similar observations obtained in some consistent fashion.

In Unit 1, we will learn how to analyze single batches of data — batches with only one particular feature.

[2]

Topics for Unit 1

1. Perceiving and Recognizing a Batch
2. Organizing a Batch Using Analytic Tools
3. Condensing a Batch for Description
4. Numerical and Graphical Summaries
5. Determining Transformations for a Batch
6. Defining a Well-Behaved Batch

QMPM

[3]

1970 Populations of Pittsburgh Census Tracts

972.	4082.	1972.	391.	631.	735.	1938.
3062.	2919.	2424.	6887.	729.	3689.	2437.
2085.	2973.	3712.	2505.	1919.	3294.	449.
4187.	1050.	1645.	3629.	453.	1645.	2447.
1867.	1359.	2855.	1876.	2915.	1405.	2386.
2471.	728.	1205.	2382.	3122.	1019.	1410.
765.	2776.	2135.	1349	3628.	4415.	3153.
5747.	2135.	1330.	4730.	2942.	3469.	4095.
2155.	4247.	1472.	3832.	1452.	3378.	1971.
1253.	2316.	3092.	4060.	3945.	2602.	848.
2744.	3228.	[illegible]769.	4858.	4014.	1645.	5300.
5269.	2979.	[illegible]148.	3268.	3133.	4520.	6003.
5319.	5435.	1212.	2041.	2068.	1577.	2084.
152[illegible].	4615.	3994.	7910.	3188.	4392.	4758.
1985.	5203.	484.	5880.	3962.	3752.	1884.
3156.	3578.	2398.	1424.	3820.	2019.	1418.
4719.	3509.	6796.	5371.	5630.	3765.	7425.
996.	2607.	2396	1870.	2574.	603.	2658.
2297.	335.	6235.	3121.	791.	1558.	1343
2670.	1227.	1159.	2579.	569.	588.	442.
3338.	2779.	345.	4327.	2987.	2254.	2569.
1792.	2932.	2125.	1056.	2[illegible]5.	1963.	719.
3103.	2487.	1193.	3291.	1044.	4561.	3609.
1289.	3853.	3905.	2857.	4437.	1399.	2144.
3812.	2612.	1640.	3921.	992.	1906.	3425.
6242.	5818.	6527.	955.	5300.	1289	2829
3297.	3413.	334.				

1.0

[4]

1970 Populations of Pittsburgh Census Tracts
Data in Thousands of Persons

0.972	4.082	1.972	0.391	0.631	0.735	1.938
3.062	2.919	2.424	6.887	0.729	3.689	2.437
2.085	2.973	3.712	2.505	1.919	3.294	0.449
4.187	1.05	1.645	3.629	0.453	1.645	2.447
1.867	1.359	2.855	1.876	2.915	1.405	2.388
2.471	0.728	1.205	2.382	3.122	1.019	1.41
0.765	2.776	2.135	1.349	3.628	4.415	3.153
5.747	2.135	1.33	4.73	2.942	3.469	4.095
2.155	4.247	1.472	3.832	1.452	3.378	1.971
1.253	2.316	3.092	4.06	3.945	2.602	0.048
2.744	3.226	3.769	4.858	4.014	1.645	5.3
5.269	2.979	5.148	3.268	3.133	4.52	6.003
5.319	5.435	1.212	2.041	2.068	1.577	2.034
1.521	4.615	3.994	7.91	3.188	4.392	6.758
1.985	5.203	0.484	5.88	3.962	3.752	1.884
3.156	3.578	2.398	1.424	3.82	2.019	1.418
4.719	3.509	6.796	5.371	5.63	3.765	7.425
0.996	2.607	2.396	1.87	2.574	0.603	2.658
2.297	0.335	6.235	3.121	0.791	1.558	1.343
2.67	1.227	1.159	2.579	0.569	0.588	0.442
1.338	2.779	0.345	4.327	2.987	2.254	2.569
1.792	2.932	2.125	1.056	2.325	1.963	0.719
3.103	2.487	1.193	3.291	1.044	4.561	3.609
1.289	3.853	3.905	2.857	4.437	1.399	2.144
3.812	2.612	1.64	3.921	0.992	1.986	3.425
6.242	5.818	6.527	0.955	5.3	1.289	2.829
3.297	3.413	0.334				

1-0

Questions to be Answered for Single Batches

1- Minimum Data Value

2- Maximum Data Value

3- Typical Data Value

4- Variability of the Data Values

5- Uniformity

6- Shape

[6.

1970 Populations of Pittsburgh Census Tracts

Maximum and Minimum Values of Batch Indicated

972.	4082.	1972.	391	631.	735.	1938.
3062.	2919	2424.	6887.	729.	3639.	2437.
2085.	2973.	3712.	2505.	1919.	3294.	449.
4187.	1050.	1645.	3629.	453.	1645.	2447.
1867.	1359.	2855.	1876.	2915.	1405.	2388.
2471.	728.	1205.	2382.	3122.	1019.	1410.
765.	2776.	2135.	1349.	3628.	4415.	3153.
5747.	2135.	1330.	4730.	2942.	3469.	4095.
2155.	4247.	1472.	3832.	1452.	3378	1971.
1253.	2316.	3092.	4060	3945.	2602.	848.
2744	3228.	3769.	4858.	4014.	1645.	5300.
5269.	2979.	5148.	3268.	3133.	4520.	6003.
5319.	5435.	1212.	2041.	2065.	1577.	2084.
1521.	4615.	3994.	$\boxed{7910.}$ max	3188.	4392.	4758.
1985.	5203.	484.	5880.	3962.	3752.	1884.
3156.	3578.	2398.	1424.	3820.	2019.	1418.
4719	3509.	6796.	5371.	5630.	3765.	7425.
996.	2607.	2396.	1870.	2574.	603.	2658.
2297.	335.	6235.	3121.	791.	1558.	1343.
2670	1227.	1159.	2579.	569.	588.	442.
3338.	2779.	345.	4327.	2987.	2254.	2569.
1792.	2932.	2125.	1056.	2325.	1963.	719.
3103.	2487.	1193.	3291.	1044.	4561.	3609
1289.	3853.	3905.	2857.	4437.	1399.	2144.
3812.	2612.	1640.	3921.	992.	1906.	3425.
6242.	5818.	6527.	955.	5300.	1289.	2829.
3297.	3413.	$\boxed{334.}$ min				

1.0

QMPM

[7]

1970 Black Populations of Pittsburgh Census Tracts

Maximum and Minimum
Values of Batch Indicated

41.	620.	114.	149.	550.	58.	1837.
2794.	2281.	404.	261.	34.	67.	37.
491.	891	3657	2403.	1838.	3120.	279.
3657.	44.	1626.	3432.	365.	180.	414.
116.	247	27	59.	109.	287.	73
42.	87.	17.	15.	50.	642.	18
6.	17.	32.	196.	9.	3	21
14.	145.	5.	34.	17.	2131.	977
17	65	60	970.	940.	131.	40.
212.	880.	2572.	2683.	2816.	1448.	582.
2513.	3129.	3686.	4611. max	3831.	1530	3941
86.	90.	66.	29.	800.	18.	102
40	34.	1.	18.	772.	46.	152.
1177.	579.	80.	39.	30.	48.	183.
1688.	3118.	18.	50.	128.	88.	0. min
18.	2088.	2027.	15.	67.	3.	1
169.	3.	232.	12.	10.	8.	1.
108.	428.	169.	7.	91	6.	162.
53	108.	0. min	272	8.	1076.	475.
2035.	636.	58.	320.	130.	24.	12.
38.	44	23.	103.	3.	0. min	501.
1061.	2081	514.	339.	140.	4.	184.
763.	159.	81	27.	21	3304.	379.
1.	22.	19.	19	4.	14.	12.
248.	8.	0. min	15.	52.	7.	6
0. min	0. min	342	2.	20.	115.	5.
5.	109.	0. min				

98

1.0

[8]

Let X denote

the single batch
of total populations of each
of the 185 census tracts in Pittsburgh in 1970:

X = population of the Pittsburgh Census Tract:

Identifying Individual Observations:

X_1 = population of first census tract = 972

X_2 = population of second census tract = 4082

⋮ ⋮

X_{185} = population of 185th census tract = 834

X_i = population of some arbitrary ith census tract

Lecture 1-1. Organization for Analysis

Organization for Analysis: The Use of Numeric and Graphic Methods for Analytical Organization of Single Batches (1)

Lecture Content:

1. Discuss methods for recording and presenting a batch of data in an organized manner.

2. Show how such tools convey various batch characteristics.

Main Topics:

1. Methods for organizing a batch

2. Questions to ask of a batch

Tools Introduced:

1. Sorted batch

2. Histogram

3. Stem-and-Leaf Display

Topic 1. Methods for Organizing a Batch

I. Basic issue: Organization of data

1. Arbitrary--the manner in which data are usually gathered, recorded or transmitted

    a. At the data collector's discretion--i.e., a matter of convenience

    b. Contextually defined--e.g., stations on transit line

    c. May depend on data gathering procedure--e.g., census

    d. Obscures behavior of batch values

    e. Makes summarization and analysis difficult

2. Analytical--the manner in which we desire to arrange data

    a. Consistent

    b. Context free

    c. Reliable

    d. Conveys behavior of batch values: shape, spread, location, outliers

    e. Simplifies continued analysis of the batch

II. Problem: Analyst often must use data which come arbitrarily organized

1. Arbitrarily organized data are unwieldy

2. Such data do not permit ready description

3. Such data do not permit conclusions to be drawn about batch behavior--cannot get a "feel" for the batch

III. Solution: Simple and understandable tools for analytical organization

1. Simplest method--Sorted batch (2)

2. Classical method--Histogram (3)

3. Exploratory method--Stem-and-Leaf display (4)

IV. Methods

Only analytical organization of data is discussed in this lecture. The techniques covered can be applied to all types of batches of data. They are visual displays that are easily appreciated cognitively and help the data analyst by addressing the problem of what to examine in a batch. Universal rules for reliable and quick construction are presented.

1. Sorted batch: a simple organization, an array of values, ordered from smallest to largest

    a. Example shows a sorted batch: 1970 populations of Pittsburgh census tracts

    b. Features

        i. Simple idea

        ii. Retains information on individual values

        iii. Operationally difficult to construct

    c. Analytic qualities

        i. Largest and smallest values identifiable

        ii. Ability to locate order statistics (explain "counting in")

    d. Procedure: arrange data in increasing order

    e. Sorted batch constructed by computer:

    In the session introducing CMU-DAP system.

2. Histogram: A bar graph which visually presents some of the information in a batch

    a. Example: Histogram of 1970 populations of Pittsburgh census tracts

    b. Features

        i. Reasonably interpretable

        ii. Common technique

        iii. Formal definition

iv. Loses information on individual values

v. Operationally difficult to construct

c. Analytic qualities

i. Shape--separation, symmetry, irregularity, and clustering of values

ii. Spread--variation of values

d. Procedure:

(Draw histogram of 1970 populations of Pittsburgh census tracts on blackboard, explaining each step.)

i. Draw vertical (y) and horizontal (x) axes on a sheet of ordinary graph paper

ii. On horizontal axis, mark off smallest data value and highest data value in batch, using the scale of the axis; in this case, 0.3 and 7.9 thousand

iii. Divide this interval into the desired number of "bins" of equal size for display. It may be necessary to round the smallest value down and the largest value up to obtain a convenient width for each subinterval. For these data, use 8 bins of width 1000.

iv. Record number of data values falling into each bin. This information is needed to determine height of each bar.

v. Mark off vertical axis to correspond to number of data values per bin

vi. Draw in bars

vii. Can also have intervals of unequal size, and can combine intervals to produce a "squeezed" version or break up intervals to produce a "stretched" version.

e. Histogram constructed by computer:

In session introducing the CMU-DAP system

3. Stem-and-Leaf display: An easy and versatile method of organizing a batch into roughly numerical order. (4)

   a. Example: Stem-and-Leaf of 1970 Pittsburgh census tract populations in thousands of persons

   b. Features

      i. "Face validity"

      ii. Retains information on individual data values (display and storage versions)

      iii. Many versions--flexible

      iv. No formal rules for "correct" version

      v. Operationally easy to construct

   c. Analytic qualities

      i. Largest and smallest data values

      ii. Location of order statistics

      iii. Shape

      iv. Spread

   d. Procedure:

      (Work through an example on the blackboard.)

      i. Choose a convenient unit, or power of ten, for the display

      ii. Every data value in the batch is cut to a whole multiple of the unit

      iii. Separate each value for the display into a stem and a leaf

      iv. Find the largest and smallest stems

      v. Write down these stems and all the intervening stems in a vertical column

      vi. Use asterisks (*) to indicate the number of digits represented in a leaf

vii. Draw a vertical line

viii. Place the leaves on the line corresponding to the correct stem

e. Another example:

Net migration for Pennsylvania counties in percent of population from 1970 to 1974 (5)

(Do Stem-and-leaf of batch on blackboard)
(Set aside the high outliers)
(Unit = 0.1%, single stems)

i. The outlying counties have been set aside--the causes of their large increase in population should be investigated. Counties are: Monroe, Pike, Wayne, Wyoming

ii. If the display appears too "squeezed", we can increase the number of lines per stem from 1 to 2 or 5

f. Another example:

Stem-and-leaf of the Pittsburgh populations where each stem now has 2 lines. (6)

Use "*" and "." to split the stem for leaves 0-4 and 5-9.

i. The first line, labelled *, holds leaves 0-4.

ii. The second line, labelled, ., holds leaves 5-9

g. Another example:

Stem-and-leaf for the Pittsburgh populations where each stem now has 5 lines. (7)

i. Line labelled * holds leaves 0-1

ii. Line labelled t holds leaves 2-3 ("two" and "three")

iii. Line labelled f holds leaves 4-5 ("four" and "five")

iv. Line labelled s holds leaves 6-7 ("six" and "seven")

v. Line labelled . holds leaves 8-9

vi. If a display appears too "stretched", change the unit by decreasing it, and decrease the number of stems per line

(Compare net migration stem-and-leaf displays on two different scales.) (8,9)

vii. Choose the "best" display by controlling the maximum number of leaves per line

Rough rule: max leaves/line $\approx 10 \log_{10} N$

h. Stem-and-Leaf display constructed by computer:

In the session introducing the CMU-DAP computing system.

106

Topic 2. Questions to Ask of a Batch

I. Basic Issue: Once organized, what can we learn from a single batch?

II. Try to answer the following questions which relate to the batch values: (10)

1. Do the values cluster or are they uniformly spread?

2. Are there any deviant values, outliers?

3. Is the batch symmetrical or asymmetrical?

4. Are the values widely spread out?

5. Are there any separations in the display?

6. What are the order statistics of the batch?

7. Where is the "center" of the batch?

III. Methods related to questions:

Stem-and-leaf displays permit us to answer all of these questions. Histograms do not answer (2), (6) or (7) completely, since individual data values cannot be identified. Sorted batch does not answer (3), (4) or (7) and (2) and (5) are difficult to answer.

(Discuss appearance of each of the next 2 slides; answer questions) (11-12)

(Suggest the use of answers to questions in II as attempts to summarize batch. Summary is facilitated by analytically organizing the batch.)

Lecture 1-1
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Title, Content, and Topics of lecture
Topic 1		
Section III		
1	2	Sorted Batch
2	3	Histogram
3	4	Stem-and-Leaf Display
Section IV		
1	2	Sorted Batch
2	3	Histogram
3	4	Stem-and-Leaf Display
3.e	5	Net migrations for Pennsylvania Counties, 1970-74
3.f	6	Stem-and-Leaf Display of Pittsburgh Populations
3.g	7	Stem-and-Leaf Display of Pittsburgh Populations
3.g.vi	8	Stem-and-Leaf Display of Net Migrations, Unit = 0.1%
3.g.vi	9	Stem-and-Leaf Display of Net Migrations, Unit = 1%
Topic 2		
Section I	10	Questions for Single Batches
Section III	11	Stem-and-Leaf Display for Average Education
	12	Stem-and-Leaf Display for Percentage in Poverty

108

Lecture 1-1 [1]

Organization for Analysis:

The use of Numeric and Graphic tools for organizing batches.

Lecture Content:

Methods for presenting a batch of data in an organized manner to convey a variety of characteristics of the batch, such as: a typical value, shape, and outliers.

Main Topics:

1) Methods for organizing a batch.
2) Questions to ask of a batch.

[2]

Simplest Method-
A Sorted Batch

334.	335.	345.	391.	442.	449.	453.
484.	569.	588.	603.	631.	719.	720.
729.	735.	765.	791.	848.	955.	972.
992.	996.	1019.	1044.	1050.	1056.	1159.
1193.	1205.	1212.	1227.	1253.	1289.	1289.
1330.	1343.	1349.	1359.	1399.	1405.	1410.
1418.	1424.	1452.	1472.	1521.	1558.	1577.
1640.	1645.	1645.	1645.	1792.	1867.	1870.
1876.	1884.	1906.	1919.	1938.	1963.	1971.
1972.	1985.	2019.	2041.	2068.	2084.	2085.
2125.	2135.	2135.	2144.	2155.	2254.	2297.
2316.	2325.	2382.	2388.	2396.	2398.	2424.
2437.	2447.	2471.	2487.	2505.	2569.	2574.
2579.	2602.	2607.	2612.	2658.	2670.	2744.
2776.	2779.	2829.	2855.	2857.	2915.	2919.
2932.	2942.	2973.	2979.	2987.	3062.	3092.
3103.	3121.	3122.	3133.	3153.	3156.	3188.
3228.	3268.	3291.	3294.	3297.	3330.	3378.
3413.	3425.	3469.	3509.	3570.	3609.	3628.
3629.	3689.	3712.	3752.	3765.	3769.	3812.
3820.	3832.	3853.	3905.	3921.	3945.	3962.
3994.	4014.	4060.	4002.	4095.	4107.	4247.
4327.	4392.	4415.	4437.	4520.	4561.	4615.
4719.	4730.	4750.	4858.	5140.	5203.	5269.
5300.	5300.	5319.	5371.	5435.	5630.	5747.
5818.	5880.	6003.	6235.	6242.	6527.	6796.
5887.	7425.	7910.				

1-1

[3]

Classical Method:

Histogram

1970 Populations of Pgh. Census Tracts

1-1

[4]

Exploratory Method:
A Stem-and-leaf Display

```
0 * * * | 93 67 [illegible] 77 [illegible] 37 55 93 79 83
1       | [illegible] 93 39 [illegible] 26 28 57 84 48 53 21 70 91 02 36 92
2 * * * | [illegible] 40 95 [illegible] 93 43 71 19 13 67 90 03 06 35 62 65 79 25 91 34 81 8
3       | [illegible] 26 16 [illegible] 83 09 27 21 [illegible] 97 15 85 71 31 26 89 89 42 4
4 * * * | 01 47 02 [illegible] 05 63 [illegible] 35 4
5       | 73 21 39 20 36 23
6 * * * | [illegible] 72 25
7       | [illegible]
```

For: Pgh. Populations of Census Tracts
Unit = 100 persons (values cut)

[5]

Net Migration, in Percent, of Population of Penna. Counties, 1970 to 1974.

County	% Net Migration	County	% Net Migration	County	% Net Migration
Adams	6.0	Elk	-2.5	Montour	-1.0
Allegheny	-5.1	Erie	0.7	Northampton	2.7
Armstrong	-0.6	Fayette	0	Northumberland	1.4
Beaver	-1.3	Forest	-0.8	Perry	6.4
Bedford	0.3	Franklin	0.8	Philadelphia	-6.8
Berks	1.6	Fulton	3.5	Pike	15.5
Blair	-0.8	Greene	1.7	Potter	0.8
Bradford	-0.2	Huntingdon	1.1	Schuylkill	0.8
Bucks	3.7	Indiana	2.3	Snyder	4.6
Butler	3.2	Jefferson	5.1	Somerset	0
Cambria	-2.3	Juniata	4.8	Sullivan	-3.3
Cameron	-5.8	Lackawanna	-0.1	Susquehanna	2.0
Carbon	3.0	Lancaster	2.5	Tioga	3.1
Centre	3.2	Lawrence	-1.2	Union	5.1
Chester	0.9	Lebanon	2.1	Venango	0.8
Clarion	3.3	Lehigh	1.8	Warren	-2.7
Clearfield	0.9	Luzerne	1.9	Washington	0
Clinton	-0.1	Lycoming	0.4	Wayne	8.4
Columbia	4.8	McKean	-2.9	Westmoreland	-1.2
Crawford	2.2	Mercer	-0.4	Wyoming	13.4
Cumberland	3.2	Mifflin	-3.2	York	1.4
Dauphin	-0.4	Monroe	13.8		
Delaware	-3.8	Montgomery	-0.2		

Stem-and-Leaf Display of 1970 Populations of Pittsburgh Census Tracts

```
(UNIT = 10**2)
   0* | 33334444
   0. | 556677777789999
   1* | 00001122222233333444444
   1. | 55566666788889999999
   2* | 00000111112233333344444
   2. | 5555666666777888999999
   3* | 001111111222222333444
   3. | 55666667777888899999
   4* | 0000123344
   4. | 5567778
   5* | 12233334
   5. | 6788
   6* | 022
   6. | 578

   HI |  7425.  7910.
```

114

[7]

Stem-and-Leaf Display
for Pittsburgh Populations

Unit = 100 persons
(values cut)

```
    t | 3333
    f | 444455
    s | 66777777
 0•   | 89999
 1*** | 000011
    t | 22222233333
    f | 444444555
    s | 66667
 1•   | 88889999999
 2*** | 0000011111
    t | 22333333
    f | 444445555
    s | 66666777
 2•   | 8889999999
 3*** | 001111111
    t | 22333333
    f | 44455
    s | 66667777
 3•   | 888899999
 4*** | 00001
    t | 233
    f | 4455
    s | 6777
 4•   | 8
 5*** | 1
    t | 223333
    f | 4

  HI  | 5600, 5700, 5800, 5800, 6000, 6200, 6200,
      | 6500, 6700, 6800, 7400, 7900
```

Stem-and-Leaf Display [8]
of Pennsylvania Counties' Net Migration

Unit = 0.1 % (values cut)

```
-6 | 8
-5 | 1 8
-4 |
-3 | 8 2 3
-2 | 5 9 7
-1 | 3 2 0 2
-0 | 6 8 2 3 1 4 8 1 4 2
 0 | 3 9 9 7 0 8 4 8 8 0 8 0
 1 | 6 7 1 8 9 4 4
 2 | 2 5 1 7 0
 3 | 7 2 0 2 3 2 5 3 1
 4 | 8 8 6
 5 | 1 1
 6 | 0 4

HI | 13.8  18.5  8.4  13.4
```

Stem-and-Leaf Display [9]
of Pennsylvania Counties' Net Migration

Unit = 1 % (values cut)

```
    s | 6
    f | 55
    t | 333222
-0  * | 1111 0000000000
 0  * | 000000000000 1111111
    t | 22222333333333
    f | 44455
    s | 66
 0  . | 8
 1  * |
    t | 33
    f |
    s |
 1  . | 8
```

116 1-1

Questions for Single Batches

1. Do the values cluster or are they uniform?

2. Are there any deviant values, outliers?

3. Is the batch symmetrical or asymmetrical?

4. Are the values widely spread out?

5. Are there any separations?

6. What are the order statistics?

7. Where is the center?

1-1

Stem-and-Leaf Display [11]
of 1970 Average Years of Education for Individuals in Pittsburgh Census Tracts

```
(UNIT = 10**-1)
   8  | 22
   8. | 5677778
   9  | 00112222333
   9. | 5555556667778888888
  10  | 0000000000111111223333
  10. | 5555555666667777778
  11  | 000000111111122222333
  11. | 55555566666777888
  12  | 0000000000000000000000011111111112222223334
  12. | 55555556666888
  13  | 12
  13. | 5
  14  | 01
  14. | 88

  HI  |     15.69    16.0
```

Stem-and-Leaf Display [12]
of Percent of Individuals with 1970 Annual Incomes below the Poverty Level in Pgh. Census Tracts

```
(UNIT = 10**0)
  0  | 11
   T | 22223333
   F | 4444444555555
   S | 6666666666666777777777777777
  0. | 88888888899999999
  1  | 0000000011111111
   T | 2222222222223333333
   F | 4444555555
   S | 66666777
  1. | 889999999
  2  | 000011
   T | 233
   F | 444555555
   S | 77
  2. | 99
  3  | 111
   T | 222
   F | 4444
   S | 66
  3. | 0889
  4  | 0011
   T | ?

  HI |     46.00    46.00    49.30    49.70    53.00    59.60    74.20
```

118

i-1

Lecture 1-2. Condensation for Description

Condensation for Description: The Use of Numeric and Graphic Methods (1) to Describe the Information Contained in Single Batches

Lecture Content:

1. Discuss methods for both condensing a batch and presenting the condensed "summaries"

2. Show how such tools effectively describe the batch

Main Topics:

1. Condensing a batch to a small set of numbers

2. Adequacy of these summaries to describe a batch

Tools Introduced:

1. 5-number summary

2. Simple schematic plot

3. Expanded number summary and schematic plot

Topic 1. Condensing a Batch to a Small Set of Numbers and a Graphic

I. Basic issue: Condensation of a batch

1. Condensation is "second-order" summarization--less information is retained than organization techniques

   a. Stem-and-Leaf display and histogram give too much detail

   b. Seek several easily obtained numbers which convey some of the detail of the organization tools

   c. More expedient to "describe" the batch with these "number summaries" than with the entire stem-and-leaf

   d. Schematic plots are efficient mnemonic devices

2. Condensation causes a loss in information except in special instances when the batch can be reconstructed with knowledge of only a few values, i.e. when the batch is well-behaved

II. Problem: Organization tools are not convenient summaries of a batch

1. Usually, organized batch retains too much information. Less may be more useful

2. Condense batch to quantify answers to these questions:

   a. What is a typical value of the batch?

   b. How much variation is present in the values of the batch?

3. Condensations must be easy to obtain, effective in their summarization, readily interpreted, communicated, and remembered

III. Solutions: Simple and expanded numeric and graphic summaries of a batch

1. Simple numerical method--5-number summary

2. Simple graphical method--Simple Schematic ("box-and-whisker") plot

3. Expanded methods--Expanded number summary and schematic plot

IV. Methods: Numeric and Graphic presentations of order statistics

1. Notion of order, depths, folding, and "counting in"

(Note: Distinguish between order statistics and actual data values)

2. 5-number summary: Simple condensation (2)

a. Example: shows a 5-number summary of 1970 populations of Pittsburgh census tracts

b. Features

i. Displays of some order statistics--median, max, min, hinges (quartiles)

ii. Adequately conveys characteristics of most batches

iii. Computable, with some difficulty, from sorted batch

iv. Easily computed from stem-and-leaf

v. Cannot be computed from histogram

vi. Does not give sufficient detail for a large or asymmetric batch

c. Analytic qualities

i. Largest and smallest data values ("extremes") contained in summary

ii. Median, or middle value of batch, included as a typical value

iii. Hinges (quarters), or medians of the two halves of batch, included

d. Procedure

i. Add a column of cumulative counts, or "depths", to a stem-and-leaf display of the batch

ii. Depths should be cumulated from first line down (3)
and from last line up, toward the middle

iii. Stop cumulating when cumulative counts in each direction are roughly equal

iv. First number in summary is smallest value in batch, minimum, which has a depth of 1. Label the minimum "E" (4)

v. Last number in summary is largest value in batch, maximum, which has a depth of 1. Also label the maximum "E"

vi. Third number in summary is middle value in batch, median. Which has a depth of (N+1)/2. Median is defined as single middle value of batch (N odd) or mean of two middle values (N even), and is labelled "M"

vii. Second and fourth numbers in summary are hinges, medians of the two halves of batch. Hinges have a depth of (Depth of M + 1)/2, and are labelled "H"

viii. Arrange the 5-number summary vertically with 3 columns: (5)

Column 1 = Depths
Column 2 = Letter Abbreviations (E, H, M)
Column 3 = Values

ix. Tukey calls 5-number summary a "letter value display"

e. Another "view": Information contained in 5-number summary (6)

i. Useful measure of "spread" of batch is midspread. (7)
Computable from number summary
Midspread = Upper Hinge (UH) - Lower Hinge (LH)
Variability of batch varies directly with midspread

ii. 25% of batch is less than LH, 25% greater than UH Hence 50% of batch lies between hinges (6)

iii. Median (M) is a typical or "central" value of the batch. Half of the batch is less than M, and half is greater. We will use the median as the "average" value of the batch

iv. Range of the batch is also a measure of spread
Range = Upper Extreme - Lower Extreme

f. Another example: Symmetric batch

i. Symmetric batch has median lying halfway between the hinges, halfway between extremes, and halfway between any other "folds" (8)

ii. Any batch where median is not exactly halfway between the hinges or extremes is not symmetric--it is asymmetric

g. Another example: Median incomes for families and unrelated individuals in Pittsburgh census tracts, 1970 (9)

(Compute 5-number summary from sorted batch) (10)
(Compute 5-number summary from stem-and-leaf of batch)
(Try to compute 5-number summary from histogram of batch)

h. 5-number summary constructed on computer:

In the session introducing the CMU-DAP computing system

2. Simple Schematic Plot: Graphical presentation of 5-number summary (Tukey calls this tool a "box-and-whisker" plot)

a. Example: Schematic plot for median incomes for Pittsburgh census tracts in 1970 (11)

b. Features

i. Extremely useful in discussing appearance of batch

ii. Some attributes of batches, such as symmetry, best conveyed by this graphical tool

iii. May be difficult to recover the exact values of the 5-number summary from the plot

c. Analytic Qualities

   i. Made on ordinary graph paper

   ii. y-axis represents values in batch

   iii. Extremes, hinges, and median clearly marked

   iv. Shape and spread of batch easily seen

d. Procedure

   i. Draw a box that stretches from hinge to hinge, crossing with a bar at the median

   ii. Draw a line, or "whisker", from the box to each extreme

   iii. Examine length of box for information on spread of batch

   iv. Examine location of bar within box, and box between extremes for information on symmetry of batch

   v. Examine length of whiskers for information on outliers

e. Another example: Pittsburgh populations with the outliers indicated (12)

   Median lies halfway between hinges, but large number of outliers makes the batch asymmetric

f. Another example: Schematic plot of net migrations of Pennsylvania counties, 1970-1974 (13)

   Schematic plots need not be made vertically on graph paper. Plots can be drawn horizontally on regular paper.

3. Expanded number summaries and schematic plots--Adequate condensation for large batches (N > 100)

   a. Example: Schematic plot of Pittsburgh median incomes (14)

b. Features

i. Emphasizes the outliers in the batch

ii. Very effective in condensing the batch

iii. Expanded summary best presented as a schematic plot

c. Analytic Qualities

i. Define fences beyond the hinges to identify outliers

ii. Outliers suitably indicated on the schematic plot

iii. Shape, spread, and outliers of batch easily seen

d. Procedure

i. Introduce further descriptive numbers (15)

Step = 1.5 x Midspread
Inner fence (f) = Hinge +/- 1 step
Outer fence (F) = Hinge +/- 2 steps
Adjacent values are data values closest to, but still inside the inner fences

ii. Data values between the inner and outer fences are "outside" and are marked on the plot with circles (16)

iii. Data values beyond the outer fence are "far out" and are marked on the plot with squares

iv. Whiskers on the plot should be dashed, ending with dashed crossbars at the adjacent values

v. Far out values should be labelled on the plot in capital letters

vi. Outside values and adjacent values should be labelled on the plot in small letters

vii. Tukey recommends the use of a "Fenced-Letter Display", to reduce clutter (17)

(Note: These definitions of outside values may not be sufficient in certain cases. Deciding whether a value is deviant is usually a subjective process. These techniques help identify outliers but should not replace common sense.)

e. Expanded number summary and schematic plot constructed on computer:

In the session introducing the CMU-DAP computer system.

126

Topic 2. Adequacy of These Summaries in Describing a Batch

I. Basic Issue: Once batch is condensed, how effectively do the summaries describe it?

II. Features of batch that must be included in condensation:

1. Identification of typical value

2. Determination of spread of batch

3. Location of outliers

4. Maximum, Minimum, and Range of batch

(Create 3 or 4 examples of schematic plots from data sets, and present them either on the blackboard or as transparencies. Discuss the appearance of each, indicating how the above necessary features are documented.)

Lecture 1-2
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 1-2 Outline
Topic 1 Section IV		
1.a	2	5-number summary
1.d	3	Stem-and-Leaf display with depths
1.d.iv	4 (overlay 3)	5-number summary located
1.d.vii	5	5-number summary for Pittsburgh populations
1.d.ix	2	Letter-value display
1.e	6	Information contained in 5-number summary
1.e.i	7	Midspread
1.e.ii	6	Information contained in 5-number summary
1.f	8	Symmetric Batch
1.g	9	1970 Pittsburgh Median Incomes
1.g	10	5-number summary of median incomes
2.a	11	Simple schematic plot
2.e	12	Schematic plot, 5-number summary, stem-and-leaf
2.f	13	Schematic plot of Pennsylvania Net Migrations
3.a	14	Schematic plot of Pittsburgh Median Incomes

3.d.i	15	Numbers for Expanded Summary
3.d.ii	16	Anatomy of a Schematic Plot
3.d.vii	17	Tukey's Fenced-Letter display

Lecture 1-2 [1]

Condensation for Description

The use of numeric and graphic methods to describe the information contained in single batches

Lecture Content

Discuss methods of condensing a batch in order to describe the information contained in the batch

These "condensed summaries" must effectively convey this information

Main Topics

1. Condensing the batch to a small set of numbers
2. Adequacy of these summaries in describing a batch

130

[2]

Letter-Value Display

Five Number Summary for 1970 Populations of Pittsburgh Census Tracts

#185

M	93	2600	
H	47	1500	3700
E	1	300	7900

[3]

Stem-and-Leaf Display of the 1970 Populations of Pittsburgh Census Tracts

Unit = 100 persons (values cut)

Depths		
8	0·	33334444
23	0***	556677777789999
46	1·	00001122222233333444444
65	1***	5556666788889999999
88	2·	00000111112233333344444
110	2***	5555666667778889999999
75	3·	00111111122222233444
56	3***	5566667777888899999
37	4·	0000123344
27	4***	5567778
20	5·	12233334
12	5***	6788
8	6·	022
5	6***	578
2	7·	4
1	7***	9

XVI.I.83

5-number Summary

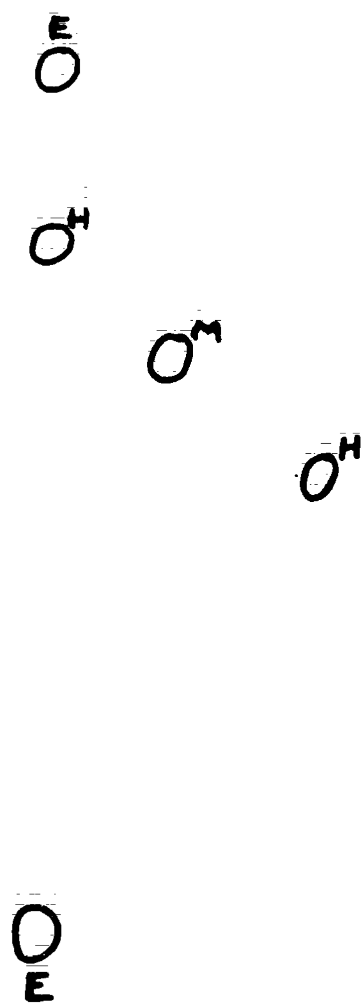
E 300

H 1500

M 2600

H 3700

E 7900



[5]

5-number Summary
for 1970 Pittsburgh Census Tracts' Populations

Depth		cut values		with all digits
1	E	500	Extreme (minimum)	534
47	H	1500	Lower Hinge	1521
93	M	2600	Median	2607
47	H	3700	Upper Hinge	3769
1	E	7900	Extreme (maximum)	7910

E = extreme

depth of one

minimum and maximum of batch

M = median

depth of $\frac{N+1}{2}$

50% of batch lies on either side

H = hinge

depth of $\frac{\text{depth of median} + 1}{2}$

halfway from each extreme to median

1-2

[6]

Information Contained in a 5-Number Summary

Midspread

Range

E	Minimum
H	Lower Hinge
M	Median
H	Upper Hinge
E	Maximum

50% of batch

50% of batch

50% of batch

[7]

<u>Midspread</u> or H-spread

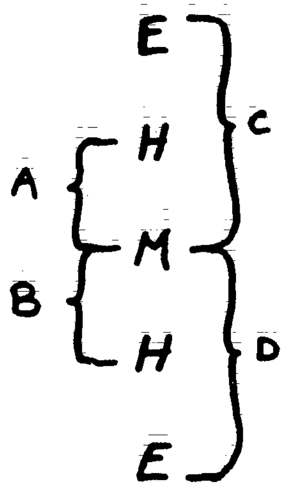
is equal to the difference between the hinges

Midspread = UH - LH

where UH = upper hinge
LH = lower hinge

Example: Midspread of 1970 Pittsburgh populations = 2200 persons

Symmetric Batch



Median located
halfway between the hinges
and
halfway between the extremes.

$A = B \; (UH - M = M - LH)$

$C = D \; (E - M = M - E)$

Example of a Symmetric Batch:

E	-40
H	-10
M	10
H	30
E	60

$\left.\begin{array}{l} UH - M = 20 \\ M - LH = 20 \end{array}\right\}$ Median halfway between hinges

$\left.\begin{array}{l} E - M = 50 \\ M - E = 50 \end{array}\right\}$ Median halfway between extremes

[9]

1970 Median Incomes of Families and Unrelated Individuals for Pittsburgh Census Tracts

1369.	1678.	4734.	5326.	4348.	10082.	2087.
2799.	2990.	961.	1711.	4516.	3537.	3605.
6060.	4915.	2829.	2964.	2622.	2566.	2644.
7000	7049.	4313.	3551.	3762.	5698.	5090.
6911.	7840.	8989.	6102.	7070.	5402.	6844.
7640.	5768.	8796.	5338	6480.	3356.	7305.
4623.	6660.	3967.	5816	7438.	6672.	6226.
7063.	10197.	11167.	8479.	12644.	4491.	6145.
8045.	7856.	6453.	6416.	5939.	10921.	5514.
5049.	8014.	8078.	6665.	4840.	2569.	2911.
5170.	5605.	5750.	4029.	4687.	5933.	7602.
1638.	6444.	11120.	13644.	6845.	9308.	8219.
9500.	8743.	7917.	9266.	8107.	6423.	6628.
3097.	6894.	8320.	9011.	4555.	7348.	7677.
2457.	3672.	4563.	5400.	6858.	8207	6696.
6444.	6449.	7524.	9489.	8211.	8335.	8250.
7458.	7369.	8465.	8815	9149.	8000.	9239.
6733.	8792.	8364.	7639.	7009	10442.	7811.
7107.	3846.	11032.	9288.	8375.	4190.	5881.
3018.	4097.	3243.	2997.	2022.	8500.	2383.
4069.	2766.	3808.	6608.	6094.	7346.	5922.
5447.	4092.	4792.	5219.	6217.	7860.	7517.
6704.	6337.	6891.	7583.	8700.	3625	7213.
9128.	7765.	8625.	7379.	6837.	7870.	7219.
7127	10318.	10027.	9233.	6917.	11674.	8247.
7830.	8449.	7597.	6300	9410.	10159.	9968.
9265.	9462.	10471.				

1-2

5 - number Summary

for Median Income of Families and Unrelated Individuals

for Pittsburgh Census Tracts

E	961	UH - M = 1373
H	4734	M - LH = 2110
M	6844	
H	8217	E - M = 6800
E	13644	M - E = 5883

This is an

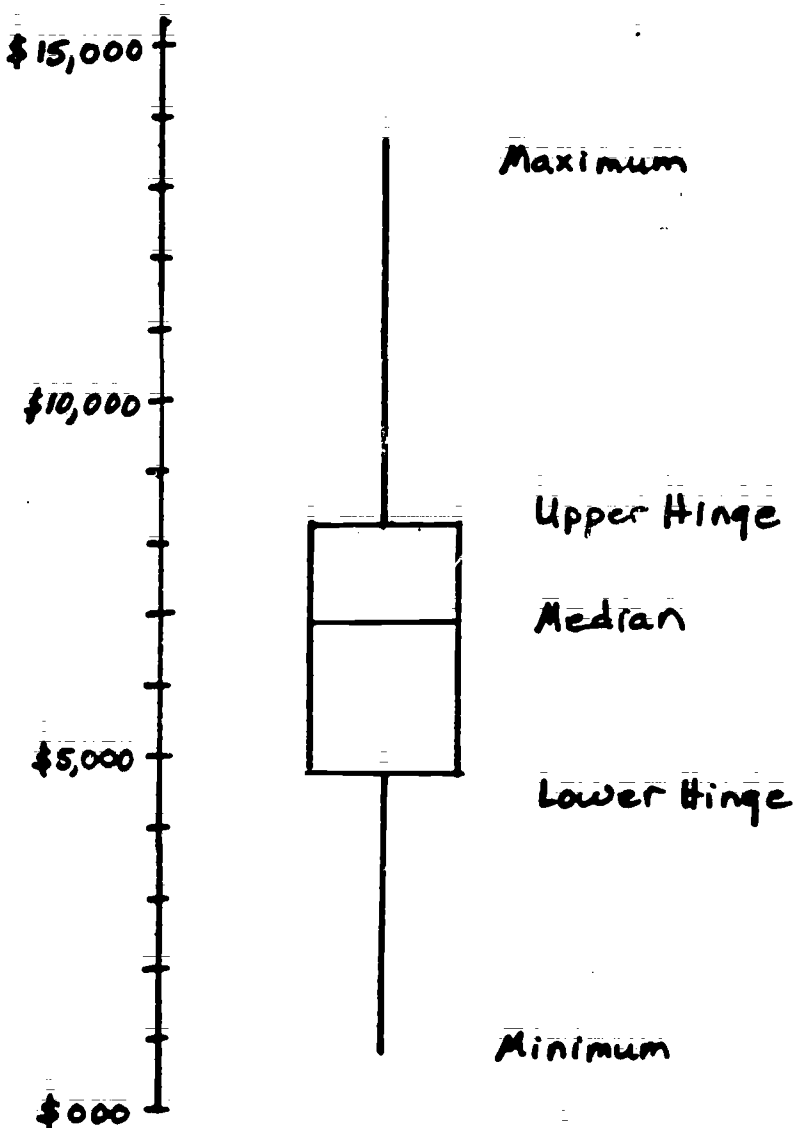
Asymmetric Batch:

Median lies closer to Upper Hinge
and closer to Lower Extreme
because of large outliers.

[11]

<u>Simple Schematic Plot</u>

for 1970 Median Incomes for Pgh. Census Tracts



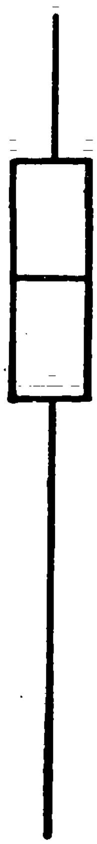
138

[12]

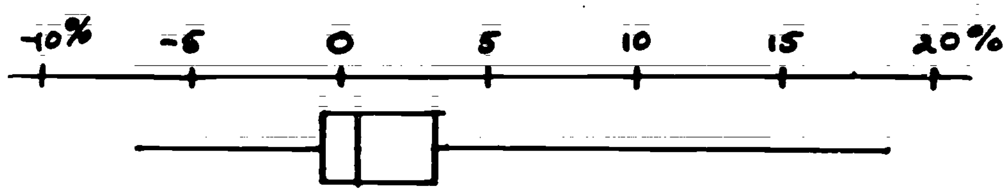
Schematic Plot, 5-number Summary, & Stem-and-Leaf Display of the 1970 populations of Pgh. census tracts

unit = 100 persons (values cut)

Schematic Plot	Number Summary		Stem-and-Leaf Display	
300	E	300	0·	333344444
			0***	556677777789999
1000			1·	0000112222223333344444
	H	1500	1***	5566667888899999999
2000			2·	0000011111122333333444444
	M	2600	2***	55556666667778889999999
3000			3·	001111111222223344 4
	H	3700	3***	55666677778888999999
4000			4·	0000123344
			4***	5567778
5000			5·	1223334
			5***	6788
6000			6·	022
			6***	578
7000			7·	4
	E	7910	7***	9



Schematic Plot of Net Migration in Population of Penna. Counties, 1970-74, in percent

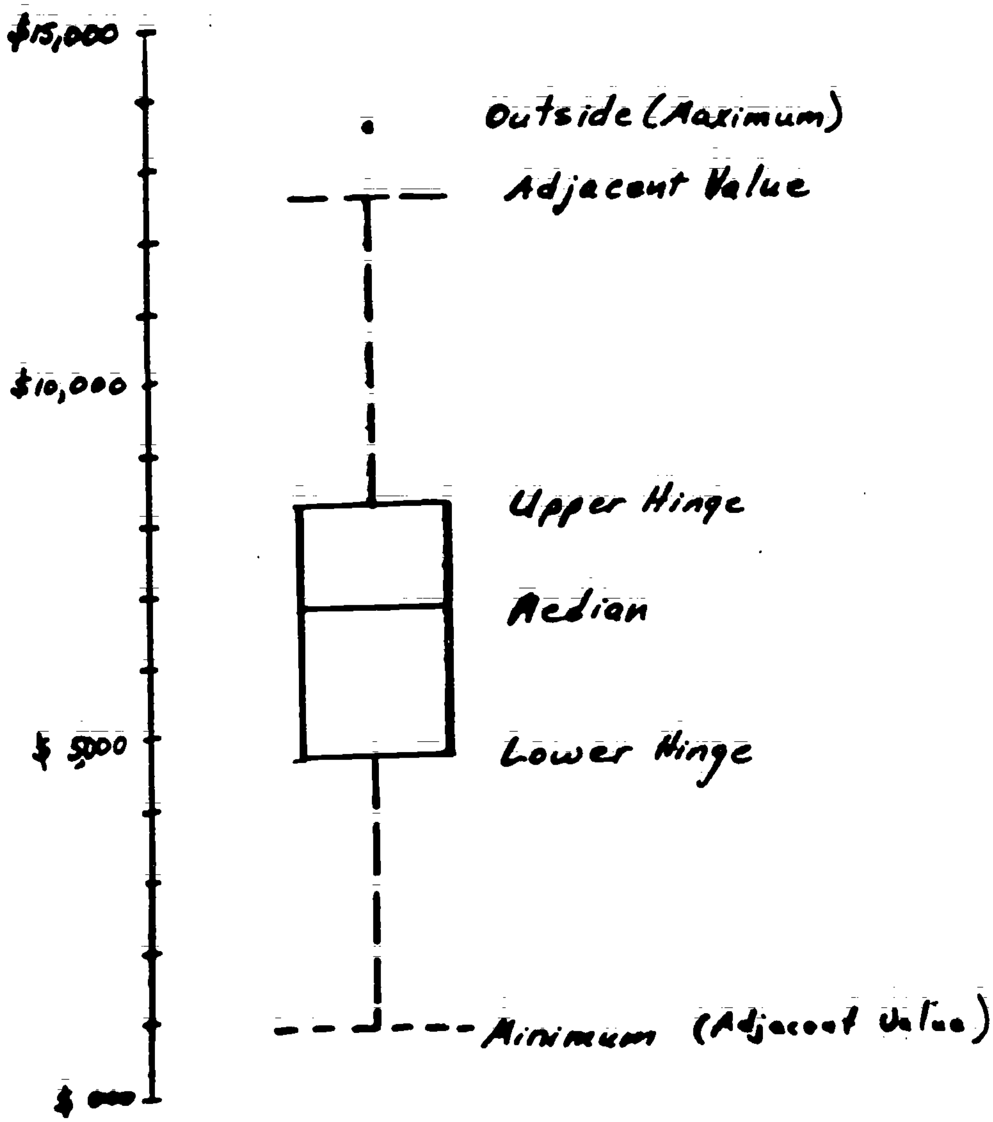


5-number Summary

Depth		Value
1	E	-6.8
17½	H	-0.4
34	M	0.9
17½	H	3.2
1	E	18.5

Schematic Plot

of Pittsburgh Median Incomes



1-2

Numbers for Condensing a Batch into an Expanded Summary

M	median	middle value of batch
E	extreme	smallest and largest value of batch
H	hinge	values halfway between M and E
ΔH	midspread	difference between values of hinges
	step	(3/2)* midspread
f	inner fence	one step beyond hinges
F	outer fence	two steps beyond hinges

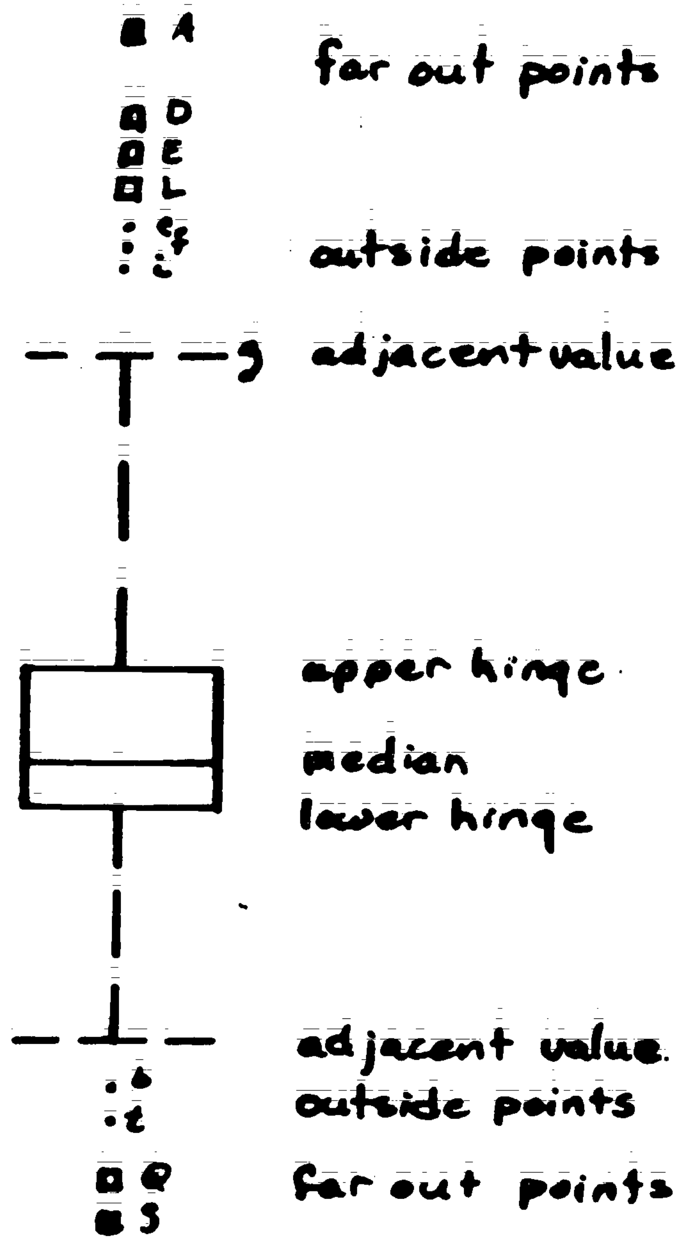
Data values at each end closest to, but still inside, the inner fence are "adjacent."

Data values falling between the inner and outer fence are "outside."

Data values beyond the outer fence are "far out"

[6]

Anatomy of a Schematic Plot



[17]

Tukey's Fenced-Letter Display

#N

M	depth	median		
H	depth	lower hinge	upper hinge	midspread
E		minimum	maximum	

Step

inner fence f	lower f	upper f	adjacent values
	#outside values	#outside values	
outer fence F	lower F	upper F	
	#outside values	#outside values	

Fenced-Letter Display for Pittsburgh Median Income Data

186

M	93	6844		
H	47	4734	8217	3488
E	1	961	13644	

5224.5

f	-490.5	13441.5	adjacent =
	xxx	1	12644
F	-5715	18666	outside =
	xxx	xxx	13644

144 1-2

Lecture 1-3. Transformations for Symmetry

Transformations for Symmetry: The Use of Various Algebraic Transformations to Promote Symmetry in a Single Batch

Lecture Content:

1. Discuss different types of data and the need for transformation
2. Introduce methods of determining a good transformation

Main Topics:

1. Units of measurement and different types of data
2. Methods of determining a good transformation

Tools Introduced:

1. Transformation Summaries

Topic 1. Units of Measurement and Different Types of Data

I. Basic issue: Batch of data has a specific, but alterable, unit of measurement

   1. Unit of Measurement

      a. Data always measured by some recording instrument, and data values in batch are given in specific units

      b. Units may not be ideal for intended analysis

      c. May need to alter a unit of measurement by transforming the data, to obtain a better unit of analysis (2)

   2. Chosen unit of analysis depends on the type of data to be analyzed

      a. Amounts--never negative, may be very large e.g. height, weight, monetary units, distances, certain ratios (3)

      b. Counts--never negative, always integer valued e.g. numbers of persons, things, or events

      c. Percentages or numbers bounded on both extremes--take values between a smallest possible number and a largest possible number e.g. percentage Black (between 0 and 100%) statistical correlations (between -1 and 1)

      d. Differences of amounts or counts ("balances")--positive or negative, unbounded e.g. profit (difference of monetary amounts) net migration (difference of counts of persons

   3. Chosen transformation should make batch more symmetric and, consequently, closer to being "well-behaved" and easily summarized

II. Problem: Need simple rules for choosing a transformation

   1. Simple rules may not always be correct

   2. Best transformation depends on type of data to be analyzed

   3. Unfortunately, even best transformation may fail to increase symmetry

4. Or, by increasing symmetry, transformation may increase variation, or produce more outliers

III. Solution: "Correct" transformation depends on type of data and on spread of data

1. If ratio of maximum to minimum value is quite large (magnitude of 2 or greater), then transformation is essential.

2. If ratio of maximum to minimum is small (less than 20), then transformation will not change the appearance of the batch.

3. Correct transformations are "theoretically" correct, but may fail in practice (3)

   a. Amounts and Counts (particularly large counts) - Logarithms most useful, so are square roots

   b. Percentages and small counts--Special "arcsine" transformation very useful

   c. Differences--Transform the counts or amounts whose difference is under consideration

IV. Examples

1. Counted Data--Pittsburgh populations

   a. Take logarithms, base 10, of observations

   b. Logarithms have not made batch symmetric. Batch is asymmetric, trailing out to the right instead of to left (4)

   c. Try square roots of observations (5)

   d. Schematic plots show relationship between the raw data and the transformations

2. Percentage Data--Percent of individuals under poverty (6) level in Pittsburgh

   a. Take Arcsine (Square Root (X)) for transformation X = proportions, between 0 and 1

b. Spread has decreased, the symmetry improved with special transformation (7)

c. Schematic plot shows increased symmetry, although outliers still present (8)

3. Amounts--Police expenditures in millions of dollars by state, 1973 (9)

a. Try square root and log transformation

b. Logs are very effective

4. Difference of Counts--Net migrations for Pennsylvania counties, 1970-1974 (10)

a. Stem-and-leaf shows symmetry but large outliers (11)

b. Net migration = Change in Population - Number of Births + Number of Deaths. Transform these three batches separately.

c. Positive and Negative values in the Change in Population batch make transformation impossible (12) (13)

148

Topic 2. Methods of Determining a Good Transformation

I. Basic Issue: Need a reliable method of finding a good transformation

1. Transformation must promote symmetry, and bring the outliers of the batch toward the median

2. Restrict ourselves to transformations from X to X^R for any value of R

3. This form of transformation includes logs (R=0)

II. Problem: How do we find the correct exponent R?

III. Solution: Examine 5-number summary of raw and transformed batch

1. Correct transformation will have median halfway between hinges and extremes

2. Simple Ladder of Powers indicates that: (14)

a. Increasing R expands the larger values of X

b. Decreasing R compresses the larger values of X

3. Ladder of Powers useful in conceptualizing how various transformations act on batches

IV. Method: Transformation Summaries

1. Example shows transformation summaries for the number of births in Pennsylvania counties, 1970-1974 (15)

a. Deaths take a similar transformation

b. "Natural" increase in population = Births - Deaths will also be symmetric with logs of births and deaths

2. Features

a. Useful if correct transformation for type of data in batch does not promote symmetry

b. Also useful if batch does not fall neatly into one of the four types

c. Easily computable from 5-number summary of raw batch

d. Helps "zero in" on the appropriate exponent, R, for transformation

3. Analytic Qualities

   a. Midhinge and Midextreme indicate whether R should be increased or decreased

   b. Correct R has median = midhinge = midextreme

   c. Upwards trend (M < midhinge < midextreme) indicates R should be decreased

   d. Downwards trend (M > midhinge > midextreme) indicates R should be increased

   e. Useful exponents:

      i. R = 1, Raw data

      ii. R = 2, Squared data

      iii. R = 1/2, Square roots

      iv. R = 0, Logarithms

      v. R = -1, Negative Reciprocals (change of sign retains order). Rarely will additional transformations be needed

4. Procedure

   a. Compute 5-number summary for batch

   b. Compute Midhinge (MidH) = 1/2 (UH + LH)
      Midextreme (MidE) = 1/2 (Max + Min)

   c. Compare MidH, MidE, and Median (M)

   d. If M < MidH < MidE, decrease R
      If M > MidH > MidE, increase R

   e. 5-number summary for transformation of batch easily found by raising 5-number summary of raw batch to the correct exponent

   f. Continue search until M = MidH = MidE

5. Transformation summaries constructed on computer:

   a. Use LET and REEX to transform batch

   b. Use SUMMARY and ESTATS to examine effect of transformation

   c. Discovering the correct symmetrizing transformation is iterative process

Lecture 1-3
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 1-3 Outline
Topic 1 Section I		
1.c	2	Need for new unit of analysis
2.a	3	Types of Data & Transformations
Section IV		
1.b	4	Stem-and-Leaf of Logs of Pittsburgh Populations
1.c	5	Stem-and-Leaf of Square Roots of Pittsburgh Populations
2.	6	Stem-and-leaf of Pittsburgh poverty
2.b	7	Stem-and-leaf of transformation
2.c	8	Schematic plot of transformation
3.	9	Police Expenditures, by State, 1973
4.	10	Net migrations of Pennsylvania counties
4.a	11	Stem-and-leaf of net migrations
4.c	12	Stem-and-leaf of Births and deaths
4.c	13	Stem-and-leaf of change in population

Topic 2
Section III

2.	14	Simple ladder of powers

Section IV

1.	15	Transformation summaries

[1]

Lecture 1-3

Transforming for Symmetry:

The use of various Algebraic Transformations to promote Symmetry in a single batch.

Lecture Content:

Discuss different types of data and the need for transformation.

Discuss methods of finding the "correct" transformations.

Main Topics:

1) Units of measurements and different types of data.

2) Methods of determining a good transformation.

154

1-3

[2]

Example of Need for New Unit of Analysis.

Batch: Number of Blacks in each of the Census Tracts in Pgh. in 1970

Units of Measurement: (old unit of analysis)
eg. Black persons

Desire to have a different unit of analysis:
eg. percentage Blacks in each census tract
therefore,
Divide each data value by the total population of the tract and multiply by 100 %.

New unit of analysis: eg. Percent Black

Example.

Tract 1 has 41 Black persons
Population of Tract 1 is 972 persons
therefore,
Tract 1 is (41/972) x 100 % =
4.21 % Black

[3]

Types of Data

a) Amounts:

never negative; may be arbitrarily large.

eg., height, weight, $, £.

→ Transformation: log or square root

b) Counts:

never negative; always integer-valued.

eg., population, number of families, number of health care facilities.

→ transformation: log or square root if large
arcsine of square root if small

c) Percentages or "bounded" numbers:

values are between a smallest and largest possible number.

eg., % black.

→ transformation: arcsine of square root

d) Differences of amounts or counts:

positive or negative; unbounded.

eg., profit and loss, "net" data or balances – net migration.

→ transformation: operate individually on the counts or amounts whose difference is under consideration

1-3

[4]

Stem-and-Leaf of Base 10 logs of Pittsburgh Populations

```
PGH_LOGPOP

      LO |        2.52       2.53       2.54

   (UNIT = 1Ø**-2)
     25  | 9
     26  | 4558
     27  | 568
     28  | Ø566689
     29  | 28899
     3Ø  | Ø122678889
     31  | 1122334445566899
     32  | 1111577778889999
     33  | ØØ1112223356667777888999
     34  | Ø11111223445556666677789999999
     35  | ØØ1111223344555566677788889999
     36  | ØØØ112234445567778
     37  | 112222335566799
     38  | 13379
```

[5]

Stem-and-Leaf display of square roots of Pittsburgh Populations

(Unit = 10++0)

1. | 8889
2 | 1112344
2. | 56677789
3 | 011112224444
3. | 555566667777788999
4 | 00002333333444444
4. | 5555666667788888899999
5 | 0000111112223333444444
5. | 555555666677777888899
6 | 0000011111122222333334
6. | 556667778889
7 | 12222233
7. | 5566789
8 | 022
8. | 68

158

1-3

[6]

Stem-and-Leaf display of Pgh. Percentage in Poverty. Percentage Data

```
PGH_PERPOV

     (UNIT = 10**0)
        0  | 11
        T  | 22223333
        F  | 4444444355555
        S  | 66656666666667777777777777777
        0. | 88888888899999999
        1  | 00000000111111111
        T  | 2222222222223333333
        F  | 44444555555
        S  | 666667777
        1. | 8899999999
        2  | 000011
        T  | 233
        F  | 4445555555
        S  | 77
        2. | 99
        3  | 111
        T  | 222
        F  | 4444
        S  | 66
        3. | 8889
        4  | 0011
        T  | 2

       HI  |   46.     46.     49.3    49.7    53     59.6    74.2
```

1-3

[7]

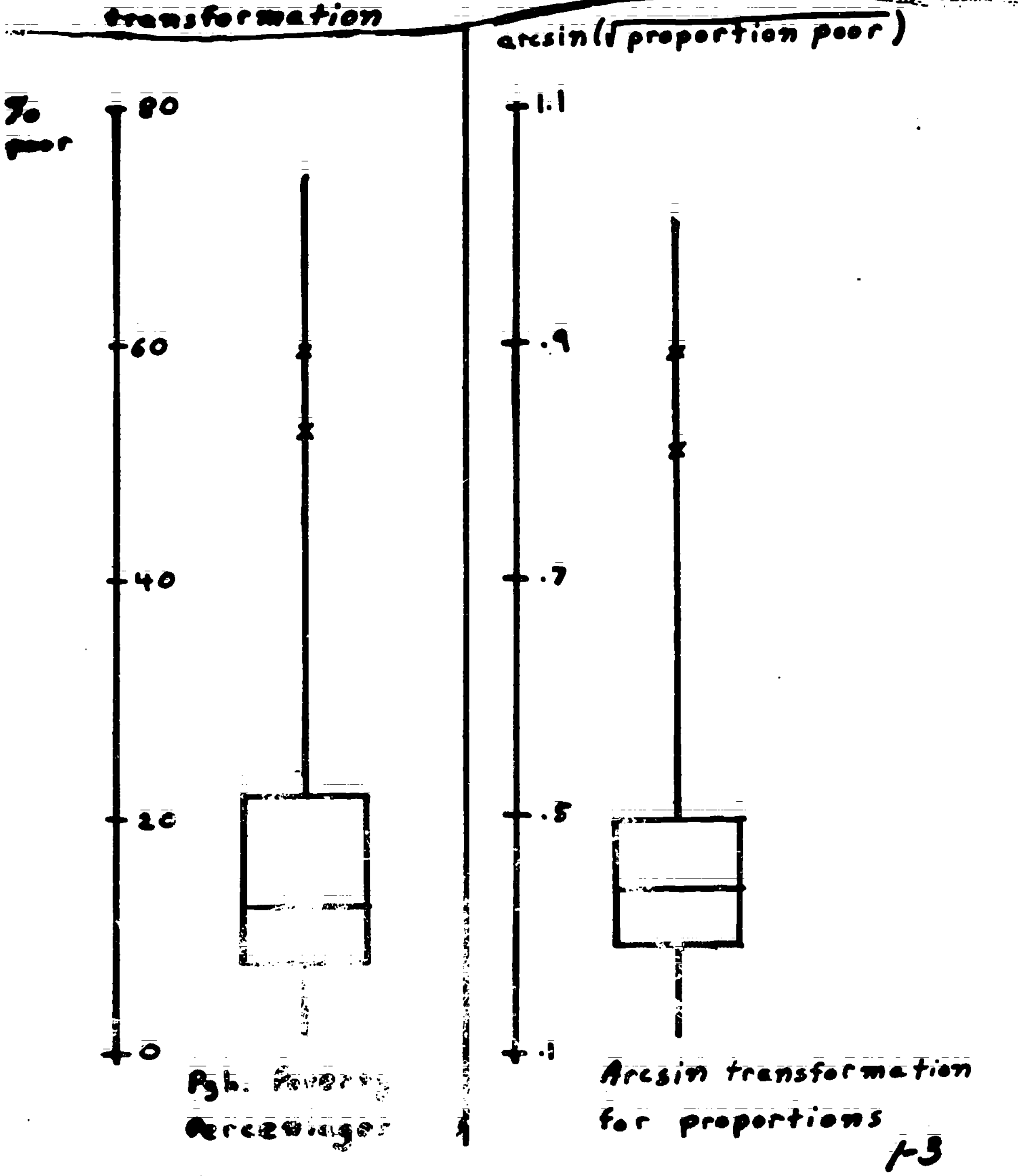
Stem-and-Leaf of transformed Pgh. Poverty Proportions.

```
(UNIT = 10**-2)
   1  | 13
   1. | 56677799
   2  | 0111122334444444
   2. | 555555566666666777777778888999999
   3  | 0000011112222223334444444
   3. | 5555556666666667777888899
   4  | 00000111222233333
   4. | 555566666666888
   5  | 0012222233334
   5. | 577999
   6  | 00022234
   6. | 6667999
   7  | 0144
   7. | 78

  HI  |     0.815   0.90     1.236
```

160

1-3

Schematic plot of Pgh. Poverty Percentages and Special [8
transformation



[9]

Police Expenditures, in Millions of Dollars, 1973
(States listed by Federal Administration Region)

Region I	
Conn.	$98
Maine	19
Mass.	214
N. H.	18
R. I.	27
Vt.	10
Region II	
N. J.	285
N. Y.	1022
Region III	
Del.	18
D. C.	113
Md.	151
Pa.	345
Va.	111
W. Va.	24
Region IV	
Ala.	61
Fla.	215
Ga.	96
Ky.	59
Miss.	40
N. C.	101
S. C.	48
Tenn.	72
Region V	
Ill.	$434
Ind.	102
Mich.	320
Minn.	89
Ohio	256
Wisc.	127
Region VI	
Ark.	31
La.	101
N. Mex.	28
Okla.	50
Tex.	241
Region VII	
Iowa	51
Kan.	43
Mo.	129
Neb.	31
Region VIII	
Col.	60
Mont.	14
N. Dak.	10
S. Dak.	11
Utah	25
Wyo.	8
Region IX	
Ariz.	$78
Calif.	[illegible]
Hawaii	
Nev.	
Region X	
Alaska	[illegible]
Idaho	15
Ore.	[illegible]
Wash.	91

1-3

[10]

Net Migrations of Pennsylvania Counties, 1970-1974, in numbers of persons, Counties in Alphabetical Order

3400.	-82100.	-500.	-2800.
100.	4900.	-1100.	-100.
15300.	4200.	-400.	-500.
1500.	3200.	2700.	1300.
700.	0.	2600.	1800.
5100.	-1000.	-22800.	-1000.
1800.	100.	NA	800.
400.	600.	400.	2600.
2200.	900.	-300.	8000.
-1300.	2000.	4600.	6500.
500.	-1400.	-500.	-1500.
6400.	-1100.	NA	5700.
1400.	1800.	-131700.	2200.
100.	1300.	1400.	0.
-100.	600.	1200.	1400.
500.	-1300.	-100.	2500.
-4500.	2500.	3900.	

1-3

[11]

Stem-and-Leaf Display
of Pennsylvania Counties, Net Migrations

```
LO | -131700.  -82100.  -22800.
(UNIT=10**2)
 -4 | 5
 -3 |
 -2 | 8
 -1 | 54331100
 -0 | 55543111
  0 | 001111445566789
  1 | 2334445888
  2 | 02255667
  3 | 249
  4 | 269
  5 | 17
  6 | 45

HI | 8000.  15300.
```

Stem-and Leaf Displays of Change, Number of Births and Deaths in Pennsylvania Counties, 1970-1974

[12]

Births (unit = 10**3)

```
0 | 0000001111
 T | 22222222222223333333
 F | 4445555
 S | 666677
0. | 88899
1 | 011
 T | 222
 F | 44
 S | 67
1. | 889
2 | 1
 T | 2

HI | 28600.00   33100.00
     33200.00   83100.00
     124100.00
```

Deaths (unit = 10**3)

```
0 | 00000000001111111111
 T | 2222222222333333
 F | 44555555
 S | 7
0. | 889999
1 | 00011
 T | 2233
 F | 5
 S |
1. | 9

HI | 24100.00   24400.00
     74000.00   100600.00
```

Stem-and-Leaf of Change in Population [13]

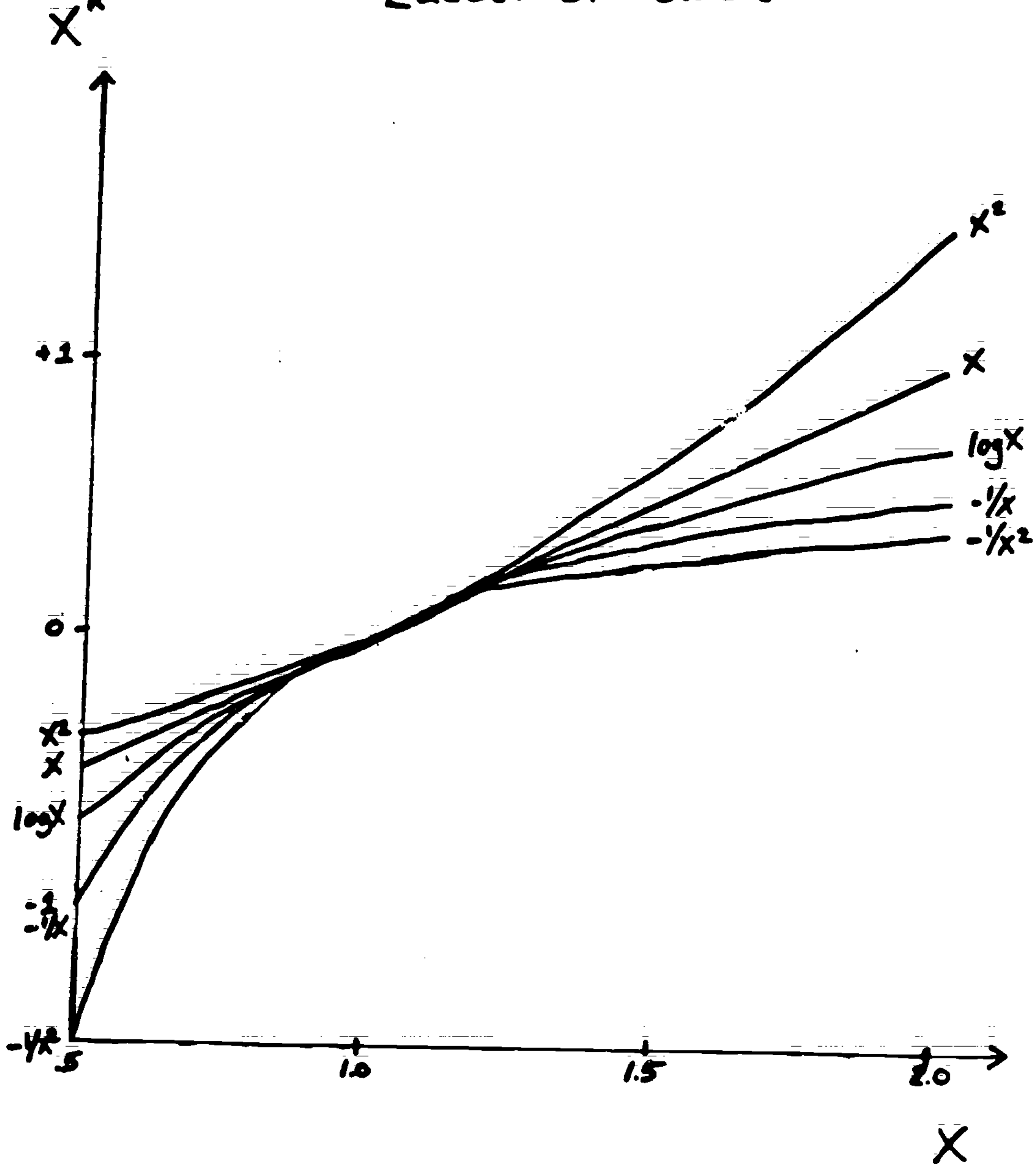
(Unit = 10 ** 2)

LO	-108200.00	-73000.00	-14100.00	
-1	1			
-0	653222			
0	12666778			
1	0223335667777			
2	1122222778			
3	11247			
4	0449			
5	4			
6	9			
7	14			
8	0001			
9	2			
HI	10000.00	10500.00	11400.00	
	17800.00	31500.00		

167

[14]

Ladder of Powers



1-3

[15]

Transformation Summaries. Number of Births in Pennsylvania Counties, 1970-74

	Raw R = 1	Sq. Root R = 1/2	Log R = 0	Neg. Reciprocal R = -1 (Unit = .001)
1 E	300	17.32	2.48	-3.33
17½ H	26[illegible]0	48.48	3.42	-0.38
34 M	5300	72.80	3.72	-0.19
17½ H	11850	108.86	4.07	-0.08
34 E	124700	352.28	5.09	-0.07
M	5300	72.80	3.72	-0.19
Mid H	7250	78.67	3.75	-0.23
Mid E	62200	1[illegible].80	3.78	-1.67
Trend	Up	Up	Flat	Down

Base 10 Logarithms Appropriate Transformation

p3

Lecture 1-4. Analysis of a Well-Behaved Batch

Analysis of a Well-Behaved Batch: Presentation of a Special Type of Batch and Examination of its Features

Lecture Content:

1. Define a well-behaved batch and discuss its characteristics
2. Introduce measures to summarize this special kind of batch

Main Topics:

1. Definition of a well-behaved batch
2. Location and scale measures for a well-behaved batch

Tools Introduced:

1. Mean
2. Variance and Standard Deviation

Topic 1. Definition of a Well-Behaved Batch

I. Basic Issue: Defining a well-behaved batch

1. Well-behaved batches are theoretical entities and are rarely observed empirically

2. Many data analysts incorrectly believe that well-behaved batches are common

3. We discuss them because of their role in regression analysis

4. The well-behaved batch presented here was artificially constructed to facilitate the introduction of the definition

II. Definition (2)

1. Well-behaved batch is: (3)

a. Symmetric: MidH = MidE = M (4)

b. Devoid of outliers

2. For a well-behaved "standard" batch with M = 0, and Midspread = 1.36: (5)

a. 50% of batch > 0; 50% < 0

b. 50% of batch is between -0.68 and 0.68

c. 80% of batch is between -1.29 and 1.29

d. 80% of batch is between -1.65 and 1.65

e. 95% of batch is between -1.96 and 1.96

f. Extremes are approximately -2.60 and 2.60, but may be larger

3. Well-behaved batch has shape that resembles (in theory) a Gaussian (or "normal") function (6)

171

Topic 2. Location and Scale Measures for Well-Behaved Batch

I. Basic Issue: Need for special summarization tools for a well-behaved batch

1. All well-behaved batches have similar appearance

2. Two well-behaved batches may differ only in:

   a. Location--where batches are positioned along the Real number line

   b. Scale--how spread out the batches are, amount of variation in the data values

3. Need to quantify these concepts to facilitate comparison of well-behaved batches

II. Problem: Which location and scale measures are appropriate?

1. The median of a batch is a measure of location, as is the mode of a batch (data value with greatest frequency of occurrence) and the arithmetic average, or mean, of the batch (7)

2. The midspread and range are measures of spread. The variance, or average of the squared differences from the mean, also measures spread (8)

3. The standard deviation, or square root of the variance of a batch, in the same unit as the data values, is also useful in measuring the scale of the batch

III. Solution: Mean and standard deviation are the correct measures of location and scale, respectively

1. In a well-behaved batch, $\overline{X}$, mean, and M, median, are equal to each other and to the mode (7)

2. In a well-behaved batch, the standard deviation, s, is approximately equal to 3/4 x Midspread (8)

IV. Methods: Mean and Standard Deviation

1. Example shows mean and standard deviation of our hypothetical well-behaved batch (9)

A well-behaved batch with $\overline{X} = 0$, and $s = 1$, is called a standard or standardized well-behaved batch

2. Features

   a. Mean and standard deviation are sufficient to describe a well-behaved batch (explain statistical sufficiency)

   b. Any well-behaved batch m~v be standardized by subtracting the mean from each data value and dividing the remainder by s, $(X - \overline{X})/s$

   c. Mean and standard deviation are not sufficient to describe batches that are not well-behaved

3. Analytic Qualities

   a. Median and midspread are more resistant, or less affected by outliers, than mean and standard deviation

   b. Nonetheless, $\overline{X}$ and s are classical measures of location and spread (for all batches)

4. Procedures

   a. $\overline{X} = (1/N) \sum_i X_i$

   b. $s = \sqrt{(1/N) \sum_i (X_i - \overline{X})^2}$

5. Another example: I.Q. scores for 100 16 year old females (10)

   a. Batch is well-behaved: (11)

      i. Symmetric

      ii. No outliers

      iii. $\overline{X} = M = 101$

      iv. $s = 3/4 \times$ Midspread $= 12$

   b. Standardize batch. Note resemblance to hypothetical standard well-behaved batch (12)
   (13)

   (This lecture should be followed by a review of the entire unit before the quiz is given.)

Lecture 1-4
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 1-4 Outline
Topic 1 Section II		
1.	2	Hypothetical Well-Behaved Batch, Sorted
1.a	3	Stem-and-Leaf and 5 Number Summary of Hypothetical Batch
1.b	4	Schematic plot of Hypothetical Batch
2.	5	Histogram of Hypothetical Batch
3.	6 (overlay 5)	Gaussian function
Topic 2 Section II		
1.	7	Measures of Location of a Batch
2.	8	Measures of Scale of a Batch
Section IV		
1.	9	Hypothetical Batch, mean and standard deviation
5.	10	I.Q. scores for 16 year old females
5.a	11	Stem-and-Leaf of I.Q. scores
5.b	12	Standardized I.Q. scores
5.b	13	Stem-and-Leaf of Standardized Scores

Lecture 1-4

Analysis of a Well-Behaved Batch

Presentation of a very special type of batch and an examination of its features

Lecture Content

Define a well-behaved batch

Discuss its characteristics

Introduce measures to summarise this special batch

Main Topics

1. Definition of a well-behaved batch

2. Location and scale measures for well-behaved batches

[2]

Hypothetical Well-Behaved Batch of Data, Sorted.

-2.60	-0.66	0.00	0.70
-2.15	-0.61	0.01	0.74
-1.99	-0.60	0.05	0.79
-1.85	-0.59	0.08	0.81
-1.73	-0.58	0.10	0.84
-1.65	-0.52	0.13	0.86
-1.60	-0.50	0.16	0.90
-1.53	-0.49	0.18	0.95
-1.45	-0.44	0.22	0.99
-1.36	-0.40	0.26	1.04
-1.30	-0.38	0.26	1.09
-1.26	-0.34	0.27	1.14
-1.23	-0.31	0.29	1.19
-1.19	-0.28	0.32	1.25
-1.15	-0.26	0.35	1.29
-1.06	-0.25	0.39	1.33
-0.98	-0.23	0.43	1.41
-0.93	-0.19	0.48	1.49
-0.87	-0.17	0.50	1.58
-0.85	-0.14	0.52	1.63
-0.83	-0.11	0.55	1.70
-0.81	-0.09	0.59	1.79
-0.76	-0.06	0.61	1.98
-0.71	-0.02	0.65	2.18
-0.69	-0.01	0.68	2.50

1-4

[3]

Stem-and-Leaf of a Hypothetical Well-Behaved Batch. values cut Unit = .10

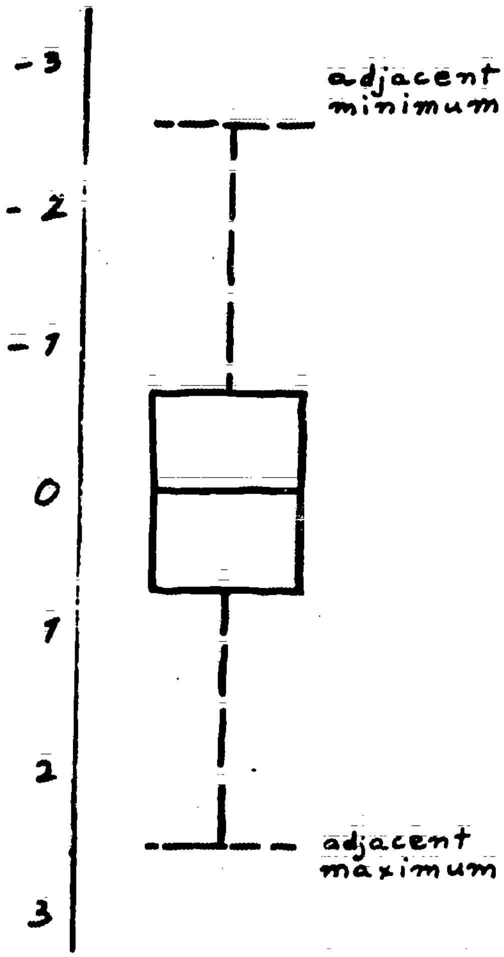
1	-2**	6
2	-2·	1
8	-1**	987665
16	-1·	43322110
32	-0**	9988887766665555
50	-0·	444333222211110000
50	0·	000011112222233344
32	0**	5555666777888999
16	1·	00112234
7	1**	56779
2	2·	1
1	2**	5

5-Number Summary of Batch

1	E	-2.60
25	H	-0.69
50½	M	-0.005
25	H	0.70
1	E	2.50
	M	-0.005
Mid	H	-0.005
Mid	E	-0.05

(Very Symmetric) 1-4

Schematic Plot of Hypothetical Well-Behaved Batch. [4]



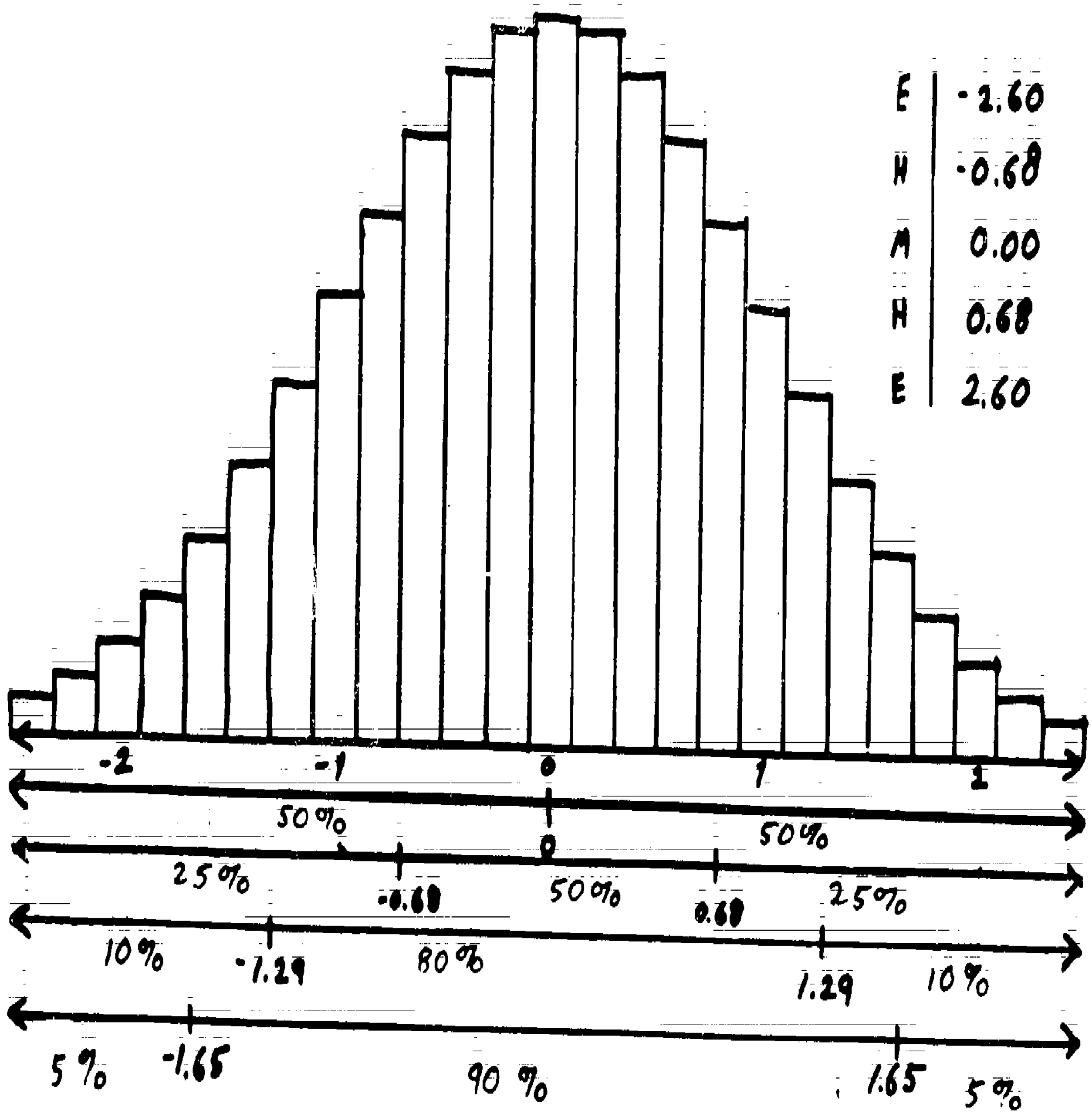
Letter-Value Display

50½	M	-0.005		
25	H	-0.69	0.70	1.39
1	E	-2.60	2.50	
		2.09		
	f	-2.78	2.79	adj = -2.60, 2.50
		**	**	
	F	-4.87	4.88	
		**	**	

No outside or far out values!

1-4

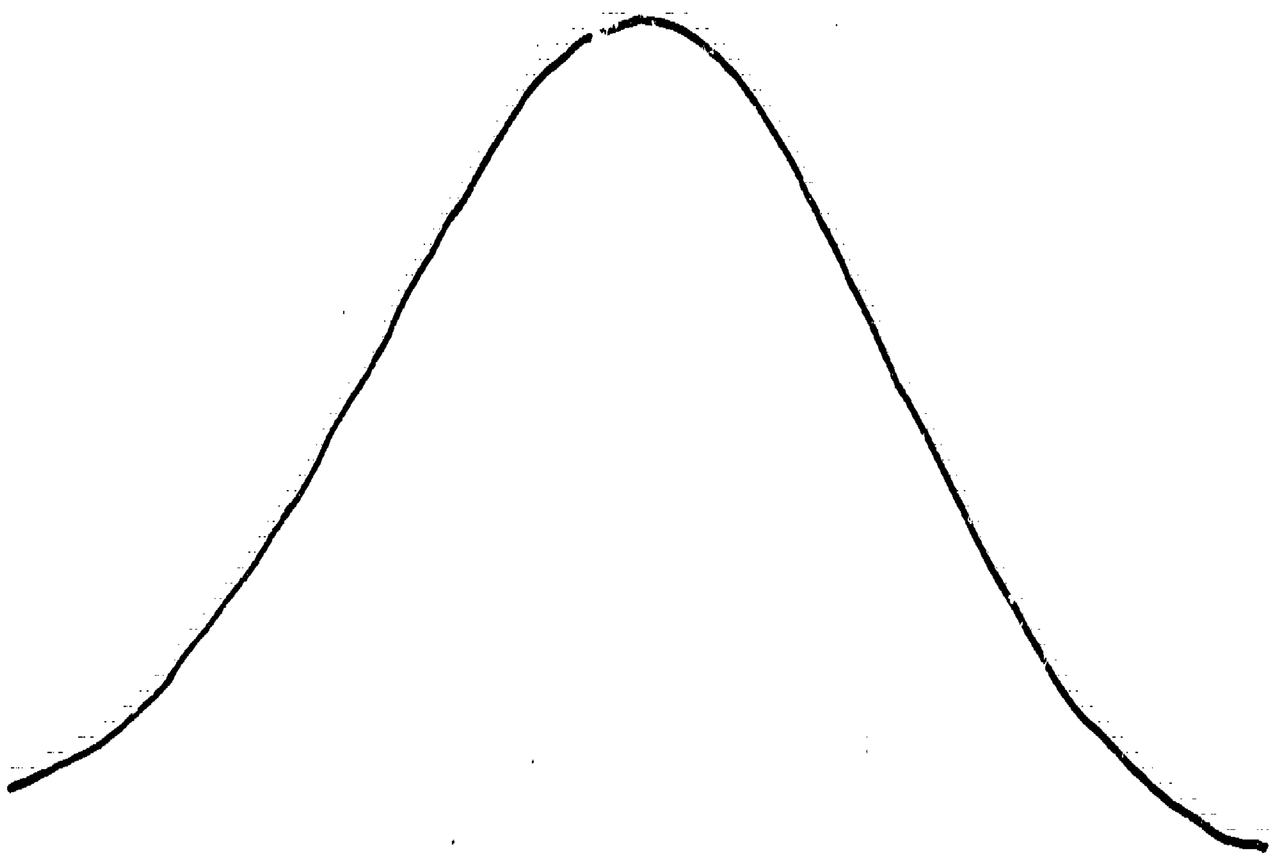
Hypothetical Histogram for a Well-Behaved Batch, with 5 number summary [5].



1-4

Gausian Function as an approximation to the hypothetical histogram for a Well-Behaved Batch.

[6]



XVI.I.131

Module I

Measures of the Location of a Batch

(The Three M's)

1) Median (M)

middle value of a batch.

2) Mode (Mo)

most frequent data value of a batch.

3) Mean ($\overline{X}$)

arithmetical average of a batch.

Measures of the Scale of a Batch

1) Midspread

difference in hinges of a batch.

2) Range

difference in extremes of a batch.

3) Variance (s^2)

average squared difference of the values from the mean of a batch (in units^2).

4) Standard Deviation (s)

square root of the variance, s^2 (in same unit as data).

[9]

Hypothetical Well-Behaved Batch

unit = .10

```
-2* * | 6
-2  • | 1
-1* * | 9 8 7 6 6 5
-1  • | 4 3 3 2 2 1 1 0
-0* * | 9 9 8 8 8 8 7 7 6 6 6 6 5 5 5 5
-0  • | 4 4 4 3 3 3 2 2 2 2 1 1 1 1 0 0 0 0
 0  • | 0 0 0 0 1 1 1 1 2 2 2 2 2 3 3 3 4 4
 0* * | 5 5 5 5 6 6 6 7 7 7 8 8 8 9 9 9
 1  • | 0 0 1 1 2 2 3 4 4
 1* * | 5 6 7 7 9
 2  • | 1
  * * | 5
```

$M = -.005 \approx 0$

Midspread = .70 + .69 = 1.39
3/4 × Midspread = 1.04

$\bar{X} = -.005 \approx 0 = \frac{1}{N} \sum_{i=1}^{N} x_i$

$S = 1.01 = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$

Our well-behav d batch has a location of $\bar{x} \approx 0$ and a scale of $s = 1.0$.

Batch of I.Q. Scores [10] for 16 year old Females.

100	98	93	106
86	95	110	97
94	120	128	87
100	110	108	97
113	115	111	91
101	111	81	95
112	101	119	93
111	115	113	104
107	105	101	103
96	90	108	104
92	92	96	101
105	114	108	91
107	77	95	95
103	88	111	88
82	89	[illegible]	92
113	93	103	92
111	126	86	101
108	92	108	102
83	98	104	76
96	100	85	107
119	92	101	90
103	99	93	113
114	94	126	108
91	98	98	111
102	116	80	105

Stem-and-Leaf Display of I.Q. Scores

(unit = 1)

4	7	*	6 6 7 9
8	8	·	0 1 2 3
16	8	*	6 6 6 7 8 8 9
33	9	·	0 0 1 1 1 2 2 2 2 2 2 3 3 3 3 4 4
45	9	*	5 5 5 5 6 6 7 7 8 8 8 9
(17)	10	·	0 0 0 0 1 1 1 1 1 2 2 2 3 3 3 4 4
38	10	*	5 5 5 6 7 7 7 8 8 8 8 8 8 9
24	11	·	0 1 1 1 1 1 1 1 2 3 3 3 3 4 4
9	11	*	5 5 6 9 9
4	12	·	0
3	12	*	6 6 8

5-number Summary

1	E	76
26	H	92
50 ½	M	101
26	H	108
1	E	126

midspread = 108 − 92 = 16
3/4 × midspread = 12

$\bar{X} = 101$
$S = 11.5$

M	101
mid H	100
mid E	102

Batch is symmetric, has no outliers,
$\bar{X} \approx M$, $S \approx 3/4 \times$ midspread.

Hence, it is well-behaved.

Standardized Values of Batch of I. Q. Scores

$$\text{standardized value} = \frac{x - 101}{17.5} \qquad [12]$$

-0.02	-0.23	-0.65	0.46
-1.23	-0.49	0.78	-0.82
-0.57	1.62	2.26	-1.16
-0.01	0.76	0.66	-0.26
1.01	1.22	0.90	-0.82
0.02	0.83	-1.63	-0.44
0.93	0.01	1.53	-0.63
0.88	1.24	1.02	0.80
0.52	0.87	0.05	0.21
-0.89	-0.88	0.64	0.74
-0.67	-0.69	-0.88	0.85
0.37	1.45	0.65	-0.77
0.50	-1.45	-0.46	-0.46
0.19	-1.00	0.91	-1.05
-1.56	-0.93	-1.76	-0.70
1.01	-0.66	0.13	-0.72
0.86	2.16	-1.23	0.02
0.66	-0.73	0.60	0.12
-1.48	-0.18	0.27	-2.0[illegible]
-2.06	-0.01	-1.80	0.55
1.57	-0.73	0.07	-0.90
0.18	-0.12	-0.64	1.05
1.05	-0.84	2.10	0.85
-0.75	-1.02	-0.17	[illegible]
0.16	1.27	-1.69	[illegible]

[13]

Stem-and-Leaf Display of Standardized Values of Batch of I.Q. Scores

(unit = .1)

```
-2**  | 0 0
-1    | 9 7 6 6 5
-1    | 4 3 2 2 1 0 0 0
-0    | 9 9 8 8 7 7 7 7 7 7 6 6 6 6 6 5 5
-0**  | 4 4 4 3 3 3 2 2 1 1 1 0 0 0
 0    | 0 0 0 0 0 1 1 1 1 1 2 2 3 3 3 3 4
 0    | 5 5 5 6 6 6 6 6 6 7 7 7 8 8 8 8 8 9 9 9
 1**  | 0 0 0 0 0 1 2 2 2
 1    | 5 5 6
 2**  | 1 1 2
```

5-number Summary

E	-2.06
H	-0.69
M	0.02
H	0.74
E	2.26

Midspread = 1.43

3/4 × Midspread = 1.07

Homework Problems
Unit 1

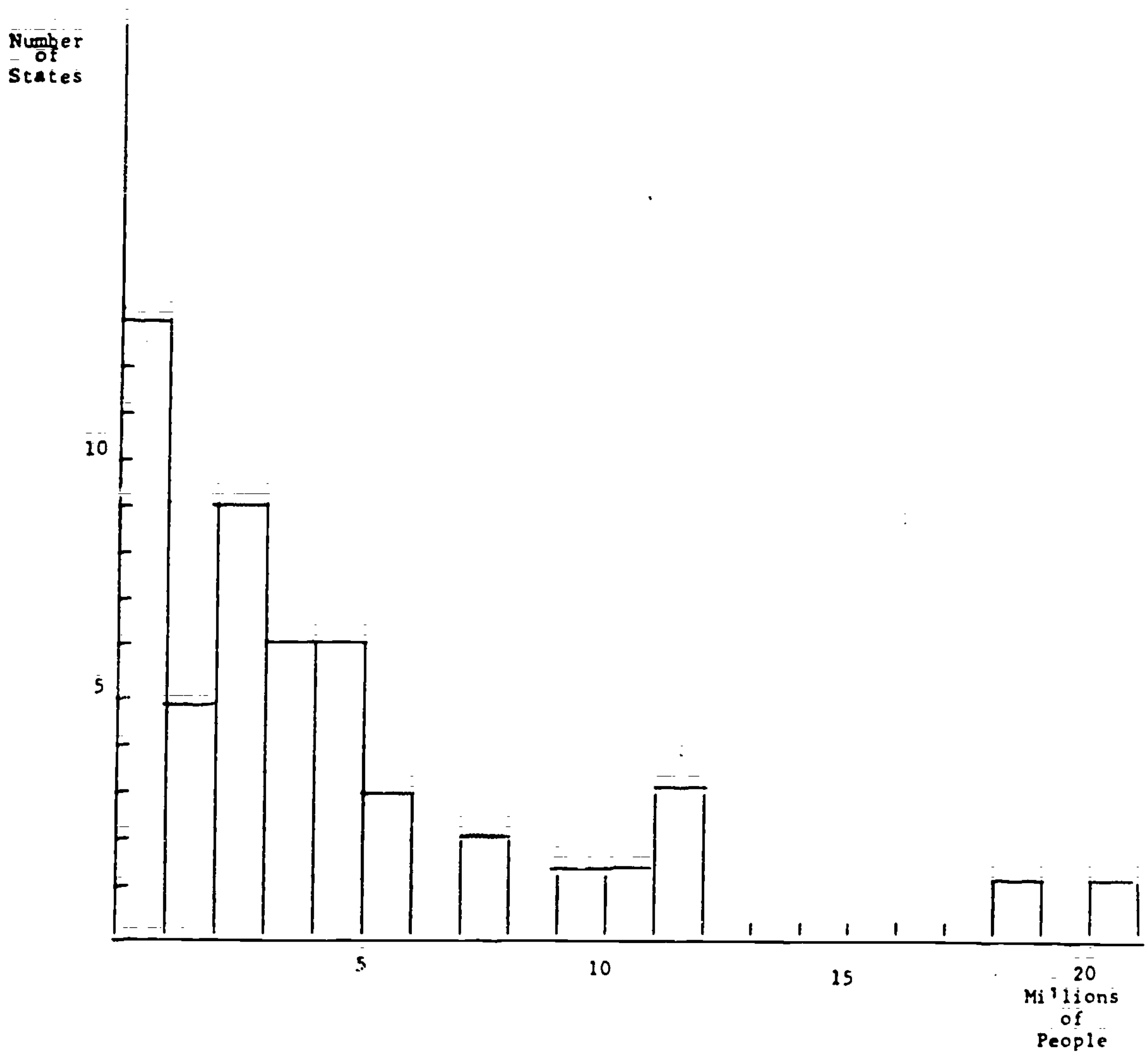
1. A municipality is trying to decide between building its own steam-electric generating plant or purchasing power from a private supplier. Data exist on the installed generating capacity of 33 plants in municipalities with similar socioeconomic and demographic characteristics. Installed generating capacity is a measure of the size of a plant. A first step in the decision process involves examining the range of plant sizes. Sort the data on installed generating capacity; then make a histogram.

   Do the data cluster or are they uniformly spread out? Are the data symmetrical? Are there any outliers? If there are clusters or outliers, where do they occur? What can the municipality infer from the sort and histogram?

Installed Generating Capacity in Megawatts

Bull Run	950	Barry	1770.8
Colbert "A"	846.6	Canal	542.5
Colbert "B"	550	Etiwanda	1069.1
Gallatin	1255.2	Astoria	1550.6
Johnsonville "A"	1485.2	Ravenswood	1827.7
Johnsonville "B"	591.2	Conemaugh	1872
Kingston	1700	Kyger Creek	1086.3
Paradise "A"	1408	Keystone	1872
Paradise "B"	1150.2	Elrama	510.3
John Sevier	823.3	Mt. Storm	1140.5
Shawnee	1750	Joppa	1100.5
Widows Creek "A"	853	Four Corners	1636.2
Widows Creek "B"	1125	Fort Martin	1152
Big Sandy	1050.8	Wabash River	908
Cane Run	1016.7	Parish	1255.4
Clifty Creek	1303.6	Sam Bertron	826.3
		Gannon	1270.4

2. Below is a histogram of the 1973 population of the U.S. for the fifty states and the District of Columbia and the original data from which the histogram was composed. What is the interval of population size into which the largest number of states fall? What is the number of states in that interval? Which states are the outliers of this batch? How would a logarithmic transformation of this batch affect the display? Data are on the next page.



State	1973 Population (in thousands)	State	1973 Population (in thousands)
Maine	1,028	N. Carolina	5,273
New Hampshire	791	S. Carolina	2,726
Vermont	464	Georgia	4,786
Massachusetts	5,818	Florida	7,678
Rhode Island	973	Kentucky	3,342
Connecticut	3,076	Tennessee	4,126
New York	18,265	Alabama	3,539
New Jersey	7,361	Mississippi	2,281
Pennsylvania	11,902	Arkansas	2,037
Ohio	10,731	Louisiana	3,764
Indiana	5,316	Oklahoma	2,663
Illinois	11,236	Texas	11,794
Michigan	9,044	Montana	721
Wisconsin	4,569	Idaho	770
Minnesota	3,897	Wyoming	353
Iowa	2,904	Colorado	2,437
Missouri	4,757	New Mexico	1,106
N. Dakota	640	Arizona	2,058
S. Dakota	685	Utah	1,157
Nebraska	1,542	Nevada	548
Kansas	2,279	Washington	3,429
Delaware	576	Oregon	2,225
Maryland	4,070	California	20,601
D.D.	746	Alaska	330
Virginia	4,811	Hawaii	852
W. Virginia	1,794		

3. Below are the test scores of fifty fifth grade students. Make a stem-and-leaf and a schematic plot of this batch. What are the mean and standard deviation of this batch? How well-behaved is this batch? What is the median of the batch? How does it compare to the mean?

72	112	56	104
67	135	97	66
76	102	97	78
77	93	63	82
92	87	53	81
85	81	112	96
79	106	100	72
65	71	49	83
75	82	77	67
83	112	93	78
89	102	86	99
102	96	90	105
118	80		

193

4. a. As a member of the mayor's task force on residential integration you have been asked to make a study of the distribution of nonwhites in Omaha, Nebraska. The data below are from the 1970 U.S. Census of Population. They give the percent of the population of each census tract in Omaha that is nonwhite. Put them in the form of a stem-and-leaf display.

0.253	0.000	3.811	0.559	2.306	10.188
27.276	21.528	43.338	44.603	37.037	42.392
52.831	35.694	98.315	69.719	1.597	6.769
6.941	1.329	0.357	1.095	0.511	0.247
0.121	0.233	0.593	0.591	0.689	4.660
0.119	6.90E-02	0.777	0.322	0.152	3.39E-02
7.27E-02	0.110	8.64E-02	0.421	0.544	0.117
1.056	0.475	0.369	0.363	0.153	0.132
0.206	0.145	0.563	0.174	7.943	23.754
17.297	5.663	9.35E-02	5.58E-02	5.33E-02	0.605
15.269	20.628	1.875	1.507	1.336	1.083
0.326	0.288	0.129	9.57E-02	0.296	4.82E-02
5.96E-02	0.446	5.94E-02	7.41E-02	2.57E-02	3.39E-02
7.05E-02	0.118	0.114	1.313	1.473	5.05E-02
0.239	0.128	14.012	0.232	0.153	0.207
0.262	5.44E-02	11.111	2.37E-02	6.79E-02	5.29E-02

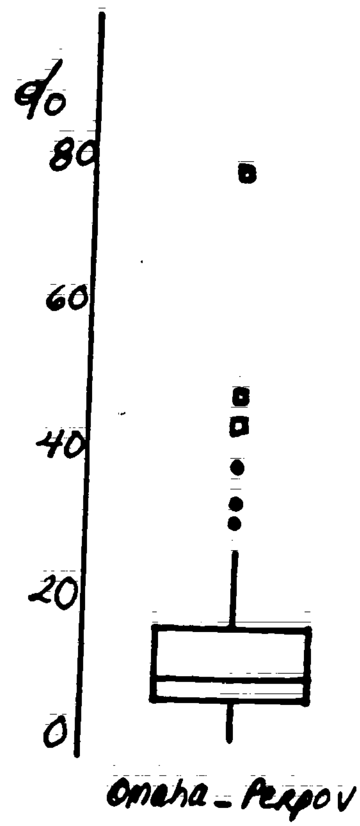
b. Prepare a five number summary of these data and present it as a letter-value display

c. Present a fenced letter display.

d. Prepare a schematic plot of these data.

e. The members of the mayor's task force are unfamiliar with stem-and-leaf display. Put the data into the form of a histogram. What information has been lost in going from one to the other?

5. The figure below is a schematic plot of the percent of families in Omaha census tracts with incomes below the poverty level in 1970. Label the different kinds of outliers, the hinges, and the median and indicate the values that these points correspond to.



195

6. Test scores on a group of children (age 10) from the same neighborhood were as follows:

   2.95, 3.22, 3.32, 3.40, 3.59, 3.73, 3.80

   To study the effect on various summaries of a change in one value in a batch, vary the value shown as 3.22. Examine the effects on the mean, median, standard deviation, and $S = 3/4 \times$ midspread, as the moving value goes from below 2.95 to above 3.80. Use intervals of .2 (i.e., moving value first equals 2.90, then 3.10, then 3.30, etc.) Also move the value to 4.90.

7. (a) Draw two different stem-and-leaf displays of the welfare data by stretching or squeezing the stem. Which do you think is preferable? Why? Do they both give the same information about the batch?

   (b) What do you infer from your analysis about the cost of welfare per inhabitant? Pay particular attention to outliers.

   (c) Summarize the data in a letter-value display. Now exclude outliers and present in a letter-value display. Comment on the differences.

   Data are on the next page.

1972 Cost of Welfare per Inhabitant by State

Alabama	$ 40.43	Nebraska	$ 27.76
Alaska	50.09	Nevada	20.36
Arizona	27.05	New Hampshire	32.73
Arkansas	49.70	New Jersey	50.27
California	97.30	New Mexico	35.98
Colorado	52.48	New York	89.37
Connecticut	38.03	North Carolina	25.59
Delaware	36.56	North Dakota	29.45
District of Columbia	100.44	Ohio	30.85
Florida	23.52	Oklahoma	54.90
Georgia	44.55	Oregon	35.94
Hawaii	52.21	Pennsylvania	50.55
Idaho	30.40	Rhode Island	50.87
Illinois	55.43	South Carolina	16.35
Indiana	23.53	South Dakota	28.75
Iowa	33.61	Tennessee	33.51
Kansas	30.30	Texas	33.61
Kentucky	35.92	Utah	33.35
Louisiana	51.36	Vermont	54.65
Maine	51.61	Virginia	25.67
Maryland	35.64	Washington	48.16
Massachusetts	71.23	West Virginia	34.01
Michigan	56.72	Wisconsin	32.39
Minnesota	45.89	Wyoming	18.50
Mississippi	48.22		
Missouri	40.43		
Montana	24.24		

(Source: 1975 World Almanac, Page 157)

8. A recent study of career choice listed the percentage of doctorate-holders who held a job in the same field as their doctorates. Prepare a stem-and-leaf display of the results.

Do the data cluster or are they uniformly spread out? Are the data symmetrical? Are there any outliers? If there are any clusters or outliers, where do they occur? What can you infer about career choice from your analysis?

Mathematics	91%
Physics, Astronomy	90%
Chemistry	84%
Earth Sciences	93%
Engineering	92%
Agriculture, Forestry	73%
Health Sciences	78%
Biochemistry, Physiology, Biostatistics	70%
Anatomy, Cytology, Genetics, Entomology	47%
Botany, General Biology, Botany	51%
Anthropology, Archaeology	NA
Sociology	79%
Economics, Econometrics	76%
Political Science, International Relations	81%
History	85%
Language, Literature	83%
Philosophy, Arts	70%
Business, Theology	73%
Education	81%
Psychology	90%

9. Identify the following batches as bounded numbers, amounts, counts, or differences.

(a) The average hourly earnings in manufacturing industries were:

1950	$1.44
1955	1.86
1960	2.26
1965	2.61
1970	3.36

(b) Grain receipts at western Canadian grain centers in 1972-73 (in Thousands of Bushels):

Wheat	633,258
Oats	32,484
Barley	236,816
Rye	9,252
Flaxseed	18,346
Rapeseed	62,949

(c) Unemployment rate for Americans aged 16 and over:

Spanish	7.5
White	4.3
Black	9.3

(d) Indians in North Dakota, 1970:

Apache	9	Kaw, Omaha	33
Cherokee	50	Lumbee	33
Chippewa	6,721	Shoshone	11
Creek	18	Sioux	3,655
Iroquois	45	Other	1,629

(e) Performances of record long run Broadway plays:

Fiddler on the Roof	3,242
Life with Father	3,213
Tobacco Road	3,182
Hello Dolly	2,844
My Fair Lady	2,717
Man of LaMancha	2,328

(f) Change in population in major Alaskan cities between 1960 and 1970 census:

Anchorage	3,844
Fairbanks	1,460
Juneau	- 747
Ketchikan	511
Spenard	9,015

(g) Sales of recreational vehicles, 1973:

Travel trailers	212,300 units
Motor homes	129,000 units
Truck campers	89,800 units
Camping trailers	97,700 units
Pickup covers	223,700 units

(h) Percent of high school seniors with no college or vocational school plans, by family income (1974)

Under $5,000	27.1
$5,000 - $7,499	23.5
$7,500 - $9,999	21.0
$10,000-$14,999	19.7
$15,000-$24,999	15.3
$25,000 and over	6.9

200

(i) Distance from home to college for first-time students in 4-year colleges, 1973:

Distance, in Miles	Percentage Distribution
10 or less	15.8
11-50	19.9
51-100	15.9
101-500	35.6
more than 500	12.8

(j) Average raise received by instructional staff in universities at beginning of 1975-76 school year:

Professor	$1,748
Associate Professor	1,045
Assistant Professor	848
Instructor	857

(k) Distribution and frequency of low-income families, by place of residence

Residence	Number in Group (Millions)
Urban	27.5
Rural non-farm	11.4
Rural farm	4.8

(l) U. S. shoreline, in statute miles:

Atlantic coast	28,673
Gulf coast	17,141
Pacific coast	40,298
Arctic coast (Alaska)	88,633

10. Below is a list of food indexes for major U.S. cities in July, 1974. Prepare a stem-and-leaf display and five number summary. How does this batch compare to the hypothetical well-behaved batch?

Atlanta	162.7	Milwaukee	154.8
Baltimore	163.1	Minneapolis	162.9
Boston	161.6	New York	165.0
Buffalo	159.9	Philadelphia	164.5
Chicago	160.4	Pittsburgh	162.9
Cincinnati	163.2	Portland	154.8
Cleveland	159.2	St. Louis	157.6
Dallas	155.7	San Diego	159.2
Detroit	162.6	San Francisco	154.8
Honolulu	156.9	Scranton	159.3
Houston	162.7	Seattle	155.3
Kansas City, MO	160.7	Washington, D.C.	164.4
Los Angeles	155.5		

202

11. Population densities by state are skewed towards low density. Select an appropriate transformation for symmetry, based on the summary numbers. Do the transformation and present the results in a stem-and-leaf display. Discuss clustering, outliers and symmetry.

Population Density by State, 1970
(people per square mile)

Alabama	67.9	Montana	4.8
Alaska	0.5	Nebraska	19.4
Arizona	15.6	Nevada	4.4
Arkansas	37.0	New Hampshire	81.7
California	127.6	New Jersey	953.1
Colorado	21.3	New Mexico	8.4
Connecticut	623.7	New York	381.3
Delaware	276.5	North Carolina	104.1
District of Columbia	12,401.8	North Dakota	8.9
Florida	125.5	Ohio	260.0
Georgia	79.0	Oklahoma	37.2
Hawaii	119.8	Oregon	21.7
Idaho	8.6	Pennsylvania	262.3
Illinois	199.4	Rhode Island	905.5
Indiana	143.9	South Carolina	85.7
Iowa	50.5	South Dakota	8.8
Kansas	27.5	Tennessee	94.9
Kentucky	81.2	Texas	42.7
Louisiana	81.0	Utah	12.9
Maine	32.1	Vermont	47.9
Maryland	396.6	Virginia	116.9
Massachusetts	727.0	Washington	51.2
Michigan	156.2	West Virginia	72.5
Minnesota	48.0	Wisconsin	81.1
Mississippi	46.9	Wyoming	3.4
Missouri	67.8		

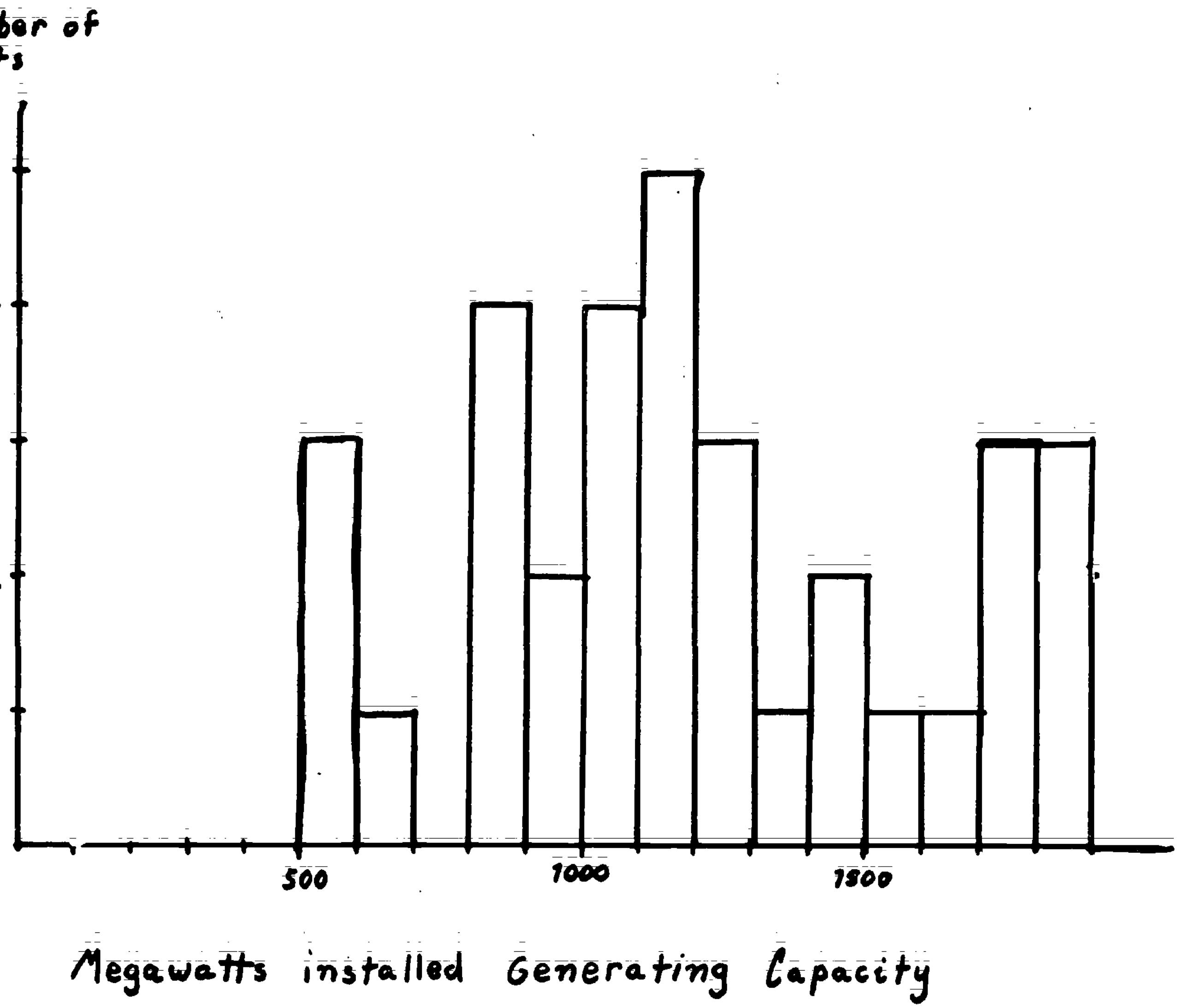
(World Almanac, p. 154)

Homework Solutions
Unit 1

1. Installed Generating Capacity in Megawatts (Sorted)

	1140.5
510.3	1150.2
542.5	1152
550	1255.2
691.2	1255.4
823.3	1270.4
826.3	1303.6
846.6	1408
853	1485.2
908	1550.6
950	1636.2
1016.7	1700
1069.1	1750
1086.3	1770.8
1096.8	1827.7
1100.3	1872
1125	1872

The data cluster between 800 and 1300, so they are not uniformly spread out. The batch is roughly symmetrical and has no outliers. The municipality will be interested in noting that plant sizes in similar municipalities range from 500-1900 megawatts installed generating capacity. Central values of that range are observed more often than the extremes and a typical value (the median) is 1140.5 megawatts.



ber of
ts
500
1000
1500
Megawatts installed Generating Capacity

XVI.I.154

2.

a) The interval of population size into which the largest number of states falls is zero to one million.

b) The number of states in that interval is thirteen.

c) California and New York are the outliers of this batch.

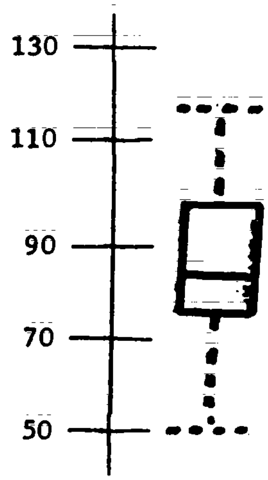
d) A logarithmic transformation of this batch would promote symmetry by compressing the larger values in the batch while stretching out the smaller values.

3.

UNIT = 1

A.

4	9
5	3 6
6	3 5 6 7 7
7	1 2 2 5 6 7 7 8 8 9
8	0 1 1 2 2 3 3 5 6 7 9
9	0 2 3 3 6 6 7 7 9
10	0 2 2 2 4 5 6
11	2 2 2 8
12	
13	5



B. MEAN = $\Sigma X_i/50 = 86.46$

STD. DEV. = $\sqrt{\dfrac{\sum^{N}(X_i-\bar{X})^2}{N}} = 1.75$ Where N=50 and $\bar{X}$ = mean

C. MEDIAN = 84 (depth - 25h)

This batch is roughly symmetric and Gaussian in shape. The median and mean are approximately equal. Midhinge = 87.5; Midextreme = 92; these values are also close to the median, but there is clearly an upward trend. There is one outside value, which could not be the case in a well-behaved batch. However, for real data, this batch comes remarkably close to being well behaved. Notice also that 3/4 x midspread = 17.25, which is close to the standard deviation.

4. (A) Unit = .01%

```
0* |02233455555566777899
1  |11111222345557
2  |0033345689
3* |22566
4  |247
5  |145699
6* |08
7  |7
8  |
9* |
```

Unit = .1%

```
0**
1  |0003334558
2  |3
3  |8
4  |6
5  |6
6  |79
7  |9
8  |
9  |
```

Unit = 1%

```
0***|
1   |01457
2   |0137
3   |57
4***|234
5   |2
6   |9
7***|
8   |
9   |8
```

(B) #96 % pop. nonwhite in Omaha census tracts

```
M 48h |      .355      |
H 25  |.11        1.8  |
   1  | 0         98   |
```

208

4. (C) #96 % nonwhite pop. in Omaha census tracts

M 48h		.355	
H 25	.11	1.8	1.69
1	0	98	

	2.54	
f	-2.43	4.34
	XXX	three
F	-4.97	6.88
	XXX	nineteen

ADJ: 0, 3.8

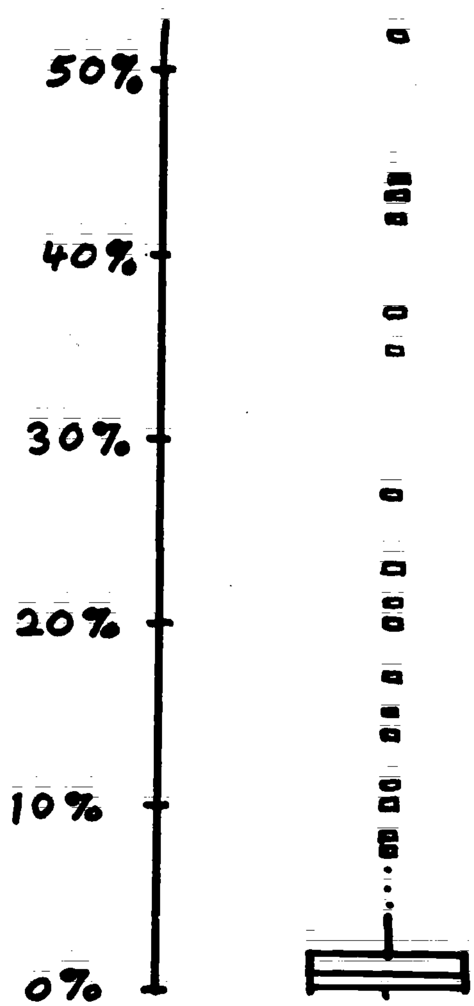
OUT: 4.6, 5.6, 6.7

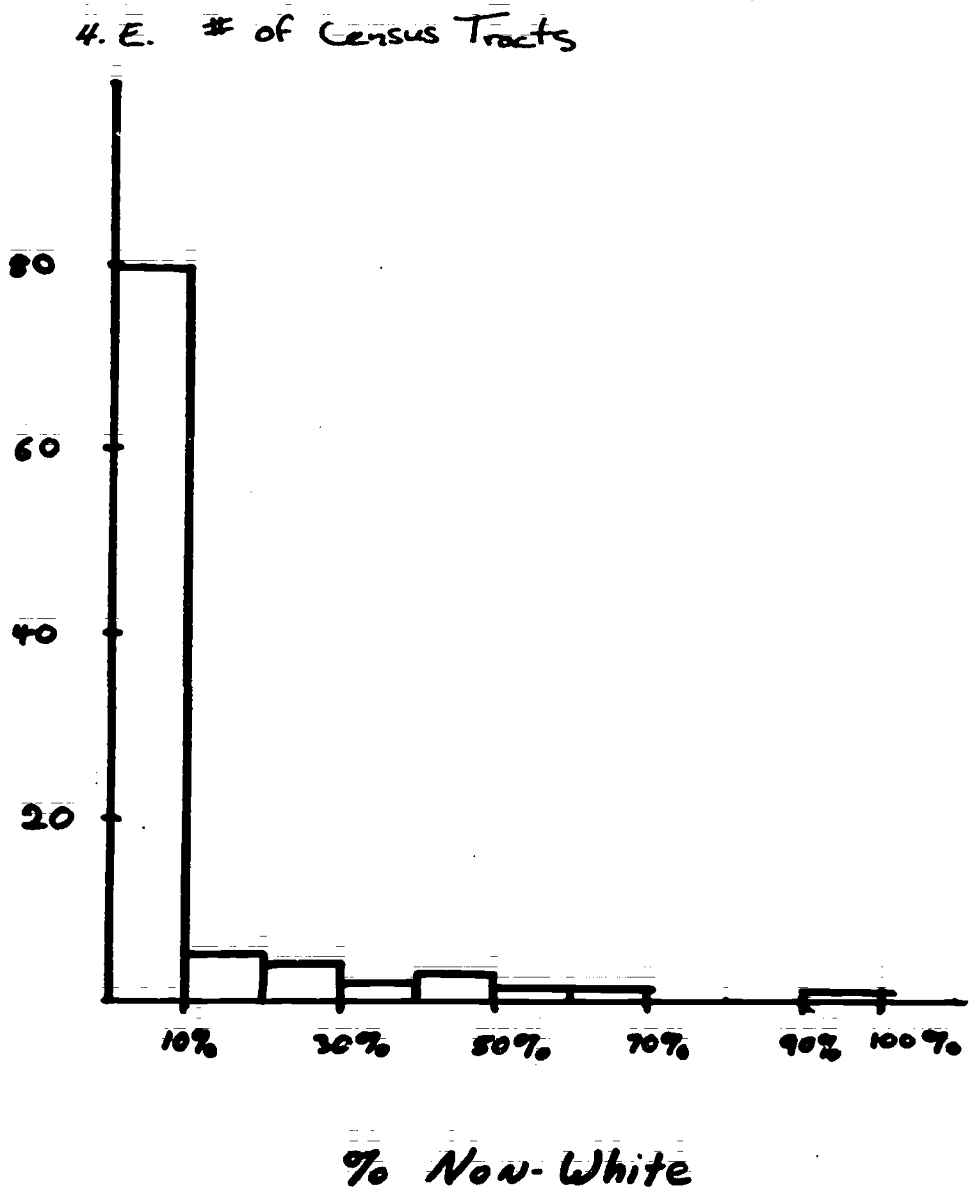
FAR: 6.9, 7.9, 10, 11, 14, 15, 17, 20, 21, 23, 27, 35, 37, 42, 43, 44, 52, 69, 98

(E) Although the histogram shows that almost 80% of the census tracts' populations are less than 10% nonwhite, you lose the information that 21 tracts are less than .1% nonwhite, that 62 tracts have less than 1% nonwhite, and so on. You also lose the specific values.

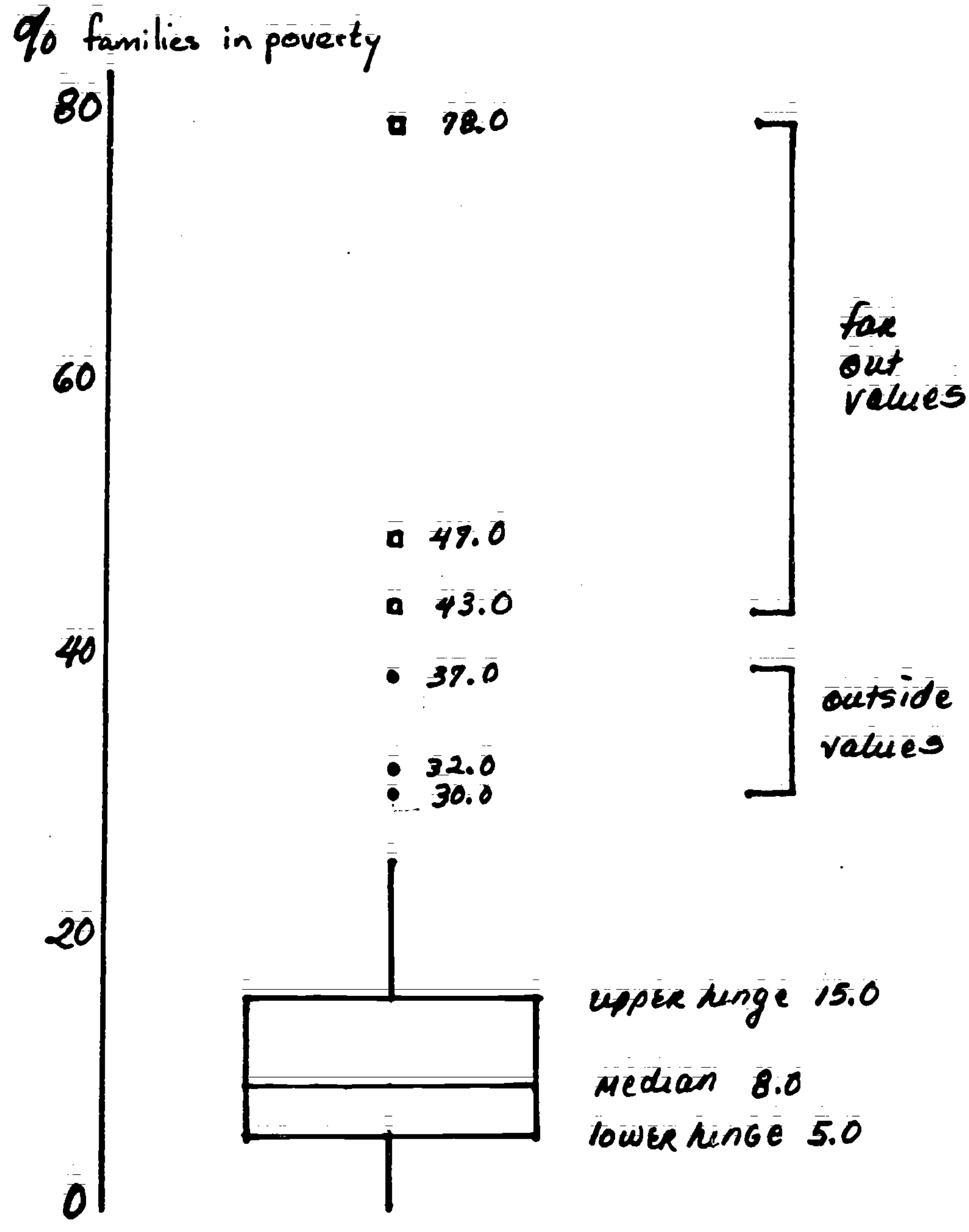
In short you have lost a lot of the detailed information.

4. D Percent Non-white per Census Tract in Omaha




4. E. # of Census Tracts
80
60
40
20
10%
30%
50%
70%
90%
100%
% Non-White

QMPM
Problem 5



6)

moving value	mean	median	$sd = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$	$sd = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$	$\hat{S} = \frac{3}{4} \times$ midspread
2.9	3.38	3.40	.35	.33	.39
3.1	3.41	3.40	.32	.29	.34
3.22	3.43	3.40	.30	.28	.29
3.3	3.44	3.40	.29	.27	.26
3.5	3.47	3.50	.29	.26	.23
3.7	3.50	3.59	.30	.28	.26
3.9	3.53	3.59	.33	.30	.30
4.9	3.67	3.59	.61	.57	.30

While the mean varies with every shift of the moving value, the median jumps twice, but remains constant for any extreme magnitude of the moving value. Similarly, while the standard deviation moves with every shift of the moving value, the midspread does shift, but remains constant for any extreme magnitude of the moving value. What is demonstrated is the resistance of the median and the midspread to extreme values of the batch.

7 (a) Two likely stem-and-leaf displays are:

1972 Cost of Welfare per Inhabitant by State

unit = $1

1	
•	68
2	3340
•	775985
3	0302033342
•	865555
4	040
•	9588
5	0221104004
•	56
6	
•	
7	1
•	
8	
•	9
9	
•	7
10	0

unit = $1

s	6
1	8
2	0
t	33
f	554
s	77
2•	98
3	000
t	323332
f	45555
s	6
3•	8
4	00
t	
f	45
s	
4•	988
5	011000
t	22
f	445
s	6
5•	
HI:	71, 89, 97, 100

There is no one preferable scale; as long as you can defend it, you may select any scale. An argument could be made in favor of the scale on the right (the stretched scale) because it gives a better idea of the shape.

Both scales give the same information about the batch; one might argue that it is easier to read the information from the stem-and-leaf on the right, or that the scale on the left emphasizes the cluster and the outliers.

(b) In 1972, the cost of welfare per inhabitant ranged from $16.35 in South Carolina to $100.44 in the District of Columbia. There is a cluster of values around $30-35 and a smaller cluster at $48-52. There are four outliers--Massachusetts, New York, California, and the District of Columbia--all of which were high. The outliers are all states with large metropolitan areas. D.C., with the highest cost per inhabitant, is exclusively urban. The lowest costs are associated with rural states: S.Carolina, Wyoming, Nevada. Thus high welfare costs per inhabitant are associated with urban areas.

7. (c)

#51

M	26	35	
H	13h	30	50
E		16	100

#47

M	24	35	
H	12h	29.5	49.5
E		16	56

When high outliers are excluded, the only number of the five number summary with a major change is the maximum. Hinges change very little. The batch is much more symmetric with outliers excluded.

8.

LO	47,51
7	3003
.	896
8	4131
.	5
9	10320
	1 value missing

unit = 1%

The data do not cluster, but are not quite uniform either, due to the gap at 86-89 followed by several values at 90-93. However, by contrast with bellshaped and skewed batches, this one may be considered uniform. By the same reasoning, the data are relatively symmetric. There are two low outliers: 47 and 51. Roughly 70-93% of doctorate-holders in various fields have jobs in the same field. Biological sciences are an exception, where fewer doctorate-holders are employed in their field.

9.

a) amount
b) amount
c) bounded numbers
d) count
e) count
f) difference
g) count
h) bounded numbers
i) bounded numbers
j) difference
k) count
l) amount

10. Two likely stem-and-leaf displays are:

unit = 1

1 5 f	5 5 4 4 4 5
s	6 7
1 5	9 9 9 9
1 6	1 0 0
t	2 3 3 2 2 2 2
f	5 4 4

and

unit = .1

1 5 4	8 8 8
5	7 5 3
6	9
7	6
8	
1 5 9	9 2 2 3
1 6 0	4 7
1	6
2	7 6 7 9 9
3	1 2
4	5 4
1 6 5	0

The five number summary is:

	n=25	
13 M	160.4	
7 H	156.9	162.9
E	154.8	165.0

To compare with the hypothetical well-behaved batch, calculate the midhinge and midextreme.

$$\text{Midhinge} = \frac{156.9 + 162.9}{2} = 159.9$$

$$\text{Midextreme} = \frac{154.8 + 165.0}{2} = 159.9$$

This batch appears well-behaved in that its median, midhinge, and midextreme have approximately the same value. However, it is clear from either stem-and-leaf display that the batch is more uniform than bell shaped; therefore, it is not an example of a well-behaved batch.

11) After a logarithmic transformation, the data are:

unit = .1

```
-0 . | 3
+0 . |
   * | 5 6 6 9 9 9 9
 1 . | 1 1 2 3 3 4
   * | 5 5 5 6 6 6 6 7 7 8 8 8 9 9 9 9 9 9 9
 2 . | 0 0 0 1 1 1 1 3 4 4 4
   * | 5 6 7 8 9 9
 3 . |
   * |
 4 . | 0
   * |
```

These data now more closely approximate the well-behaved batch. They are roughly symmetric, with a cluster around the center (median 1.8). There are two outliers: one high at 4.0 (D.C.) and one low at -.3 (Alaska).

218

Unit 1 Quiz

I. Answer the following questions briefly and generally.

1. What is a batch?
2. How are the median and mean affected by deviant values in a batch?
3. What is the midspread?
4. What is the midhinge?
5. What techniques are there for condensing a batch?
6. What are the possible advantages of condensing a batch?
7. What are the most common transformations?
8. What is the simple ladder of powers?
9. What are the possible advantages of transforming single batches?
10. What is a well-behaved batch?
11. How may two well-behaved batches differ?
12. How is a well-behaved batch standardized?

II. Below is a list of the infant mortality rates (deaths per 1000 live births) for the sixteen Eastern Montana counties.

Carter	28.4	Powder River	15.0
Custer	14.2	Prairie	15.7
Daniels	26.3	Richland	16.6
Dawson	17.2	Roosevelt	42.2
Fallon	21.1	Rosebud	40.8
Garfield	12.1	Sheridan	19.7
McCone	24.2	Valley	26.2
Phillips	28.1	Wibaux	27.2

Do the following with the data:

1. Sort the batch.

2. Prepare a stem-and-leaf display.

3. Make a five-number summary.

4. Make a schematic plot.

5. Discuss the data, based on your work in parts 1-4.

(By now you should know what questions to ask of a batch.)

Unit 1 Quiz
Solutions

I. 1. A batch is a set of similar numbers obtained in some consistent fashion.

2. The median is affected very little by deviant values. Extremely large values may increase the mean a lot, while extremely small values may greatly lower the mean.

3. The midspread is the distance between the hinges (upper hinge − lower hinge). It is a measure of spread.

4. The midspread is the value halfway between the hinges $\left(\frac{\text{upper hinge} + \text{lower hinge}}{2}\right)$. It is a measure of central location.

5. Schematic plots, expanded schematic plots, five number summaries, expanded number summaries are techniques for condensing a batch.

6. The purpose of condensing a batch is to summarize it by describing a typical value and variation of the values, and identifying outliers.

7. Common transformations are of the form $X \rightarrow X^R$ where, $R = -1$, 0, 1/2, 2 (that is, negative reciprocals, logarithms, square roots, and squares). The arcsin of the square root of x is another common transformation.

8. The simple ladder of powers is a plot of x against transformations of x showing the direction and to some extent the rapidity with which the transformation changes the batch.

9. Transformations reexpress the batch in units that are desirable for intended analysis. This usually means increasing symmetry; reducing outliers and variance is also desirable.

10. A well behaved batch has median = mean = midhinge = midextreme, has $s = 3/4\ \Delta H$, has no outliers and resembles a Gaussian function in shape.

11. Two well-behaved batches may differ only in location and scale.

12. To standardize a well-behaved batch, subtract the mean from each value and divide by the standard deviation.

II. 1.

12.1	24.2
14.2	26.2
15.0	26.3
15.7	27.2
16.6	28.1
17.2	28.4
19.7	40.8
21.1	42.2

2. Unit = 1 death per 1000 live births

```
1 * | 24
  . | 55679
2 * | 14
  . | 66788

HI  | 40.8, 42.2
```

3. #16

M 8h	22.6		midspread 11.5
H 4h	16.1	27.6	
E .	12.1	42.2	

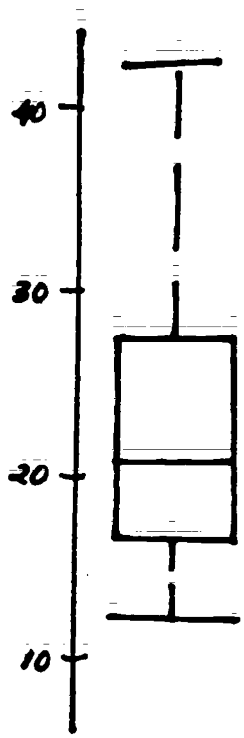
optional:

$\boxed{17.3}$

f	-1.2 ***	44.9 ***	adjacent 12.1 42.2
F	-18.4 ***	62.1 ***	

222

4.



5. Infant mortality rates range from 12.1 to 42.2 deaths per 1000 live births in Eastern Montana. Although the step is large enough that there are no outside values, there are clearly two high outliers. It is worth looking more deeply into the populations and standards of living in Roosevelt and Rosebud counties. The data do not particularly cluster, and a typical infant mortality rate is 22.6.

Quantitative Methods for Public Management

Lecture 2-0. Introduction to Unit 2

Introduction to Unit 2, Analysis of Multiple Batches of Data, Non-Ordered

Lecture Content:

Introduction to the objectives, problem, and notation of Unit 2

Main Topics:

1. Specific Introduction to the Objectives of Unit 2
2. Presentation of General Problem of Unit 2
3. Notation for Unit 2

Topic 1. Specific Introduction to the Objectives of Unit 2

I. Questions to be answered in Unit 2

1. What is a non-ordered multiple batch? (1)

   a. A collection of two or more batches related in some qualitative way

   b. There is no quantifiable ordering of the batches in the collection (2)

2. What analyses can be done on a collection of batches?

   a. How can we best examine or contrast the batches?

      (Note: Since we are studying more than 1 batch, we can discuss comparison of batches)

   b. What is, if one exists, the best unit of analysis for the examination

II. Skills to be mastered in Unit 2 (3)

1. Perceiving and recognizing multiple batches that are non-ordered

2. Organizing the batches to facilitate comprehension, presentation, and analysis

3. Comparison of the batches in the collection

4. Transformations to stabilize variation across the batches

Topic 2. Introduction to the Problems of Unit 2

I. What is a non-ordered multiple batch?

1. Example: 1970 population of the 185 census tracts in Pittsburgh, and the 96 census tracts in Omaha, Nebraska (4)

   a. Relation: 1970 populations, by census tract

   b. Qualitative aspect: 2 major SMSAs

2. Ordered batches are associated in a quantitative manner--we can measure the relationship between the batches in some unit. These will be considered in future lectures.

II. How can we compare the batches? (5)

1. Minimum values--Which batch has the smallest minimum?

2. Maximum values--Which batch has the largest maximum?

3. Median values--How do the typical values of the batches compare?

4. Spreads--Which batch has the smallest midspread? Which has the largest?

5. Shape--Are the batches symmetric? Do the batches have similar stem-and-leaf displays?

6. Units--Are the batches measured in the same units?

III. Is there a better unit of analysis?

1. Are the extremes roughly equal?

2. Do the batches have similar ranges?

IV. Examples

1. Population data (6)

   a. Minimums: 334 (Pittsburgh), 12 (Omaha, why so small?)

b. Maximums: 7910 (Pittsburgh), 12458 (Omaha)

c. Medians: 2602 (Pittsburgh), 3402 (Omaha)

d. Spreads (Midspreads): 2248 (Pittsburgh), 3051 (Omaha)

e. Both batches measured in numbers of persons

f. Omaha batch has larger range than Pittsburgh

g. Cannot compare shape

2. Achievements Pretest Scores for incoming students, by undergraduate studies (7)

a. Minimum (21) and Maximum (48) equal

b. Medians: 31.5 engineering and science, 33 humanities and social science

c. Midspreads: 16 engineering, 13 humanities

d. Batches measured in number of correct answers

e. Batches appear quite similar

3. Life expectancies for various countries by 5 national groupings (8)

Industrial (20 countries)
Petroleum Exporting Countries (9)
High-Income Countries (24)
Middle-Income Countries (19)
Lower-Income Countries (33)

a. Minimums, Maximums, Medians vary greatly

b. Midspreads vary roughly from 3 to 13 years

c. Batches appear quite dissimilar

V. Conclusion

1. Need methods of comparing batches

2. Need methods of determining whether transformation is warranted

Topic 3. Introduction to the Notation of Unit 2

I. Conventions

1. Capital letter ("X") denotes entire data set (9)

2. First subscript (X_i) denotes specific batch

3. Second subscript (X_{ij}) denotes specific element in a specific batch

II. Example: Life expectancies

1. Let X = Life expectancies for countries

2. Let X_1 = Life expectancies for Industrial countries

   X_2 = Life expectancies for Petroleum exporting countries

   .
   .
   .

   X_5 = Life expectancies for Lower-income countries

3. Let X_{11} = Life expectancy for Australia

   X_{12} = Life expectancy for Austria

   .
   .
   .

   $X_{5,33}$ = Life expectancy for Zaire

   Let X_{ij} = Life expectancy for country j in batch i

228

Lecture 2-0
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Topic I		
Section I		
1.a	1	Multiple Batch
1.b	2	Non-ordered Batches
Section II		
1.	3	Topics for Unit 2
Topic 2		
Section I		
1.	4	1970 populations of Pittsburgh & Omaha
Section II		
1.	5	Questions to be answered for Multiple Batches
Section IV		
1.	6	Populations, several values indicated
2.	7	Achievement Pretest Scores
3.	8	Life Expectancies for Various Countries
Topic 3		
Section II		
1.	9	Notation

Multiple Batch

A non-ordered multiple batch of data is a collection of two or more batches related in some qualitative way.

In Unit 2:

We learn to analyze multiple batches of data that are unordered.

230

There is no quantifiable ordering of the batches when the collection is non-ordered [2]

Examples of Non-Ordered Multiple Batches:

1. 1970 populations of the 185 census tracts in Pittsburgh and the 96 census tracts in Omaha
2. Test scores of men and women in this class

Examples of Ordered Multiple Batches:

(discussed in unit 3)

1. Test scores for this year's and last year's Quantitative Methods classes on the fall final

   Quantifiable by year

2. Life expectancies for countries with GNP above $10 billion and for countries with GNP below $10 billion

   Quantifiable by GNP

Topics for Unit 2: [3]

1. Perceiving and recognizing multiple batches, non-ordered
2. Organizing the batches using analytic tools
3. Comparison of the batches
4. Transformations to stabilize the variation or spread

2-0

1970 Populations of the Census Tracts of Pgh. and Omaha. [4]

Pittsburgh

972	4082	1972	391	631	735	1930
3062	2919	2424	6087	729	3689	2437
2085	2973	3712	2505	1919	3294	449
4187	1050	1645	3629	453	1645	2447
1067	1359	2855	1876	2915	1405	2388
2471	728	1205	2382	3122	1019	1410
765	2776	2135	1349	3628	4415	3153
5747	2135	1330	4730	2942	3469	4095
2155	4247	1472	3832	1452	3378	1971
1253	2316	3092	4060	3945	2602	840
2744	3228	3769	4058	4014	1645	5300
5269	2979	5148	3268	3133	4520	6003
5319	5435	1212	2041	2068	1577	2084
1521	4615	3994	7910	3188	4392	4758
1985	5203	484	5880	3962	3752	1004
3156	3578	2398	1424	3820	2019	1418
4719	3509	6796	5371	5630	3765	7425
996	2607	2396	1870	2574	603	2658
2297	335	6235	3121	791	1558	1343
2670	1227	1159	2579	569	500	442
3330	2779	345	4327	2987	2254	2569
1792	2932	2125	1056	2325	1963	719
3103	2487	1193	3291	1044	4561	3609
1289	3853	3985	2857	4437	1399	2144
3812	2612	1640	3921	992	1986	3425
6242	5018	6527	955	5300	1289	2829
3297	3413	334				

Omaha

5524	12	3254	3040	2290	3573	3142
4004	1959	2177	2530	2241	1440	720
653	1212	2755	1566	1700	2400	3357
2648	2542	3244	3312	3004	2359	2540
3620	5400	7501	4350	2703	3110	4602
2954	5501	5476	3473	5457	2756	2573
1326	1094	3240	2201	3912	2269	2912
5522	5059	5173	4079	3410	3197	4379
6414	5374	5627	5702	3471	3054	5972
3450	6139	923	6130	9366	6952	7315
5401	12456	5035	2466	6733	4049	7703
8054	9926	7644	5267	830	1033	11074
4109	3114	992	1725	3269	4347	1520
135	4213	5000	7566	7356		

232

2-0

Questions to be Answered for Multiple Batches.

1. Minimum Data Values
2. Maximum Data Values
3. Medians
4. Spreads
5. Shape
6. Units

Population Data, several data values indicated.

[6]

Pittsburgh

972	4082	1972	391	631	735	1938
3062	2919	2424	6887	729	3689	2437
2085	2973	3712	2505	1919	3294	449
4187	1050	1645	3629	453	1645	2447
1867	1359	2855	1876	2915	1405	2388
2471	728	1205	2382	3122	1019	1410
765	2776	2135	1349	3628	4415	3153
5747	2135	1330	4730	2942	3469	4095
2155	4247	1472	3832	1452	3378	1971
1253	2316	3092	4060	3945	2602 med	848
2744	3228	3769	4858	4014	1645	5300
5269	2979	5148	3268	3133	4520	6003
5319	5435	1212	2841	2068	1577	2084
1521	4615	3994	7910 Max	3188	4392	4758
1985	5203	484	5880	3962	3752	1884
3156	3578	2398	1424	3820	2019	1418
4719	3509	6796	5371	5630	3765	7425
996	2607	2396	1870	2574	603	2658
2297	335	6235	3121	791	1558	1343
2670	1227	1159	2579	569	588	442
3338	2779	345	4327	2987	2254	2569
1792	2932	2125	1056	2325	1963	719
3103	2487	1193	3291	1044	4561	3609
1289	3853	3905	2857	4437	1399	2144
3812	2612	1640	3921	992	1906	3425
6242	5818	6527	955	5300	1289	2829
3297	3413	334 min				

Omaha

5524	12 min	3254	3040	2298	3573	3142
4004	1959	2177	2538	2241	1448	720
653	1212	2755	1566	1700	2408	3357
2648	2542	3244	3312	3004	2359	2540
3628	5408	7591	4350	2703	3110	4602
2954	5501	5476	3473	5457	2756	2573
1326	1894	3248	2201	3912	2269	2912
5522	5859	5173	4079	3410 med	3197	4379
6414	5374	5627	5782	3471	3854	5972
3450 med	6139	923	6130	9366	6952	7315
5401	12458 max	5035	2466	6733	4049	7783
8854	9920	7644	5267	838	1833	11074
4189	3114	992	1725	3269	4347	1528
135	4213	5888	7566	7356		

2-0

[7]

Achievement Pretest Scores for Incoming Students, by Undergraduate Major

Engineering and Science		Humanities and Social Science
21.	min	21.
21.		22.
23.		24.
24.		25.
25.		25.
26.		27.
26.		27.
27.		28.
27.		29.
28.		30.
30.	med	33.
33.		33.
35.		33.
37.		33.
41.		37.
41.		37.
42		40
43.		40.
43.		41.
46.		42.
47.		47.
48.	Max	48.

midspread [16] (Engineering and Science)

midspread [13] (Humanities and Social Science)

Life Expectancies for Various Countries by 5 National Groupings, 197[illegible] [8]

QMPM

Industrial	
Australia	71.1
Austria	70.5
Belgium	70.6
Canada	72.0
Denmark	73.3
Finland	69.3
France	72.3
W. Germany	70.3
Ireland	70.8
Italy	70.7
Japan	73.2
Netherlands	73.8
New Zealand	71.1
Norway	74.0
Portugal	68.[illegible]
S. Africa	65.0
Sweden	74.7
Switzerland	72.[illegible]
Great Britain	72.0
United States	71.[illegible]

Petroleum	
Algeria	50.7
Ecuador	52.4
Indonesia	47.5
Iran	50.[illegible]
Iraq	5[illegible].6
Libya	52.1
Nigeria	37.0
Saudi Arabia	47.3
Venezuela	66.4

Higher Income	
Argentina	67.2
Brazil	60.7
Chile	63.3
Colombia	45.1
Costa Rica	63.4
D. Republic	57.3
Greece	69.1
Guatemala	49.0
Israel	71.5
Jamaica	64.7
Lebanon	NA
Malaysia	56.0
Mexico	61.4
Nicaragua	49.9
Panama	59.3
Peru	54.1
Singapore	67.6
Spain	69.6
Taiwan	68.0
Trinidad	64.[illegible]
Tunisia	51.7
Uruguay	68.6
Yugoslavia	67.[illegible]
Zambia	43.5

Middle-Income	
Bolivia	49.7
Cameroon	41.0
Congo	41.0
Egypt	52.7
El Salvador	58.5
Ghana	37.1
Honduras	49.1
Ivory Coast	30.0
Jordan	52.3
S. Korea	61.9
Liberia	44.9
Morocco	50.5
Papua N. Guinea	46.8
Paraguay	59.4
Philippines	51.1
Syria	52.8
Thailand	56.2
Turkey	53.7
S. Vietnam	50.0

Lower-Income	
Afghanistan	37.5
Bangladesh	NA
Burma	42.3
Burundi	36.8
Cambodia	43.8
C. Africa Rep.	34.5
Chad	32.0
Dahomey	37.3
Ethiopia	38.5
Guinea	27.0
Haiti	32.6
India	41.3
Kenya	49.1
Laos	47.5
Madagascar	36.0
Malawi	39.5
Mali	37.2
Mauritania	41.0
Nepal	40.6
Niger	41.0
Pakistan	51.3
Rwanda	41.0
Sierra Leone	41.0
Somalia	38.5
Sri Lanka	65.9
Tanzania	47.6
Togo	40.5
Uganda	35.1
U. Volta	47.5
S. Yeman	31.6
Yemen	42.3
Zaire	42.3
	38.8

2-0

[9]

Let X denote the entire data set

X = Life Expectancies

Let X_1 = Life expectancies for Industrial Countries

X_2 = Life expectancies for Petroleum Countries

$\vdots$

X_5 = Life expectancies for Lower Income Countrie

Let $X_{1,1}$ = Life expectancy of Australia = 71.1 years

$X_{1,2}$ = Life expectancy of Austria = 70.5 years

$\vdots$

$X_{1,20}$ = Life expectancy of U.S.A. = 71.3 years

$X_{2,1}$ = Life expectancy of Algeria = 50.7 year

$\vdots$

$X_{5,33}$ = Life expectancy of Zaire = 38.8 year

20

Quantitative Methods for Public Management

Lecture 2-1. Comparison of Batches

Comparison of Batches: The use of Numeric and Graphic Methods for Comparison of Multiple Batches (1)

Lecture Content:

1. Discuss extensions of Unit 1 tools for analyzing two or more batches simultaneously
2. Show how these methods convey characteristics of the collection of batches

Main Topics:

1. Comparing several batches of data
2. Effectiveness of these comparison tools

Tools Introduced:

1. Parallel Stem-and-Leaf Display
2. Parallel Schematic Plot

Topic 1. Comparing Several Batches of Data

I. Basic Issue: Comparison of data

1. We know how to organize and condense single batches effectively
2. Often interesting data sets contain qualitatively related multiple batches
3. Need techniques to examine them simultaneously
4. Need to organize the batches in a consistent, reliable, and effective manner to facilitate comparison and analysis

II. Problem: Can the tools of Unit 1 be used to analyze two or more batches?

1. Develop simple rules for extending the elementary techniques of previous unit
2. First step in analysis should be organization of the batches
3. Organization should be followed by a condensation of information
4. Specific questions to be answered:
   a. How do extremes of the batches compare?
   b. Are the medians of the batches similar?
   c. Are the midspreads of the batches equal?
   d. How do the shapes of the batches compare?
5. Remember the batches must be non-ordered. Ordered batches are discussed in Unit 3 where we concentrate on the relationship between the batches and the appropriate ordered scale

III. Solution: organization and condensation tools computed in parallel

1. Parallel stem-and-leaf display

2. Parallel schematic plot

IV. Methods

1. Parallel stem-and-leaf display: Organization tool

   a. Example shows a parallel stem-and-leaf display of the 1970 populations of Pittsburgh and Omaha census tracts (2)

   b. Features

      i. Simple idea

      ii. Same features as with single batch:

         A. "Face validity"

         B. Retains information on individual data values

         C. Flexible

      iii. Easy to construct

   c. Analytic Qualities

      i. Extremes easily located

      ii. 5-number summaries found using depths for each batch

      iii. Shapes of batches

   d. Procedure

      i. Choose a convenient unit, one for all batches together

      ii. Separate every data value into a stem and a leaf

      iii. Find smallest minimum and largest maximum for the entire batch

      iv. Write down the stems, one set for all batches

      v. For each batch, place leaves on correct stem

      vi. Batches are separated in the display, with leaves placed in parallel groups

e. Example: 1970 populations for Pittsburgh and Omaha

i. Convenient unit for display is 100 persons

ii. We take separate stem-and-leafs and put them side by side, with common set of stems (2)

iii. Parallel displays shows:

A. Difference in shape

B. Difference in spread

C. Omaha outliers (11874, 12458)

iv. Square root transformation improves symmetry of both batches (3)

f. Another example: Undergrad Cumulative Average for most incoming masters students by undergraduate background

(Make parallel stem-and-leaf on board)
(Unit = 1)
(Note resemblance or lack thereof)

g. Parallel stem-and-leaf displays on computer:

Use STEM once per batch, specify same UNIT, LPS, HICUT, LOCUT, for each batch. Paste stems together

2. Parallel Schematic Plot: Graphical Condensation

a. Example: 1970 populations for Pittsburgh and Omaha (4)

b. Features

i. Useful in discussing appearance of batches

ii. Adequate comparison tool for nearly all collections

iii. Computable from parallel stem-and-leaf

c. Analytic qualities

i. Made on ordinary graph paper

ii. y-axis is common scale for all values in all batches

iii. Extremes, hinges, and medians clearly marked

d. Procedure

i. Determine smallest and largest values in data set to make scale

ii. Compute 5-number summaries

iii. Draw a simple schematic plots, one per batch, in parallel, using common scale

e. Another <u>example</u>: Life expectancies for countries, (5)
classified as to their "wealth" (6)

i. Schematic shows differences in spread and (7)
location

ii. Petroleum similar to middle income, but not in midspread

iii. Downward trend evident

f. Schematic plots, in parallel, on computer:

Use function BOX with all data files as arguments

213

Topic 2. Effectiveness of these Comparison Tools

I. Basic Issue: once condensed into parallel schematic plot how much can we learn about the batches ?

II. Try to answer:

1. Are there any outliers in the data set ?
2. How do the batches compare with respect to shape ?
3. Is there any obvious relation among the medians or midspreads ?

II. Methods

Parallel stem-and-leaf displays and schematic plots answer these questions

(Present several other examples of unordered multiple batches and discuss appearance of each)

Lecture 2-1
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 2-1 Outline
Topic 1		
Section IV		
1.a	2	Parallel stem-and-leaf of populations
1.e.ii	2	Parallel stem-and-leaf of populations
1.e.iv	3	Parallel stem-and-leaf of square roots of populations
2.a	4	Parallel schematic plot of populations
2.f	5	Life expectancies for countries
2.f	6	Parallel stem-and-leaf of life expectancies
2.f.i	7	Parallel schematic plot of life expectancies

[1].

Lecture 2-1

Comparison of Batches:

The use of Numeric and Graphic methods for comparing batches.

Lecture Content:

Extensions of the tools of Unit 1 to facilitate the analysis of two or more batches simultaneously.

Main Topics:

1. Comparing several batches
2. Questions to ask of a batch.

2-1

Parallel Stem-and-Leaf Display 1970 Populations, by Census Tract [2]

Stem	Pgh	Omaha
0***	33334444	01
0·	55667777778999	67899
1	0000112222223333344444	234
1·	55566667888899999999	5577899
2	000001111122333333444444	12222344
2·	55556666677788899999999	5555677799
3	00111111122222233444	0011112222334444
3·	556666777788889999	5689
4***	0000123344	00012333
4·	5567778	6
5	12233334	01234444
5·	6788	55567889
6	022	114
6·	578	79
7***	4	33
7·	9	5567
8***		
8·		8
9		3
9·		9

2-1

217

[3]

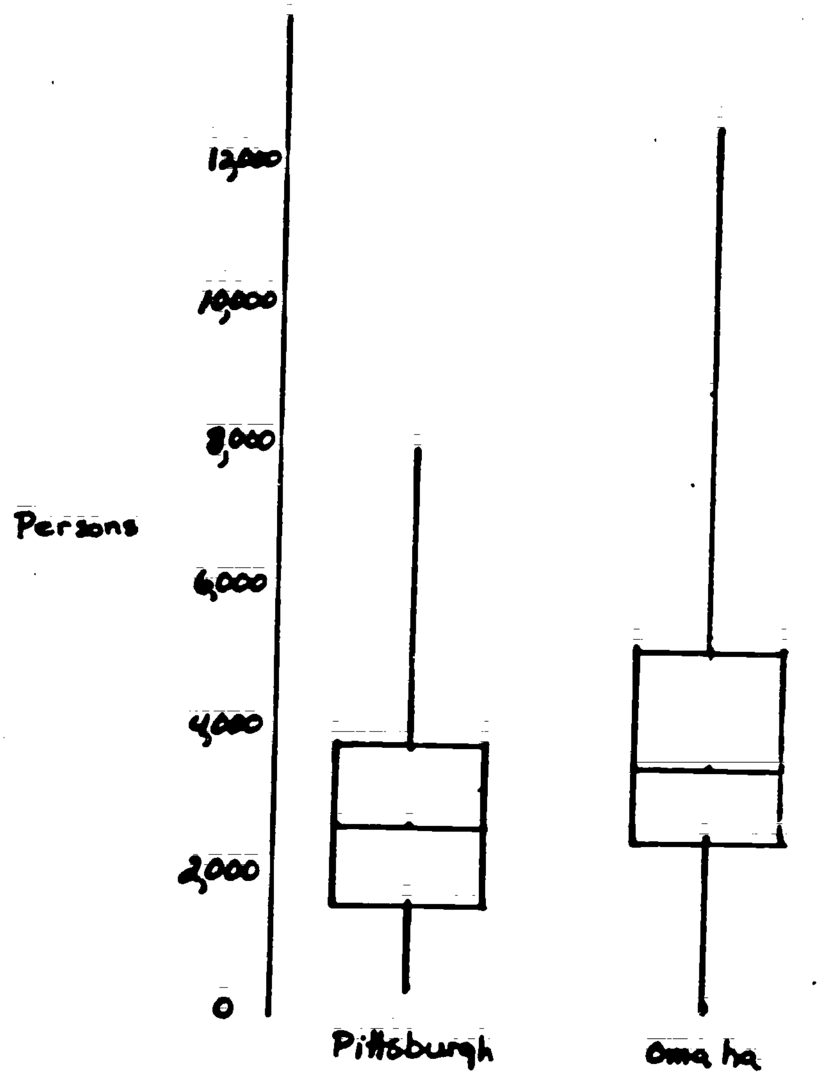
Square Roots of 1970 Pittsburgh & Omaha Populations
Parallel Stem-and-Leaf
(unit = 1)

0		3
0·		
1		1
1·	8889	
2	1112344	
2·	56677789	568
3	011 112224444	014
3·	555566667 777788999	6899
4	000023 333 334 444 44	112 34
4·	555 566 666 778 888 889 999 9	667 77899
5	000 011 111 222 333 344 4444	000 011 223 44
5·	555 556 666 677 777 888 899	555 666 677 778 8889
6	000 001 111 112 222 233 333 4	022 33344
6·	556667 778 889	556 7
7	122 22233	012 333 34444
7·	5566789	566 6788
8	022	023
8·	68	556 778
9		4
9·		69
High		108.0, 111.6

2-1

Parallel Schematic Plots, [4]

1970 Populations, by Census Tract, Pgh. & Omaha



Life Expectancies for Countries in 5 Classes. [5]

NAT_IND_LIFE

71.1	70.5	70.6	72.	73.3	69.8	72.
70.3	70.8	70.7	73.2	73.8	71.1	74.
68.2	NA	74.7	72.1	72.	71.3	

NAT_PETRC_LIFE

50.7	52.4	47.5	50.	51.6	52.1	37
42.3	66.4					

NAT_HIINC_LIFE

67.2	60.7	63.3	45.1	63.4	57.9	69
49.	71.5	64.7	NA	56.	61.4	49
59.3	54.1	67.6	69.6	68.	64.3	51
68.6	67.7	43.5				

NAT_MIDINC_LIFE

49.7	41.	41.	52.7	58.5	37.1	45
30.	52.3	61.9	44.9	50.5	46.8	59
51.1	52.8	56.2	53.7	50.		

NAT_LOWINC_LIFE

37.5	NA	42.3	36.8	43.8	34.5	32.
37.3	38.5	27.	32.6	41.3	49.1	47.5
36.	38.5	37.2	41.	40.6	41.	51.3
41.	41.	38.5	65.9	47.6	40.5	35.1
47.5	31.6	42.3	42.3	38.8		

2-1

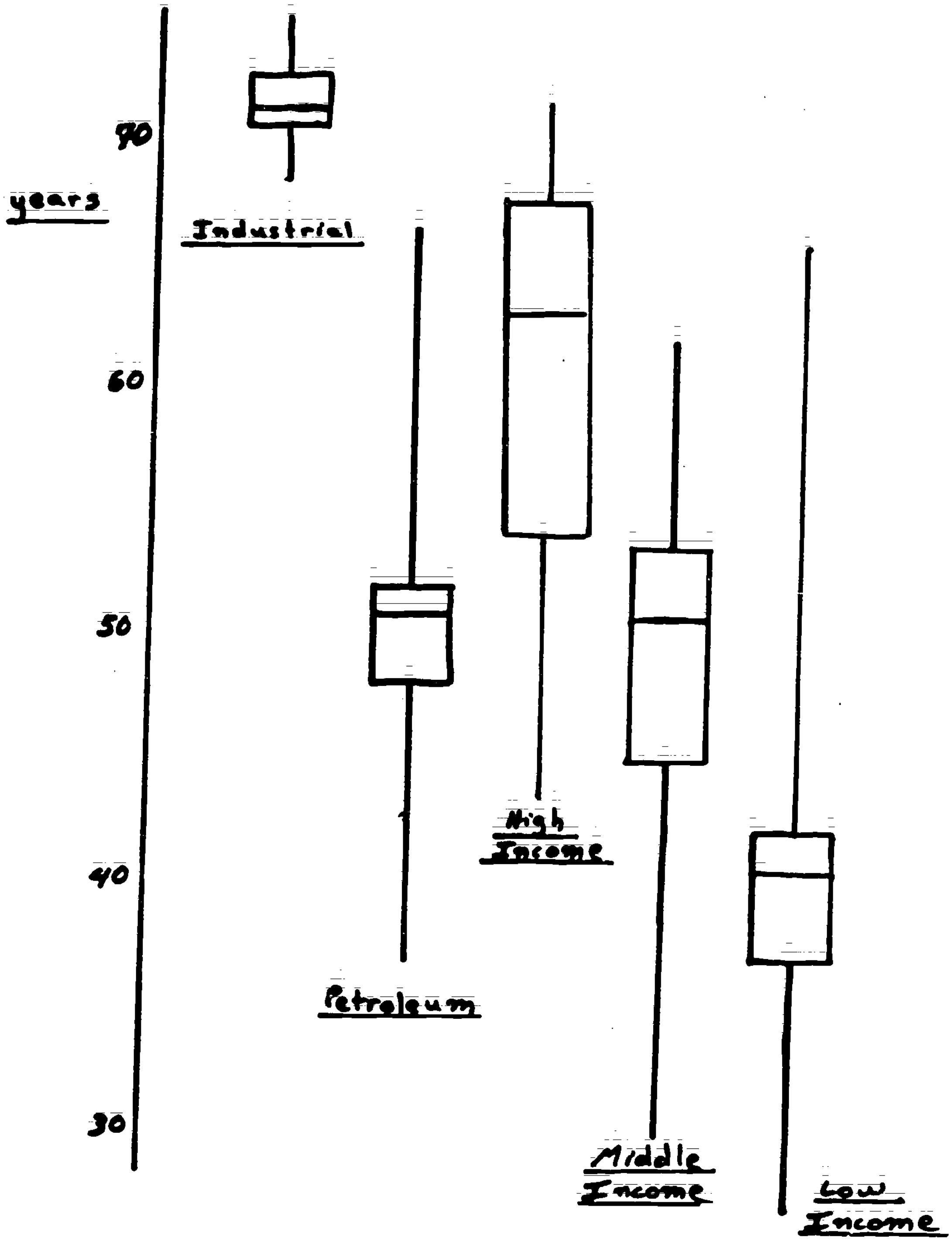
Life Expectancies of Countries, Parallel Stem-and-Leaf [6]

	Industrial	Petroleum	Higher Income	Middle Income	Lower Income
2*					
2·					7
3*				0	1224
3·		7		7	5667778888
4*		2	3	114	00111112228
4·		7	599	699	7779
5*		00122	14	0012223	1
5·			679	689	
6*			013344	1	
6·	89	6	7778899		5
7*	0000011122233344		1		

QMPM

Life Expectancies by Nation, Classified into 5 Batches

[7]



2-1

Quantitative Methods for Public Management

Lecture 2-2. Transformation for Stabilization of Spread

Transformations for Stabilization of Spread: The Use of Various Algebraic Transformations to Equalize Spread Among Batches (1)

<u>Lecture Content</u>:

1. Discuss need for transformation
2. Introduce method of determining a good transformation

<u>Main Topics</u>:

1. Necessity of transforming a multiple batch
2. Use of medians and midspreads in finding a good transformation

<u>Tools Introduced</u>:

1. Median/Midspread Plot

Topic 1. Necessity of Transforming a Multiple Batch

I. Basic Issue: Comparison of batches is difficult if batches differ greatly in spread

1. We know transformations are helpful in changing the shape of single batches

2. When more than 1 batch is being analyzed, comparisons are easier if batches are similar in spread

   a. Example: Parallel Schematic Plot of life expectancies for nations (2)

      i. Difference in spreads in the 5 batches makes conclusions concerning location difficult

      ii. Spreads are roughly equal, except for industrialized nations

   b. Example: Parallel Schematic of Infant Mortality for nations (3)

      i. Locations similar, spreads vary enormously

      ii. If we balance the spread, will locations still be similar ?

   c. Example: Parallel Schematic of Per capita Income for nations (4)

      Note relationship between location and spread

3. If comparisons of location are to be made, task is easier if spreads are equalized

4. We transform batches to equalize or "balance" the spread

5. If comparisons of spread are to be made, transformation is unnecessary; merely "line up" plots so that medians are equal, and compare spreads

6. In conclusion, how much of the difference in location is due solely to location, and how much is due just to difference in spread ?

II. Problem: Want our schematic plots to tell their story as clearly and simply as they can

1. Symmetry of spread within batches is <u>helpful</u> for summarizing single batches

2. Balance of spread between batches is <u>essential</u> for comparisons

III. Solution: Choose transformation to achieve equalization in spread

1. The transformation will usually promote symmetry within batches

2. As with transformations for symmetry, the search for a good transformation is exploratory, and even the best transformation may fail to equalize spread completely.

Topic 2. Use of Medians and Midspreads in Finding a Good Transformation

I Basic Issue: How do we find the best transformation ?

1. We understand that transformation may be essential in comparing batches

2. Since transformation affects the relationship between the medians and the midspreads of the batches, how do we use these values to find the best transformation ?

II. Problem: How do we let the medians and midspreads tell us the correct transformation

1. We are searching for a consistent relation between medians and midspreads

2. The best way to study the relationship of the medians and midspreads is with a scatterplot

3. Could line up schematics, but a scatterplot is more clear

4. Best to look at a scatterplot of log (Median) versus log (Midspread), one ordered pair per batch

5. A linear scatter implies transformation is necessary

III. Solution: Examine slope of the log(Median) vs. log(Midspread) scatterplot

1. Suppose scatterplot was close to linear with an "eyeball" slope of p

2. Correct exponent for the transformation $Y = X^R$ is r = (1-p)

3. Slope tells how far down the "ladder of powers" to move

4. Slope of: 1 = logs (i.e., 1-p = 0)
   2 = negative reciprocals (1-p = -1)
   1/2 = square roots (1-p = (1/2))
   0 = no transformation (1-p = 1)
   -1 = squares (1-p = 2)

IV. Method: Log (Median) / Log (Midspread) Plot

1. Example: Log median / Log midspread plot for Per capita incomes of countries

2. Features

   a. Useful in determining a good transformation to compare batches

   b. Scatterplot made from 5-number summaries of the batches

   c. Relationship of th[illegible] medians and log midspreads determine the co[illegible]nent, r, for the transformation

3. Analytic Qualities

   a. Slope of the [illegible] determines r

   b. Relationship between r and slope, p, is r = (1-p)

   c. Random scatter (p=0) implies no transformation necessary

   d. Plot made on ordinary graph paper

   e. If collection has fewer than 4 batches, may not have enough points to determine slope

   f. Log(median) vs Log(midextreme) (or other measures of spread) plot may be used to determine transformation

   g. For "well-behaved" batches, log(mean) vs. log(standard deviation) is acceptable - note that standardizing obscures differences in level

4. Procedure

   a. Compute 5-number summaries for the batches and find midspreads (5)

   b. Compute the logarithms of the medians and midspreads

   c. On a piece of ordinary graph paper, plot log(median) as x and log(midspread) as Y, one point for each batch; or use log-log paper and plot median vs. midspread directly (6)

   d. Find a slope by choosing two representative points, one at left end of scatter (X_L, Y_L) and one at right end of scatter (X_R, Y_R). Slope = $(Y_R - Y_L)/(X_R - X_L) = p$

e. If preferred, an "eyeball slope" may be used--a slope fit to the data by eye

f. Correct exponent of transformation, r, is (1-p)

g. With new exponent, find transformation of 5-number summaries

h. Make new schematic plot of transformed data to compare batches (7)

5. Another example: Percentage of Individual Tax Returns audited in Fiscal 1974, by state

   a. 4 regions in the U.S. (8)

   b. Parallel stem-and-leaf shows slight difference in spread

   c. Log(median) vs. log(midspread) plot indicates $r = -5\ 2/3$, a strange transformation

   d. Best left in original unit

6. Another example: Percentage of population illiterate in 1960 by state

   a. Same 4 regions used (9)

   b. Parallel displays show differences in both location and spread

   c. Plot has slope of -0.60. 1 - (-0.60) = 1.6, about 2. Try squares

   d. Schematic transformed data shows equalization of spread (except for Atlantic) (10)

7. Another example: Percentage of population illiterate in 1900, by state (11)

   a. Batches differ greatly in spread

   b. Unable to determine slope; is it 1 or 3 ?

   c. Try both $-1/(X^2)$, and logs

   d. Negative reciprocals squares fail miserably (12)

   e. Logs quite good (13)

8. Log Median/Log Midspread plots constructed on computer:

   a. Use SUMMARY to obtain Medians and Midspreads

   b. Input these into 2 separate files

   c. Take logs with REEX

   d. Plot with PLOT

260

Lecture 2-2
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 2-2 Outline
<u>Topic 1</u>		
Section I		
2.a	2	Parallel Schematic Plot of Life Expectancies
2.b	3	Parallel Schematic Plot of Infant Mortality
2.c	4	Parallel Schematic Plot of Per capita Incomes
<u>Topic 2</u>		
Section IV		
4.a	5	5-number summaries of Per capita Incomes
4.c	6	Log Median vs Log Midspread and Eyeball Slope for Per capita Incomes
4.h	7	Parallel Schematic Plot of Logs of Per capita Incomes
5.	8	Percentages of Individual Tax Returns Aud[illegible]ted by State in 1974
6.	9	Percent illiterate by State, 1960
6.d	10	Parallel Schematic Plot of Squares of % Illiterate, 1960
7.	11	Percent Illiterate by State, 1900
7.d	12	Parallel Schematic plot of Negative Reciprocals
7.e	13	Parallel schematic plot of LOGS

Lecture 2·2 [1]

Transformations for Stabilization of Spread:

The use of various algebraic transformations to equalize spread among batches.

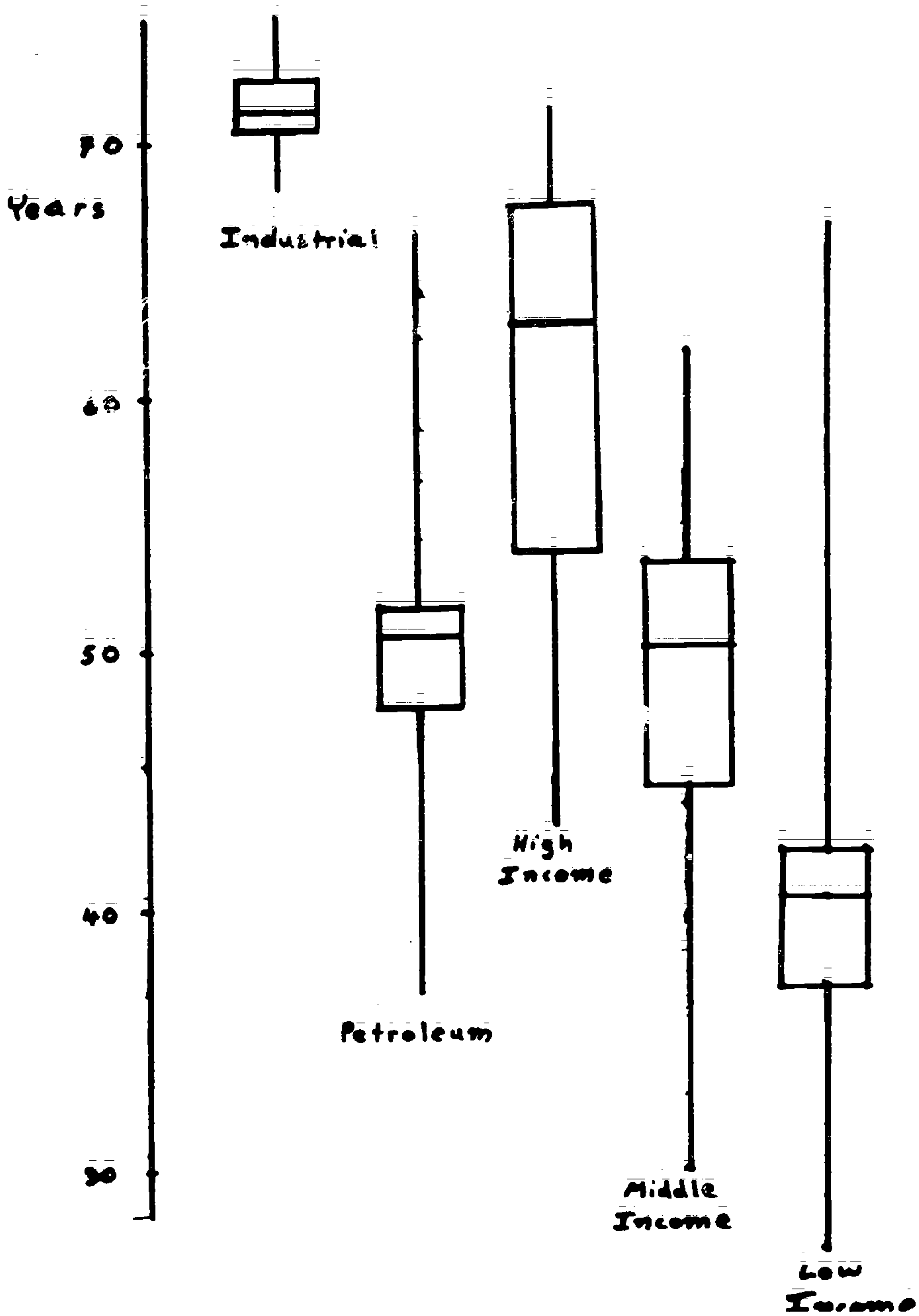
Lecture Content

Discuss need for various transformations and introduce a method of determining a good transformation for a batch.

Main Topics

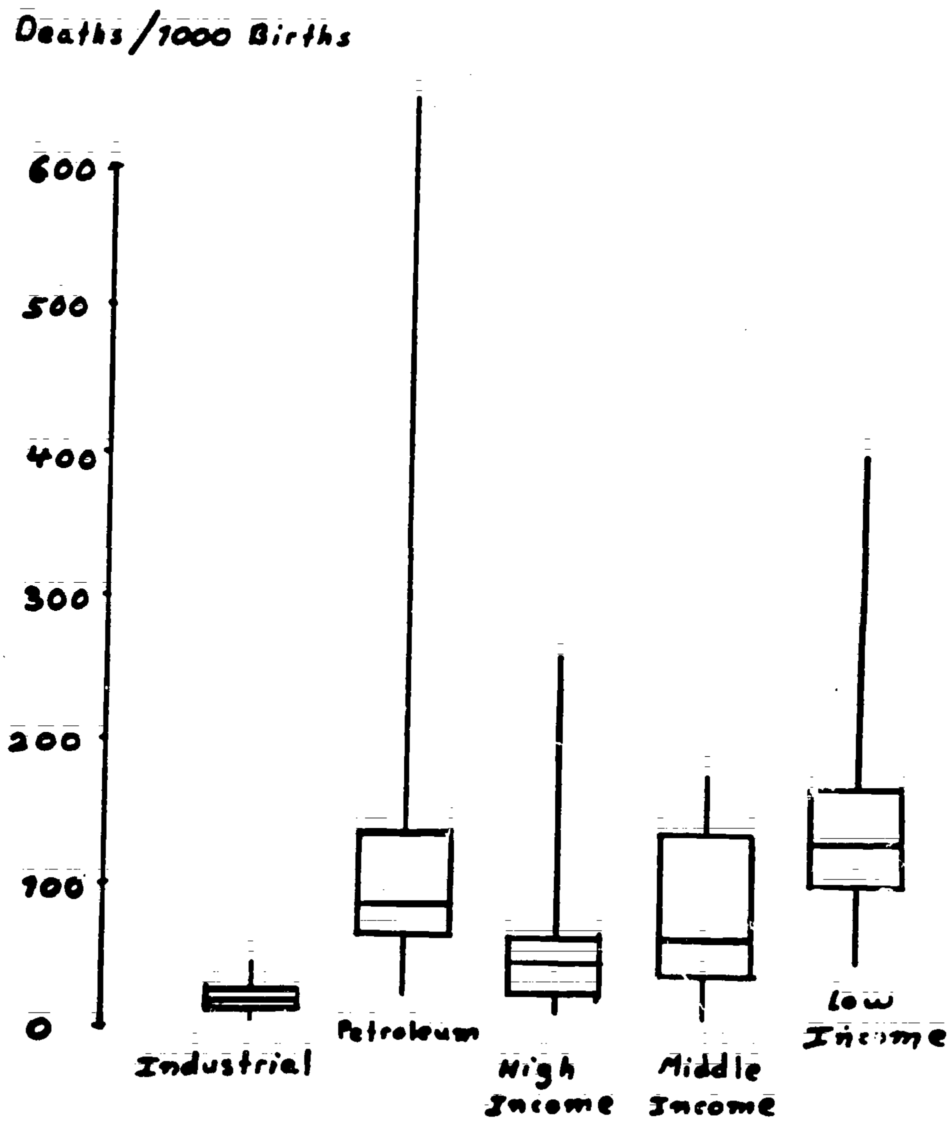
1. Necessity of Transforming a multiple batch.
2. Use of medians and mid-spreads to find transformations.

Life Expectancies by Nation, Classified into 5 Batches. [2]

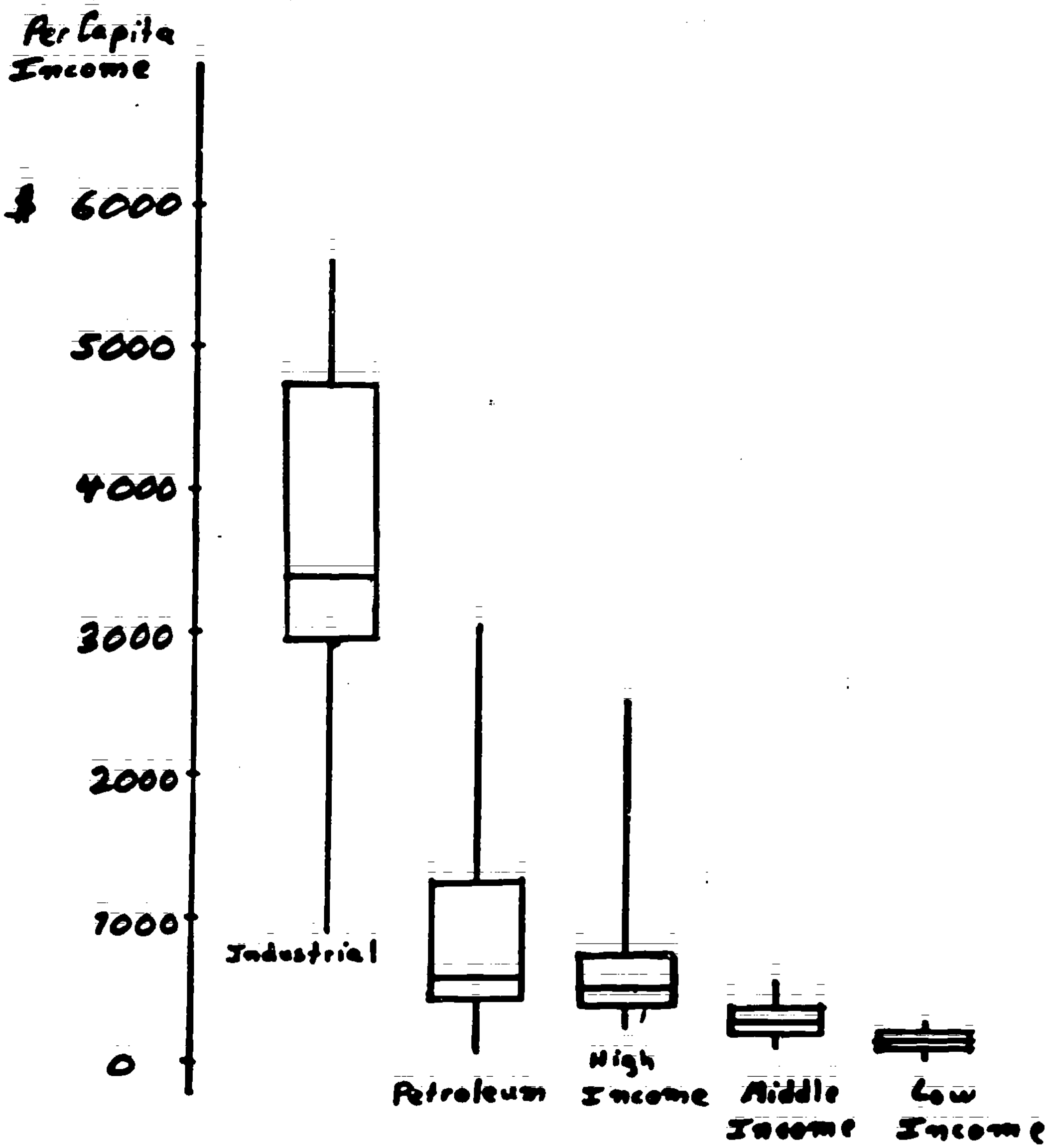


2-2

Infant Mortality by Countries, Classified into Batches. [3]



Per Capita Income for Countries, Classified into Batches [4]

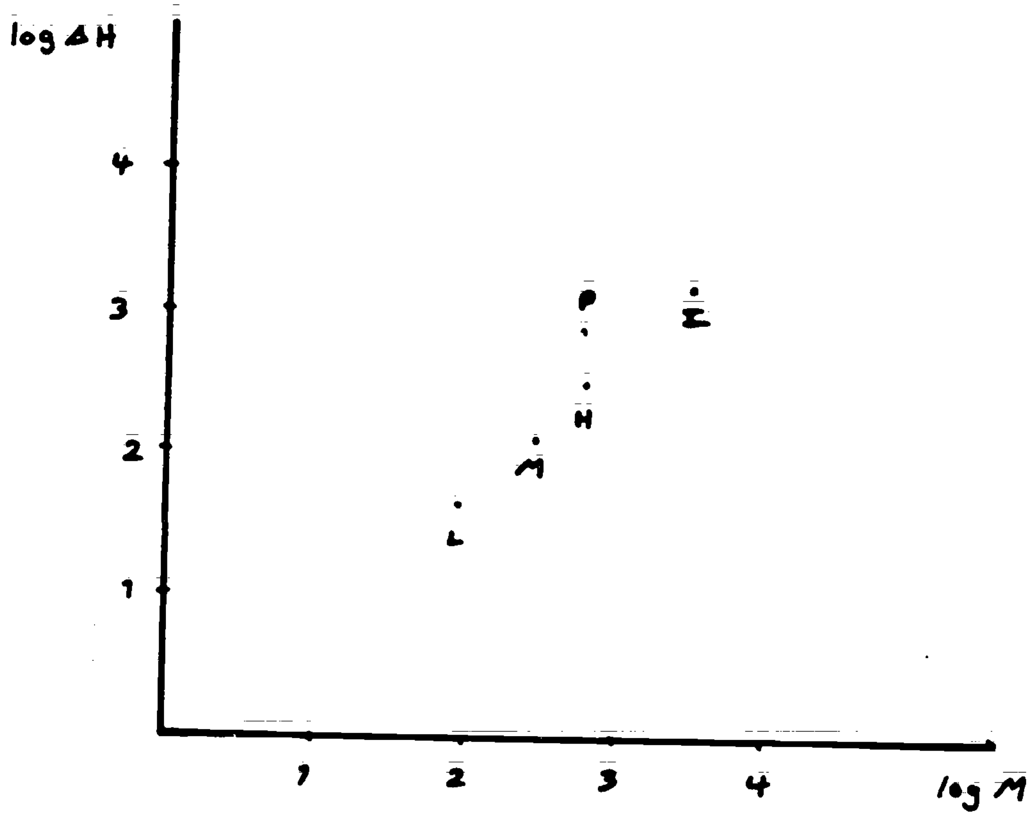


Per-Capita Incomes for Countries [5]

5 Number Summaries

	Industrial	Petroleum	Higher Income	Middle Income	Lower Income
E	956	110	261	130	50
H	2963	360	406	210	77
M	3403	560	610	281	96
H	4751	1280	776	344	125
E	5596	3010	2526	497	186
midspread	1788	920	370	134	48
Log M	3.53	2.75	2.79	2.45	1.98
Log ΔH	3.25	2.96	2.57	2.13	1.68

Log Median vs. Log Midspread for Per Capita Incomes [6]



Representative Points

$L = (1.98, 1.68)$ and $I = (3.53, 3.25)$

$$\text{Slope} = \frac{3.25 - 1.68}{3.53 - 1.98} = 1.01 = p$$

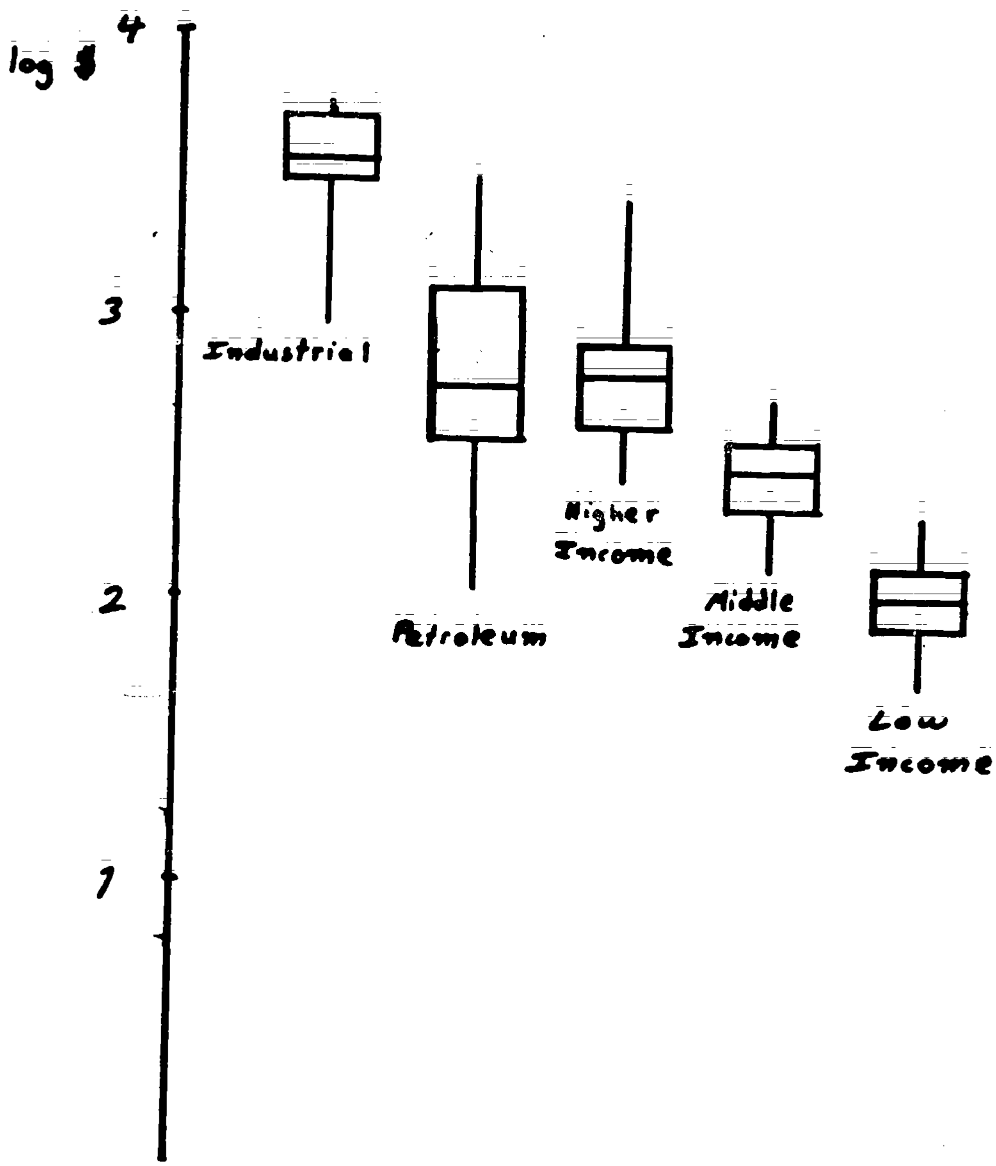
R, exponent for transformation, $= 1 - 1.01 \approx 0$

take logs

2-2

Per Capita Incomes for Countries, Parallel Schematic Plot; Log Scale.

[7]



268

2-2

Percentage of Individual Tax Returns Audited in Fiscal 1974.

[8]

Atlantic (13 states)

New York	3.0 %
Maine	2.1
Massachusetts	1.6
Vermont	2.1
Connecticut	1.8
New Hampshire	2.2
Rhode Island	1.8
Maryland & D.C.	2.1
New Jersey	2.1
Pennsylvania	1.6
Virginia	1.9
Delaware	2.2

South East & Central (12)

Gorgia	2.3 %
Alabama	2.3
South Carolina	2.3
North Carolina	2.7
Mississippi	2.8
Florida	2.7
Tennessee	1.7
Ohio	1.4
Michigan	2.0
Indiana	1.2
Kentucky	1.4
West Virginia	1.3

Mid-and Southwest (16)

South Dakota	1.5 %
North Dakota	1.8
Illinois	2.0
Iowa	1.3
Wisconsin	2.7
Nebraska	2.3
Missouri	2.1
Minnesota	1.4
New Mexico	2.1
Wyoming	1.8
Colorado	1.9
Texas	3.1
Arkansas	2.2
Louisiana	2.6
Oklahoma	2.3
Kansas	2.5

Western (10)

Alaska	2.7 %
Idaho	2.0
Montana	2.7
Hawaii	1.9
California	2.5
Arizona	1.9
Oregon	1.5
Nevada	3.4
Utah	2.2
Washington	2.0

2-2

Illiteracy of the Population, by State, 1960: Percentages. [9]

Atlantic

New York	2.9 %
Maine	1.3
Massachusetts	2.2
Vermont	1.1
Connecticut	2.2
New Hampshire	1.4
Rhode Island	2.4
Maryland & D.C.	1.9
New Jersey	2.2
Pennsylvania	2.0
Virginia	3.4
Delaware	1.9

Southeast & Central

Georgia	4.5
Alabama	4.2
South Carolina	5.5
North Carolina	4.0
Mississippi	4.9
Florida	2.6
Tennessee	3.5
Ohio	1.5
Michigan	1.6
Indiana	1.2
Kentucky	3.3
West Virginia	2.7

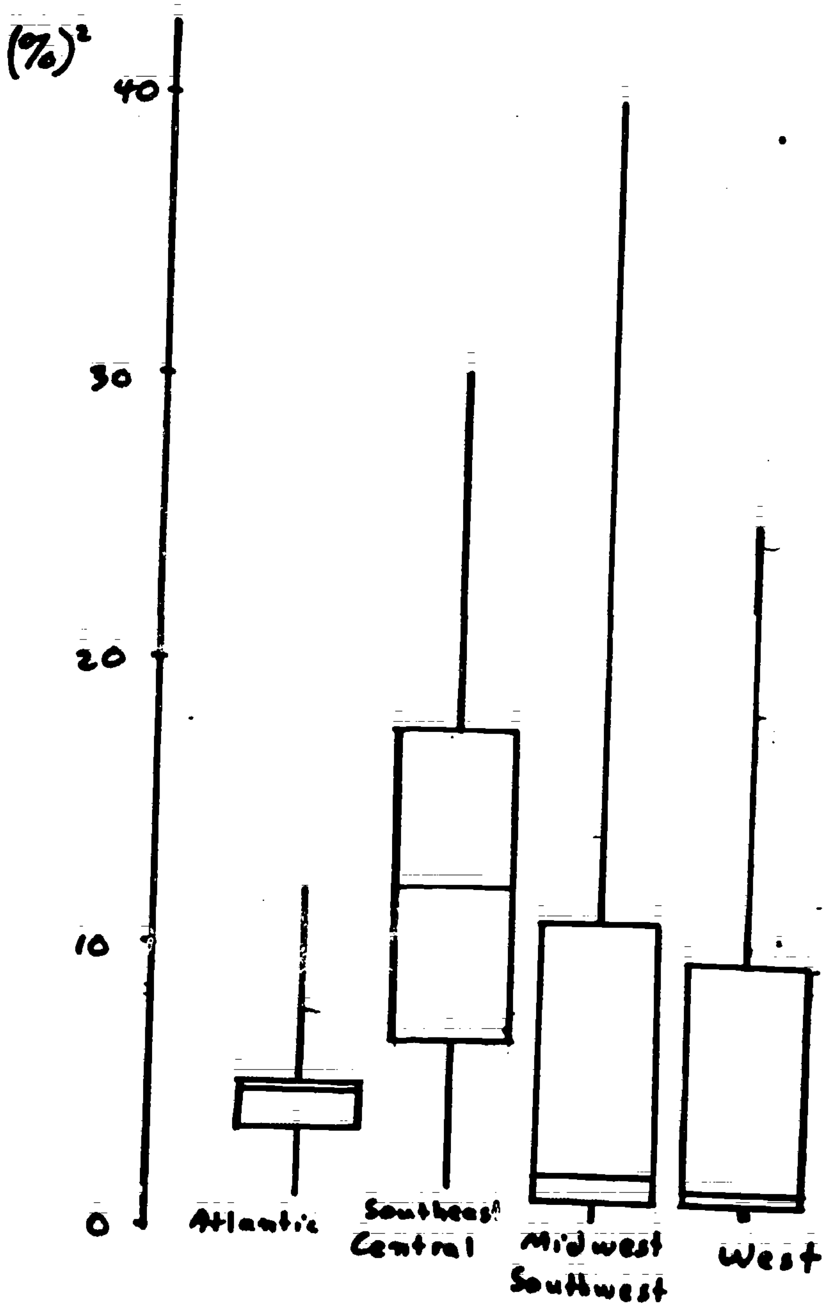
Midwest & Southwest

South Dakota	0.9
North Dakota	1.4
Illinois	1.8
Iowa	0.7
Wisconsin	1.2
Nebraska	0.9
Missouri	1.7
Minnesota	1.0
New Mexico	4.0
Wyoming	0.9
Colorado	1.3
Texas	4.1
Arkansas	3.6
Louisiana	6.3
Oklahoma	1.9
Kansas	0.9

Western

Alaska	3.0
Idaho	0.8
Montana	1.0
Hawaii	5.0
California	1.8
Arizona	3.8
Oregon	0.8
Nevada	1.1
Utah	0.9
Washington	0.9

Parallel Schematic Plots for Squares of Percent Illiterate Data [10]



2-2

Illiteracy of the Population, Percentages by State 1900.

[11]

Atlantic

New York	6.1 %
Maine	5.5
Massachusetts	6.5
Vermont	6.4
Connecticut	6.5
New Hampshire	6.7
Rhode Island	9.2
Maryland & D.C.	11.5
New Jersey	6.5
Pennsylvania	6.9
Virginia	24.3
Delaware	13.2

Southeast & Central

Georgia	32.1
Alabama	35.1
South Carolina	37.4
North Carolina	30.1
Mississippi	34.1
Florida	23.4
Tennessee	21.9
Ohio	4.5
Michigan	4.8
Indiana	5.2
Kentucky	187
West Virginia	12.6

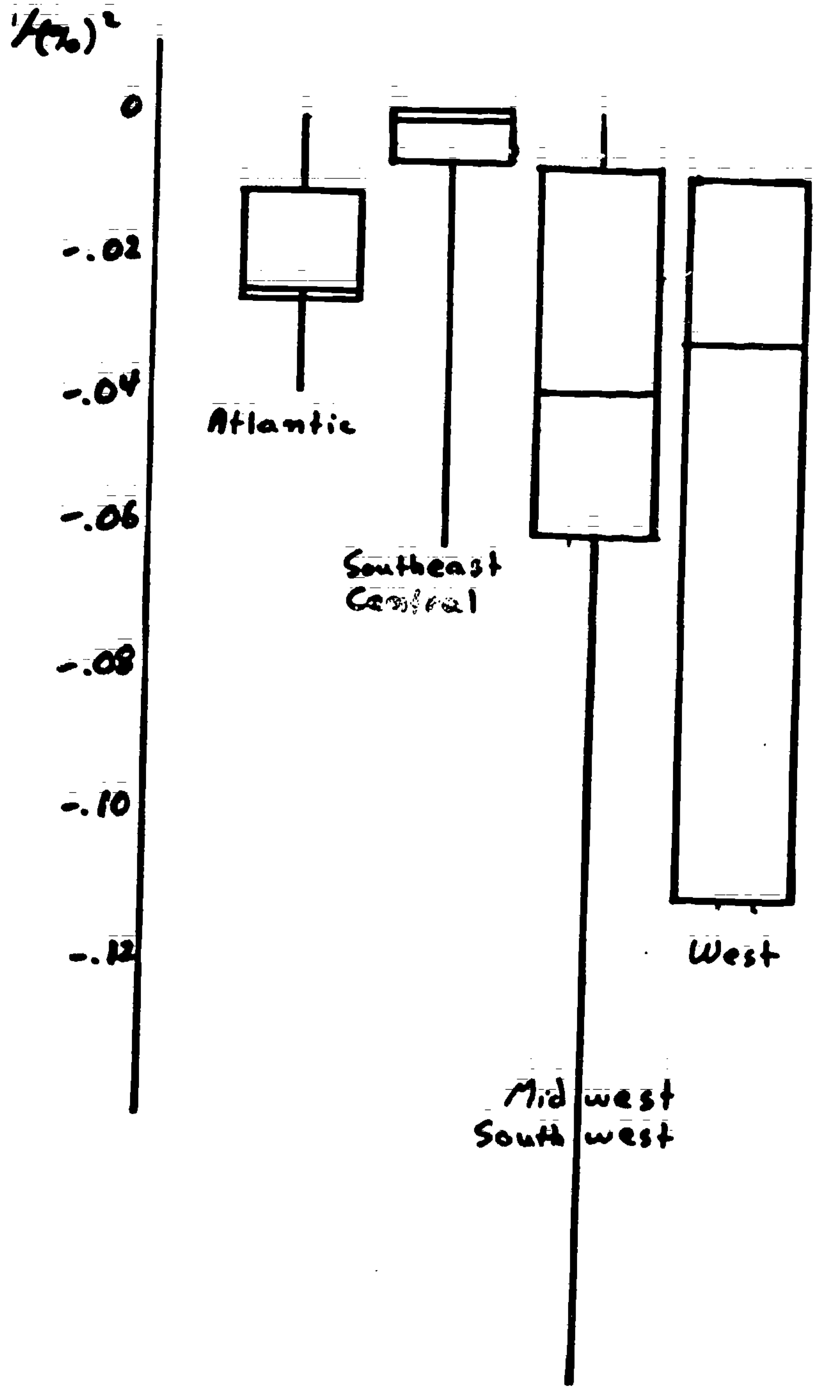
Midwest & Southwest

South Dakota	5.8
North Dakota	6.1
Illinois	4.8
Iowa	2.7
Wisconsin	5.4
Nebraska	2.6
Missouri	7.0
Minnesota	4.6
New Mexico	35.7
Wyoming	4.4
Colorado	4.5
Texas	15.6
Arkansas	21.3
Louisiana	39.6
Oklahoma	11.7
Kansas	3.3

West

Alaska	40.6
Idaho	5.1
Montana	6.6
Hawaii	35.2
California	5.3
Arizona	30.0
Oregon	3.7
Nevada	13.8
Utah	3.6
Washington	3.4

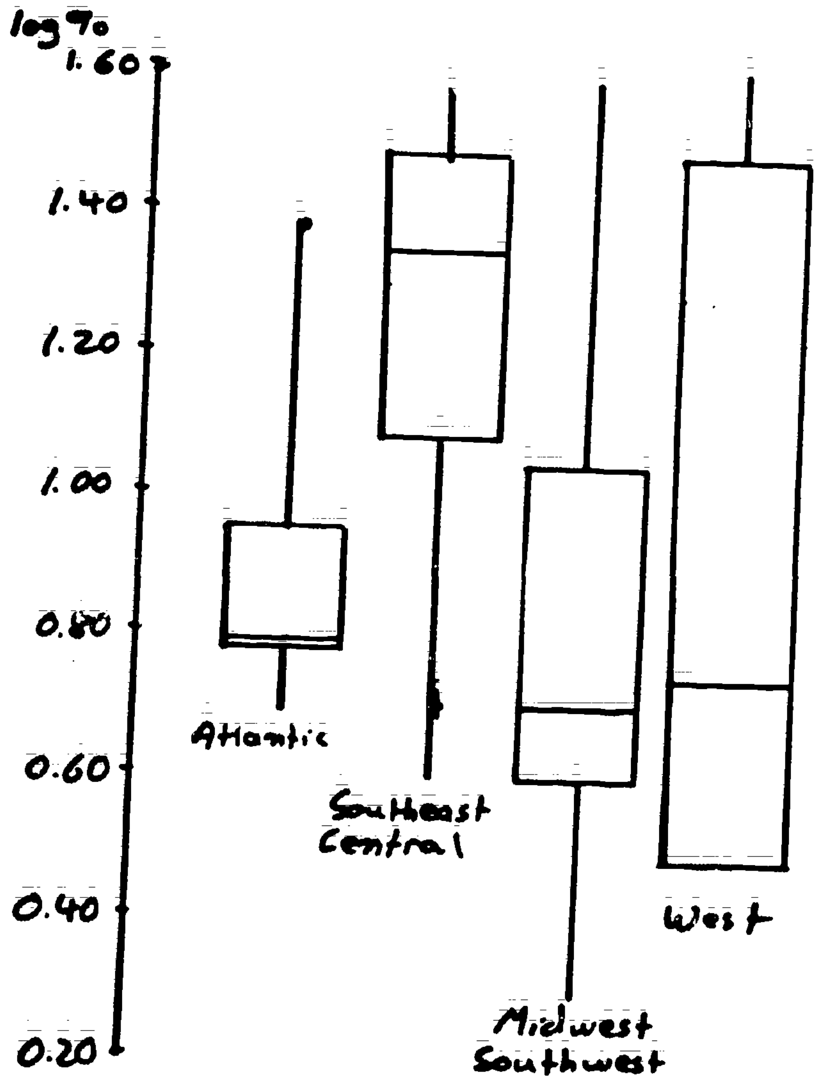
Negative Reciprocal Squares Transformation for 1900 Percent Illiteracy. [12]



2-2

Log Transformation for 1900 Percent Illiteracy

[13]



274

2-2

Homework
Unit 2

1. These data are from a study designed to determine whether varying the report of the results of a controversial psychological study can influence judgements about the ethics of the research.

Three groups of subjects (high school teachers) read summaries of the Milgram (1963) obedience study [Milgram, S. "Behavioral Study of Obedience", Journal of Abnormal and Social Psychology, Vol. 7, 371-378]. These summaries were identical except for the reporting of the results. One group of teachers read the actual results of the Milgram study (Actual Results group). One group read that nearly all of Milgram's subjects delivered the highest shock available to the confederate (Many Comply group). A third group read that nearly all of Milgram's subjects refused to deliver the highest shock to the confederate (Many Refuse group).

After reading the report, the teachers answered a number of questions. Among these questions, there was a seven point scale on which the teachers were asked to rate the ethics of the study. (The higher the rating, the more ethical the study was believed to be).

Compare the three groups and summarize the differences among them.

THE DATA:

Actual Results:	6, 1, 7, 2, 7, 1, 7, 3, 4, 1, 1, 1, 6
Many Comply:	3, 1, 3, 7, 6, 7, 4, 3, 1, 1, 2, 5, 5
Many Refuse:	5, 7, 7, 7, 6, 6, 6, 6, 7, 2, 6, 3, 6

2. An experiment on nipples for baby bottles compared different nipple designs--the conventional one having a medium circular hole and a new one having a terminal slot .11 inches long. A special bottle permitted an unrestricted flow of milk, and the new nipple was positioned horizontally and vertically to determine the effect of orientation. For 24 babies, the volume (in milliliters per suck) was as follows:

Medium Hole	Slot Vertical	Slot Horizontal
0.81	1.33	0.92
0.50	2.10	0.78
0.78	1.50	1.20
0.43	1.60	1.00
0.50	1.70	0.65
0.71	2.00	NA
0.71	1.21	0.80
0.34	1.35	0.55

(a) Compare these batches with parallel schematic plots.

(b) Transform the batches to stabilize the spread of the values.

3. Suppose we collect measurements of FEV (Forced Expiratory Volume) from individuals that work at the same factory and are of the same age, sex, and height. FEV is a measure of pulmonary function. We subdivide these individuals by smoking status into the groups: A=never a smoker; B= exsmoker; c= present smoker, currently smoking less than two packs per day, D= present smoker, currently smoking at least two packs per day.

The data are as follows:

A: 260, 275, 260, 290
B: 232, 230, 246, 245
C: 224, 202, 262, 225
D: 180, 195, 202, 175

Compare the Groups.

4. When the trial of Dr. Benjamin Spock and his associates began in 1968, the defense challenged the list of prospective jurors because only 9% were women. A more detailed examination of jury venires in the U. S. District Court for the District of Massachusetts revealed that in venires summoned for trials before the six colleagues or the trial judge between 4 April, 1966 and 22, October 1968, the percentages of women were:

Judge A: 40, 30, 16, 35, 50

Judge B: 36, 32, 32, 27, 29, 45

Judge C: 34, 30, 32, 29, 24, 28, 20, 35

Judge D: 24, 30

Judge E: 33, 36, 28, 20, 18, 22, 40

Judge F: 22, 21, 31, 27, 17, 29, 26, 29, 34

While those for the trial judge were:

Trial Judge: 16, 18, 14, 6, 18, 15, 9, 24

(a) Compare these batches of percentages both numerically and graphically.

(b) Combine Judges A-F into one batch and compare the trial judge with it.

(c) Which comparison is most effective ? Why ?

5. Four groups of students were subjected to different teaching techniques and tested at the end of a specified period of time. Their scores are shown below:

Techniques

1	2	3	4
65	75	59	94
87	69	78	89
73	83	67	80
79	81	62	88

Compare batches (transformation is unneccessary) to determine whether the teaching techniques differ.

6. The stem-and-leaf display below gives the percentage of families in each Manhattan police precinct where combined income in 1970 was less than $4,000. (Data from New York Times, March 30, 1973)

   (a) Write down the five number summary for these data, and calculate

   $$\hat{S} = 3/4 * \text{Midspread}$$

   (b) What evidence (if any) is there in your answer to (a) that this batch could be made more symmetric by transformation.

   (c) The lower hinge, median, and upper hinge for precincts in the Bronx, Brooklyn, Queens, and Staten Island are given below. Combine these data with that from Manhattan to find a transformation that would equalize the variability in the five batches.

Percent Families with Income < $4,000

Manhattan	0	5 6 7 9
Unit =	1	0 2 2 4 5 5 6 7 9
10%	2	0 3 6 7 7
	3	1 5

	Number of Precincts	Lower Hinge	Median	Upper Hinge
Bronx	11	11-1/2	17	30
Brooklyn	23	12-1/2	19	24-1/2
Queens	14	6	9	11
Staten Island	3	5	6	8

Homework Unit 2
Solutions

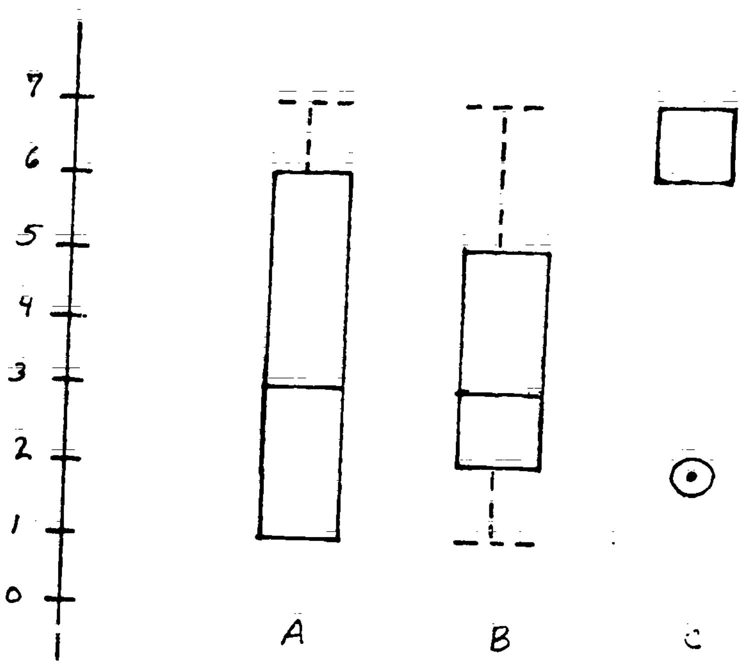
1. Letter Value Displays

#13	Actual (A)		
M7			
H4	1	6	5
E	1	7	

#13	Many Comply (B)		
M7	3		
H4	2	5	3
E	1	7	

#13	Many Refuse (C)	
M7	6	
H4	6	7
E	2	7

Schematic Plots



For the group that read the actual results, there were a wide range of opinions; some thought it was ethical, and others thought that it was not.

For those who were told that many complied, opinions still were split, but more people rated the experiment with middle values (3's, 4's, and 5's) indicating that they questioned or were uncertain about the ethics of the experiment.

The most interesting result was for the group that was told that most people refused to administer the shock. Almost all of this group felt that the experiment was very ethical.

Thus, as long as participants refuse to administer a shock, the teachers felt the experiment was ethical; but when some were told that shocks were administered, they began to question and disapprove of the experiment.

2. Letter Value Displays

Medium Hole

#8			
M4h	.61		
H2h	.47	.75	.28
E	.34	.81	

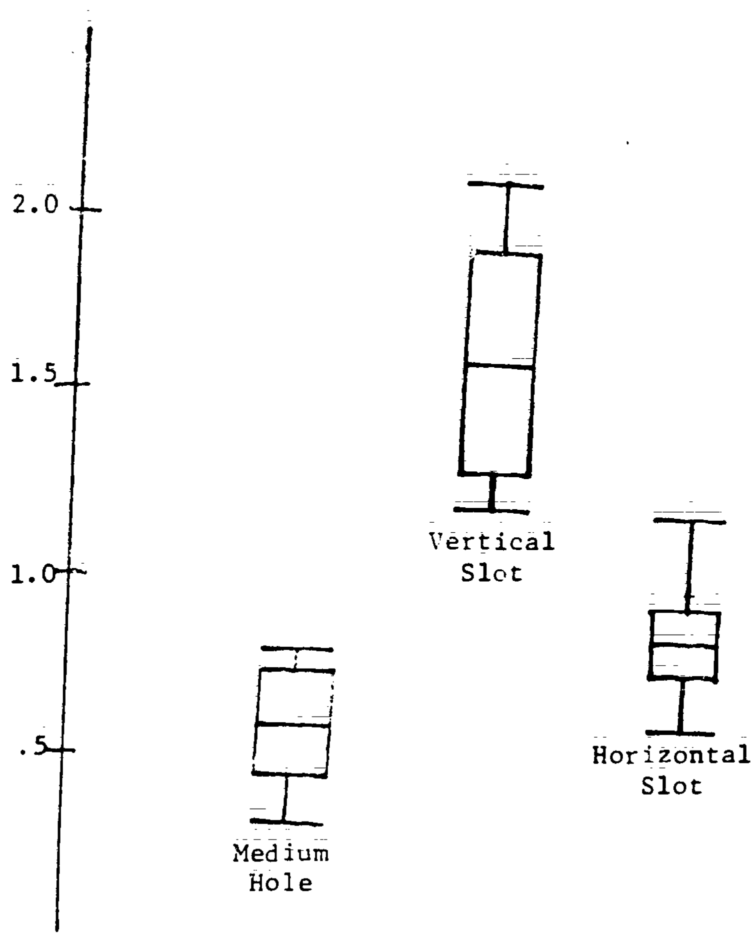
Vertical Slot

#8			
M4h	1.6		
H2h	1.3	1.9	.6
E	1.2	2.1	

Horizontal Slot

#7			
M4	.88		
H2h	.79	.96	.17
E	.65	1.2	

A) Parallel Schematic Plots



B. Mid Summaries

Median	.61	1.6	.88
Midhinge	.61	1.6	.875
Midextreme	.57	1.65	.925

From the parallel schematic plots and the midsummaries we can see that no transformation is called for.

Clearly, the nipples with a terminal slot are better, with a vertical orientation being the best.

etter Value Displays

A

#4			
M2h	268		
H1h	260	275	
E	2[illegible]0	290	

B

#4			
M2h	239		
H1h	231	246	15
E	230	246	

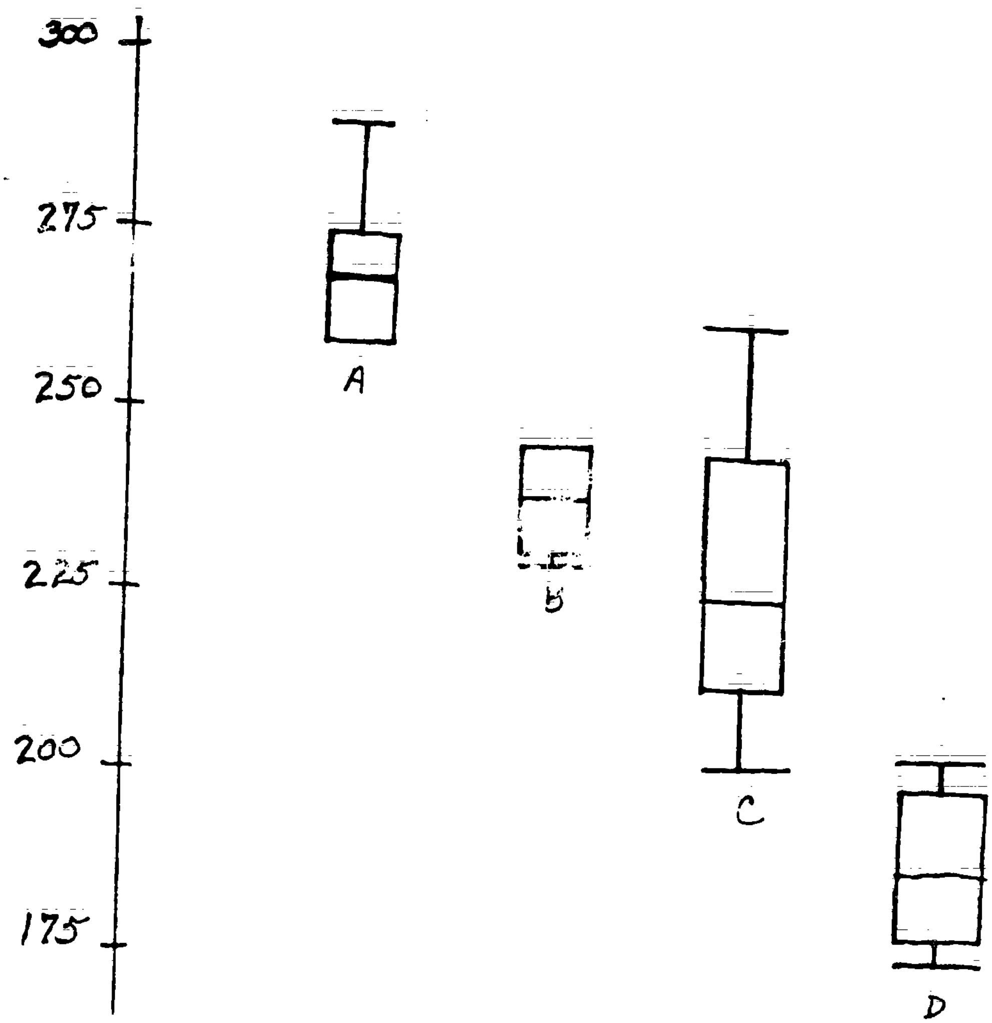
C

#4			
M2h	225		
H1h	213	244	31
E	202	262	

D

#4			
M2h	188		
H1h	178	199	21
E	175	202	

Parallel Schematic Plots



There appears to be a definite difference between the non-smokers, ex-smokers, less than two pack smokers, and at least two pack smokers. The more one smokes, the lower the F.E.V. However, before you could say much about how of a difference there is, more data should be obtained. Four observations is not enough.

4. Sort

A	B	C	D	E	F	T
16	27	20	24	18	17	6
30	29	24	30	20	21	9
35	32	28		22	22	14
40	32	29		28	26	15
50	36	30		33	27	16
	45	32		36	29	18
		34		40	29	18
		35			31	24
					34	

Numerical comparison

		min	max	range
A:	median = 35	16	50	34
B:	median = 32	27	45	18
C:	median = 29.5	20	35	15
D:	median = 27	24	30	6
E:	median = 28	18	40	22
F:	median = 27	17	34	17
T:	median = 15.5	6	24	18

Stem-and-Leaf Displays Judges A-F combined

1*	
1·	678
2*	0012244
·	6778899
3*	0001222344
·	5566
4*	00
·	5
5*	0

unit = 1

N = 37

M	18		29	
H	9	24		34
E	1	16		50

outside value - 50

Stem-and-leaf   all judges A-F&T

```
0* |
 · | 69
1* |
1· | 566788
2* | 00122444
2· | 677889999
3* | 0001222344
3· | 5566
4* | 00
4· | 5
5* | 0
```

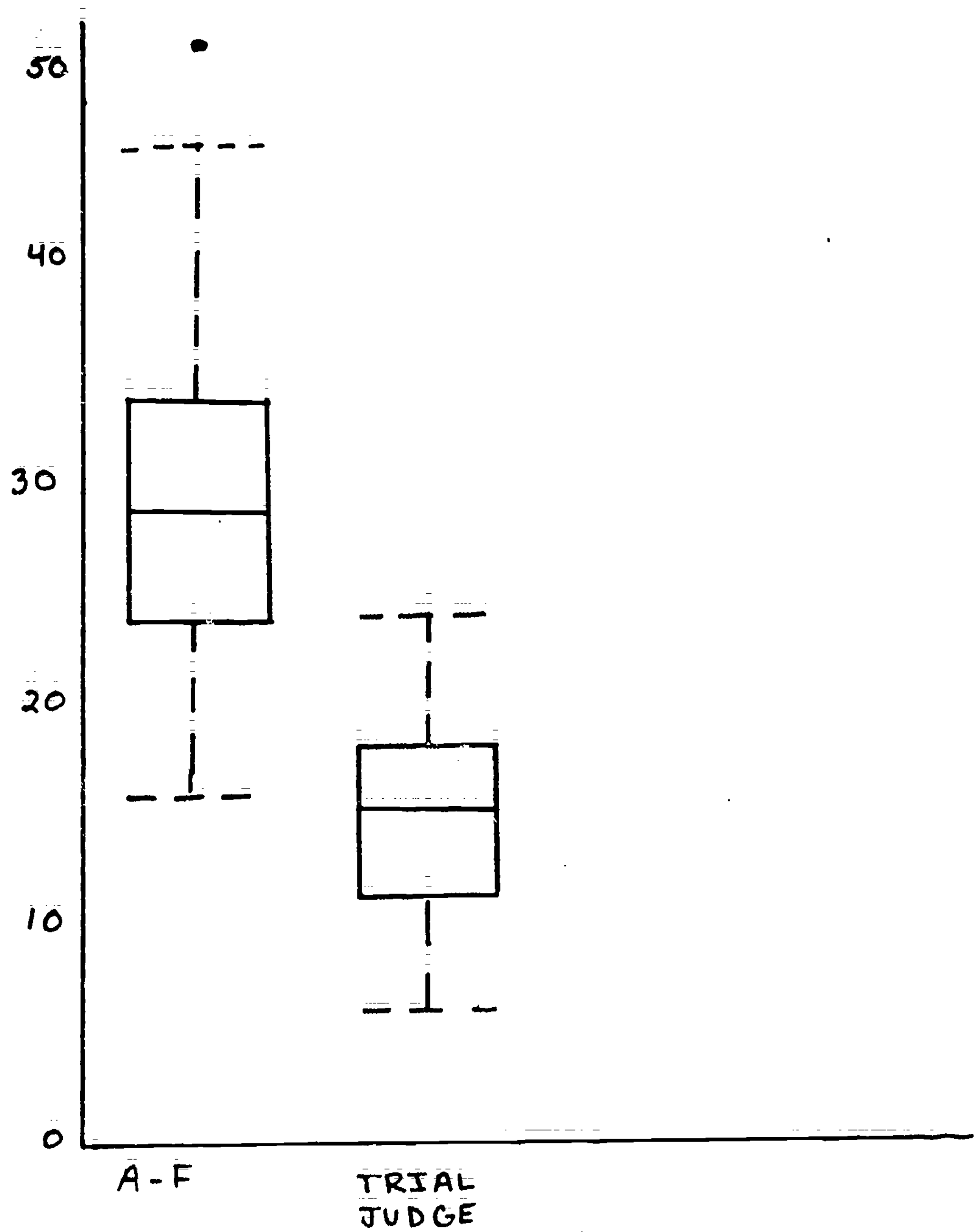
n = 45

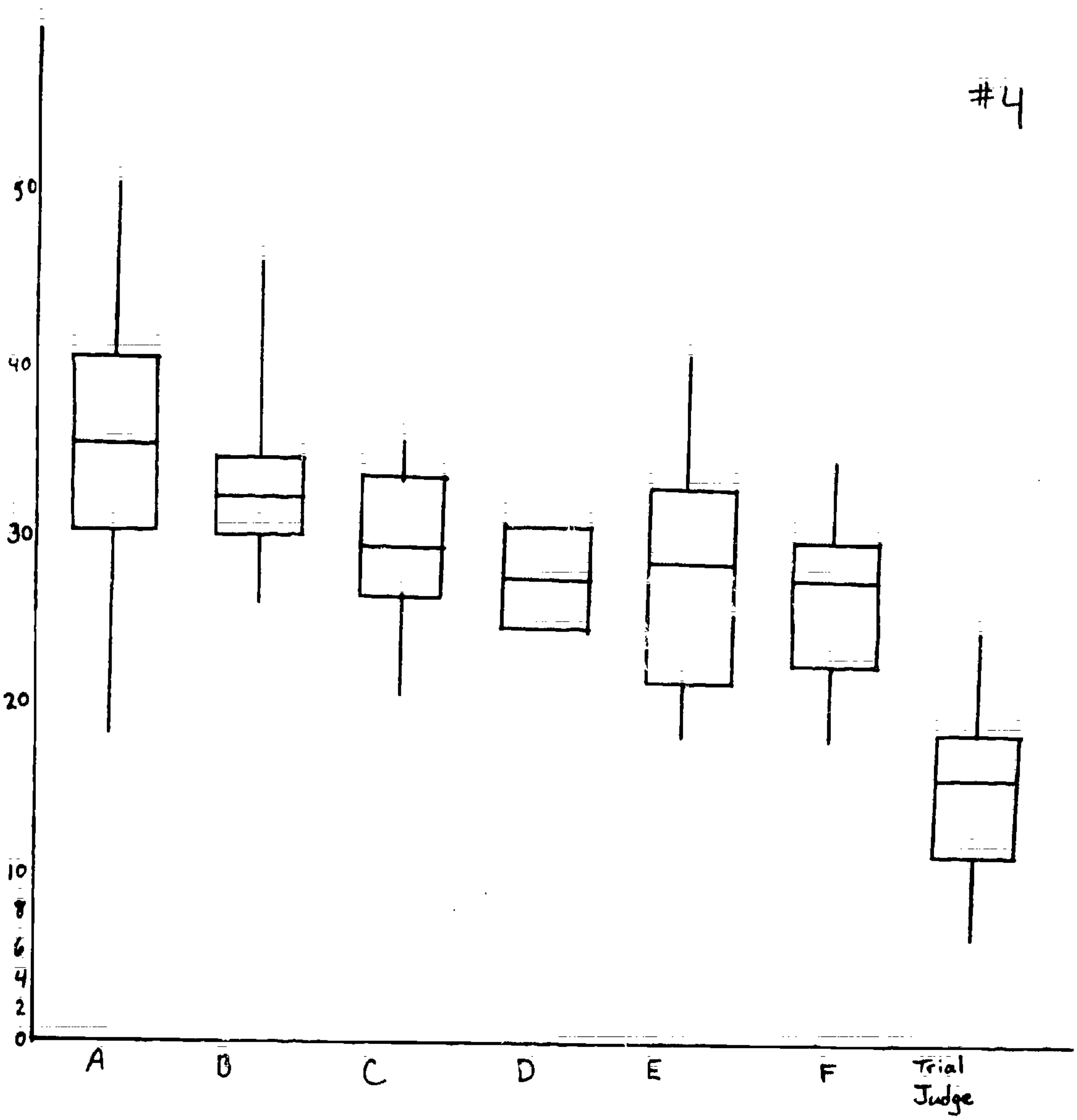
```
median 23  |        28        |
hinge 12   |  20           32 |
extreme 1  |   6           50 |
```

While the individual comparison of each judge with the trial judge shows that the trial judge's typical percentage of women was lower than the other judges, I think the larger group comparison is more valid. This is because of the total numbers involved.

The comparison of the two boxplots clearly shows that about 75% of the trial judges venire's had a lower percentage of women than the combined group of judges. It also shows that even when the greatest percengage of women were in the trial judge's venire, 75 percent of the combined grouping had more women.

Both comparisons raise questions concerning how juries are chosen, since mose of the venires had less than 40% women.


50
40
30
20
10
0
A-F
TRIAL
JUDGE

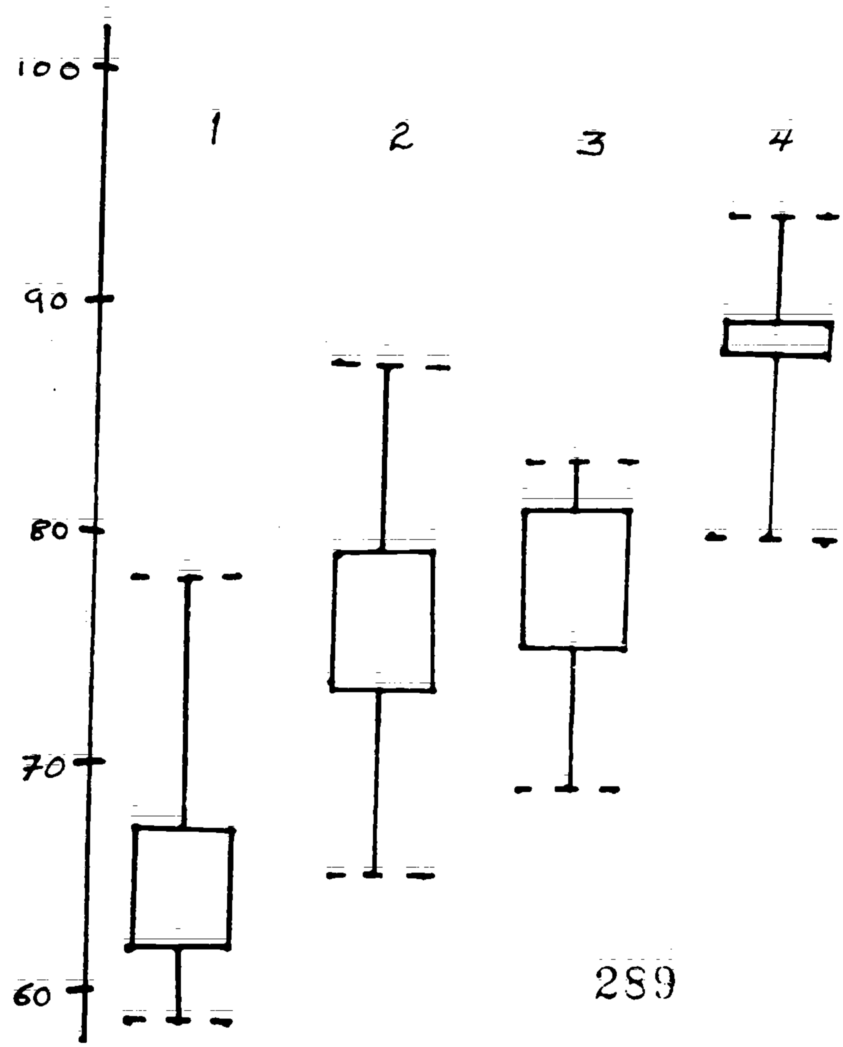


#4
50
40
30
20
10
8
6
4
2
0
A
B
C
D
E
F
Trial Judge

5.

1	2	3	4
65	69	59	80
73	75	62	88
79	81	67	89
87	83	78	94

	median	min	max	range
1	76	65	87	22
2	78	69	83	14
3	64.5	59	78	19
4	88.5	80	94	14



6. n = 20

(a)

M 10h		15.5	
H 6h	11		24.5
E 1	5		35

$\hat{S}$ = 3/4 (24.5-11) = 3/4 (13.5) = 10.125

	Bronx	Brooklyn	Queens	Staten Is.	Manhattan
UH	30	24h	11	8	24.5
Median	17	19	9	6	15.5
LH	11h	12h	6	5	11
Midspread	18.5	12	5	3	13.5
Midhinge	20.75	18.5	8.5	6.5	17.75

(b) midhinge ≠ midextremes ≠ median ≠ mean

17.75 ≠ 20 ≠ 15.5 ≠ 17.3

batch doesn't trail off at both extremes, so it might be made more symmetric--however the ratio of maximum to minimum value is less than 20 which would seem to indicate that transformation might not help. Also, though there were differences in the different measures of typical value, they are not very large differences.

(c) Transformation to equalize variability in the batches--negative reciprocal square root of x_i.

Bronx	Brooklyn	Queens	Staten Is.	Manhattan	
-.18	-.20	-.30	-.35	-.20	UH
-.24	-.23	-.33	-.41	-.25	Median
-.29	-.28	-.41	-.45	-.30	LH
.11	.08	.11	.10	.10	Midspread

Quiz, Unit 2

WRITE ALL ANSWERS ON A CLEAN SHEET OF PAPER

Part I. Answer the following questions briefly and generally.

1. What is an ordered multiple batch ?

2. How do we best compare a collection of related single batches ?

3. Why would we consider a transformation of a multiple batch ?

4. How do we determine the "best" transformation for a multiple batch ?

5. If a multiple batch consisted of 2 well-behaved batches, and it was determined that a transformation was necessary, what statistics of the batches would we use to find the "best" transformation ?

Part II.

1. Given below is a data set of median annual incomes of individuals with doctorates employed in education (academia), government, and industry in 1964.

	Area of Employment		
Area of Doctorate	Education	Government	Industry
Agriculture	$11,100	$11,500	$12,000
Biology	10,500	11,900	14,000
Earth Sciences	9,900	11,700	13,500
Mathematics	10,300	15,100	17,000
Chemistry	10,000	12,700	14,000
Physics	11,000	13,800	16,000
Psychology	10,000	11,500	15,900

Using the information on the batches given below, determine the best transformation for the batches. You need not carry out the transformation.

	Education	Government	Industry
E	$ 9,900	$11,500	$12,000
H	10,000	11,600	13,750
M	10,300	11,900	14,000
H	10,750	13,200	15,950
E	11,100	15,100	17,000
midspread	750	1,600	2,200
log H	4.00	4.06	4.14
log M	4.01	4.08	4.15
log H	4.03	4.12	4.20
log midspread	2.88	3.20	3.34

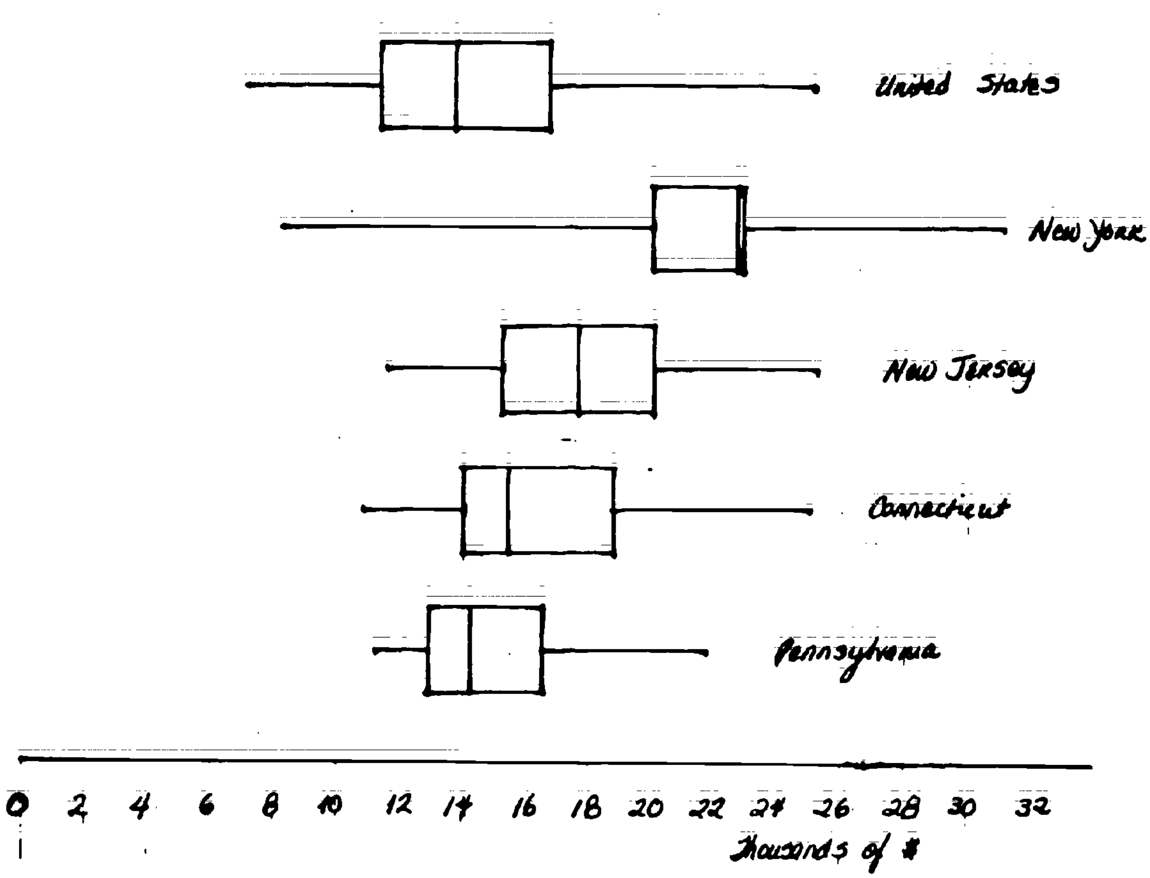
2. On the next page is a detail, or small section of, a display given in the book Profiles in School Support, 1969-1970.

   A. Briefly discuss the "analytic" features of the display: what kind of display is it, what do the various lines of each box mean, etc., as explained in the aforementioned book.

   B. Compare the 4 states among themselves.

   C. Compare each state separately with the United States.

Expenditures per Classroom unit, 1969-70.



XVI.I.238

Unit 2 Quiz
Solutions

PART ONE

1. An unordered multiple batch is a set of batches which have each been collected in a consistent manner, containing similar values having a non-quantitative relationship to one another.

2. We can best compare related single batches through the use of parallel stem-and-leaf diagrams and parallel schematic plots. We can also use the five number summaries. But we must be cautious to control spread via a transformation if necessary.

3. Transformation in a multiple batch is used to equalize the spread and remove a possible consistent relationship between spread and typical value in the batches.

4. The method for finding the "best" transformation consists of taking the logarithm of the median and midspread (sometimes the Δ between the extremes) of each batch and then plotting these as points in an x-y (logM, Log$_\Delta$H) plane. A line is drawn to approximately fit these points with slope = p. The best transformation for the data will be found by subtracting p from one and raising (or lowering) the original data by a power equal to that difference.

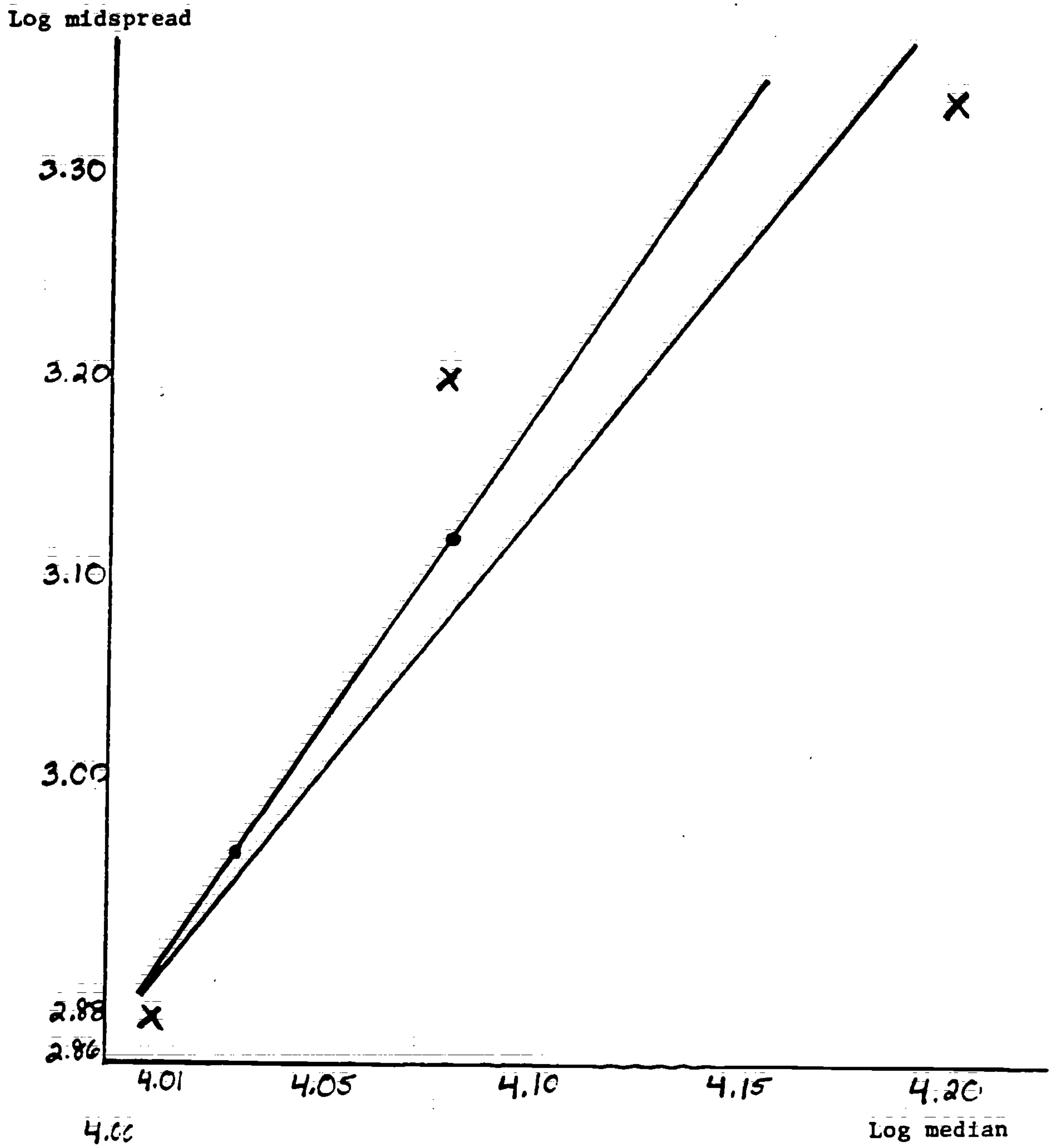
   i.e. $X \rightarrow X^R$ where $R = 1 - p$

5. The mean and the standard deviation

PART TWO

1. The slope of possible lines to fit those points vary from -2 to -3. This indicates that transformations could range on the ladder of powers from R = -1 to R = -2 (negative reciprocals or negative reciprocals of the square root.) Take $(x_1,y_1) = (4.01, 2.88)$; $(x_2,y_2) = (4.15, 3.34)$

$$\frac{y_2-y_1}{x_2-x_1} = \frac{3.34-2.88}{4.15-4.01} = \frac{.46}{.41} = 3.29 \qquad r = 1-3 = -2$$



slope = $\frac{3.12-2.96}{4.08-4.03} = \frac{.16}{.05} = 3.2$ approximately 3

R = 1-p = 1-3 = -2

transformation: negative reciprocal of the square = $\frac{1}{x^2}$

2. A. These styles represent boxplots, but they are not exactly as we have defined them. The solid line extends from the 2nd percentile to the 98th percentile. The outer edges of the boxes represent the level of expenditures for the 25th and 75th percentiles. The middle line in the box is the level of expenditures at the 50th percentile.

Percentiles involve dividing a distribution into 100 sections with an equal number of observations in each section. Therefore, the 50th percentile is very similar to a median in that half of the observations are found on either side of it So the internal box is like a boxplot but the whiskers are not.

B. No consistent increase in spread with typical value. Distributions quite asymmetric. Nonetheless, seems to be a clear trend of increasing typical value.

C. New York, New Jersey and Connecticut all have greater expenditures in 75% of their classrooms than the national 50th percentile expenditures Pennsylvania expends more in slightly over 1/2 of that state's classrooms than the 50 percent of the nation as a whole. One rather interesting thing to note is that some percentage ($2<x<25$) of New York's classrooms expend more than 98% of the national number of classrooms.

The distance from the 25th percentile to the 75th percentile is greater for the national figure than for any of the state figures. New York is very different from the Nation. New Jersey and Connecticut are too, with Pennsylvania most similar.

Some Principles of Graphics for Tables and Charts

This brief handout discusses some ideas on the effective use of graphics in technical papers and presentations. Some of these principles are due to Edward Tufte, whose lecture on 23 April 1976, given to the Statistics Department at Harvard University, is the basis for this discussion.

We will discuss the 7 principles:

1) Less is more

2) The 3 purposes of graphics for communications

3) Small multiples are useful

4) Think about page arrangement

5) Integrate text and graphics

6) Three-dimensional graphics are special

7) Graphics should have "rough drafts"

These principles will be introduced by means of various examples of graphics taken from many sources, including The Wall Street Journal, The New Yorker and Scientific American. The principles are partly subjective--what we think constitutes a good graphic may not agree with your conception of a good display. After all, graphics are visual and works of art; there is a subjective aspect to their appreciation. However, we believe that these principles are sound and can turn bad displays into good ones, if they are followed.

Principle 1: Less is More

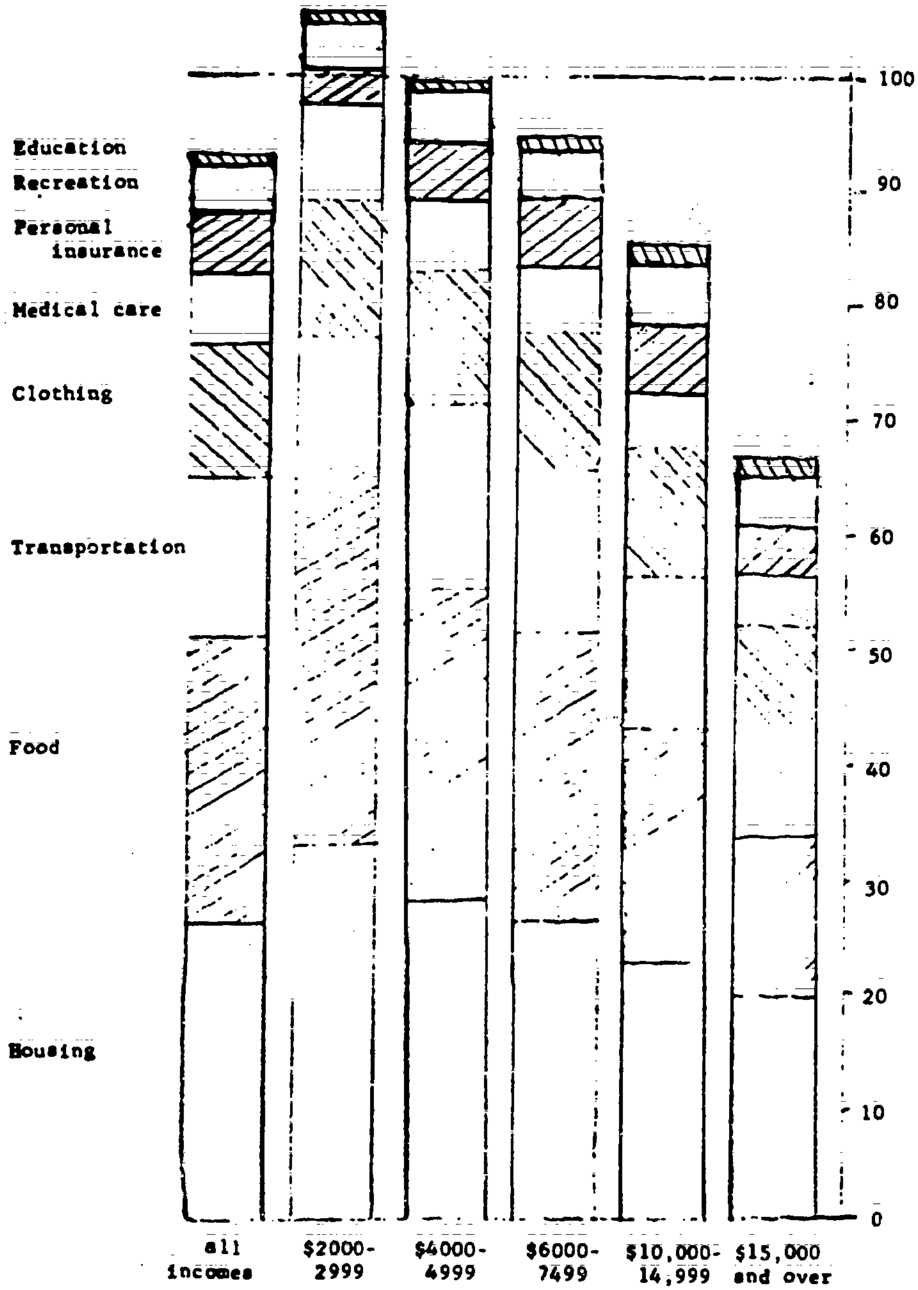
Never try to crowd too much information into a display. Two or three graphics are much easier on the eye than one graphic. If you feel that the display under development contains too much information and might overload your readers' circuits, make two or three displays from the original. Or, if you strive for simplicity, merely take the most important features from the original display and discard the remainder. Remember, graphics must be interpretable by the average fellow. A reader should not spend the majority of his/her time trying to decipher the tables and charts contained within your paper.

Figure 1 is a histogram, where the bars are broken into various components of personal expenditures, by percentages. There is just the right amount of information in this display. Any additional bars, or additional categories of expenditures would make this display uninterpretable. In contrast, we present Figure 2, a bar chart, with the same construction as the histogram in Figure 1. The wild plaids of lines in the bars of the display make it difficult to read. There are too many cities included here. Can you find additional disagreeable features.

Figure 3 is an example of a bar chart, a display similar to a histogram, but with a horizontal axis referring to various characteristics about the data set. The axis does not have a scale as with a histogram. This bar chart is difficult to examine because of the curved bars, although it may be pleasing to the advertising firm that constructed it. The moral of the figure is: Do not try to make your display too ornate if this excessiveness detracts from its comprehension.

We have included 2 other displays from The Wall Street Journal that are quite good. Figures 4 and 5 are both bar charts that are pleasing because of their simplicity, and their effectiveness in conveying their message. Notice, however, that the border around Figure 5 is unnecessary--the arrow is catchy, but the numbers should speak for themselves.

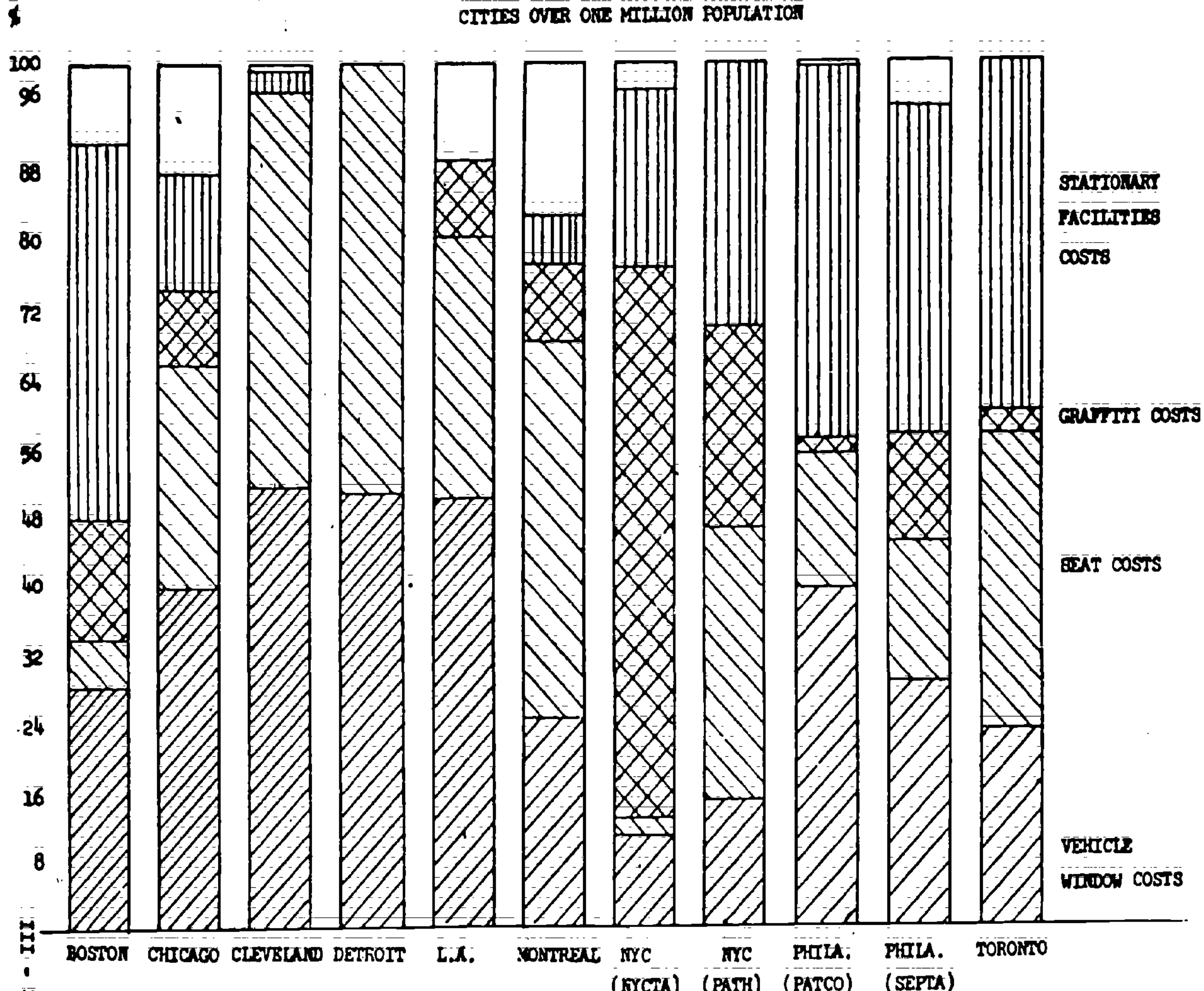
Figure 1
Expenditures as a percent of disposable personal income



300

Figure 2

WINDOW COSTS, SEAT COSTS, GRAFFITI COSTS & STATIONARY FACILITIES COSTS AS A % OF TOTAL VANDALISM COSTS
CITIES OVER ONE MILLION POPULATION



XVI.I.245

Figure 3

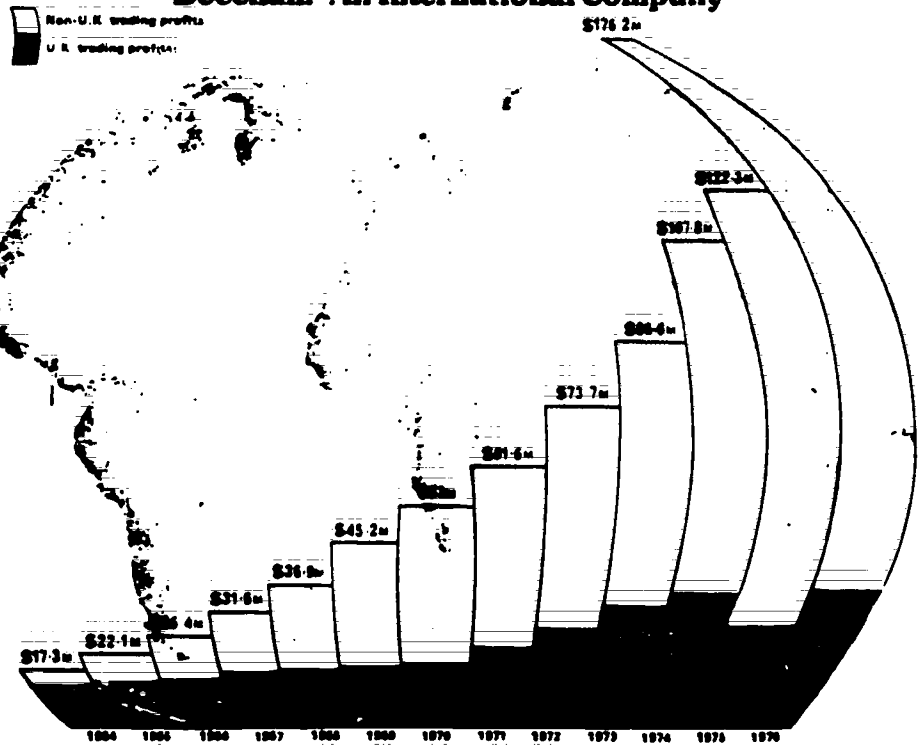
A global view of Beecham helps to account for 13 years of record profits

Beecham is an international company. Not only have its trading profits been increasing continuously for 13 years. Just as important, the number of countries in which these profits are earned has been increasing at the same time. Last year the largest share earned in any one country – which happened to be the U.K. – was only 19.9 per cent of the total.

So what are the highlights of 1975/76?

- World-wide sales $1,011.4 million. Up by $232.4 million or 29.8 per cent. on 1974/75
- Trading profit $196.2 million. Up by $53.9 million, or 44.1 per cent. on 1974/75
- Pre-tax profit $162.8 million. Up by $52.3 million, or 47.3 per cent. on 1974/75

Beecham-An International Company



BEECHAM GROUP LIMITED, BRENTFORD, MIDDLESEX, ENGLAND

Human and veterinary prescription medicines
toiletries cosmetics proprietary medicines, food and drink products, animal health and animal nutritional products adhesives

Sales and profit figures have been converted from sterling at the rate of U.S. $1.765 to £1

XVI.I.247

Figure 4

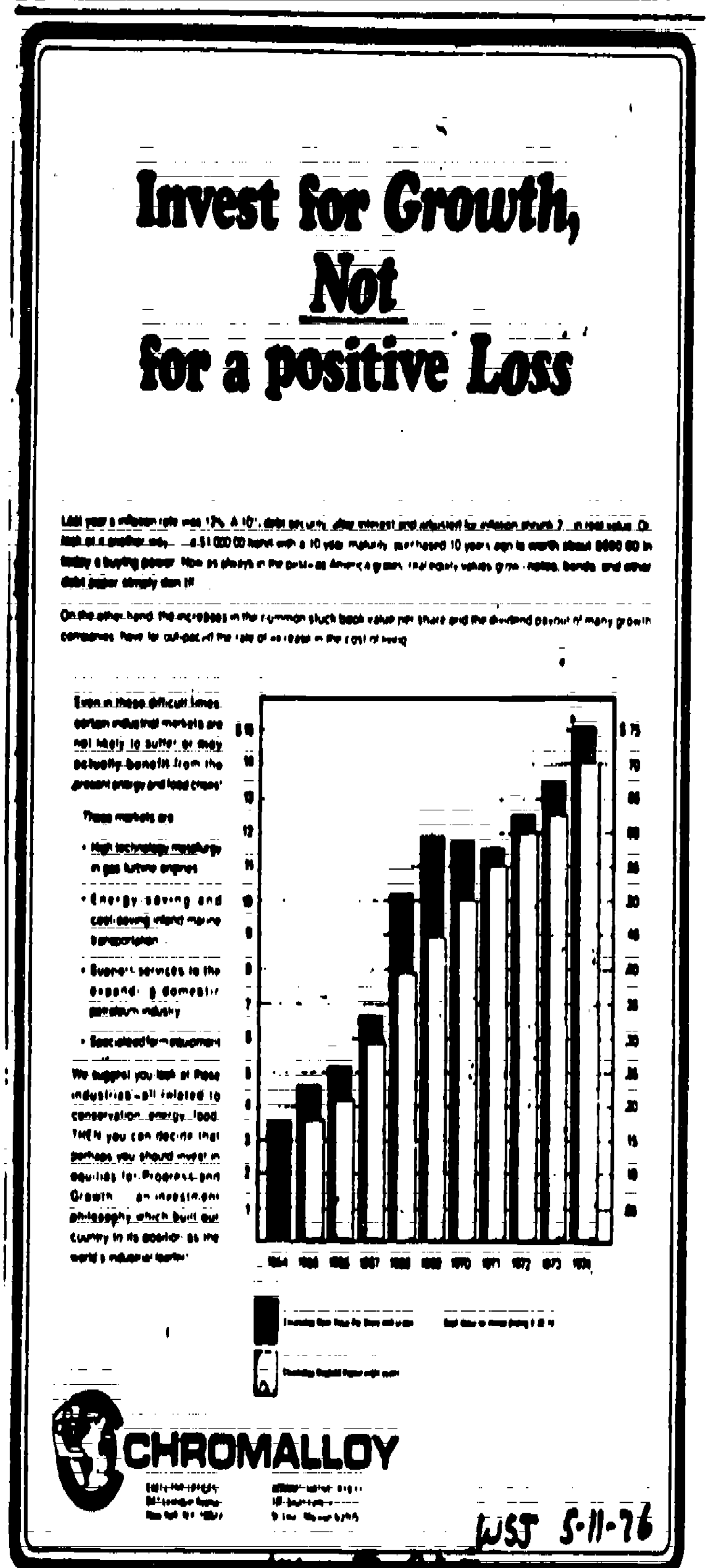
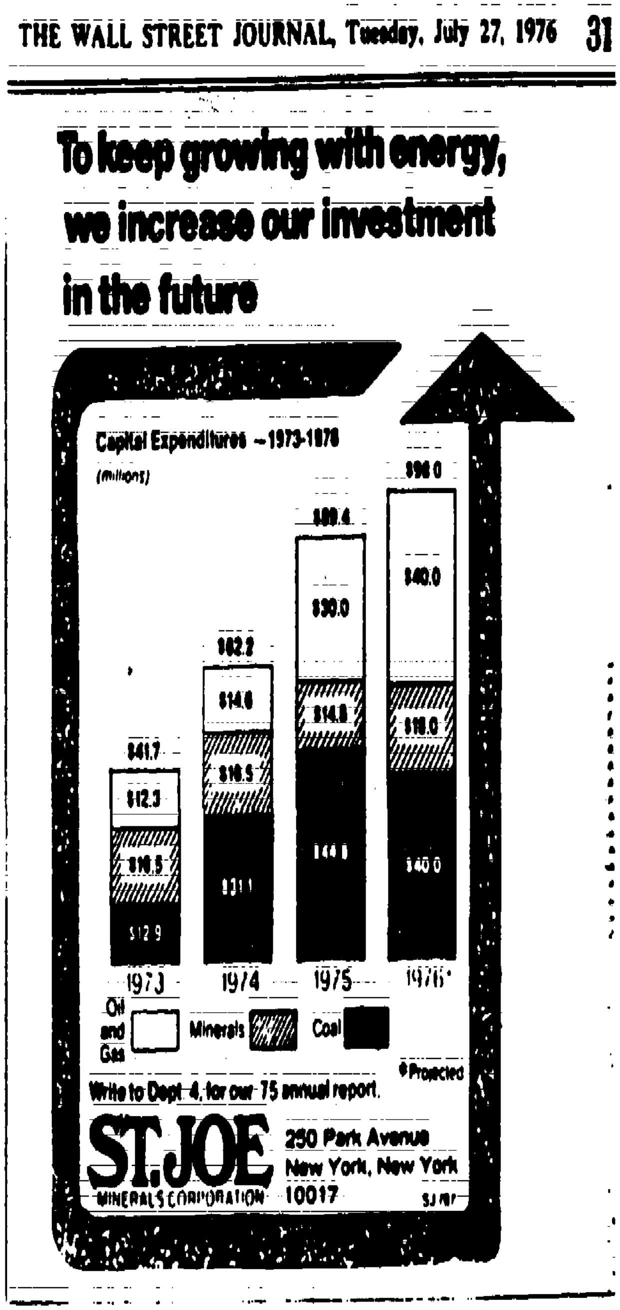


Figure 5.



Module I

Principle 2: Graphics are for communication

Graphic displays, as a substitute for oral or textual communication have 3 purposes: exploration, reconstruction, and decoration. Displays should be truthful and not misleading--see How to Lie With Statistics by Darrell Huff for some very "dishonest" graphics. Graphic tools should attempt to reconstruct reality and allow the reader to explore more fully the underlying situation in addition to decorating otherwise "dull" presentations.

Figure 6 is an example of a blot map, occasionally a very deceptive graphic device. In a blot map, we darken all counties or states that possess a certain characteristic. The blot map reproduced here was taken from The New York Times, and presents all the counties with 15 percent or more positive net migration of persons 60 years and older between 1960 and 1970. These 206 counties are supposedly the fastest growing retirement communities. The encircled counties in California, Arizona, Nevada, Utah, and Wyoming, listed at the bottom, include 40% of the shaded-in area on the map: however, only 0.14% of the people over 60 years live in these counties! The title of the article "More Elderly are Retiring in the North", is not at all verified by this map. One draws the incorrect conclusion that the Southwest U.S. is more popular than the remainder of the country, with the possible exception of the retirement haven, Florida. The moral is: Blot maps based on counties are misleading because of the large number of empty counties.

We also include a very good display, Figure 7, taken from Scientific American, which very effectively communicates information about nuclear devices.

Principle 3: Small Multiples are Useful

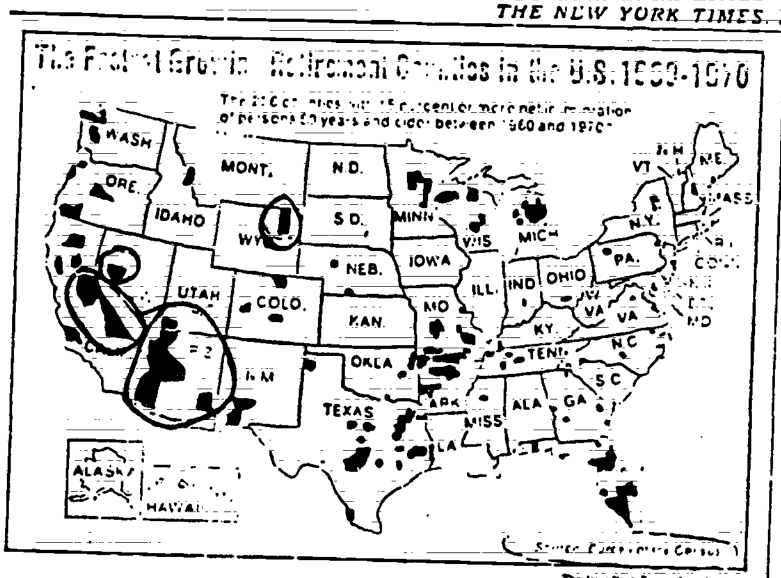
Graphic displays can be quite small. Many small displays, arranged on a page, can be quite effective in communicating your message. Figure 8 shows a 12 x 5 array of histograms presented in a good manner.

In a humorous vein, Figure 9 is an example of multivariate "faces", developed by Herman Chernoff. These small figures are used to differentiate observations from a larger population when more than one measurement on each observation is available. In a faces display each physical feature of a face is controlled by the value of a measurement. This is quite different from a display which puts faces on figures simply to portray the author's feelings about displayed values (see the light bulb example in the section "Principle 5").

Figure 6

New York Times, February 1, 1976

THE NEW YORK TIMES, SU



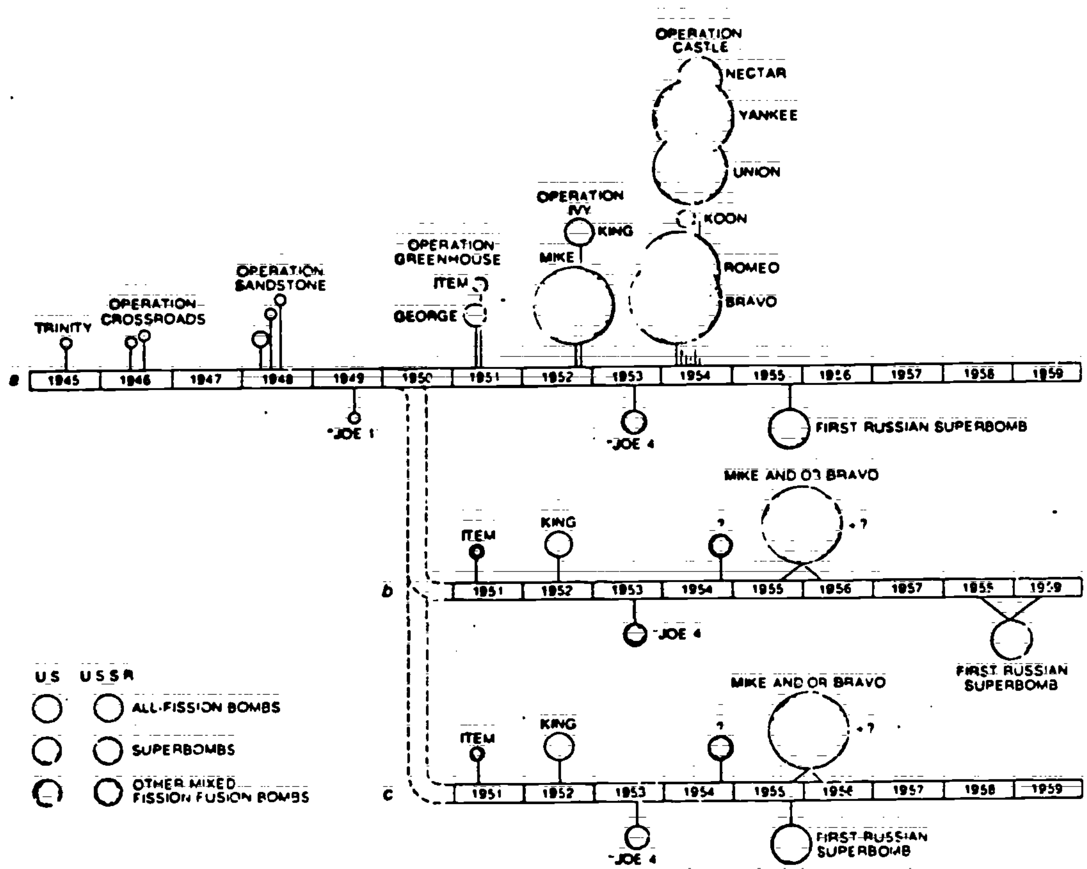
More Elderly Are Retiring in the North

California	Arizona	Utah	Nevada
Amador	Mohave	Washington	Chuuchill
Calaveras	Yavapai		
Inyo	Yuma		
Mariposa	Graham		
Mono			
Tuolumne			

Wyoming

Campbell

Figure 7



TWO HYPOTHETICAL OUTCOMES are postulated in an effort to evaluate how much risk would have been involved in a U.S. decision not to proceed with the superbomb. They are depicted in this historical chart as branches of the time line representing the actual world (a). The first branch is referred to by the author as the "most probable alternative world" (b), the second as the "worst plausible alternative world" (c). Both branches originate at January, 1950, the date President Truman announced his decision to go ahead with the superbomb. The circles denote nuclear-test explosions, the labels are U.S. code names. Area of each circle is proportional to the region that could be destroyed by that bomb. Bombs of "nominal" size (less than 50 kilotons) have been omitted after 1950.

110

Source: Herbert F. York, "The Debate Over the Hydrogen Bomb," Scientific American, 233 (October, 1975), p. 110.

308

Figure 8

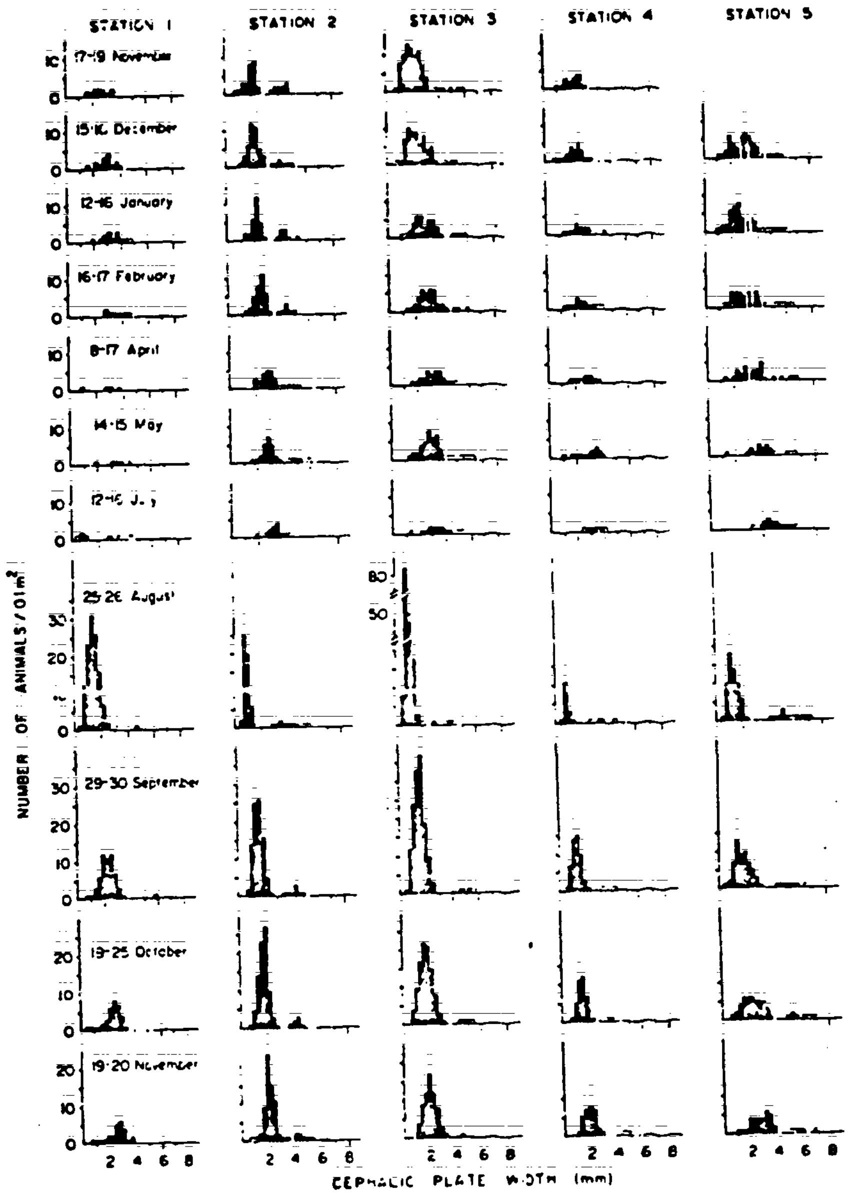
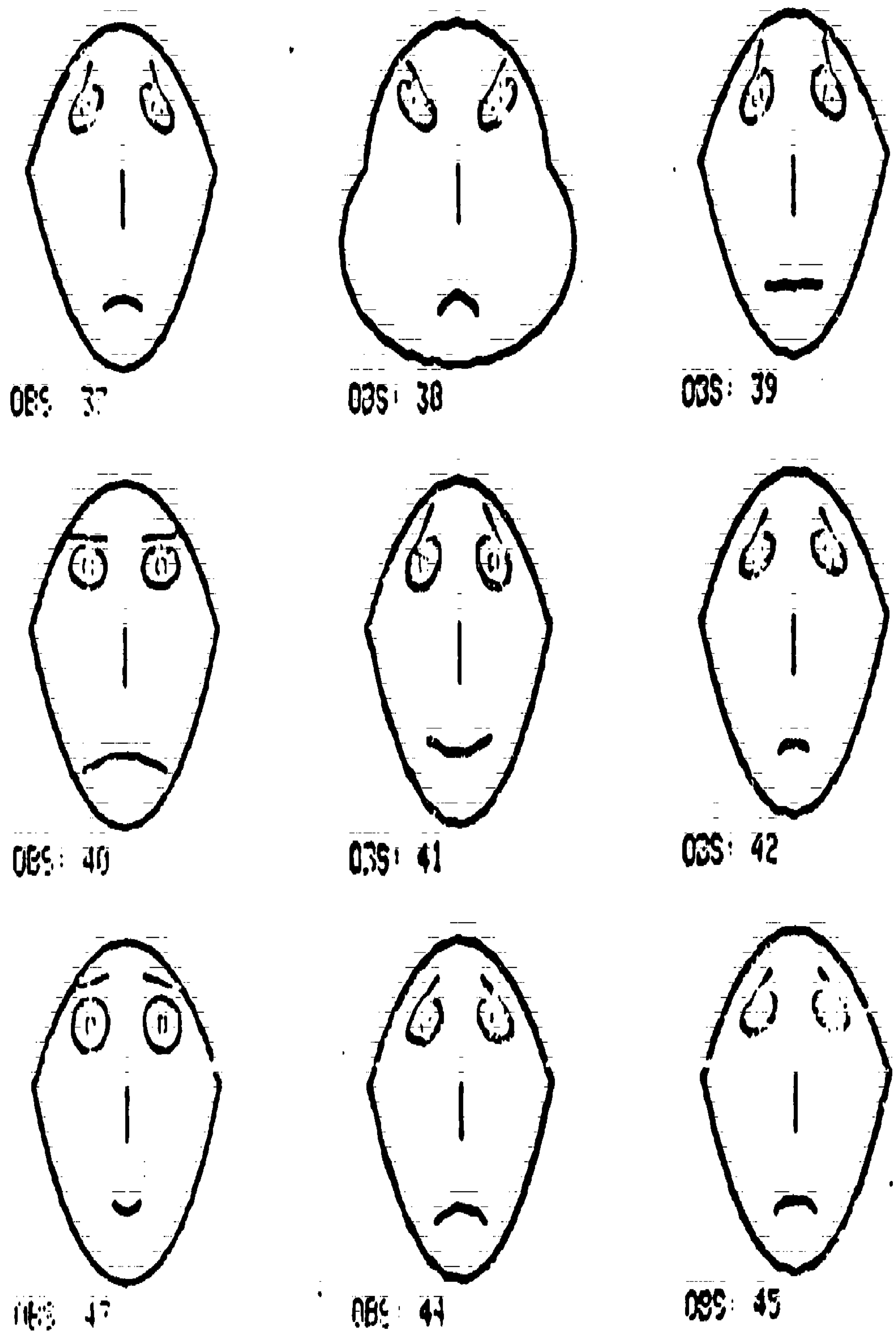


Fig. 1. Size frequency distributions of ... determined from mean grab-sample data. Open squares represent occurrence of specimens in densities less than one per 0.1 m²

Source: Frederic H. Nichols, "Dynamics and Energetics of Three Deposit-Feeding Benthic Invertebrate Populations in Puget Sound, Washington," Ecological Monographs, 45 (Winter, 1975), p. 66.

Figure 9



Principle 4: Think about Page Arrangement

When preparing graphics for publication, or written presentation, it is worth spending several minutes considering the arrangement of your displays on the printed page. The following display nicely summarizes the fourth principle of good graphics.

THINK
AHEAD

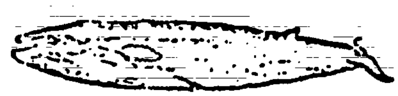
(Figure 10 is an example of a cute blunder from The New York Times.)

Figure 10

Correction

In last week's Review, a drawing of the finback whale appeared this way:

The drawing should have appeared this way:



The Review regrets the error. Whales, however, do spend just about as much of their time swimming on their backs as they do right side up.

New York Times, July 6, 1975.

Principle 5: Integrate Text and Graphics

Modern production of printed material has forced written text and graphics to become divorced from each other. The placement of all figures and tables at the end of the paper with wonderful announcements such as

---

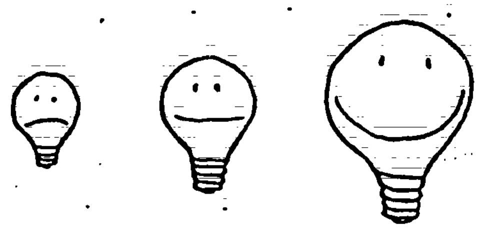
Table 983 goes about here.

---

does not aid the confused reader. Having to leaf through an entire paper to find a figure essential to the development of a hypothesis can be quite detrimental. Why do figures and tables have to be alienated from the text? This example from The New Yorker shows good integration.

PROFITS

Over the past fifteen years our profits have increased by a modest 75 million dollars. The diagram below



depicts, from left to right, three light bulbs of steadily increasing size.

Principle 6: Three-Dimensional Graphics

Graphics are in essence two-dimensional beasts since they must be reproducible on the printed page. However, as Figures 11 and 12 verify, graphics can be drawn as two-dimensional approximations to three-dimensional figures. The thing to remember, and we can label this Principle 6a, is: Professional artists can help by making good drawings, especially figures that are not easily drawn by hand.

But be careful! Figure 13 is a poor example. In this figure, ordinary histograms have uselessly been made three-dimensional. This is also a poor example of a histogram. One must ask whether it really is a histogram.

Principle 7: Rough Drafts

This principle is simply stated: Produce as many drafts of your graphic displays as you do of your text. Throughout this discussion we have equated graphics with the written word; consequently, it is to your advantage to polish your figures and tables as you polish your text.

Figure 11

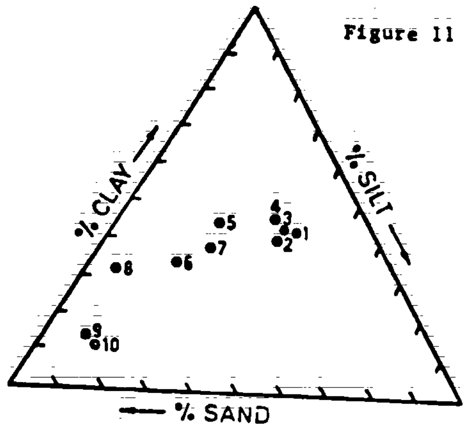
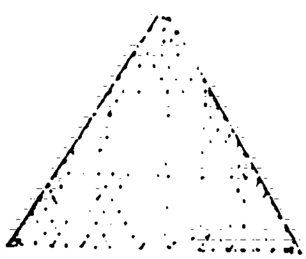


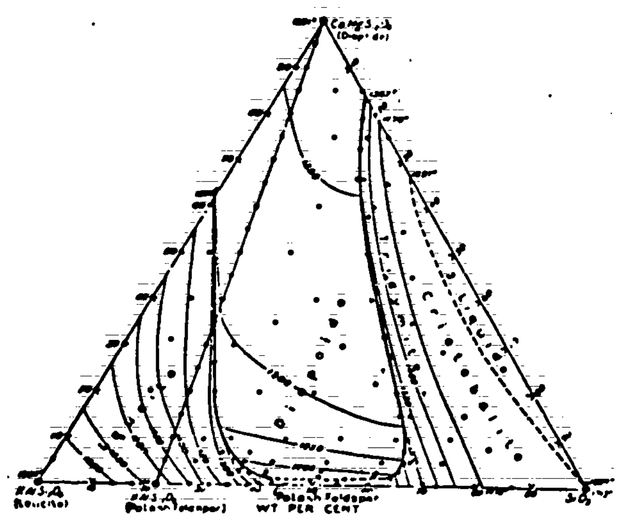
FIG. 5. Soil texture of *A* horizon under mixed grassland communities in western North Dakota. Numbers refer to stands described in text.

Source: Robert E. Redmann, "Production Ecology of Grassland Plant Communities in Western North Dakota," Ecological Monographs, 45 (Winter, 1975), p. 93.



A. Triangular Coordinate Graph Paper.

The trilinear chart was first used for investigation on strength of concrete mixtures. This form lends itself to the demonstration of problems involving a mixture of three ingredients, such as alloys containing three metals and food rations containing three dietetic elements.



J. F. Schairer and N. L. Bowen, "The System Leucite—Diopside—Silica," American Journal of Science, 1938. Geophysical Laboratory, Carnegie Institution of Washington.

B. Equilibrium Diagram of the Ternary System, Leucite—Diopside—Silica.

Figure 12

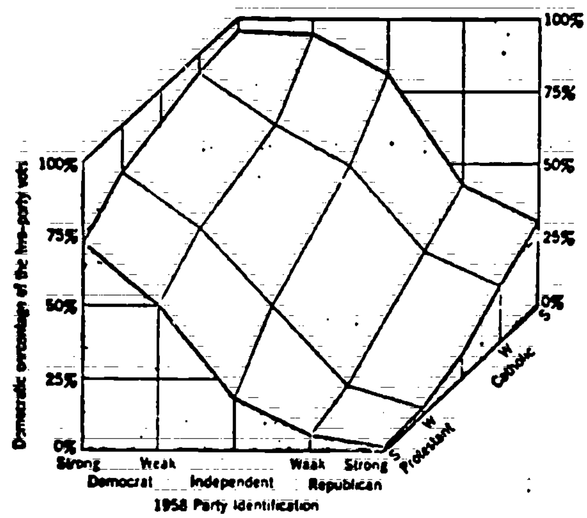


Figure 6-2a. 1960 presidential vote by party identification (1958) and by religious identification (1960).

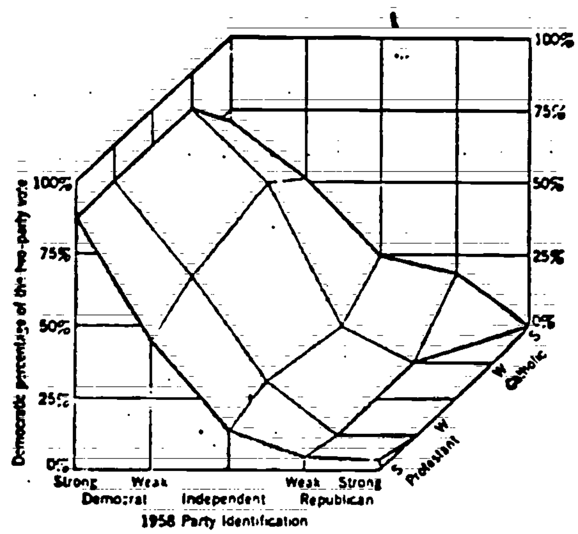
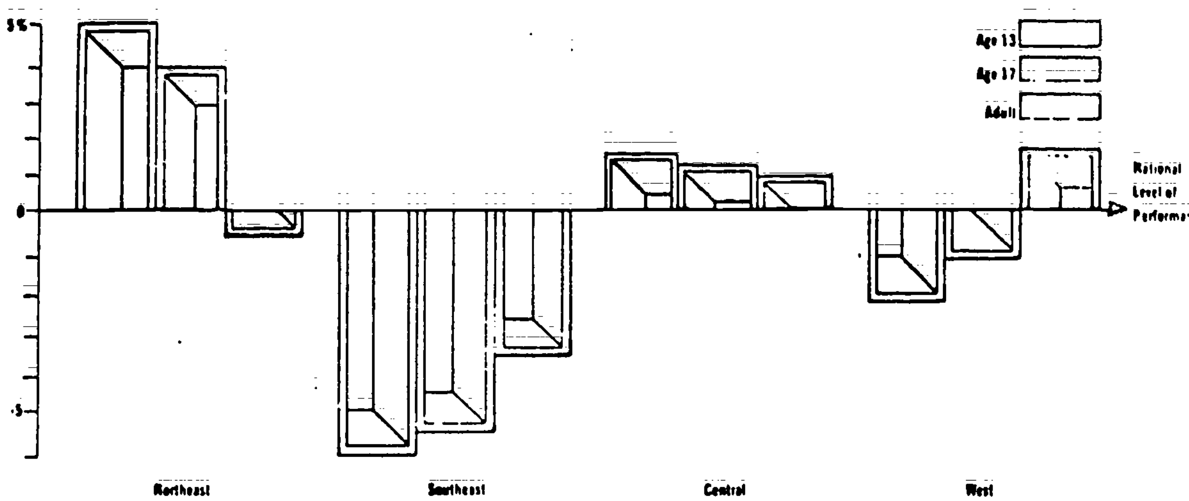


Figure 6-2b. 1956 presidential vote by party identification (1958) and by religious identification (1960).

Source: Philip E. Converse, "Religion and Politics: The 1960 Election," in Angus Campbell, Philip E. Converse, Warren E. Miller, and Donald E. Stokes, Elections and the Political Order (New York: Wiley, 1966), pp. 102-103.

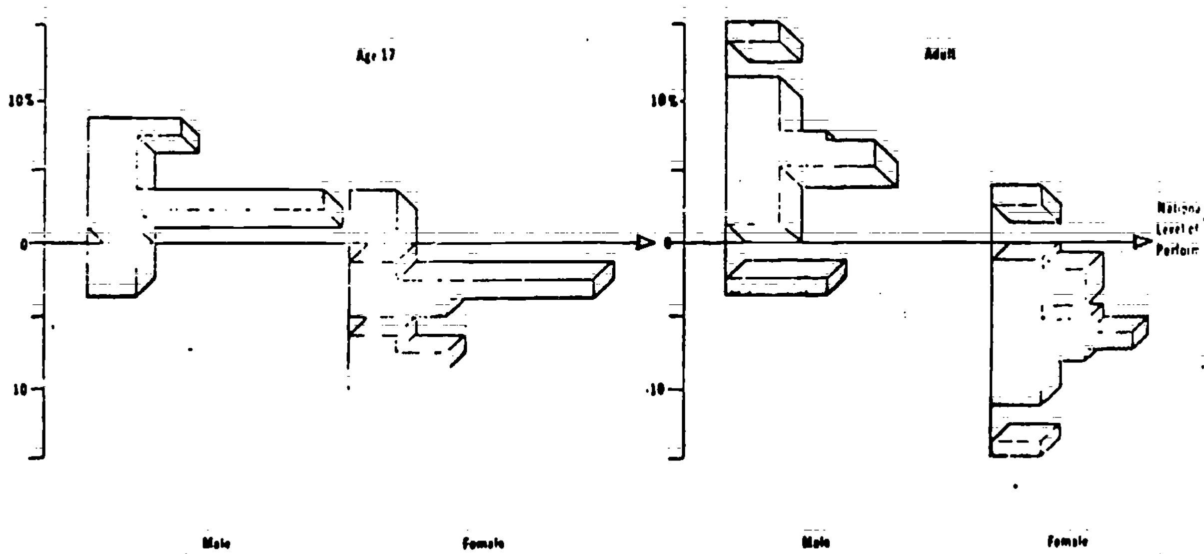
Figure 13

MEDIAN DIFFERENCES FROM NATIONAL PERFORMANCE BY REGION



Source: Office of Education, HEW, American Education, October, 1975, p.13.

MALE-FEMALE DIFFERENCES: CONSUMER COST PROBLEMS



OCT 8 - 1980

QUANTITATIVE METHODS FOR PUBLIC MANAGEMENT

MODULE II, REVISED

Developed by

SCHOOL OF URBAN AND PUBLIC AFFAIRS
CARNEGIE-MELLON UNIVERSITY

SAMUEL LEINHARDT, PRINCIPAL INVESTIGATOR
and
STANLEY S. WASSERMAN

Under Contract to

THE URBAN MANAGEMENT CURRICULUM DEVELOPMENT PROGRAM
THE NATIONAL TRAINING AND DEVELOPMENT SERVICE
5028 Wisconsin Avenue, N.W.
Washington, D.C. 20016

Funded by

The Office of the Assistant Secretary
for Policy Development and Research
U.S. Department of Housing and Urban Development

Package XVI

Acknowledgements

Assistance in the preparation of this package was provided by Blaine Aikin, Larry Albert, Joseph Chmill, Steve Clark, Marjorie Farinelli, Janice Greene, Gretchen Hemmingsen, Paul W. Holland, J. Michael Hopkins, Gaea Leinhardt, Richard Sandusky, Christine Visminas, Diane Warriner, and Tammar Zeheb.

TABLE OF CONTENTS

Material intended solely for the instructor is denoted by a (I). Material that should also be distributed to the students is denoted by a (S).

Introduction to Module II

Overview

Module II of the Quantitative Methods for Public Management package provides students with experience in handling complicated data sets describing policy relevant issues and, thus, promotes the development of analytically oriented managerial skills. Two kinds of skills are emphasized: performance and criticism. The module contains two units, numbers 3 and 4. Unit 3, y versus one X, introduces the student to modeling a relationship between two variables: a carrier variable, X, and a response variable, y. The general strategy is to use a linear model of the relationship and explore the utility of various transformations on X or y or both in improving the fit of a linear model. Fitting, modeling, finding equations for data, and evaluating a fit are all specific technical skills taught in this unit. Some simple procedures are introduced for determining a good transformation and for fitting a line to transformed data. All procedures can be performed without the aid of a computer.

Unit 4, the second unit in Module II, introduces the student to modeling relationships between one response variable, y, and multiple carrier variables, X_i. Transformations to improve the reasonableness of a linear model are again stressed. In this unit the fitting technique is least squares regression, and the student receives an extensive exposure to the mathematical principles of of the least squares fitting procedure as well as numerous examples of applications with special emphasis on the pitfalls and dangers of simple, mechanical application of regression analysis to multivariate data.

Specific Objectives

Unit 3

Upon successful completion of unit 3 a student will be able to perform graphical analyses of multiple ordered batches of quantitative data, summarize these batches using the notions of a conditional typicalvalue, construct scatterplots of X,y data sets in which X and y are quantitative variables, use a line fitted through the conditional typicals to model an X,y data set, use least squares to fit a line to X,y data set, find a transformation of X and/or y to improve the linearity of a fitted model for the data, and analyze X,y data in which X is a variable indicating time. The critical skills a student will obtain include the ability to evaluate how well typical conditionals summarize batches, evaluate the ability of a linear fit to summarize a X,y data set, evaluate the comparative advantage of least squares versus other fitting procedures, evaluate the need for a transformation, and evaluate the need for smoothing of a data set in preparation for an analysis of time series data.

Unit 4

Upon successful completion of Unit 4 a student will be able to construct a model for continuous multivariate data using the least squares procedure, find transformations that improve a least squares fit, interpret coefficient values in a regression model, use indicator variables and splines in model construction, perform inference on coefficients, perform regression analyses on a computer and by hand, and evaluate a fitted model.

In this unit the critical skills the student will learn center around comprehension of the effectiveness of the least squares procedure as a fitting technique and the problems that arise when nonlinearity is present, when overfitting occurs, when residuals are not normally distributed, and when carrier variables are collinear. Students will be able to evaluate the appropriateness of using the least squares fitting procedure for specific data sets and be able to determine whether results are due to relationships in the data or to the peculiarities of the fitting procedure. Since this technique is one of the most common analytic procedures appearing in quantitative policy studies, the "doing" and "criticizing" skills learned in Module II will be very important to the practitioner.

Unit 3
Reading Assignments

Lecture	Reading
3-0	Prerequisite Inventory
3-1	Tukey, Chapter 5
3-2	Tufte, DAPP, Chapter 1
Workshop	"Graphics for Scatterplots"
3-3	McNeil, Chapter 3 Tufte, DAPP, pp. 65-108 Tukey, Chapter 6
3-4	McNeil, Chapter 6 Tukey, Chapter 7

In addition, read the following articles in Tanur et al.:

pp. 120-129
153-161
195-202
354-361

and the following articles in Tufte, QASP:

pp. 37-67
113-125

Texts:

McNeil, Donald R., Interactive Data Analysis, New York: John Wiley & Sons, 1977.

Tanur, Judith, et al., editors, Statistics: A Guide to the Unknown, San Francisco: Holden-Day, 1972.

Tufte, Edward R., Data Analysis for Politics and Policy, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1974.

Tufte, Edward R., editor, The Quantitative Analysis of Social Problems, Reading, Massachusetts: Addison-Wesley Publishing Co., 1970.

Tukey, John W., Exploratory Data Analysis, Reading, Massachusetts: Addison-Wesley, 1977.

Prerequisite Inventory
Unit 3

Unit 3 of Module 1 focuses on the analysis of ordered multiple batches and paired batches of data, i.e., data in which every element of one data vector is associated with an element in another data vector. As in the prior two units, the skills to be learned in this unit presuppose mastery of several elementary concepts and procedures. Before proceeding to Unit 3, you should assure yourself that you are familiar with these basics.

This inventory is divided into the following four sections:

1. Review of Units 1 and 2
2. Functions---Paired Observations, Notation, Plotting
3. Special Types of Functions---Linear, Absolute Value, Exponential, Inverse, Logarithm, and Polynomial
4. Properties of Functions--Minimization

Additional references to these topics appear in the Appendix. Homework problems have been assigned which require use of these concepts. If you discover that you are weak in areas which will not be covered in class, you should consult appropriate course personnel to arrange for tutorial assistance and/or a reading guide to background material.

Section 1. Review of Units 1 and 2

Data acquired by data analysts are usually organized in arbitrary fashion. While arbitrarily organized data may make retrieval of specific values easy (e.g., an alphabetical organization of test grades for students in a class) it obscures the behavior of the batch of values and makes continued analysis of the batch difficult. The stem-and-leaf display is one tool the data analyst may use to organize data analytically. This type of display possesses features of a numerically ordered sort of the values and of a histogram simultaneously. While it permits retrieval of individual values it also provides a picture of the shape of the batch and permits one to obtain the order statistics by counting in. In constructing a stem-and-leaf display one first notes the extreme values of the batch and makes a choice of unit for the leaves. These are placed to the right of a vertical line which breaks the original values in the batch into stems, which are multiples of the unit, and leaves. Thus, the numbers -30 through 30 would appear below in one possible stem-and-leaf display.

```
-3 | 0
-2 | 9876543210
-1 | 9876543210
-0 | 987654321
 0 | 0123456789
 1 | 0123456789
 2 | 0123456789
 3 | 0
```

(Notice that the location of zero on the number line has been split into -0 and +0). It does not matter whether smaller values appear towards the top of the display or towards the bottom; the choice is up to the data analyst. Stretched versions of the display are possible by using two lines per stem with leaves with values from 0 to 4 on one line and 5 to 9 on another (using * and . as reminders) or five lines per stem (using *, t, f, s, . as reminders). It is sometimes necessary to change the unit in the middle of a display. By using asterisks as place holders and placing leaves on the set of stems with the correct number of asterisks, such compound stem-and-leaf displays can be created. The integers from 80 to 200 illustrate this implicit increase in unit.

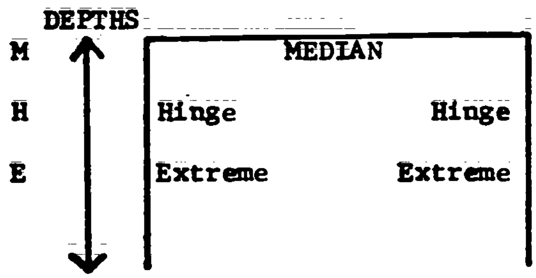
```
 8* |0123456789
 9* |0123456789

1** |0123456789
2** |0
```

(Note the inclusion of a blank row to separate the parts of the display which are based on different units.) Also note that the change in unit means that not all the integers from 100 to 200 are displayed. Rather, the units shift leads to representing only the integers 110, 120,. . . , 190, 200. Remember that when making stem-and-leaf displays free hand, care must be taken to line up leaves under one another; otherwise the display's ability to give an accurate impression of the shape of the batch may be compromised.

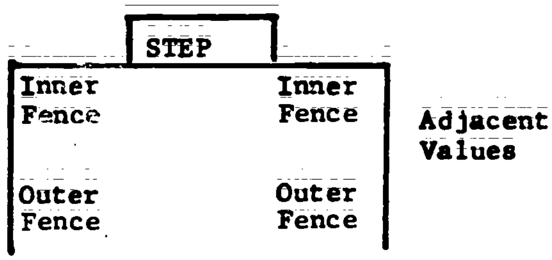
If the stem-and-leaf display still seems a bit confusing Chapter 1 of McNeil or Chapter 1 of Tukey should be reviewed.

While the stem-and-leaf display is a handy and effective data organizing tool, for some purposes it may retain too much information. Putting the information that is in a batch into numeric and graphic summaries is called condensation. While some information is lost in this condensing process, it reduces the number of distracting factors to a small, easily appreciated set of values or aspects of pictures. These summaries are usually more easily manipulated and contrasted than are stem-and-leaf displays. But remember, they are not as informative as a stem-and-leaf display. It is usually wise to examine a stem-and-leaf display of a batch before condensing it.

The five number summary contains the median, hinges and extremes. The median is the value obtained by counting in the sorted batch halfway. It is located at the "depth" (N+1)/2 where N is the total number of values in the batch. If N is even, then the median is the mean of the two middle values. The hinges are located at: (depth of the median +1)/2. The extremes are simply the largest and smallest values in the batch. They are located at either end of the batch at depth 1. Tukey suggests the following letter value display for this summary.

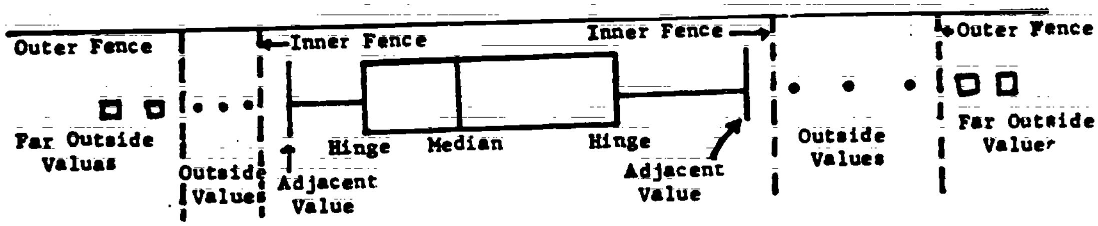


An extended version of this display, a <u>fenced letter display</u>, gives the value of the midspread (the difference in the hinges), a step (1.5 times the midspread), the inner fences (1 step beyond each hinge), the outer fences (2 steps beyond), and the adjacent values (the last values before each inner fence).



(Note that for some batches adjacent values and hinges may be equal.)

The schematic plot is a graphical summary. It represents the values from a fenced letter display as a picture.



Note that the inner fences and outer fences are not actually drawn in the display.

The schematic plot can be drawn vertically or horizontally at the discretion of the analyst.

The utility of this picture rests in its schematic quality; it is a structural outline of the batch. Of course, there will be batches whose structures are not well conveyed by this form of display. Batches with separations between values may fit in this class. Consequently, graphical summaries as well as numerical summaries should be relied upon only after the entire batch has been examined in a stem-and-leaf display.

It is easier to think about and summarize batches which are <u>symmetric</u> than those which are not. In a symmetric batch the median will be in a position around which the batch could be folded with one half of the batch reflected by the other half. Consequently, the hinges and extremes will be equally spaced from the median. When the batch is symmetrical, the mid-hinge (the mean of the hinges, MidH), and the mid-extremes (the mean of the extremes, MidE), will have the same value as the median. This fact can be used to test for symmetry. (Note that in some batches which we will call symmetric these values will be only approximately equal.)

Some batches, while not symmetric in the original unit, become symmetric after a simple power transformation of the form X^r where r is a simple power, a rung on the ladder of powers and where $r = 0$ implies logarithms. To find the transformation that best symmetrizes the batch we need only investigate the midsummary values derived from the batch's five number summary. When $M < \text{Mid } H < \text{Mid } E$ we go down the ladder of powers; when $M > \text{Mid } H > \text{Mid } E$ we go up. It should be noted that there may not be a convenient r which symmetrizes the batch in question. In that case, the raw values must suffice. The midsummary array of a transformed batch is called a transformation summary.

The usual values of r are 1/2 for square roots, 2 for squares, -1 for reciprocals (negative to preserve order) and 0 for logs. Sometimes we can achieve easy transformation simply by using the following rules for types of data. For amounts and large counts use logs, for percentages use the arc sine of the square root, and for balances transform before subtracting to obtain the balance.

A special kind of symmetric batch that has no outliers and closely approximates a theoretical Gaussian or Normal curve, is called a well behaved batch. It is mathematically convenient to summarize such batches with the mean and standard deviation. The mean, $\bar{X}$, is equal to:

$$\frac{\sum_{i=1}^{N} x_i}{N}$$

and the standard deviation, s, is the square root of the variance and is equal to

$$\sqrt{\frac{1}{N}\Sigma(X_i-\bar{X})^2}$$

Another way of thinking about the standard deviation is to view it as the average squared deviation about the mean. An important property of a well behaved batch is that s ≈ 3/4 midspread. (Remember the symbol "≈" means "approximately equal to".) We can transform any well behaved batch into a standard well behaved batch by subtracting the mean of the batch from each value and dividing each difference by the batch standard deviation. The resulting batch of standardized values has mean = 0 and variance = standard deviation = 1.

The importance of this standardization process is that a great deal is known about the properties of standard well behaved batches. In particular, we know what percent of the batch lies between various values. For example, between -1.96 and +1.96 lies 95% of the values.

Well behaved batches play particularly important roles in statistical inference for regression by least squares, a topic to be covered later in this course. Any well behaved batch is completely summarized when its mean and standard deviation are known. For ill behaved batches, which are by far the more common variety, the mean and standard deviation are very rarely sufficient summaries.

When we have only one batch of values there is little additional analysis that can be performed. Once we have answered questions concerning typical value, spread, shape, and separations and have searched for a symmetrizing transformation, we have obtained just about as much information as we can. However, when the data come in the form of multiple batches, we can expand our inquiry by contrasting the batches with one another. Unordered multiple batches are a set of batches (2 or more) which have no quantitative relation between them, i.e., the batches cannot be located on a scale. Ordered batches, to be considered in unit 3, possess this property. By contrasting batches we mean that we can compare typical values, shapes, etc.--all of the features which, for single batches, we simply noted.

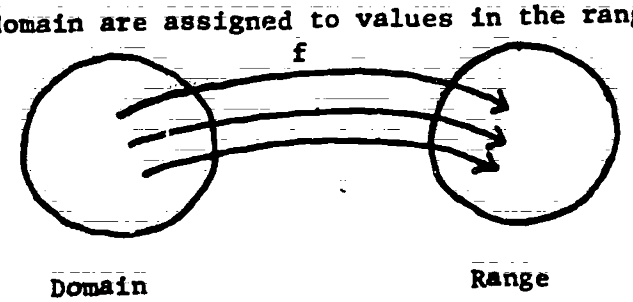
To perform contrasts on unordered multiple batches we can use the same tools employed earlier but in parallel fashion. That is, we can draw stem-and-leaf displays and schematic plots side by side. We must be careful here to have the plots on the same scales and to use the same units. We can also use side by side number summaries for contrasts.

Often, a consistent relationship between typical value and spread may interfere with our ability to make an effective comparison. This occurs when there is an obvious trend in which typical value increases with spread or vice versa. When this happens it is very difficult to determine how much of apparent differences in typical values are due to differences in spread. An appropriately chosen transformation can often effectively <u>equalize</u> <u>spread</u> and permit us to perform contrasts with the confounding influence of changes in spread eliminated. We can usually find this transformation by first calculating the median and midspread for each batch and then making a scatterplot of log(median) against log(midspread). If a clear line seems to fit these points we crudely estimate the slope of this line, m, and transform all the batches by taking X^r where $r = 1-m$, following the rules for the ladder of powers. One may also view this procedure as a way of obtaining the appropriate unit for all the data. (Sometimes it may be necessary to examine log(median) against log(difference in extremes).)

Section 2. Functions

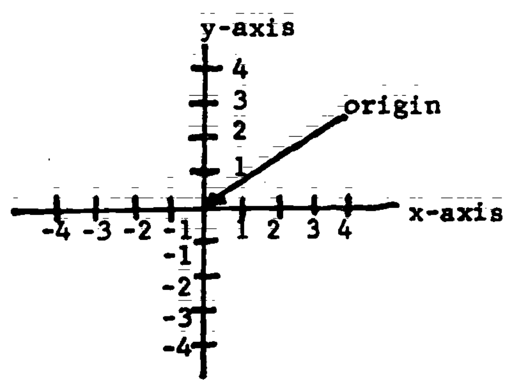
One of the most fundamental concepts in modern mathematics is that of <u>functions</u>. A function is an operation involving two sets of numbers, the input values which are usually denoted by x, and the output values, usually denoted by y or f(x). (Note this functional notation. We read notation f(x) as "f of x" or " a function of x". Other letters in upper or lower case may be used instead of f. Do not confuse the use of parentheses here with their usual usage in an equation where they imply a multiplicative operation. Brackets, [ ], and other separators may be used for the same function notation purposes.) To each input value, a function assigns exactly one output value. The set of all input values for a function is called its domain and the set of all output values is its range.

The mathematical operations that we use most often, such as square root, square, and logarithm, are all functions. When we use a variable to represent the domain of a function, we call it an independent variable. The variable representing the range is called the dependent variable; it is functionally dependent on the independent variable. Functions are usually indicated by letters preceding the independent variable which is enclosed in parentheses. Thus, $f(x) = \log_b x$ is the logarithmic function; $f(x) = x^{1/2}$ is the square root function, etc. A useful way of thinking about functions is to view them as rules of correspondence. A graphical representation of this assignment process, in which values in the domain are assigned to values in the range, is shown below.

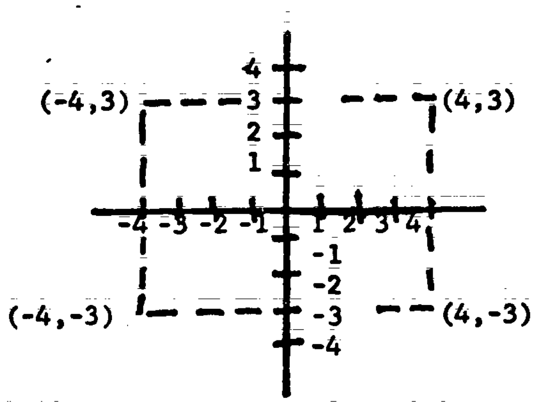


We deal with functions whose domain and range consist of real numbers, i.e., functions of real variables. A subset of these functions are functions which have as their domains only the integers. These are called functions of discrete variables.

We can obtain geometric representations of functions of a variable by graphing or plotting the function on a rectangular or Cartesian coordinate system. In this system two real lines (coordinate axes) are drawn at right angles on a plane so that they share a common origin. The convention is to label the vertical line as the y-axis and the horizontal line as the x-axis.



When a rectangular coordinate system is drawn on a plane the plane is called a rectangular coordinate plane or xy-plane. Points may be plotted on this plane in the following way. An ordered pair of values, one for the x variable and one for the y, in that order (by convention), is determined. These are usually given as a pair of values enclosed in parentheses. To graph the point we move along the x-axis to the location on the x-scale that equals the first member of the ordered pair of values. We then move vertically above or below this point parallel to the y-axis until we reach an imaginary horizontal line intersecting the y-axis at a point on the y-scale equal to the second of the ordered pair of values. An illustration appears below for the points (-4,-3), (4,-3), (-4,3) and (4,3)



Obviously, the procedure can be performed in reverse order by first finding the y-axis location and then the x. The dashed lines in the illustration are provided for clarity and are not drawn in practice. To distinguish one point from another we use subscripts. Each ordered pair receives a subscript value specifying its sequence in the set of ordered pairs. If we think of the four plotted points as having come to us in the sequence (-4,3), (4,3), (4,-3), (-4.-3) then we can indicate these four points as follows:

$$(x_1,y_1) = (-4,3)$$
$$(x_2,y_2) = (4,3)$$
$$(x_3,y_3) = (4,-3)$$
$$(x_4,y_4) = (-4,-3)$$

In general, points are denoted (x_i, y_i) where i runs from 1 for the first point to n, the last point. When (x_i, y_i) appears alone it means "some arbitrary point".

Formally, we say that the rectangular coordinates of a point are given by the ordered pair (x,y) and we use the terms point and ordered pair interchangeably. The terms for the x and y values in the ordered pair are <u>abscissa</u> and <u>ordinate</u>, respectively. We also name the <u>quadrants</u> into which the rectangular coordinate system divides the plane as follows.

II	I
III	IV

Quadrants of the Cartesian Plane

We can graph functions by recording in ordered fashion values of the domain and corresponding values of the range. If the function is defined on a continuous variable then the number of possible points that can be graphed will be infinite regardless of whether the

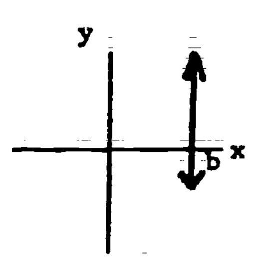
339

function is bounded. We represent the graph of a continuous function as a smooth curve and a discrete function as a set of distinct points.

The procedure involves constructing a table for x and y values. One chooses x values in the domain of the function and solves the equation for the corresponding individual values of y. Usually, a few ordered pairs of values are sufficient to allow us to make a geometric picture of all the ordered pairs that represent solutions to the equation. Obviously, if the domain and range extend over all the real numbers, we can't graph the function all the way to infinity. By convention, we place arrow heads on the end of graphed curves to indicate that the function continues similarly beyond the last plotted point (although sometimes the arrowheads are left out). Examples of these procedures appear below and in the next section.

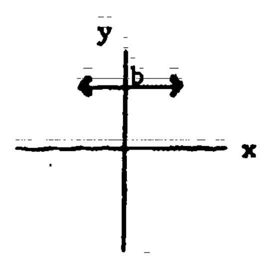
Note that not all equations in x and y define functions of x. Nor are all curves that can be drawn on the xy-plane functions of x. The critical quality of a function is the assignment of a of a single value in the domain to a single value in the range. Compare the following graphs.

Not functions of x.



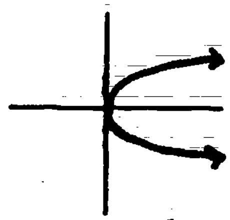
x = b

Functions of x.

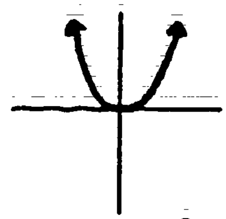


y = b

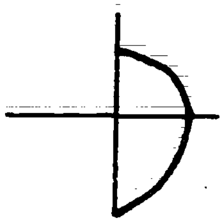
Not functions of X | Functions of X

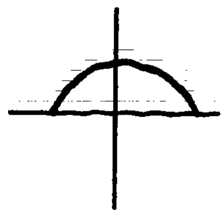


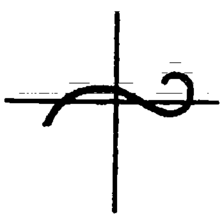
$x = ay^2$

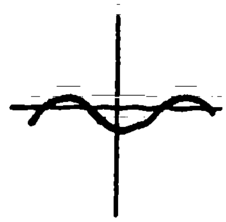


$y = ax^2$









In general, the graph of a function, $f(x)$, contains all points $(x,f(x))$, where x is in the domain of f. The procedure for plotting equations which are not functions of x requires obtaining all the multiple values for y which represent solutions to the equation for a given value of x.

Section 3. Special Types of Functions

A. Linear functions

Functions of the form

$$f(x) = ax + b$$

where a and b are real numbers are called linear functions. Their graphs are straight lines which intercept the y axis at the point (0,b) and have slope = a. The slope is often denoted by an "m" and is the ratio of the vertical change to the horizontal change, or the number of unit changes in y for a unit change in x. The Greek upper case delta,"Δ", by convention, is used to represent change. Thus, slope is

$$m = \frac{\Delta y}{\Delta x}$$

and can be computed from any two points that lie on a line, (x_1, y_1), (x_2, y_2) by the following definition:

$$m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

For any pair of points satisfying a given linear function, this ratio is constant. The slope of the vertical line $x = a$ is not defined (note that this equation is not a function of x).

Some important facts about lines follow.

a. Two lines with slopes $m_1 = m_2$ are parallel.

b. A horizontal line has slope $m = 0$.

c. If $m > 0$ the line rises from left to right, i.e., it slopes upward.

d. If $m < 0$ the line falls from left to right, i.e., it slopes downward.

e. The form $y = mx + b$ is called the <u>slope-intercept</u> form of the line.

f. The form $y-y_1 = m(x - x_1)$ is the <u>point-slope</u> form of the equation for a line.

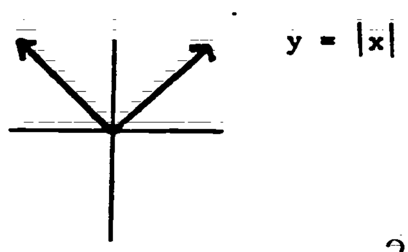
g. The form $ax + by + c = 0$ is the <u>general linear</u> form of the equation for a line.

B. <u>Absolute value</u>

The function with the form

$$f(x) = |x|$$

is called the <u>absolute value</u> function. The two vertical bars surrounding the x on the right side of the equation are a notational convention indicating that only nonnegative values of x are to be returned. Thus, the domain of the absolute value function is all the real numbers while its range is the nonnegative reals. Its graph appears below.



The Absolute Value Function

343

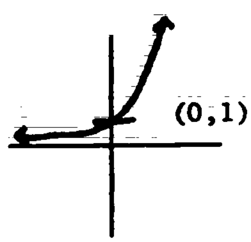
C. Exponential

The function with the form

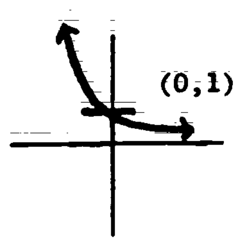
$$f(x) = b^x$$

where $b > 0, \neq 1$ with all the real numbers in its domain is called the *exponential* function. Its range is all the positive numbers.

The exponential function has a graph with the shape:



when $b > 1$ and a shape



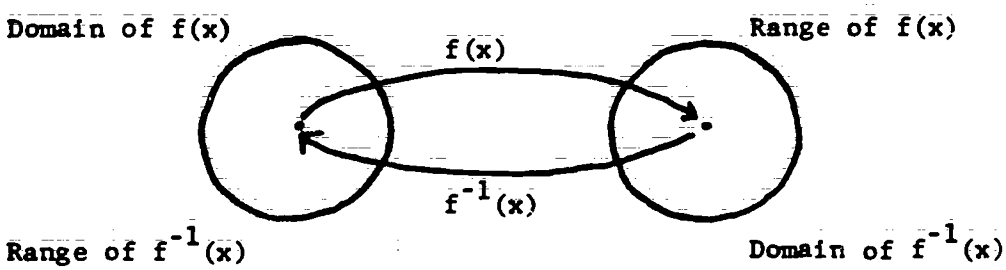
when b is between 0 and 1. Regardless of the actual value of b, the graph has one of these two basic shapes. For $b > 1$, as b increases the curve becomes more steep. For $b < 1$, as b decreases the curve becomes more steep.

The constant, b, is called the <u>base</u> of the exponential. Two common bases are the irrational number, $e \approx 2.71828$, and the number 10. Note that graphs of all exponential functions pass through the point (0,1). Furthermore, although the curve approaches the x axis, it never actually intersects it.

D. <u>Inverse</u>

The inverse of a function is that function which when applied to a function of x returns the original value of x. The inverse, by convention, is denoted by a superscript -1 placed at the upper right hand side of the function just before the left parenthesis, e.g., the inverse of $f(x)$ is $f^{-1}(x)$ and $f^{-1}[f(x)] = x$.

If we think of a function as a rule of correspondence which assigns a value in the function's range to every value in the domain, then the inverse function takes as its domain the function's range and assigns to these values the corresponding values in the function's domain. For example, if $f(x)$ is a function with a domain value of 5 assigned to a range value of 25 then $f^{-1}(x)$ has a domain value of 25 to which 5 is the assigned value in its range. A graphical picture of this process appears below.



The inverse of the linear function, $f(x) = ax$ is $f^{-1}(x) = (1/a)x$ since, by substitution, $f^{-1}f((x)) = f^{-1}(ax) = (1/a)(ax) = x$. The function $f(x) = x$ is its own inverse. In general, if $f(x) = ax+b$ then $f^{-1}(x) = (1/a)(x-b)$.

E. <u>Logarithm</u>

Another important inverse function is the inverse of the exponential, the <u>logarithmic</u>. Recall that the logarithm of x to the base b is defined as

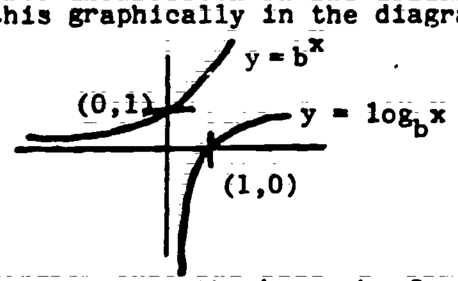
$$y = \log_b x$$

which is simply a notational way of saying that

$$b^y = x.$$

To see that the logarithm is the inverse of the exponential we must determine $f^{-1}[f(x)]$ for $f(x) = b^x$. Assuming that $f^{-1}(x) = \log_b[f(x)]$ we have $f^{-1}[f(x)] = \log_b b^x = x \ (\log_b b) = x$. In other words, the logarithmic function is defined as the inverse of the exponential, i.e., it is that function which reverses the ordering of the pairs of points that represent assigned values of the domain and range for exponential functions.

We can see this graphically in the diagram below.



We see from the diagram that the base, b, for the logarithm is the same constant that is the base of the exponential. Where the point (0,1) must be on the exponential plot, the reversed point (1,0) must be on the logarithmic. While the exponential approaches the x-axis (i.e., the point where $y = 0$) but never reaches it (except at $x = -\infty$ in the example) the logarithmic approaches the y-axis (i.e., the point where $x = 0$) but never reaches it (except at $y = -\infty$ in the example).

E. <u>Polynomials</u>

Functions of the form

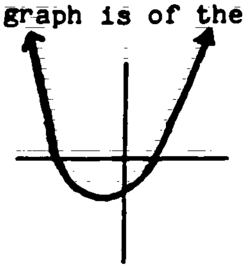
$$y = f(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$$

are called <u>polynomials</u> in x of degree n where n is the largest exponent of x for $a \neq 0$. The exponent of x is always nonnegative and the constants, a_i, are real numbers. Polynomials in x can be plotted on the Cartesian plane.

The linear function is a polynomial in x of degree 1. An important polynomial is the polynomial of degree 2, called the <u>quadratic</u> function. It is usually written as

$$y = ax^2 + bx + c$$

The domain of the quadratic function is all the real numbers and, in general, the graph is of the following form.



$y = ax^2 + bx + c$

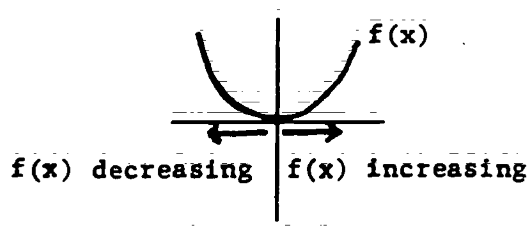
317

Note that the graph of the quadratic function may be shifted to the right or left or up or down depending on the values of the constants. A graph of a quadratic is called a parabola. Another form is $y = g(x-h)(x-k)$ where g, h and k are constants.

A polynomial of degree 3 is called a cubic. Graphs of cubics have two bends in them. Polynomials of degree higher than three do not have special names. In general, the graph of a polynomial of degree n has n-1 bends.

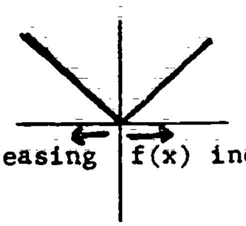
Section 4. Properties of Functions--Minimization

Functions of x can be evaluated over their entire domains or over portions of their domains. Within any interval on the x axis, a function is said to be increasing or decreasing depending on whether, as x increases, f(x) is strictly increasing or strictly decreasing, respectively. For some functions, such as linear functions, the function will be either increasing or it will be decreasing over its entire domain. This is also true of the exponential and logarithmic functions. However, polynomials have bends in them and are strictly increasing or strictly decreasing only within some interval on the x-axis.



A Parabola

The parabola in the example decreases in the interval $-\infty<x<0$ and is increasing in the interval $0<x<\infty$. The absolute value function, $f(x) = |x|$, behaves similarly.



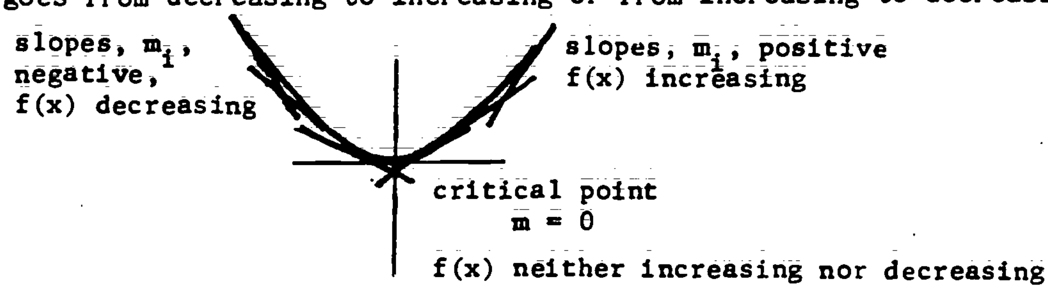
Absolute Value Function

Functions that are strictly increasing have minimum values of f(x) at their left-most bound. In other words, linear functions with positive slope have values of f(x) which are less than any other value in their range when x is the lowest value in their domain. Linear functions that are decreasing over their domain, i.e., returning smaller and smaller values for f(x) as x increases, have a minimum value in their range that corresponds to the maximum value in their domain. If these functions have domains which are the real numbers then their minima are not defined. Exponentials in which b is > 1 are strictly increasing and when $b < 1$ (but > 0) they are strictly decreasing. The same will be true of logarithmic functions.

Linear functions that have positive slope are increasing; those with negative slope are decreasing. To evaluate a function which is not linear we may place lines tangent to points on the function and examine the slopes of these tangents. If the slopes are always positive, then the function is increasing; if the slopes are always negative, then the function is decreasing.

319

Some non-linear functions have points where a tangent to the curve will have slope = 0. These are called <u>critical points</u>. If the directionality of a function changes within some interval in its domain, i.e., if it goes from decreasing to increasing or increasing to decreasing, then the slope of tangents to the curve must go from positive to negative or negative to positive, respectively. To do so they must at some point have slope = 0, i.e., they must have a critical point in this same interval. If a switch in directionality occurs in an interval then the critical point is a relative minimum or relative maximum depending on whether the change in directionality goes from decreasing to increasing or from increasing to decreasing.



If there are no other relative minima or relative maxima then the critical point identifies an <u>absolute minimum</u> or <u>absolute maximum</u>.

Homework
Prerequisite Inventory, Unit 3

1. On a cartesian coordinate system graph the function

$$f(x) = |x|+10$$

for the values of $x=-10$ through $+10$. Label the point $(0,10)$.

2. On a cartesian coordinate system graph the function

$$f(x) = \frac{1}{x}$$

for the values of $x=-10$ through $+10$. Label the points $(-1,-1)$ and $(1,1)$.

3. Make a plot of the function

$$f(x) = \sqrt{x}$$

for integer values of x from 0 to 10.

4. Graph $x=4y^2$. Is this a function of x?

5. What are the domain and range of the following functions?

   a. $f(x) = |2x-10|$

   b. $f(x) = \frac{16}{x^2}$

   c. $f(x) = \sqrt{x-5}$

6. Locate and label the following points on a rectangular coordinate system and give the quadrants in which each point lies.

   $(2,7)$, $(8,-3)$, $(-\frac{1}{2},-2)$, $(0,0)$

7. The following table indicates the number of widgets that are purchased each week for four weeks and the price widgets sold for in each week. Plot price as a function of quantity sold. Sketch in the wave. What is the name for this type of curve (in Economic jargon!)? What is the relationship between widget price and widget sales? Is this an increasing or decreasing function?

week	1	2	3	4
price/widget	20	10	5	4
quantity/week sold	5	10	20	25

351

8. Find the slope of the straight line which passes through the following points.

   a. (5,2), (7,5)

   b. (-2,3), (3,-1)

   c. (-2,4), (-2,8)

   d. (5,-2), (4,-2)

9. Give in functional form the equations for linear functions that have the indicated properties.

   a. passes through (1,2) with slope 6.

   b. passes through (-2,5) with slope -1/4.

   c. passes through (1,4) and (8,7).

   d. passes through (3,-1) and (-2,-9).

10. For the following give the slope and y intercept of the line.

   a. $x = -2y + 4$

   b. $4x + 9y - 5 = 0$

   c. $\frac{3}{4}x = \frac{7}{3}y + \frac{1}{4}$

   d. $\frac{x}{2} - \frac{y}{3} = -4$

11. The forecasted population, P_F, of a city is given by

$$P_F = P_c e^{at}$$

where P_c is the current population, a is a constant, e is Euler's number and t is the number of years after 1976.

If the city's current population is 100,000 give the forecasted population in 1996 (assume a = .05). What interpretation can you give to a?

12. If $6y = e^{2r}$ express r as a function of y.

13. Solve the following equation for x.

$$x + 1 = \log_4 16$$

14. Simplify the expression $10^{2 \log x}$

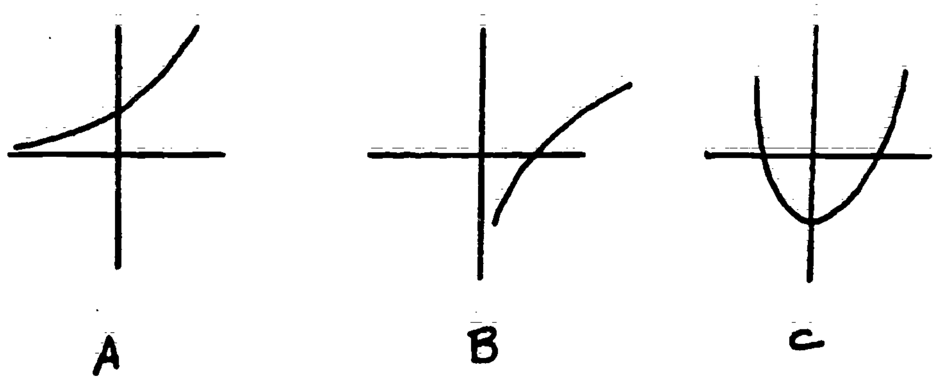
15. Solve the equation $y = e^{(\ln 3 + 2 \ln 4)}$

16. The demand equation for a product is defined by

$$p=12^{1-.1x}$$

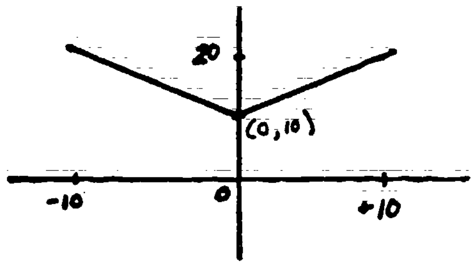
Express x as a function of p.

17. The following graphs are typical of what kind of function?

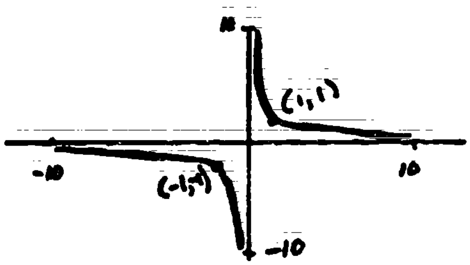


Homework Solutions
Prerequisite Inventory, Unit 3

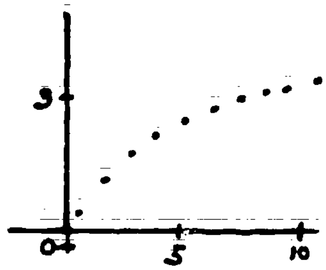
1.



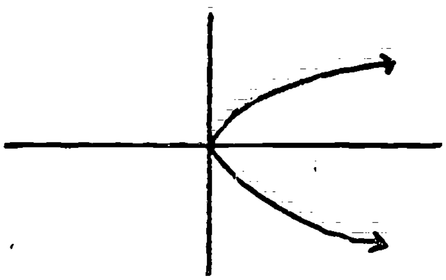
2.



3.



4.



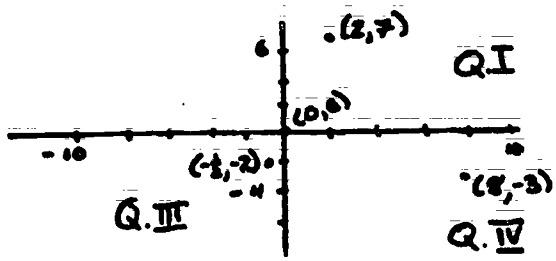
No.

5. a. Domain : all reals
Range : all nonnegative reals

b. Domain : all non-zero reals
Range : all positive reals

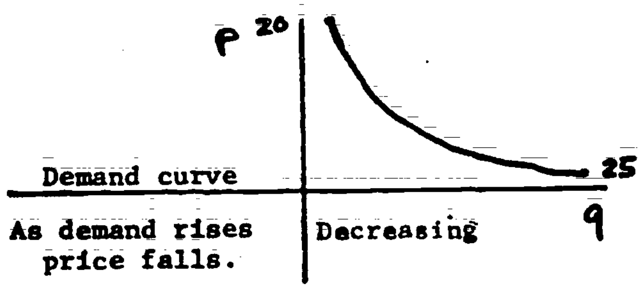
c. Domain : all reals ≥ 5
Range : all nonnegative reals

6.



(0,0) does not lie in a quadrant.

7.



8. a. 3/2

   b. -4/5

   c. not defined

   d. 0

9. a. $f(x) = 6x - 4$

   b. $f(x) = -\frac{x}{4} + \frac{9}{2}$

   c. $f(x) = \frac{3}{7}x + \frac{25}{7}$

   d. $f(x) = \frac{8}{5}x - \frac{29}{5}$

10. a. $-\frac{1}{2}$, 2

    b. $-\frac{4}{9}$, $\frac{5}{9}$

    c. 9/28, -3/28

    d. 3/2, 12

355

11. $\approx 271{,}828$ a is the yearly percentage increase

12. $r = \frac{1}{2}\log_e(6y)$

13. $x = 1$

14. x^2

15. 48

16. $x = f(p) = 10(1 - \frac{\log p}{1 \cdot 0792})$ or $10(1 - \frac{\log p}{\log 12})$

17. a. exponential

    b. logarithmic

    c. parabolic

QMPM

Lecture 3-0. Introduction to Unit 3

Introduction to Unit 3, Analysis of (X,Y) data.

Lecture Content:

Introduction to the objectives, problem, and notation of Unit 3

Main Topics:

1. Specific Introduction to the Objectives of Unit 3
2. Presentation of General Problem of Unit 3
3. Notation for Unit 3

Note to Instructor:

Unit 3's primary pedogogic role is that of a precursor to regression. The lectures are set up to provide students with an intuitive grasp of fitting lines to (X,Y) data so that the application of specific fitting algorithms do not seem like arbitrary operations. The essential notion remains that of finding a model that fits the data and quantifies the effect on Y of movement along the X dimension. Exploratory procedures learned in earlier units are applied throughout.

Topic 1. Specific Introduction to the Objectives of Unit 3

I. Questions to be answered in Unit 3

1. What is an ordered multiple batch?

   a. A collection of batches related in some quantitative way (as opposed to unordered multiple batches which are <u>qualitatively</u> related) (1)

   b. The ordered relation between batches is defined on some scale and used in the analysis

   c. Examples: life expectancies for countries, classified by per capita income of country; number of vehicles for transit systems, classified by the population served by system

2. What analyses can be done on an ordered collection of batches?

   a. How can we best examine the batches by using the ordered scale of the batches

   b. How can we summarize the information in the batches and the relationship between each batch, and the value for the batch on the scale

   c. How can we transform both the batches and the scale relation

3. What is an (X_i, Y_i) paired observational batch?

   a. Data set consisting of two batches of equal size (2)

   b. The ith observation of the first batch, called X_i, is related to the ith observation of the second batch, Y_i

   c. We thus have a batch of paired observations, or ordered pairs (X_i, Y_i)

   d. Examples: IQ scores of twins; achievement pretest score (X) and fall final exam score (Y) for each member of this class

4. What analyses can be done on a batch of paired observations?

   a. How can we best examine the scatterplot of (X_i, Y_i) values

b. How can we best summarize the relationship between the X variable and Y variable

c. How do we determine whether a transformation of either X or Y or both would improve the summarization

5. What is a batch of time series data?

a. (X,Y) paired data set, where X is time (months, years, decades etc.) (3)

b. One Y will be associated with each X i.e., impossible to have two or more observations at a single point in time

c. Examples: Gross National Product of the U.S. for the years 1940-1976; daily reported cases of swine flu, January-September 1976

6. What analyses can be done on time series data?

a. How can we smooth the data to remove irregularities

b. How and when can we extrapolate beyond the current time range, and interpolate between two adjacent time points

c. What can we say about any periodicities within the time series

II. Skills to be mastered in Unit 3 (4)

1. Perceiving and analyzing ordered multiple batches

2. Looking at scatterplots of (X,Y) data

3. Summarizing scatterplots by fitting lines

4. Smoothing the irregularities in time series data

5. Extrapolating, interpolating, and studying the periodicities of time series data

Topic 2. Introduction to the Problems of Unit 3

I. What is an ordered multiple batch?

1. Example: Average net interest cost, in percent, for bond sales for public schools, by Bond Moody rating, for various years (5)

   a. Relation: Percent interest for bonds, issued for public schools

      Quantitative aspect: Bonds classified by their Moody Rating, Aaa-Ba

2. The Quantitative ordering is extremely important. We can associate for each batch in the collection a value X_i on the ordered scale.

II. How can we best analyze the batches? (6)

1. Obvious questions:

   a. Minima

   b. Maxima

   c. Spreads

   d. Medians

   e. Shape

   f. Units

2. Subtle questions:

   a. What is a good typical value for each batch?

   b. Conditional on being in a specific batch, what is the typical value for the batch? We call these "conditional typical values"

   c. How are these conditional typicals used to summarize the entire batch?

III. What is a batch of (X,Y) data?

Example: Number of vehicles and vehicle miles for transit systems serving populations over 1 million people, in 1971 (7)

a. X variable: Number of Vehicles

b. Y variable: Transit system vehicle miles, in millions

c. 11 observations, 1 per transit system

IV. How can we best summarize this batch of paired observations?

1. What can we learn from looking at the (X,Y) scatterplot? (8)

2. Do the data have a linear point cloud?

3. Or does the point cloud have a peculiar shape?

4. How do we effectively summarize linear point clouds?

5. Can we transform nonlinear point clouds to make them more linear, and hence more easily summarized?

V. What is a batch of time series data?

<u>Example</u>: Total expenses for Community Hospitals (9)

a. X time variable: Year, 1950, 1955, 1960-1972

b. Y variable: Costs (in million $)

c. 15 time points

VI. How can we better understand this time series?

1. What curve is traced by the time plot?

2. What curve remains after the data have been smoothed?

3. Can we extrapolate beyond the current range? What will (10) expenses look like in 1975? 1980? What were they in 1940?

4. Can we interpolate between two consecutive data points? What were expenses in 1953? 1959?

5. Are there any periodicities in the data set?

VII. Conclusion: We need specific tools to use in analyses of each of these three data forms.

381

Topic 3. Introduction to the Notation of Unit 3

I. Ordered Batches

1. Capital letter ("Y") denotes data set
2. First subscript (Y_i) denotes specific batch
3. X_i denotes the value on the quantitative scale associated with batch Y_i
4. Second subscript (Y_{ij}) denotes specific observation in a specific batch

II. (X,Y) paired observation

1. Capital letters (X and Y) denote each batch. Pairing of batches is an underlying concept of multiple regression, in which one dependent variable (Y) is explained by (paired with) several independent variables (X's)
2. A specific ordered pair is denoted by (X_i, Y_i).

III. Time series data

1. Same notation as paired observations

Lecture 3-0
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Topic 1		
Section I		
1.a	1	Ordered multiple batch
3.a	2	(X,Y) paired observational data
5.a	3	Time series data
Section II		
1.	4	Topics for Unit 3
Topic 2		
Section I		
1.	5	Average bond interest costs
Section II		
1.	6	Plot of average school bond interest costs
Section III		
1.	7	Vehicles and vehicle miles for transit system
Section IV		
1.	8	Plot of vehicles and vehicle miles
Section V		
1.	9	Total expenses for community hospitals
Section VI		
3.	10	Plot of hospital expenses

Ordered Multiple Batch

An ORDERED multiple batch of data is a collection of two or more batches related in a quantitative way.

In Unit 3:

We learn to analyze multiple batches of data that are ordered by examining "conditional typical values" -- a typical value for each batch that is representative of the batch.

Since the batches are ordered on a quantitative scale, we study how these "conditional typicals" vary with the scale.

(3-0)

(X_i, Y_i) Paired Observational Data

__*Paired Observational Data*__ *is a data set consisting of two batches, such that the ith observation of the first batch is related to the ith observation of the second batch. We label the ith observation of the first batch X_i, and the ith observation of the second batch Y_i and write the paired observation (X_i, Y_i).*

In Unit 3:

We plot the (X_i, Y_i) data, examine the resulting scatterplot, and summarize the scatterplot with conditional typical values, transforming if necessary.

[3]

Time Series Data

Time Series Data is a data set consisting of two batches. The X batch is a time scale (days, months, years, etc.) and we have one Y data value associated with each X value.

In Unit 3:

We plot the time series data, smooth out unnecessary irregularities, extrapolate beyond the range of the data, interpolate between data observations, and study any periodicities, if present.

(3-0)

Topics for Unit 3:

1. Perceiving and analyzing ordered multiple batches.

2. Looking at Scatterplots of (x,y) data.

3. Summarizing (x,y) scatterplots by fitting lines.

4. Smoothing the irregularities in time series data.

5. Extrapolating, Interpolating, and studying the periodicities of time series data.

[5]

Average Net Interest Cost, in Percent, for Bond Sales for Public Schools.

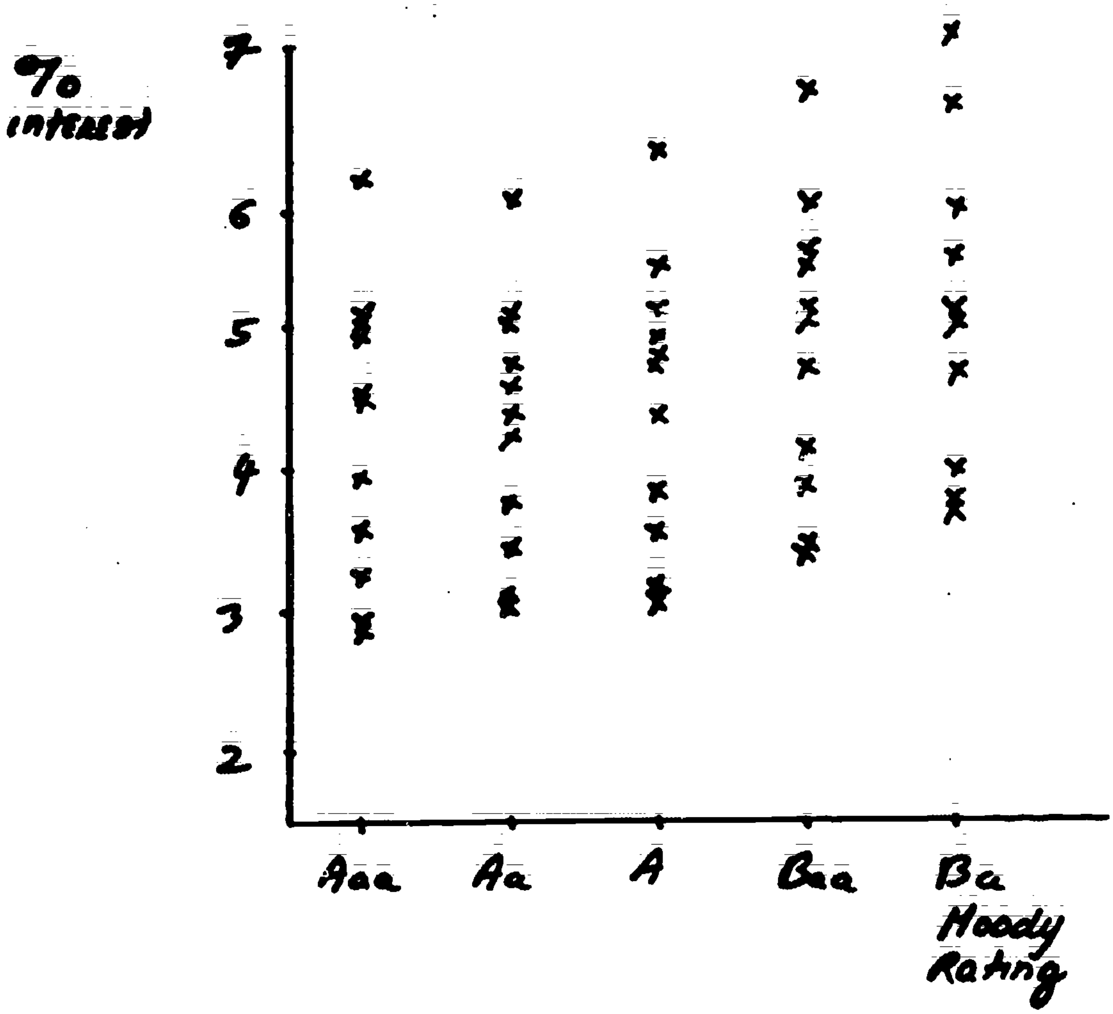
Moody Rating

Aaa	Aa	A	Baa	Ba
2.88	3.07	3.17	3.43	3.80
2.93	3.11	3.16	3.44	3.76
3.26	3.48	3.56	3.86	4.01
3.56	3.79	3.86	4.17	4.68
3.96	4.23	4.40	4.74	5.05
5.05	4.41	4.73	5.07	5.53
6.04	5.90	6.28	6.71	7.09
5.10	5.02	5.14	5.93	6.60
4.54	4.60	4.92	5.48	5.84
4.53	4.77	4.79	5.18	6.17
4.97	5.04	5.48	5.59	

Each row is a year (1964 - 1974)

[6]

Average Net Interest Cost, in %, for Public School Bond Sales; Points are for Years 1964-1974.

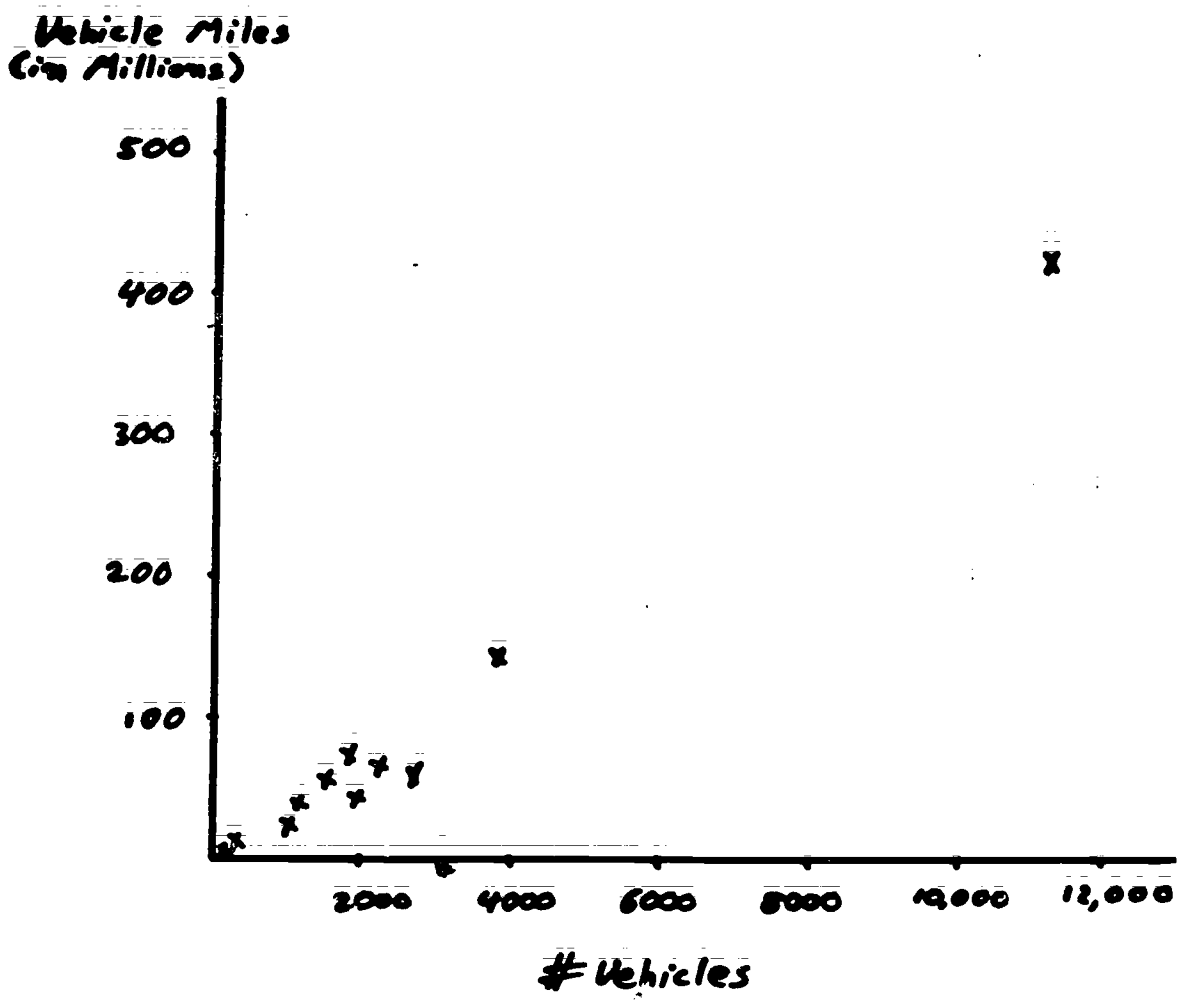


[7]

Numbers of Vehicles and Vehicle Miles for Transit Systems Serving Populations Over 1 Million People; 1971.

Transit System	No. of Vehicles	Transit System Vehicle Miles (in Millions)
Boston	1983	43.5
Chicago	3824	146.2
Cleveland	1011	25.2
Detroit	1171	35.1
Los Angeles	1511	58.8
Montreal	2221	64.5
New York (NYCTA)	11270	428.5
New York (PATH)	252	9.7
Philadelphia (PATCO)	75	3.7
Philadelphia (SEPTA)	2793	57.6
Toronto	1886	72.4

Graph of Numbers of Vehicles and Vehicle Miles for Transit Systems Serving Populations over 1 million; 1971.



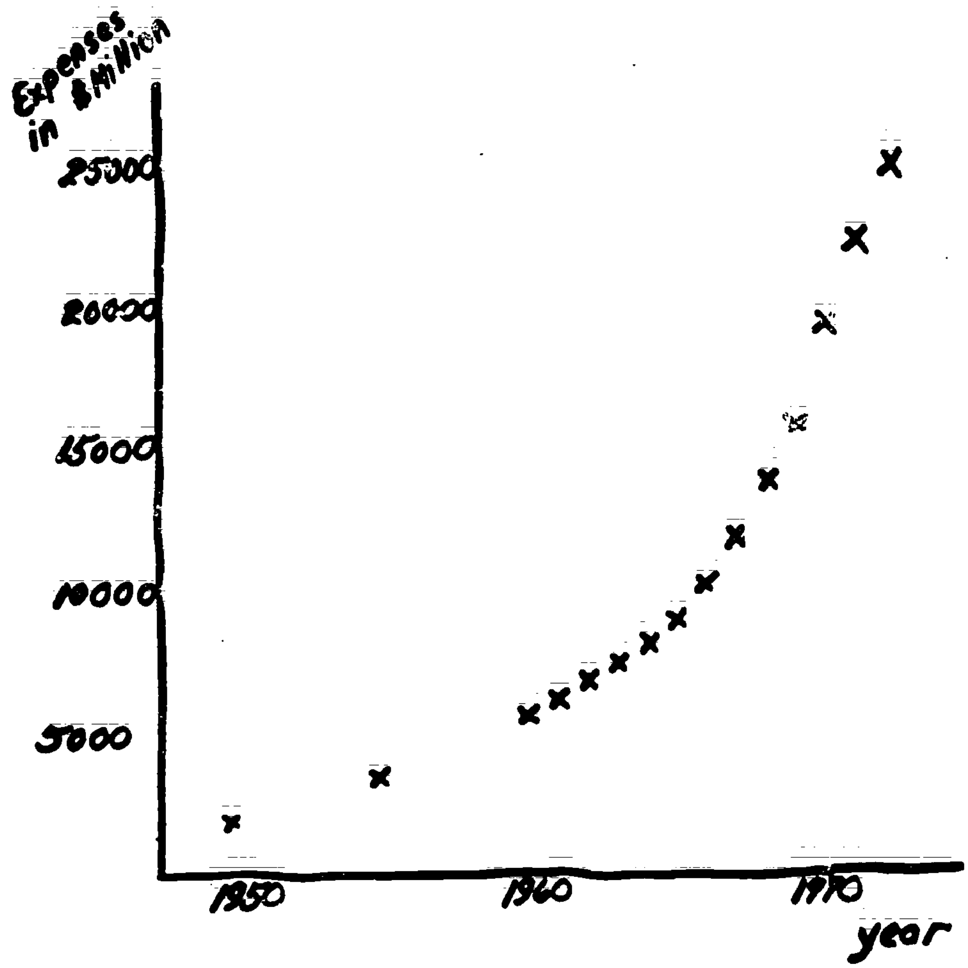
[9]

Total Expenses for Community Hospitals.

Year	Amount (in Millions)
1950	$ 2120.
1955	3434.
1960	5617.
1961	6250.
1962	6841.
1963	7532.
1964	8349.
1965	9147.
1966	10,275.
1967	12,081.
1968	14,162.
1969	16,613.
1970	19,560.
1971	22,400.
1972	25,549.

(3-0)

Total Expenses for Community Hospitals



(3-0)

Lecture 3-1. Analysis of Ordered Batches

Analysis of Ordered Batches: The perception, display, and summarization of a collection of ordered batches

Lecture Content: (1)

1. Discuss the techniques for displaying two or more batches simultaneously

2. Introduce new measures for the summarization of the relationship between the multiple batch and the ordered scale

Main Topics:

1. Display of several batches ordered on some scale

2. Introduction of "conditonal typical values" to summarize the batches

3. Discussion of the effectiveness of conditional typical values in summarization

Topic 1. Display of several batches ordered on some scale

I. Basic Issue: Comparison of ordered multiple batches, using the natural scale

1. We know how to compare and transform unordered multiple batches

2. We are interested in analyzing ordered batches in an effective, consistent, and reliable manner

3. We need techniques to examine the batches, using the scale associated with the collection

II. Problem: Can we simply use the comparison tools of Unit 2 for unordered batches?

1. Specific questions to be answered are similar to those for unordered batches

2. What do we do about the ordered nature of the batches?

3. As usual, a condensation of the information in the batches should follow from an organization of the collection

4. We organize the batches as in Unit 2, but our condensation utilizes the natural scale

III. Solution: Organize Parallel Schematic Plots of the batches with positioning determined by the scale

IV. Method

1. We familiarize ourselves again with the definition of an ordered multiple batch: a collection of two or more batches that are related in a quantitative way

2. We look at some hypothetical examples: life expectancies for countries classified by per capita income; number of vehicles per transit system, classified by population served

3. Here is a real <u>example</u> that we shall examine:

a. Number of Live Births, classified by the age of the mother at the time of the birth (2a)

b. Batch observations are various years, 1950-1967 (2b)

4. We plot the observations on an (X,Y) plane (3)

   a. X_{ij} = Value on scale for batch i

   b. Y_{ij} = Observation j in batch i

   c. X_{ij} is constant over all j

   d. Scatterplot for Live Birth data

      i. X_{ij} not well defined--given as range; e.g., 24-29 years

      ii. Let X_{ij} be the midpoint of each interval; e.g., 25-29 years interval has X_{ij} = 27

      iii. X for Over 45 and Under 15? Arbitrary; use 47 and 13

5. Next draw a schematic plot for each batch--centered at (4) the correct X for each batch

6. Width of box = width of interval associated with the corresponding X

7. Thus we have organized each batch, using the position of each batch on the X scale

8. "Ordered" Parallel Schematics with CMU-DAP

   a. Unfortunately the plots cannot be positioned properly

   b. Treat each batch separately, and cut and paste each schematic on a piece of graph paper, in the proper place

9. Plotting the raw data with CMU-DAP

   a. Create a X_{ij} data file, constant for a given i, to PLOT against the Y_{ij} multiple batch values.

Topic 2. Conditional Typical Values to summarize the batches

I. Basic Issue: Once organized into parallel schematics, how can we summarize each batch

1. The "pattern" of the schematic display is very important in the analysis

a. Do the plots increase? If so, is the increase roughly linear, or is the functional relation of higher degree

b. Do the plots decrease? Again, what is the functional form of the decrease?

2. We want to pick one value from each batch to study further the pattern of the batches

II. Problem: What value do we use for our summarization?

1. The value should be representative

2. If the spread of each batch was zero, we would have no problem in choosing a set of typical values

III. Solution: Use medians, our good friend!

1. The typical value for Y_{ij} depends on the batch X_{ij} value

2. We compute typical values of Y_{ij}, "Conditional" on being located in batch i--"conditional typicals"

3. Conditional typical value of Y_{ij}, given scale value X_{ij} = Median of batch i = median $(Y_{i1}, Y_{i2}, \ldots, Y_{in_i})$ where n_i = # observations in batch i.

4. For our live births example--here are the conditional typicals (5)

5. We can locate each conditional typical within each batch on the (X,Y) scatterplot, and connect them

6. We study the form of the line segments on this connected plot (6)

7. Hinges also help in our study--we can locate the hinges, and connect them (7)

8. Specific question: Do the line segments connecting the conditional typicals form a line?

377

9. Secondary question: Are the spreads of the batches constant?

10. In a later lecture, we transform both X and Y to

    a. Promote linearity of the conditional typicals

    b. Equalize spread within the batches

11. Conditional Typicals constructed with CMU-DAP

    a. Merely use SUMMARY to find medians, and draw them in on your scatterplot

pic 3. Effectiveness of Conditional Typical Values in Summarization

I. Basic Issue: Assessing how well conditional typicals describe the data set

1. The breaking up of data into Fit + Residual has been discussed
2. For ordered batches: Y_{ij} data value = Conditional Typical for batch i + Residual
3. Fit = Conditional Typical for batch i
4. How much is left after we subtract the fit from each data value?

II. Problem: How do we analyze the batch of Residuals from the fit

1. Residuals should not be large relative to the fit
2. The batch of residuals should be
   a. Symmetric
   b. No obvious outliers
   c. Close to well-behaved

III. Solution: Analyze the residuals as a single batch using the tools of Unit 1.

IV. Methods

1. Back to our example--residuals from conditional typicals for live birth data (8)
2. Stem-and-Leaf Display of Residuals. Note large number of zeros, and a few outliers (9)
3. Schematic plot and number summary very helpful--note symmetry and outliers (10)
4. Another example: Average net interest costs, in percent, for bond sales for public schools. Entries are for years, 1964-1974 (11)
5. Find conditional typicals, plot the values, and find residuals (12) (13)

Lecture 3-1
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 3-1 Outline
Topic 1		
Section IV		
3.	2a 2b	Number of Live Births by age of mother, 1950-67
4.	3	Plot of Live Birth data
5.	4	Parallel Schematic plot of live birth data
Topic 2		
Section III		
4.	5	Conditional typical values for live births
5.	6	Conditional typical values connected
7.	7	Hinges and conditional typicals connected
Topic 3		
Section IV		
1.	8	Residuals from fits for live birth data
2.	9	Stem-and-leaf of residuals
3.	10	Schematic plot of residuals
4.	11	Average interest rates for school bonds
5.	12	Conditional typicals for school bond interest rates
5.	13	Conditional typicals connected

Lecture 3-1

Analysis of Ordered Batches: Perceiving, Displaying and Summarizing a collection of ordered batches.

Lecture Content:

1) Discuss the techniques for displaying two or more ordered batches simultaneously.

2) Introduce new measures for summarizing the relationship among the batches, using the ordered scale.

Main Topics:

1) Display several batches ordered on some scale.

2) Introduce "conditional typical values" to summarise the batches.

3) Discuss how well the conditional typicals convey the characteristics of each batch.

Number of Live Births, by Age of Mother for 1950-1967.

Age of Mother, in years.

	Under 15	15 - 19	20 - 24	25 - 29
1950	5,021	419,535	1,131,234	1,021,902
1951	5,086	443,872	1,198,966	1,072,374
1952	5,032	438,046	1,212,010	1,104,012
1953	5,316	455,878	1,220,532	1,110,768
1954	6,058	477,880	1,257,104	1,122,050
1955	5,883	484,097	1,273,908	1,119,279
1956	6,356	520,422	1,325,444	1,131,3[illegible]
1957	6,960	550,212	1,361,396	1,140,[illegible]
1958	6,648	554,184	1,367,826	1,108,[illegible]
1959	6,776	571,048	1,406,200	1,09[illegible]
1960	6,780	586,966	1,426,912	1,092,81[illegible]
1961	7,462	601,720	1,445,054	1,081,706
1962	7,340	600,298	1,444,978	1,045,086
1963	7,594	586,454	1,453,790	1,023,942
1964	7,826	585,710	1,439,486	1,007,362
1965	7,768	590,894	1,337,350	925,7[illegible]2
1966	8,128	621,426	1,297,990	872,786
1967	8,593	576,445	1,310,598	867,426

(3-1)

[2b]

Number of Live Births, by Age of Mother for 1950-1967

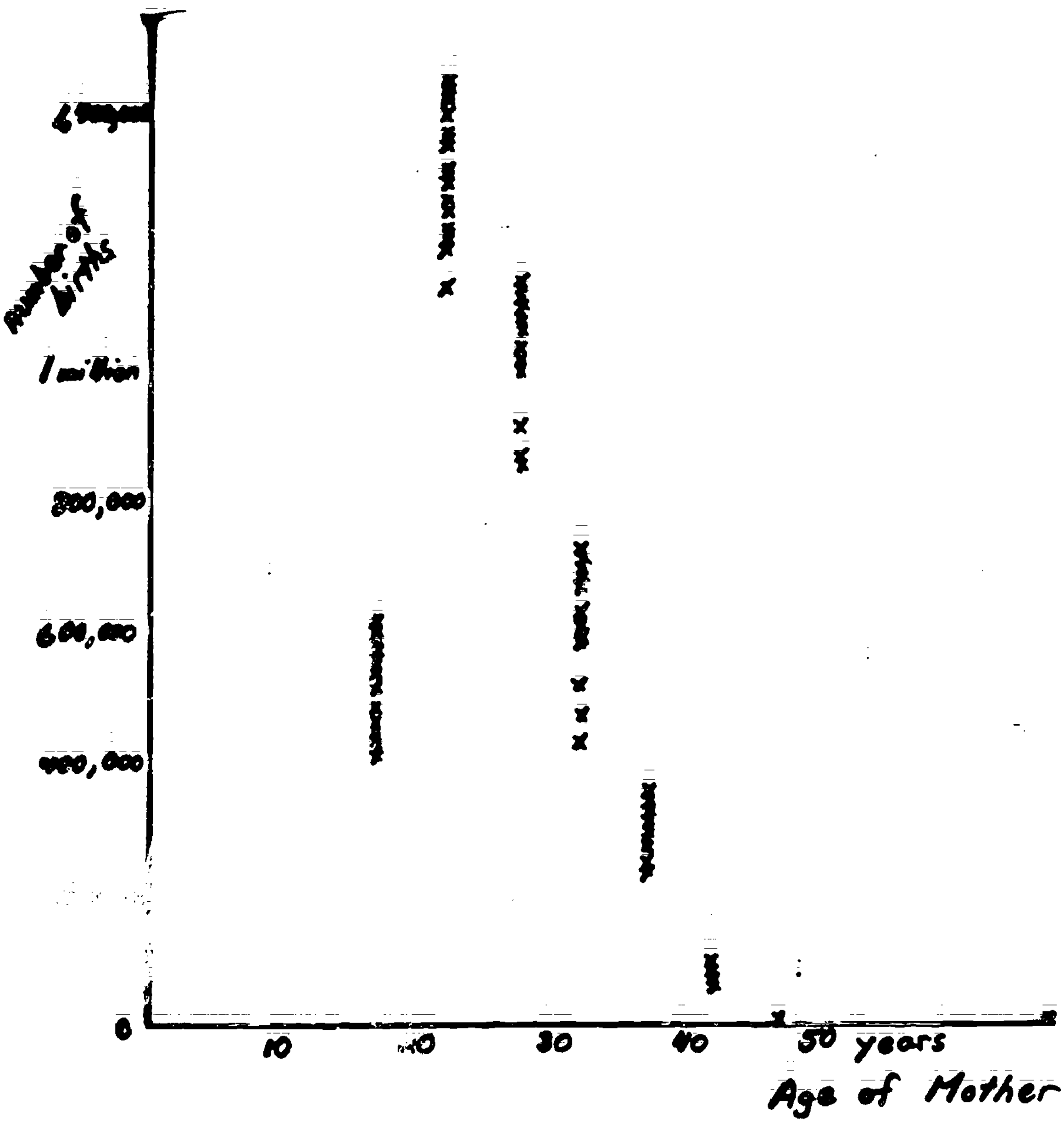
Age of Mother, in years

	30-34	35-39	40-44	45 and over
1950	677,821	293,440	79,804	4,830
1951	637,238	304,878	78,224	4,932
1952	679,220	318,338	80,494	5,170
1953	691,090	326,102	83,290	5,004
1954	720,820	337,078	86,766	5,106
1955	[illegible],277	345,305	87,587	5,111
1956	[illegible],790	355,158	89,734	5,140
1957	730,218	365,298	90,808	5,272
1958	721,680	[illegible]	[illegible],702	5,116
1959	[illegible],826	[illegible],120	89,626	5,246
1960	687,922	359,904	92,564	5,182
1961	677,264	355,700	94,114	5,256
1962	638,382	334,708	91,490	5,080
1963	610,186	322,182	88,982	4,930
1964	588,006	309,814	87,626	4,670
1965	529,376	293,908	81,716	4,614
1966	474,542	252,526	79,440	4,436
1967	438,373	227,323	67,053	4,158

383

(3-1)

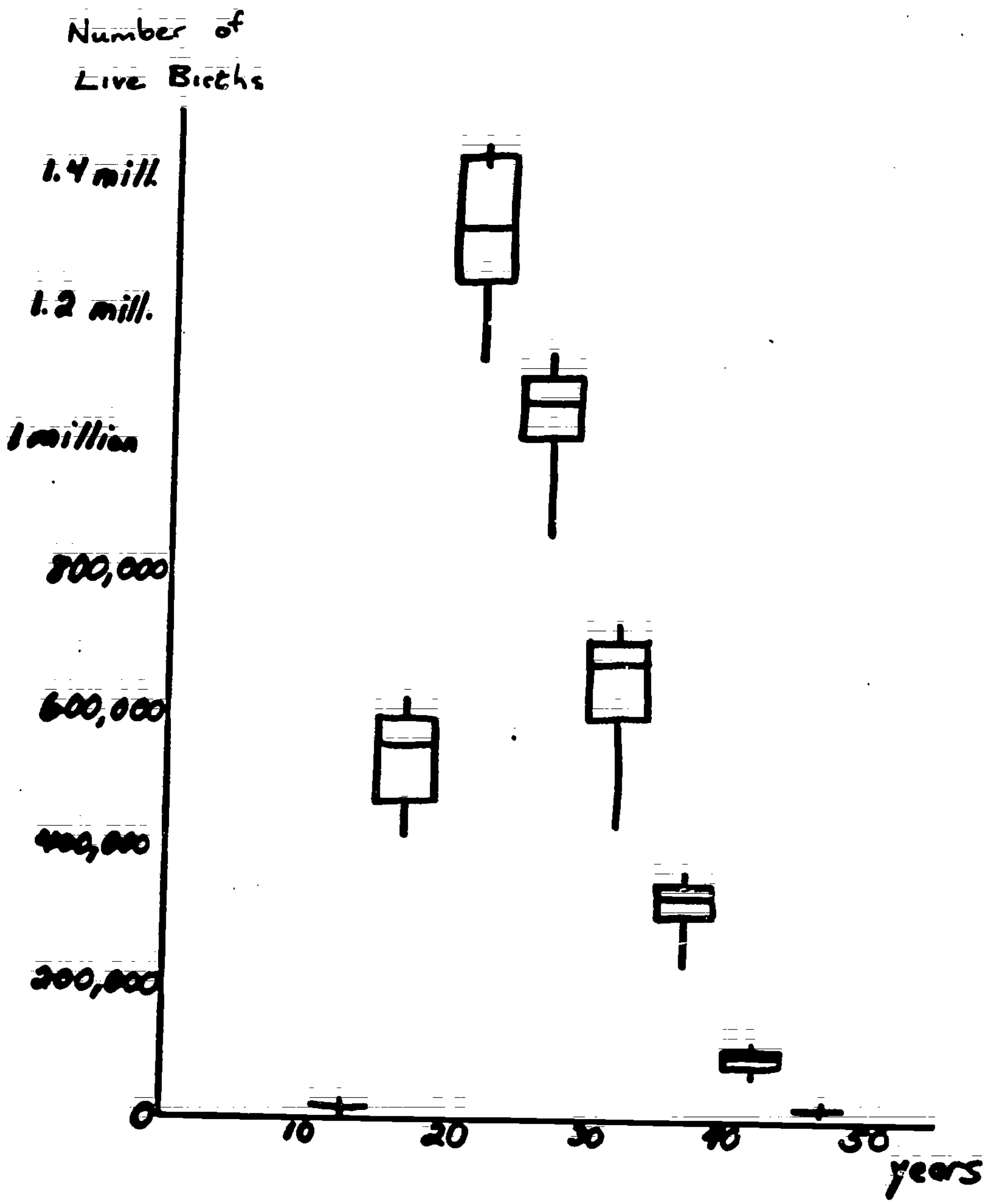
Number of Live Births by Age of Mother, 1950-1967



(3-1)

[4]

Parallel Schematic plot of Live Births by Age of Mother



385

(9-1)

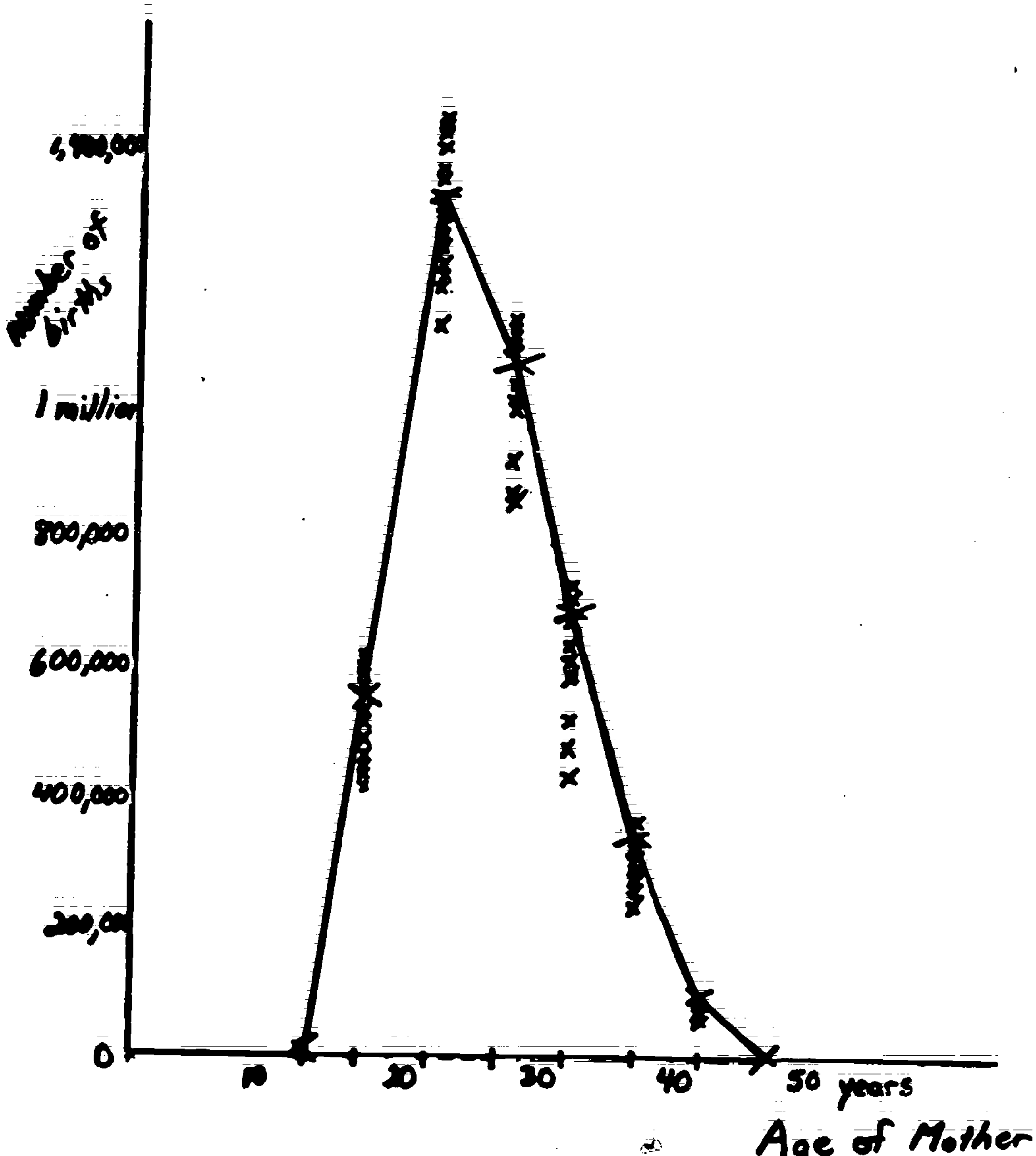
Live Births by Age of Mother
Conditional Typical Values, or "Fits" for Each Age Class

"X" Age Class	Typical value of "Y", Given "X"
Under 15	6,700 births
15-19	560,000 births
20-24	1,310,000 births
25-29	1,065,000 births
30-34	680,000 births
35-39	330,000 births
40-44	85,000 births
Over 45	5,000 births

(3-1)

[6]

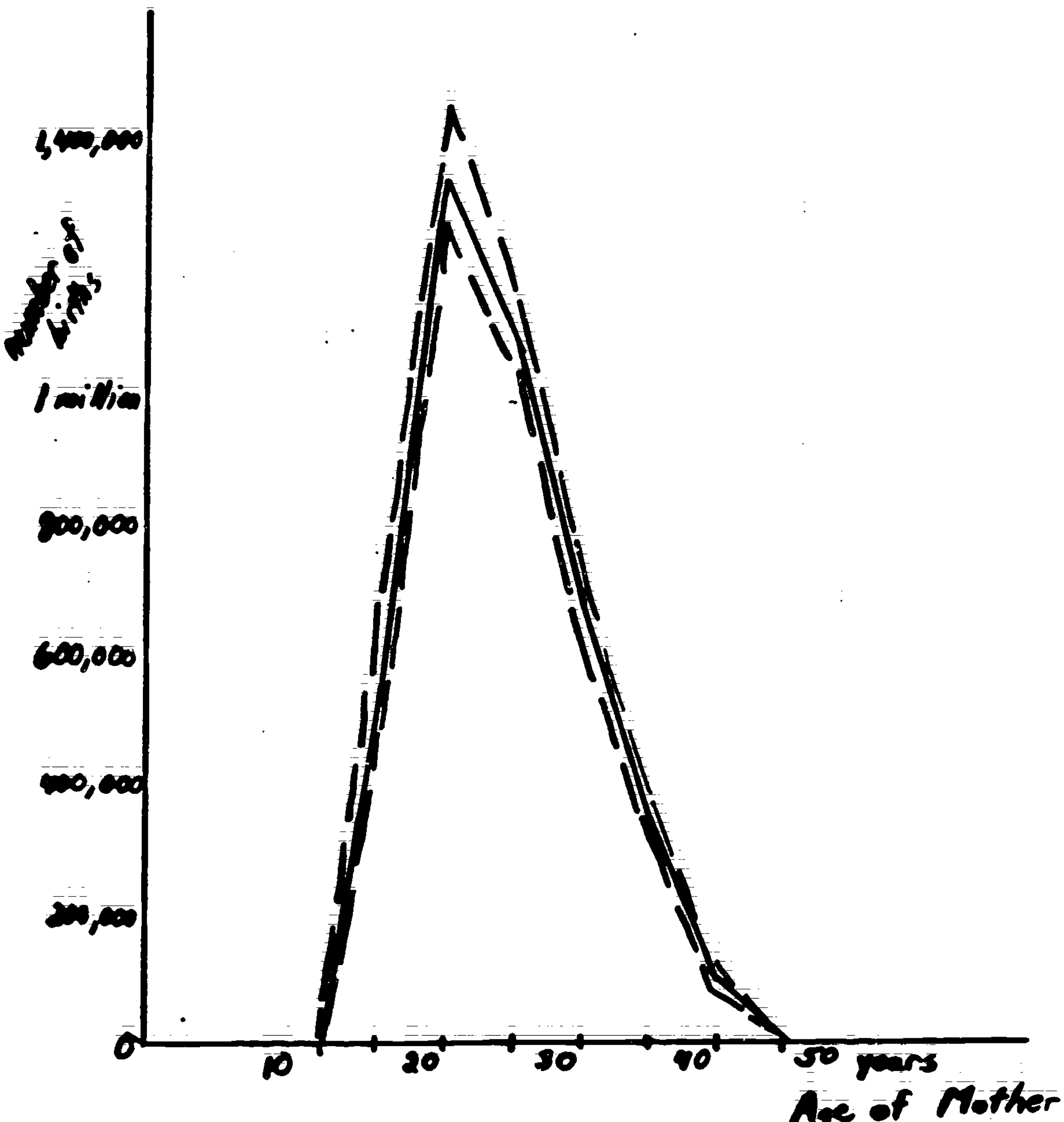
Number of Live Births by Age of Mother, 1950-1967
Conditional Typical Values located and connected



(3-1)

387

Number of Live Births by Age of Mother, 1950-1967
Medians and Hinges Connected.



(B-1)

[8]

Live Births by Age of Mother
Residuals from Conditional Typicals

Residual = Data Value − Conditional Typical

(Residuals unit = 10 thousand births)

0	-140	-180	-45	-90	-40	-10	0
0	-120	-140	5	-50	-25	-5	0
0	-120	-120	40	0	-10	-5	0
0	-100	-90	50	10	-5	0	0
0	-80	-50	55	90	10	0	0
0	-80	-40	50	40	15	0	0
0	-40	15	65	45	25	5	0
0	-10	50	75	50	35	5	0
0	-5	55	40	30	30	5	0
0	10	95	30	20	30	5	0
0	25	115	30	10	30	5	0
0	40	135	15	0	25	10	0
0	40	135	-20	-40	5	5	0
0	30	145	-40	-70	-10	0	0
0	25	130	-60	-95	-20	0	0
0	30	20	-140	-150	-50	-5	0
0	60	-20	-190	-190	-80	-10	0
0	35	0	-200	-250	-100	-15	0

[9]

Stem-and-Leaf Display of Residuals from Conditional Typicals for Live Birth Data.

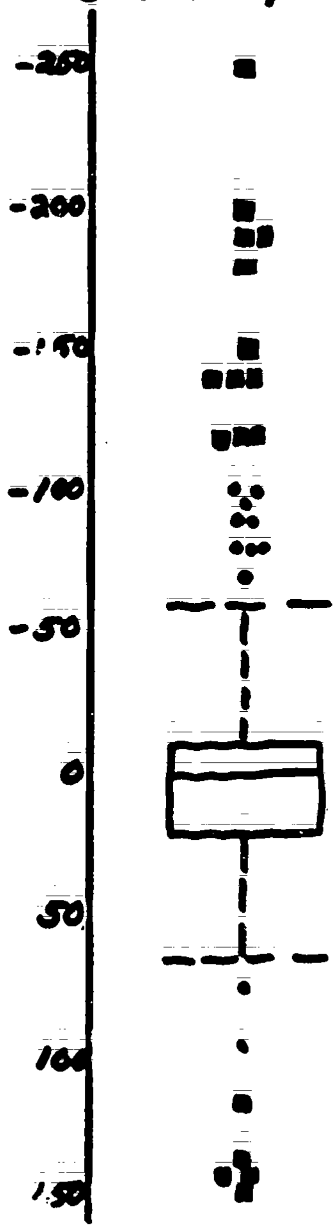
Unit = 10,000 Live Births

```
 LO | -250,-200,-190,-190,-180,-150
-14 | 000
-13 |
-12 | 000
-11 |
-10 | 00
 -9 | 500
 -8 | 000
 -7 | 0
 -6 | 0
 -5 | 000
 -4 | 500000
 -3 |
 -2 | 5000
 -1 | 500000
 -0 | 55555
  0 | 0000000000000000000000000000000000000000000000000055555555
  1 | 00000555
  2 | 005555
  3 | 00000000055
  4 | 000005
  5 | 000055
  6 | 05
  7 | 5
  8 |
  9 | 5
 Hi | 115,130,135,135,145
```

(3-1)

[10]

Schematic plot and Number Summary of Residuals for Live Birth data.



#144

0		midspread
-10	20	30
-150	-250	

$\boxed{45}$

-55	65	adjacent
9 values	2 values	-60
-100	110	65
12 values	5 values	

[11]

Average Net Interest Cost, in Percent, for Bond Sales for Public Schools, Medians circled.

Moody Rating

Aaa	Aa	A	Baa	Ba
2.88	3.07	3.17	3.43	3.80
2.93	3.11	3.16	3.44	3.76
3.26	3.48	3.56	3.86	4.01
3.56	3.79	3.86	4.17	4.68
3.96	4.23	4.40	4.74	5.05
5.05	4.41	4.73	5.07	5.53
6.04	5.90	6.28	6.71	7.09
5.10	5.02	5.14	5.93	6.60
4.54	4.60	4.92	5.48	5.84
4.53	4.77	4.79	5.18	5.17
4.97	5.04	5.48	5.59	

Rows represent years, 1964-1974

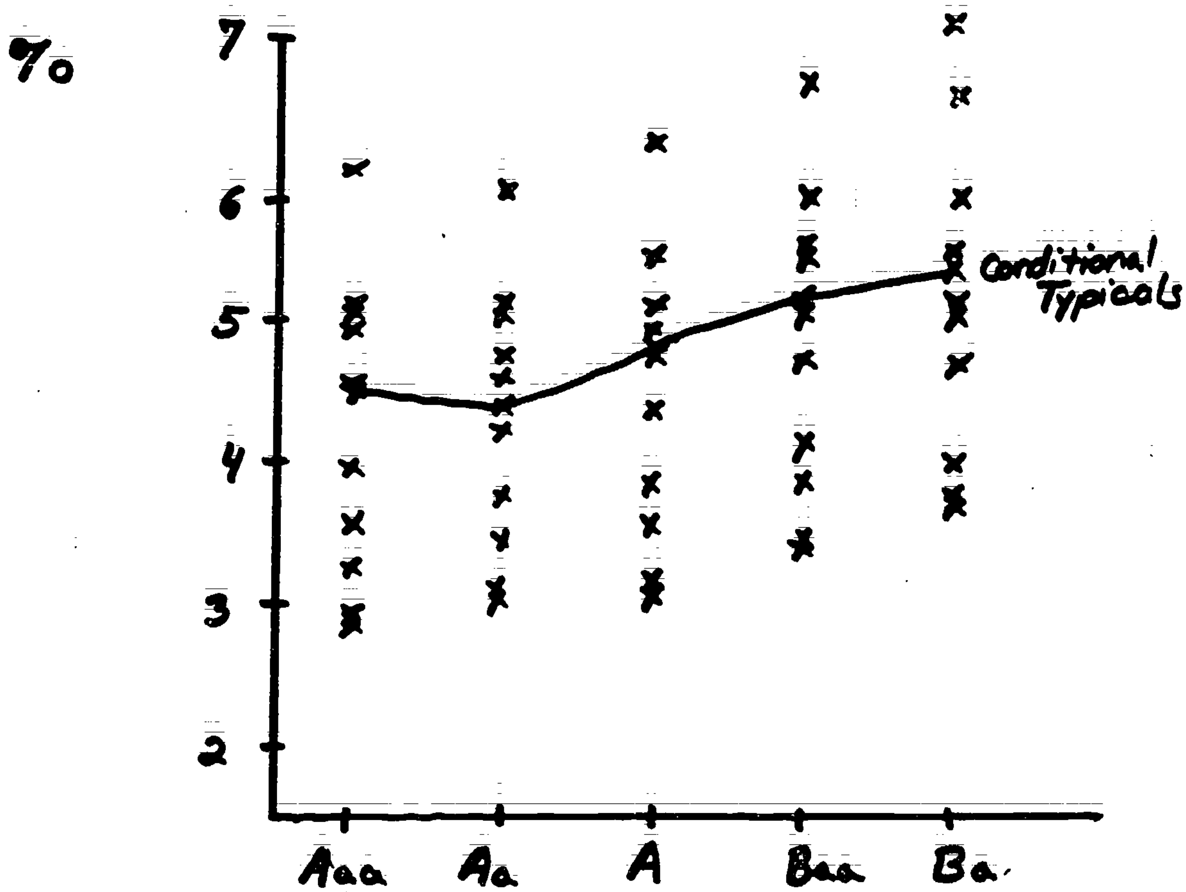
(3-1)

Average Net Interest Costs for School Bonds

Conditional Typical Values

Moody Rating "X"	Typical Value of "Y", Given "X"
Aaa	4.53
Aa	4.41
A	4.73
Baa	5.07
Ba	5.11

Average Net Interest Cost, in %, for Public School Bond Sales; Entries are for Years 1964-1974; Conditional Typical Values Given.



Lecture 3-2. Looking at (X,Y) data

Looking at (X,Y) data: Analysis by reorganization of (X,Y) paired observation data (1)

Lecture Content:

1. Discussion of how (X,Y) paired observation data may be viewed as an ordered multiple batch
2. Summarization of the ordered batch representation of the (X,Y) data set by fitting a line to the conditional typical values

Main Topics:

1. Viewing an (X,Y) data set as an ordered collection of "mini-batches"
2. Fitting a line to the conditional typical values by using three mini-batches

Tool Introduced:

Resistant Line

Topic 1. Viewing an (X,Y) data set as an ordered collection of "mini-batches"

I. Basic Issue: Consideration of a data set of paired observations as an ordered multiple batch

1. We know the characteristics of an (X,Y) paired observation data set: two batches of which the ith observation of one is related to the ith observation of the other

2. We have already presented various examples of these data sets: IQ scores of twins; scores on the pretest and the final exam for each member of the class

3. We now have a good feel for ordered multiple batches and the summarization of the batches with conditional typical values

4. Can we analyze an (X,Y) data set as an ordered multiple batch and thus condense it by the use of conditional typicals?

5. <u>Example</u>: Percent illiterate in the population, by state, in 1930 (X) and 1960 (Y) (2)

   We shall use this data set in future discussions

II. Problem: How do we break up an (X,Y) data set into multiple batches?

1. We use the X variable as the ordered multiple batch scale

2. The number of batches is of course arbitrary--depends on the number of observations, n, in the data set

3. As limiting cases:

   a. Use n mini-batches: 1 batch per X (or distinct X) value

   b. Use only 1 batch--Y becomes a single batch of numbers

4. We choose the number of batches so that the corresponding intervals on the X axis are:

   a. Bounded by integers

   b. Approximately equal width (if possible)

   c. Containing equal numbers of Y values

5. A scatterplot of the (X,Y) data is always the first step in the analysis

   a. The plot helps to determine where to break up the X axis

   b. Here is the scatterplot for our illiteracy data--(3) note linear pattern (important)

III. Solution: The number of "mini-batches" to use is arbitrary, and their location along the X axis should be determined by a scatterplot of the observations

IV. Method: Using the Illiteracy data

1. Here are the mini-batches

   a. If X_i is less than 2%, Y_i is in batch 1

   b. If X_i is between 2% and 4%, Y_i is in batch 2

   c. If X_i is between 4% and 6%, Y_i is in batch 3

   d. If X_i is between 6% and 10%, Y_i is in batch 4

   e. If X_i is greater than 10%, Y_i is in batch 5

2. Thus have 5 batches, 3 of equal width 2%, 1 of width 4%, 1 of width 10.5%

3. The inequality in width was forced by the clustering of the data points at the left end of the plot

4. Here is the data set arranged into our mini-batches. Batch observations are the 1960, % illiterate (Y) values (4)

5. We merely analyze this rearranged data set as an ordered multiple batch

   a. Parallel Stem-and-Leaf shows increasing pattern, few outliers (5)

   b. Parallel schematics drawn so that width of box width of interval. Spreads increase (6)

   c. Compute conditional typical values (4)

Batch 1	0.9%
Batch 2	1.6%
Batch 3	2.2%
Batch 4	3.45%
Batch 5	4.35%

d. Plot these values and the hinges, connected on a separate plot. Very linear, eyeball slope ≈ .3 (7)

6. In conclusion, the connected conditional typical plot is very informative

7. However, if this plot is linear, we would formally like to fit a line as a final summarization

Topic 2. Fitting a line to the conditional typical values

I. Basic Issue: Formalization of the analysis of (X,Y) paired observational data by fitting a line

1. We want to know exactly how Y varies with X; i.e., if $Y = f(X)$, what is f?

2. We hope that $f(X) = a + bX$, a line

3. If f is not a line, perhaps we can transform X and/or Y to make it so. We discuss these transformations in the next lecture

4. Note that Y is a function of X. In some cases this is obviously so. But X could also be a function of Y!

5. In some ways, which variable to use as the dependent variable (which variable is a function of the other) is arbitrary

II. Problem: How do we find the a and b in the equation $Y = a + bX$

1. We would like to use the conditional typical values in the fitting process

2. How many mini-batches do we use?

3. Which two points in the connected conditional typical plot do we use to draw the line?

III. Solution: Use three mini-batches of roughly equal size and connect the first and last conditional typicals

IV. Method: Resistant Line

1. The line is known as a resistant line, due to Tukey. It is a fitting procedure that is resistant to outliers in the data

2. Procedure: applied to Illiteracy Data

a. Break the data into thirds according to the X (% illiterate in 1930) values--easy rule to apply to find endpoints of our 3 intervals. If the number of observations is not divisible by 3, put the extra 1 or 2 in the middle mini-batch. That is not necessary in this case. (8)

b. Find Median X and Median Y in each third
Median X = midpoint of interval
Median Y = conditional typical of the batch

c. Label these three median pairs

$(X_{(1)},Y_{(1)}),\ (X_{(2)},Y_{(2)}),\ (X_{(3)},Y_{(3)})$

Note that it is unlikely that the $(X_{(i)},Y_{(i)})$ pairs will actually be paired together in the original data.

d. Locate these three points on the scatterplot, and connect $(X_{(3)},Y_{(3)})$ and $(X_{(1)},Y_{(1)})$. This is the fitted line (9)

e. Formally calculate,

$$b = (Y_{(3)} - Y_{(1)})/(X_{(3)} - X_{(1)}) = \frac{3.2}{9.6} = 0.33$$

$$a = \text{Median } (Y_{(i)} - bX_{(i)}) = \text{Median } (0.43, 0.63, 0.43) = 0.43$$

f. Examine fitted line on the scatterplot. Note how well it fits (except Alaska) (10)

g. Line may need to be "polished" or adjusted slightly for a better fit.

3. To determine how well the line fits the data, we calculate residuals:

$$Y_i - a - bX_i \qquad (11)$$

4. This batch of residuals is extremely important in assessing the fit. Treated as a single batch, residuals should be symmetric about 0, with no outliers. In other words, residuals should be well behaved, mean 0, standard deviation indeterminate. (12)

5. Line constructed with CMU-DAP

Use function LINE. Options to save fitted values and residuals.

400

Lecture 3-2
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 3-2 Outline
Topic 1 Section I		
5.	2	Illiteracy Data, per State, 1930 and 1960
Section II		
5.b	3	% Illiterate 1930 vs. % Illiterate 1960
Section IV		
4.	4	% Illiterate 1960 classified into Mini-batches
5.a	5	Stem-and-Leaf displays of Illiteracy Mini-batches
5.b	6	Schematic Plot of Illiteracy Mini-batches
5.d	7	Connected Conditional Typicals for 1960 Illiteracy Data
Topic 2 Section IV		
2.a	8	Illiteracy Data, broken up into thirds
2.d	9	Thirds of Illiteracy data and connected conditional typicals
2.f	10	% Illiterate 1960 vs % Illiterate 1930
3.	11	Illiteracy Data, Residuals
5	12	Residuals, Stem-and-Leaf

[1]

Lecture 3-2

Looking at (X,Y) data: Analysis of (X,Y) data sets by reorganizing the data into a collection of ordered batches.

Lecture Content:

1) Discussion of how to arrange an (X,Y) data set into an ordered multiple batch.

2) Summarization of the ordered multiple batch data set by fitting a line to the condition typical values.

Main Topics:

1) Viewing an (X,Y) data set as an ordered collection of "mini-batches."

2) Fitting a line to the conditional typical values by using only three mini-batches.

Illiteracy Data [2]

X = % of population illiterate in 1930
Y = % of population illiterate in 1960

	X	Y
Alabama	14.0	4.2
Alaska	20.5	3.0
Arizona	11.0	3.8
Arkansas	7.6	3.6
California	2.8	1.8
Colorado	3.1	1.3
Connecticut	5.1	2.2
Delaware	4.4	1.9
D.C.	1.7	1.9
Florida	7.7	2.6
Georgia	10.4	4.5
Hawaii	17.5	5.0
Idaho	1.2	0.8
Illinois	2.7	1.8
Indiana	1.8	1.2
Iowa	0.9	0.7
Kansas	1.4	0.9
Kentucky	7.3	3.3
Louisiana	15.1	6.3
Maine	3.0	1.3
Maryland	4.2	1.9
Massachusetts	4.0	2.2
Michigan	2.2	1.6
Minnesota	1.4	1.0
Mississippi	14.8	4.9

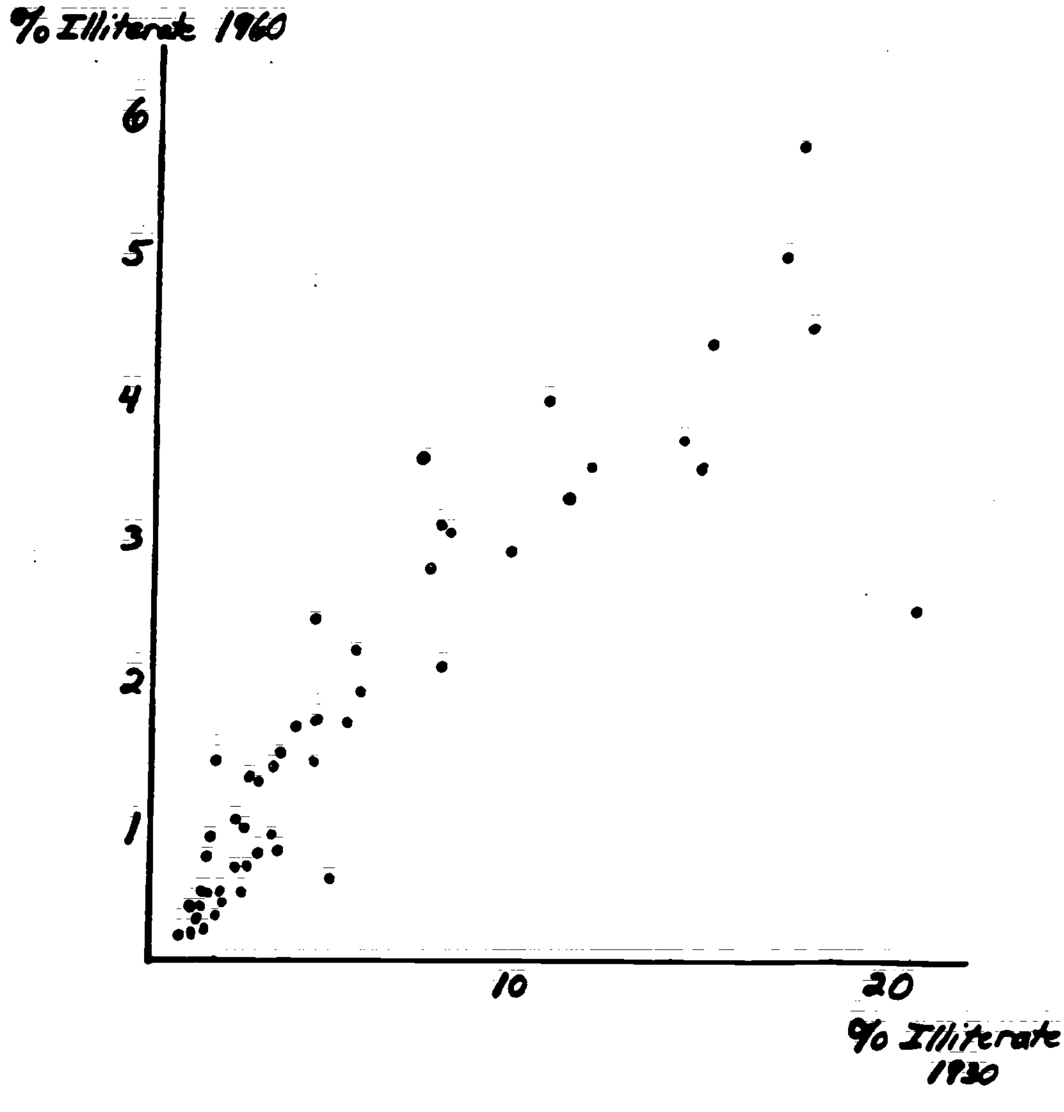
	X	Y
Missouri	2.5	1.7
Montana	1.9	1.0
Nebraska	1.3	0.9
Nevada	4.8	1.1
New Hampshire	3.0	1.4
New Jersey	4.3	2.2
New Mexico	14.9	4.0
New York	4.1	2.9
N. Carolina	11.5	4.0
North Dakota	1.7	1.4
Ohio	2.5	1.5
Oklahoma	3.1	1.9
Oregon	1.1	0.8
Pennsylvania	3.5	8.0
Rhode Island	5.5	2.4
S. Carolina	16.7	5.5
South Dakota	1.4	0.9
Tennessee	8.0	3.5
Texas	7.3	4.1
Utah	1.4	0.9
Vermont	2.4	1.1
Virginia	9.7	3.4
Washington	1.1	0.9
West Virginia	5.5	2.7
Wisconsin	2.1	1.2
Wyoming	1.8	0.9

4[illegible]3

3-2

%Illiterate 1960 (Y) plotted against % Illiterate 1930(X).
One point per state.

(3)



(3-2)

Percent Illiterate, 1960, classified into mini-batches by 1930 percent illiterate, for states and D.C. (4)

Batch 1		Batch 2		Batch 3		Batch 4		Batch 5	
.C.	1.9%	California	1.8%	Connecticut	2.2%	Arkansas	3.6%	Alabama	4.2%
aho	0.8	Colorado	1.3	Delaware	1.9	Florida	2.6	Alaska	3.0
diana	1.2	Illinois	1.8	Maryland	1.9	Kentucky	3.3	Arizona	3.8
wa	0.7	Maine	1.3	Massachusetts	2.2	Tennessee	3.5	Georgia	4.5
nsas	0.9	Michigan	1.6	Nevada	1.1	Texas	4.1	Hawaii	5.0
nesota	1.0	Missouri	1.7	New Jersey	2.2	Virginia	3.4	Louisiana	6.3
ntana	1.0	New Hampshire	1.4	New York	2.9			Mississippi	4.9
braska	0.9	Ohio	1.5	Rhode Island	2.4			New Mexico	4.0
Dakota	1.4	Oklahoma	1.9	West Virginia	2.7			N. Carolina	4.0
egon	0.8	Pennsylvania	2.0					S. Carolina	5.5
Dakota	0.9	Vermont	1.1						
ah	0.9	Wisconsin	1.2						
shington	0.9								
oming	0.9								
$n_1 = 14$		$n_2 = 12$		$n_3 = 9$		$n_4 = 6$		$n_5 = 10$	
930 % iterate s than 2%		1930 % illiterate 2%-4%		1930 % illiterate 4%-6%		1930 % illiterate 6%-10%		1930 % illiterate Greater than 10%	

(5)

Parallel Stem-and-Leaf Displays of Mini-batches of 1960 Illiteracy Data

unit = 1%

	<2%	2% - 4%	4% - 6%	6%-10%	>10%
0*	788999999				
1	00249	12334567889	199		
2*		0	222479	6	
3				3456	08
4*				1	00259
5					05
6*					3

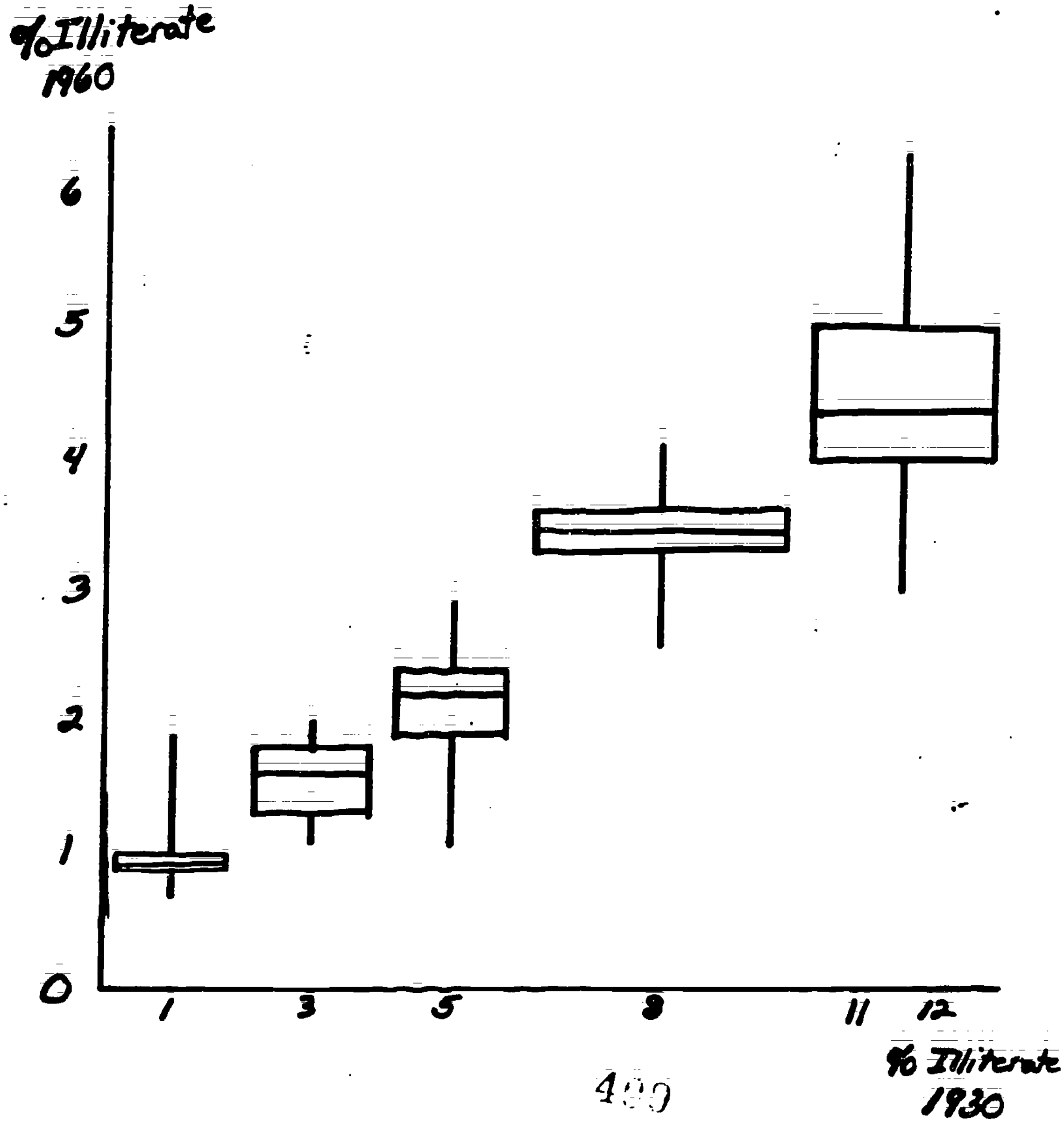
1930 % Illiterate

Module II

(3-2)

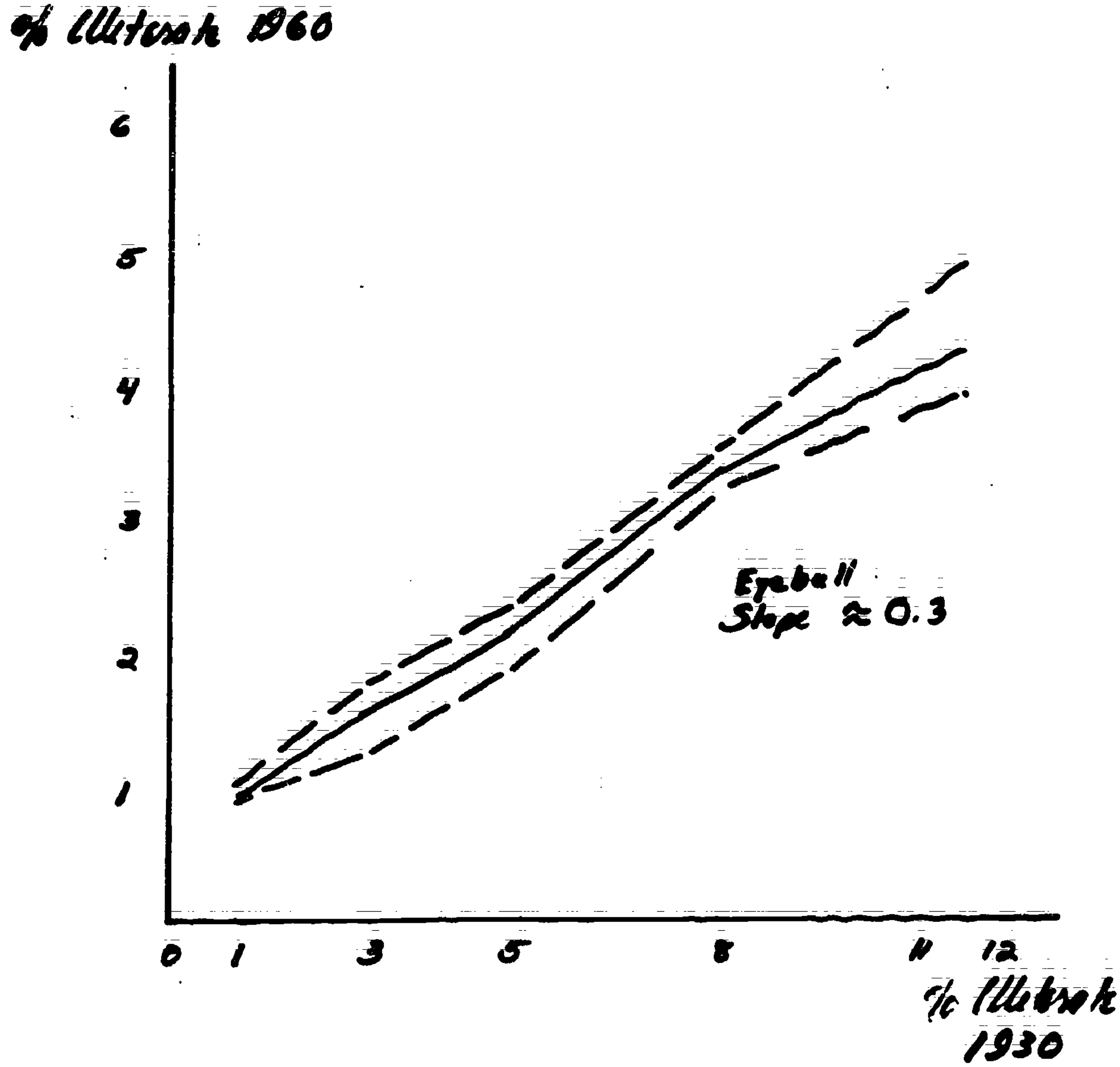
(6)

Schematic Plot of Mini-Batches of % Illiterate, 1960 Data



(7)

Illiteracy Data, Conditional Typicals for Mini-Batches Connected (solid Line) and Hinges Connected (Dashed Line).



(3-2)

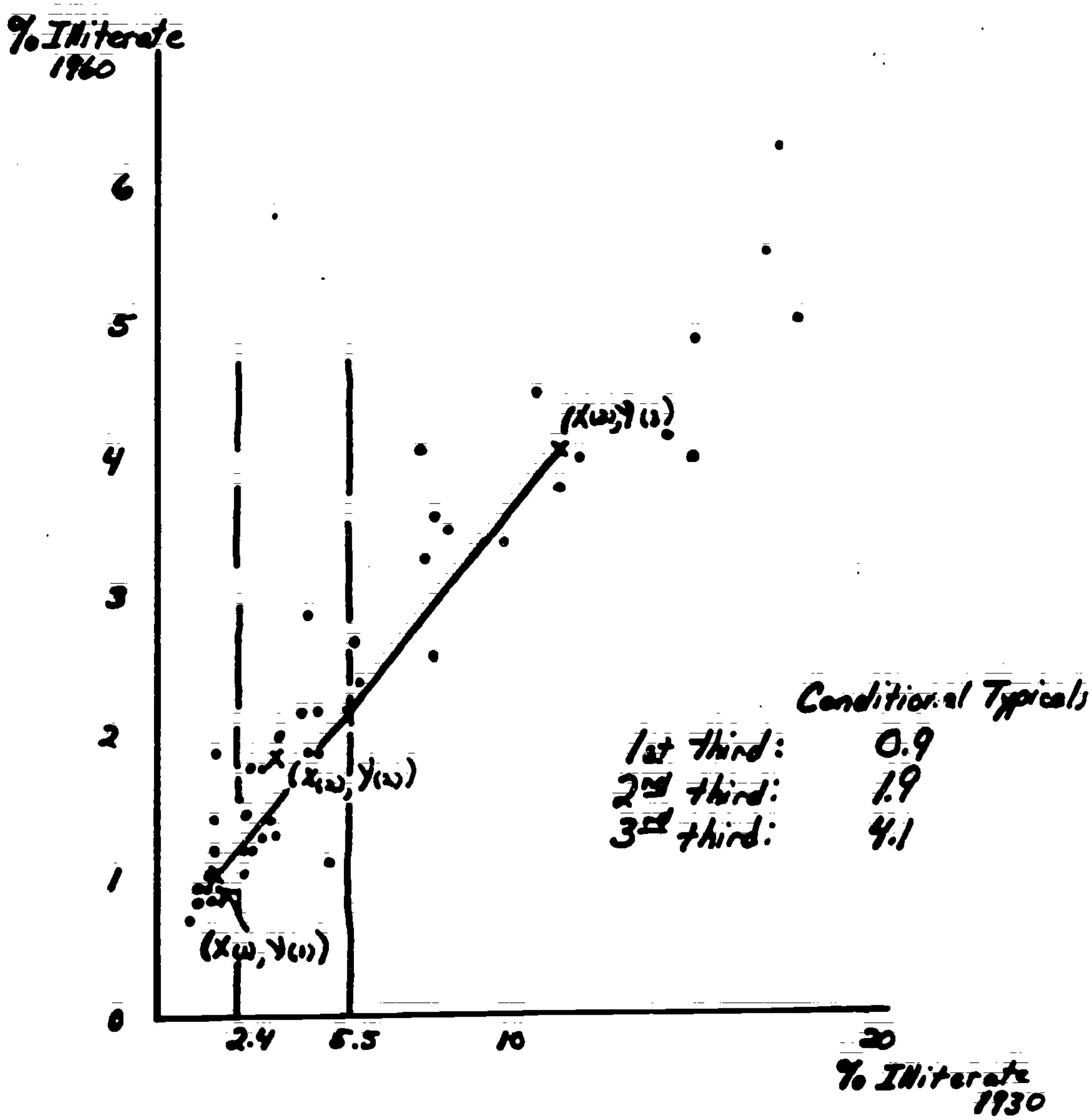
(8)

Illiteracy Data, broken up into thirds, using X

X	Y	X	Y	X	Y
0.9	0.7	2.5	1.5	5.5	2.7
1.1	0.8	2.5	1.7	7.3	3.3
1.1	0.9	2.7	1.8	7.3	4.1
1.2	0.8	2.8	1.8	7.6	3.6
1.3	0.9	3.0	1.3	7.7	2.6
1.4	0.9	3.0	1.4	8.0	3.5
1.4	0.9	3.1	1.3	9.7	3.4
1.4	0.9	3.1	1.9	10.4	4.5
1.4	1.0	3.5	2.0	11.0	3.8
1.7	1.4	4.0	2.2	11.5	4.0
1.7	1.9	4.1	2.9	14.0	4.2
1.8	0.9	4.2	1.9	14.8	4.9
1.8	1.2	4.3	2.2	14.9	4.0
1.9	1.0	4.4	1.9	15.1	6.3
2.1	1.2	4.8	1.1	16.7	5.5
2.2	1.6	5.1	2.2	17.5	5.0
2.4	1.1	5.5	2.4	20.5	3.0
$X_{(1)} = 1.4$		$X_{(2)} = 3.5$		$X_{(3)} = 11.0$	
$Y_{(1)} = 0.9$		$Y_{(2)} = 1.9$		$Y_{(3)} = 4.1$	
First Third		Second Third		Third Third	

(9)

Illiteracy data, broken up into thirds.
Conditional typicals given as median y in each third. First and third conditional typical connected.
(Theory behind Resistant Line).

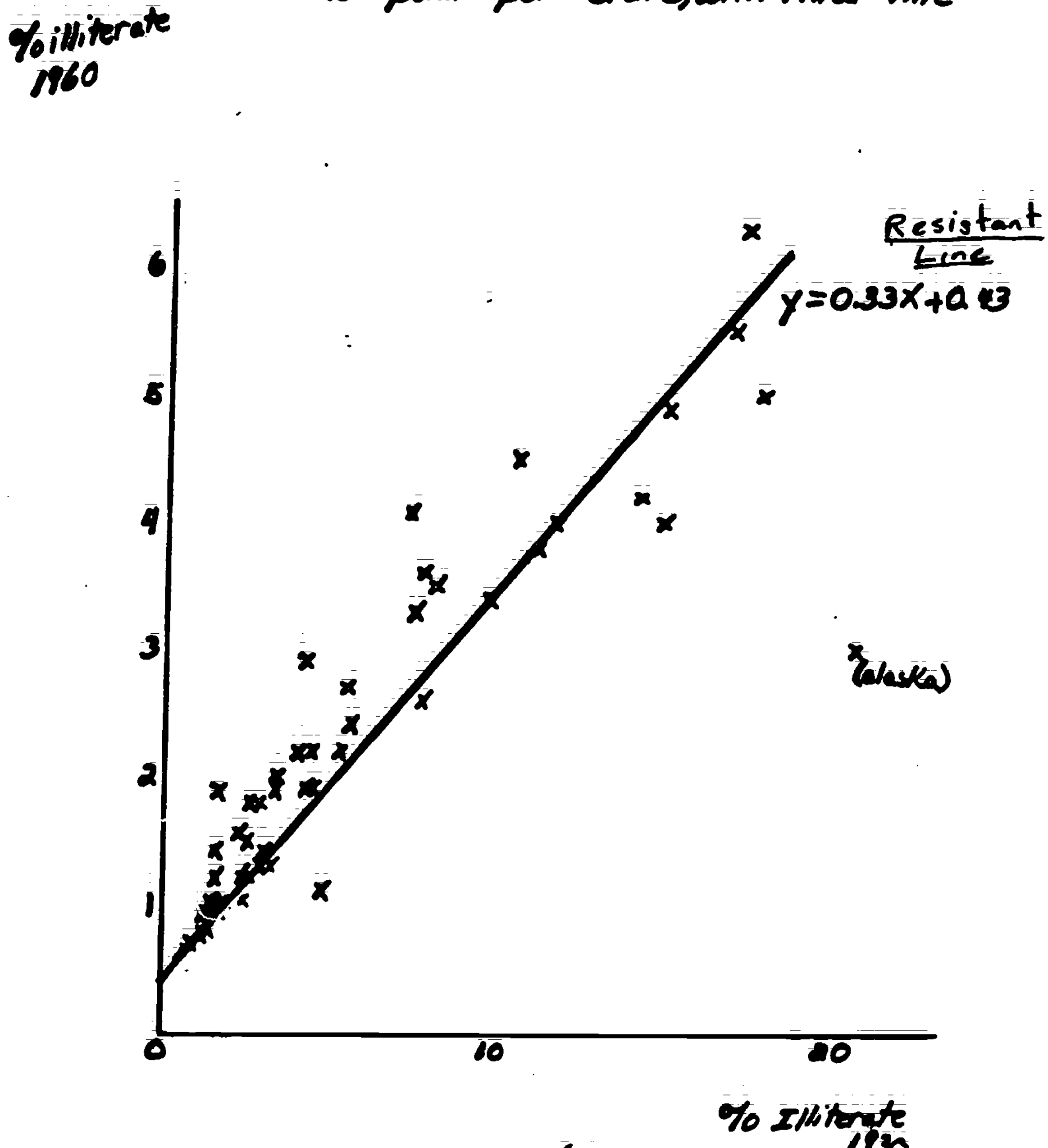


412

(3-2)

(10)

% illiterate 1960 (y) plotted against
% illiterate 1930 (x)
One point per state, with fitted line



413

(3-2)

Illiteracy Data, Residuals from Resistant Line (11)

14.0	4.2	-0.90	1.3	0.9	0.04
20.5	3.0	-4.26	4.8	1.1	-0.93
11.0	3.8	-0.30	3.0	1.4	-0.03
7.6	3.6	0.64	4.3	2.2	0.34
2.8	1.8	0.44	14.9	4.0	-1.40
3.1	1.3	-0.16	4.1	2.9	1.10
5.1	2.2	0.07	11.5	4.0	-0.26
4.4	1.9	0	1.7	1.4	0.40
1.7	1.9	0.90	2.5	1.5	0.24
7.7	2.6	-0.40	2.1	1.9	0.44
10.4	4.5	0.60	1.1	0.8	0
17.5	5.0	-1.26	3.5	2.0	0.40
1.2	0.8	-0.03	6.5	2.4	0.14
2.7	1.8	0.47	16.7	5.5	-0.50
1.8	1.2	0.17	1.4	0.9	0
0.9	0.7	-0.03	8.0	3.5	0.40
1.4	0.9	0	7.3	4.1	1.24
7.3	3.3	0.44	1.4	0.9	0
15.1	6.3	0.84	2.4	1.1	-0.13
3.0	1.3	-0.13	9.7	3.4	-0.26
4.2	1.9	0.07	1.1	0.9	0.10
4.0	2.2	0.44	6.5	2.7	0.44
2.2	1.6	0.44	2.1	1.2	0.07
1.4	1.0	0.10	1.8	0.9	-0.13
14.8	4.9	-0.46			
2.5	1.7	0.44			
1.9	1.0	-0.06			

(12)

Stem-and-Leaf of Residuals from Resistant Line

unit = .10

Lo	-4.26
f	4
t	
-1	
-0**	9 9
s	
f	5 4 4
t	3 2 2
-0	1 1 1 1 0 0 0 0
0**	0 0 0 0 0 0 0 0 0 1 1 1 1
t	2 3
f	4 4 4 4 4 4 4 4 4 4 4 4
s	6
0	8 9
1**	1
t	2 2

415

(3-2)

Lecture 3-3. Summarizing Scatterplots

Summarizing Scatterplots of (X,Y) data: Transforming (X,Y) data sets to improve the linear fit, and fitting lines by least squares.

Lecture Content:

1. Transformations of (X,Y) data sets
2. Least Squares Principle and coefficient estimates
3. Assessing the fit

Main Topics:

1. Transformations to improve linearity and equalize spread
2. Fitting a line using least squares
3. Looking for patterns in the residuals

Tools Introduced:

1. Least Squares
2. Residual Plots

416

Topic 1. Transformations to improve Linearity and Equalize Spread

I. Basic Issue: Transforming data to make a fitted line a good summary

1. Paired observational data are rarely linear in the raw form

2. In addition to this nonlinearity, the spreads of the constructed mini-batches may not be equal

3. We seek to transform the data to:

   a. Improve Linearity

   b. Equalize Spread

4. Both these goals are important, and should be sought whenever possible and necessary

II. Problem: How do we achieve these 2 goals?

1. Linearity is (usually) increased by transforming the X variable

   a. Transforming X to higher powers has the effect of stretching the X axis, which promotes linearity in plots that resemble exponential functions (e^x) or ($-e^x$)

   b. Transforming X to small powers has the effect of shrinking the X axis, which promotes linearity in plots that resemble negative exponential functions (e^{-x}) or ($-e^{-x}$)

2. Spread is often equalized by transforming Y; similar to transformations to equalize spread with multiple batches

3. Our conditional typical values should be useful in choosing good transformations, since the plot of the values "mimics" the patterns of the (X,Y) scatterplot

III.. Solution: Use (median X, median Y) points from the three thirds for resistant lines

1. We divide the data into thirds on the basis of the X-values, keeping each Y with its paired X-value

2. We then have 3 sample points

$$(X_{(1)},Y_{(1)}),\ (X_{(2)}.Y_{(2)}),\ (X_{(3)},Y_{(3)})$$

which are the medians of the 3 mini-batches

3. The effect of transformations on the data set can be seen by merely transforming the 3 sample points

4. Seek a transformation so that the slopes:

$$\frac{Y_{(2)}^{R_1} - Y_{(1)}^{R_1}}{X_{(2)}^{R_2} - X_{(1)}^{R_2}} = \frac{Y_{(3)}^{R_1} - Y_{(2)}^{R_1}}{X_{(3)}^{R_2} - X_{(2)}^{R_2}}$$

are equal:

or $\frac{S_1}{S_2} \approx 1$

IV. Method

1. Example: Per capita Income (X) and Infant Mortality (Y) for nations (2) (2a)

2. Scatterplot shows both nonlinearity (curve has e^{-x} shape) and disparity in spread (3)

3. Mini-batches of data are constructed

4. Mini-batches are plotted via Parallel Schematic Display--discrepancies from ideal situation are evident (4)

5. Connecting the Conditional Typicals and Hinges is quite useful in studying the relationship of the raw data (5)

6. Recall Tukey's diagram for determining which direction to move in our transformations. Transforming Y can also help improve linearity

7. However, we first concentrate on transforming X for linearity. If necessary, we then transform Y to equalize spread and possibly promote increased linearity

8. As mentioned, we take the three resistant line summary points, and examine the line connecting the first and second, and the line connecting the second and third

9. When the slopes of these lines are equal, we have the appropriate transformation of X, and Y

10. The calculations for our example: log, log appears best (6)

11. Scatterplot of log (infant mortality) vs log (income) is very linear (7)

12. Fitted resistant line· $Y = 3.33 - .59X$ (8)

13. Residuals are nice and tight around zero, except for 2 large values (Libya and Saudi Arabia) (8)

14. In conclusion, we study the effect on the schematic plots of the mini-batches of the log X and log Y transformations

    a. log(X) has shrunk the X scale, and increased linearity (9)

    b. log(Y) has definitely equalized spread (10)

    c. Put them both together, and plot looks very good (11)

419

Topic 2. Fitting a line using least squares

I. Basic Issue: Presentation of the Least Squares Principle

1. Resistant line is just one method of fitting a line to an (X,Y) point cloud

2. We prefer it because it is resistant to outlying or deviant points

3. The "classical" fitting procedure is known as "least squares"

4. In recent years least squares has come under attack because it is very sensitive to outliers

II. Method

1. The least squares principle finds the line which has the minimum value of the quantity: (12)

$$\sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)^2$$

2. This line, $Y_i = \hat{a} + \hat{b}X_i$, minimizes the sum of the squared residuals

3. Since the residuals are used in the procedure, they are very important in assessing how well the line fits

4. Here is the geometrical interpretation of least squares: Note that we minimize the squared distances of the points from the line (13)

5. The least squares line will be a good fit when:

   a. Data are linearly related

   b. Spread about the line is constant

   c. No outliers

6. How do we assess the fit? (14)

   a. Examine the residuals--stem-and-leaf, plot vs X (see Topic 3)

   b. Examine variance about the line:

$$s^2_{y|x} = \frac{\Sigma(Y_i - \hat{a} - \hat{b}X_i)^2}{n-2}$$

   We want this as small as possible

c. Examine one minus the ratio of residual variation to the total variation of Y:

$$r^2 = 1 - \frac{\Sigma(Y_i - \hat{a} - \hat{b}X_i)^2}{\Sigma(Y_i - \bar{Y})^2} = 1 - \frac{S^2_{y|x}}{S^2_y}$$

This is the "percent of variance" explained.

7. Least Squares line for Infant Mortality data: (15)

$Y = 3.11 - .512X$

Slope differs from resistant slope

8. Residuals slightly more tight around 0 than with resistant line (16)

9. Least Squares with CMU-DAP

Use function MREG.

421

Topic 3. Looking for Patterns in the Residuals

I. Basic Issue: What should the batch of residuals resemble?

1. Batch of residuals should be:

   a. Symmetric about zero

   b. Devoid of outliers

2. That is: batch should be well-behaved

3. Plotted against X, residuals should be a random swarm of points, with no pattern

II. Method: Residual Plots

1. Plot of residuals (Y) vs X for Infant Mortality data--no (17) pattern evident, two high outliers are apparent

2. Patterns to look out for

   a. Trigonometric (Sinusoidal)

   b. Sign patterns

   c. Wedge shape

   d. Linear

   e. Curves

   f. Deviants

Lecture 3-3
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 3-3 Outline
Topic 1 Section IV		
1.	2 2a	Income and Infant Mortality Rate for Nations
2.	3	Scatterplot of Income and Infant Mortality
4.	4	Schematic plots of Infant Mortality
5.	5	Conditional Typicals for Infant Mortality data
10.	6	Determination of Transformation for Infant Mortalities
11.	7	Scatterplot of Logged Infant Mortality data, with Fitted Line
13	8	Stem-and-Leaf of Residuals
14.a	9	Schematic Plot--X transformed
14.b	10	Schematic Plot--Y transformed
14.c	11	Schematic Plot--X and Y transformed
Topic 2 Section II		
1.	12	Least Squares principles
4.	13	Geometrical Least Squares
6.	14	How well does the Least Squares Line fit

423

LECTURE 3-3

Summarizing Scatterplots of (X,Y) data: Transforming (X,Y) data sets to improve the linear fit, and fitting lines by least squares.

Lecture Content:

1) Discussion of transformations of (X,Y) data sets.

2) The Least Squares principle and estimates of the coefficients.

3) Assessing the fit.

Main Topics:

1.) Transformations to improve linearity and equalize spread.

2) Fitting a line using least squares.

3) Looking for patterns in the residuals.

(3)

INCOME & INFANT MORTALITY RATE FOR NATIONS

INCOME	INFANT MORTALITY	INCOME	INFANT MORTALITY
$3426	16.7	$1760	27.8
3350	23.7	302	79.1
3346	17	2526	22.1
4751	16.8	727	26.2
5029	13.5	631	13.6
3312	10.1	295	32
3403	12.9	684	60.9
5040	20.4	507	46
2009	17.8	754	34.1
2298	25.7	334	65.1
3292	11.9	1268	20.4
4103	11.6	1256	15.1
3723	16.2	261	19.1
4102	11.3	732	26.2
956	44.8	434	76.3
NA	NA	799	40.4
5596	9.6	406	43.3
2963	12.8	310	259
2503	17.5	430	86.3
5523	17.6	360	78.5
1191	59.6	110	125
425	170	1280	139
510	78	560	28.1
426	62.8	3010	300
725	54.4	180	58
406	48.8	1530	650
		1240	51.7
		193	60.4

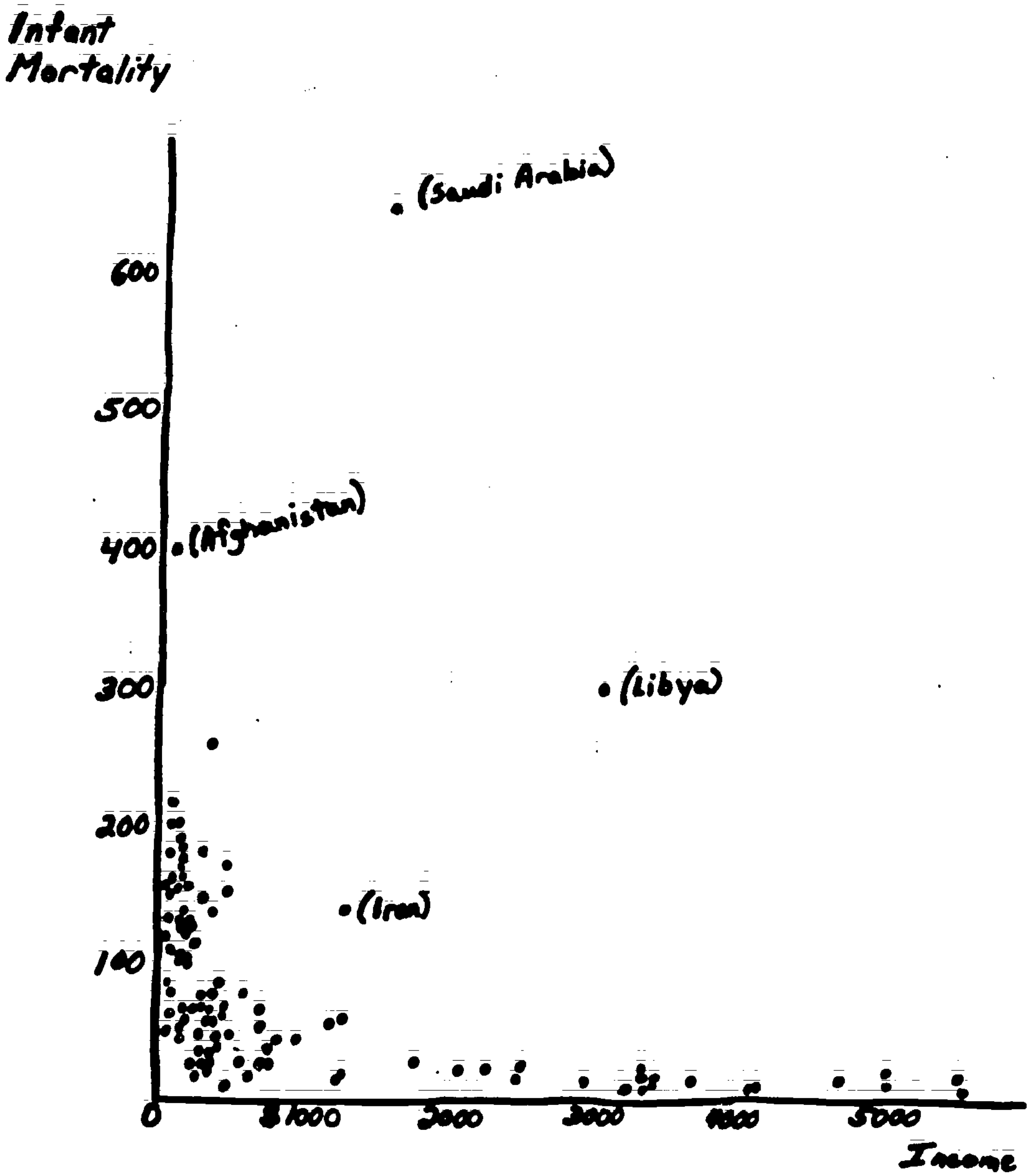
(2a)

INCOME & INFANT MORTALITY RATE FOR NATIONS

INCOME	INFANT MORTALITY
$ 165	137
281	180
210	114
319	58.2
217	63.7
284	39.3
387	138
334	21.3
344	58
197	159.2
279	149
477	10.2
347	38.6
230	67.9
334	21.7
210	27
435	153
130	100
83	400
111	124.3
73	200
68	150
123	100
122	190
70	160
81	109.6
79	84.2
79	216
$ 100	NA
93	60.6
169	55
71	NA
120	102
130	148.3
50	120
174	187
90	NA
70	200
102	124.3
61	132.9
148	170
85	158
162	45.1
125	129.4
120	162.5
160	127
134	160
62	180
96	80
77	50
186	104

(3)

Scatterplot of Per capita Income vs. Infant Mortality for Nations

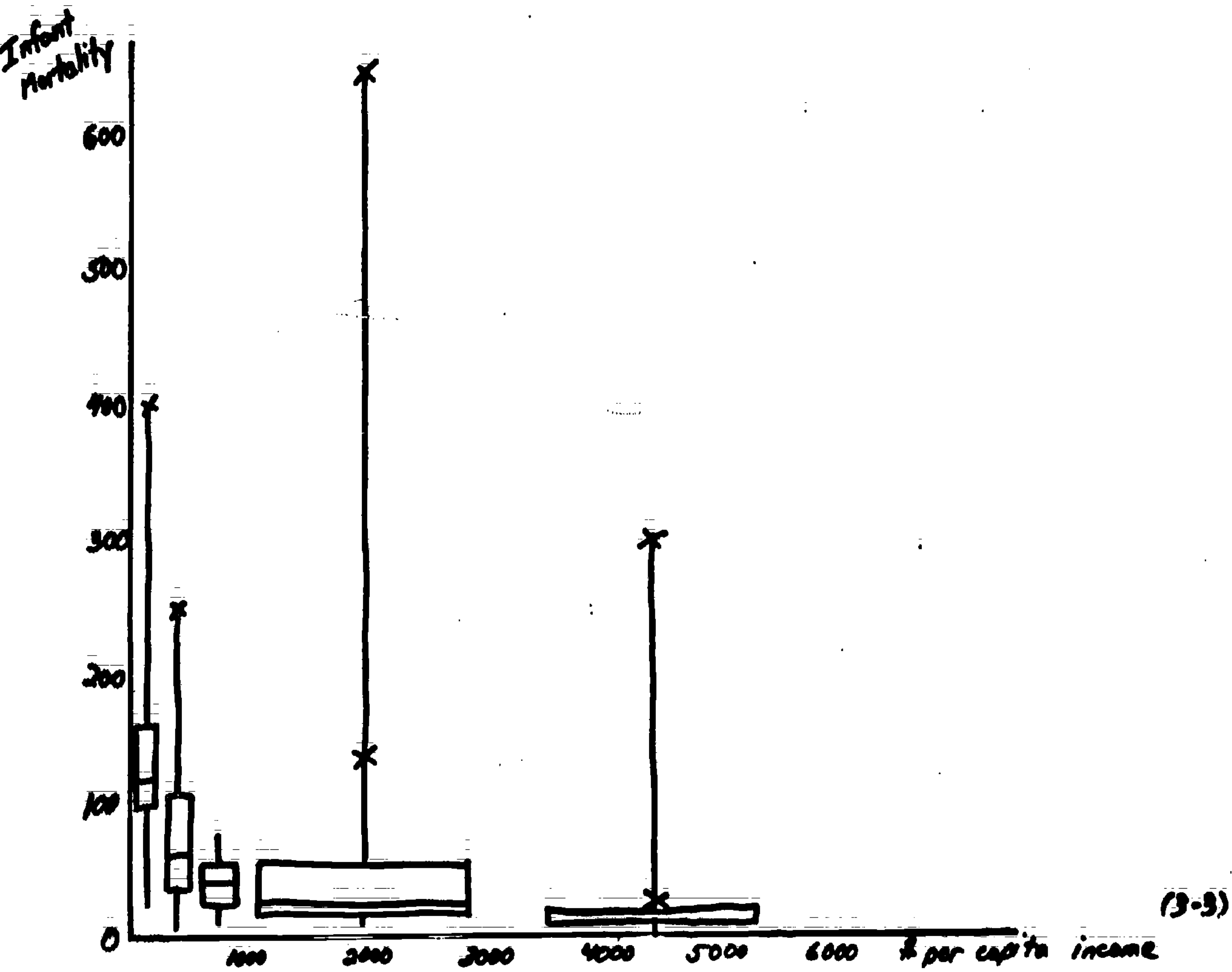


428

(3-3)

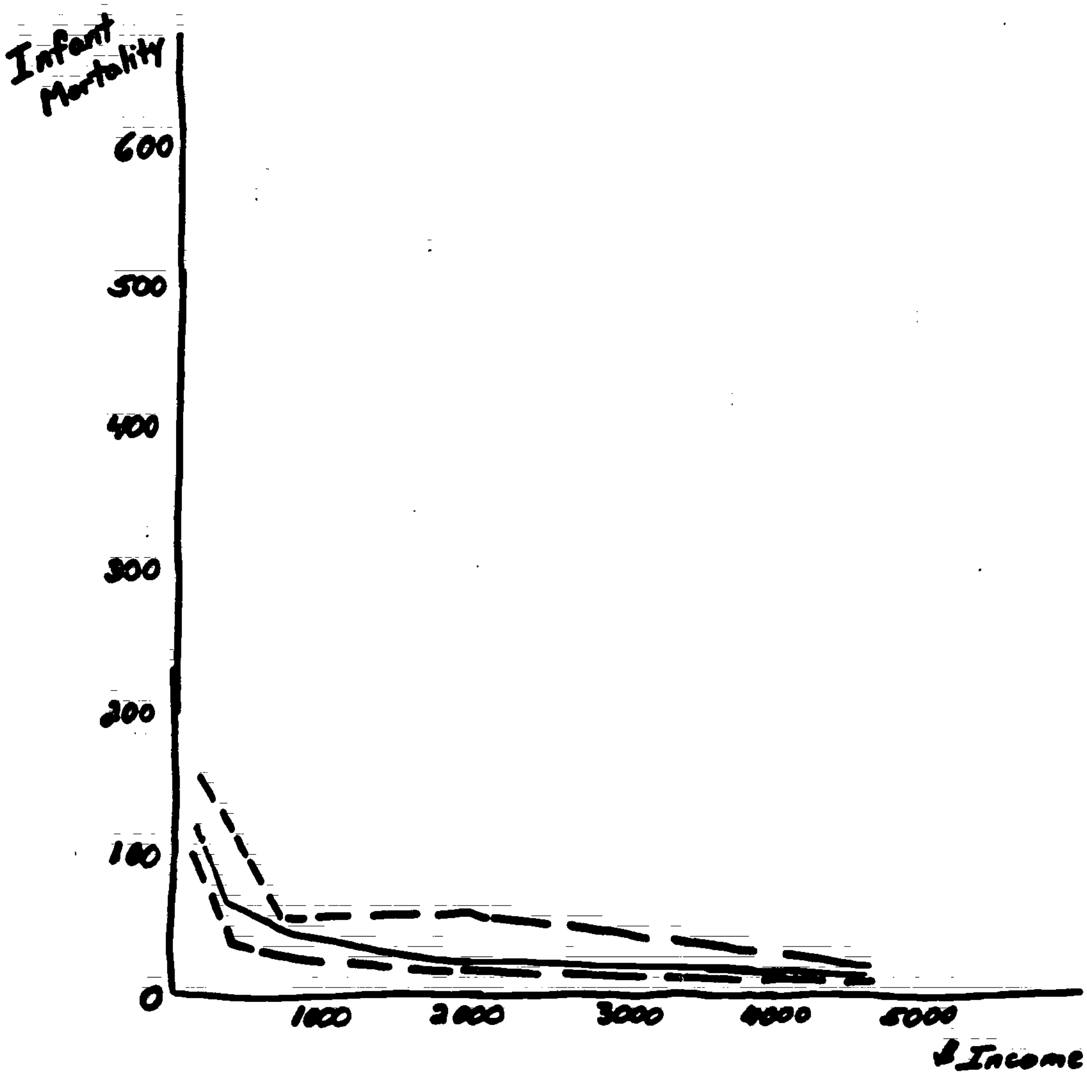
(4)

Schematic plots of batches for Infant Mortality Rate



(5)

Conditional Typicals and Hinges for Batches of Infant Mortality data



(6)

Determination of Transformation of Infant Mortality Data

X = National Income in Dollars
Y = Infant Mortality Rate

Summary Points

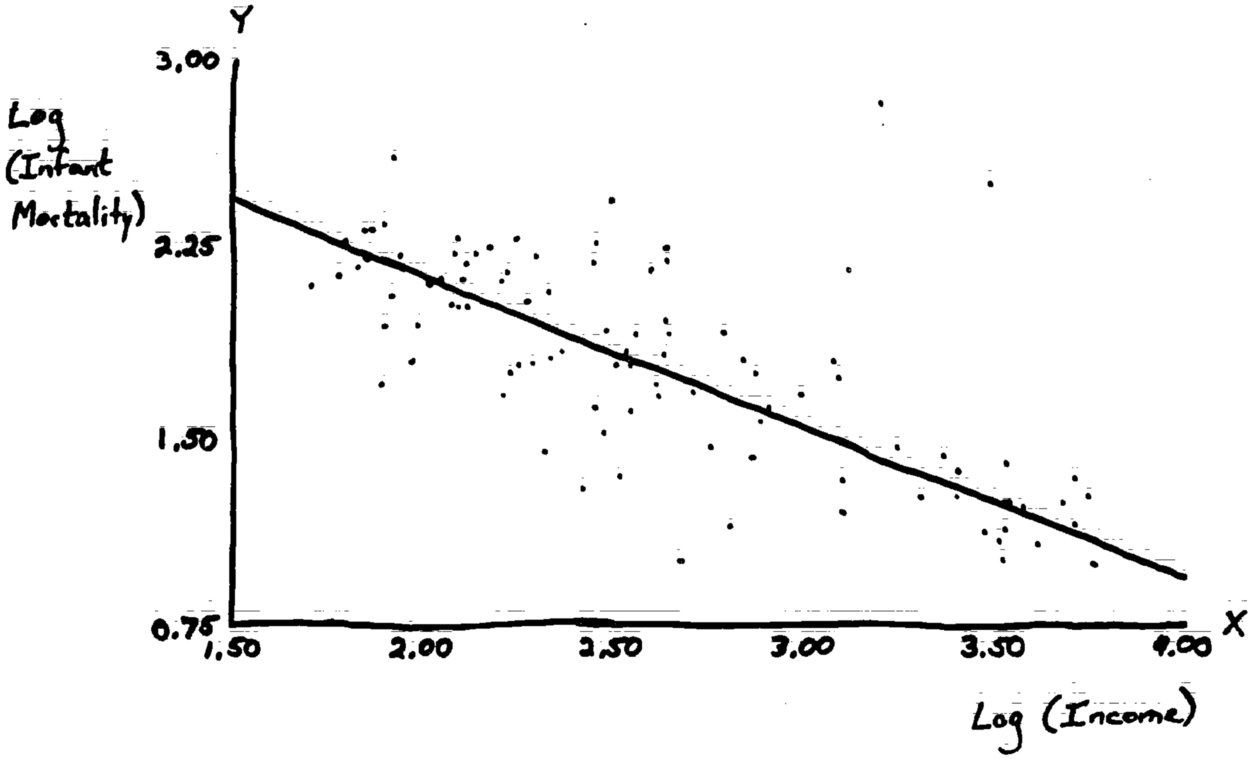
(100, 125)	(334, 58)	(2526, 20.4)
$(X_{(1)}, Y_{(1)})$	$(X_{(2)}, Y_{(2)})$	$(X_{(3)}, Y_{(3)})$

Plot suggests transforming X down and/or Y down.

SLOPES

	$\frac{Y_{(2)} - Y_{(1)}}{X_{(2)} - X_{(1)}}$	$\frac{Y_{(3)} - Y_{(2)}}{X_{(3)} - X_{(2)}}$		Apprx. ratios
X, Y	-0.29	-0.02		$\frac{1}{15}$
$\sqrt{X}, \sqrt{Y}$	-0.43	-0.10		$\frac{1}{4}$
$\log(X), \sqrt{Y}$	-6.81	-3.53		$\frac{1}{2}$
$\sqrt{X}, \log(Y)$	-0.04	-0.01		$\frac{1}{4}$
$\log(X), \log(Y)$	-0.64	-0.51	* * *	.827
$-\frac{1}{\sqrt{X}}, \log(Y)$	-7.36	-13.03		2
$\log(X), -\frac{1}{\sqrt{Y}}$	-0.081	-0.102		1.25

Resistant Line Fitted to Log (Infant Mortality) for Nations (7)



Resistant Line: $Y = -0.59487 * X + 3.32525$

433

(3-3)

Module II

(8)

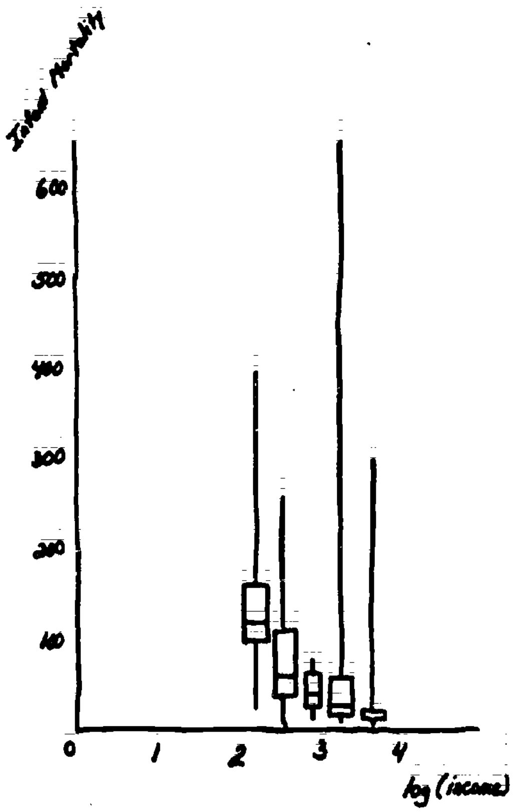
Stem-and-Leaf of Residuals from Resistant Line

Fit of Log(Infant Mortality) vs. Log(National Income)

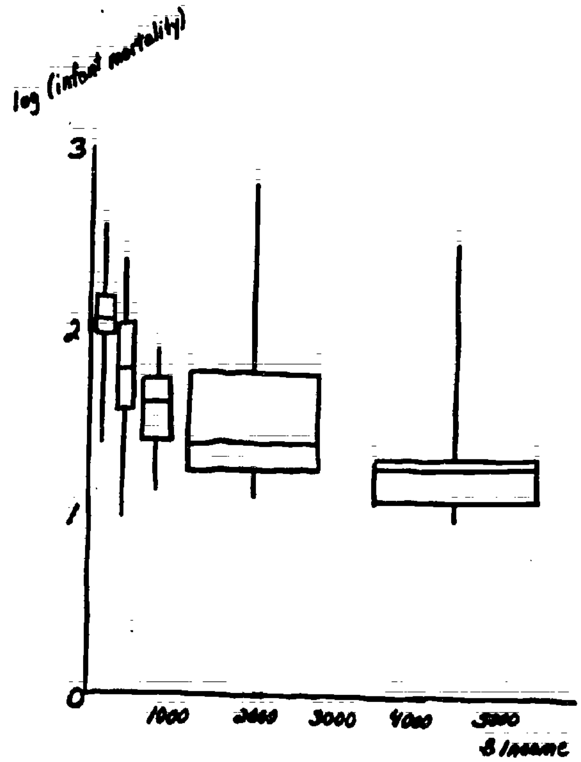
unit = .1

```
-0 ·  |
    s | 6 7
    f | 5 4 4 5 5
    t | 2 2 3 3 2 2 2 2 2 2 3 3 2 2
-0 *  | 1 1 1 1 1 1 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 0 0 0 1 0 0 1 0
 0 *  | 0 1 0 0 0 1 0 0 0 1 0 1 0 0 0 1 1 0 1 0 1 1 0 1 1 1 0 1 0 1 0
    t | 2 2 2 3 3 2 3 2
    f | 4 5 4 4
    s | 6
 0 ·  |
 1 *  |
    t | 2 3
```

Schematic Plot of Batches. X transformed. (9)



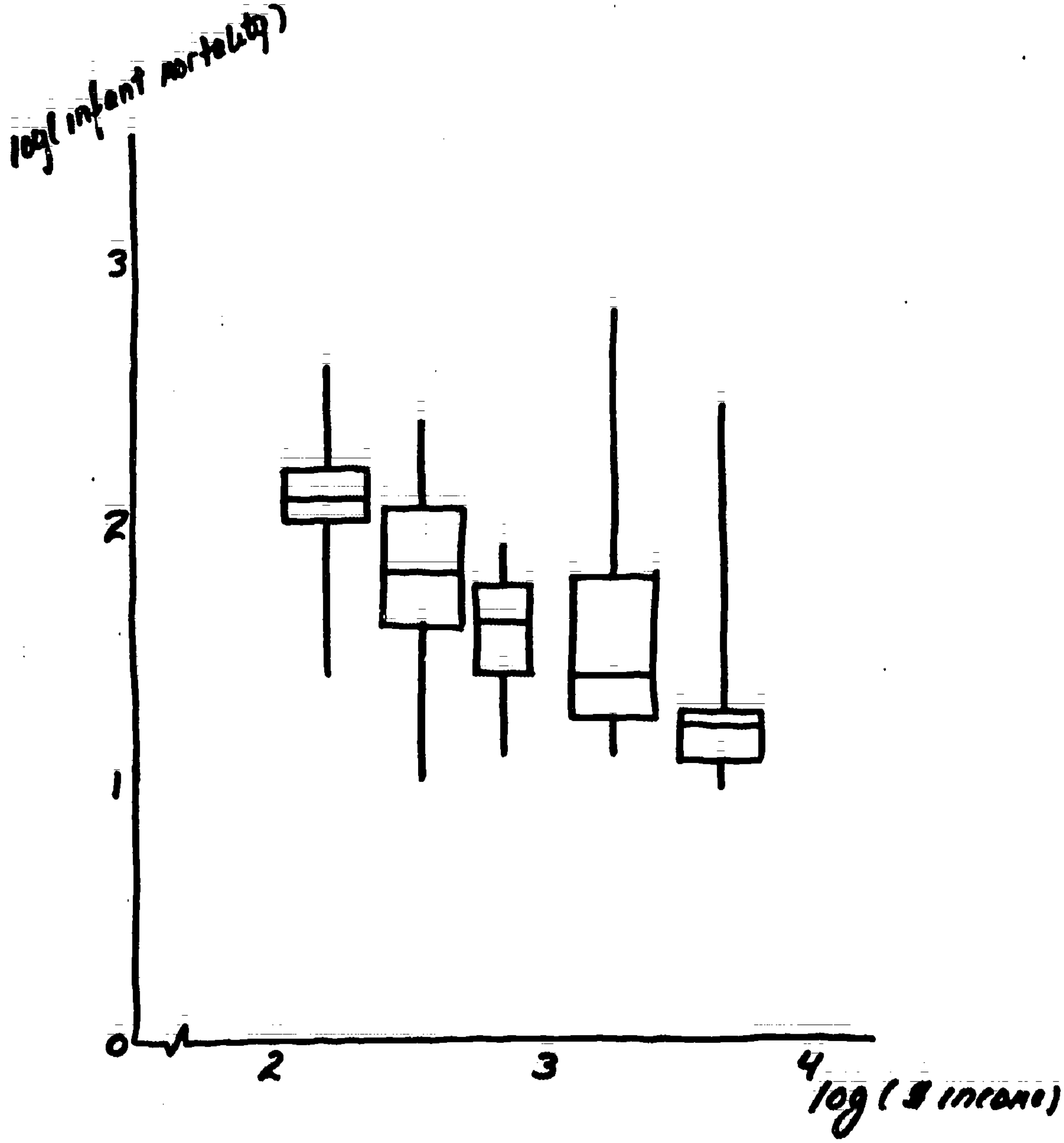
Schematic Plot of Batches. Y transformed. (10)



(3-3)

XVI.II.111

Schematic Plot of Batches. (11)
Both X and Y transformed.



(3-3)

(12)

LEAST SQUARES

Hypothesized relationship:

$$y = a + bX$$

Least squares principle:

FIND the $\hat{a}$ and $\hat{b}$ such that the quantity

$$\sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b} X_i)^2 \text{ is at a minimum.}$$

i.e., we find the line that minimizes the sum of the residuals squared.

The least squares estimates:

$$\hat{b} = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2}$$

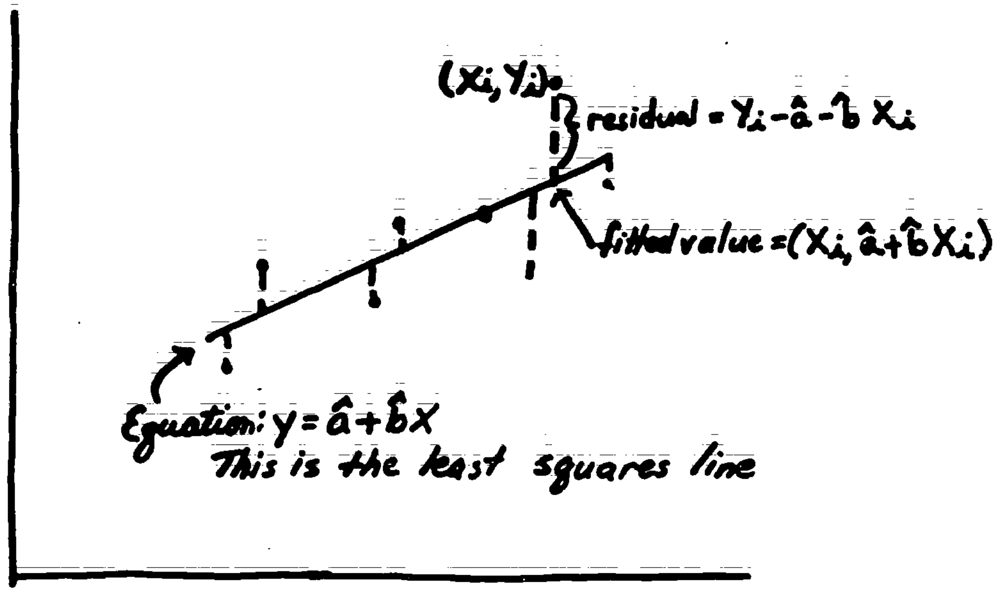
$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

The residuals from the least squares line:

$$r_i = Y_i - \hat{a} - \hat{b} X_i$$

help us assess how well the line fits the data.

Geometrical interpretation of Least Squares. (13)



The least squares line minimizes the squared distances of the data points from the line.

We have the "decomposition":

data value = fitted value + residual

$$Y_i = \hat{Y}_i + (Y_i - \hat{Y}_i)$$

The least squares line minimizes the sum:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

440 (3-3)

(14)

How well does the least squares line fit?

We examine:

1) All n residuals from the line

2) The residual variation, or variance about the line:

$$S^2_{y \cdot x} = \frac{\sum (Y_i - \hat{a} - \hat{b} X_i)^2}{n-2}$$

<u>NOTE</u> that over all possible lines, $S^2_{y \cdot x}$ is at a minimum.

3) 1 minus the ratio of the residual variation to the total variation of Y:

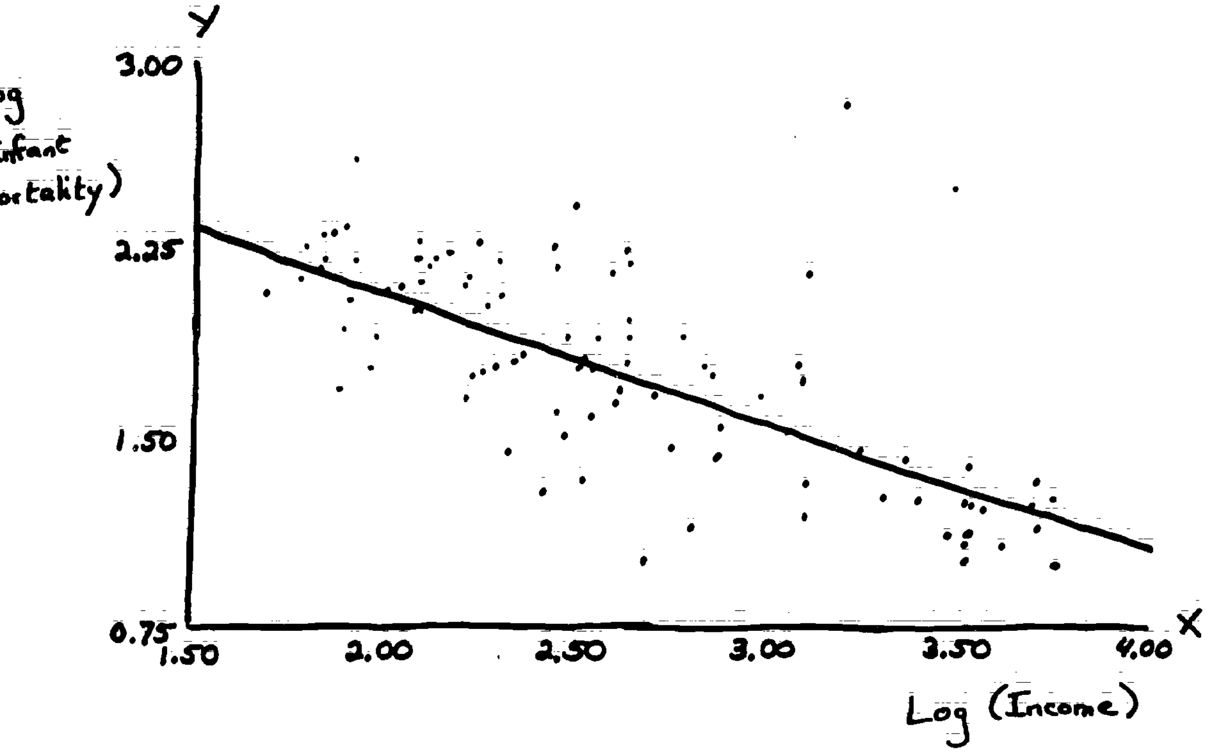
$$r^2 = 1 - \frac{\sum (Y_i - \hat{a} - \hat{b} X_i)^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{s^2_{y \cdot x}}{s^2_y}$$

This is interpreted as the percentage of total variation that we have "explained" by fitting the line.

(3-3)

(15)

Least Squares Line for Infant Mortality Data



Least Squares Line: Y= -0.511798* X -- 3.10715

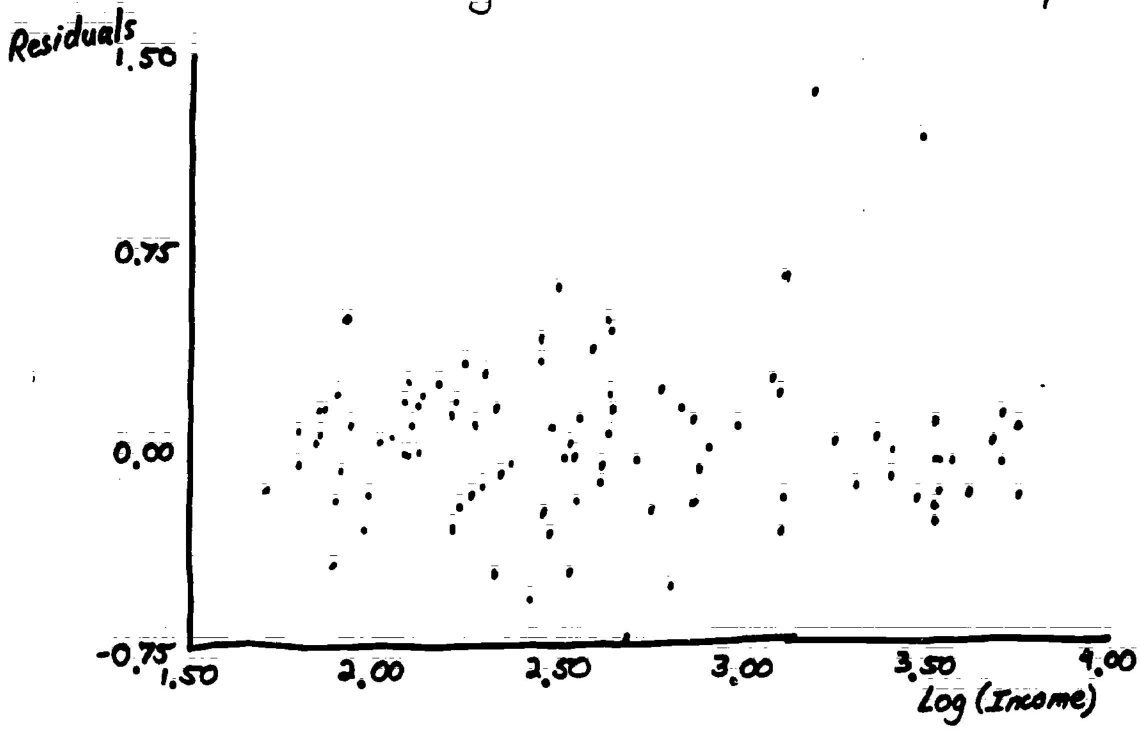
XVI.II.116

Residuals from the Least Squares Line (16)
Infant Mortality Data

unit = .1

```
-0 · |
   s | 7
   f | 5 5 4 4 4 4
   t | 3 2 2 2 2 2 2 3 2 3 2 2 2 2 2 3 2 3
-0 * | 0 0 0 1 1 1 0 1 0 0 0 0 1 0 0 1 1 0 1000000101
 0 * | 0 0 0 0 0 0 0 0 0 1 1 1 0 0 1 1 1 0100111 10001100
   t | 2 2 3 2 3 2 3 2
   f | 4 5 4 4 4
   s | 6
 0 · |
 1 * | 1
   t | 3
```

Plot of Residuals from the Least Squares Line vs. log (Income) for Infant Mortality Data



Lecture 3-4. Analysis of Time Series Data

Analysis of Time Series Data: Smoothing time series data with little Structure, and studying and summarizing time series data with substantial structure

Lecture Content: (1)

1. Smoothing time plots to remove irregularities, and identifying any periodicities in the data
2. Fitting lines to time plots, and extrapolating and interpolating apparent trends

Main Topics:

1. Smoothing Time Plots
2. Summarizing Time Plots

Tools Introduced:

Running Medians of 3 Smoother

Topic 1. Smoothing Time Plots

I. Basic Issue: Time Series Data may have quite a bit of "noise"

1. Time series data consist of paired data, where the X value is a time scale--days, months, years, etc.

2. We generally have only one Y observation for each X value

3. Such data sets can be quite irregular, having many peaks and troughs, when plotted

4. We need to be able to find the pattern of the data (if present) by filtering out the irregularities, or "noise"

II. Problem: How do we best identify any patterns in the data

1. Time series data are commonly collected. We can think of many examples: U.S. Gross National Product for the years 1946-1976; daily reported number of swine flu cases, January-September 1976; Dow Jones averages in a 30 day period

2. We need techniques applicable to all these instances

3. We would like to <u>average</u> a time series data set to remove noise

4. There are 2 distinct methods of averaging

   a. Monthly Averages

   b. Running Averages

5. Monthly averages occur when data are collected daily and then are summed or averaged so that only one data value is reported for each month

6. Similarly, we can average monthly data to get yearly data, yearly data to obtain decade data, etc.

7. Such averaging is quite helpful and often used; however there is a great reduction in number of observations (30 → 1, 12 → 1, etc.)

8. We prefer the use of running averages, since such loss does not occur

III. Solution: "Smooth" the data by taking running medians of three

1. Smoothing has become quite popular in the last 10 years because it is easy to do and is effective

2. We have chosen a very simple smoother--running medians of 3--which works well even though it is simple

IV. Method

1. Example: Emergency Room registrations at D.C. General Hospital, 1970-1975. Monthly data (2)

2. The first step is always make a time plot of the data (3)

   a. Note the many peaks and troughs

   b. Data appear to increase until 1973, then fall

   c. Very difficult to compare years because of irregularities

3. We shall smooth the data to remove these

   a. Write down the data in one column on the left margin of a page (4a) (4b)

   b. Take 3 values consecutively and record their median; for the second data value 8120, record med(7476, 8120, 7706) = 7706

   c. Continue through the data, taking 3 at a time

   d. Endpoints? Merely copy the end value. Tukey has other suggestions

   e. Continue the smoothing until the ith smooth is identical to the (i-1)st smooth. These data required 3 smooths

   f. From one smooth to the next, we need only record those values that change

   g. Plot the smoothed data, and study it

      i. Hospital data similar from year to year

      ii. Rises, peaks in summer, then falls

      iii. 1973 distinctly higher; 1971, 1975 distinctly lower (5)

4. Is this the whole story? Suppose we consider the total number of registrations, and divide to obtain % of all registrations that are emergencies (6)

   a. These data are more similar, and their plot has less patt (7)

b. We smooth these percentages and plot them (8)

   i. Shape is similar to before; peak in summer, low in winter

   ii. Hence conclusions are similar to conclusions from raw data

5. Note that these data had little trend or linear increase; we could not fit a line to them. In the next section, we analyze data with more pattern

6. Smoothing with CMU-DAP:
   Use function SMOOTH

450

Topic 2. Summarizing Time Plots

I. Basic Issue: Describing the trends in time series data

1. After smoothing the data, if there is evidence of time trends, we should transform the data (if necessary) and fit a line

2. Specific issues are

   a. Extrapolation: can we say anything about the data beyond the range that we have?

   b. Interpolation: Can we estimate a Y value for a time point lying between 2 time points for which we have data?

3. Data with substantial structure are much easier to extrapolate and interpolate than data that are mostly noise

4. Are there any monthly, seasonal, etc. trends? These are called periodicities, and if present, should be noted

5. Finally, how does (X_i, Y_i) relate to (X_{i-1}, Y_{i-1})?

II. Problem: Are there any problems unique to time series data?

1. With only one Y value for every X, equally spaced X's, trends are much more evident than with ordinary (X,Y) data

2. The study of periodicities and extrapolation and interpolation presents no difficulties; however one must use caution, because drawing conclusions from a data set is a "delicate" matter

III. Methods

1. We study another <u>example</u>: per capita expenditures for household electricity in the U.S., 1929-1972. Data are in hundred $/person (9)

2. Data reveal an exponential trend

3. Take log (Y) and find a reasonably linear trend (10)

4. Fitted line has equation: (11)
   $\log Y = 0.06\,X - 7.29$

451

5. Note cyclical pattern of residuals from line--this pattern is an example of a 20 year period: high in ≈ 1935, low in 1945, high in 1955, low in 1965

6. Residuals as a schematic plot look fine (12)

7. However, plotted against X, the periodicity is revealed! (13)

8. Extrapolation is reasonably easy, because the line is an adequate summary (11)

9. Interpolation is also straightforward, but remember periodicities!

10. A further <u>example</u>: Number of physicians in the United States, 1850-1973 (14)

    a. Data on U.S. population reveal that increase has not been constant with the population

    b. Suppose we plot (X_i, Y_i) as a function of (X_{i-1}, Y_{i-1}): we let our Y variable be Y_i and our X variable Y_{i-1}

    c. Plot is amazingly linear! (15)

        i. Y(t-1) can be used to make good predictions of Y(t)

        ii. The regression of Y(t) on Y(t-1) is a "lagged regression", and the knowledge that this regression is good is quite useful

452

Lecture 3-4
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 3-4 Outline
Topic 1		
Section IV		
1.	2	D.C. Hospital Emergencies
2.	3	D.C. Hospital Emergencies Scatterplot
3.a	4a 4b	Smoothed D.C. Hospital Emergency Registrations
3.g.iii	5	Smoothed Emergencies Plotted
4.	6	% Emergency Registrations
4.a	7	Plot of % Emergency Registrations
4.c	8	Plot of smoothed % Emergency Registrations
Topic 2		
Section III		
1.	9	Per Capita Expenditures for Household Electricity
3.	10	Log Per Capita Electricity Expenditure
4.	11	Resistant Line for log Expenditures
6.	12	Residuals from Resistant Line
7.	13	Residuals plotted against time
10.	14	Number of Physicians in US
10.c	15	Logged Regression Plot of Physicians

453

[1]

Lecture 3-4

Analysis of Time Series Data: Smoothing time series data with little structure, and studying and summarizing time series data with substantial structure.

LECTURE CONTENT:

1) Smoothing time plots, and identifying any periodicities in the data

2) Fitting lines to time plots, and extrapolating and interpolating apparent trends.

MAIN TOPICS:

1) Smoothing time plots.

2) Summarizing time plots.

451

(3-4)

[2]

D.C. General Hospital

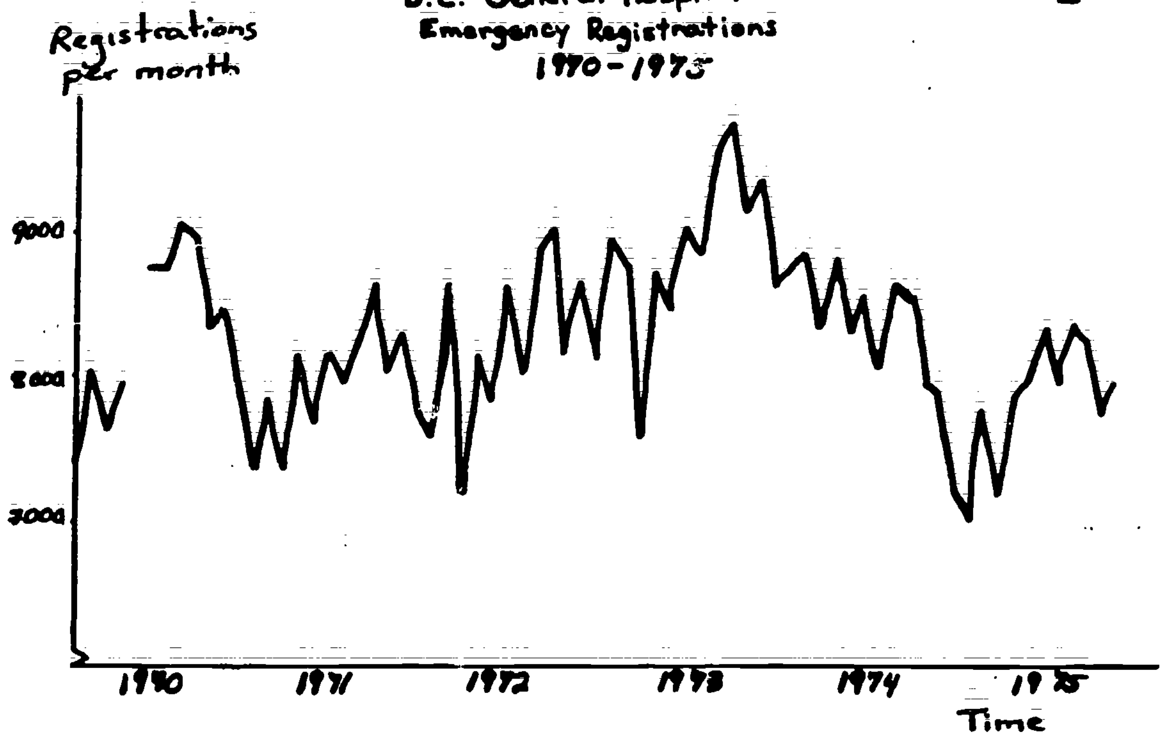
Emergency Registrations

(Calendar Year)

	1970	1971	1972	1973	1974	1975
Jan.	7476	7412	7620	9025	7463	7049
Feb.	8120	7979	8706	8815	8796	7874
Mar.	7706	7980	8214	7656	8449	7253
Apr.	8003	8288	8264	8849	8965	7912
May	NA	7753	7999	8545	8400	8069
June	8829	8256	8700	9166	8677	8448
July	8841	8074	8151	9257	8155	8035
Aug.	9103	8324	8977	9653	8737	8415
Sept.	9032	8778	9142	9852	8671	8323
Oct.	8443	8148	8294	9272	8071	7800
Nov.	8562	8846	8712	9425	7994	8096
Dec.	7914	8722	8239	8714	7306	NA

[3]

D.C. General Hospital
Emergency Registrations
1970 - 1975



QMPM

Running Medians of 3 for Smoothed D.C. General Hospital Emergency Registrations [4a]

1970	1st smooth	2nd smooth	3rd smooth	final smoothed
7476	7476			7476
8120	7706			7706
7706	8003	7854.5		7854.5
8003	7854.5	8003		8003
NA	8416			8406
88[illegible]	8835			8835
8841	8841			8841
9103	9032			9032
9032	9032			9032
8443	8562			8562
8562	8443	8562		8562
7914	7914			7914
1971				
7412	7914			7914
7979	7480	7914		7914
7480	7979	7753	7914	7914
8288	7753	7979		7979
7753	8256	8074		8074
8256	8074	8256		8256
8074	8256			8256
8324	8324			8324
8778	8324			8324
8148	8346	8324		8324
8346	8148			8148
7822	7822			7822
1972				
7620	7822			7822
8706	7620	7822		7822
7214	8706	7999		7999
8264	7999	8264	8151	8151
7999	8264	8151	8264	8264
8700	8151	8264		8264
8151	8700			8700
8977	8977			8977
9142	8977			8977
8294	8792			8792
8792	8294	8792		8792
8239	8792			8792

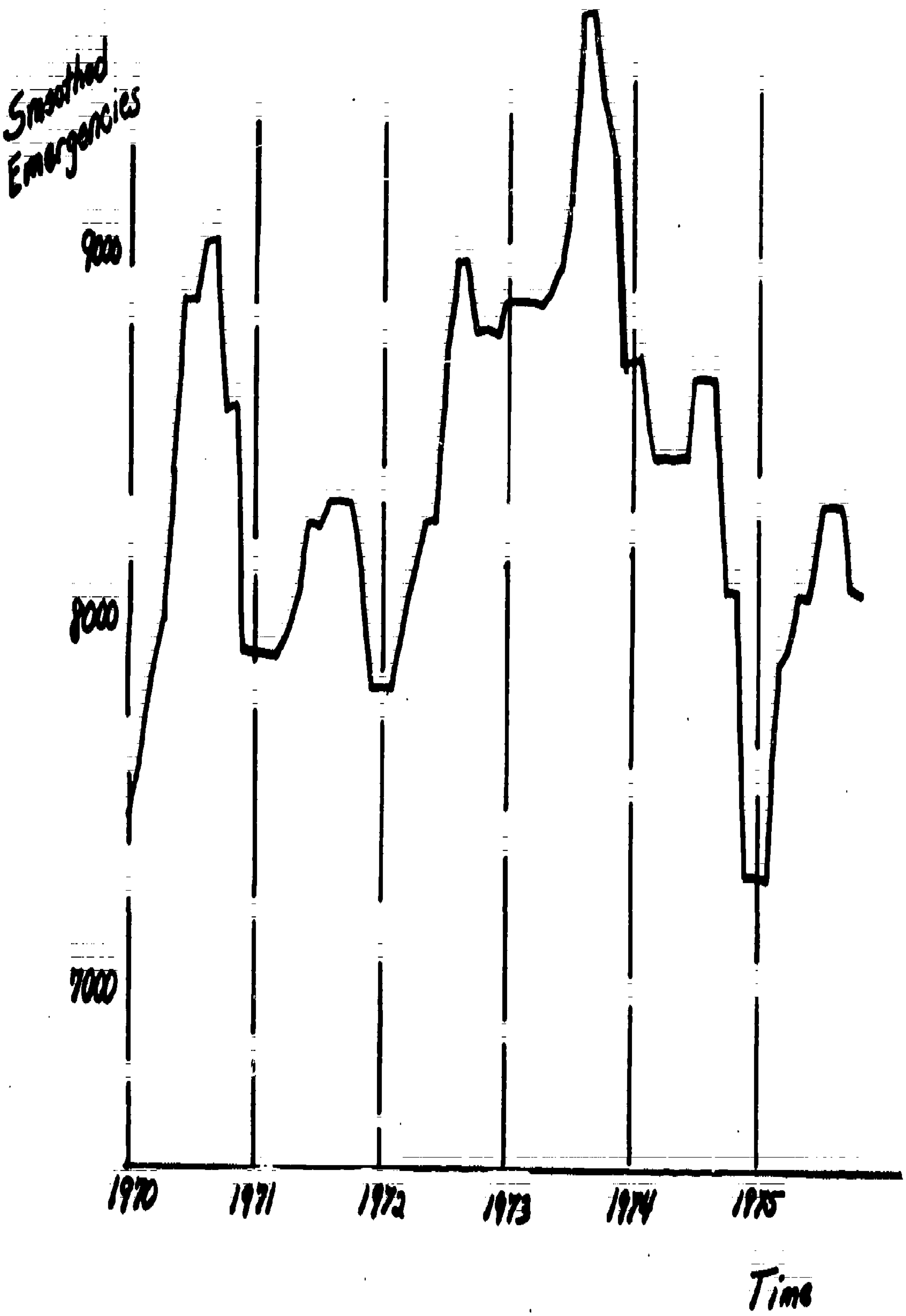
3-4

1973			
9025	8815		8815
8815	8815		8815
7656	8815		8815
8849	8545	8815	8815
8545	8849		8849
9166	8957		8957
8957	9166		9166
9653	9653		9653
9852	9653		9653
9272	9425		9425
9425	9272		9272
8714	8714		8714
1974			
7463	8714		8714
8996	8447	8714	8714
8447	8965	8447	8447
8965	8447		8447
8400	8447	8447	8447
8677	8400	8671	8447
8155	8677		8671
8737	8671		8671
8671	8671	8071	8671
8071	8071		8071
7994	7994		8071
7307	7307		7307
1975			
7049	7307		7307
7874	7253	7307	7307
7253	7874		7874
7912	7912		7912
8069	8069		8069
8448	8069		8069
8035	8415	8323	8323
8415	8323		8323
8323	8323		8323
7800	8096		8096
8096	8096		8096

3-4

D.C. General Hospital, Smoothed Emergencies

[5]



[J-4]

D.C. General Hospital [6]

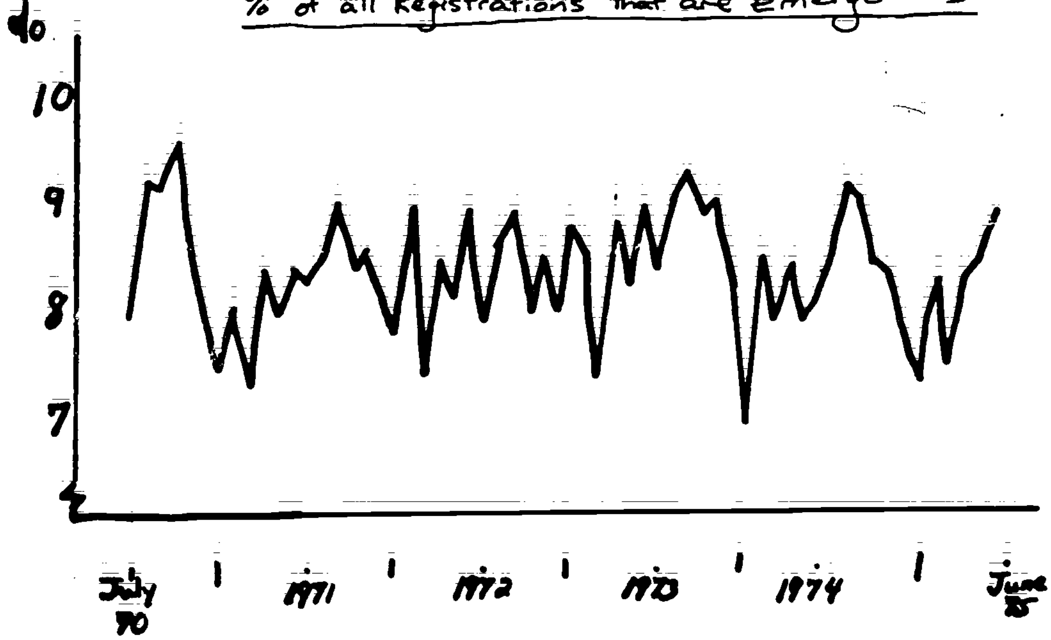
% of all Registrations that are Emergencies

(Fiscal Year)

	1971	1972	1973	1974	1975
July	8.93	8.24	7.86	8.38	8.54
Aug.	9.18	8.49	8.66	9.04	9.14
Sept.	9.12	8.96	8.82	9.22	9.07
Oct.	8.52	8.31	8.88	8.68	8.45
Nov.	8.64	8.52	8.48	8.82	8.37
Dec.	7.99	7.98	7.95	8.16	7.64
Jan.	7.48	7.77	8.70	6.98	7.37
Feb.	8.05	8.88	8.50	8.42	8.24
Mar.	7.55	7.36	7.38	7.91	7.59
Apr.	8.37	8.43	8.53	8.39	8.28
May	7.83	8.16	8.24	7.86	8.45
June	8.33	8.88	8.84	8.12	8.84

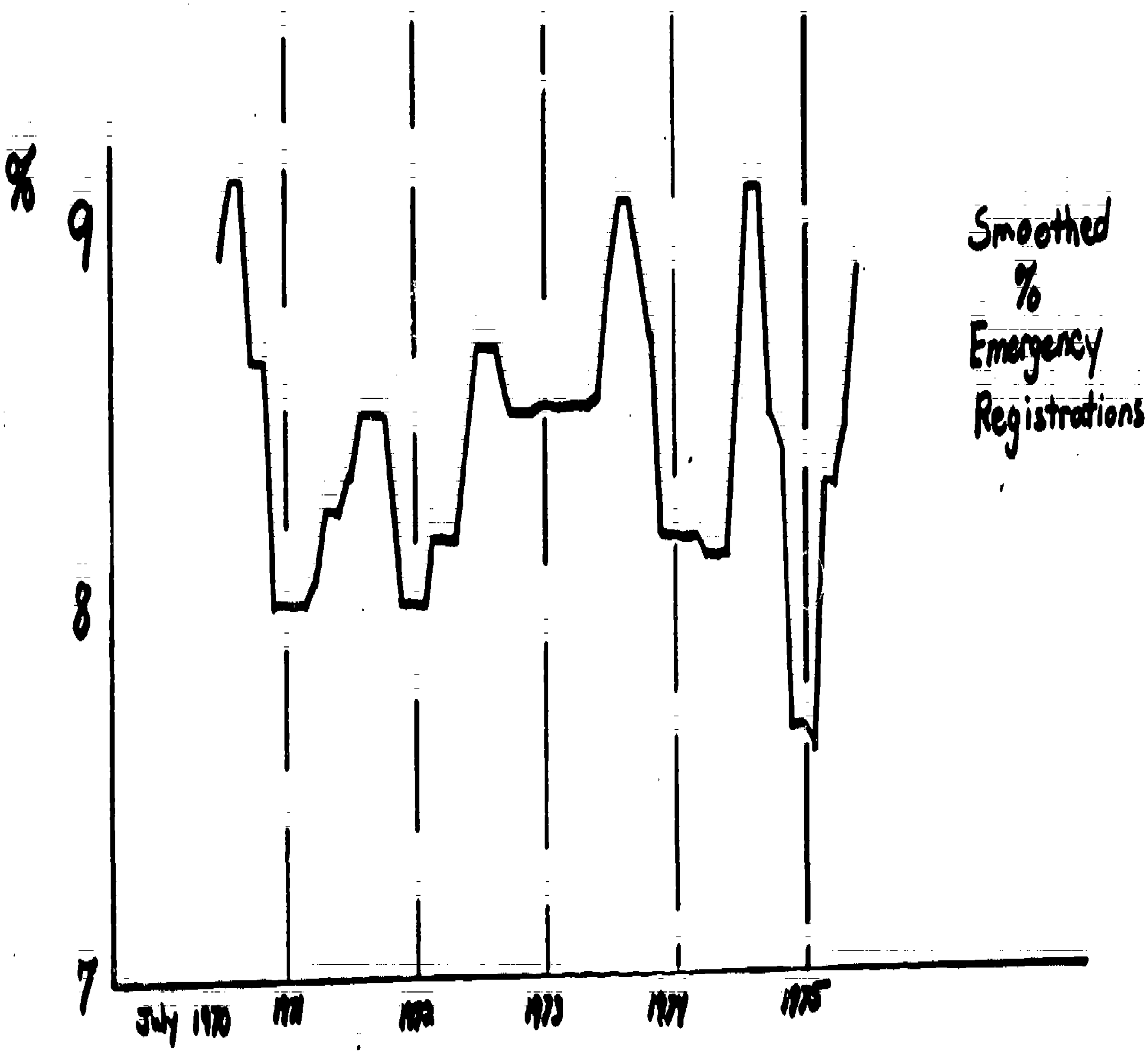
D.C. General Hospital [7]

% of all Registrations that are Emergencies



469

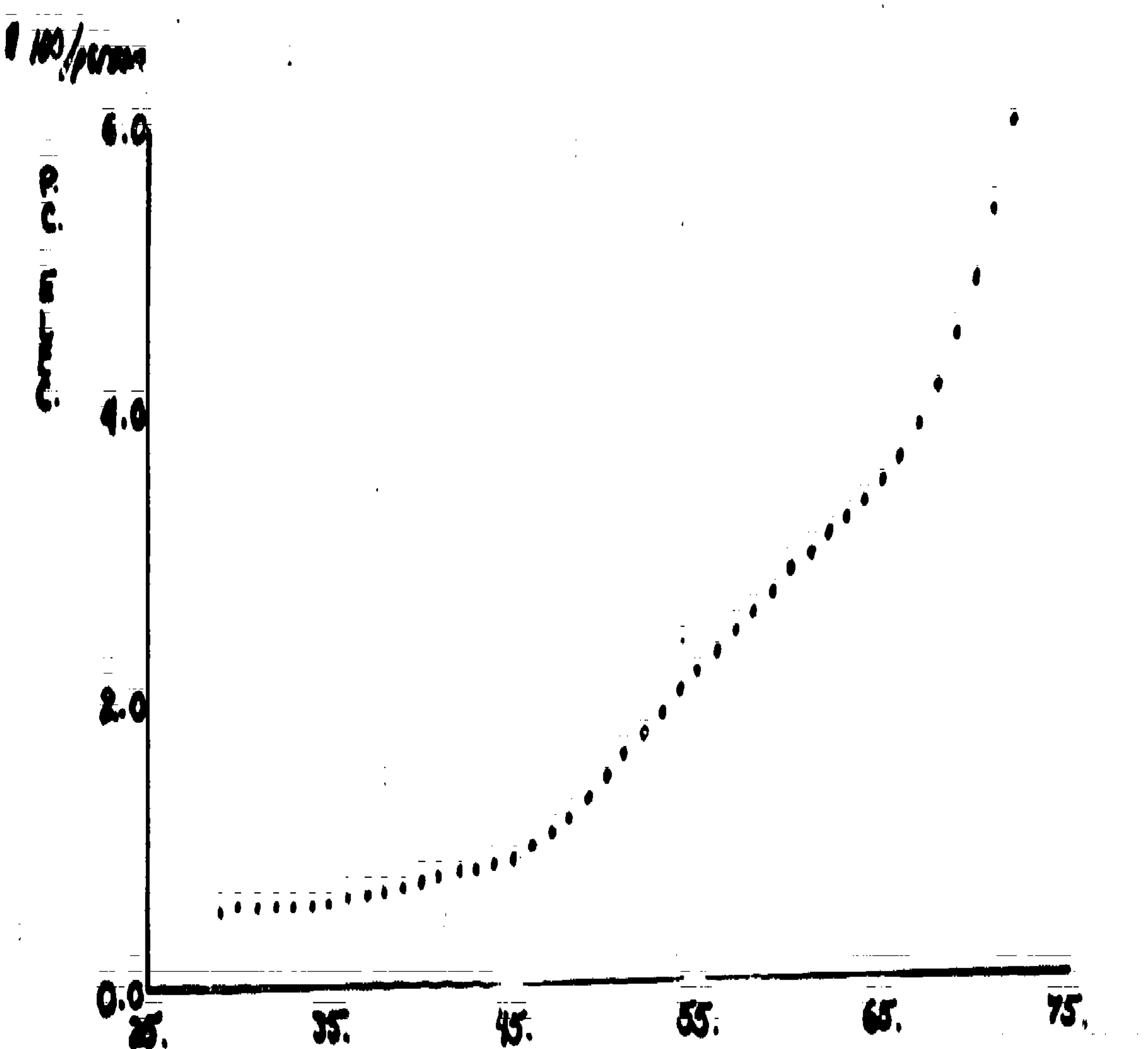
3-4



XVI.III.132

[3-4]

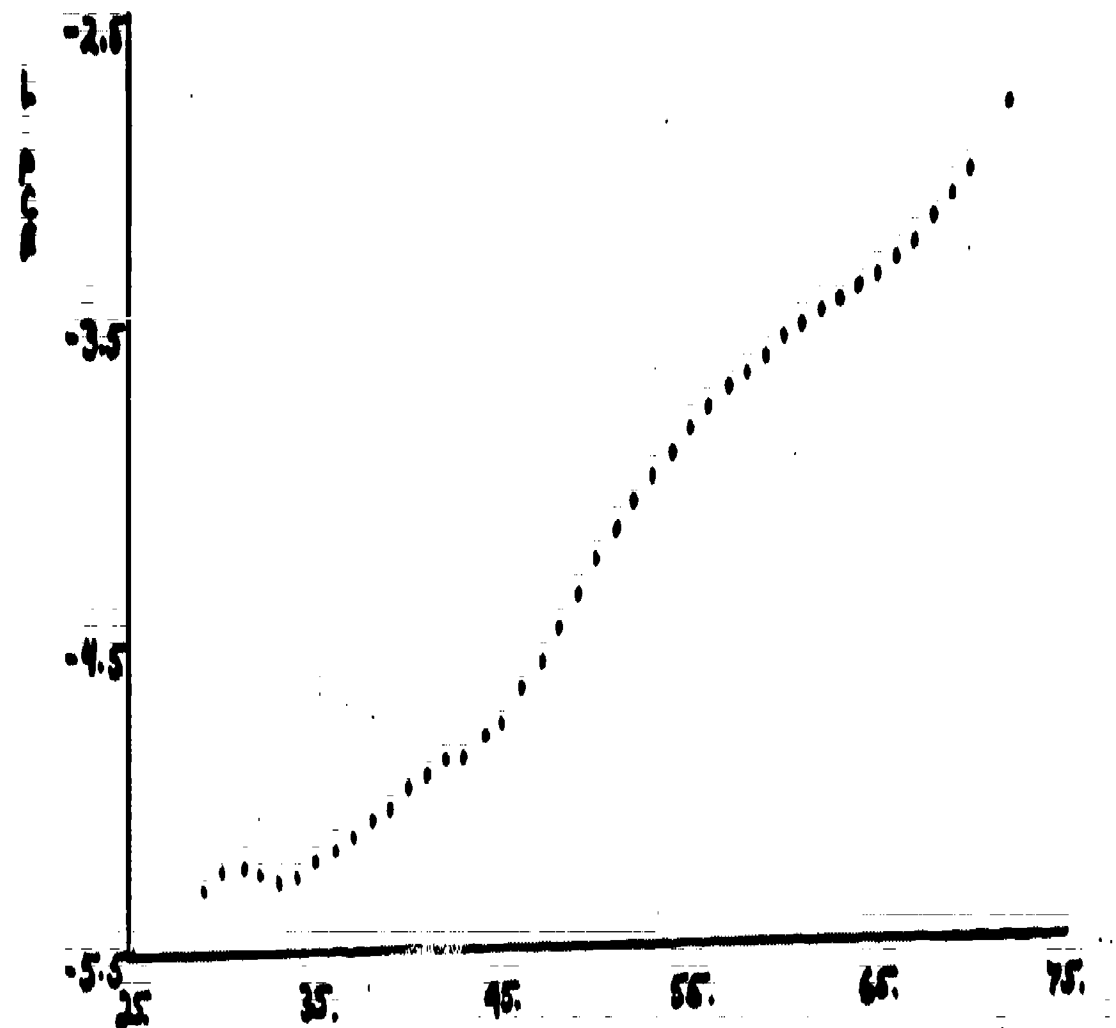
[9]



PER CAPITA EXPENDITURES FOR HOUSEHOLD ELECTRICITY
US 1929-1972

XVI.II.133

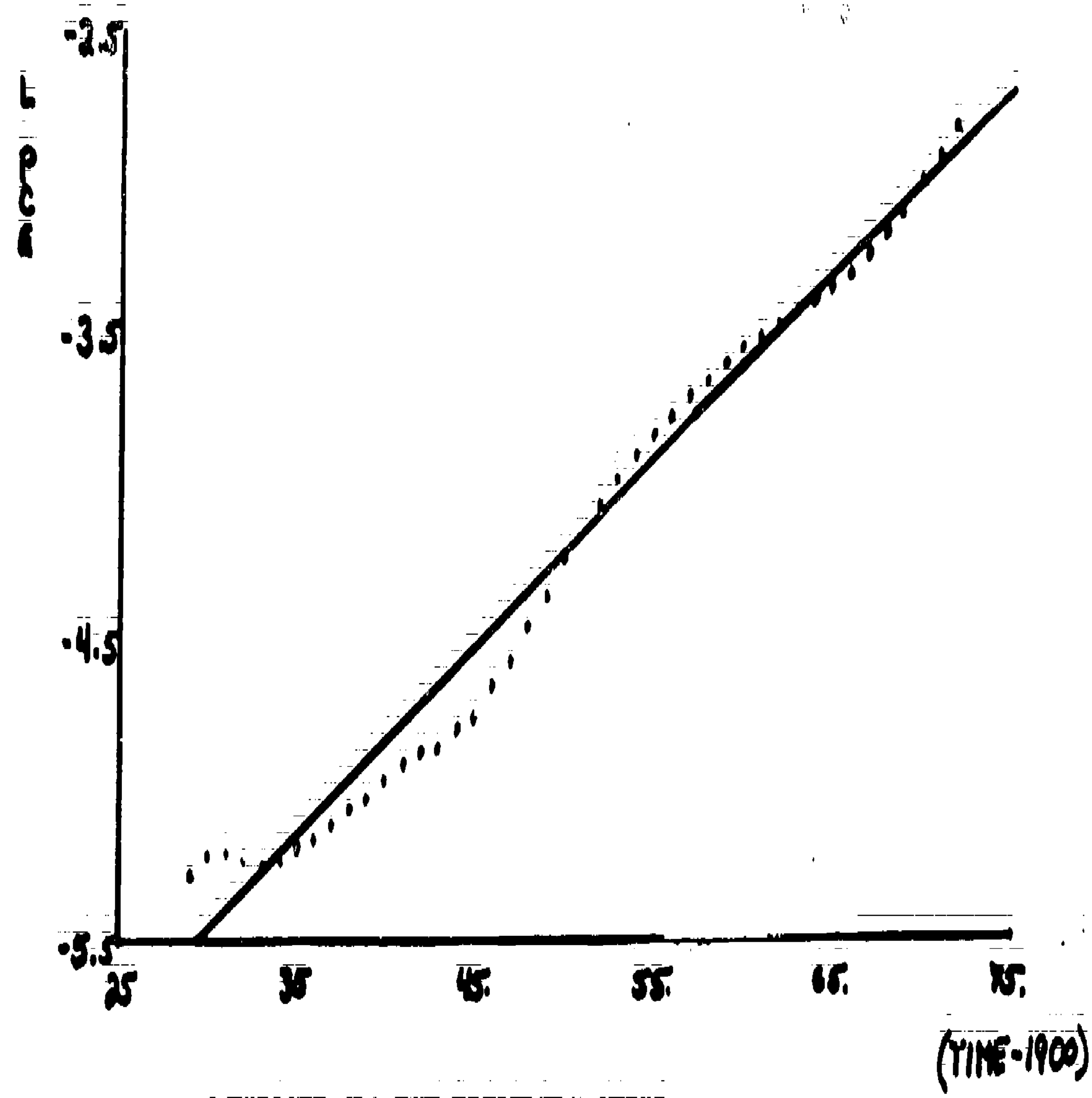
[3-4]



PLOT OF LOG OF PER CAPITA HOUSEHOLD ELECTRICITY EXPENDITURES (TIME-1900)

XVI.II.134

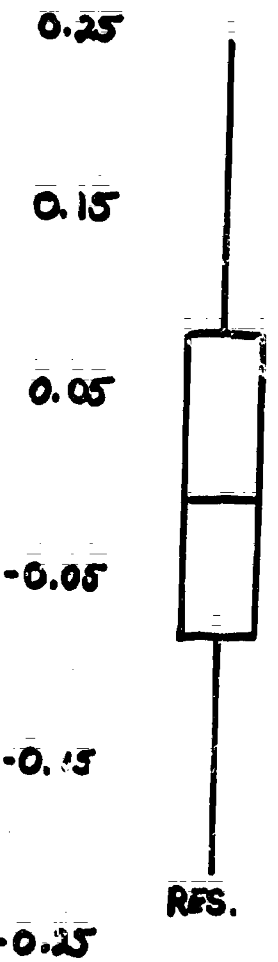
[11]



EQUATION IS: Y=0.060[illegible]X-7.2[illegible]

RESISTANT LINE FITTED TO LOG OF PER CAPITA HOUSEHOLD ELECTRICITY EXPENDITURES

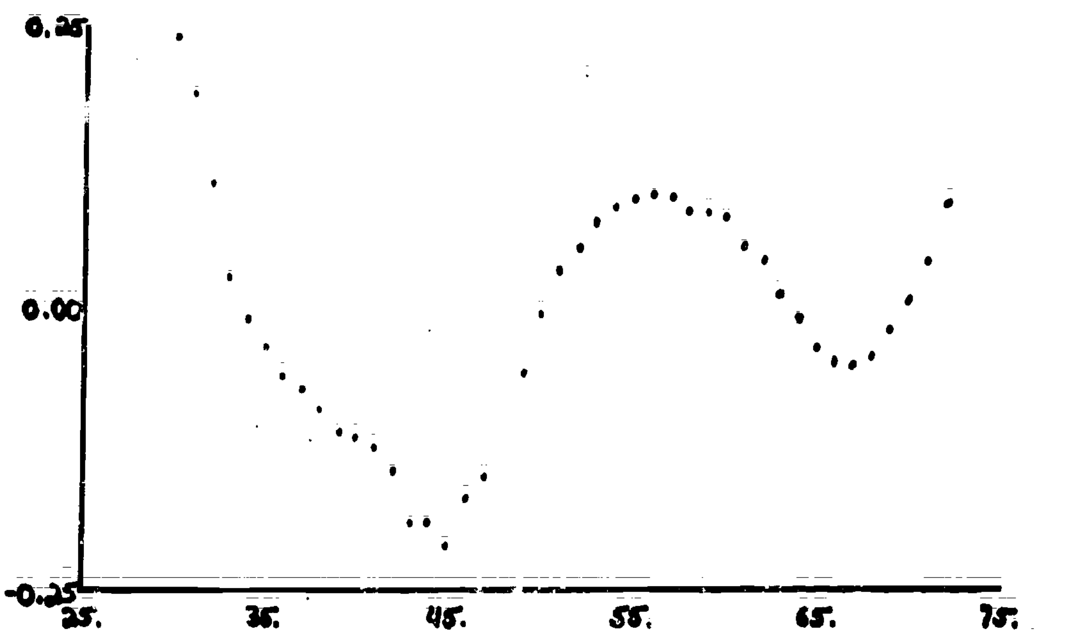
XVI.II.135



RESIDUALS FROM RESISTANT LINE FOR LOG ELEC EXPENDITURES

XVI.III.136

[13]

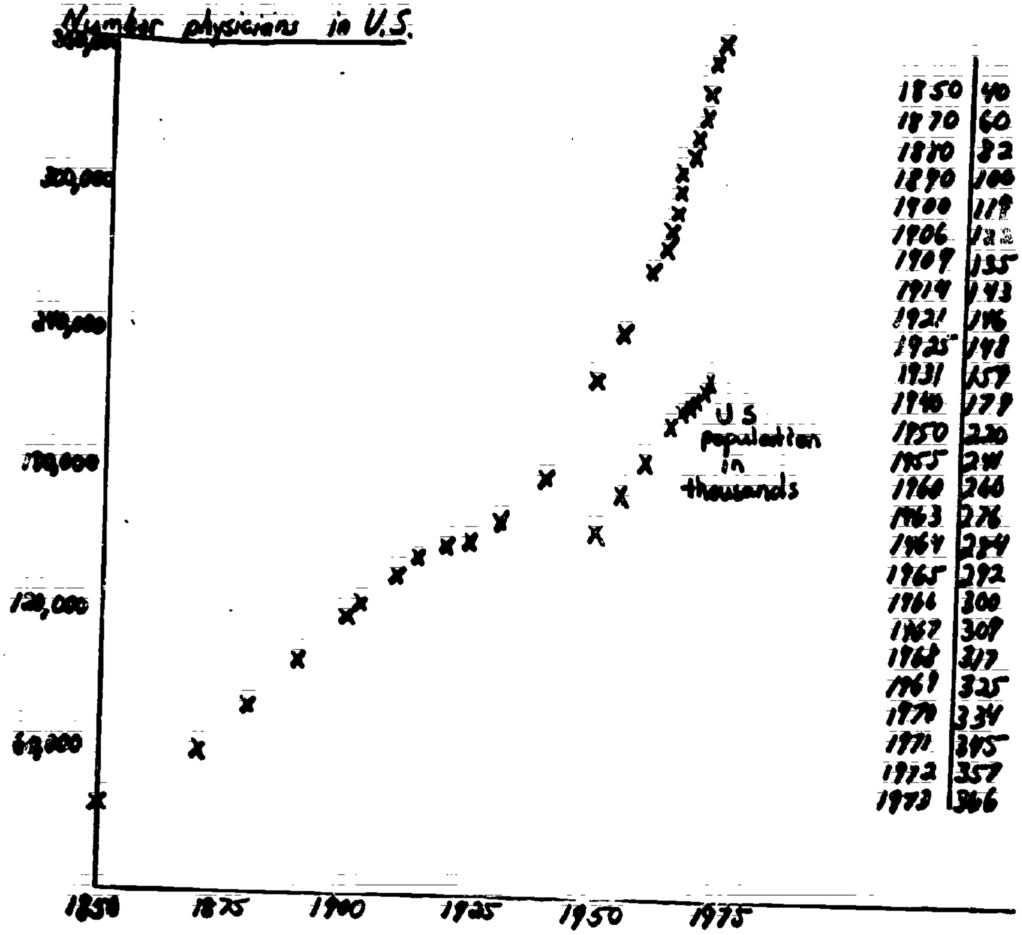


PLOT OF RESIDUALS AGAINST TIME (X VARIABLE)

XVI.II.137

[J-4]

Number physicians in U.S.

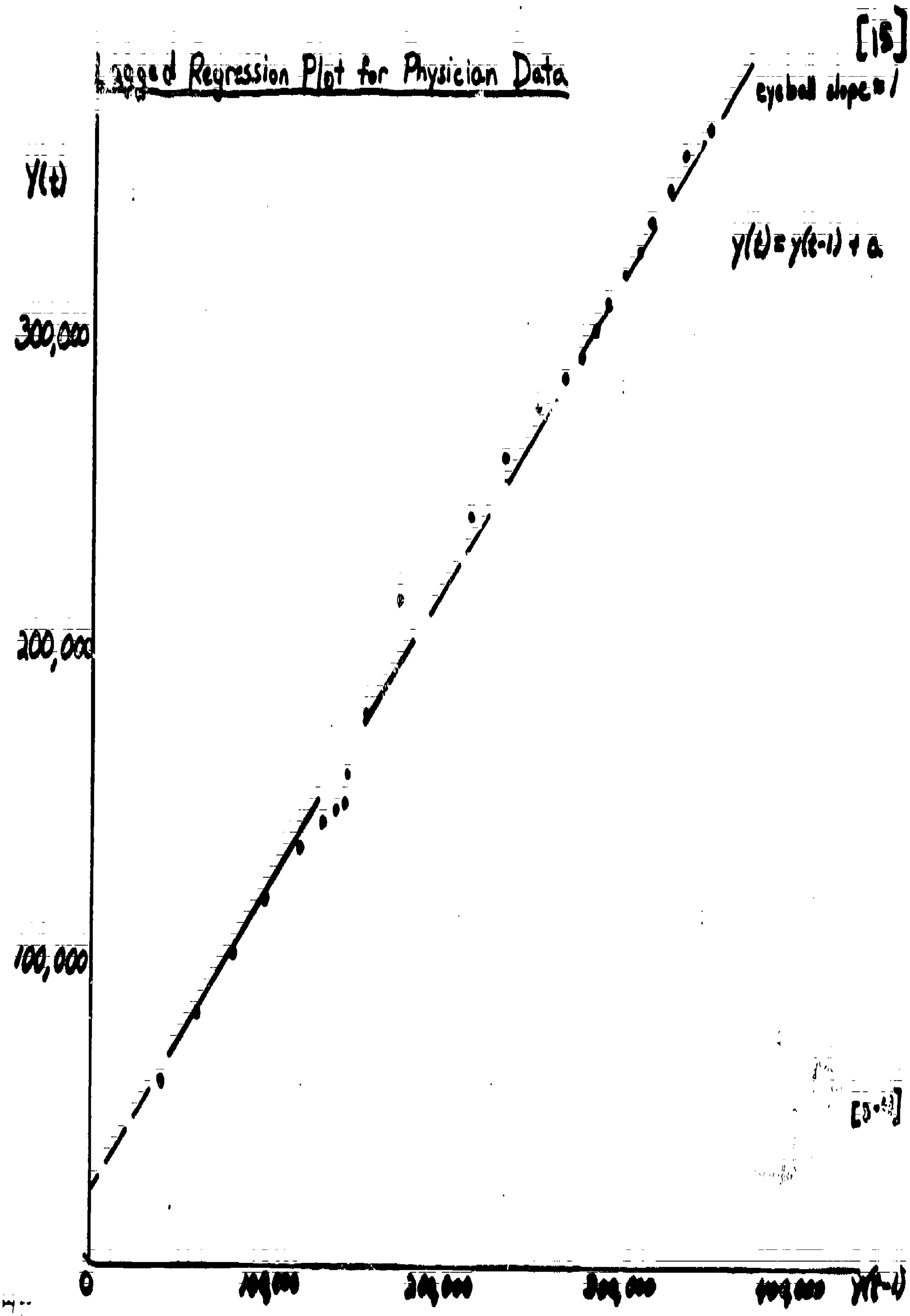


in '000s

1850	40
1870	60
1880	82
1890	100
1900	119
1906	[illegible]
1909	135
1914	143
1921	146
1925	148
1931	157
1940	179
1950	220
1955	241
1960	260
1963	276
1964	284
1965	292
1966	300
1967	309
1968	317
1969	325
1970	334
1971	345
1972	357
1973	366

XVI.II.138

[3-4]


[15]
Lagged Regression Plot for Physician Data
eyeball slope = 1
Y(t)
y(t) = y(t-1) + a
300,000
200,000
100,000
0
100,000
200,000
300,000
400,000
Y(t-1)

XVI.III.139

QMPM

Homework, Unit 3

1. The following data were collected for a study of intended, expected, desired, and ideal family size. The data below show the total number of births expected by married women 18-39 years old by age group and race for the years 1965-1972.

   (a) Considering each age group as a separate batch, construct parallel stem and leaf displays for each of the four batches. Calculate the five number summaries for each batch and display these under the corresponding stem-and-leaf display. What patterns (if any) do you observe?

   (b) Draw parallel schematic plots for the four batches. Do you find any additional differences in expected births?

   (c) Redraw the parallel schematic plot and connect the medians, hinges, and extreme values on this graph as done with the ordered multiple batches in class. What can you learn from the plot? Which graphic presentation ((a), (b), or (c)) makes these patterns (or lack of them) most obvious)?

   (d) Look again at the raw data. Suggest two other possible ways in which the data might be examined. (Hint: Above we explored a relationship between expected births and [illegible] of wives. What other possibilities are suggested by the raw data?)

   (e) What do you conclude about the relationship between expected births and a married woman's age? What implications can you draw from your analysis regarding the demand for elementary school teachers in the next 10 years?

2. The following data show the number of employees on non-agricultural payrolls for the years 1951-1973 for the 11 states comprising the "old south."

   (a) Consider each year as a batch. Calculate the 5 number summary for each batch. You may wish to cut the data values to tens of thousands. If you order the data (cut or raw) what unusual fact stands out? (Hint: identify the ordered data by state). Why might this be so?

   (b) Draw a parallel schematic plot for the twelve batches. What trends (if any) do you observe?

   (c) How does the number of employees in non-agriculture jobs change over time? What implications does your analysis have for the employment structure of the "old south?"

QMPM

(1) Total number of Births Expected by Married Women 18-39 Years Old by Age Group and Race 1965 to 1972

Age Group (yrs.)	21	27	32	37
Year & Race				
White				
1972	2.2	2.4	2.8	3.2
1971	2.4	2.6	2.9	3.2
1970	2.6	2.7	3.0	3.2
1967	2.9	3.0	3.2	3.2
1965	3.1	3.3	3.5	3.3
Black				
1972	2.4	2.8	3.7	4.0
1971	2.6	3.1	3.7	4.2
1970	2.9	3.2	3.8	4.1
1967	2.8	3.4	4.3	4.2
1965	3.4	4.0	4.4	4.1

(2) #Employees on Non-agricultural payrolls by state for the "old south", in thousands

State/Year	1951	1953	1955	1957	1959	1961	1963	1965	1967	1969	1971	1973
Alabama	662.8	692.7	702.9	754.8	764.4	774.6	812.5	886.5	951.8	1000.2	1021.9	1135.6
Arkansas	319.0	319.6	321.0	337.4	359.4	376.0	414.9	455.3	497.9	530.7	549.2	619.9
Florida	759.7	848.8	965.9	1152.7	1273.0	1333.9	1447.4	1619.1	1816.4	2069.9	2249.2	2756.5
Georgia	872.3	929.7	959.5	997.4	1030.1	1050.7	1139.7	1257.1	1394.7	1531.7	1602.9	1799.7
Louisiana	669.5	711.4	725.5	802.6	789.1	780.6	817.0	905.5	1005.0	1041.0	1064.3	1172.9
Mississippi	333.7	344.1	354.0	366.9	397.2	408.7	443.7	485.3	531.9	567.0	593.5	678.5
North Carolina	987.2	1023.7	1059.4	1101.3	1163.7	1209.1	1298.6	1431.2	1600.9	1747.0	1818.4	2014.7
South Carolina	505.8	543.8	533.0	545.0	566.8	587.0	630.6	686.1	754.4	819.8	862.6	984.0
Tennessee	805.9	852.6	867.6	886.8	907.0	934.0	1002.5	1108.5	1218.8	1309.8	1356.8	1534.8
Texas	2103.5	2224.7	2291.2	2450.2	2513.0	2544.1	2700.1	2925.3	3251.7	3599.2	3692.1	4146.4
Virginia	869.4	903.2	912.0	972.0	1000.5	1034.8	1123.8	1218.9	1330.2	1438.1	1558.0	1747.4

3. The following data show total school budgets for various towns in the Pittsburgh area and the total number of pupils in the corresponding school systems. (The Pittsburgh district is deleted since it is considerably larger than any of the others).

   (a) Considering the number of pupils as the X variable, make a scatterplot of the data. What can you learn from the plot?

   (b) Group the X values into mini-batches. (One possible grouping would be 0-1.9, 2.0-2.9, 3.0-3.9, 4.0-5.9, 6.0+. Draw a parallel schematic plot. What patterns (if any) do you observe?

   (c) Determine the conditional typical values. Plot these values on a separate graph. What trends (if any) do you see? How well do the conditional typicals summarize the information in the mini-batches?

   (d) Calculate the residuals from the conditional typicals and analyze as a single batch. Do the residuals indicate any "lack of fit?"

   (e) What relationship between school budget and number of pupils does your analysis suggest?

482

(3)

Total School Budget and Total Number of Pupils
For Area School Districts

District	Budget	Pupils
Allegheny Valley	$ 4,866,760	2,415
Avonworth	3,185,000	1,865
Babcock	4,795,500	2,922
Baldwin-Whitehall	14,248,615	8,515
Bethel Park	15,001,891	8,776
Brentwood	3,284,971	1,807
Carlynton	5,288,313	2,910
Chartiers Valley	10,658,743	5,715
Churchill	8,468,331	4,799
Clairton	3,430,498	1,956
Cornell	3,408,583	1,416
Deer Lakes	4,937,707	3,082
Duquesne	2,514,094	1,530
East Allegheny	6,310,000	3,479
Edgewood	1,655,497	859
Elizabeth-Forward	7,360,243	5,071
Fox Chapel	11,876,197	6,074
Gateway	14,717,000	8,508
General Braddock	4,851,796	2,668
Hampton	5,470,700	3,229
Highlands	8,781,000	5,338
Keystone Oaks	8,140,801	4,930

District	Budget	Pupils
McKeesport	$11,825,045	8,139
Montour	7,436,496	4,269
Moon	9,098,485	5,183
Mt. Lebanon	14,992,283	7,815
North Allegheny	13,308,965	7,665
North Hills	13,793,726	8,104
Northgate	4,822,199	2,693
Penn Hills	20,242,552	13,480
Plum	8,997,000	6,020
Quaker Valley	4,984,400	2,406
Riverview	3,382,934	1,986
Shaler	14,636,498	9,205
South Allegheny	4,522,672	3,171
South Fayette	2,502,741	1,283
South Park	3,932,238	2,382
Steel Valley	6,915,120	3,514
Sto-Rox	5,185,295	2,999
Swissvale	3,686,749	2,040
Turtle Creek	2,739,724	1,505
Upper St. Clair	9,862,247	5,349
West Allegheny	5,126,134	3,410
West Jefferson Hills	7,093,055	4,551
West Mifflin	9,706,972	5,808
Wilkinsburg	6,496,542	3,570

Source: Pittsburgh Press, July 10, 1976.

4. The following data show the violent crime index rate per 100,000 population for various American urban rapid transit systems in 1970.

   (a) Scatterplot the data (consider population the X variable). What pattern (if any) do you see?

   (b) A common transformation for populations is log. Looking at the raw data, do you think that this is an appropriate transformation? Why or why not?

   (c) Scatterplot the transformed data. What pattern (if any) do you see?

   (d) Fit a resistant line to the transformed data and plot it on the scatterplot from (c). Is it a good fit?

   (e) Calculate, plot (vs X), and examine the residuals. What do they tell you about the fitted line?

   (f) Polish the line once. Plot the polished line on the scatterplot from (c). Does it appear to be a better fit than the line fitted in (d)? Calculate, plot, and examine the residuals from the polished line.

   (g) Compare this residual plot to that in (e). Which shows a better fit?

   (h) What conclusions can you draw about the relationship between violent crime rate on transit systems and city population?

(4) Violent Crime Index Rate per 100,000 Population for Various American City Rapid Transit Systems in 1970

City/System	Violent Crime Index Rate	Center City Pop.	Log. Center City Pop.
Boston	861	628,215	5.798
Cleveland	1,077	738,956	5.869
Detroit	1,958	1,492,914	5.174
Oakland	1,078	358,486	5.554
Albany	105	280,032	5.447
Atlanta	779	497,426	5.697
Baltimore	2,112	895,222	5.952
Columbus	540	540,025	5.732
DC Metro	2,203	746,169	5.873
Denver	822	512,691	5.710
Ft. Worth	453	393,476	5.595
Indianapolis	481	746,613	5.873
Milwaukee	213	709,537	5.851
New Orleans	1,066	593,471	5.773
Portland	708	375,161	5.794
St. Louis	1,502	622,236	5.794
San Antonio	550	650,188	5.813
San Diego	286	695,790	5.842
Seattle	603	524,623	5.720
Ann Arbor	394	99,797	4.999
Billings	153	52,851	4.723
Chattanooga	579	113,003	5.053
Concord	30	30,022	4.477
Dayton	1,200	239,591	5.379
Everett	229	51,926	4.715
Lafayette	71	44,955	4.653
Orlando	792	97,565	4.989
Pueblo	271	96,746	4.986
Schenectedy	162	77,134	4.887
Syracuse	357	197,297	5.295
Tacoma	408	151,061	5.179

5. The following data show the number of reported cases of mumps per month for 1972 and 1973.

   (a) Scatterplot the data (remember, the X variable here is TIME). What trends (if any) do you see?

   (b) Smooth the data with running medians of 3.

   (c) Discuss the periodicity of the time series.

   (d) How many cases of mumps do you think were reported in March 1974? August 1974? December 1974? February 1975? September 1975?

   (e) What implications can you draw from your analysis regarding when during a year spot commercials should be run on TV to convince parents to get mumps vaccine shots for children?

Reported Cases of Mumps in the US

	1972	1973
January	9184	7160
February	8921	7349
March	10806	8306
April	9663	6434
May	9929	7404
June	5483	5045
July	2634	2030
August	1799	1357
September	1480	1068
October	2641	2456
November	5418	4759
December	6205	5751

Source: Center for Disease Control, Morbidity & Mortality, Vol. 23.

6. The following data give median incomes of persons 25 years old or older by years of school completed and by sex for the US in 1973.

(a) Make a scatterplot of root median income (Y) against years of school completed (X). What relationship (if any) do you observe?

(b) Calculate the regression line for income vs. years in school for MEN ONLY. Plot this line on the scatterplot you drew for (a). Does this line confirm or contradict the relationship you observe in (a)? Calculate and plot the residuals. What do you conclude about this model from the plot of the residuals?

(c) Calculate the regression line for income vs. years in school for WOMEN ONLY. Plot this line on the scatterplot you drew for (a). Does this line confirm or contradict the relationship you observe in (a)? Calculate and plot the residuals. What do you conclude about this model from the plot of the residuals?

(d) Calculate the regression line for income vs. years in school for BOTH MEN AND WOMEN (i.e., combine all the data into one batch). Plot this line on the scatterplot you drew for (a). Calculate and plot the residuals. How "well" does this line fit the data? Do you prefer this single line or the two individual lines found in (b) and (c), and why? How important is the person's sex in fitting a model to this data?

(e) What policy implications do you derive from your analysis concerning the differential status of men and women in the US?

Median Annual Income of Persons 25 years old and over by Years of School completed and by Sex for the USA 1973

Years of School Completed	Median Income	
	Men	Women
4	$4463	$1873
8	6371	2220
10	8622	2836
12	10832	3970
14	11670	4564
16	13939	6214
20	16027	8936

Source: U.S. Dept. of Commerce, Bureau of Census <u>Current Population Reports</u>, Series P-60, No. 97.

7. The following data show the number of reported cases of Venereal Disease (Gonorrhea and Syphilis) per year for 1962-1974.

(a) Scatterplot the data (remember, the X variable here is TIME). What trends (if any) do you see?

(b) Interpolate to find the number of cases in 1970 of Syphilis and Gonorrhea.

(c) Plot Y(t) vs Y(t-1). What does this plot tell you?

(d) How many cases of gonorrhea do you estimate will be reported in 1976? How many of syphilis?

(e) If you were designing new public health programs to reduce the incidence of venereal disease what comparative emphasis would you make over the next ten years regarding syphilis and gonorrhea? What aspects of your analysis would you use to convince a group of concerned lay people of the correcteness of your policy?

Reported Cases of VD per year for 1962-1974

Year	# cases syphilis (in 1000s)	# cases gonorrhea (in 1000s)
1962	126	264
1963	124	278
1964	114	301
1965	113	325
1966	105	352
1967	102	405
1968	96	465
1969	92	535
1971	96	670
1972	91	767
1973	87	843
1974	84	899

Source: Center for Disease Control, Morbidity & Mortality vols. 20, 23 #53

Homework Solutions
Unit 3

1. # Births expected by wives 18-39 years old by age.

(a)

```
21 yrs. old        27 yrs. old        32 yrs. old        37 yrs. old

Unit = .1          Unit = .1          Unit = .1          Unit = .1

2 | 244            2 | 4              2 |                2 |
  | 66899            | 678              | 89               |
3 | 14             3 | 01234          3 | 02             3 | 22223
                                        | 5778
4 |                4 | 0              4 | 34             4 | 01122

      2.7               3.05               3.6                3.65
 2.4       2.9     2.7       3.3     3.0       3.8      3.2        4.1
 2.2       3.4     2.4       4.0     2.8       4.4      3.2        4.2
```

Apparent trend towards <u>increase</u> in expected births with <u>increasing</u> age.

(b) (See plot) Same trend is more apparent. (Rate of increase levels off. Variability decreases.)

(c) (See plot) The plot with the connecting lines shows the same trends. It is a matter of opinion which plot makes these trends most obvious; probably the best case can be made for plot (c), with the connecting lines.

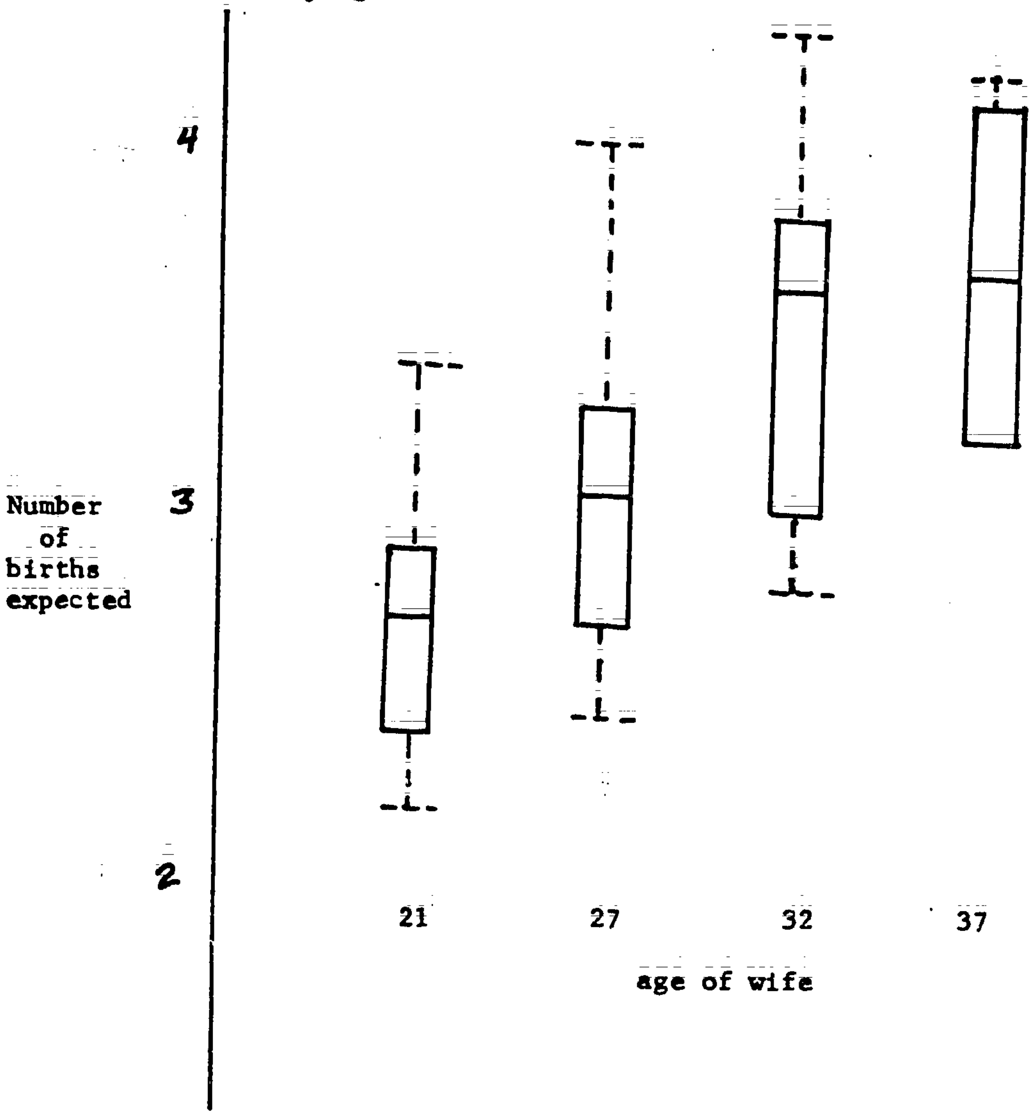
(d) Expected births vs. race (Black > White trend).
Expected births vs. year (<u>Decreases</u> with year).

(e) As a woman's age increases (and also the number of children <u>she already has</u>) the total number of children she <u>expects</u> to have increases.

Since we do not know from these numbers whether the number of children of elementary school age is increasing (over time) we cannot use these numbers for predicting the demand for elementary school teachers in the future.

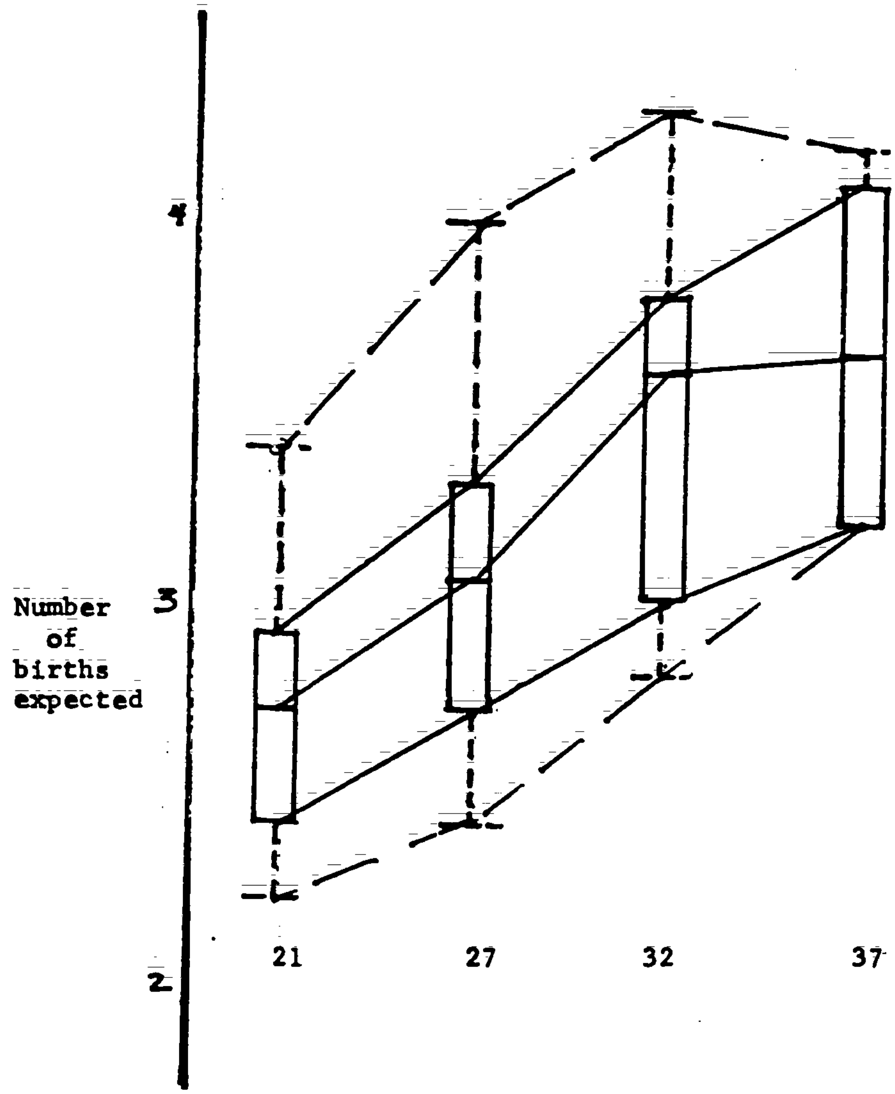
But note that expected family size is certainly less in 1972 than in earlier years; hence, the number of elementary school age children will probably decrease in the future, causing a decrease in demand for teachers.

1.b) Number of births expected by wives 18-39 years old by age



491

1.c) Number of births expected by wives 18-39 years old by age



hinges and median	solid line
extremes	dashed line

2(a) # Employees on Non agricultural payrolls values ordered and cut (tens of thousands)

Year	1951	1953	1955	1957	1959	1961	1963	1965	1967	1969	1971	1973	State
extreme	31	31	32	33	35	37	41	45	49	53	54	61	(Ark)
	33	34	35	36	39	40	44	48	53	56	59	67	(Miss)
hinge	50	54	53	54	56	58	63	68	75	81	86	98	(SC)
	66	69	70	75	76	77	81	88	95	100	102	113	(Ala)
	66	71	72	80	78	78	81	90	100	104	106	117	(La)
median	75	84	86	88	90	93	100	110	121	130	135	153	(Tenn)
	80	85	91	97	100	103	112	121	133	141	155	174	(Va)
hinge	86	90	95	99	103	105	113	125	139	153	160	179	(Ga)
	87	92	96	110	116	120	129	143	160	174	181	201	(NC)
	98	102	105	115	127	133	144	161	181	206	224	275	(Fla)
extreme	210	222	229	245	251	254	270	292	325	359	369	414	(Tx)

5 number summaries

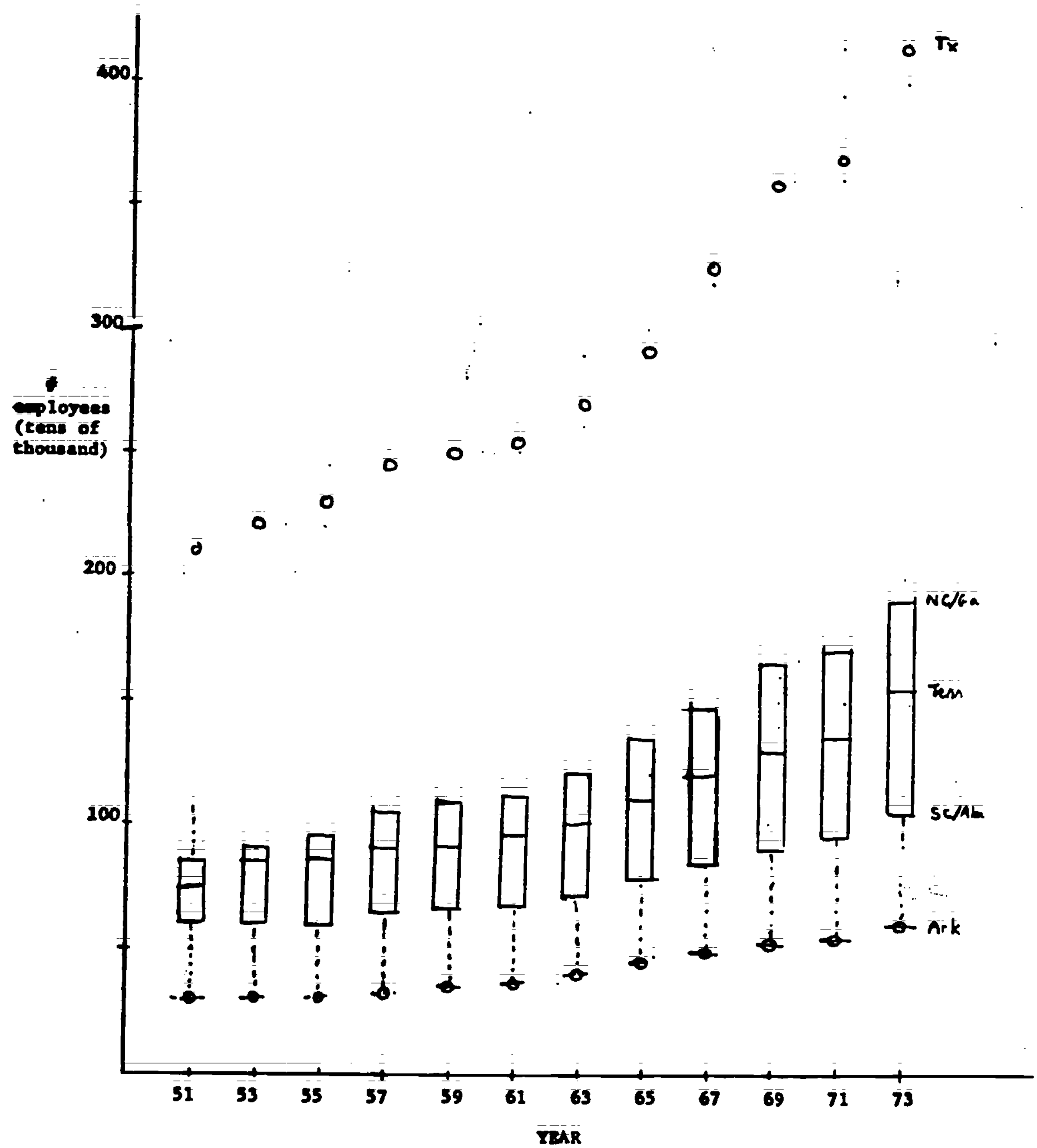
extreme	31	31	32	33	35	37	41	45	49	53	54	61
hinge	58	61.5	61.5	64.5	66	67.5	72	78	85	90.5	94	105.5
median	75	84	86	88	90	93	100	110	121	130	135	153
hinge	86.5	91	95.5	104.5	109.5	112.5	121	134	149.5	163.5	170.5	190
extreme	210	222	229	245	251	254	270	292	325	359	369	414
midspread	28.5	29.5	34.0	40.0	43.5	45.0	49	56	64.5	73.0	76.5	84.5
3/2 H	43	45	51	60	65	67	74	84	96	109	115	127

Note that relative ranking of states remains same (possibly reflects total population) each year.
Also that spread increases.

QMPM

2) b

Employees on NON agricultural payrolls
(upper whiskers and adj. values not shown)



2(b) Note that the # employees on non-agricultural payrolls increases over time for all of the states (in fact the # employees on non-agricultural payrolls roughly doubles between 1951 and 1973).

(c) See also (b) above. Since the rise in # employees on nonagricultural payrolls may be due "only" to the general rise in population, we have insufficient data to draw conclusions about the changing employment structure (if indeed it is changing) in the old south. We might wish to examine possible changes by comparing the annual rate of increase of # employees on nonagricultural payrolls with either rate of increase of total population (by state) or rate of increase of # employees on agricultural payrolls, or even both.

3(a) School budget clearly increases with the number of pupils in the school system. There is a strong linear pattern.

(b) The trend noted in (a) above is also striking in this plot.

(c) Use the median of each mini-batch. The plot looks like that in (b) but with less information (we lose information on the spread of each mini-batch). The pattern is still apparent.

(d)

Lo	-52.5
-2	43
-1	51521
-0**	664909534194
0**	0122200134081403477
1	0402
2	0
Hi	60.0

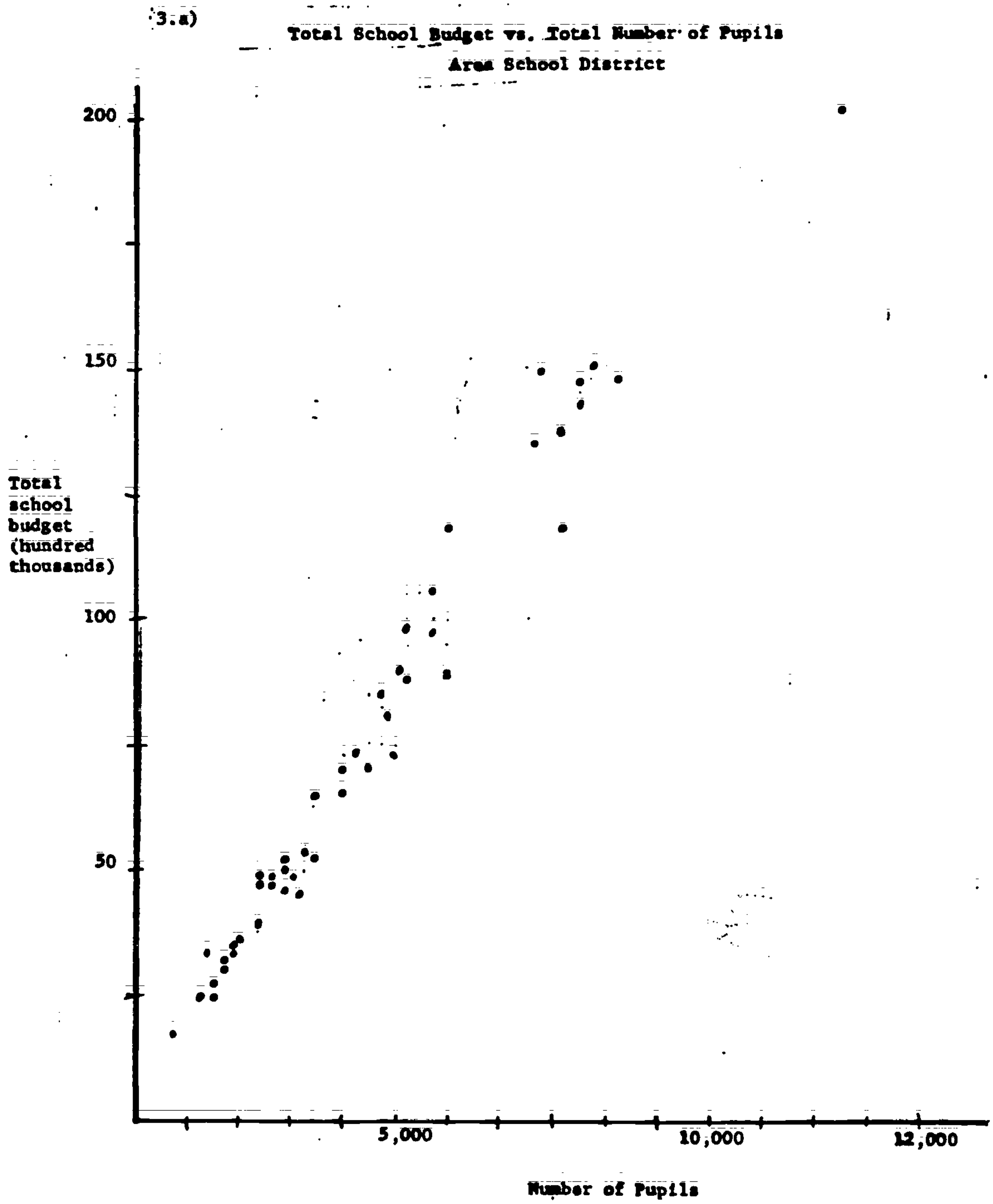
5 Number Summary of Residuals

1	E	-52.5
12	H	-6.0
23	M	0
12	H	4.0
1	E	60.0

Residuals are relatively symmetric. They cluster around zero, except for one high and one low outlier.

(e) Clearly budget increases with the number of students. Using a resistant line with one step of polish, the relationship is summarized as approximately y = 20000 X

3.a)


Total School Budget vs. Total Number of Pupils
Area School District
200
150
100
50
Total school budget (hundred thousands)
5,000
10,000
12,000
Number of Pupils

499

3.b,c)

Minibatches, ordered & summarized

Batch	1	2	3	4	5
# pupils	0-1.9	2.0-2.9	3.0-3.9	4.0-5.9	6.0-up
median # pupils (thousands)	1.0	2.5	3.5	5.0	8.1
budgets (ordered)	16.5	36.8	45.2	70.9	89.9
	25.0	39.3	49.3	73.6	118.2
	25.1	47.9	51.2	74.3	118.7
	27.3	48.2	54.7	81.4	133.0
	31.8	48.5	63.1	84.6	137.9
	32 8	48.6	64.9	87.8	142.4
	33.8	49.8	69.1	90.9	146.3
	34.0	51.8		97.0	147.1
	34.3	52.8		98.6	149.9
				106.5	150.0
					202.4
extreme	16.5	36.8	45.2	70.9	89.9
hinge	25.1	47.9	50.25	74.3	125.7
median	31.8	48.5	54.7	86.2	142.4
hinge	33.8	49.8	64.0	97.0	148.5
extreme	34.3	52.8	69.1	106.5	202.4
midspread	8.85	7.2	13.75	22.7	22.8

Residuals from Medians:

	1	2	3	4	5
				-15.3	-52.5
	-15.3	-11.7	-9.5	-12.6	-24.2
	- 6.8	- 9.2	-9.5	-11.9	-23.7
	- 6.7	- 0.6	-5.4	- 4.8	- 9.4
	- 4.5	- 0.3	-3.5	- 1.6	- 4.5
	0	0	0	+ 1.6	0
	+ 1.0	+ 0.1	+8.4	+ 4.7	+ 3.9
	+ 2.0	+ 1.3	+10.2	+10.8	+ 4.7
	+ 2.2	+ 3.3	+14.4	+12.4	+ 7.5
	+ 2.5	+ 4.3		+20.3	+ 7.6
					+60.0

490

4. (a) The crime rate index increases with center city population. There is, however, quite a bit of variation--even within cities of similar size.

(b) The data have a shape which suggests that we should transform X down. Log (X) is not unreasonable.

(c) The crime rate index increases with log center city population in a very linear fashion.

(d) The unpolished resistant line is

$$\text{Crime} = -4381.6211 + (945.2319)\ (\text{Log Population})$$

(e) A very noticeable trend in the residuals is that the absolute value of the residual increases with the size of the log center city population. (Note the increasing divergence of the data from the fitted line as log(pop) increases. Otherwise the fit is "pretty good")

(f) The once-polished resistant line is

$$\text{Crime} = -3629.6418 + 805.4307\ (\text{Log Population})$$

There is a slight difference between the unpolished and polished lines, but the fit appears to be about the same.

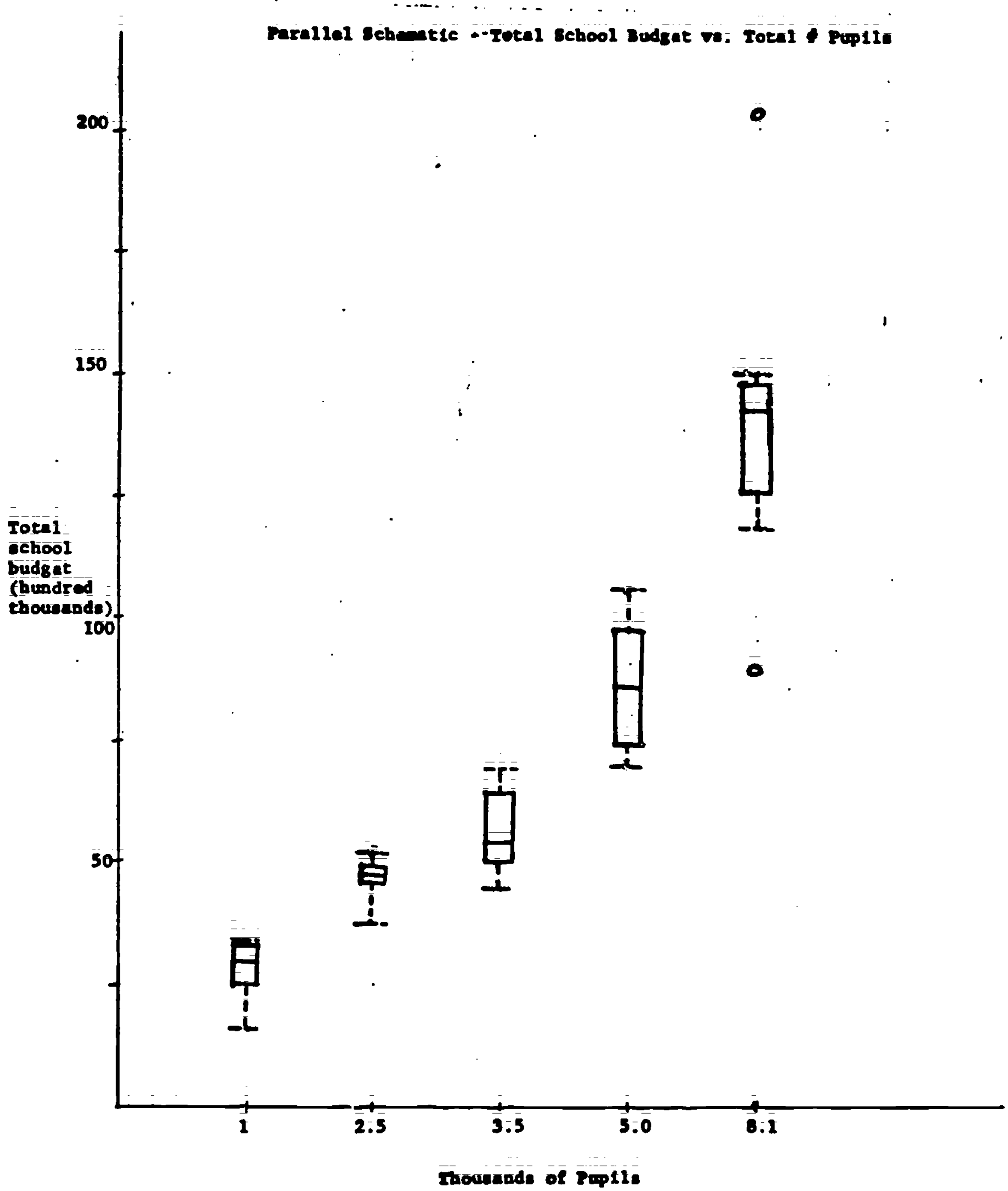
(g) Same comments as (e). The two residual plots appear to be very similar.

(h) The violent crime index rate on transit system increases with population. This relationship can be described approximately by the fitted model (resistant line)

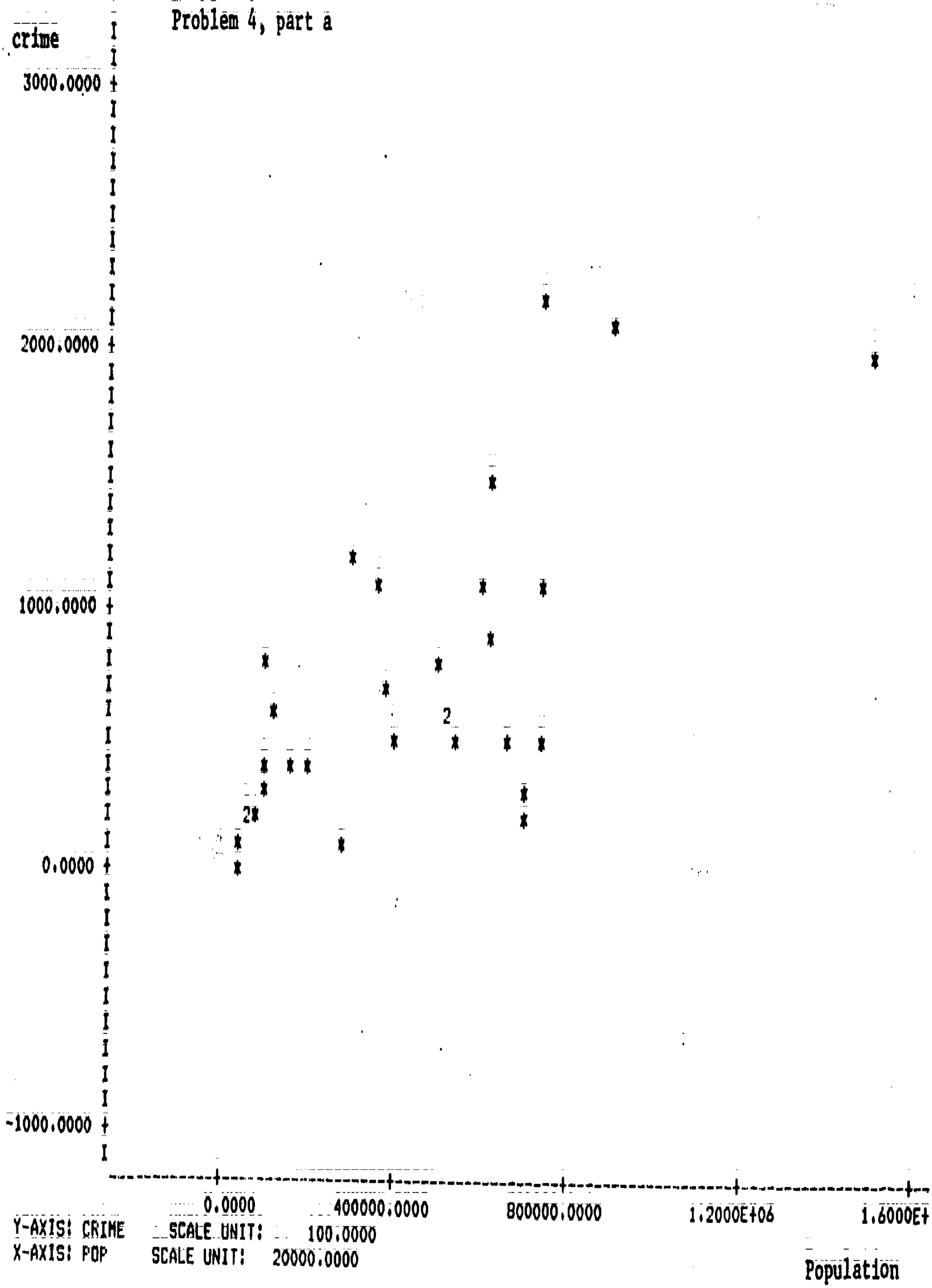
$$\text{Crime} = -3629 + 805\ (\text{Log(pop)}).$$

The model fits best for areas with smaller populations.

509


Parallel Schematic --Total School Budget vs. Total # Pupils
200
150
100
50
Total school budget (hundred thousands)
1
2.5
3.5
5.0
8.1
Thousands of Pupils

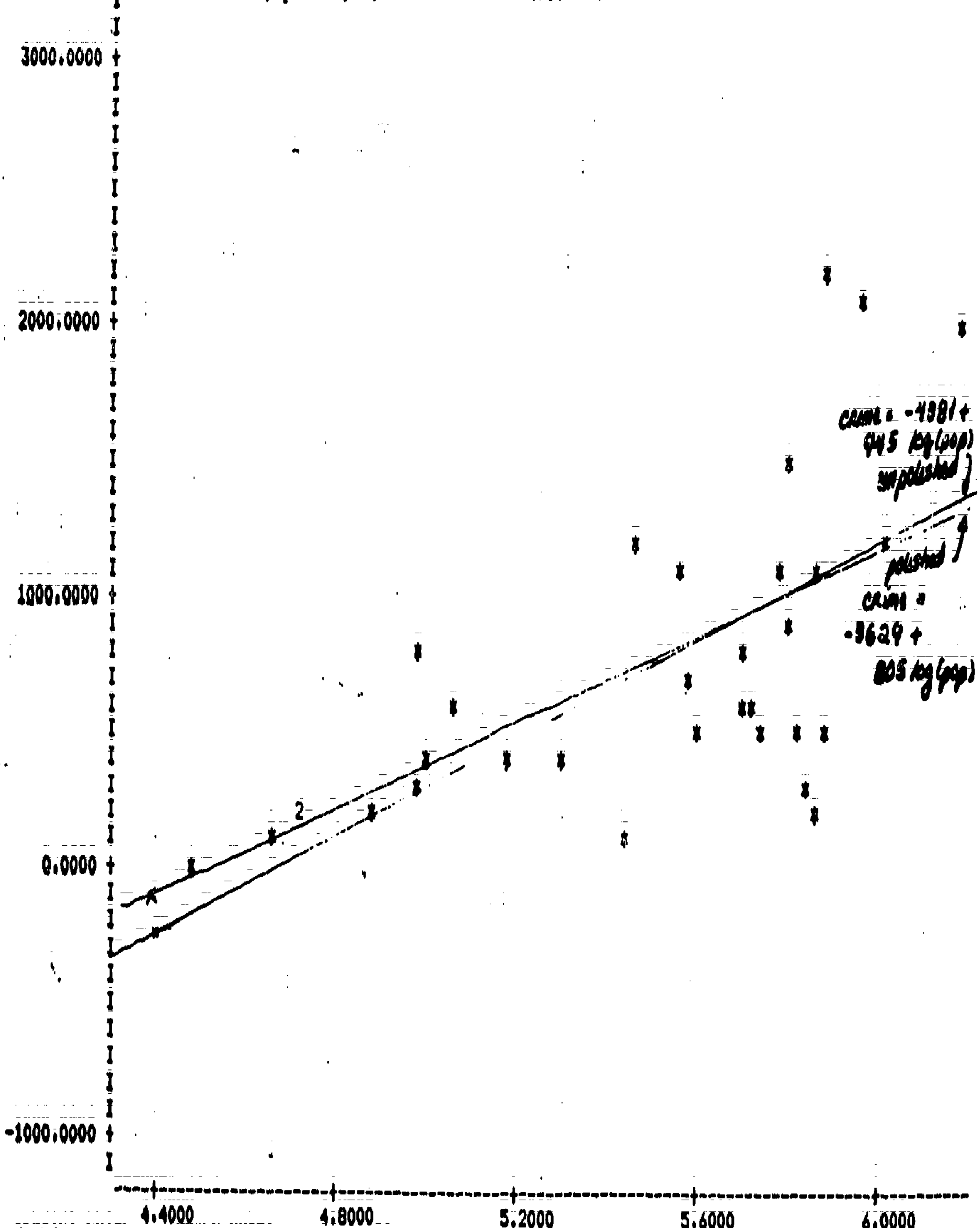
:PLOT CRIME VS POP



Y-AXIS: CRIME SCALE UNIT: 100.0000
X-AXIS: POP SCALE UNIT: 20000.0000

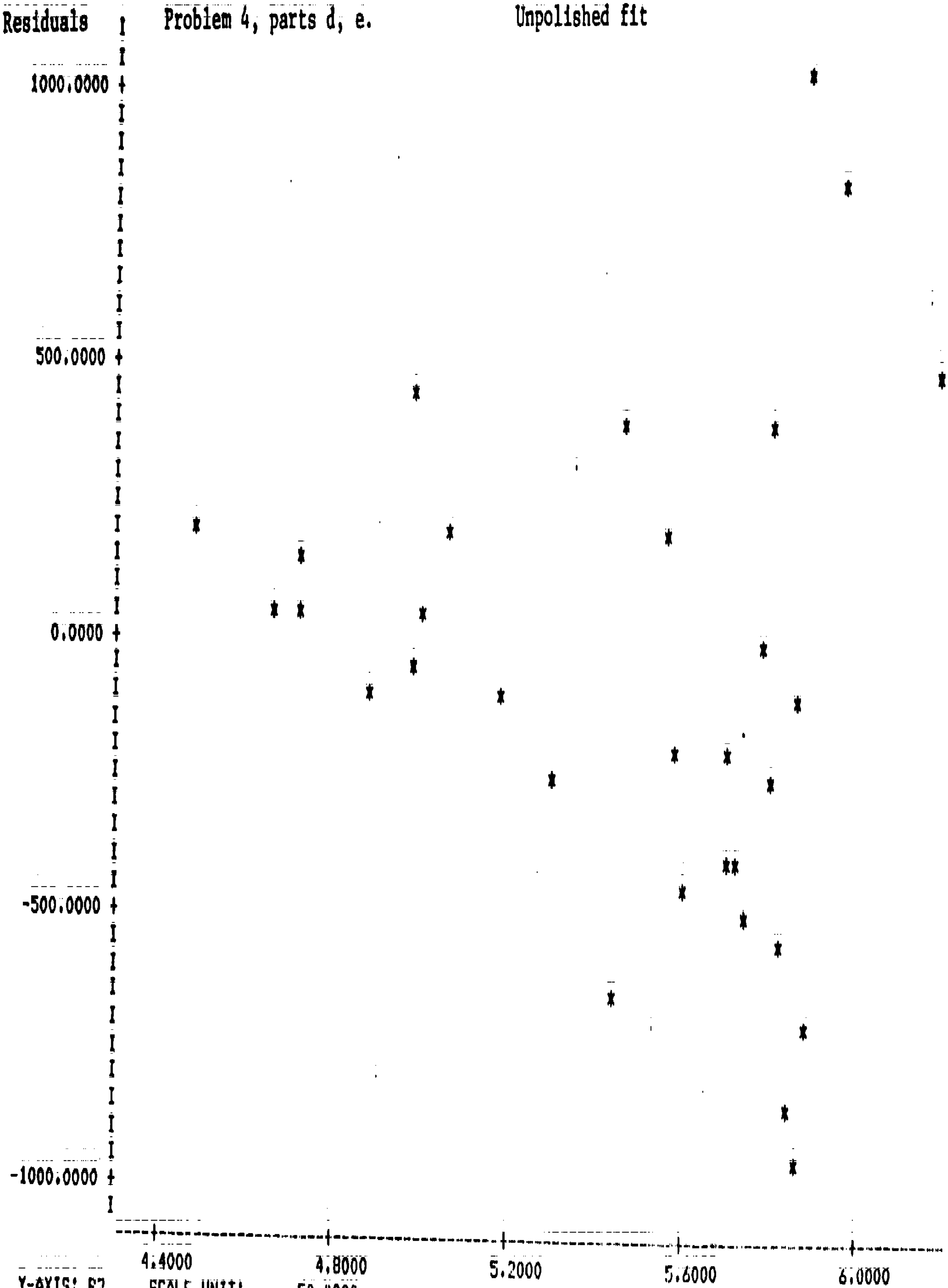
:PLOT CRIME VS LPOP

Problem 4, part c, d, f   Fitted Lines



Y-AXIS: CRIME   SCALE UNIT: 100.0000
X-AXIS: LPOP   SCALE UNIT: 0.0200

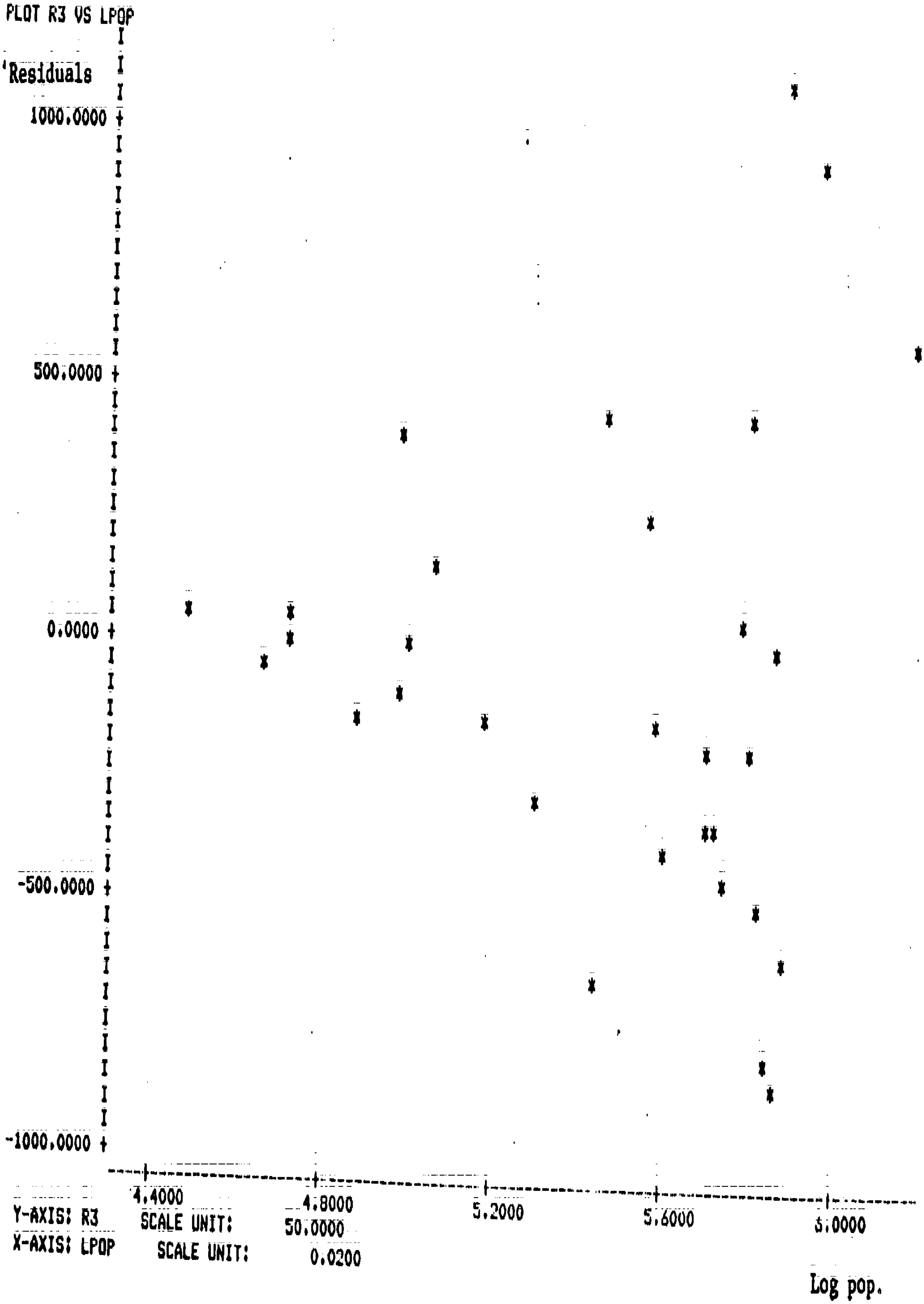
PLOT R2 VS LPOP



Y-AXIS: R2 SCALE UNIT: 50.0000
X-AXIS: LPOP SCALE UNIT: 0.0200

Log pop.

Problem 4, part f Polished fit



5. (a) There are fewer cases of mumps in 1973 than in 1972, and there are monthly or seasonal effects.

(b). Reported Cases of Mumps.

Month	Data	Smoothed Once
J 72	9,184	9,184
F	8,921	9,184
M	10,806	9,663
A	9,663	9,929
M	9,929	9,663
J	5,483	5,483
J	2,634	2,634
A	1,799	1,799
S	1,480	1,799
O	2,641	2,641
N	5,418	5,418
D	6,205	6,204
J 73	7,160	7,160
F	7,349	7,349
M	8,306	7,349
A	6,434	7,404
M	7,404	6,434
J	5,045	5,045
J	2,039	2,039
A	1,357	1,357
S	1,068	1,357
O	2,456	2,456
N	4,759	4,759
D	5,751	4,751

(c) Lows occur in summer months (July, August, September, (October))

Highs occur in winter/spring months (December, January, February, (March), (April).

Note the decreasing overall trend.

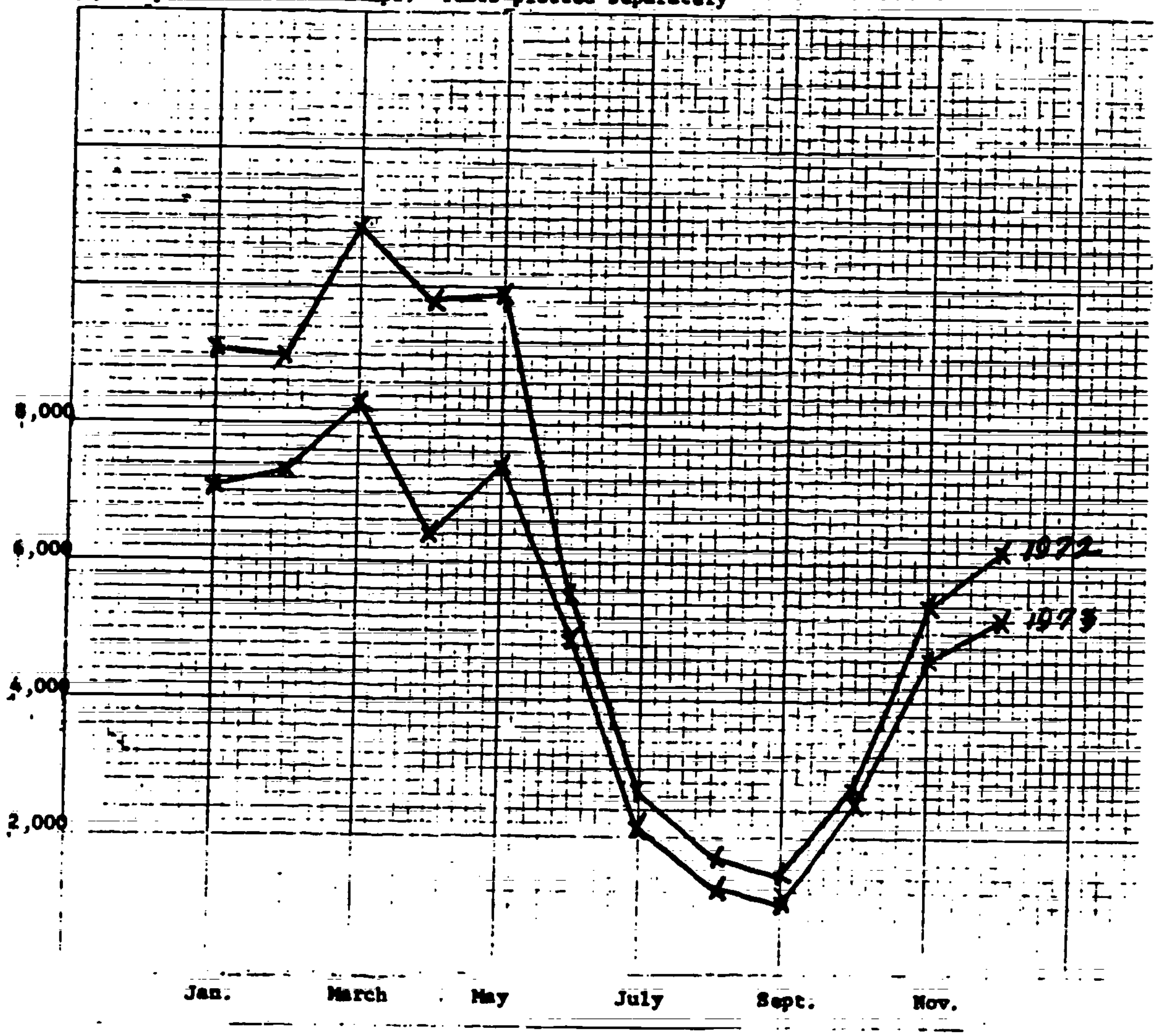
(d) March 74 ~ 7000 Feb. 75 ~ 5000(?)
Aug. 74 ~ 1100 Sept. 75 ~ 1000(?)
Dec. 74 ~ 5000

(e) One should run commercials a month or two prior to the expected onset of mumps (to allow time for parents to get their children vaccinated and to allow time for the children to develop the immunity from the disease.

The data suggest midsummer (July) for these spots to begin, with an exhortation to get children vaccinated (before school starts in September).

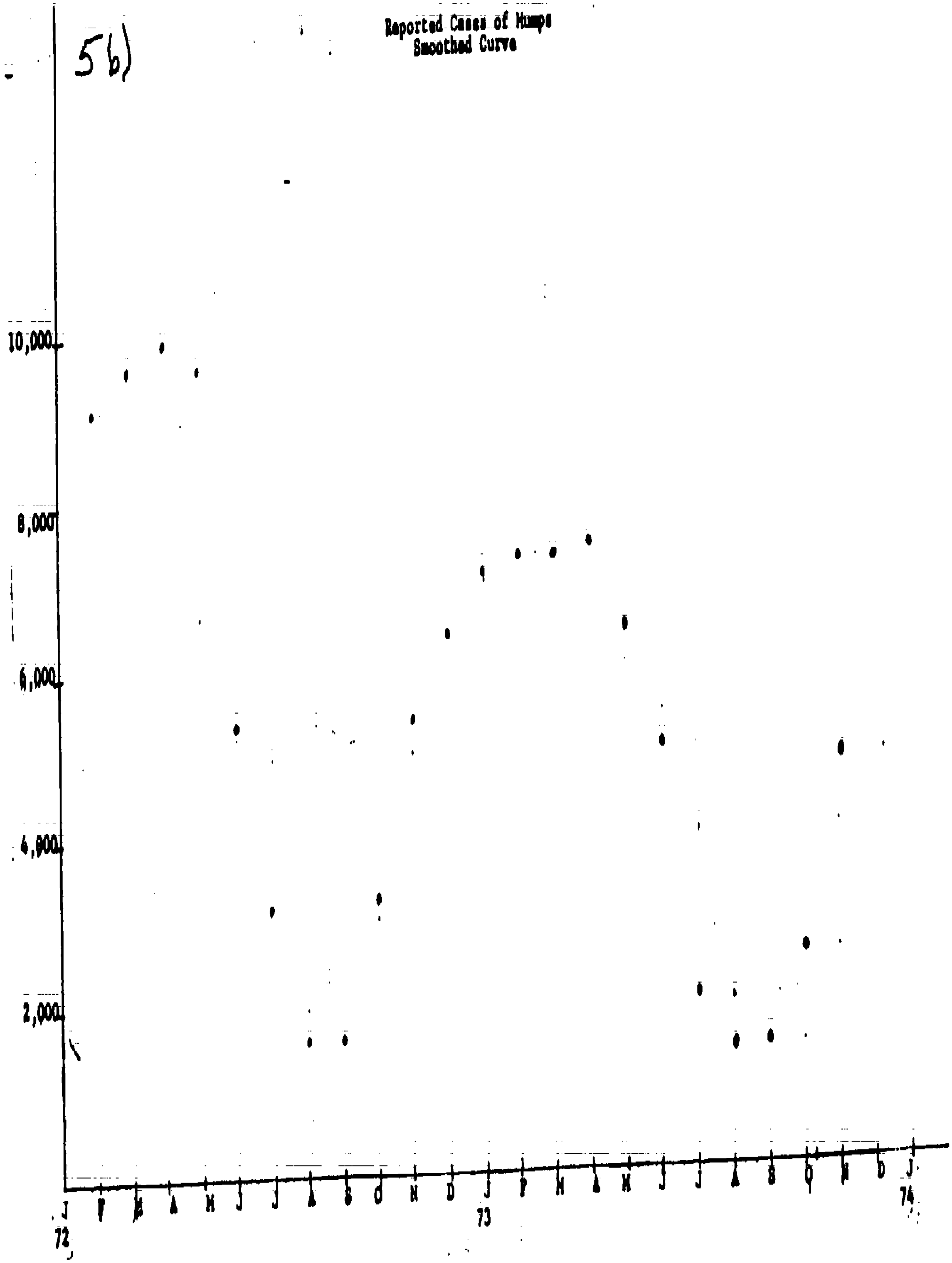
510

5a) Reported cases of mumps. Years plotted separately



511

Reported Cases of Mumps
Smoothed Curve



XVI.II.168

QMPM

6 a) The scatterplot suggests a trend of increasing median income with increasing education. If we code the points for men and women, we can readily see that men's incomes are greater than women's incomes at each level.

b) Root (Median Income $_{(Men)}$) = 52.20 + 3.94 (Years of School Completed)

The line confirms the increasing trend in men's median income with years of school completed. It also accents the apparent differences in men's and women's salaries for comparable levels of education.

Residuals: -1.16
-3.90
1.25
4.60
.67
2.82
- .44

There is no noticeable pattern in the plot of the residuals. The least squares line seems to fit quite well.

c) Root (Median Income $_{(women)}$) = 23.55 + 3.37 (Year of School Completed)

The line confirms the increasing trend in women's median income with years of school completed. The slope, being less than the slope of the line for men, indicated that the average additional income from one more year of education is less than that for males.

Residuals: 6.25
-3.39
-4.00
- .98
-3.17
1.36
3.58

There is no noticeable pattern in the plot of the residuals. The least squares lines seems to fit well but the residuals are larger than in the case of men's incomes.

d) Root (Median Income $_{(all)}$) = 37.87 + 3.65 (Years of School Completed)

Residuals:	Men	Women
	14.82	- 9.2
	12.73	-19.97
	18.45	-21.15
	22.38	-18.69

19.02	-21.45
21.75	-17.48
15.68	-16.39

The residuals are strongly positive for those points corresponding to men's incomes and strongly negative for those points corresponding to women's incomes. The two lines clearly provide a superior summary of the data. However, if we are asked to predict average income, regardless of sex, we would want to use the regression line for both sexes.

If we are given the sex of an individual we can predict median income with much greater accuracy than if we are forced to use the equation derived from the combined data (barring blind luck). Compare the sums of the absolute values of the residuals from the two regression lines as opposed to the one: 41.53 versus 238.16.

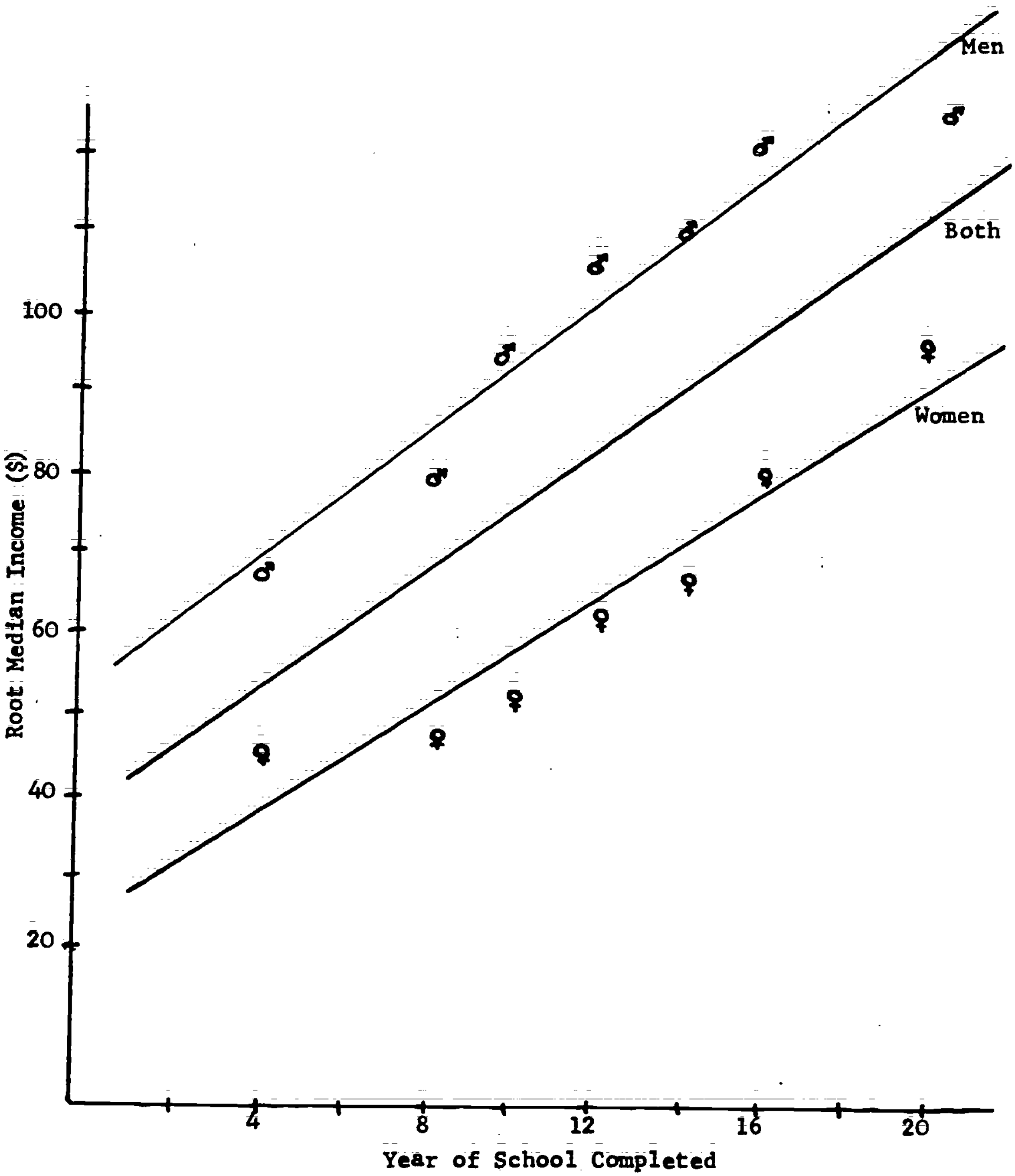
e) There are several possible reasons for the disparity in median incomes between men and women.
   - There may be job discrimination against women (e.g. lower level of job assignments despite equal training, lower pay for the same job).
   - Women may tend to concentrate their studies and take jobs in fields which typically pay less than those fields which interest men.
   - More women than men of comparable education may decide to not market their skills (e.g. become home-makers)

These three explanations, as well as others, all probably contribute to the disparity. Further study is needed to confirm or deny each. The job discrimination possibly is of particular concern since this is an illegal practice and could be dealt with in the courts in specific cases. The policy implications of the latter two explanations are less clear since they may or may not involve a sex-related choice. It may be that if women tend to stay out of particular fields of work or study some type of effort, such as a publicity campaign directed towards women should be undertaken to bring their talents into the field (e.g. the military has recently been trying to attract female volunteers).

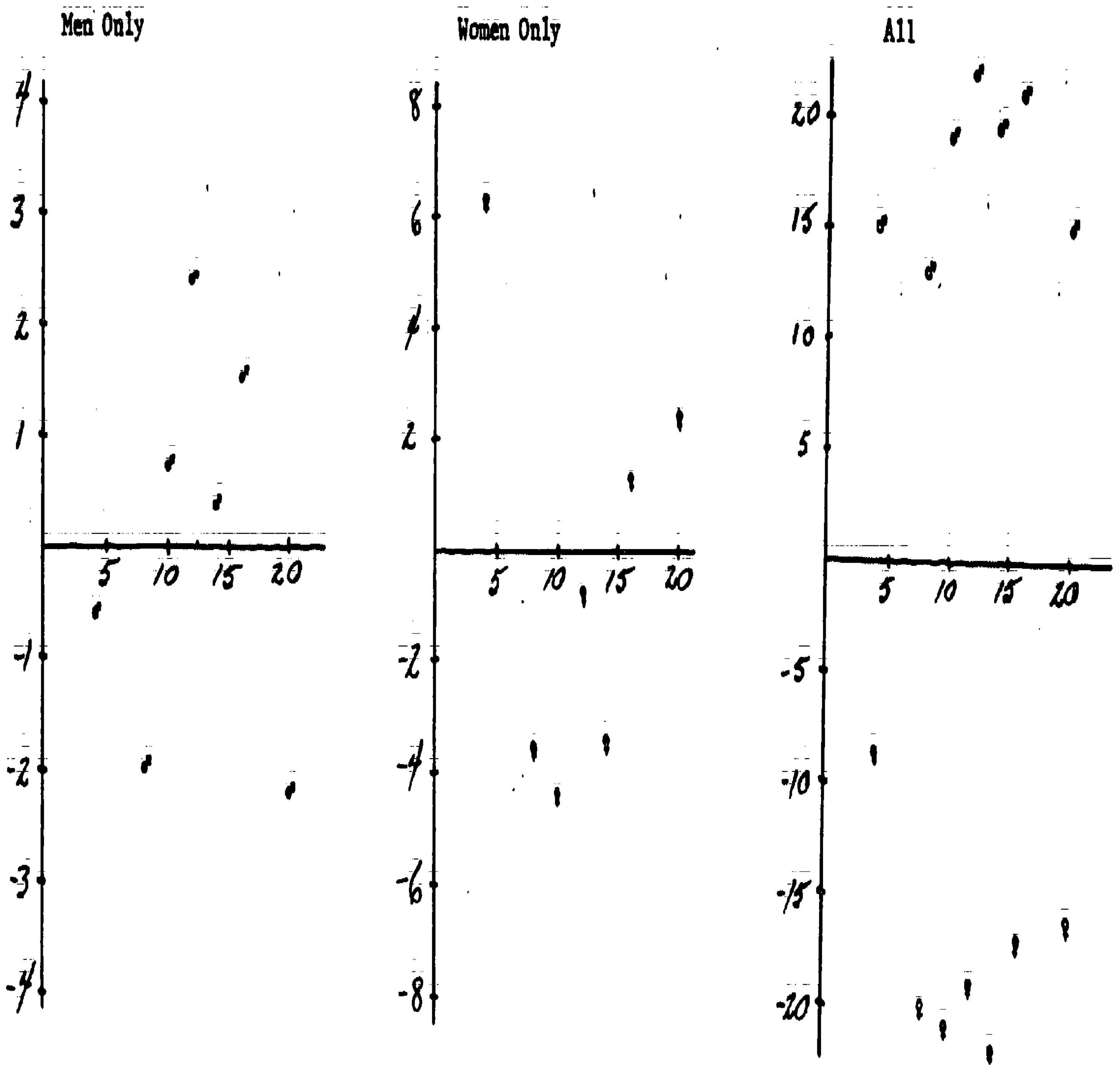
515

#6 a, b, c, d.

Root (Total Annual Income) vs Years of School Completed



6.b,c,d PLOTS OF RESIDUALS VERSUS AGE



XVI.III.172

7. (a) Gonorrhea has been increasing over time, while syphilis has been declining.

(b) For 1970: $\frac{92 + 96}{2} = 94$ thousand (Syphilis)

$$\frac{535 + 670}{2} = 602.5 \text{ thousand (Gonorrhea)}$$

(c) The linear relationship implies that a good prediction for the number of cases in a given year can be based on the number of cases during the previous year.

(d) Syphilis: 78-80 thousand (decrease of 2-3 per year)

Hence:	1972-1971	97
	1973-1972	76
	1974-1973	56
	1975-1974	3/4 (56) = 45
	1976-1975	3/4 (45) = 36
		45 + 36 = 81

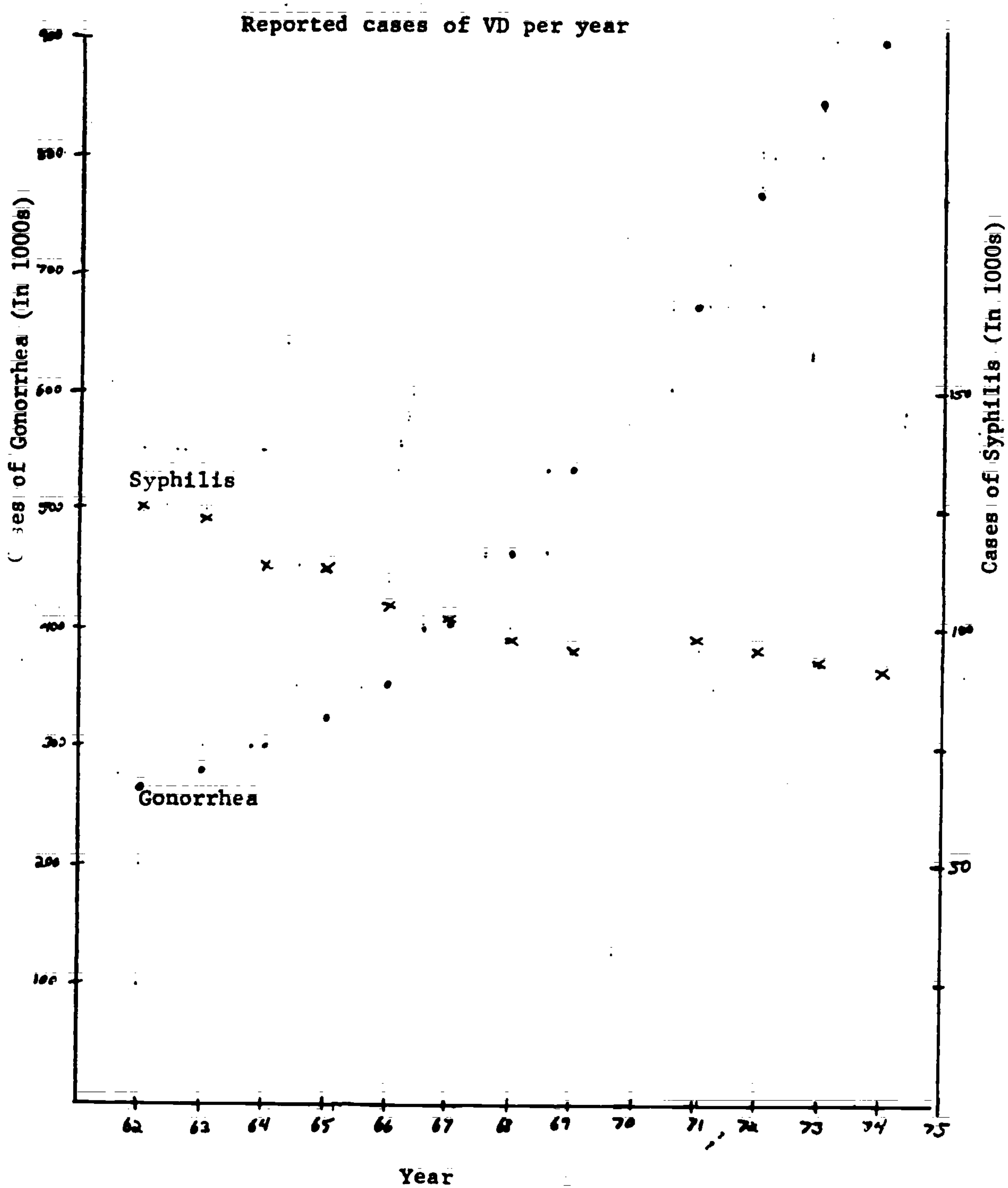
Gonorrhea: 980 thousand (# increases by about 3/4 of previous years increase)

(899 + 81) = 980

(e) Clearly, Gonorrhea is far more prevalent. Further, while the incidence of Gonorrhea is increasing ("out of control"), that of Syphilis appears to be on the decline ("under control"). Note, however, that the data <u>appear</u> to indicate that the <u>rate of increase</u> of Gonorrhea is leveling off.

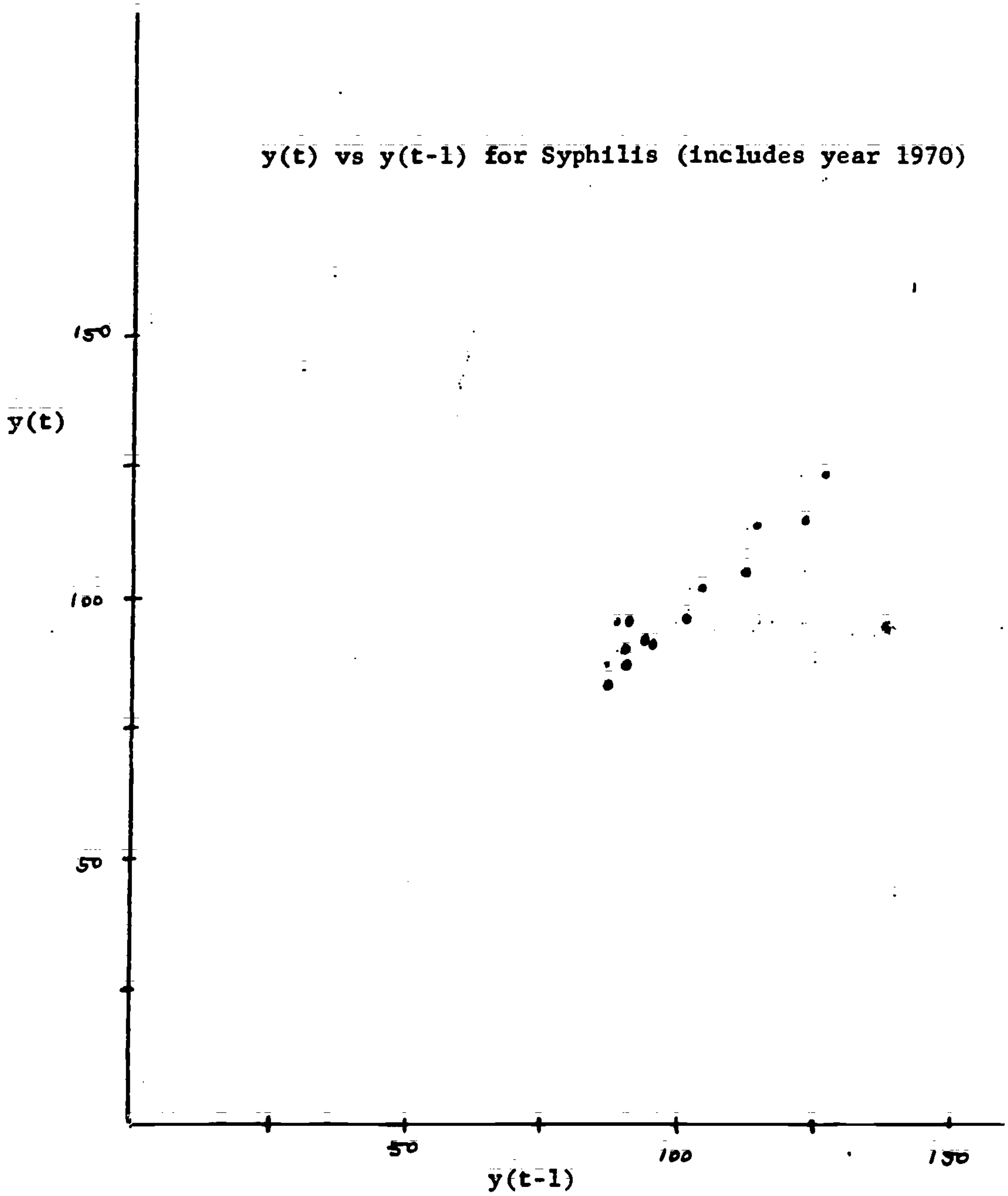
Since the spread of VD is crucially dependent on the number of individuals infected, clearly Gonorrhea will require a greater amount of effort to bring under control.

7a)

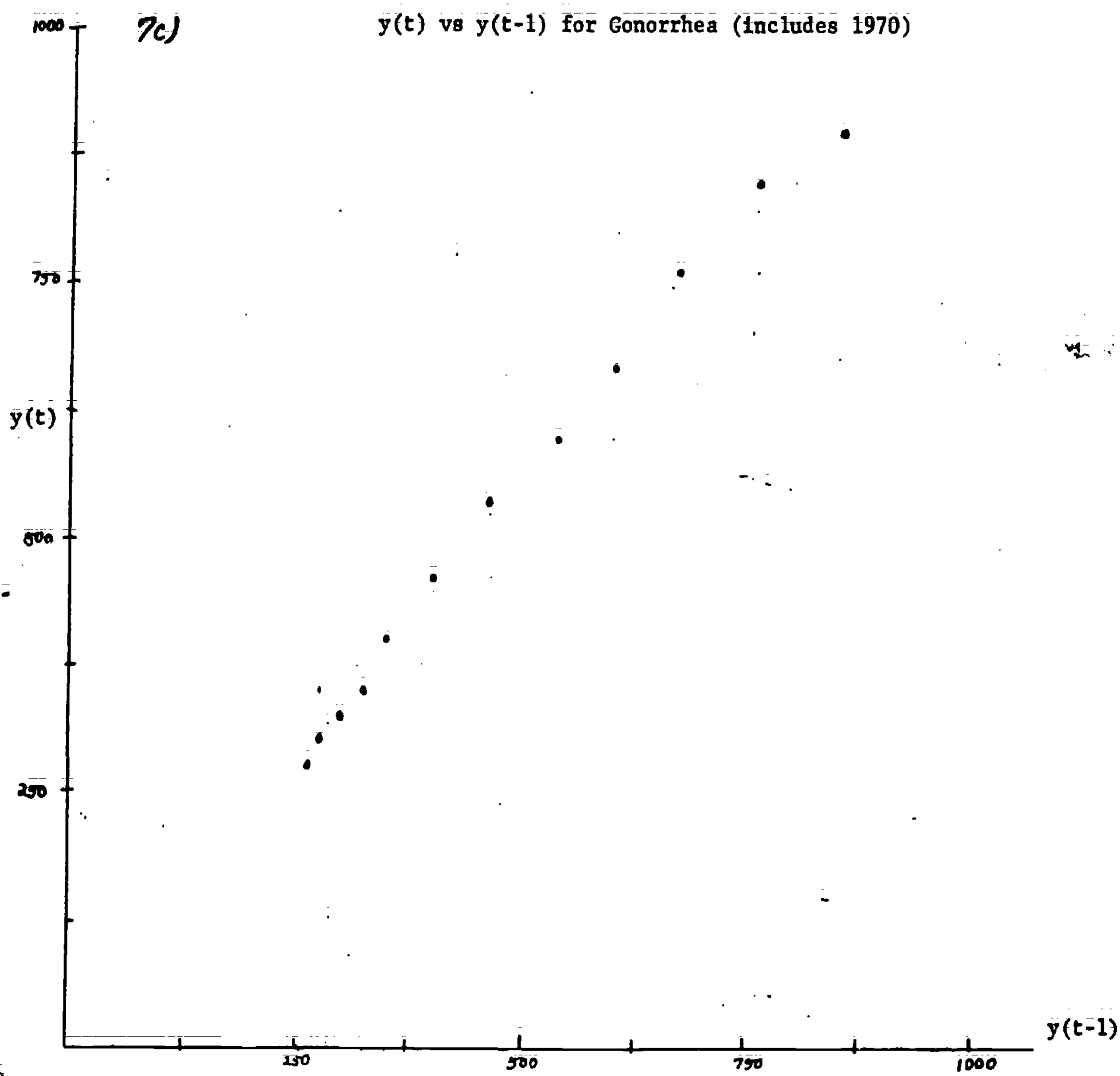

Reported cases of VD per year
Syphilis
Gonorrhea
Cases of Gonorrhea (In 1000s)
Cases of Syphilis (In 1000s)
Year
62
63
64
65
66
67
68
69
70
71
72
73
74
75

529

7 c)


y(t) vs y(t-1) for Syphilis (includes year 1970)
y(t)
150
100
50
y(t-1)
50
100
150

7c) y(t) vs y(t-1) for Gonorrhea (includes 1970)



Quiz Unit 3

Time = 60 minutes
Suggested problem times given. Credit is roughly proportional to these times.

Part I. Answer 5 of the 6 following questions (15 minutes)

(1) Give two possible reasons why the Y observations in an (X,Y) data set should be transformed.

(2) How does an ordered multiple batch differ from an unordered multiple batch?

(3) What representative point is used as a "conditional typical value" in a mini-batch?

(4) How do "outlying" data values affect a least squares line and a resistant line?

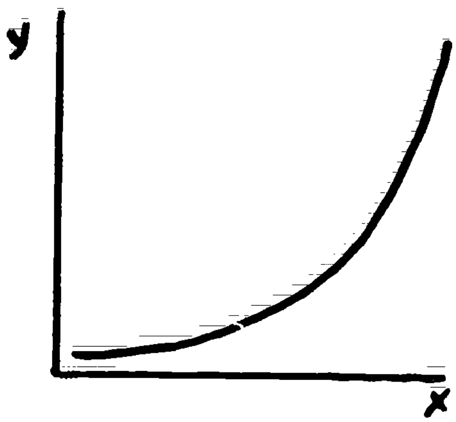
(5) What are "residuals" from a line fitted to an (X,Y) data set?

(6) According to the study cited by Tufte, what factors affect voting rates in American cities?

Part II. Answer 2 of the following 3 questions. (20 minutes)

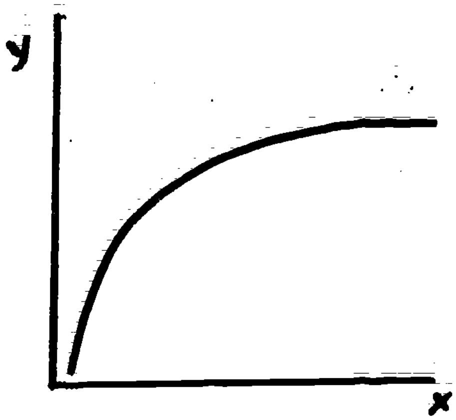
(1) Which direction do we move on the ladder of powers and for what variables, if the scatterplot of the raw data resembles:

a)

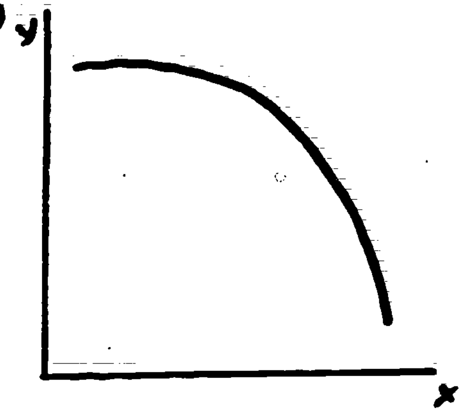


b)

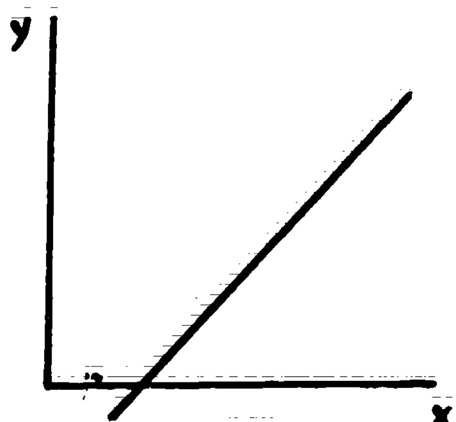
c)



d)

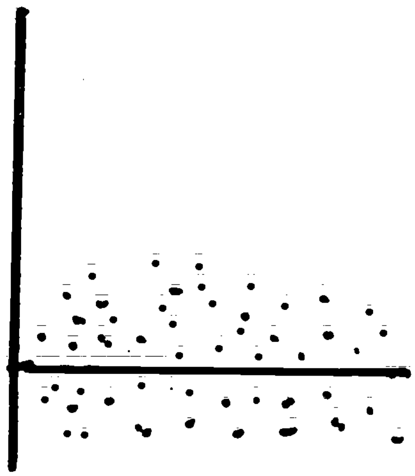


e)

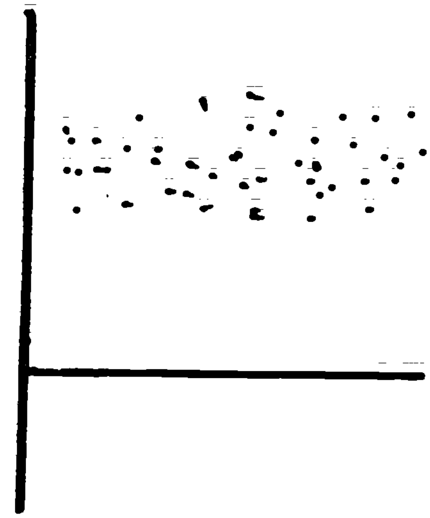


(2) What do these residuals from a polished fit imply about how well the linear model fits the data (x-variable = X, y - variable = residuals)?

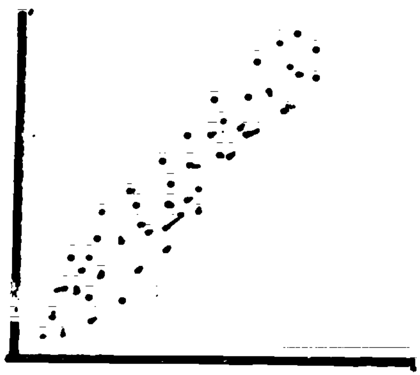
a)



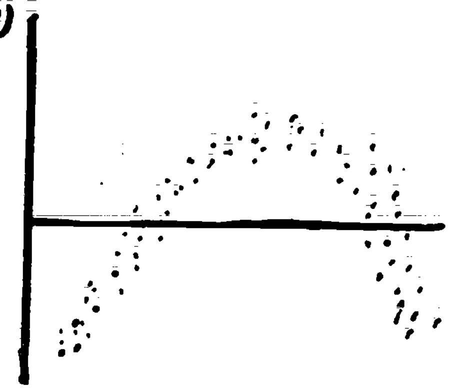
b)



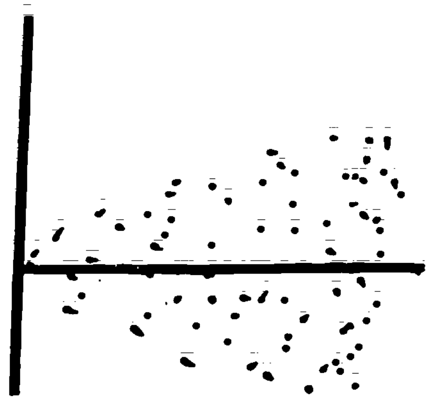
c)



d)

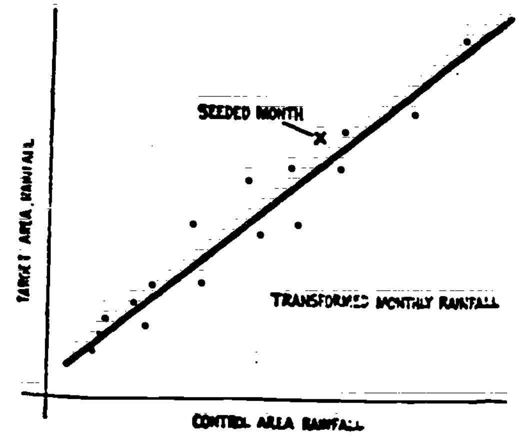


e



QMPM

(3)



This graph is from Louis Battan's article, "Cloud Seeding and Rainmaking." (Statistics: A Guide to the Unknown, pp. 354-361).

(a) What problem does Battan discuss? What solution does he propose?

(b) Interpret the graph. (A complete answer will include a description of the plotted points, a discussion of goodness of fit, and the implications for Battan's hypothesis.)

527

Part III. Answer 1 of the following 2 questions (25 minutes)

(1) Consider the following data set on U.S. Coal Production (from J. Tukey, Exploratory Data Analysis, Chapter 7).

year	coal production
1955	467 million tons
6	500
7	493
8	410
9	412
1960	416
1	403
2	422
3	459
4	467
1965	512
7	552
8	545

Note that the production for 1966 is not available.

(a) Smooth these data until the ith smooth is identical to the (i-1)th smooth. Merely use the actual data values for end points.

(b) Interpolate to find coal production in 1966.

(c) Extrapolate to find coal production in 1969.

(d) Are there any apparent trends in these data?

528

(2) Consider the following data set on 1976 and pre-1976 malpractice insurance premiums for Greater Boston hospitals and health centers (from D. Hoaglin, A First Course in Data Analysis, chapter 5).

Institution	old premium (thousands of $)	new premium (thousands of $)
Boston Hospital for Women	418	304
Peter Bent Brigham	1220	839
Robert Breck Brigham	169	120
Children's	866	646
Beth Israel	833	635
Massachusetts General	2263	1682
McLean	255	218
Harvard Health Service	136	114
Mt. Auburn	162	148
Sidney Farber Cancer Center	66	60
Massachusetts Eye and Ear	350	258
Harvard Community Health Plan	192	136
New England Deaconess	435	348

Let x = old premium, and y = new premium. We have calculated

$\Sigma x_i = 7365$ $\quad\bar{x} = 566.5$

$\Sigma x_i^2 = 8719409$ $\quad\bar{y} = 423.7$

$\Sigma y_i = 5508$

$\Sigma y_i^2 = 4752590$

$\Sigma x_i y_i = 6432511$

(a) Find least squares estimates of the coefficients for the linear model relating y to x.

(b) Would a resistant line be very different from the least squares line for these data? Why or why not?

529

Quiz Unit 3
Solutions

I. 1. Y observations in an (X,Y) data set may be transformed to equalize variance or to promote linearity.

2. The batches in an ordered multiple batch are related in a quantitative way, while those in an unordered batch are related only in a qualitative way.

3. We use the median of a mini-batch as the "conditional typical value".

4. An outlying value pulls a least squares line towards it but has very little effect on a resistant line.

5. A residual from a fitted line is the observed Y value minus the fitted Y value.

6. The factors which affect voting rates are:

   (a) closeness of election
   (b) ease of voting
   (c) socioeconomic factors

II. 1. (a) Up on X, down on Y
   (b) Down on X and Y
   (c) Down on X, up on Y
   (d) Up on X and Y
   (e) No transformation necessary

2. (a) Good fit
   (b) There is still level to be removed
   (c) There is still tilt to be removed
   (d) The data needed to be transformed
   (e) Slope and level have been removed, but spread of the residuals increases as X increases. A transformation of y would be appropriate

3. (a) Battan discusses potential drought and proposes cloud seeding as a way to increase rainfall.
   (b) The plot is of the square root of rainfall for target area (Y) and control area (X). Each point represents a month; the "x" point represents the month during which cloud seeding took place in the target area. The fitted line describes the data well, implying that the target and control areas are well-matched. Since the seeded month is not an outlier, Battan's experiment provides <u>no</u> proof that cloud seeding does increase rainfall.

III. 1. (a)

Data	First Smooth	Final Smooth
467	467	467
500	493	493
493	493	493
410	412	412
412	412	412
416	412	412
403	416	416
422	422	422
459	459	459
467	467	467
512	489.5	489.5
NA	532	532
552	548.5	545
545	545	545

(b) Interpolated value is 532 million tons.

(c) There are several extrapolated values for 1969 which make sense. You may argue that coal production is leveling off at 545 million tons. You may argue that coal production is gradually increasing and will be approximately 551 million by 1969.

(d) Coal production is increasing over time. There may be periodicities, but we don't have enough data to tell.

2. (a)
$$\hat{\beta} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} = \frac{\Sigma x_i y_i - \bar{x}\Sigma y_i - \bar{y}\Sigma x_i + n\bar{x}\bar{y}}{\Sigma x_i^2 - 2\bar{x}\Sigma x_i + n\bar{x}^2}$$

$$= \frac{\Sigma x_i y_i - \bar{x}\Sigma y_i - \bar{y}n\bar{x} + n\bar{x}\bar{y}}{\Sigma x_i^2 - 2\bar{x}\Sigma x_i + \bar{x}\Sigma x_i}$$

$$= \frac{\Sigma x_i y_i - \bar{x}\Sigma y_i}{\Sigma x_i^2 - \bar{x}\Sigma x_i}$$

$$= \frac{6,432,511 - 566.5(5508)}{8,719,409 - 566.5(7365)}$$

$$= .7284$$

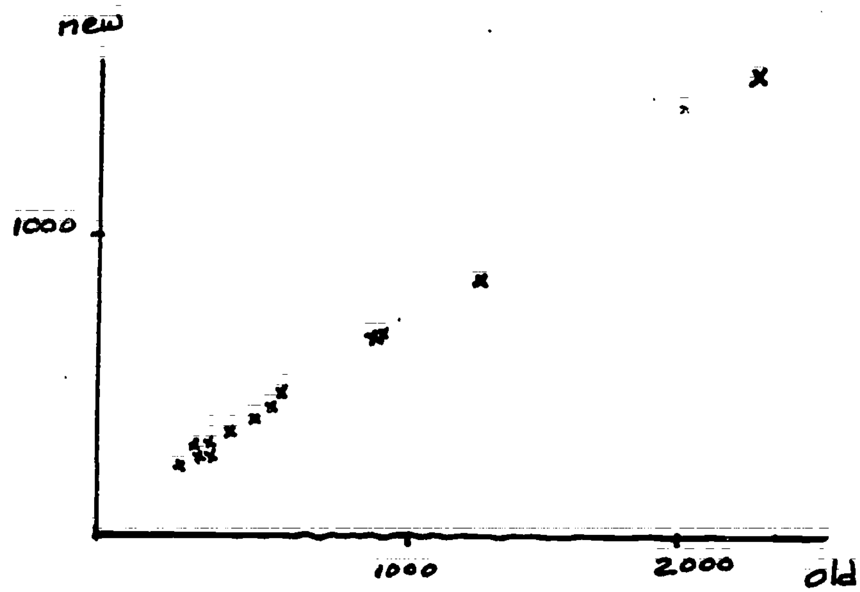
531

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

$= 423.7 - .7284(566.5)$

$= 11.06$

(b) No, a resistant line would not be very different. A plot of the data shows that there is a linear relationship between old and new premiums, with no outliers.



532

Unit 4
Reading Assignments

Lecture	Reading
4-0	Prerequisite Inventory
4-1	Tufte, pp. 135-148
4-2	Wonnacott & Wonnacott, pp. 1-24, 53-67
4-3	Tufte, pp. 156-163
4-4	No reading
Workshop	Handout: "What to Look for in Reading Technical Reports"
4-5	Handout: "Covariances and Independence in the Bivariate Multiple Regression Model"
4-6	Tufte, pp. 148-155
4-7	No reading

In addition, read the following articles:

Kaplan, Robert, and Samuel Leinhardt, "Determinants of Physician Office Location," Medical Care, Vol. II, No. 5, Sept.-Oct. 1973, pp. 406-415.

Kaplan, R., and S. Leinhardt, "The Spatial Distribution of Urban Pharmacies," Medical Care, Vol. XIII, No. 1, Jan. 1975, pp. 37-46.

Lave, Judith R., and Samuel Leinhardt, "The Cost and Length of a Hospital Stay," Inquiry, Vol. XIII, Dec. 1976, pp. 327-343.

Lave, J.R., and S. Leinhardt, "An Evaluation of a Hospital Stay Regulatory Mechanism," AJPH, Vol. 66, No. 10, 1976, pp. 959-967.

Texts:

Tufte, Edward R., Data Analysis for Politics and Policy, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1974.

Wonnacott, R.J., and T.H. Wonnacott, Econometrics, New York: John Wiley and Sons, 1970.

Prerequisite Inventory, Unit 4

Unit 4 of Module II is concerned with multiple linear regression, i.e., the fitting of linear models relating many X variables to a single Y variable. As in the previous three units, the ability to master the concepts and techniques in Unit 4 is dependent upon the mastery of several simple mathematical ideas. Before proceeding to Unit 4, you should be very familiar with the topics discussed in this inventory.

This inventory is divided into the following sections:

1. Review of Units 1 and 2, Batches of Data.
2. Review of Unit 3, Univariate Regression.
3. Representation of a data set as a matrix.
4. Matrix manipulations.

This unit depends heavily on Unit 3. If you feel that you do not have a good understanding of this prior unit, please consult a member of the course's teaching staff for additional tutoring.

Section 1. Review of Units 1 and 2, The Analysis of Batches of Data

A good review of the concepts and techniques of the first module of QMPM is given in Section 1 of Prerequisite Inventory, Unit 3. Detail concerning the construction of number summaries, schematic plots. and stem-and-leaf displays is presented there, as well as a review of important terminology. You should reread this section since this material is important for the proceedures and concepts of Unit 4.

Number summaries, schematic plots, and stem-and-leaf displays may be drawn in parallel when analyzing multiple batches of data. We merely use one scale (or one set of stems) for all the batches.

When are these concepts employed in the analysis of complex multi-variable data sets? Since such data sets are a collection of batches, related in some complicated fashion, the tools of Unit 1 and 2 are useful in "getting a feel for the data". One should analyze the multi-variable data set as single batches, or a multiple batch, before fitting the desired linear model.

These tools are also very helpful in evaluating how well the model fits the data. Since data = fit + residual, the single batch of residuals from the fit is extremely important. Residuals as a batch are occasionally assumed to be well-behaved, a powerful assumption, not often justified. We postpone the discussion of residuals and well-behaved batches to the next section.

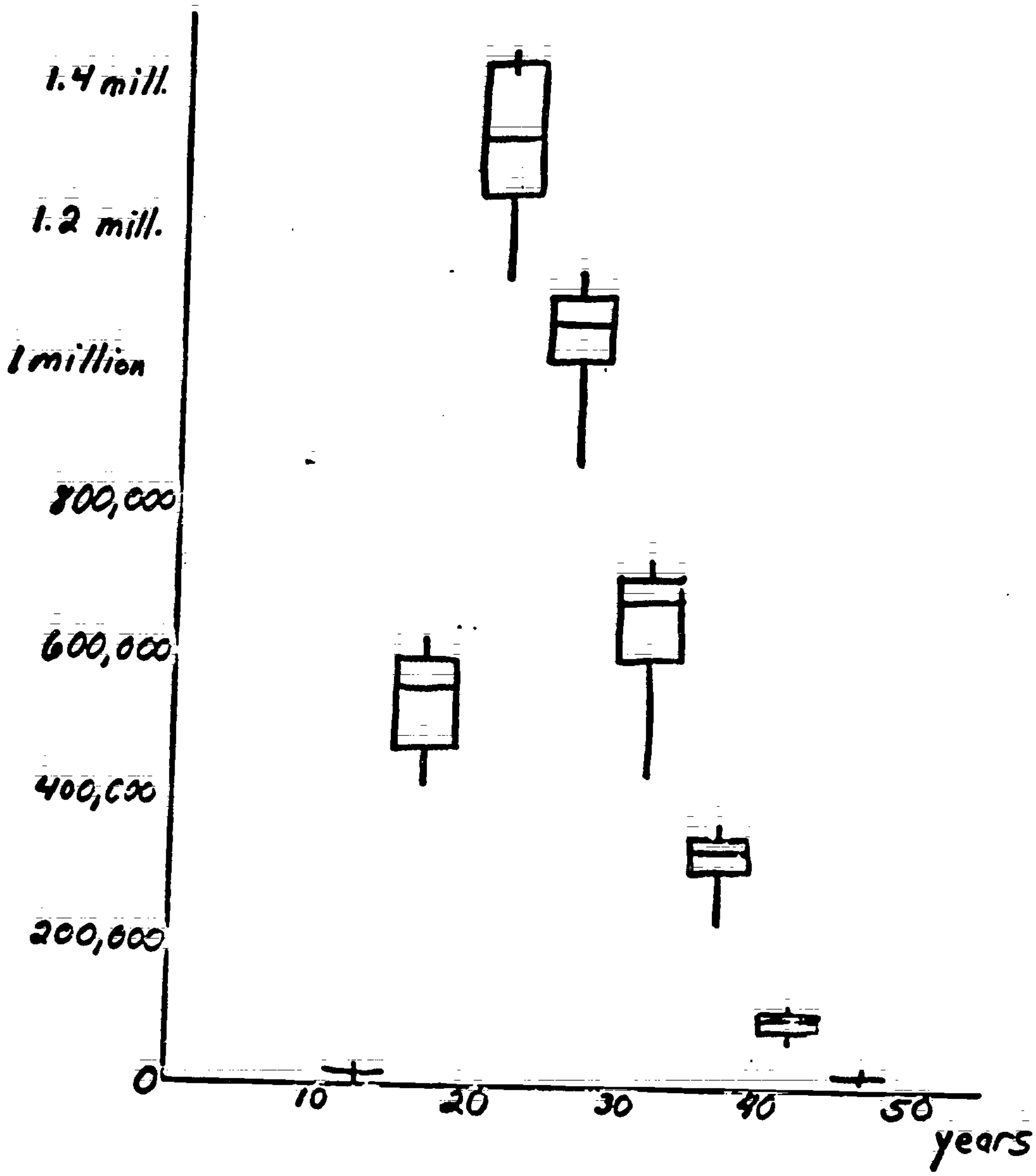
Section 2. Review of Unit 3, Univariate Regression

Unit 3 discussed the analysis of ordered multiple batches, a collection of batches with an associated scale. For example, a data set of the number of live births of women, classified by the age of the mother at time of birth is an ordered multiple batch. We have one batch for women under 15 years of age, one batch for women 15-19 years, one for 20-24 years, 25-29 years, 30-34 years, 35-39 years, 40-44 years, 45 and over. There are 8 batches, each with values for total number of live births, one datum for each year from 1950 to 1967. Associated with each batch is the midpoint of the age interval. These midpoints, ≈ 14, 17, 22, 27, 32, 37, 42, ≈ 46, constitute the age scale for the multiple batch.

We showed how parallel schematic plots are drawn for ordered multiple batches. Each plot is centered at the correct value on the scale for the batch, and the width of the plot is made equal to the

Parallel Schematic plot of Live Births by Age of Mother

Exhibit 1



536

width of the interval on the scale associated with the batch. As can be seen from the live birth schematic plot, exhibit I, this display summarizes the relationship between the data and the scale values quite well.

To summarize further these ordered batches, we compute typical values for the values, conditional on the values being located in a specific batch: conditional typical values. The conditional typical value for a data value in batch i is defined to be the median of the batch. The conditional typicals for the live birth data are given in Exhibit 2. Note how the values rise and fall as age increases, similar to the raw data. The conditional typicals are representative

Exhibit 2

Live Births by Age of Mother
Conditional Typical Values, or "Fits" for Each Age Class

"X" Age Class	Typical Value of "Y", Given "X"
Under 15	6,700 births
15-19	560,000 births
20-24	1,310,000 births
25-<9	1,065,000 births
30-34	680,000 births
35-39	330,000 births
40-44	85,000 births
Over 45	5,000 births

values for each batch, and reflect the relationship between the raw data and the values along the ordered scale. Thus, we have the decomposition of data values into conditional typicals + residuals. If the conditional typicals provide a good fit to the data, residuals will be small; otherwise, they will be large. Exhibit 3 displays the residuals from the conditional typical values for the live birth data. The many far out points indicate some lack of fit.

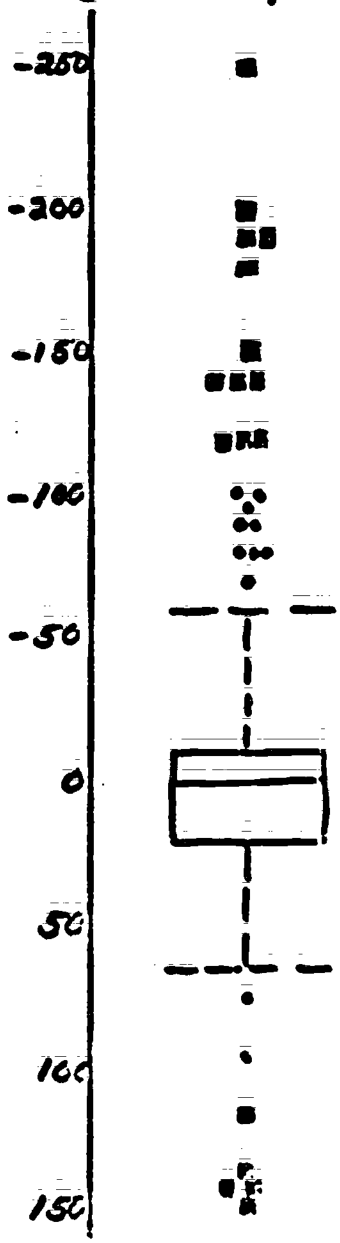
The analysis of (X,Y) paired observational data begins with a consideration of these ordered pairs as a collection of mini-batches. We use the X variable to break up the X axis into several intervals, and then group together the Y values of the ordered pairs falling in each interval into a single mini-batch. This "chopping up" of the X axis follows an examination of the scatterplot of the (X,Y) data. Exhibit 4 is such a scatterplot of percent of the population illiterate in 1930 in a state (X) and percent of the population illiterate in 1960 (Y). There are 51 points, one per state and the District of Columbia.

The scatterplot is used to break the data into mini-batches, such that the intervals on the X axis are bounded by integers approximately of equal width, and contain equal numbers of Y values. It may not be possible to achieve all three of these goals, but we must rely on our professional judgment when working with real data.

Once we have achieved the reorganization of an (X,Y) data set into batches, it can be analyzed as a collection of ordered batches. The important question is: How linear is the relationship between the

Schematic plot and Number Summary of Residuals for Live Birth data.

Exhibit 3



#144

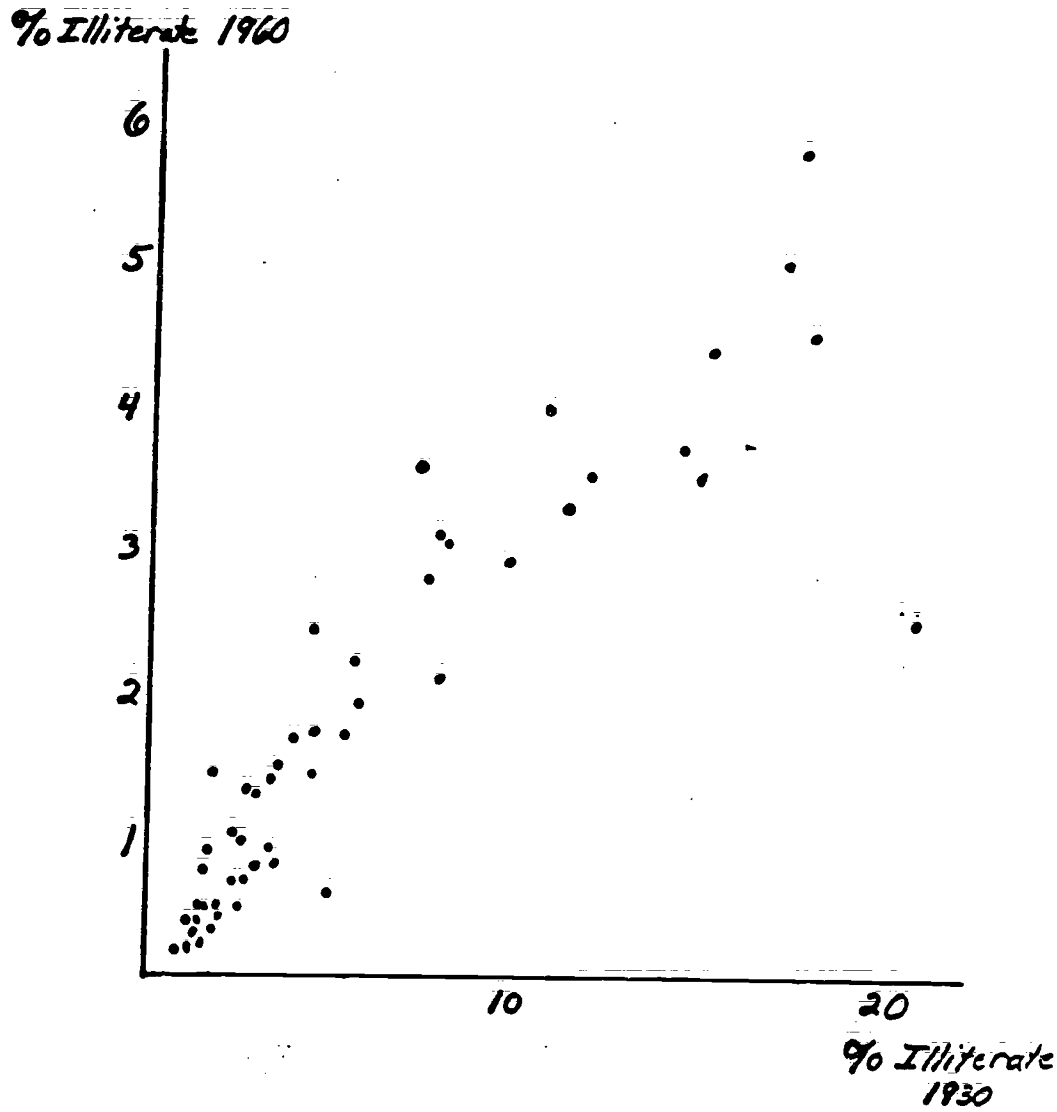
0		mid pred
-10	20	30
-150	-250	

$\boxed{115}$

-55	65	adjacent
9 values	2 values	-60
-100	110	65
12 values	5 values	

539

% Illiterate 1960 (Y) plotted against % Illiterate 1930 (X).
One point per state.
Exhibit 4



conditional typical values and the X scale? To answer this we plot the conditional typicals and the hinges of the mini-batches on a separate plot and connect them. If the relationship is close to linear <u>and</u> if the three lines are roughly parallel--as are the lines in Exhibit 5 of the illiteracy data--then we are in good shape. If the plot lacks these qualities, then we might want to transform our (X,Y) data set.

With regards to transformation we have 2 goals: (1) increase linearity, and (2) equalize spread. The plot of the conditional typicals and hinges is quite useful in assessing how far we must go to achieve these goals. If the lines connecting the medians, upper hinges, and lower hinges are not straight, then a transformation on the X variable to increase linearity is needed. If the lines connecting the summary quantities are not parallel, and diverge or converge as X increases, then the midspreads of the batches are not constant. To equalize these spreads we transform Y. How do we determine how far up or down the ladder of powers to move with X and Y? Exhibit 6 is useful in this determination. Identify the shape of the scatterplot as one of the 4 functional forms in this display, and transform accordingly. Finding the <u>best</u> transformation is an iterative process. Try several.

Once we have successfully transformed the (X,Y) data set, we are now ready to summarize formally the relationship of Y to X. We fit a line, either <u>resistantly</u> or by <u>least squares</u>, to the (X,Y) data set. We hypothesize

511

$$Y_i = a + bX_i$$

Exhibit 5

Illiteracy Data, Conditional Typicals for Mini-Batches Connected (solid Line) and Hinges Connected (Dashed Line).

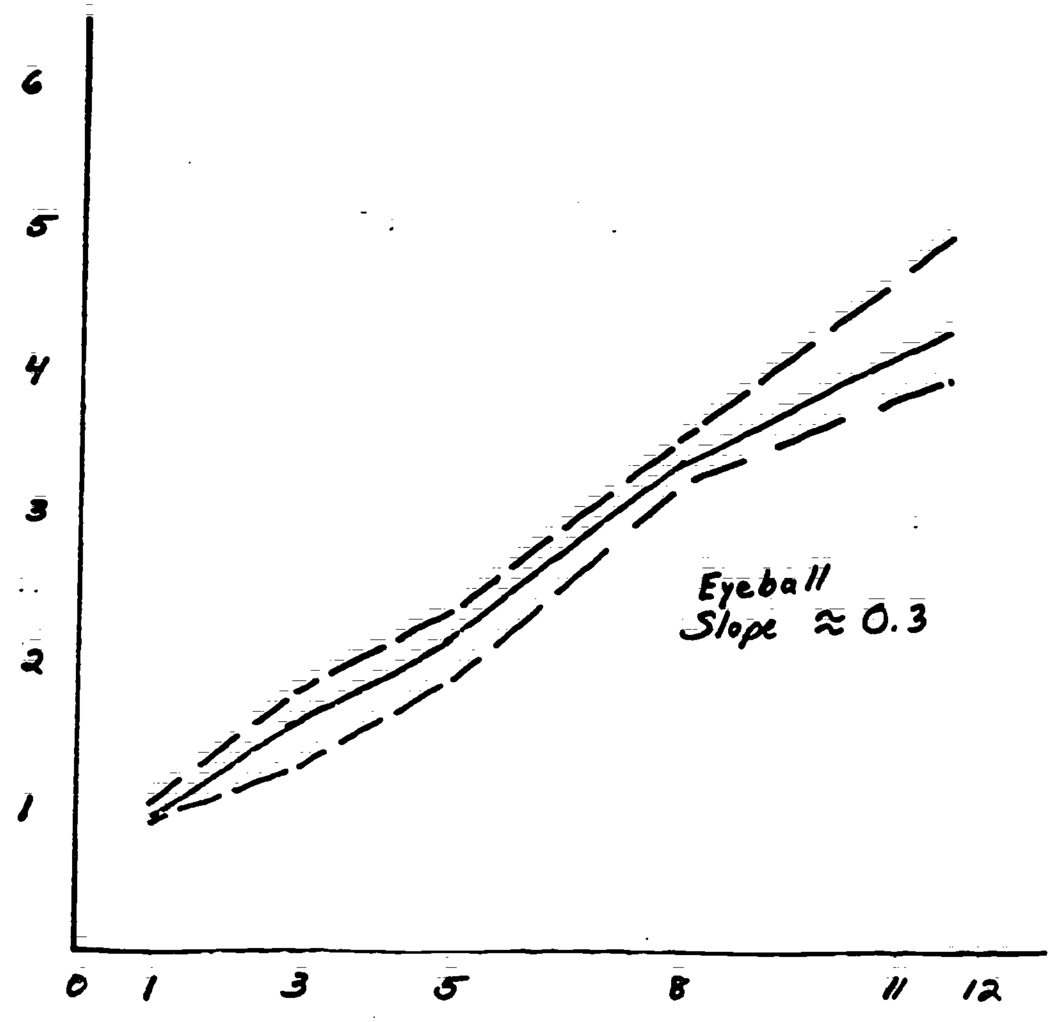
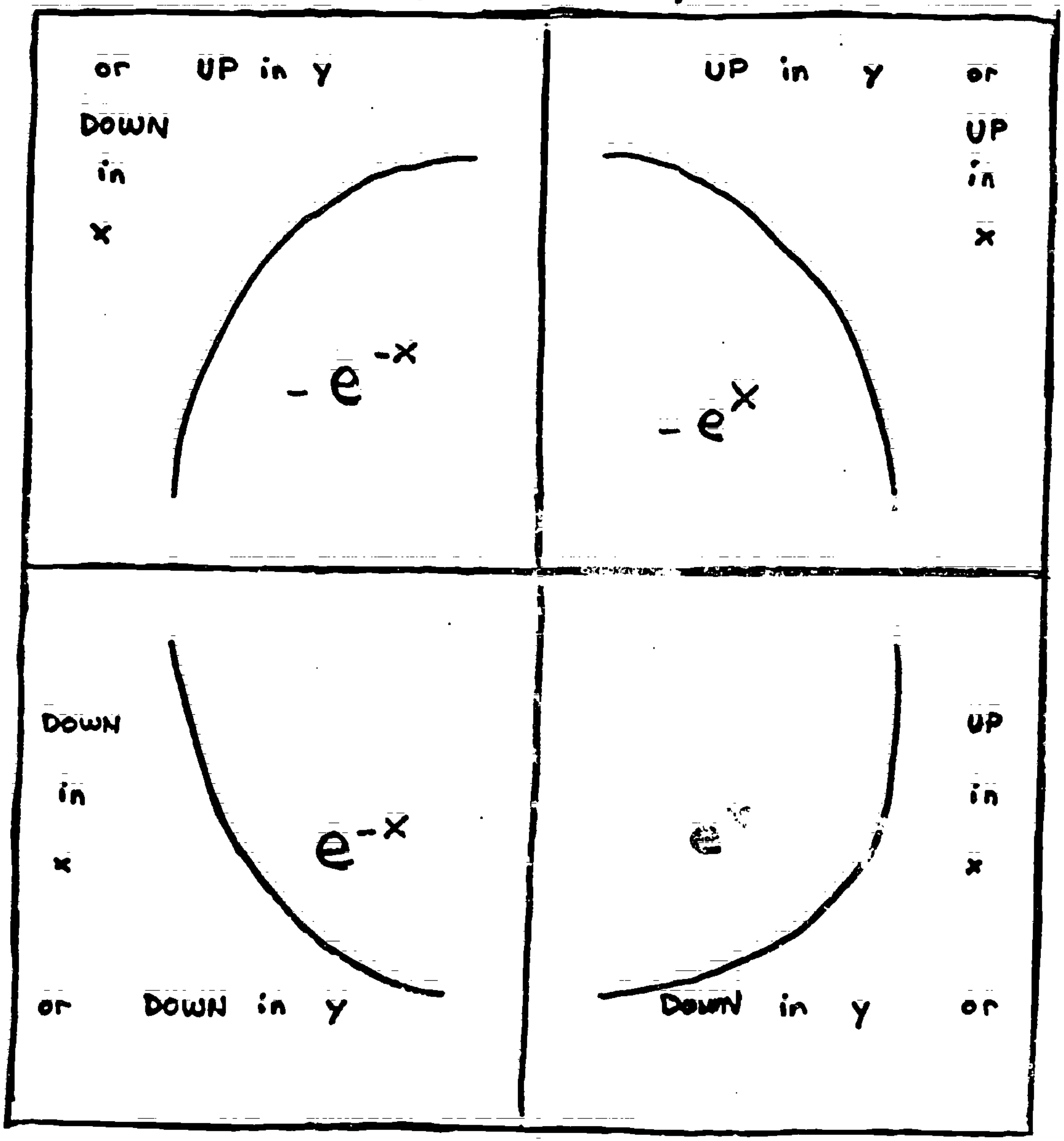


Exhibit 6

How to Move in Re-expressing x or y Alone
(the four different shapes)



UP: squares, cubes, etc.
DOWN: logarithms, reciprocals, etc.

as the function relating Y to X. We call the model fitting process "regression", and state that "Y is regressed on X". Since we have only one variable X to be regressed upon, the regression is a "univariable" or univariate regression. Y is called the dependent variable, X the independent variable.

Resistant lines are calculated by breaking the data into thirds--or 3 equal sized mini-batches--and computing the median of the X's and the median of the Y's--the conditional typical value--within each third. Label these three summary points

$(X_{(1)}, Y_{(1)})$ = median of first third

$(X_{(2)}, Y_{(2)})$ = median of second third

$(X_{(3)}, Y_{(3)})$ = median of third third.

We compute

$$b = \frac{(Y_{(3)} - Y_{(1)})}{(X_{(3)} - X_{(1)})}$$

and

$$a = \frac{1}{3}\left[(Y_{(1)} - bX_{(1)}) + (Y_{(2)} - bX_{(2)}) + (Y_{(3)} - bX_{(3)})\right].$$

The resistant line may need several steps of polish to remove all the tilt and level from the residuals. To polish, we fit a line to the residuals from the previous fit, and add the a and b calculated from the polishing to the a and b from the previous fit.

Least squares minimizes the sum of the squared residuals. We seek the a and b that minimize

$$\sum_{i=1}^{n} (Y_i - a - bX_i)^2 .$$

Least squares provides a fit very similar to a resistant line if the data are linear, the spread about the line is constant, and there are no outliers. If any of these conditions are violated, then the least squares line will not fit the data well, and the resistant line is preferable. Resistant lines are "resistant" to violations of these assumptions.

We compute

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

and

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

as least squares coefficients estimates. To evaluate how well the least squares line fits the data, we calculate

$$S^2_{y|x} = \frac{\sum (Y_i - a - bX_i)^2}{n - 2}$$

and

$$r^2 = 1 - \frac{\sum (Y_i - a - bX_i)^2}{\sum (Y_i - \bar{Y})^2}$$

$S^2_{y|x}$ is the variance about the line and should be as small as possible. r^2 is 1 minus the ratio of residual variation to total variation, and is interpreted as the "percent of the total variation" explained by the line. The closer this quantity is to 1 the more completely the line "explains" the data.

Residuals are defined as

$$r_i = Y_i - a - bX_i$$

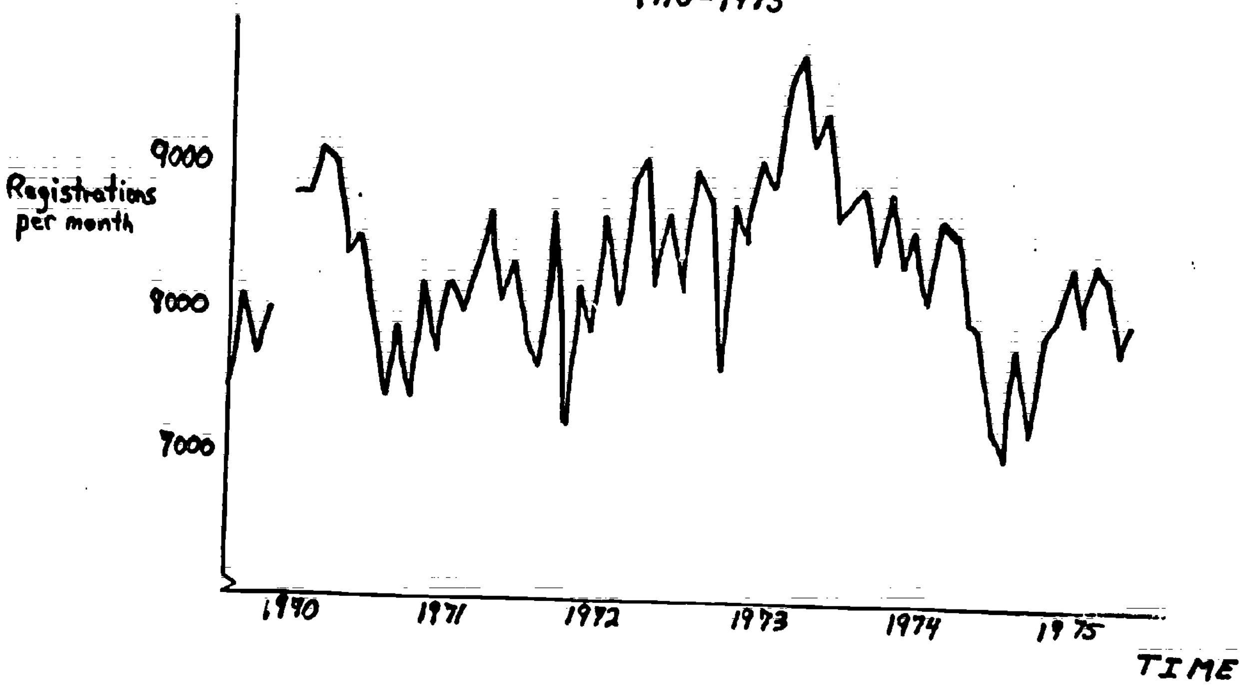
and are very important in evaluating the least squares and resistant fit. Residuals, treated as a single batch, should be well-behaved. A well-behaved batch is symmetric about the mean of the batch; approximately 64% of the batch values are within one standard deviation of the mean, and approximately 95% of the batch values are within 2 standard deviations of the mean. Such a batch has no outliers. This well-behaved assumption is crucial to least squares lines and will be discussed further in Unit 4.

A plot of the residuals versus X is also important. Such a plot should be a random swarm of points, devoid of any pattern. Any pattern, such as trigonometric, wedge, linear, or curvilinear, is an indication that the line does not fit.

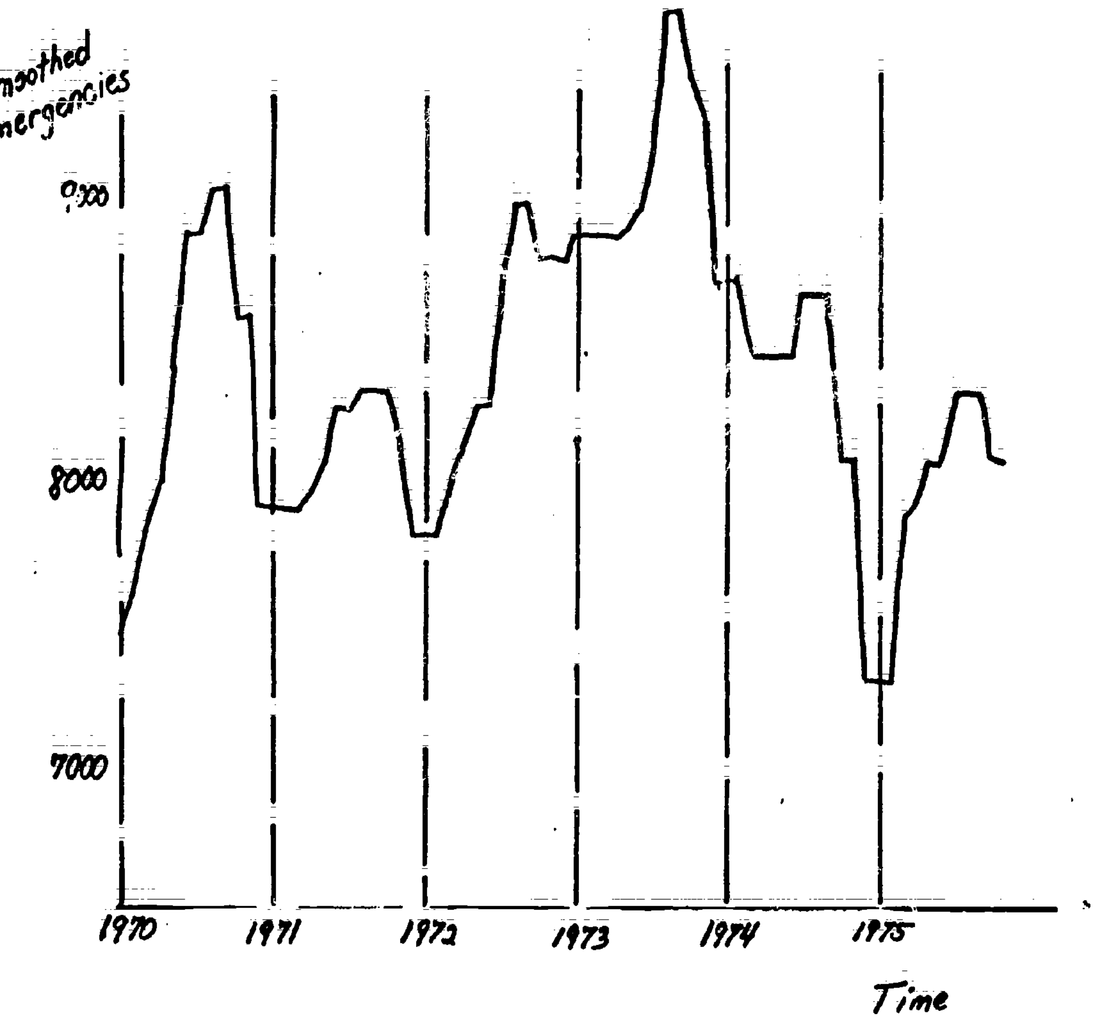
Time series data are a special kind of (X,Y) data. The X variable refers to time (months, weeks, days, etc.) and there is one Y_i associated with each X_i. Time series data contain quite a bit of noise, and it is usually necessary to smooth these data sets to filter out the irregularities. Running medians of 3 is one smoother, and involves taking the median of 3 consecutive data values, beginning at the first time point and working down to the last. The data are smoothed several times until the smoothed values from the ith iteration are identical to those from the (i-1)th. Exhibit 7 is a time plot of emergency registrations at D. C. General Hospital, and exhibit 8 is the smoothed time plot. Note how many of the peaks and troughs have been removed by the smoothing.

Exhibit 7

D.C. General Hospital
Emergency Registrations
1970-1975



D.C. General Hospital, Smoothed Emergencies
Exhibit 8



Module II

If the time plot shows sufficient trends, then we may extrapolate, estimate beyond the range of the data, and interpolate, estimate between two consecutive time points. The identification of periodicities such as seasonal highs and lows is also important.

Section 3. Representation of a data set as a matrix.

Consider an (X,Y) data set. This data set contains 2 related batches X and Y of equal size N. The observations in X are denoted x_i, and those in Y, y_i. In fact,

$$X = (x_1, x_2, \ldots, x_{N-1}, x_N)$$
$$Y = (y_1, y_2, \ldots, y_{N-1}, y_N)$$

i.e., the data vector X, of length N, can be represented as an N-tuple of values x_1 through x_N. Similarly for Y. We have written X and Y horizontally; henceforth, we shall represent these vectors as vertical columns;

$$\underset{\sim}{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} \qquad \underset{\sim}{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}$$

We call X and Y column vectors and represent them with little letters underscored with "tildas": $\underset{\sim}{x}$ and $\underset{\sim}{y}$. All vectors will be written as little letters with tildas: $\underset{\sim}{a}$, $\underset{\sim}{b}$, etc. The length of a vector is equal to the number of observations, N. $\underset{\sim}{y}$, in the linear model, is called the dependent variable, or vector of dependent observations.

Unit 4 is concerned with data sets containing a data vector $\underset{\sim}{y}$ and more than one $\underset{\sim}{x}$ vector. We relate the variable Y to variables $X_1, X_2, \ldots, X_p$, i.e., we seek to "describe" Y as a function of p dependent variables $X_1, X_2, \ldots, X_p$. We need a convenient mathematical representation of the variables $X_1, X_2, \ldots, X_p$.

We present an example. For N = 10 eastern states, we have data on the transportation equipment industry in 1957.

State	(million $) Aggregate Value Added	(million $) Aggregate Capital Service Flow	(million man-hours) Aggregate Man-Hours Worked
Connecticut	690	39	124
Maine	29	2	6
Maryland	415	18	69
Massachusetts	242	15	39
New Jersey	667	33	83
New York	940	73	190
Ohio	1611	158	260
Pennsylvania	618	34	98
Virginia	174	7	31
West Virginia	23	2	4

Exhibit 14: Regression Data

We seek to estimate Aggregate Value Added (Y), as a function of Aggregate Capital Service Flow (X_1) and Aggregate Man-Hours Worked (X_2). This functional relationship is known in economics as the Cobb-Douglas production function.

The dependent variable, $\underset{\sim}{y}$, is

$$\underset{\sim}{y} = \begin{pmatrix} 690 \\ 29 \\ 415 \\ 242 \\ 667 \\ 940 \\ 1611 \\ 618 \\ 174 \\ 23 \end{pmatrix}$$

The 2 independent variables ($p = 2$) are

$$\underset{\sim}{x}_1 = \begin{pmatrix} 39 \\ 2 \\ 18 \\ 15 \\ 33 \\ 73 \\ 158 \\ 34 \\ 7 \\ 2 \end{pmatrix} \qquad \underset{\sim}{x}_2 = \begin{pmatrix} 124 \\ 6 \\ 69 \\ 39 \\ 83 \\ 190 \\ 260 \\ 98 \\ 31 \\ 4 \end{pmatrix}$$

The elements in $\underset{\sim}{x}_1$ are denoted x_{i1}, $i = 1,\ldots,10$, and the elements in $\underset{\sim}{x}_2$ are denoted x_{i2} $i = 1,\ldots,10$. Hence, $x_{11} = 39$, $x_{21} = 2$, ..., $x_{10,1} = 2$, $x_{12} = 124$, ..., $x_{10,2} = 4$.

Suppose we place the vectors $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$ side by side, and label this "entity" $\underset{\sim}{X}$. We have:

$$\underset{\sim}{X} = \begin{pmatrix} 39 & 124 \\ 2 & 6 \\ 18 & 69 \\ 15 & 39 \\ 33 & 83 \\ 73 & 190 \\ 158 & 260 \\ 34 & 98 \\ 7 & 31 \\ 2 & 4 \end{pmatrix}$$

553

We call $\underset{\sim}{X}$ a <u>matrix</u> (plural: <u>matrices</u>). A matrix is merely a collection of p vectors. It is symbolized by a capital letter underscored with a tilda. A matrix is a 2 dimensional quantity, characterized by first dimension = number of rows and second dimension = number of columns. Our data matrix $\underset{\sim}{X}$ has dimensions 10 and 2, and is a (10 x 2) matrix. Note that a matrix with only one column is a vector. If N = p, $\underset{\sim}{X}$ is called <u>square</u>; otherwise it is <u>rectangular</u>.

In general, the data matrix $\underset{\sim}{X}$ of independent variables will have dimensions N and p. The elements of $\underset{\sim}{X}$ are x_{ij} where $i = 1, 2, \ldots, N$, and $j = 1, 2, \ldots, p$. In multiple regression, a column of $\underset{\sim}{X}$ is a single variable, $\underset{\sim}{x}_j$, and a row of $\underset{\sim}{X}$ is a single observation--a multivariable observation. The observations in the data matrix formed from exhibit 14 refer to the 10 eastern states. On each observation (state) we record capital service flow (X_1) and man-hours worked (X_2). Remember that value added is not part of the $\underset{\sim}{X}$ data matrix; it is the $\underset{\sim}{y}$ vector of dependent observations.

An (N x p) data matrix $\underset{\sim}{X}$ is:

$$\underset{\sim}{X} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdot & \cdot & \cdot & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdot & \cdot & \cdot & x_{2p} \\ \cdot & \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & & & & \cdot \\ x_{N1} & x_{N2} & x_{N3} & \cdot & \cdot & \cdot & x_{Np} \end{pmatrix}$$

This representation will be used throughout Unit 4.

Section 4. Matrix manipulation

In this section we define

(1) Matrix addition: $\underset{\sim}{X} + \underset{\sim}{Y}$

(2) Null matrix: $\underset{\sim}{Z}$

(3) Matrix multiplication: $\underset{\sim}{X}\,\underset{\sim}{Y}$

(4) Matrix transposition: $\underset{\sim}{X}^t$

(5) Identity matrix: $\underset{\sim}{I}$

(6) Matrix inversion: $\underset{\sim}{X}^{-1}$

Matrix addition is a simple operation. To add 2 matrices $\underset{\sim}{X}$ and $\underset{\sim}{Y}$, they must be of the same dimensions, (N x p). Let $\underset{\sim}{C} = \underset{\sim}{X} + \underset{\sim}{Y}$. If $\underset{\sim}{X}$ has elements x_{ij} and $\underset{\sim}{Y}$ has elements y_{ij}, then the (i,j) element of $\underset{\sim}{C}$, c_{ij}, equals $x_{ij} + y_{ij}$. We merely add the corresponding entries of $\underset{\sim}{X}$ and $\underset{\sim}{Y}$. An example illustrates this. If

$$\underset{\sim}{X} = \begin{pmatrix} 3 & 2 & 0 \\ 9 & 1 & 6 \end{pmatrix} \quad \text{and} \quad \underset{\sim}{Y} = \begin{pmatrix} 14 & 17 & 1 \\ 2 & 0 & 9 \end{pmatrix}$$

then

$$\underset{\sim}{C} = \begin{pmatrix} 3+14 & 2+17 & 0+1 \\ 9+2 & 1+0 & 6+9 \end{pmatrix} = \begin{pmatrix} 17 & 19 & 1 \\ 11 & 1 & 15 \end{pmatrix}$$

Subtraction is defined as follows: if $\underset{\sim}{C} = \underset{\sim}{X} - \underset{\sim}{Y}$, then $\underset{\sim}{C} = \underset{\sim}{X} + (-\underset{\sim}{Y})$; i.e. it is addition of $\underset{\sim}{X}$ to the negative of $\underset{\sim}{Y}$.

The null matrix $\underset{\sim}{Z}$ plays a special role in addition. It is an (N x p) matrix of zeros: $z_{ij} = 0$ for all i and j. If $\underset{\sim}{X}$ and $\underset{\sim}{Z}$ are (N x p) matrices, and $\underset{\sim}{Z}$ is the null matrix, then

$$\underset{\sim}{X} + \underset{\sim}{Z} = \underset{\sim}{X} - \underset{\sim}{Z} = \underset{\sim}{X}.$$

Matrix multiplication is slightly more complicated than matrix addition. It is <u>not</u> a term by term operation of multiplying corresponding entries! This is important to remember. In order to multiply matrices $\underset{\sim}{X}$ and $\underset{\sim}{Y}$ we require that the number of <u>columns</u> of $\underset{\sim}{X}$ must equal the number of <u>rows</u> of $\underset{\sim}{Y}$.

If $\underset{\sim}{X}$ has dimension (N x p) and $\underset{\sim}{Y}$ has dimension (p x q), then the product $\underset{\sim}{X}\underset{\sim}{Y}$ is an <u>(N x q)</u> matrix, $\underset{\sim}{C}$, whose entries, c_{ij}, are obtained by summing the products formed by multiplying, in order, each entry in the ith row of $\underset{\sim}{X}$ with each corresponding entry in the jth column of $\underset{\sim}{Y}$. Formally,

$$c_{ij} = \sum_{k=1}^{p} x_{ik}\, y_{kj}$$

Matrix multiplication is defined as multiplying the rows of the matrix on the left with the columns of the matrix on the right. In general, $\underset{\sim}{X}\,\underset{\sim}{Y}$ does not equal $\underset{\sim}{Y}\,\underset{\sim}{X}$: matrix multiplication is not commutative.

An example helps. Let

$$\underset{\sim}{X} = \begin{pmatrix} 2 & 1 & -6 \\ 1 & -3 & 2 \end{pmatrix} \quad \text{and} \quad \underset{\sim}{Y} = \begin{pmatrix} 1 & 0 & -3 & 0 \\ 0 & 4 & 2 & 0 \\ -2 & 1 & 1 & 1 \end{pmatrix}.$$

Since the number of columns of $\underset{\sim}{X}$, 3, equals the number of rows of $\underset{\sim}{Y}$, the operation $\underset{\sim}{C} = \underset{\sim}{X}\,\underset{\sim}{Y}$ is defined. $\underset{\sim}{C}$ will have dimension (2 x 4). The first element of $\underset{\sim}{C}$, c_{11}, is formed by the summing the products of the <u>first</u> row of $\underset{\sim}{X}$ with the <u>first</u> column of $\underset{\sim}{Y}$: $c_{11} = 2\cdot1 + 1\cdot0 + (-6)\,(-2) = 14$. c_{12} is formed with the <u>first</u> row of $\underset{\sim}{X}$ and the <u>second</u> column of $\underset{\sim}{Y}$: $c_{12} = 2\cdot0 + 1\cdot4 + (-6)\cdot1 = -2$. The matrix $\underset{\sim}{C}$ is

$$\underset{\sim}{C} = \begin{pmatrix} 2\cdot 1+1\cdot 0+6\cdot 2 & 2\cdot 0+1\cdot 4-6\cdot 1 & -2\cdot 3+1\cdot 2-6\cdot 1 & 2\cdot 0+1\cdot 0-6\cdot 1 \\ 1\cdot 1-3\cdot 0-2\cdot 2 & 1\cdot 0-3\cdot 4+2\cdot 1 & -1\cdot 3-3\cdot 2+2\cdot 1 & 1\cdot 0-3\cdot 0+2\cdot 1 \end{pmatrix}$$

$$= \begin{pmatrix} 14 & -2 & -14 & -6 \\ -3 & -10 & -7 & 2 \end{pmatrix}$$

A square matrix $\underset{\sim}{X}$ has equal numbers of rows and columns. It has dimension (N x N). A square matrix is symmetric if

$$x_{ij} = x_{ji} \quad \text{for all } i \text{ and } j.$$

For example, the matrix

$$\underset{\sim}{X} = \begin{pmatrix} 9 & 6 & -3 & 14 \\ 6 & 3 & 0 & 2 \\ -3 & 0 & 2 & 4 \\ 14 & 2 & 4 & 1 \end{pmatrix}$$

is symmetric. With a square matrix, we call the terms x_{ii}, $i = 1, \ldots, N$, the diagonal of the matrix. The diagonal of the above matrix is (9,3,2,1). Note that the diagonal is not well-defined in a rectangular matrix.

The transposition, or transpose, of a matrix, $\underset{\sim}{X}^t$, is defined as a reversal of the (i,j) elements with the (j,i) elements. If $\underset{\sim}{Y} = \underset{\sim}{X}^t$, then

$$y_{ij} = x_{ji} \quad .$$

If $\underset{\sim}{X}$ is a (N x p) matrix, then $\underset{\sim}{Y}$ is (p x N). Consider the matrix $\underset{\sim}{C}$ given above.

$$\underset{\sim}{C}^t = \begin{pmatrix} 14 & -3 \\ -2 & -10 \\ -14 & -7 \\ -6 & 2 \end{pmatrix}$$

If $\underset{\sim}{X}$ is a square symmetric matrix, $\underset{\sim}{X} = \underset{\sim}{X}^t$; i.e., transposition does not change the matrix.

The matrix $\underset{\sim}{X}^t\underset{\sim}{X}$ is quite important in regression. $\underset{\sim}{X}^t\underset{\sim}{X}$ is the matrix multiplication of the transpose of an (N x p) data matrix $\underset{\sim}{X}$ with the data matrix. $\underset{\sim}{X}^t\underset{\sim}{X}$ is a square matrix of dimension (p x p). Let $\underset{\sim}{Y} = \underset{\sim}{X}^t$, and $\underset{\sim}{C} = \underset{\sim}{X}^t\underset{\sim}{X} = \underset{\sim}{Y}\,\underset{\sim}{X}$. $\underset{\sim}{C}$ has elements

$$c_{ij} = \sum_k y_{ik}\, x_{kj}$$

$$= \sum_k x_{ki}\, x_{kj}$$

since $y_{ik} = x_{ki}$. The diagonal elements of $\underset{\sim}{C}$, c_{ii}, are the sums of the squares of the columns of $\underset{\sim}{X}$:

$$c_{ii} = \sum_k x_{ki}^2 \quad .$$

The off-diagonal elements, elements with $i \neq j$, are the sums of the ith column of $\underset{\sim}{X}$ multiplied by the jth column of $\underset{\sim}{X}$ and are called "cross-products". Note that $\underset{\sim}{C}$ is symmetric:

$$c_{ji} = \sum_k y_{jk} \cdot x_{ki} = \sum_k x_{kj} \cdot x_{ki} = \sum_k x_{ki} \cdot x_{kj} = c_{ij} \quad .$$

The matrix $\underset{\sim}{X}^t\underset{\sim}{X}$ is called the matrix of sums of squares and cross-products.

Just as multiplication has a unique identity element, 1, matrix multiplication has an identity matrix $\underset{\sim}{I}$. $\underset{\sim}{I}$ is a square (pxp) matrix, with ones on the diagonal, and zeros elsewhere:

$$\underset{\sim}{I} = \begin{pmatrix} 1 & 0 & 0 & . & . & . & 0 & 0 \\ 0 & 1 & 0 & . & . & . & 0 & 0 \\ 0 & 0 & 1 & . & . & . & 0 & 0 \\ . & . & . & & & & & \\ . & . & . & & & & & \\ . & . & . & & & & & \\ 0 & 0 & 0 & . & . & . & 1 & 0 \\ 0 & 0 & 0 & . & . & . & 0 & 1 \end{pmatrix} .$$

Multiplication of an (N x p) matrix $\underset{\sim}{X}$ by $\underset{\sim}{I}$ yields $\underset{\sim}{X}$:

$$\underset{\sim}{X}\,\underset{\sim}{I} = \underset{\sim}{I}\,\underset{\sim}{X} = \underset{\sim}{X}.$$

Division of matrices is quite complicated. The process is known as matrix inversion and is defined only for square matrices. If $\underset{\sim}{Y}$ is a (p x p) matrix, the inverse of $\underset{\sim}{Y}$ is denoted $\underset{\sim}{Y}^{-1}$, such that

$$\underset{\sim}{Y}\,\underset{\sim}{Y}^{-1} = \underset{\sim}{Y}^{-1}\,\underset{\sim}{Y} = \underset{\sim}{I} \quad .$$

Inverting a large matrix cannot be done without the aid of a computer. For small matrices, we have the following result:

If $\underset{\sim}{Y}$ is a (2x2) matrix, then

$$\underset{\sim}{Y}^{-1} = \begin{pmatrix} y_{22}/(y_{11}y_{22} - y_{12}y_{21}) & -y_{12}/(y_{11}y_{22} - y_{12}y_{21}) \\ -y_{21}/(y_{11}y_{22} - y_{12}y_{21}) & y_{11}/(y_{11}y_{22} - y_{12}y_{21}) \end{pmatrix}$$

Determining the inverse of the matrix $\underset{\sim}{X}^t\underset{\sim}{X}$, $(\underset{\sim}{X}^t\underset{\sim}{X})^{-1}$, is the "key computation" in multiple regression.

References:

Section 3.

Snedecor, G. W. and Cochran, W. G. Statistical Methods, Sixth Edition, Iowa State University Press, Ames, Iowa, 1967, Chapters 6 and 13.

Section 4.

Green, P. E. and Carroll, J. D. Mathematical Tools for Applied Multivariate Analysis, Academic Press, New York, 1967, Chapter 2.

Haeussler, E. and Paul, R. Introductory Mathematical Analysis: For Students of Business and Economics, Second Edition, Reston Publishing Co., Reston, Va., 1967, Chapter 14.

Nering, E. D. Linear Algebra and Matrix Theory, Second Edition, John Wiley & Sons, New York, 1970.

Homework
Prerequisite Inventory, Unit 4

Let:

$$A = \begin{pmatrix} 1 & 6 & 9 \\ 3 & 2 & 1 \\ 4 & 7 & 0 \\ 0 & 1 & 3 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 7 & 2 \\ 9 & 1 & 14 \\ 8 & 5 & 2 \\ 7 & 4 & 3 \end{pmatrix} \quad C = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \end{pmatrix} \quad D = \begin{pmatrix} 9 & 3 \\ 1 & 6 \end{pmatrix}$$

State whether the following operations are valid, and if so, compute the resulting matrix.

(1) $A + D$

(2) $A\ C$

(3) A^t

(4) C^tC

(5) $A + B$

(6) $C^{-1}\ C$

(7) D^{-1}

(8) B^{-1}

(9) A^tB

(10) $A + I$

(11) If Z is a (5x5) null matrix, what is ZI?

(12) What is the diagonal of the matrix C given above?

(13) Are the off diagonal terms of the matrix $B + B^t$ well defined?

(14) Are any of the above matrices symmetric?

(15) Prove, that for any square matrix F, $F + F^t$ is symmetric.

(16) Compute: A^tA.

Homework Solutions
Prerequisite Inventory, Unit 4

1. Invalid operation, must have equal dimensions.

2. $$\begin{pmatrix} 1 & 6 & 9 \\ 3 & 2 & 1 \\ 4 & 7 & 0 \\ 0 & 1 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 40 & 26 & 24 \\ 10 & 10 & 14 \\ 18 & 15 & 26 \\ 11 & 7 & 5 \end{pmatrix}$$

3. $$\begin{pmatrix} 1 & 3 & 4 & 0 \\ 6 & 2 & 7 & 1 \\ 9 & 1 & 0 & 3 \end{pmatrix}$$

4. $$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 14 & 10 & 10 \\ 10 & 9 & 10 \\ 10 & 10 & 14 \end{pmatrix}$$

5. $$\begin{pmatrix} 2 & 13 & 11 \\ 12 & 3 & 15 \\ 12 & 12 & 2 \\ 7 & 5 & 6 \end{pmatrix}$$

6. $$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

7. $$\begin{pmatrix} \frac{6}{9\cdot 6 - 3} & -\frac{3}{9\cdot 6 - 3\cdot 1} \\ -\frac{1}{9\cdot 6 - 3\cdot 1} & \frac{9}{9\cdot 6 - 3\cdot 1} \end{pmatrix} = \begin{pmatrix} \frac{6}{51} & -\frac{3}{51} \\ -\frac{1}{51} & \frac{9}{51} \end{pmatrix} = \begin{pmatrix} \frac{2}{17} & -\frac{1}{17} \\ -\frac{1}{51} & \frac{3}{17} \end{pmatrix}$$

8. Invalid operation, must be a square matrix.

9. $$\begin{pmatrix} 1 & 3 & 4 & 0 \\ 6 & 2 & 7 & 1 \\ 9 & 1 & 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 7 & 2 \\ 9 & 1 & 14 \\ 8 & 5 & 2 \\ 7 & 4 & 3 \end{pmatrix} = \begin{pmatrix} 60 & 30 & 52 \\ 87 & 83 & 57 \\ 39 & 76 & 41 \end{pmatrix}$$

10. Invalid operation, the identity matrix, $\underset{\sim}{I}$, is a square matrix.

11. $\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ (that is, $\underset{\sim}{Z}$)

12. $x_{11} = 1$
$x_{22} = 1$
$x_{33} = 1$

13. No, because only in a square matrix is the diagonal well defined.

14. Yes, $\underset{\sim}{C}$

15.

$$\underset{\sim}{F} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \qquad \underset{\sim}{F}^t = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & & & \vdots \\ \vdots & & & \vdots \\ a_{1n} & \cdots\cdots & & a_{nn} \end{pmatrix}$$

$$\underset{\sim}{F} + \underset{\sim}{F}^t = \begin{pmatrix} a_{11} + a_{11} & a_{12} + a_{21} & \cdots & a_{1n} + a_{n1} \\ a_{21} + a_{12} & a_{22} + a_{22} & \cdots & a_{2n} + a_{n2} \\ \vdots & & & \\ a_{n1} + a_{1n} & \cdots\cdots\cdots\cdots & & a_{nn} + a_{nn} \end{pmatrix}$$

In $\underset{\sim}{F} + \underset{\sim}{F}^t$, the (i,j) element is $a_{ij} + a_{ji}$, which is equal to the (j,i) element of $a_{ji} + a_{ij}$.

16. $\begin{pmatrix} 26 & 40 & 12 \\ 40 & 90 & 59 \\ 12 & 59 & 91 \end{pmatrix}$

Lecture 4-0. Introduction to Unit 4

Introduction to Unit 4, Multiple Regression

Lecture Content:

1. Introduction to objectives, problem, and notation of Unit 4
2. Introduction to the geometric representation of multiple regression

Main Topics:

1. Specific introduction to objectives of Unit 4
2. Notation for Unit 4
3. Introduction to general problem of Unit 4

Topic 1. Specific Introduction to Objectives of Unit 4:

I. Questions to be answered in Unit 4 (1)

1. What is an (x_{ij}, y_i) observational batch?

   a. Data set consisting of p+1 batches each containing N observations

   b. Data set containing p independent or x variables and 1 dependent or y variable

   c. The ith observation of the y batch, y_i, is associated with the ith observation of each of the p x batches

   d. We have, thus, a batch of N associated observations on p+1 variables

2. What analyses can be done on a batch of 1 y and multiple x variable data?

   a. What kind of summary can we use to describe the data?

      ans: Express conditional typical y as linear function of x's.

   b. How do we estimate fit.

      ans: Use least squares in multiple regression

   c. How do we determine whether transformations would improve the summary?

      ans: Examine individual x,y batches

   d. How do we adjust summarization to handle special situations in x's?

      ans: Indicator (dummy variables), splines, interactions, quadratic terms

   e. How do we judge whether the summary summarizes the data effectively?

      ans: Inference on least squares

   f. How do we judge whether the individual x variables are related to the y variable in important ways?

      ans: (t statistic) Inference on coefficients

g. **How can we determine whether our evaluation of the summary was appropriate?**

ans: Study residuals

h. **How can we determine whether the fitting procedure itself was appropriate for the data?**

ans: For knowledge of problems in least squares and awareness of alternative techniques

Topic 2. Notation

1. The ith observations on each of the p + 1 variables are associated. Thus we can represent them as an arrangement and a point in p + 1 space: (2)

$$(x_{i1}, x_{i2}, \ldots\ldots x_{ij} \ldots\ldots x_{ip}, y_i)$$

We adopt the convention that the first subscript indicates the observation, the second indicates the variable

2. Piling these arrangements on top of one another and separating the independent from the dependent variable yields an Nxp matrix of x values and a column vector, px1, of y values

$$\begin{array}{cc} & \text{p variables} \\ \begin{array}{c}\text{N} \\ \text{observations}\end{array} & \begin{bmatrix} x_{11} & x_{12} & \cdots\cdots & x_{1j} & \cdots\cdots & x_{1p} \\ x_{21} & x_{22} & \cdots\cdots & x_{2j} & \cdots\cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots\cdots & x_{ij} & \cdots\cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots\cdots & x_{Nj} & \cdots\cdots & x_{Np} \end{bmatrix} \end{array} \quad \text{and} \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_p \end{bmatrix}$$

3. In matrix notation this can be written

$$\underset{\sim}{X} \quad \text{and} \quad \underset{\sim}{y}$$

where $\underset{\sim}{X}$ is Nxp and $\underset{\sim}{y}$ is px1

(Recall that upper case letters denote matrices and lower case denote vectors.)

(Note also that $\underset{\sim}{X}$ may contain a column of 1s for least squares.)

4. We will be studying equations with multiple x's. Use b_k notation for coefficients.

Thus, $y = a + bx$, $\quad y = b_0 + b_1x$

and generally,

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots\ldots b_px_p$$

We refer to the b_k as coefficients or parameters; b_0 is the "constant" term. The equations are linear in the coefficients.

Topic 3. Introduction to general problem of Unit 4

1. What are some examples of one Y and multiple X data?

   a. Cost of accident and speed, weight of car, age of driver.

      Y is cost in dollars (continuous)

      X_1 is speed in mph (continuous)

      X_2 is weight in pounds (continuous)

      X_3 is age in years (cont... uous)

      Q: What is typical cost of an accident given speed, weight, age?

      What are the marginal effects of speed, weight and age on the typical cost?

   b. IQ scores and average pupil ... in a school system, age of pupil, number of siblin... ...rder, sex.

      Y is IQ score continuous)

      X_1 is outlay in dollars (continuous)

      X_2 is age in years (continuous)

      X_3 is sibling count (discrete)

      X_4 is birth order (discrete)

      X_5 is sex (0/1) (indicator)

      Q: What is typical IQ score given: outlay, age, siblings, order, sex?

      What are their marginal effects?

      Are all important?

   c. Median years of education in a Pittsburgh census tract and population density, median age, percent poor, percent nonwhite.

      Y is education

      X_1 is population density

      X_2 is age

X_3 is poverty

X_4 is nonwhite

(X_5 is poverty x nonwhite) interaction

2. How do we construct the summary?

Data = Fit + Residual

= Conditional Typical + Residual

$= C(y|X_1X_2 \ldots X_p) +$ Residual

We assume:

$$C(y|X_1X_2 \ldots X_p) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

and for the ith observation:

$$C(y_i|X_{i1}X_{i2} \ldots X_{ip}) = \hat{y}_i = b_0 + b_1X_{i1} + b_2X_{i2} + \ldots + b_pX_p$$

thus:

$$y_i = \hat{y}_i + R_i$$

Note: the equation is linear since exponents of the bs are all 1. The Xs may have any exponents. Just read X^k as Z.

3. <u>An example</u>: Nations data

a. Representation

Y = life expectancy

X_1 = per capita income

X_2 = infant mortality

We have for each nation (99 observations)

$$\hat{y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2}$$

Pile on top of one another

$$\hat{y}_1 = b_0(1) + b_1 X_{11} + b_2 X_{12}$$
$$\hat{y}_2 = b_0(1) + b_1 X_{21} + b_2 X_{22}$$
$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$
$$\hat{y}_{99} = b_0(1) + b_1 X_{99\,1} + b_2 X_{99\,2}$$

or in matrix notation

$$\hat{\underset{\sim}{y}} = \underset{\sim}{X}\underset{\sim}{b} \qquad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{99} \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & & \\ 1 & X_{99\,1} & X_{99\,2} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$$

b. Visualization

Equation $\hat{y} = b_0 + b_1 X_1 + b_2 X_2$ is plane in 3-space (3)

c. What does summary involve geometrically?

i. Simplified case: X variables High or Low (4)
Schematize
Display in 3-space
Fit summary--Connect medians? (5)
Fit plane?

ii. General situation: continuous data (6)
point cloud (plotting) (7)
consider solid (8)
fit plane (9)
equation $\hat{y} = b_0 + b_1 X_1 + b_2 X_2$
residuals from y
Interpretation

d. How do we choose the plane?

e. What about more X's?

f. Transformations?

Lecture 4-0
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Topic 1 Section A. 1.	1	Questions for Unit 4
Topic 2 1.	2	Data Representation
Topic 3 Section 3.b	3	Plane defined by X_1 and X_2
c.i	4 overlay 3	Schematic plots
c.i	5 overlay 4	Plane connecting Conditional typicals
c.ii	6 overlay 5	Point cloud
c.ii	7 overlay 6	Ellipsoid defined by Point Cloud
c.ii	8 overlay 7	Regression plane passing through ellipsoid
c.ii	9 overlay 8	Residuals from ellipsoid to plane

Questions for Unit 4: [1]

1. What is an (X_{ij}, Y_i) observational batch?
2. What analyses can be done on such batches?
3. How do we interpret and evaluate these analyses?

Data Representation: [2]

Associated observations on p independent variables and one dependent variable:

$$(X_{i1}, X_{i2}, \ldots X_{ij}, \ldots\ldots X_{ip}, Y_i)$$

Data matrix of N observations on p independent variables and vector of N observations on one dependent variable:

$$\begin{array}{cc} & \text{p variables} \\ \begin{array}{c}\text{N} \\ \text{observations}\end{array} & \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{ij} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{i1} & X_{i2} & & X_{ij} & & X_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{Nj} & \cdots & X_{Np} \end{pmatrix} \text{ and } \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_N \end{pmatrix} \end{array}$$

Matrix representations:

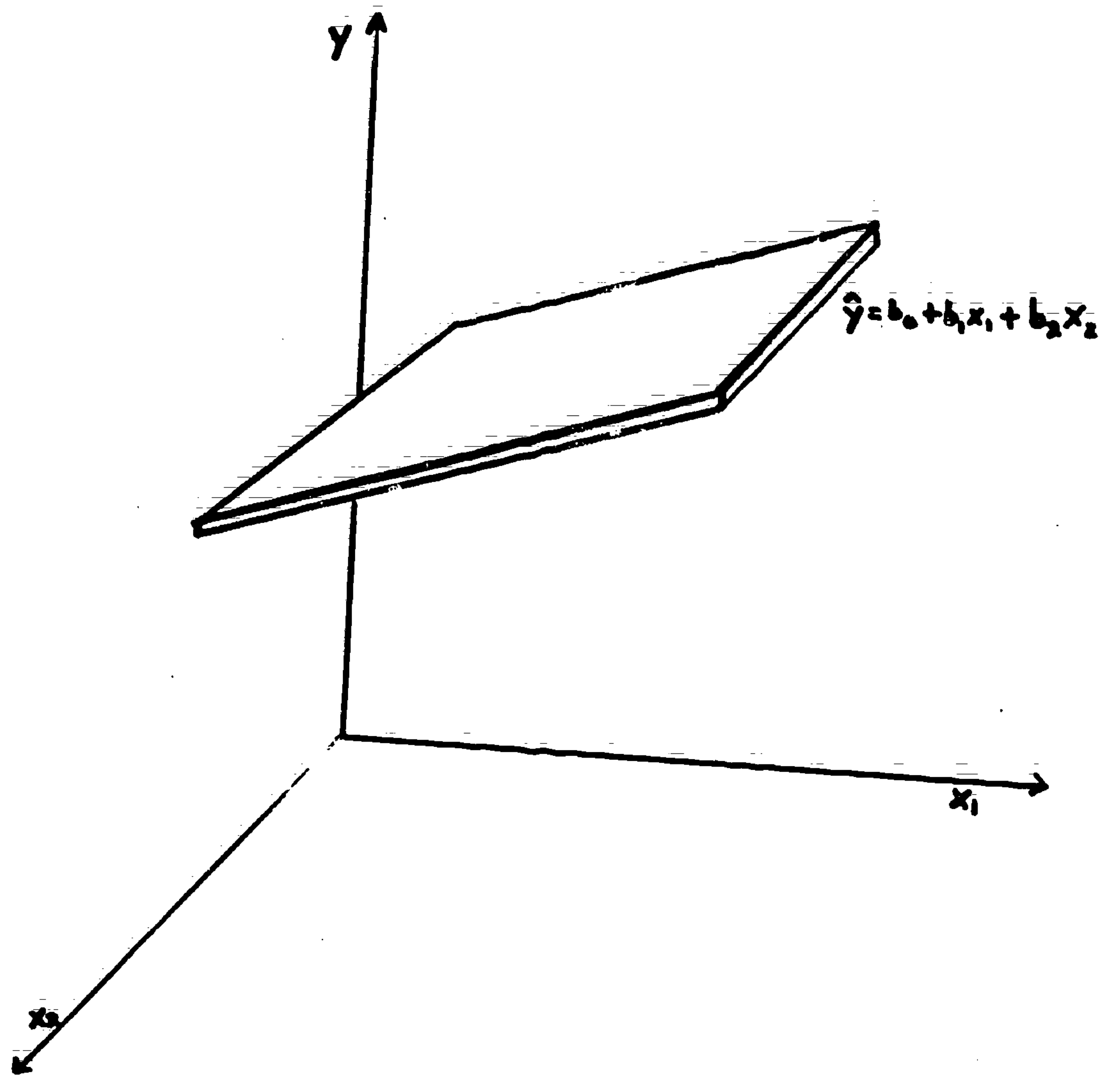
$\underset{\sim}{X}$ and $\underset{\sim}{Y}$

Linear Equation:

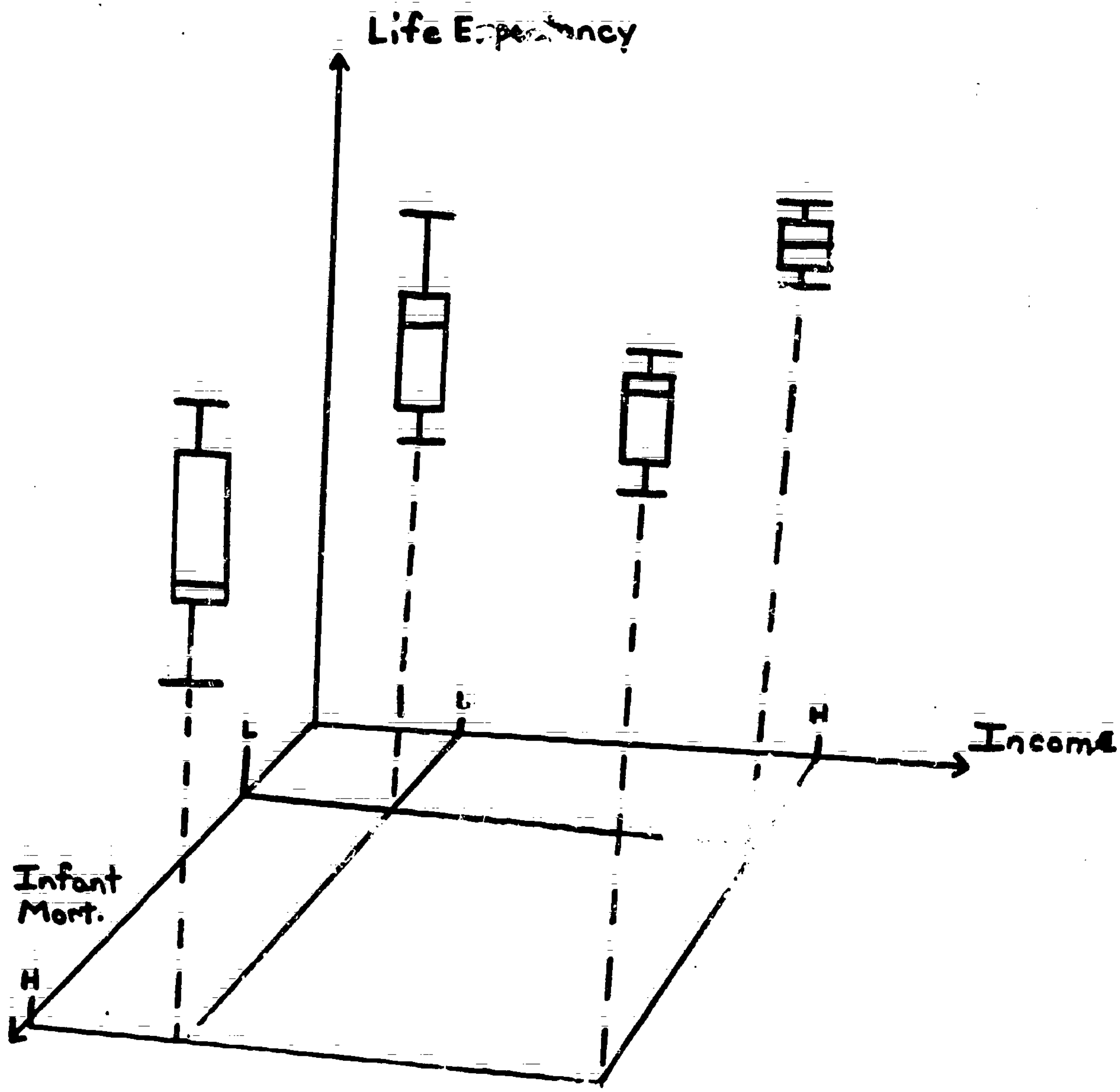
$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots\ldots + b_p X_p$$

QMPM

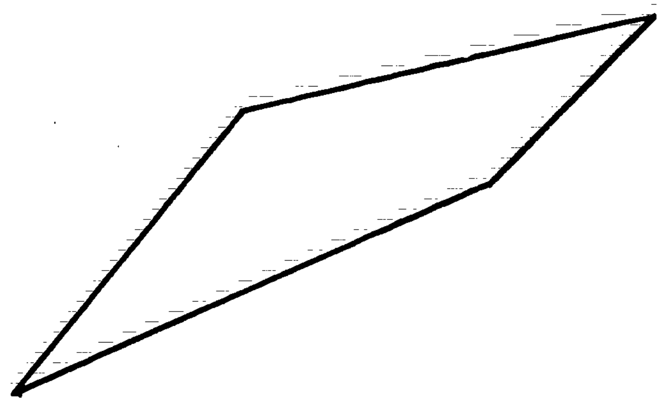
[3]

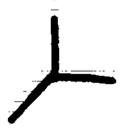


[4]



[5]

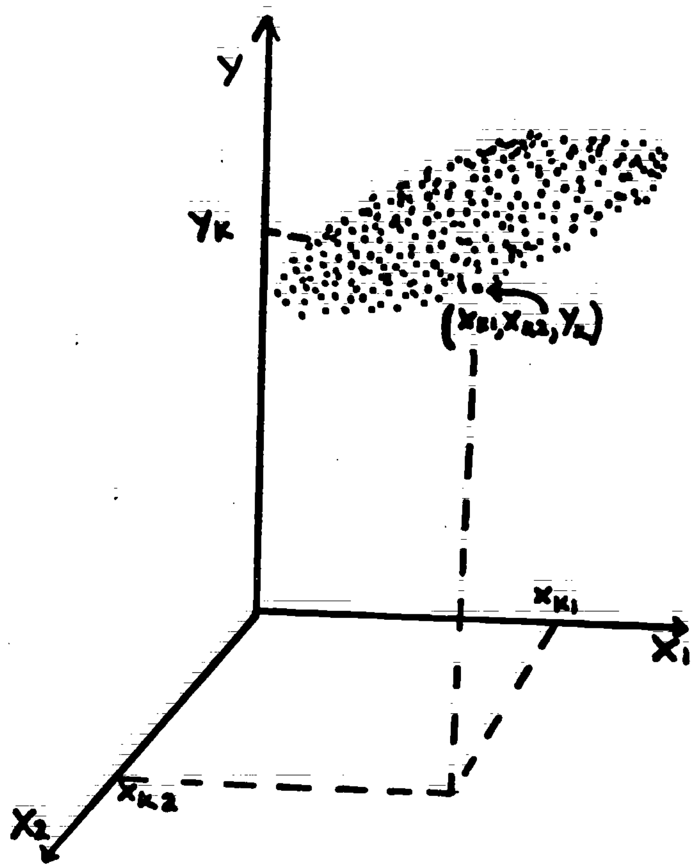




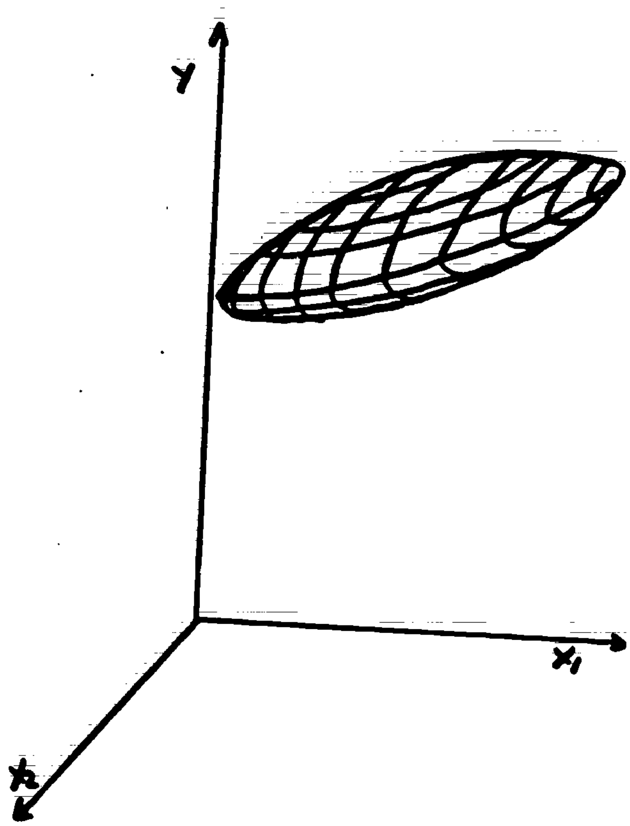
575

4-0

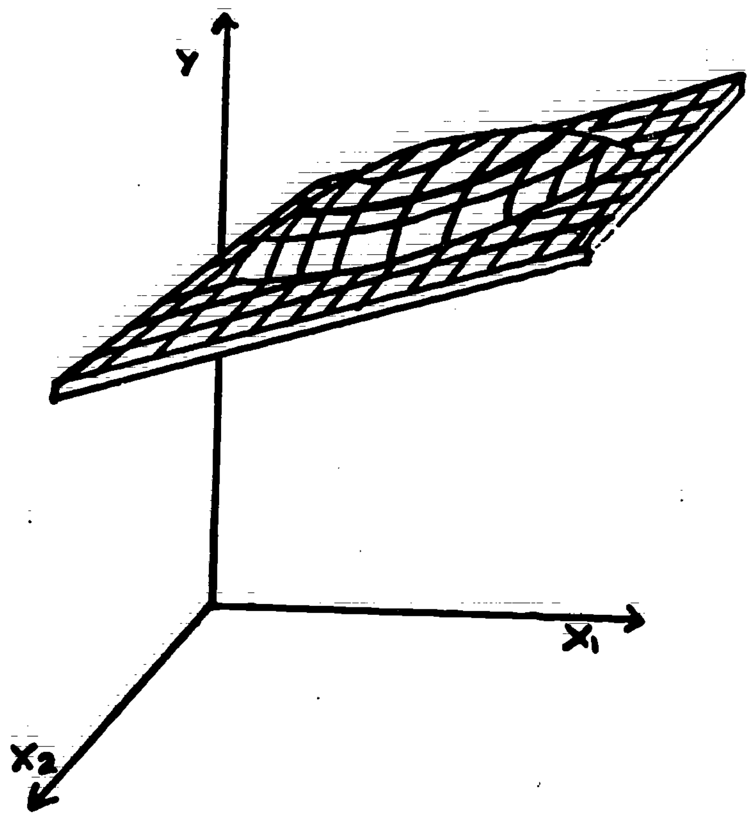
[6]

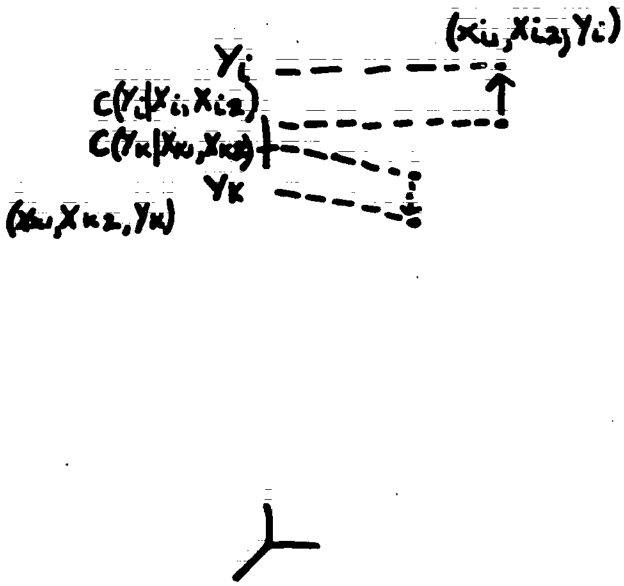


[7]



[8]





Plane has equation

$$C(Y|X_1, X_2) = b_0 + b_1 X_1 + b_2 X_2$$

Residuals

$$r_i = y_i - c(y_i | x_{i1} x_{i2})$$

$$r_k = y_k - c(y_k | x_{k1} x_{k2})$$

Lecture 4-1. Multiple Regression Using Least Squares

Multiple Regression using Least Squares: Algebraic Computations

Lecture Content:

1. The Model
2. Least Squares Estimation

Main Topics:

1. Algebraic Representation of the Model
2. Matrix Version
3. Least Squares Solution -- General
4. Least Squares Solution--Univariate
5. Examples of Computer Generated Fits

(There are no transparencies for this lecture. Material should be developed on blackboard.)

Topic 1. The Model

I. Basic Issue--Presentation of Model

1. General case: N observations, p variables

2. ith Equation

$$\hat{y}_i = C(y_i | X_{i1}, X_{i2}, \ldots, X_{ip})$$
$$= b_0 + b_1 X_{i1} + \ldots + b_p X_{ip}$$

3. Note that this equation is linear in b_i

4. Matrix notation

$$(\hat{y}_i) = (1 \; X_{i1} \; \ldots \; X_{ip}) \begin{pmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_p \end{pmatrix}$$

or $\hat{y}_i = \underset{\sim}{X}_i \underset{\sim}{b}$

5. We can "stack" the $\hat{y}_i$ into a vector and the $\underset{\sim}{X}_i$ into a matrix. Remember, there are N rows.

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \cdot \\ \cdot \\ \hat{y}_N \end{pmatrix} = \begin{pmatrix} 1 & X_1 & X_{12} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots & X_{2p} \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ 1 & X_{N1} & X_{N2} & \ldots & X_{Np} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_p \end{pmatrix}$$

or $\hat{\underset{\sim}{y}} = \underset{\sim}{X} \underset{\sim}{b}$

II. Problem--This is only a conceptual model

1. It defines a surface in p+1 dimensions

2. General equation for surface:

$$C(y | X_1, X_2, \ldots, X_p) = b_0 + b_1 X_1 + \ldots + b_p X_p$$

3. The actual data points (y_i's) do not lie exactly on it

4. How do we choose the b's such that the surface is a reasonable summary of the point cloud

5. We also desire the $\hat{y}_i$'s to be "reasonable" typical values of y_i given $\underset{\sim}{X}_i$.

582

Topic 2. Least Squares Estimation

I. Basic Issue--Minimize sum of squared residuals

1. Choose b_i so that

$$\Sigma(y_i - \hat{y}_i)^2 \text{ is minimized}$$

2. Solution:

$$\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{y}$$

3. $X'X$ must be "non-singular"

4. $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{b}$

II. Solution--Least squares calculations

1. $(\underset{\sim}{X}'\underset{\sim}{X}) =$

$$\begin{pmatrix} N & \Sigma X_{i1} & \Sigma X_{i2} & \cdots\cdots & \Sigma X_{ip} \\ \cdot & \Sigma X_{i1}^2 & \Sigma X_{i1} X_{i2} & \cdots & \Sigma X_{i1}X_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ & & & & \Sigma X_{ip}^2 \end{pmatrix}$$

Symmetric matrix. Sums of squares and cross-products

2. $\underset{\sim}{X}'\underset{\sim}{Y} =$

$$\begin{pmatrix} \Sigma Y_i \\ \Sigma X_{i1} Y_i \\ \Sigma X_{i2} Y_i \\ \vdots \\ \Sigma X_{ip} Y_i \end{pmatrix}$$

583

3. These calculations are straightforward computationally

4. Difficult task is inverting $(\underset{\sim}{X}'\underset{\sim}{X})$

III. Method--Univariate Situation (p=1)

1. $\underset{\sim}{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \qquad \underset{\sim}{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_N \end{pmatrix} \qquad \underset{\sim}{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$

2. $C(\underset{\sim}{y}|X) = \underset{\sim}{\hat{y}} = \underset{\sim}{X}\underset{\sim}{b} = \begin{pmatrix} b_0 + b_1X_1 \\ b_0 + b_1X_2 \\ \vdots \\ b_0 + b_1X_N \end{pmatrix}$

3. To determine $\underset{\sim}{b}$ and hence calculate $\underset{\sim}{\hat{y}}$, we must compute

$$\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{y}$$

4. Calculation

a. First evaluate

$$(\underset{\sim}{X}'\underset{\sim}{X}) = \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_N \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_N \end{pmatrix}$$

$$= \begin{pmatrix} N & \Sigma X_i \\ \Sigma X_i & \Sigma X_i^2 \end{pmatrix}$$

584

b. Secondly, evaluate

$$(\underset{\sim}{X}'\underset{\sim}{y}) = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_N \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{pmatrix}$$

$$= \begin{pmatrix} \Sigma\, y_i \\ \Sigma\, y_i X_i \end{pmatrix}$$

c. Thirdly, we must compute $(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$

i. If $\underset{\sim}{M} = \begin{pmatrix} ab \\ cd \end{pmatrix}$, then

$$\underset{\sim}{M}^{-1} = \begin{pmatrix} d/ad-bc & -b/ad-bc \\ -c/ad-bc & a/ad-bc \end{pmatrix}$$

ii. But $\underset{\sim}{X}'\underset{\sim}{X}$ is symmetric, and $b=c$

iii. Hence $\underset{\sim}{M}^{-1} = \dfrac{1}{ad-b^2}\begin{pmatrix} d & -b \\ -b & a \end{pmatrix}$

iv. Note: If $ad-b^2 \approx 0$, $\underset{\sim}{M}^{-1}$ cannot be computed

v. This occurs when the X's are nearly constant

vi. We have

$$\begin{aligned} a &= N \\ b &= \Sigma X_i \\ d &= \Sigma X_i^2 \\ ad - b^2 &= N\,\Sigma X_i^2 - (\Sigma X_i)^2 \\ &= N\,\Sigma(X_i-\bar{X})^2 \end{aligned}$$

vii. Hence

$$(\underset{\sim}{X}'\underset{\sim}{X})^{-1} = \frac{1}{N\,\Sigma(X_i-\bar{X})^2}\begin{pmatrix} \Sigma X_i^{\,2} & -\Sigma X_i \\ -\Sigma X_i & N \end{pmatrix}$$

585

d. Lastly,

$$\underset{\sim}{b} = \frac{1}{N \Sigma (X_i - \bar{X})^2} \begin{pmatrix} \Sigma X_i^2 & -\Sigma X_i \\ -\Sigma X_i & N \end{pmatrix} \begin{pmatrix} \Sigma y_i \\ \Sigma X_i y_i \end{pmatrix}$$

$$= \frac{1}{N \Sigma (X_i - \bar{X})^2} \begin{pmatrix} \Sigma X_i^2 \Sigma y_i - \Sigma X_i y_i \Sigma X_i \\ N \Sigma X_i y_i - \Sigma X_i \Sigma y_i \end{pmatrix}$$

e. Thus,

$$b_0 = \frac{\Sigma X_i^2 \Sigma y_i - \Sigma X_i y_i \Sigma X_i}{N \Sigma (X_i - \bar{X})^2}$$

$$= \frac{\bar{y} \Sigma X_i^2 - x \Sigma X_i y_i}{N \Sigma (X_i - \bar{X})^2}$$

$$b_1 = \frac{N \Sigma X_i y_i - \Sigma X_i \Sigma y_i}{N \Sigma (X_i - \bar{X})^2}$$

$$= \frac{\Sigma X_i y_i - \frac{1}{N} \Sigma X_i \Sigma X_i}{\Sigma (X_i - \bar{X})^2}$$

$$= \frac{\Sigma (X_i - \bar{X})(y_i - \bar{y})}{\Sigma (X_i - \bar{X})^2} = \frac{\text{cor } (X_i y)}{\text{Var}(X)}$$

$$b_0 = \bar{y} - b_1 \bar{X}$$

5. If $p > 1$, Solution is more complicated

6. Examples

a. Life Expectancy = Per capita income + infant mortality

i. $\hat{LE} = b_0 + b_1 \text{ PCI} + b_2 \text{ IM}$

$y = b_0 + b_1 X_1 + b_2 X_2$

ii. Income in \$
Mortality in deaths/1000 live births
Life expectancy in years
N = 99

iii. Model

$$\hat{y}_i = 53.36 + .005X_{1i} - .058X_{2i}$$

b. Interpretation

i. Typical increment in life expectancy for \$100 increment in income is .5 years

ii. Typical decrement in life expectancy for 100 infant deaths is 5.8 years

iii. Relate these results to batches and fitted plane

c. Is this interpretation similar to that obtained via 2 separate regressions?

No:

$$\hat{y} = 46.88 + .007X_1$$

$$\hat{y} = 61.51 - .086X_2$$

Only true when $Cov(X_1, X_2) = 0$

Lecture 4-2 Transformations

Using least squares procedures to estimate alternative functional forms for the conditional typical summary: Part I-Transformations

Lecture Content:

1. Introduction--Data Analysis and Theory Testing
2. Transformations--$X_i^{r_i}$

Main Topics:

1. Purposes of transformations
2. Constructing models with transformed variables
3. Interpreting transformed models

(There are no transparencies for this lecture.)

Topic 1: Introduction to using transformed variables in multiple linear regression

I. Basic Issue: Estimate $\underset{\sim}{b}$ in $\hat{y} = \underset{\sim}{X}\underset{\sim}{b}$ when the exponent $X_i^{r_i}$ may vary across X_i.

1. Recall $\hat{\underset{\sim}{y}} = \underset{\sim}{X}\underset{\sim}{b}$ is linear in $\underset{\sim}{b}$ and is conceptual model

2. We may want to summarize the data with

$$C(y|\, Z_1, Z_2 \ldots Z_p) = b_0 + b_1 Z_1^{r_1} + b_2 Z_2^{r_2} + \ldots b_p Z_p^{r_p}$$

we can set $Z_i^{r_i} = X_i$

and get

$$C(y|X_1, X_2 \ldots X_p) = b_0 + b_1 X_1 + b_2 X_2 \ldots + b_p X_p$$

Now we can use least squares to estimate $\underset{\sim}{b}$.

3. We may want to summarize the data with

$$\hat{y} = C(y|X_1 X_2 \ldots X_p) = X_1^{b_1} \cdot X_2^{b_2} \cdot \ldots \cdot X_p^{b_p}$$

We need transformation to linearize. Suppose we need a log transformation. Then

$$\hat{y} = \log(C|X_1, X_2) = b_1 \log X_1 + b_2 \log X_2 \text{, when } p = 2.$$

This is now linear in $\underset{\sim}{b}$.

II. Distinction between testing theory and doing data analysis

1. Theory may impose functional form for conditional typical

a. Distance in free fall: $(V_o = 0)$

$$D = \frac{1}{2} gt^2$$

Then

$$\log D = \log .5 + \log g + 2 \log t$$

But set: $y = \log D$ $\quad X_1 = \log g$ $\quad \log t = X_2$

$b_0 = \log .5$ $\quad b_1 = 1$ $\quad b_2 = 2$

Thus, $\hat{y} = b_0 + b_1 X_1 + b_2 X_2$

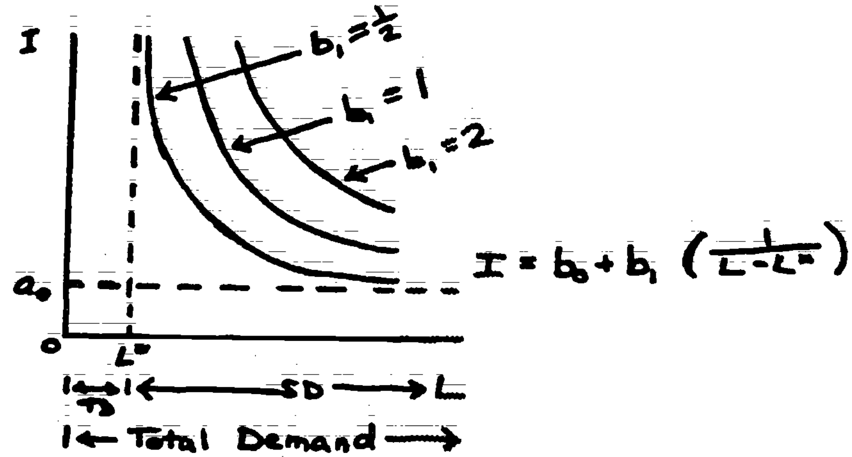
Analysis involves determining whether empirical data yield coefficients that are close to theoretical prediction.

b. Liquidity preference function.

$$I = b_0 + b_1 \left(\frac{1}{L - L^*}\right)$$

where I is interest rate
L is quantity of money
L* constant "transactions demand for money"
L-L* is speculative demand for money
b_0 is minimum level of interest

Function is



Graph is rectangular hyperbola with curvature depending on b_1.

Let $X = \frac{1}{L - L^*}$

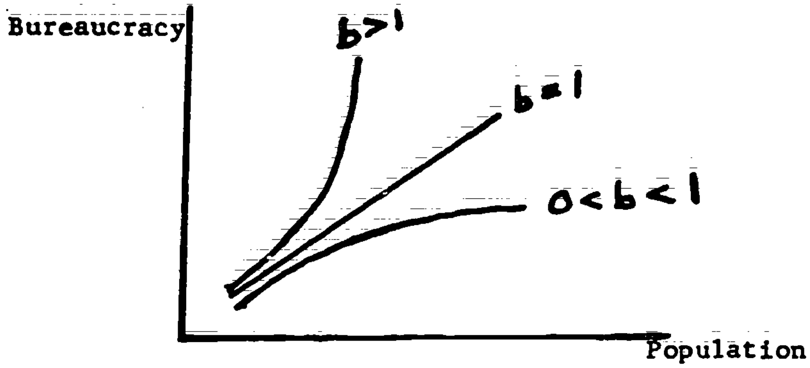
We get

$I = b_0 + b_1 X$ as model for conditional typical we have

$\hat{y} = b_0 + b_1 X$

Go backwards, i.e., untransform once b_0 and b_1 are estimated from data.

c. Several possible forms: government bureaucracy and population size



$b = 1$ constant proportion between B and P

$b < 1$ economies of scale

$b > 1$ Parkinson's law

log both sides

$\log B = \log C + b_1 \log P$

let $\log B = \hat{y}$, $\log C = b_0$, $\log P = X$

$$\hat{y} = b_0 + b_1 X$$

2. Data may require exploratory analysis to find scales (dimensions) for variables: do each variable pair (X_i, y) individually.

   a. Life expectancy vs. Infant mortality and per capita income

	b_0	b_1	b_2	R^2
LF = PCI	46.88	.007		.52
Mort	61.51		-.086	.35
PCI + Mort	53.36	.005	-.058	.65
log PCI	1.30	.158		.67
log Mort	2.08		-2.04	.59
log PCI + log Mort	1.62	.106	-.102	.75

3. Distinction between variable and regressor (carrier). May have fewer independent variables than terms in equation $\underset{\sim}{X}$ may include functions of $\underset{\sim}{X}$ and cross products of X_i

   a. Polynomial $X, X^2, X^i \ldots$ i is "order"

   b. Cross products $X_i X_j$, $X_i X_j X_k$, $X_i^2 X_j$ etc.

591

c. Interpretation of cross product terms as "interactions"

i. Mathematical

$$\hat{y} = b_0 + b_1X_1 + (b_2 + b_3X_1)X_2 \qquad X_2 \text{ has varying slope}$$

$$= b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2 \qquad \text{"different slopes for different folks"}$$

ii. Substantive--multiplicative effect

		X_1: $<$	X_1: $>$
X_2	$<$	I	II
	$>$	III	IV

	Additive	Multiplicative
I	$= \; < + <$	$< \cdot <$
II,III	$= \; < + >$	$< \cdot >$
IV	$= \; > + >$	$> \cdot >$

592

Topic 2. Transformations

I. $X_i^{r_i}$ with r positive integer

1. First order models: $r = 1$

$$\hat{y} = b_0 + b_1X_1 + b_2X_2 + \ldots\ldots b_pX_p$$

2. Second order models: $r_i = 1$ or 2

a. $\hat{y} = b_0 + b_1X_1 + b_2X_1^2$

Set $X_1^2 = Z$

$\Rightarrow \hat{y} = b_0 + b_1X_1 + b_2Z$

(One variable; two regressors. Use OLS)

b. $\hat{y} = b_0 + b_1X_1 + b_2X_1^2 + b_3X_2 + b_4X_2^2 + b_5X_1X_2$

Set $X_1^2 = Z_1$, $X_2^2 = Z_2$, $X_1X_2 = Z_3$

Get $\hat{y} = b_0 + b_1X_1 + b_2Z_1 + b_3X_2 + b_4Z_2 + b_5Z_3$

(Two variables: X_1, X_2; five regressors: X_1, X_2, X_1^2, X_2^2, X_1X_2)

3. Third order models, etc. involve 3 or more Xs multiplied (can be $X_iX_iX_i$, $X_iX_jX_k$, $X_iX_iX_k$, etc.)

4. Note: forms may be suggested by theory or residuals

II. r not a positive integer

1. Reciprocal $r = -1$

If $\hat{y} = b_0 + b_1 \frac{1}{X_1} + b_2 \frac{1}{X_2} + \ldots$

Set $Z_1 = \frac{1}{X_1}$ $\quad Z_2 = \frac{1}{X_2} \ldots$

Get $\hat{y} = b_0 + b_1Z_1 + b_2Z_2 + \ldots$

2. Logarithmic

$$\hat{y} = b_0 + b_1 \log X_1 + b_2 \log X_2 + \ldots$$

4. Square root

$$\hat{y} = b_0 + b_1X_1^{1/2} + b_2X_2^{1/2} + \ldots$$

C. In general

1. $\hat{y} = b_0 + b_1X_1^r + b_2X_2^r + \ldots$

2. $\hat{y} = b_0 + b_1X_1^{r_1} + b_2X_2^{r_2} + \ldots$

3. Use exploratory tools or theory to find each r_i

Topic 3. Interpreting transformations--Some transformations have useful substantive or mathematical interpretations

I. Univariate case--logarithms

Four cases:

		Independent Variable X: Not Logged	Independent Variable X: Logged
Dependent Variable Y	Not Logged	I	III
	Logged	II	IV

1. Case I Both dependent and independent variables linear.

Conceptual model: $\hat{y} = b_0 + b_1 x$

Differentiate both sides with respect to x:

$$\frac{d\hat{y}}{dx} = b_1$$

Thus, b_1, the slope of the line, is the amount by which $\hat{y}$ changes for a unit change in x.

2. Case II Dependent variable logged, independent variable linear. (Called "log-linear.")

Conceptual model: $\hat{y} = b_0 e^{b_1 x}$

log version (taking logs of both sides)

$$\log \hat{y} = \log b_0 + b_1 x$$

Differentiate both sides with respect to x:

$$\frac{dy}{dx}\frac{1}{\hat{y}} = b_1$$

But the left side of this is the ratio of the proportionate change in $\hat{y}$ to a unit change in x:

$$\frac{d\hat{y}}{dx}\frac{1}{\hat{y}} = \frac{d\hat{y}/\hat{y}}{dx} = b_1$$

So b_1 can be interpreted as the proportion $\hat{y}$ changes for a unit change in x or 100 x b_1 gives the percentage change in $\hat{y}$ for a unit change in x.

3. Case III Dependent variable linear, independent variable logged. (Called "linear-log.")

Conceptual model: $e^{\hat{y}} = b_0x^{b_1}$

log version:

$$\hat{y} = \log b_0 + b_1 \log x$$

Differentiate both sides with respect to x:

$$\frac{d\hat{y}}{dx} = \frac{b_1}{x}$$

Multiply both sides by x:

$$\frac{d\hat{y}}{dx}x = b_1$$

But the left hand side is the same as

$$\frac{d\hat{y}}{dx}x = \frac{d\hat{y}}{\frac{dx}{x}} = b_1$$

So b_1 can be interpreted as the ratio of the amount $\hat{y}$ changes to a proportionate change in x. Thus $b_1/100$ can be read as the amount $\hat{y}$ changes when x doubles, i.e., increases by 100%

4. Case IV Both dependent and independent variables logged (Called "log-log.")

Conceptual model: $\hat{y} = b_0x^{b_1}$

log version:

$$\log \hat{y} = \log b_0 + b_1 \log x$$

Differentiate with respect to x:

$$\frac{d\hat{y}}{dx}\frac{1}{\hat{y}} = \frac{b_1}{x}$$

596

Multiply both sides by x:

$$\frac{d\hat{y}}{dx}\frac{x}{\hat{y}} = b_1$$

But left side is ratio of proportionate change in $\hat{y}$ to proportionate change in x:

$$\frac{d\hat{y}}{dx}\frac{x}{\hat{y}} = \frac{d\hat{y}/\hat{y}}{dx/x} = b_1$$

Or it can be read as the ratio of the percentage change in $\hat{y}$ for a percentage change in x--but this is elasticity. So b_1 can be interpreted as the elasticity of y with respect to x.

II. Multiple X stiuations

1. Exponential model

$$\hat{y} = e^{b_o} + b_1X_1 + b_2X_2 + \dots$$

$$\log \hat{y} = b_0 + b_1X_1 + b_2X_2 + \dots$$

2. Reciprocal

$$\hat{y} = \frac{1}{b_0 + b_1X_1 + b_2X_2 + \dots}$$

$$\frac{1}{\hat{y}} = b_1 + b_1X_1 + b_2X_2 + \dots$$

III. Notes:

1. The least squares estimates apply to the transformed models only

2. Avoid transforming $\hat{y}$ if possible. This may have consequences for inference

3. Discuss problems in expanding the number of parameters to fit the data. Issues of parsimony, complexity, and the substantive context of the problem

4. Always redefine variables as variables, parameters as parameters

597

Lecture 4-3. Indicator Variables

Using least squares procedures to estimate alternative functional forms for for the conditional typical summary: Part II--Indicator Variables (1)

Lecture Content:

1. Constructing variables and data sets for indicator variables
2. Interpreting models containing indicator variables

Main Topics:

1. Introduction to indicator variables
2. Simple 0/1 indicator variables
3. Linear and other functional forms for indicators
4. Splines--Shifts in intercept and slope

Tools Introduced:

1. Indicator Variables
2. Splines

Topic 1: Introduction to indicator variables

I. Basic issue: Effective summary of data may require the construction of "variables" and "data" to yield an appropriate form for the conditional typical for certain data sets

1. An X may be categorical rather than continuous. Example: Race, sex, region, season

2. An X may take a known or hypothesized functional form but data may be lacking. Example: Ordered categories such as income, educational attainment

3. Time trends may be suspected Example: Growth in population, sales volume, salaries, property values, inflation rate

4. Curves, cyclic behavior, or other consistent changes in intercept and slope may be evident. Example: Admittance volume for emergencies in a hospital, number of enrolled participants in training programs

II. How can we construct alternative functional forms for the conditional typical which will permit us to use least squares estimation procedures in these special situations?

1. Use indicator (dummy or switching) variable which takes on value of 1 when special condition holds and is 0 otherwise for categorical variables

2. Use indicator variables which take on linear or other forms (quadratic, etc.) when trends or functional forms are expected and slope is different

3. Use linear splines (connected or disconnected) to track a special curve where shifts in intercept and slope are expected

599

Topic 2: Indicator variables which give rise to shifts in the intercept

I. Categorical variables only (for heuristic purposes)

The typical value may depend upon the state of the categorical variable

1. Dichotomous--0/1: Two levels or states

   a. Example: Life expectancy for industrial and nonindustrial nations

   b. We can construct a model for the conditional typical (mean) as follows: (2)

      $$C(LE \mid \text{National status}) = \hat{y} = b_0 + b_1X_1$$

      where $X_1 = \begin{cases} 1 \text{ if nation is industrial,} \\ 0 \text{ otherwise} \end{cases}$

   Interpretation:

      Effect of model is to estimate:

      $\hat{y} = b_0$ when nation is not industrial

      $\hat{y} = b_0 + b_1$ when nation is industrial

      I.e., when not industrial, model estimates typical as line horizontal to X axis with intercept = b_0. When industrial, typical, conditional on being industrial, has intercept $b_0 + b_1$. The means of the nonindustrial will be b_0, of the industrial, $b_0 + b_1$. The value of b_1 indicates how different the two groups are. (3)

      Thus X is the variable that indicates which category of nation is being considered.

   c. Can estimate model in b. using OLS:

      $\hat{y} = 49.49 + 22.18X \qquad R^2 = .44$

   Interpret result

2. Polychotomous--More then two levels or states

   a. Example: Life expectancy for industrial, nonindustrial and petroleum exporting countries

600

b. Conditional typical model

$$C(L\ E|\text{National Status}) = \hat{y} = b_0 + b_1X_1 + b_2X_2 \qquad (4)$$

where

X_1	X_2	National Status
0	0	Petroleum Exporting
0	1	Nonindustrial
1	0	Industrial

c. OLS estimates

$$\hat{y} = 50 + 21.67X_1 - .57X_2 \qquad R^2 = .43 \qquad (5)$$

3. Generally: If categorical variable has r states, r-1 indicator variables are required (6)

II. Continuous and categorical variables combined. We can combine continuous and categorical variables to yield summaries of data. (Cf. analysis of covariance.)

1. Dichotomous indicator

   a. <u>Example</u>: Life expectancy by per capita income and industrial status

   b. Conceptual model

$$C(LE\,|\text{status}) = \hat{y} = b_0 + b_1X_1 + b_2X_2$$

where

X_1 is per capita income

$$X_2 = \begin{cases} 1 \text{ if nation is industrial} \\ 0 \text{ otherwise} \end{cases}$$

   c. Interpretation of b_i

$\hat{y} = b_0 + b_1X_{i1}$ when nation i is not industrial

$\hat{y} = (b_0 + b_2) + b_1X_{i1}$ when nation i is industrial

i.e., b_2 is typical shift in LE conditional on being industrial. Note that the slope of the line relating life expectancy and per capita income is the same for both industrial and non-industrial nations; only the level is different.

601

d. OLS estimates:

$$\hat{y} = 47.15 + .005X_1 + 4.72X_2$$

2. Another <u>example</u>: Life expectancy by log(per capita income) and industrial status (7)

   a. Conceptual model:

   $$\hat{y} = b_0 + b_1 \log X_1 + b_2X_2$$

   where X_1 and X_2 are defined in 1.

   b. OLS estimates:

   $$\hat{y} = 4.74 + 18.51X_1 + 1.61X_2 \qquad R^2 = .72 \qquad (8)$$

3. Polychotomous indicator

   a. <u>Example</u>: Life expectancy by per capital income, infant mortality, industrial and petroleum exporting status

   b. Conceptual model:

   $$C(LE|PCI, IM, Status) = \hat{y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

   where X_1 is per capita income
   X_2 is infant mortality
   X_3 and X_4 are indicator variables for status

X_3	X_4	status
1	0	industrial
0	1	nonindustrial
0	0	petroleum exporting

   c. Interpreting b_i

   $$\hat{y} = b_0 + b_1X_1 + b_2X_2 \text{ when petroleum exporting}$$

   $$= (b_0 + b_3) + b_1X_1 + b_2X_2 \text{ when industrial}$$

   $$= (b_0 + b_4) + b_1X_1 + b_2X_2 \text{ when nonindustrial}$$

   d. OLS estimates (9)

   $$\hat{y} = 54.74 + .005X_1 - .059X_2 - 1.77X_3 - 1.36X_4 \qquad R^2 = .65$$

III. More complex situations

Indicator, continuous, transformed, polynomial, and interaction variables may be combined to construct effective summary of data

1. <u>Example</u>: Continue Life Expectancy

a. Is an interaction relevant?

i. Estimate: $\hat{y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$

where X_1 is per capita income
X_2 is infant mortality
X_1X_2 is multiplicative interaction term

ii. OLS estimates:

$$\hat{y} = 55.95 + .004X_1 - .088X_2 + 2.69 \times 10^{-5}X_1X_2$$

(all are significant) $R^2 = .68$

b. $\hat{y} = b_0 + b_1\text{INC} + b_2\text{Mort} + b_3\text{IND} + b_4\text{NonIND} + b_5\text{INC}\cdot\text{Mort}$

$$\hat{y} = 53.89 + .003X_1 - .09X_2 + 5.97X_3 + 2.63X_4 + 3.31 \times 10^{-5}X_1X_2$$

(discuss change in sign in national status variables)

$R^2 = .68$

c. $\hat{y} = b_0 + b_1\log\text{INC} + b_2\log\text{Mort} + b_3\text{Ind} + b_4\text{NonInd} + b_5\log\text{Inc}\cdot\log\text{Mort}$

$$\hat{y} = 46.53 + 9.33\log X_1 - 19.87\log X_2 + 4.6X_3 + 3.0X_4 + 3.1\,(\log X_1 \log X_2)$$

$R^2 = .80$

2. Another application: Seasonal Shifts

a. <u>Example</u>: Smoothed D.C. General Hospital % emergency admits by month 1970-1975
(Note--recall data from lecture on smoothing)

603

b. Conceptual model:

$$C(PEA|Season) = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

where:

$X_1 = 1$ iff summer (June, July, Aug.);
$X_2 = 1$ iff fall (Sept., Oct., Nov.);
$X_3 = 1$ iff winter (Dec., Jan., Feb.);

b_0 gives baseline for spring

c. OLS estimates:

$$\hat{y} = 8.2 + .4X_1 + .4X_2 - .2X_3 \qquad \bar{R}^2 = .50$$

Discuss shape of effects; comparative level (10)

3. More than 1 indicator variable

a. <u>Example</u>: Income by race and sex

b. Conceptual model:

$$C(Inc|Race,Sex) = b_0 + b_1X_1 + b_2X_2 \qquad (11)$$

where X_1 indicates race (two categories)
X_2 indicates sex (two categories)

c. Variable definitions:

		Race (X_1) B 0	W 1
Sex(X_2)	M 0	MB 00	MW 01
	F 1	FB 10	FW 11

d. Interpreting b_i:

b_0 is level for male black

$b_0 + b_1$ is level for male white

$b_0 + b_2$ is level for female black

$b_0 + b_1 + b_2$ is level for female white

e. Alternative race-sex indicator using interaction (12)

i. Assume single variable with four levels:

X_1	X_2	X_3	
0	0	0	male black
0	0	1	male white
0	1	0	female black
1	0	0	female white

This structure does not assume additivity of race and sex effects

ii. Model:

$$\hat{Inc} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

iii. Interpreting b_i

b_0 is level of male black

$b_0 + b_3$ is level of male white

$b_0 + b_2$ is level of female black

$b_0 + b_1$ is level of female white

Topic 3: Numeric indicator variables--known or presumed functional forms for an X variable--estimates of slopes

I. Constructed variables

1. Presumed linear

a. Example: Life expectancy by income category--low, middle, high

b. Conceptual model:

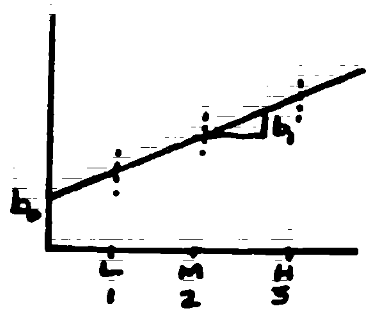
$$C(LE|Income) = b_o + b_1X_1$$

where: $X_1 = \begin{cases} 1 & \text{if low} \\ 2 & \text{if middle} \\ 3 & \text{if high} \end{cases}$

N = 29
N = 19
N = 23
Σ = 71

c. OLS estimates:

$$\hat{Y} = 30.35 + 9.96X \qquad R^2 = .55$$



2. Presumed logarithmic

a. $\hat{y} = b_0 + b_1 \log X$ where $X = \begin{cases} 1 & \text{low} \\ 2 & \text{middle} \\ 3 & \text{high} \end{cases}$

b. OLS

$$\hat{y} = 39.94 + 40.34\log_{10}X$$

3. Other forms: quadratic, etc.

4. Time trends

a. Linear time--Example: DC general (13)

b. Model:

$$C(PEA|Season, Year) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

606

where X_1, X_2, X_3 are seasonal indicators

X_4 is:

Year	X_4
1970	0
71	1
72	2
73	3
74	4

c. OLS estimates

$$\hat{Y} = 8.2 + .4X_1 + .4X_2 - .2X_3 + .01X_4 \qquad \bar{R}^2 = .49$$

d. Other functional forms possible: quadratic, logarithmic, etc.

607

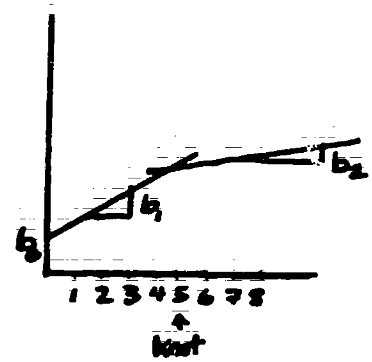
Topic 4: Combining shifts in intercept and slope: Splines

I. We can construct variables and data to handle combinations or special cases. Here function is continuous but derivative is discontinuous.

1. Two linear time trends: intersection known, slope unknown $b_0 + b_1X_1 + b_2X_2$ (14)

X_1	X_2
-4	0
-3	0
-2	0
-1	0
0	0
0	1
0	2
0	3
.	.
.	.
.	.

X_1	X_2
0	0
1	0
2	0
3	0
4	0
5	0
5	1
5	2
.	.
.	.
.	.



2. Two linear trends: intersection unknown (15)
slopes unknown

Need third indicator variable to handle intersection

$b_0 + b_1X_1 + b_2X_2 + b_3X_3$

Obs	X_1	X_2	X_3
	Data Structure		
1	1	0	0
2	2	0	0
3	3	0	0
4	4	0	0
5	5	0	1
6	5	1	1
7	5	2	1
8	5	3	1
9	5	4	1

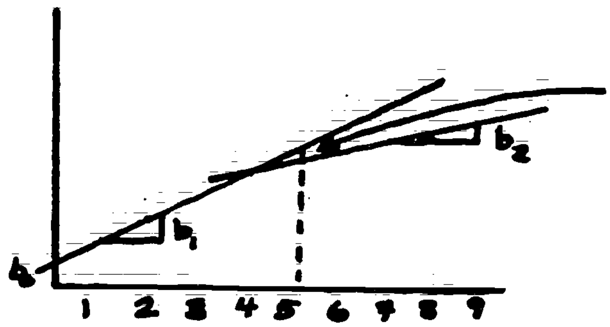
b_0 = intercept of line #1

b_1 = slope of line #1

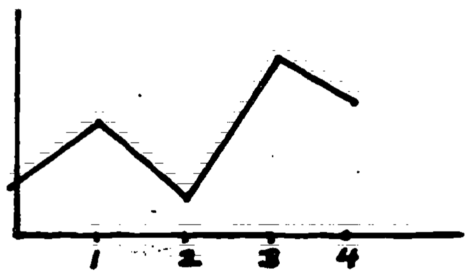
b_2 = slope of line #2

b_3 = vertical distance between line #2 and #1 at fifth observation point

608



3. Multiple peaks (16)



<u>Data Structure</u>

$X_1 = 1$ $X_2 = 2$ $X_3 = 3$ $X_4 = 4$

X	X_1	X_2	X_3	X_4	Z_1	Z_2	Z_3	Z_4
0	1	2	3	4	0	0	0	0
1	1	2	3	4	1	0	0	0
2	1	2	3	4	2	1	0	0
3	1	2	3	4	3	2	1	0
4	1	2	3	4	4	3	2	1

Model

$$\hat{y} = b_0 + b_1Z_1 + b_2Z_2 + b_3Z_3 + b_4Z_4$$

where

$$Z_1 = X$$

$$Z_2 = \max(X-X_1,\ 0)$$

$$Z_3 = \max(X-X_2,\ 0)$$

$$Z_4 = \max(X-X_3,\ 0)$$

609

Then b_1 is slope over first segment. Other b_i represent change in slope from preceeding segment.

I.e., slope for $X_1 < X < X_2$ is $b_1 + b_2$

$X_2 < X < X_3$ is $b_1 + b_2 + b_3$ etc.

Lecture 4-3
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 4-3 Outline
Topic 2		
Section A		
1.b	2	Conceptual Model
1.b	3	Scatterplot of life expectancies
2.b	4	Model of life expectancies for nations
2.c	5	Scatterplot of life expectancies
Section B		
1.b	6	Combining Indicator and continuous variables
2	7	Scatterplot of life expectancies vs per capita income
2.b	8	Scatterplot of life expectancies vs log (per capita income)
3.d	9	Geometrical representation
Section C		
2.c		Smoothed Seasonal effects
3.b		More than 1 indicator
3.f		Another structure
Topic 3		
Section A		
4.a		D.C. General Hospital conceptual model
Topic 4		
Section A		
1.		Linear Spline--Intersection known
2.		Linear Spline--Intersection unknown
3.		Multiple peaks

Lecture 4-3

Indicator Variables: Using Least Squares procedures to estimate alternative functional forms for the conditional typical summary. Part II

Lecture Content:

1.) Constructing variables and data sets for indicator variables.
2.) Interpreting models containing indicator variables.

Main Topics:

1.) Introduction to indicator variables.
2.) Simple 0/1 indicator variables.
3.) Linear and other functional forms for indicators.
4.) Splines - Shifts in intercept and slope.

2.

Conceptual Model:

$E(LE \mid \text{National Status}) = b_0 + b_1 X_1$

where: $X_1 = \begin{cases} 1 & \text{if nation is industrial.} \\ 0 & \text{otherwise.} \end{cases}$

Data Matrix

	Nation	X_1	Y_1
Industrial	1	1	Y_1
	2	1	Y_2
	3	1	Y_3
	⋮	⋮	⋮
	19	1	Y_{19}
Non-Industrial	20	0	Y_{20}
	21	0	Y_{21}
	⋮	⋮	⋮
	99	0	Y_{99}

Interpreting b_i:

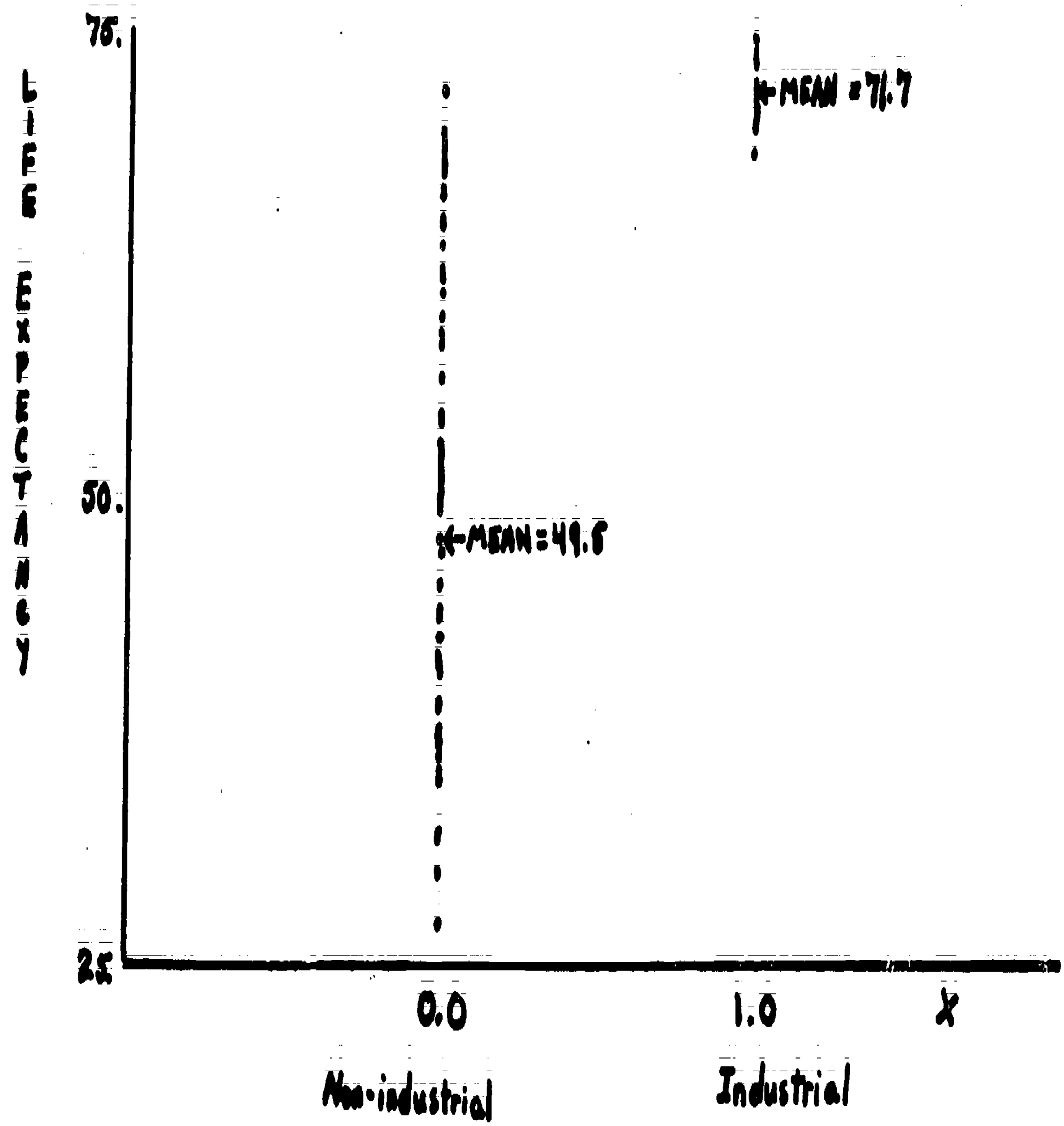
$\hat{Y}_i = b_0$ when nation i is nonindustrial.

$\hat{Y}_i = b_0 + b_1$ when nation i is industrial.

613

4-3

Scatterplot



(4)

Life Expectancy By National Status

Conceptual Model:

$$E(LE/\text{National Status}) = b_0 + b_1x_1 + b_2x_2$$

Variable Definitions

x_1	x_2	National Status
0	0	Petroleum Exporting
0	1	Non-Industrial
1	0	Industrial

Interpreting b_i:

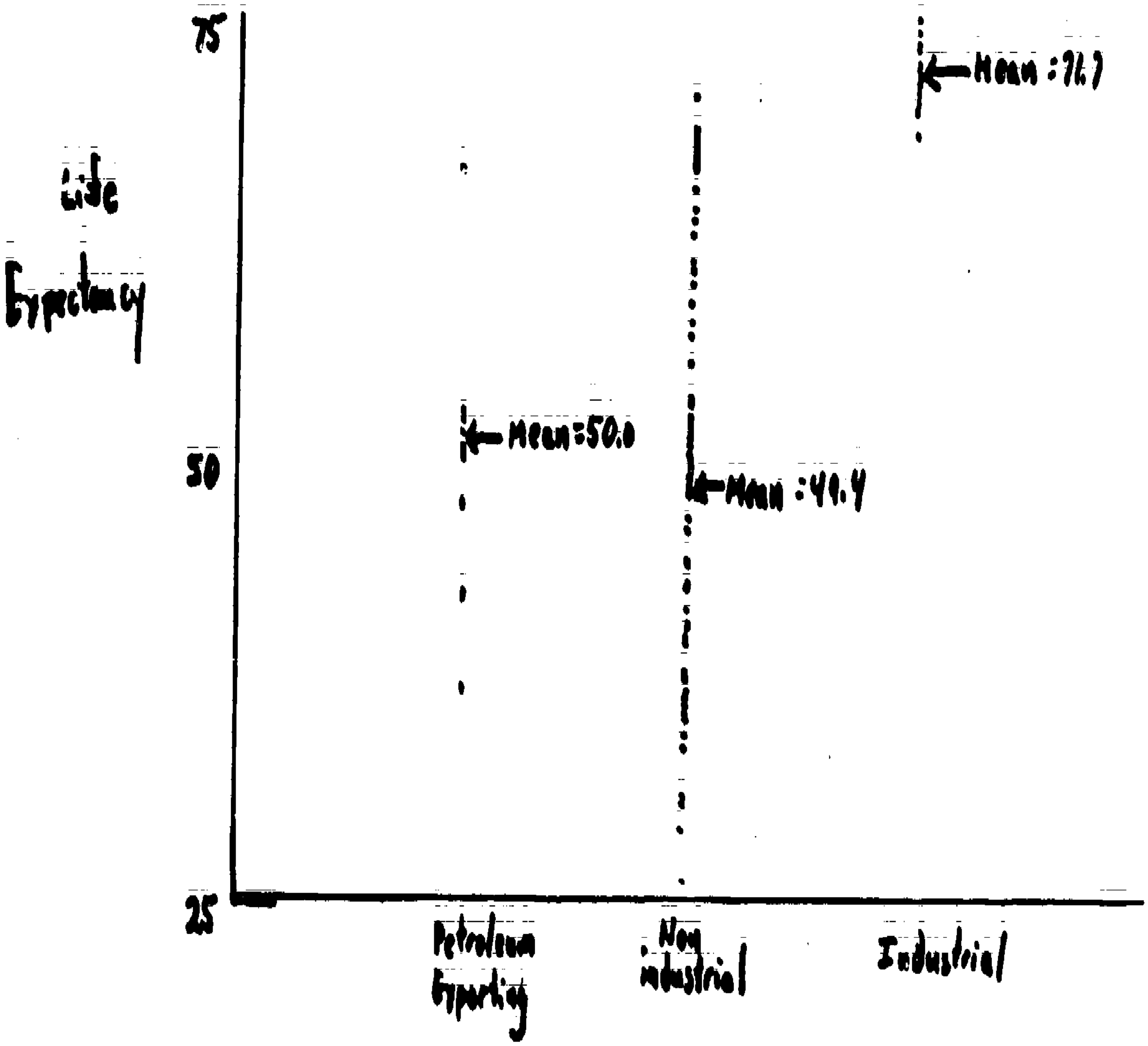
$\hat{Y}_i = b_0$ when nation i is petroleum exporting.
$= b_0 + b_1$ when nation i is industrial.
$= b_0 + b_2$ when nation i is non-industrial

Data Matrix:

observation	x_0	x_1	x_2	y	
1	1	0	0	Y_1	Petroleum Exporting
2	1	0	0	Y_2	
⋮	⋮	⋮	⋮	⋮	
⋮	1	0	1	$Y.$	Nonindustrial
⋮	1	0	1	$Y.$	
⋮	⋮	⋮	⋮	⋮	
⋮	1	1	0	$Y.$	Industrial
⋮	1	1	0	$Y.$	
99	1	1	0	$Y.$	

4-3

[5]


75
50
25
Life
Expectancy
Mean=50.0
Mean :49.4
Mean :71.7
Petroleum
Exporting
Non
industrial
Industrial

617

618

[6]

Combining Indicator and Continuous Variables

Example: Life Expectancy by per capita income and nation's industrial status

Conceptual model:

$$C(LE \mid \text{Industrial Status}) = \hat{y} = b_0 + b_1 X_1 + b_2 X_2$$

where:

X_1 is per capita income

$$X_2 = \begin{cases} 1 & \text{if nation is industrial,} \\ 0 & \text{otherwise} \end{cases}$$

Interpretation of b_i:

$\hat{y}_i = b_0 + b_1 X_{i1}$ if nation i is not industrial

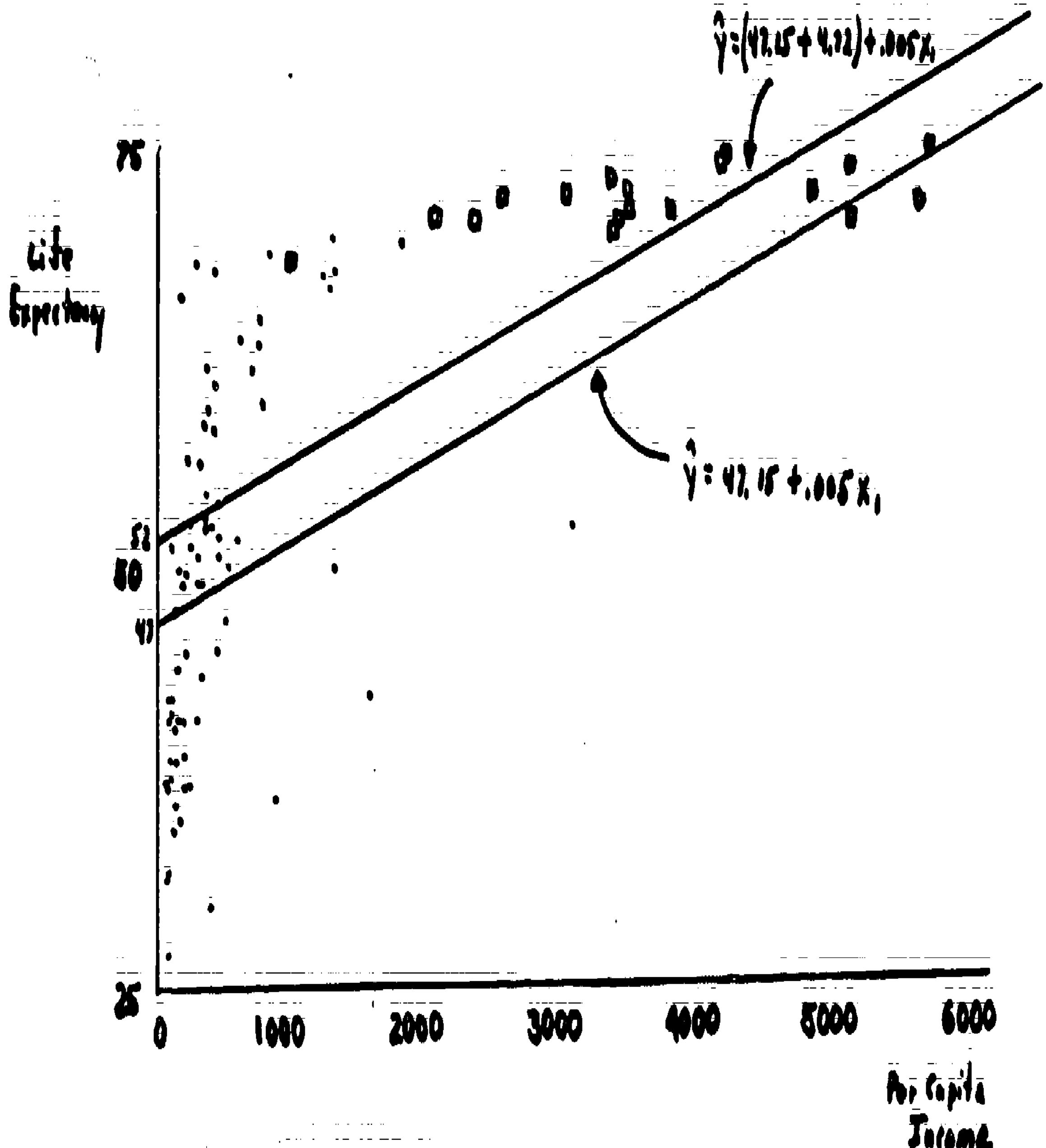
$\hat{y}_i = (b_0 + b_2) + b_1 X_{i1}$ if nation i is industrial

Thus, b_2 is shift in life expectancy conditional on being industrial.

OLS estimates:

$$\hat{y} = 47.15 + .005 X_1 + 4.72 X_2 \qquad R^2 = .52$$

4-3

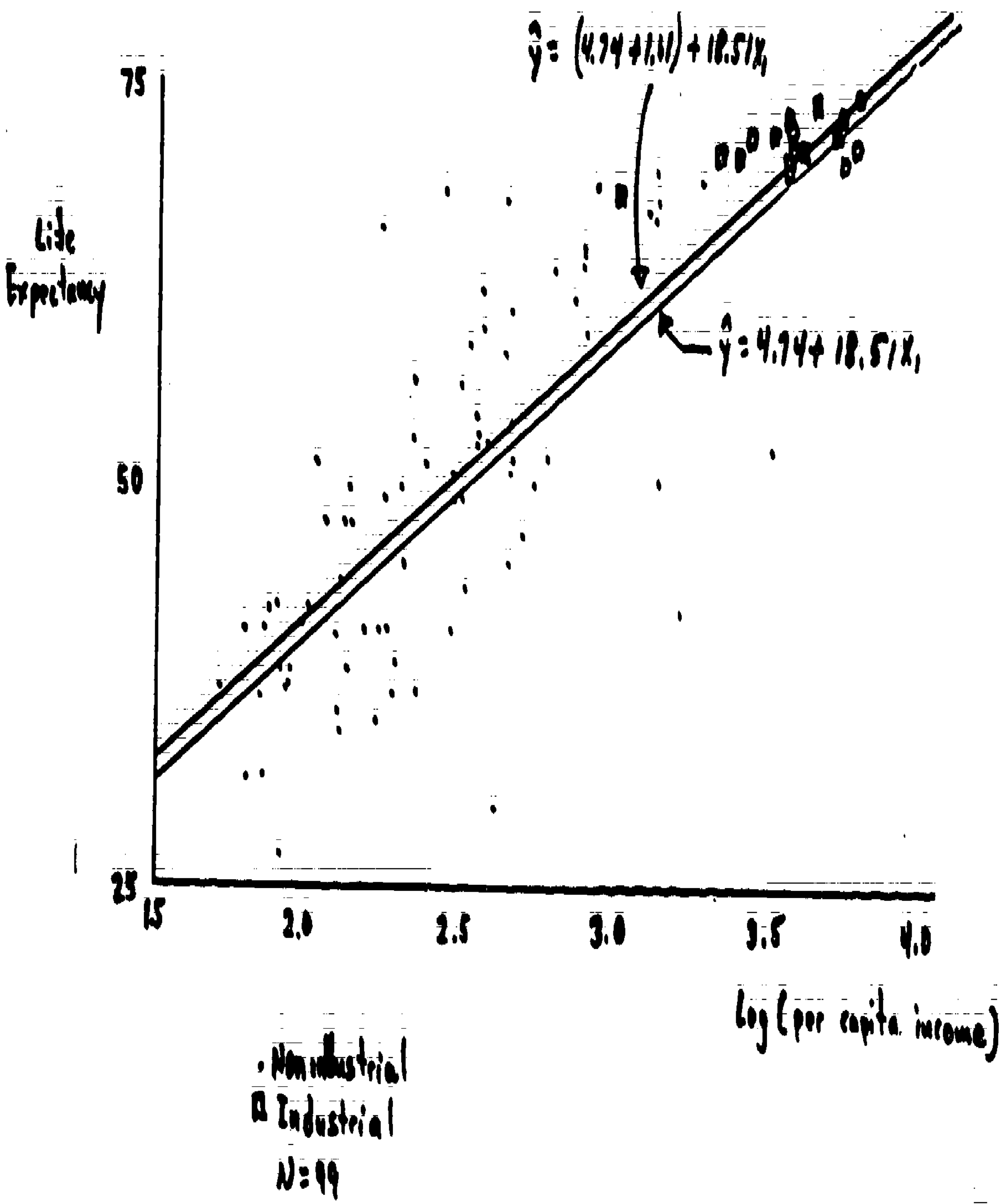


· NONINDUSTRIAL

O INDUSTRIAL

N = 99

QMPM



XVI.II.270

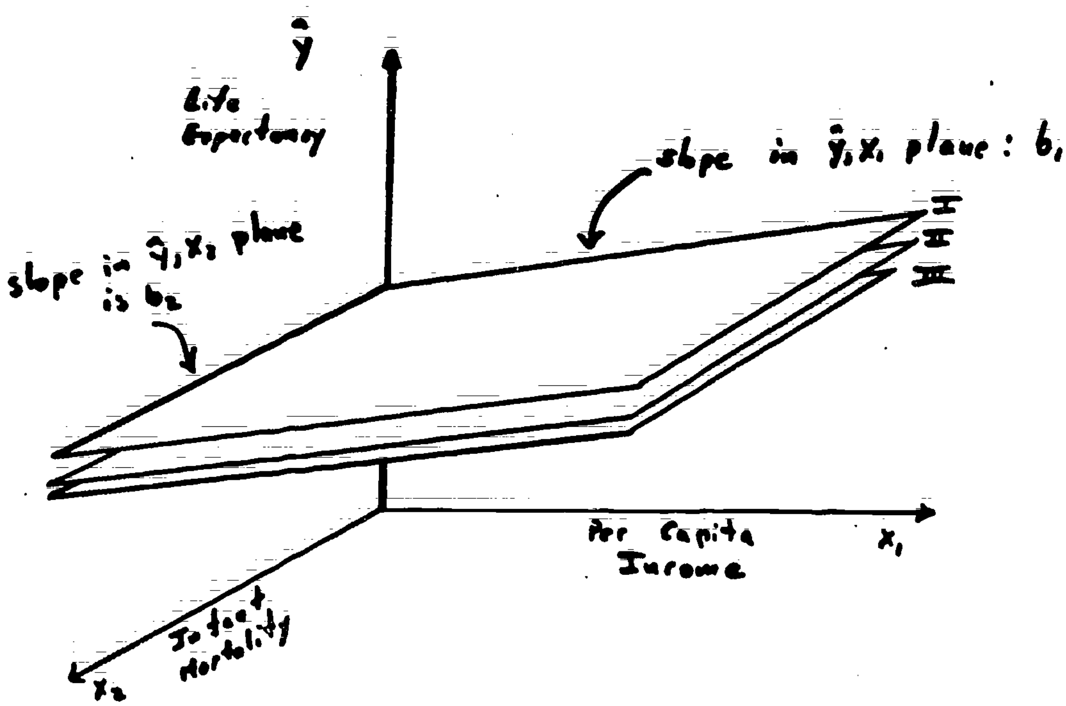
[8]

[9]

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4$$

$$\hat{y} = 54.74 + .005 X_1 - .059 X_2 - 1.77 X_3 - 1.36 X_4$$

$$R^2 = .65$$



Plane I : $\hat{y} = b_0 + b_1 X_1 + b_2 X_2$ Petroleum Exporting
Plane II : $\hat{y} = (b_0 + b_4) + b_1 X_1 + b_2 X_2$ Non-Industrial
Plane III : $\hat{y} = (b_0 + b_3) + b_1 X_1 + b_2 X_2$ Industrial

《OLS estimates indicate that b_3 and $b_4 < 0$》

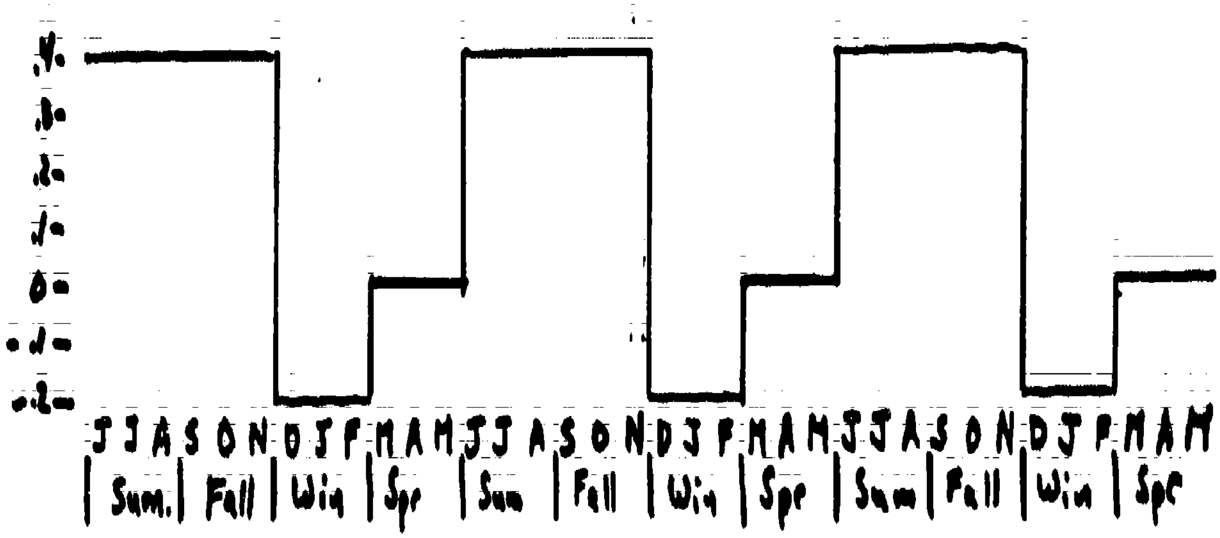
4-3

XVI.II.272

D.C. General Hospital: % Emergency Admits (Smoothed Data)

$b_0 = 8.2$

$X_1 = 1$ iff summer
$X_2 = 1$ iff fall
$X_3 = 1$ iff winter



$$\hat{y} = 8.2 + .4X_1 + .4X_2 - .2X_3$$

$R^2 = .50$

626

[07]

[11]

More than 1 indicator

Income by race, sex

Conceptual model: $E(INC \mid Race, sex) = b_0 + b_1 X_1 + b_2 X_2$

where X_1 is race indicator

X_2 is sex indicator

Variables:

Sex (X_2)	Race (X_1): B 0	W 1
M 0	MB 00	MW 01
F 1	FB 10	FW 11

Interpreting b_i:

b_0 is typical level for male black.

$b_0 + b_1$ is typical level for male white.

$b_0 + b_2$ is typical level for female black.

$b_0 + b_1 + b_2$ is typical level for female white.

(other variables may be added to model.)

9-3

Another structure for race, sex indicators.

Assuming no additivity of effects.

X_1	X_2	X_3	
0	0	0	male black
1	0	0	male white
0	1	0	female black
0	0	1	female white

Model:

$$\widehat{Inc} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$$

Interpreting b_i:

b_0 is typical level of male black
$b_0 + b_1$ is typical level of male white
$b_0 + b_2$ is typical level of female black
$b_0 + b_3$ is typical level of female white

628

[13]

D.C. General Hospital

Conceptual Model:

$$E(\text{PEA}\,|\,\text{Season, year}) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

Where:

X_1, X_2, X_3 are season indicators

and X_4 is linear time (year)

i.e.,

Year	X_4
1970	0
1971	1
1972	2
1973	3
1974	4

Results:

$$\hat{y} = 8.2 + .4X_1 + .9X_2 - .2X_3 + .01X_4$$

$$R^2 = .49$$

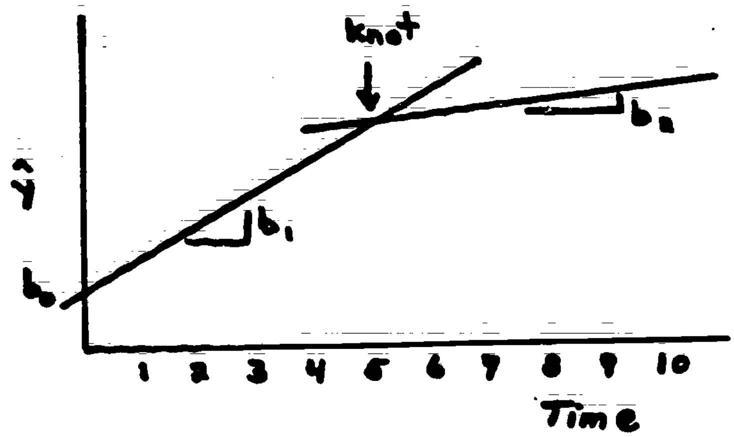
4-3

[14]

Linear Spline Intersection Known

Two Linear Time Trends

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$



X_1	X_2
1	0
2	0
3	0
4	0
5	0 ← Knot
5	1
5	2
5	3
5	4
⋮	⋮

<or>

X_1	X_2
-4	0
-3	0
-2	0
-1	0
Knot → 0	0
0	1
0	2
0	3
0	4
⋮	⋮

Linear Spline:

Intersection and slopes unknown

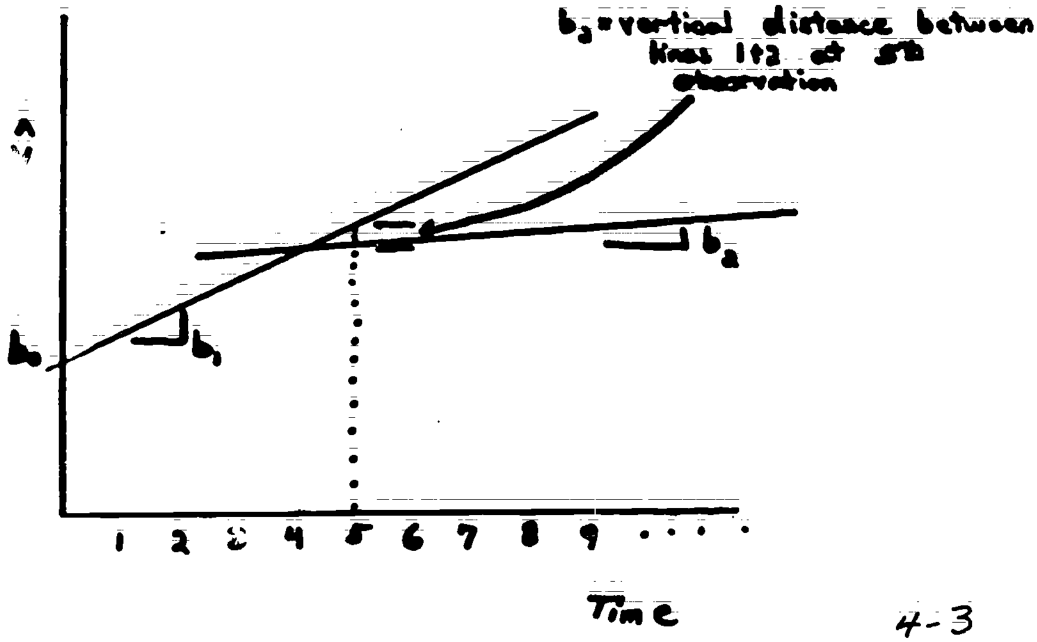
Two time trends

Data Structure

Obs.	X_1	X_2	X_3
1	1	0	0
2	2	0	0
3	3	0	0
4	4	0	0
5	5	0	1
6	5	1	1
7	5	2	1
8	5	3	1
⋮	⋮	⋮	⋮

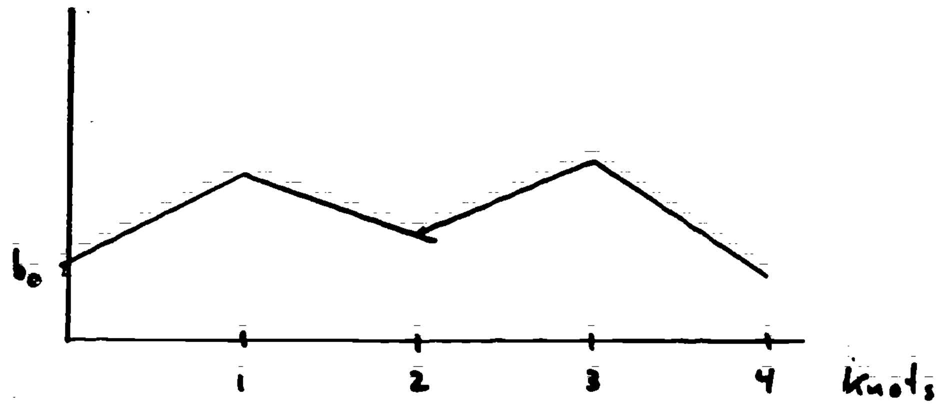
$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$$

Linear Spline Scatterplot



Multiple peaks

[16]



Data Structure

$X_1 = 1$, $X_2 = 2$, $X_3 = 3$, $X_4 = 4$

X	X_1	X_2	X_3	X_4	z_1	z_2	z_3	z_4
0	1	2	3	4	0	0	0	0
1	1	2	3	4	1	0	0	0
2	1	2	3	4	2	1	0	0
3	1	2	3	4	3	2	1	0
4	1	2	3	4	4	3	2	1

Model: $\hat{y} = b_0 + b_1 z_1 + b_2 z_2 + b_3 z_3 + b_4 z_4$

where:

$z_1 = x$

$z_2 = \max(x - X_1, 0)$

$z_3 = \max(x - X_2, 0)$

$z_4 = \max(x - X_3, 0)$

Then b_1 is slope over first segment, other b_i gives change in slope over last segment

I.e., slope for $X_1 < x < X_2$ is $b_1 + b_2$
for $X_2 < x < X_3$ is $b_1 + b_2 + b_3$, etc.

V-3

Lecture 4-4. Inference about Least Squares Coefficients

Inference about Least Squares Coefficients: A sampling experiment involving a univariate regression to study the random nature of coefficient estimates (1)

Lecture Content:

1. Discuss the sampling experiment and its purpose of studying variability

2. Assumptions made in a multivariate linear regression model

Main Topics:

1. The sampling experiment and analysis of coefficient estimates

2. Regression model with "well-behaved" data

3. Variance of least squares coefficient estimates

Topic 1. The sampling experiment and analysis of coefficient estimates

I. Basic Issue: Regression coefficient estimates depend entirely on the $(\underset{\sim}{X}_i, \underset{\sim}{y}_i)$ observations

1. We calculate the vector of coefficient estimates, $\underset{\sim}{b}$, using only the (Nxp) $\underset{\sim}{X}$ matrix and dependent variable $\underset{\sim}{y}$

2. If we add or delete observations, or use a different set of N observations, then the estimates will most certainly differ

3. Hence coefficient estimates of the model parameters depend on the "sample" of N observations

II. Problem: How do the estimates of the b_i's vary with the chosen sample (2)

1. Consider the following example:

   a. X_i = number of surgical procedures for patient i

      y_i = length of stay (LOS) for patient i, in days

   b. Data from a hospital on i = 1, 2, ..., 2435 patients

2. Note that patients have between 0 and 6 surgeries, so that we have a natural classification into batches

3. We will take small "samples" from the large data set, fit least squares lines, and study how a and b vary over the samples

III. Solution: A sampling experiment

1. Number of surgeries and length of stay is more linear in log(y) scale (3)

   a. Increase in log(LOS) as NSURG increases

   b. Plot shows number of patients with j surgeries, j = 0, 1, ..., 6.

2. We study the 7 batches of data, looking at box plots of log(LOS) with NSURG fixed (4)

   a. Each batch seems symmetric about a modal value

   b. Interesting to note that average log(LOS) is less for patients with 1 surgery than for patients with 0 surgeries

3. Parallel schematic plots show equal spreads, general increase in log(LOS)

4. Since there are only 39 patients with 4, 5, or 6 surgeries, we combine these three batches. This schematic shows linear trend better than earlier plot (5)

5. Number summaries computed. Note how $\bar{X} = M$ indicating symmetry, and how $\sigma \doteq .75$, reasonably constant (6)

6. Plot of conditional typical values shows trend (7)

7. We select a sample of 2142 patients from the 2435 to use as a sampling base (8)

   a. Note the scatterplot of 2142 patients

   b. Slope of LS line is .157, slightly less than .21 for entire data set

   c. Intercept is similar in both regressions, 2.1

8. We now draw 100 samples, with varying sample size ($\doteq$ 25) where the percentages of patients with k surgeries is same in each sample as in the entire data set

   a. Stem-and-Leaf display of the 100 estimates of a and b (9)

      i. Note how the intercepts are symmetric about $\doteq$ 2.0 or 2.1

      ii. Slope intercepts appear very well behaved

   b. Number summaries of the estimates show "well-behavedness" (10)

      i. Intercepts symmetric about 2.1, with $\sigma = .239$

      ii. Slopes symmetric about .12, slightly less than the .157 for the sample base, although $\sigma = .228$ shows that .12 is not too small

9. How do the LS estimates compare with resistant line estimates? (11)

   a. Data have large spread, so that resistant line estimates may be slightly closer to b = .157

   b. Intercept estimates have larger spread than with LS, more outliers

c. Slope estimates also have more spread, although mode ≈ .15 is more apparent

d. Number summaries show that b's have mean of .155, very close to the "correct" value (12)

IV. Conclusion: Sampling shows how estimates of regression parameters vary

1. When the observations used in the model do <u>not</u> include all the observations, i.e. when "sample size" ≠ "population size", estimates will vary around the "true values"

2. Hence, there is a degree of randomness inherent in the estimates

3. Estimates appear quite well-behaved

4. LS estimates not quite as accurate as Resistant Line estimates, although spread is certainly less

Topic 2. Regression model with "well-behaved" data

I. Basic Issue: Definition of well-behaved regression data

1. Since Data = Fit + Residual, the definition of well-behaved regression data begins with the residuals

2. Compute Residual = Data - Fit = $y_i - \hat{y}_i$ and examine as a batch

II. Problem: What is a well-behaved batch of residuals?

1. First of all, the linear model must "fit", so that R^2 is large (near 1), and residuals are small

2. Residuals should be:

   a. Homoscedastic--variance of residuals should remain constant as $\underset{\sim}{X}$ increases--easy to envision in 2 dimensions

   b. A well behaved single batch--symmetric about 0, with 95% between 2 and -2

   c. Devoid of <u>all</u> patterns

3. Let σ^2 = Variance of Residuals. σ^2 estimated by

$$\frac{1}{N-P} \sum_{i=1}^{N} (y_i - y_i)^2$$

4. [illegible] the standard error of the residuals ($\sqrt{\sigma^2}$ for the [illegible]00 regressions (13)

   a. Mean of the σ's = .785

   b. σ for entire sample base = .777 so that σ's are roughly symmetric about the "correct" value

5. With well-behaved univariate data, resistant line estimates should equal LS estimates

Topic 3. Variance of least squares coefficient estimates

I. Basic Issue: Theoretical form of variance of least squares estimates

1. Now that we know that the batch of residuals has variance σ^2, how do the parameter estimates vary?

2. Let $\underset{\sim}{b}_{LS}$ be the vector of least squares coefficient estimates

3. $\underset{\sim}{b}_{LS}$ is a "random" vector, varying about $\underset{\sim}{b}$, the "true" regression coeffients

4. Var $(\underset{\sim}{b}_{LS})$ depends on $\underset{\sim}{X}$ and σ^2

II. Problem: Interpretation of $\mathrm{Var}(\underset{\sim}{b}_{LS})$

1. Definition:

$$\mathrm{Var}(\underset{\sim}{b}_{LS}) = \sigma^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$$

2. Diagonal terms of $\mathrm{Var}(b_{LS})$ are variances of individual b_i

3. Off diagonal terms, (i,j), are covariances of b_i and b_j

4. Transparency shows output from LS regression of log(LOS) on NSURG for entire sampling base (14)

   a. Note $\sigma^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$

   b. Various other quantities, R^2, σ = standard error, "t" will be discussed in detail in next lecture

638

Lecture 4-4
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 4-4 Outline
Topic 1 Section II		
1.	2	[illegible] ...ot of Length of Stay vs Number [illegible] ...ical Procedures
Section III		
1.	3	[illegible] ...ot of Least Squares Line
3.	4	Sc...atic plots of Log Length of Stay
4.	5	Schematic plots of Log Length of Stay, 4, 5, 6 surgeries combined
5.	6	Number Summaries of Log Length of Stay
6.	7	Plot of conditional typicals
7.	8	Scatterplot of Sample Data
8.a	9	Stem-and-Leaf of Regression Coefficients from Least Squares
8.b	10	Number Summaries of Regression Coefficients, Least Squares
9.	11	Stem-and-Leaf of Regression Coefficients from Resistant Line
9.d	12	Number Summaries of Regression Coefficients, Resistant Line
Topic 2 Section II		
4.	13	Standard Errors of the Residuals
Topic 3 Section II		
4.	14	Variation of Coefficients

639

Lecture 4-4

Inference about Least Squares Coefficients:
A sampling experiment to study the random nature of coefficient estimates.

Lecture Content:

1.) Discussion of the sampling experiment and the study of variability.

2) Assumptions made in a multivariate linear regression model.

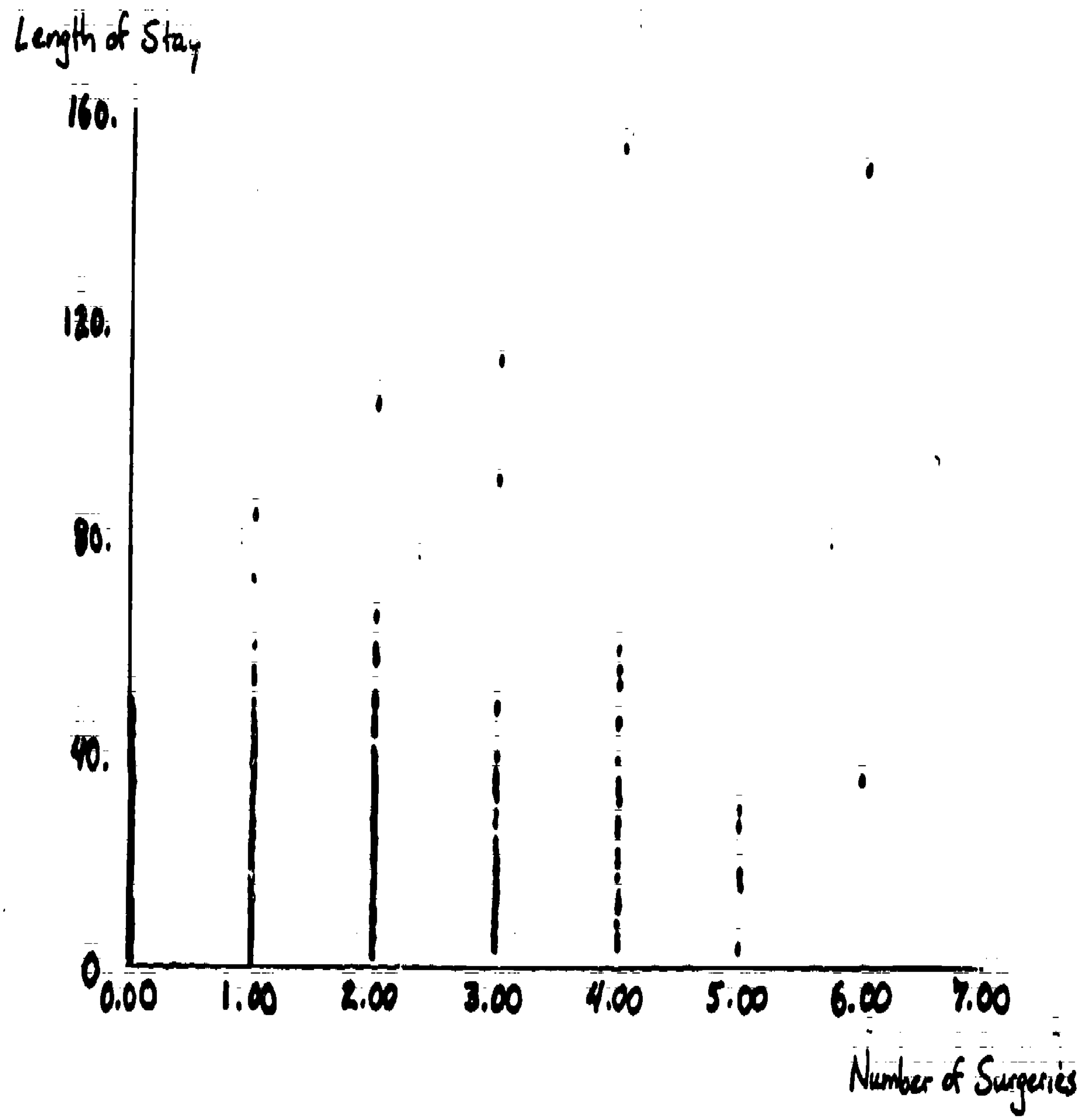
Main Topics:

1.) The sampling experiment and analysis of coefficient estimates.

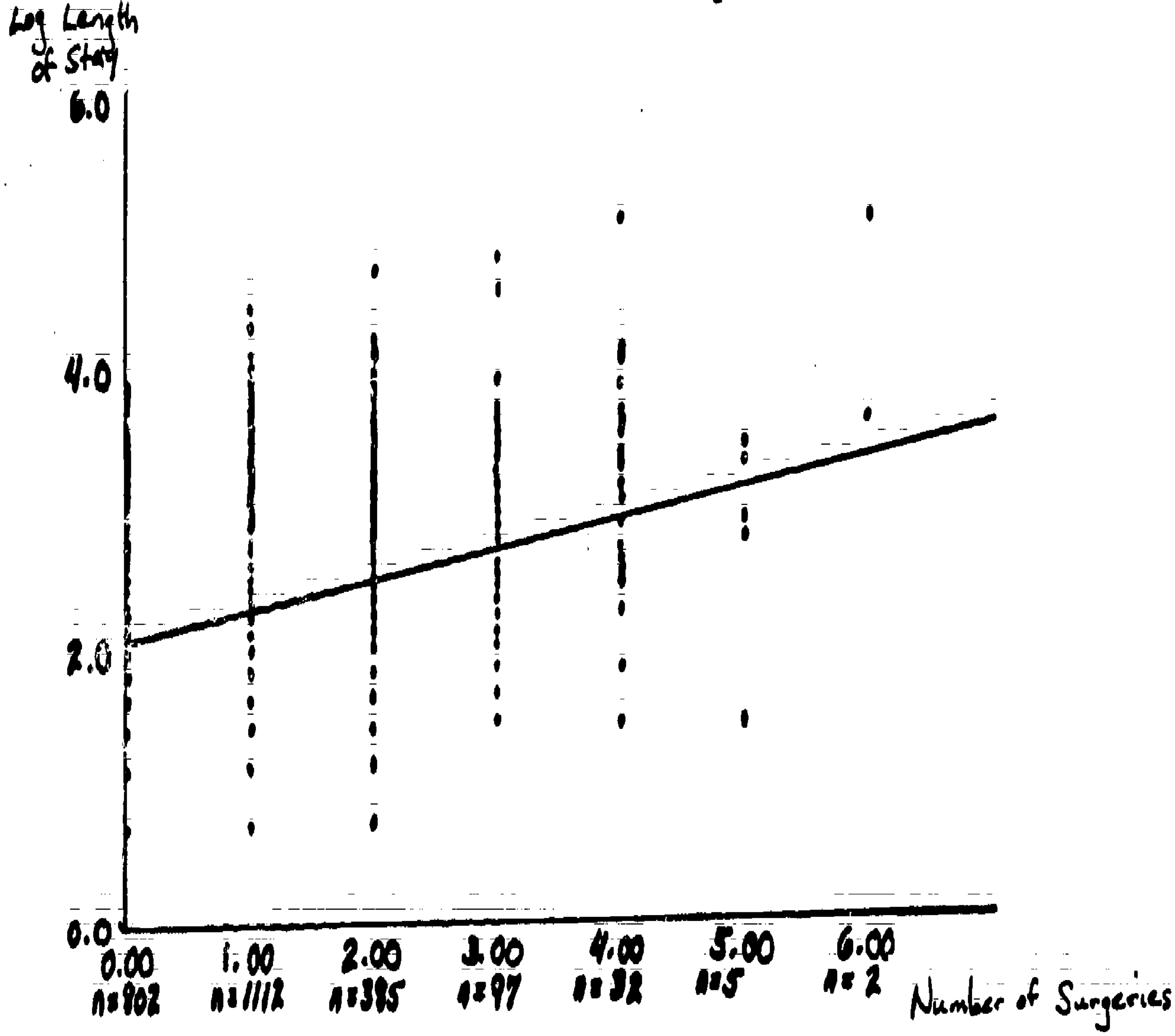
2) Regression model with "well-behaved data

3.) Variance of least squares coefficient estimates.

Scatterplot of Length of Stay in days vs. Number of Surgical Procedures for Hospital Data [2]



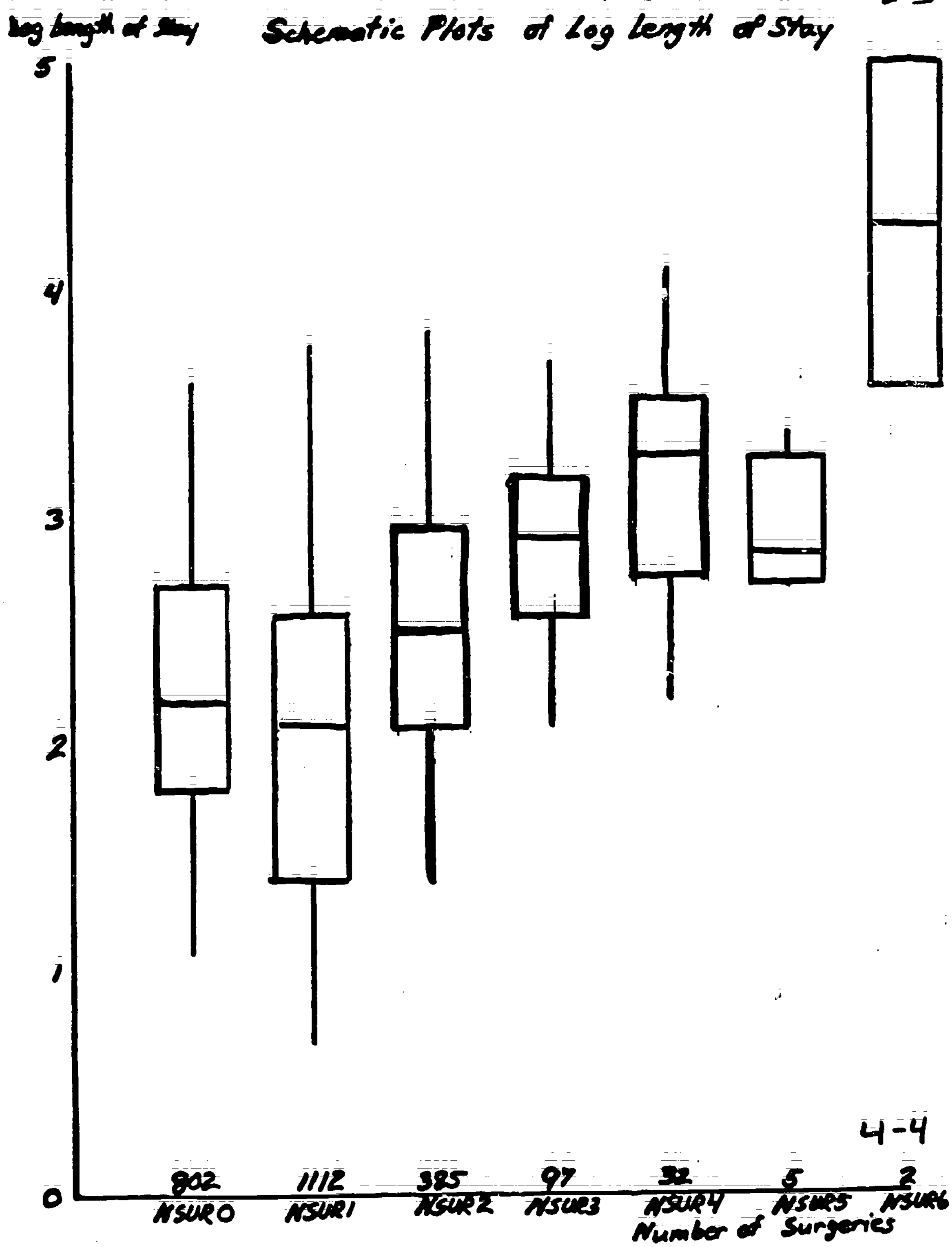
Scatterplot and Least Squares Line of Log Length of Stay vs. number of Surgical Procedures for Hospital Data
Observations in each "batch" given.



Least Squares Line is: Y = 0.210X+2.01

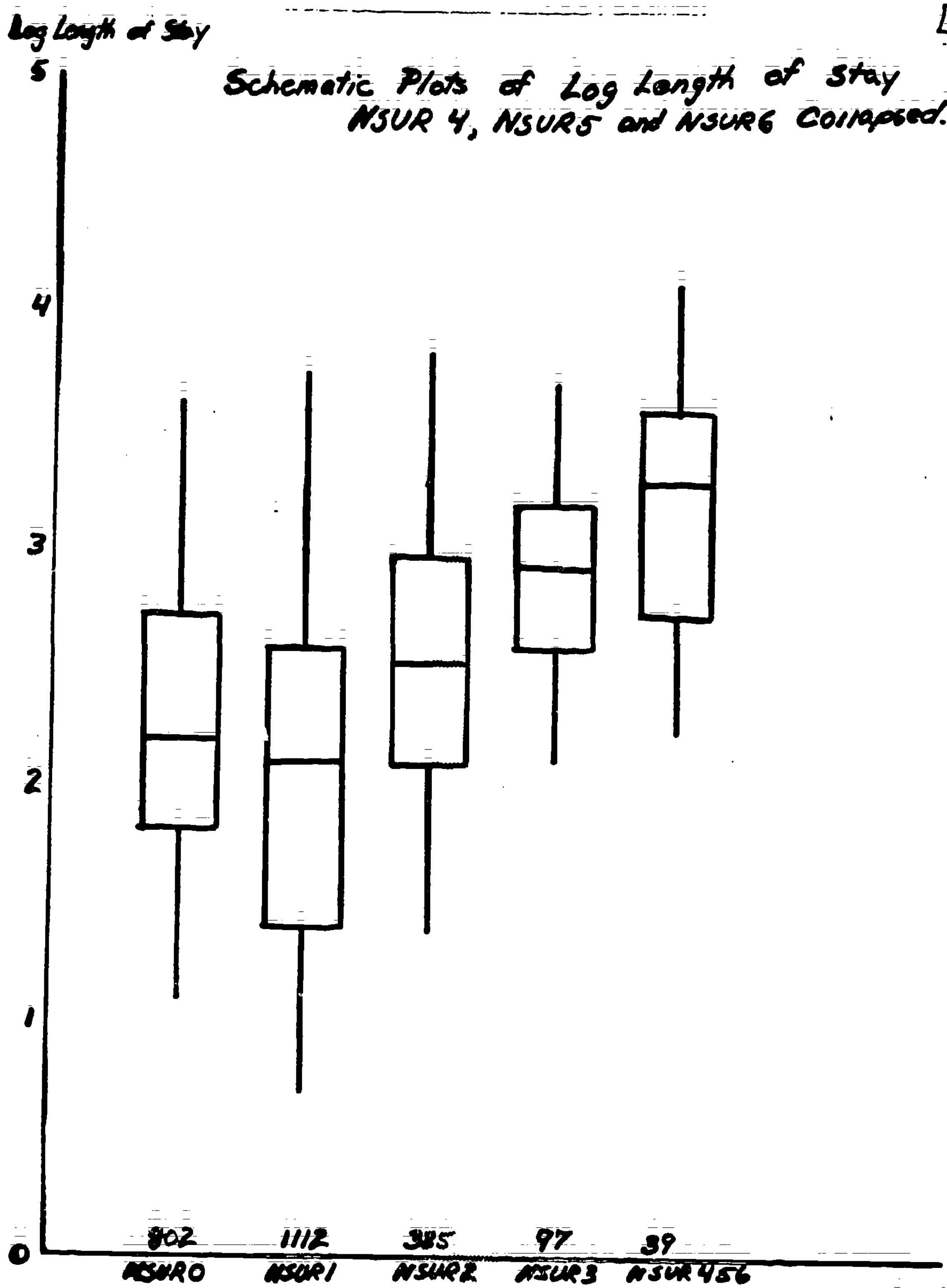
XVI.II.288

[4]



Log length of Stay
Schematic Plots of Log Length of Stay
5
4
3
2
1
0
802
NSUR0
1112
NSUR1
385
NSUR2
97
NSUR3
32
NSUR4
5
NSUR5
2
NSUR6
Number of Surgeries
4-4

[5]


Log Length of Stay
Schematic Plots of Log Length of Stay
NSUR 4, NSUR5 and NSUR6 Collapsed.
5
4
3
2
1
0
802
1112
385
97
39
NSUR0
NSUR1
NSUR2
NSUR3
NSUR456

4-4

Number Summaries for Log Length of Stay - Data Classified Into Batches by Number of Surgical Procedures

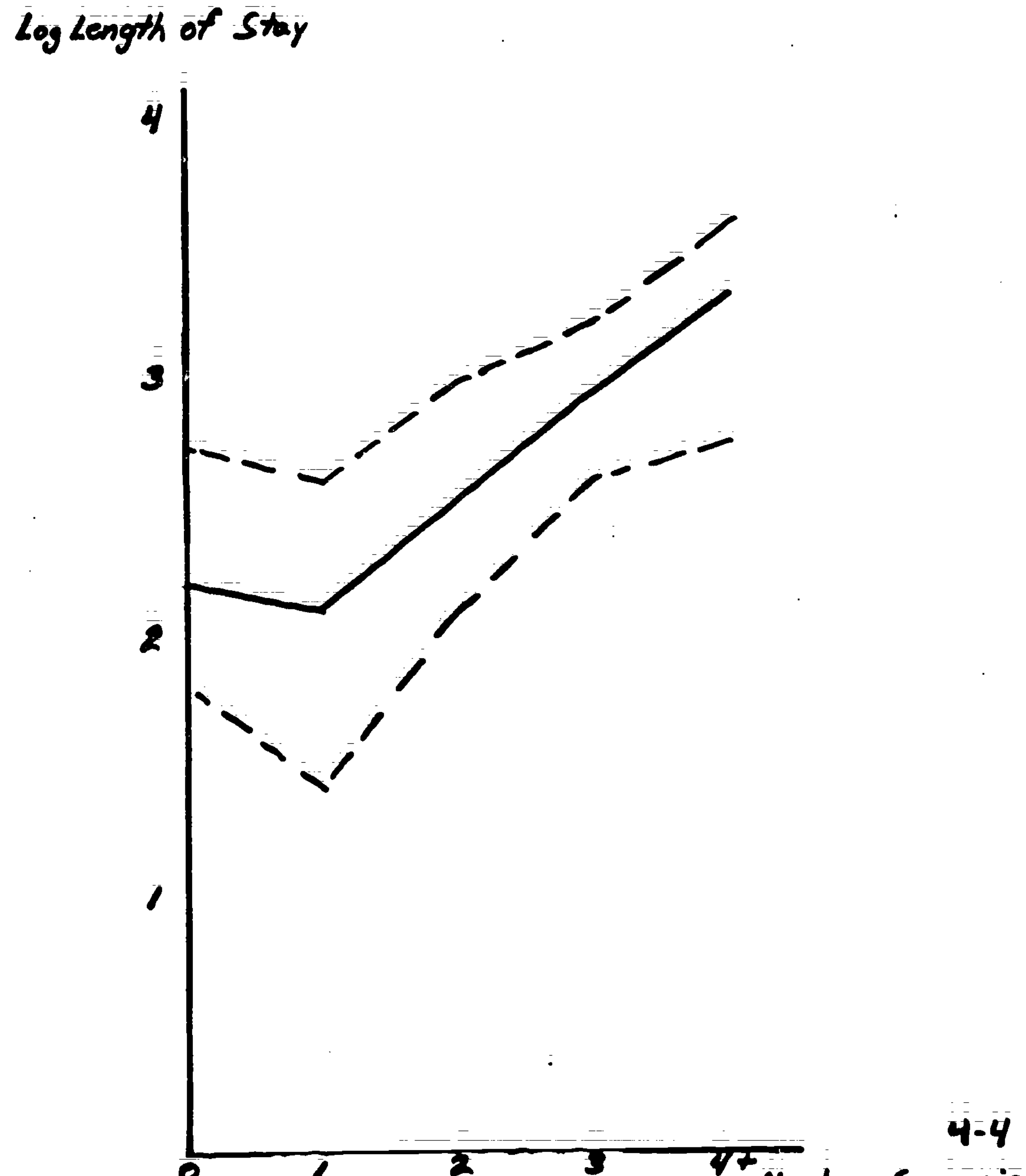
[6]

	NSUR 0	NSUR 1	NSUR 2	NSUR 3	NSUR 4	NSUR 5
MIN	0	0	0.693	1.386	1.386	1.386
LH	1.792	1.386	2.079	2.565	2.736	2.708
MED	2.197	2.079	2.485	2.890	3.277	2.833
UH	2.708	2.565	2.944	3.178	3.526	3.258
MAX	3.871	4.431	4.654	4.787	5.037	3.367
$\bar{X}$	2.154	2.048	2.522	2.865	3.179	2.711
σ	0.813	0.780	0.637	0.874	0.707	0.791
M	2.197	2.079	2.485	2.890	3.277	2.833
ΔH	0.916	1.179	0.865	0.613	0.790	0.550
N	802	1112	385	97	32	5

Module II

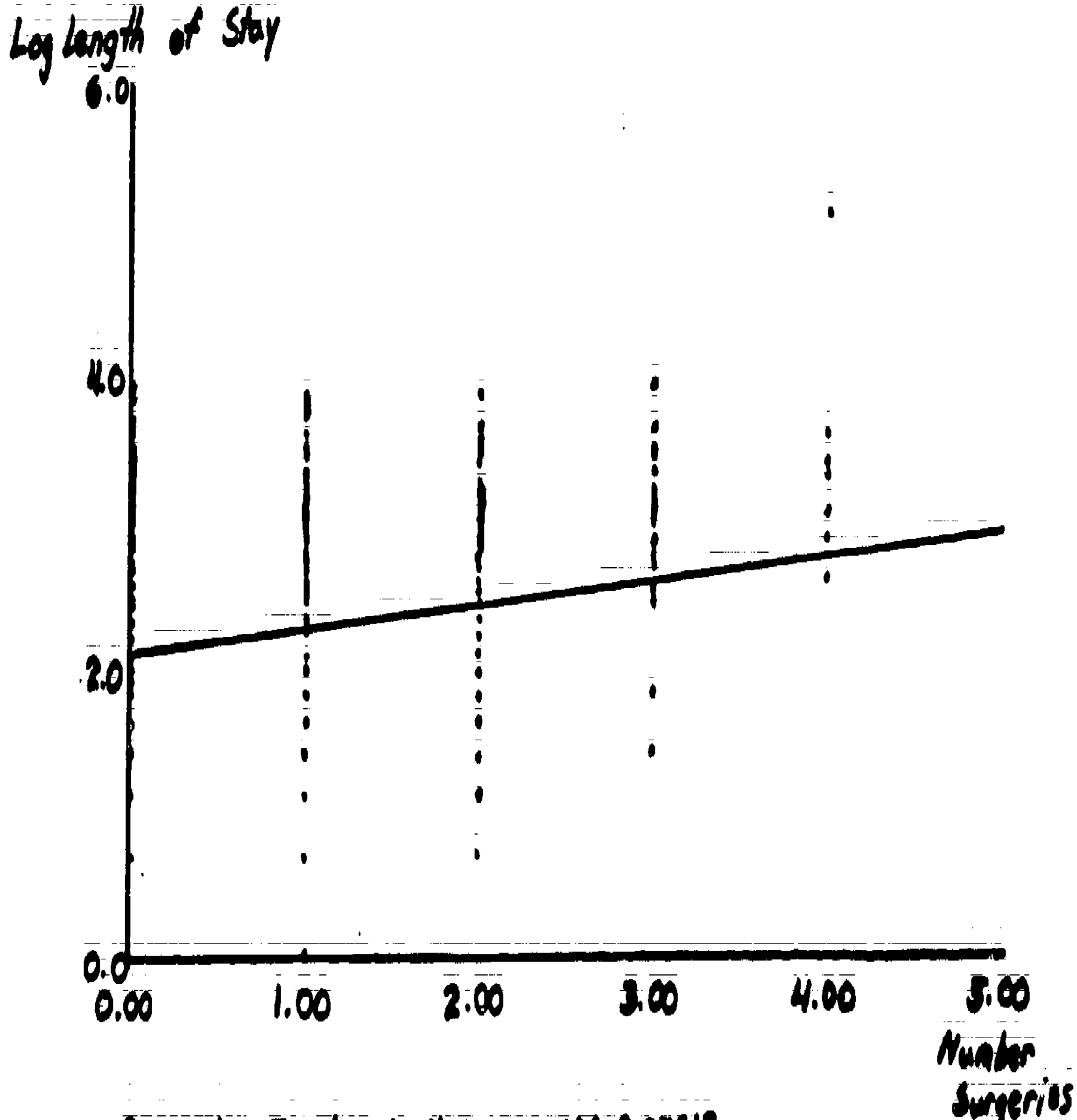
XVI.II.291

Plot of Conditional Typical Values and Hinges [7]



4-4

Scatterplot of sample of Hospital Data, N = 2142



Regression Equation is Y=0.157348X+2.07312

Stem-and-Leaf Displays of Regression Coefficients from the 100 Samples. Least Squares Estimates.

a

Unit = 10^{-2}

```
LO   1.31171 ,  1.45958

15 | 7
16 | 5 9 9
17 | 3 3 9
18 | 0 2 2 3 3 5 6 6 7 7 8
19 | 0 1 1 3 3 4 4 4 5 5 5 7 8 8 9 9 9
20 | 0 1 1 1 3 5 6 6 7 8 9 9 9 9
21 | 2 3 3 4 4 4 5 6 6 7 8 8 9
22 | 0 1 2 2 2 3 4 5 5 5 6 6 6 6 7 7 8 9
23 | 2 2 4 4 5 6 9 9 9
24 | 0 3 4 4 7 9 9 9
25 | 6
```

b

Unit = 10^{-2}

```
LO   -.528729 ,  -.522186

-3 | 8 5 2
-2 | 9 7 1
-1 | 5 3 3 2 2 2 1 0
-0 | 8 7 5 5 4 3 3 2 2 2 2 1 1 0 0
 0 | 1 1 4 5 5 6 6 7 7 8 8 8 8 8 9 9 9
 1 | 1 2 2 3 3 4 4 4 5 5 6 6 8 8 8 8 9
 2 | 0 1 1 2 2 5 5 6 8 8 8 9
 3 | 0 0 1 1 2 2 3 4 5 6 8 9
 4 | 0 0 1 3 4 5 5 6
 5 | 4 7
 6 | 5
```

[10]

Number Summaries of Regression Coefficients for 100 Samples. Least Squares Estimates

a

NOB	100	
MEAN		2.0914
MED		2.09629
TRI		2.09837
MID		2.10184
MIN		1.31171
LH		1.93786
MED		2.09629
UH		2.[illegible]106
MAX		2.[illegible]958

SD	0.238836
MIDSP	0.3212
RANGE	1.25787

b

NOB	100	
MEAN		0.117006
MED		0.124183
TRI		0.127072
MID		0.122965
MIN		-0.528729
LH		-0.024043
MED		0.124183
UH		0.283964
MAX		0.656161

SD	0.227827
MIDSP	0.308007
RANGE	1.18489

4-4

Stem-and-Leaf Displays of Regression Coefficients for 100 Samples. Resistant Line Estimates. [11]

UNIT = 10^{-2}

LO: 0.462098 0.527964 0.562532 0.84042 1.0536 1.0911 1.10288 1.2666 1.28512

```
14 | 3
15 | 7
16 | 6
17 | 1 2 8
18 | 6 7 8 8
19 | 0 1 3 4 5 6 8 8 9 9
20 | 0 1 4 5 5 9 9
21 | 1 1 3 3 3 5 6 7 7 8
22 | 0 0 1 3 3 3 3 4 6 6 7
23 | 0 2 4 4 4 6 6 6 9 9
24 | 0 0 0 1 2 2 2 2 4 4 7 8 8 8
25 | 0 1 2 3 4 5 5 5 6 7 9
26 | 8
27 | 1 9
28 | 1 2 7
29 | 1 8
```

UNIT = 10^{-2}

LO: -0.980829

```
-5 | 5 3
-4 |
-3 | 7 5 5
-2 | 9 8 7 5 2
-1 | 7 7 2 1 1 0 0
-0 | 8 7 7 6 5 4 4 3 1
 0 | 0 0 2 2 2 3 3 3 3 3 4 6 6 6 6 8
 1 | 0 0 1 1 1 2 2 3 3 4 4 5 5 5 5 5 6 7 7 9 9
 2 | 0 1 1 2 2 4 5 5 5
 3 | 0 0 2 3 4 6 7 8 9
 4 | 1 2 3 5 6 8 9
 5 | 9 9
 6 | 9 9
 7 | 1 5 5
```

HI: 0.981055 1.05562 1.09861 1.30644

Regression Coefficients for 100 Samples: Resistant Line Estimates.

NOB	100			
MEAN		2.13985	SD	0.493916
MED		2.23161	MIDSP	0.461728
TRI		2.21772	RANGE	2.52708
MID		2.22661		
MIN		0.462098		
LH		1.97298		
MED		2.23161		
UH		2.4347		
MAX		2.98918		

NOB	100			
MEAN		0.154819	SD	0.340975
MED		0.130556	MIDSP	0.352622
TRI		0.133794	RANGE	2.28726
MID		0.12779		
MIN		-0.980829		
LH		-0.03928		
MED		0.130556		
UH		0.313342		
MAX		1.30644		

Standard Error of the Residuals, for each of the 100 fitted lines

Stem-and-Leaf Displays

unit = 10^{-2}, values cut

1	5 **	4
4	5 ·	6 7 9
13	6 **	3 4 2 1 0 2 4 3 3
[illegible]	6 ·	9 6 6 8 7 6 8 9 8 5 6 6 6 8
[illegible]	7 **	1 1 0 1 2 3 4 4 0 1 3 5
(13)	7 ·	9 6 7 8 7 5 7 9 9 5 9 5 5
[illegible]	8 **	2 1 2 4 2 3 4 0 0 0 4 1 1 4 2 1 0 3 2 1
[illegible]	8 ·	9 8 5 8 6 5 6 5 9 9 8 5 5
13	9 **	3 0 3 1 0 0 0 3 0
4	9 ·	9 5 5
1	10 **	4

5 Number Summary

50½	M	.785	
26	H	.680	.850
1	E	.590	1.040

For sample of size 2142, $\sigma = .777$

Variation of Coefficients [14]

$\sigma^2 = .6042$
$\sigma = .7773$

$$(X^t X)^{-1} = \begin{pmatrix} .0009947 & -.0005892 \\ -.0005892 & .0006587 \end{pmatrix}$$

$$(X^t X) = \begin{pmatrix} 2142 & 1917 \\ 1917 & 3235 \end{pmatrix}$$

$$\sigma^2 (X^t X)^{-1} = \begin{pmatrix} .000601 & -.000356 \\ -.000356 & .000398 \end{pmatrix}$$

Variance - Covariance Matrix of the Coefficients

$a = 2.073 \qquad \sigma_a = \sqrt{.000601} = .0245$

$b = 0.157 \qquad \sigma_b = \sqrt{.000398} = .0199$

$-.000356 = \text{Cov}(a, b)$
covariance of a and b

Lecture 4-5 Model with Least Squares Estimates in Well-Behaved Batches

Model with Least Squares Estimates in Well-Behaved batches: Evaluating the model (1)

Lecture Content:

1. Discuss assumptions of linear model and the optimality of the estimates

2. Testing how well the model "fits"

Main Topics:

1. Covariances and variances of variables

2. Evaluating the model

Topic 1. Covariances and Variances of Variables

I. Basic Issue: How do we measure how "related" a set of variables are

1. Develop "pair-wise" measures of relations. There are $\binom{p}{2}$ measures for a set of p variables

2. The measure comparing X_i and X_j tells how similar these 2 variables are, independently of all other variables

II. Problem: What is the best measure?

1. Seek a dimensionless quantity--no units

2. Measure should have a maximum and a minimum value to aid us in our assessments

III. Solution--Covariances, Variances, and Correlations

1. Covariances, in (units of X_i) x (units of X_j), tell by how much X_{ik} and X_{jk} simultaneously vary from their means, for all $k = 1, 2, \ldots, N$

2. Variances, in units of X_i^2, tell by how much the observations of X_i differ from $\overline{X}_i$, in squared deviations

3. Correlations--our desired measure of relationships--are ratios of covariances and variances

IV. Method

1. Definition of Covariances and Var.

$$\text{Cov}(X_i, X_j) = \frac{1}{N} \sum_{k=1}^{N} (X_{ik} - \overline{X}_i)(X_{jk} - \overline{X}_j)$$

$$\text{Var}(X_i) = \frac{1}{N} \sum_{k=1}^{N} (X_{ik} - \overline{X}_i)^2$$

2. The (pxp) matrix of covariances (off-diag.) and variances (diagonal) is called $\Sigma = (\sigma_{ij})$, the "Variance-Covariance" matrix

3. Correlations

$$r_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}} = \frac{\Sigma(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)}{\sqrt{\Sigma(X_{1i} - \overline{X}_1)^2 (X_{2i} - \overline{X}_2)^2}}$$

$$-1 \leq r_{ij} \leq 1$$

4. R, the matrix which has ones along the diagonal and r_{ij} as off-diagonal elements, is called the diagonal matrix.
5. Correlation of -1 or 1 $\Rightarrow$

   X_i and X_j are linearly related

6. Correlation of 0

   if X_i and X_j are well-behaved, $r_{ij} = 0 \Rightarrow X_i$ and X_j are "independent", i.e., unrelated

Topic 2. Evaluating the Model

I. Basic Issue: When does least squares produce "good" estimates

1. Remember that LS is just one technique--there are others--for estimation

2. LS works well when the data adhere to several assumptions

II. Assumptions necessary for LS

1. Model: $\hat{y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_p X_p$ (2)

2. Residuals: $y_i - \hat{y}_i$

3. Four assumptions:

   a. y_i must be linearly related to the n independent variables

   b. Batch of residuals must be well behaved

   c. Residuals must be independent of X_i

   d. Variance of batch of residuals is a constant σ^2
   These assumptions must be true

4. If (a)-(d) are true, then by the Gauss-Markov Theory, LS estimates are optimal in the sense that the quantity $\Sigma(y_i - \hat{y}_i)^2$ is minimized i.e., estimates have minimum variance

5. One implication of these assumptions is that the coefficients are "well-behaved", i.e., if we could fit each model n times, with n different samples, the batch of b_i values would be well-behaved--unbiased and consistent

6. Note how well-behaved the standardized b_1 coefficients from the sampling experiment are (3)

III. Problem: How do we determine whether these assumptions hold?

1. We have various measures at our disposal:

   a. R^2 multiple correlation coefficient

   b. t-statistics

   c. Thorough examination of residuals

2. We discuss a and b here, leaving c until next lecture

IV. Method

1. R^2 = square of the multiple correlation coefficient (4)
2. R = correlation between Y and the linear combination of X's which maximizes R; the combination is the regression equation
3. t-statistic--determines whether a variables belongs in the model

$$t_i = \frac{b_i}{\sigma_{b_i}}$$

4. If $|t_i| > 3$, variable X_i is important; if t_i near 0, variable X_i can be ignored
5. R^2 and t statistics for the 100 samples (5) (6)
6. [illegible]ical regression output (7)

662

Lecture 4-5
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 4-5 Outline
Topic 2 Section II		
1.	2	Least Squares Linear Model Assumptions
6.	3	Standardized b_1 coefficients from sampling experiment
Section IV 1.	4	t-statistics and R^2
5.	5	R^2 values from sampling experiment
5.	6	statistics from sampling experiment
6.	7	Regression of Length of Stay on number surgeries

663

Lecture 4-5

Linear Model and Least Squares in Well-Behaved batches of data.

Lecture Content:

1) Assumptions of the linear model and optimality of the estimates.

2) Testing the "fit" of the model.

Main Topics:

1.) Covariances and Variances

2.) Evaluating how well the model fits.

[3]

Least Squares Linear Model Assumptions

MODEL:

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_n X_{in} \; ; \; i = 1, 2, \ldots, m$$

$$r_i = Y_i - \hat{Y}_i \quad \text{residuals}$$

1) Dependent Variable must be linearly related to the n independent variables.

2) Batch of residuals must be well behaved.

3) Residuals must be independent.

4) Variance of batch of resid..ls is a ..nstant σ^2

IF assumptions 1-4 are true, ~~then~~ the Least Squares coefficient estimates are optimal.

665

4-5

Standardized b_1 Coefficients From Sampling Experiment

2	LO	-2.834, -2.806
4	-2***	1 0
6	-1·	9 8
7	s	7
8	f	4
8	t	
15	-1***	1 0 0 0 0 0 0
18	-0·	9 8 8
27	s	7 7 6 6 6 6 6 6 6
33	f	5 5 5 5 4 4
38	t	3 2 2 2 2
49	-0***	1 1 1 1 1 1 1 0 0 0 0
(9)	0***	0 0 0 0 1 1 1 1 1
42	t	2 2 2 2 2 2 3 3
34	f	4 4 4 4
30	s	6 6 6 7 7 7 7
23	0·	8 8 8 8 8 9 9
16	1***	0 0 0 1
12	t	2 2 2 3 3
7	f	4 4 4 5
3	s	
3	1·	8
2	2***	0
1	t	3

Number Summary

M	50½	.031	
H	26	-.619	.733
Eighths	13	-1.00	1.10
E	1	-2.834	2.366

$\bar{X}$	0
M	.031
Mid H	.057
Mid E	-.234

6% > 1.96 in abs. val.

4-5

[4]

t statistics

$$t_i = \frac{b_i}{\sigma_{b_i}} = \frac{\text{coefficient estimates}}{\text{std. err of coefficient estimates}}$$

A large positive or small negative value of t_i (>3, <-3) indicates that the ith variable, X_i, is important in the regression.

Multiple Correlation Coefficient R

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \text{Percent "explained" by the regression}$$

R is the correlation between the independent variable Y and a linear combination of X variables. Which linear combination? That combination which maximizes R!

$$R^2 = \left[\max_b \operatorname{Corr}[Y, b_0 + b_1 X_1 + \ldots + b_n X_n]\right]^2$$

R^2 should be near unity.

[5]

R^2 Values for Sampling Experiment

UNIT = 10^{-3}, VALUES CUT

23	0**	0000000000 11 2222 333 4444
31	0·	5566 7888
45	1**	0112455666 7788
49	2	1239
(4)	3	0156
47	4	189
44	5**	239
41	6	134
38	7	457
35	8	1445789
28	9	6
27	10**	0
26	11	3
25	12	1 [illegible]
23	13	[illegible] [illegible] 668
18	14	48
16	15**	16
14	16	3578
10	17	789
7	18	1
6	19	45

HI | .2049, .3035, .3?31, .3291

Number Summary

M	30½	.0305	
H	26	.006	.113
Eighths E	13	.002	.165
E	1	0	.3291

R^2 for entire data base = .0283

668 4-5

[7]

Least Squares Regression of Log(los) on Nsurg. Sample of 2142 patients

NOB = 2142 NOVAR = 2 DF = 2
Multiple R squared = 0.02827
σ (SER = RMS of Residual) = 0.777263

	Coefficient	St. Err Coef.
Constant	2.07346	0.02451
NSURG	0.15734	0.01994

Variance - Covariance Matrix

$$\begin{pmatrix} 0.000601 & -0.000356 \\ -0.000356 & 0.000398 \end{pmatrix} = \sigma^2(x'x)^{-1}$$

t - statistics = coef./SE Coef.

Constant = 2.07346/.02451 = 84.596

NSURG = .15734/.01994 = 7.8907

4-5

Lecture 4-6. Evaluating the Model

Evaluating the Model: A thorough examination of the Residuals to determine how well the model fits

Lecture Content:

1. Analysis of Residuals
2. Analysis of Length of Stay of patients in a hospital

Main Topics:

1. Looking for patterns in the residuals via scatterplots
2. Applying our inferential procedures to an example

(There are no transparencies.)

Topic 1. Analysis of Residuals

I. Basic Issue: How can we use the residuals as a batch to find violations of assumptions

1. We have measures for overall assessment

   a. R^2: Coefficient of Determination

      i. Function of the Squared Residuals

      ii. Tells what fraction of the total variation is "explained" by the fitted line

      iii. Comprehensive measure of fit

   b. t statistics: Ratio of estimate to its standard error

      i. Indirect function of Residuals depends on σ^2 = Var Residuals

      ii. Tells whether a given regression coefficient is non-zero

      iii. If $-3 < t < 3$, coefficient is essentially 0, i.e., the variable does not help to "explain" y

      iv. Note: to increase t statistics, one usually decreases S.E. of b

      $$S.E.(b_i) = (\sigma^2 (\underset{\sim}{X}^t \underset{\sim}{X})^{-1}{}_{ii})^{1/2}; \text{ (i,i) element}$$

         A. We can either decrease σ^2

         B. Or increase $(X^t \underset{\sim}{X}) \rightarrow$ spread X's out over wider range

2. These measures are "gross" in that an assessment of violations of assumptions is not allowed

3. We need to examine residuals further to determine whether they:

   a. Are well behaved

   b. Are homoscedastic

   c. Are independent

672

II. Problem: How should we examine them

1. We can only conclude that either

   a. The LS assumptions appear to be violated in some specified way

   b. The LS assumptions do not appear to be violated

2. Note that b. does not mean the assumptions are correct-- It means that given the data that we have seen, we have no reason to conclude that they are violated

3. We examine the residuals graphically

4. Scatterplots are easy to make, and quite revealing

III. Solution: Principal ways of plotting residuals

1. Stem-and-leaf

2. In time sequence (if relevant)

3. Against the fitted (conditional typical) values $\hat{y}$

4. Against the independent variables

5. Any other sensible ways

IV. Methods

1. Stem-and-Leaf

   a. Stem-and-leaf the residuals $r_i = y_i - \hat{y}_i$ as a single batch

   b. Display should resemble a well-behaved batch

   c. If not, the well-behavedness assumption is violated, and one should determine exactly how by

      i. Looking at $\overline{X}$ and σ^2

      ii. Examining outliers carefully

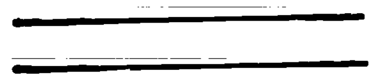
2. Time sequence plots

   a. If data are time series, i.e. gathered over time, we should plot the residuals (y) vs corresponding value on the time scale (X)

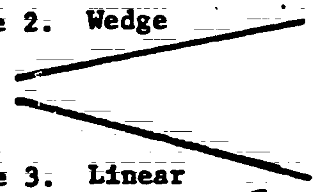
673

b. There are 5 characteristic shapes to these plots:

Shape 1. Random--Desired pattern



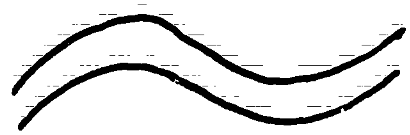
Shape 2. Wedge



Shape 3. Linear

Shape 4. Quadratic

Shape 5. Trignometric-- Sine patterns



c. These shapes imply the following

i. Shape 2. Heteroscedasticity
Use weighted LS (next lecture)

ii. Shape 3. Linear time term needed in model

iii. Shape 4. Quadratic time term <u>and</u> linear time term needed

iv. Shape 5. Tough luck--Seasonalities
Try indicator variables
Residuals are <u>not</u> independent

3. Plots against the fitted values

a. Shape 2. Heteroscedasticity
Transform y

b. Shape 3. Error in Analysis
Model Incorrect--did you leave out b_0?

c. Shape 4. Model Inaccurate
Need additional (square or interaction) terms

d. Shape 5. ? Residuals not independent

4. Plots against the Independent Variables
1 per X_i

a. Shape 2. Heteroscedasticity
Transform y or X_i

b. Shape 3. Probable error in calculations
Linear effect not removed

c. Shape 4. Need extra higher order terms in X_i

d. Shape 5. ? Residuals not independent

5. Other residual plots

a. If residuals come from different processes (1-10: Machine 1) (11-20: Machine 2) examine as separate batches

b. If considering a new independent variable, plot it against the residuals

675

Topic 2. Applying our inferential procedures to an example

I. **Example**: "The cost and Length of hospital stay" Lave & Leinhardt

Examine Tables 2, 3, 4, 5

676

Lecture 4-7 Problems with Least Squares Estimation

Problems with Least Squares Estimation: Effect of invalidated assumptions on the coefficient estimates

Lecture Content:

1. Small number of degrees of freedom
2. Robust and LAR regression
3. Ridge regression

Main Topics:

1. Overfitting--more variables than observations
2. Non-well behaved batches of residuals
3. Collinear independent variables

(There are no transparencies)

Note: This lecture covers advanced topics and should be considered optional.

677

Topic 1. Overfitting--more variables than observations

I. Basic Issue: Effects on LS estimates when N is small relative to p

1. Define "degrees of freedom" as N-p
   Desire N-p >> 0

2. Suppose N-p < 0

3. When N = p we have a "perfect fit": regression line completely describes the relationship between $\underset{\sim}{X}$ and $\underset{\sim}{y}$, residuals are 0, $R^2 = 1$.

4. If N < p, then we are in trouble. We will not be able to estimate all the coefficients, only N functions of them

5. Statistical methodology for handling this situation is underdeveloped, and quite unsatisfactory

II. Solutions

1. We can always delete variables until N is larger than p

2. Or, we can forget about fitting multiple regression models, and examine each of the p variables as a single batch

3. We then try to combine these single, separate analyses to form some impression

4. Or we can combine variables, for instance by forming interactions, so that p is reduced

678

Topic 2. Non-well behaved batches of residuals

I. Basic Issue: How do departures from "well-behavedness" affect LS estimates

1. Most common problem in least squares regression is outlying values; skewness is another frequently observed departure from well-behavedness

2. Heteroscedastic residuals best treated via transformation or Weighted Least Squares (see below)

3. Lack of independence in the errors also best treated with Weighted Least Squares

4. Outliers? We also use a "fancy" version of weighted least squares

II. Problem: How do we use weighted least squares to improve LS estimates?

1. We assign a <u>weight</u> to each observation that tells how important that observation is

2. Ordinary Least Squares assigns weights of unity to every observation; hence large outliers receive the same weight as observations which have small residuals

3. We would like to assign smaller weights to these larger outliers, so that they become less important

4. We generate a matrix W, that is diagonal, with weights lying between 0 and 1. W is (NxN), the (i,i)th diagonal element is the weight that we assign the ith residual

   a. If ith residual is small, w_{ii} near 1 : full weight

   b. If ith residual is quite large in absolute value, w_{ii} near 0 : little weight

5. How do we determine these weights?

   a. If we know what they should be, we have no problem. Merely form W, perform the WLS calculations given below, and everything will be fine

   b. If we have no idea, we use "Robust Regression" techniques, or Least Absolute Residual Regression

III. Solutions

1. Robust Regression

a. We find the matrix W in an <u>iterative</u> procedure by minimizing

$$\sum_{i=1}^{N} \rho\left(\frac{y_i - \underset{\sim}{X}_i \underset{\sim}{\beta}}{S}\right)$$

i. $\underset{\sim}{X}_i$ = ith row of $\underset{\sim}{X}$

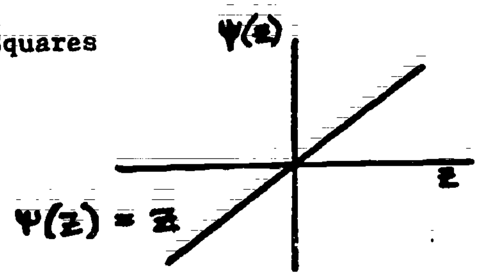
ii. S is an estimate of σ

iii. ρ' is our weighting function $\rho' = \Psi$
We describe our functions in terms of ψ, the derivative of ρ.

iv. Our weights w_{ii} are $\psi(Z)/Z$. These are the values we place on the diagonal in W

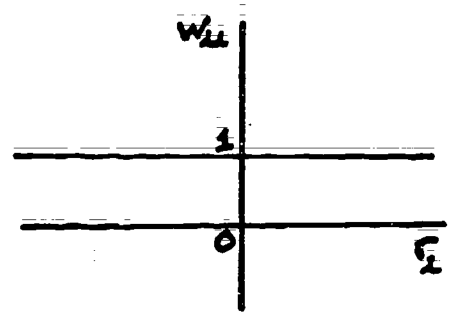
b. Examples of $\psi(\cdot)$

i. Least Squares



note how large residuals are <u>not</u> weighted downward

$w_{ii} = \dfrac{\psi(r_i)}{r_i}$, r_i = ith residual from some "initial" fit



weights of unity

a. We find the matrix W in an iterative procedure by minimizing

$$\sum_{i=1}^{N} \rho\left(\frac{y_i - \underset{\sim}{X}_i \underset{\sim}{\beta}}{S}\right)$$

i. $\underset{\sim}{X}_i$ = ith row of $\underset{\sim}{X}$

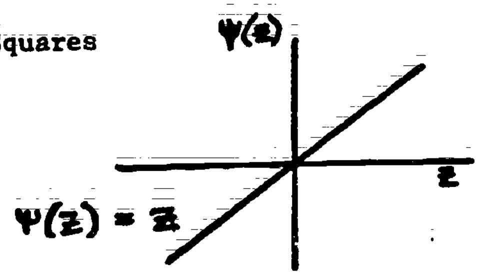
ii. S is an estimate of σ

iii. ρ' is our weighting function $\rho' = \Psi$
We describe our functions in terms of ψ, the derivative of ρ.

iv. Our weights w_{ii} are $\psi(Z)/Z$. These are the values we place on the diagonal in W

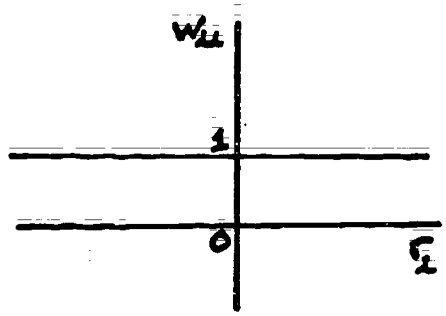
b. Examples of $\psi(\cdot)$

i. Least Squares



note how large residuals are not weighted downward

$w_{ii} = \frac{\psi(r_i)}{r_i}$, r_i = ith residual from some "initial" fit

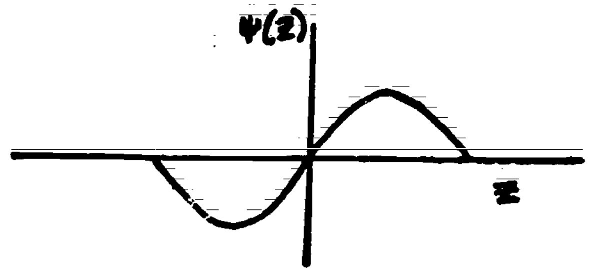


weights of unity

680

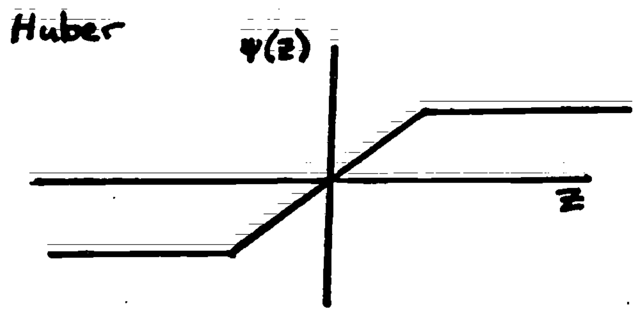
ii. Sine

$$\psi(Z) = \begin{cases} C\ \mathrm{Sin}(Z/C), & |Z| \leq \pi C \\ 0 & \text{, otherwise} \end{cases}$$



Note how large residuals get weighted downward to 0.

iii. Other functions include Huber



c. We find w_{ii} and make the W matrix and compute

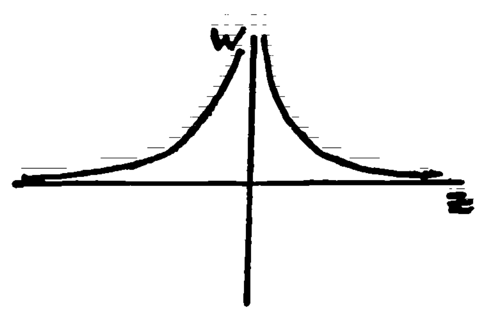
$$\underset{\sim}{\beta} = (\underset{\sim}{X}'\underset{\sim}{W}\underset{\sim}{X})^{-1}\ \underset{\sim}{X}'\underset{\sim}{W}\ \underset{\sim}{y}$$

d. We then find the residuals from this fit, weight them again, and recompute $\underset{\sim}{\beta}$. Stop when coefficients converge

2. Least Absolute Residual Regression

$\psi(Z) = \mathrm{Sign}(Z)$

$W = |Z|^{-1}$



LAR minimizes

$$\sum_{i=1}^{N} |y_i - \underset{\sim}{X}_i \underset{\sim}{\beta}|$$

absolute residuals (like finding medians)

Also an iterative procedure

3. Weighted Least Squares

a. If we can determine the covariance structure of the residuals, Σpxp, we let $W = \Sigma^{-1}$. This will get rid of any lack of independence in the residuals or any heteroscedastic tendencies

b. We then estimate

$$\underset{\sim}{b}_{WLS} = (\underset{\sim}{X}'\underset{\sim}{W}\underset{\sim}{X})^{-1} \underset{\sim}{X}'\underset{\sim}{W}\underset{\sim}{y}$$

$$Var(\underset{\sim}{b}_{WLS}) = \sigma^2(\underset{\sim}{X}'\underset{\sim}{W}\underset{\sim}{X})^{-1}$$

This is not an iterative procedure, since we assume $\underset{\sim}{\Sigma}$ known

682

Topic 3. Collinear Independent Variables

I. Basic Issue: What effects do nonindependent "independent variables" have on LS estimates

1. We should always examine $\underset{\sim}{R}$, the correlation matrix of the X's

2. If any off-diagonal element of $\underset{\sim}{R}$ is near -1 or 1, then we have problems

3. We can remedy this situation--called <u>multicollinearity</u>--only by deleting one of the 2 offending variables

4. If the situation is more complicated--many correlations >.5 or <-.5, and we cannot delete the necessary variables then we can use "Ridge Regression"

5. When there are many related X vectors, $\underset{\sim}{X}'\underset{\sim}{X}$ will be difficult to invert

II. Solution: Ridge Regression

1. Ridge regression operates directly on $\underset{\sim}{X}'\underset{\sim}{X}$, making it more "stable" and hence invertable

2. Model: $\underset{\sim}{\beta} = (\underset{\sim}{X}'\underset{\sim}{X} + kI)^{-1}\underset{\sim}{X}'\underset{\sim}{y}$

3. k lies between 0 and 1 and is added to the diagonal of $\underset{\sim}{X}'\underset{\sim}{X}$ to increase its stability

683

Homework, Unit 4

1. In the CMU-DAP library, you will find an archive named "OMAHA-Y70". This archive contains 9 variables, collected on each of the 96 census tracts in the city of Omaha, Nebraska in the 1970 census.

Let $\underset{\sim}{y}$ = MFU = Median income of families and unrelated individuals

$\underset{\sim}{X}_1$ = POP = Population of each tract

$\underset{\sim}{X}_2$ = NWHITE-PCT = Percent of population of each tract that is nonwhite

$\underset{\sim}{X}_3$ = YEARS-ED-C = Average number years of education per individual

$\underset{\sim}{X}_4$ = PCT-GT65Y = Percent of population greater than 65 years of age

(a) Plot each X against $\underset{\sim}{y}$. Comment on these 4 plots. Are the point clouds linear, or do any of the independent variables require transformation? If so, transform and plot again.

(b) Fit a multiple regression line to these data. Does the model make sense? Comment with respect to:

(i) Units of analysis of the X's
(ii) "Causality" and the underlying theory of the situation being modeled.

(c) Calculate and examine the residuals as a single batch of data. What do they tell you about your fitted line?

(d) Calculate and examine R^2. Comment.

(e) Using the linear model with any additional information gathered in (b), (c), and (d), comment on the policy implications of the model relating median family income to the several X variables.

(f) Could we have fit another line using a subset of these 4 independent variables and still have obtained a good fit? If so, what subset?

(g) If you could have <u>any</u> set of X variables (not necessarily those included in the 9 in DAP) to predict median family income, which variables would you choose and why?

2. Stored in CMU-DAP is an archive named COBBDOUGLAS with 4 variables, one observation per state.

   i) VALADD

   ii) CAPITAL

   iii) LABOR

   iv) ESTABL

These data are economic data used by many economists to "predict" the total value added (in dollars) to a subset of the economy via the Cobb Douglas production function model:

$$\left(\frac{V}{N}\right) = A \left(\frac{L}{N}\right)^{\alpha} \left(\frac{K}{N}\right)^{\beta}$$

where V = Value added (millions of \$)

L = Labor (millions of man-hours)

K = Capital Services Flow (millions of \$)

N = Number of Establishments

A, α, β = Parameters

(a) Reexpress this model in the usual multiple regression form. Plot each (transformed) independent variable against the (transformed) dependent variable. Comment on these plots.

(b) Calculate the LS regression (fit a multiple regression to the transformed data). Comment on this model with respect to:

   (i) causality
   (ii) possible dependence among the independent variables.

(c) Calculate and plot the residuals. What do they tell you about your fitted line?

(d) Calculate R^2. What does this tell you about your fitted line?

(e) What are the elasticities of value added with respect to Labor/establishment, and Capital Services/establishment, and what do they mean?

(f) Based on this model, should the new (incoming) administration concentrate more on decreasing unemployment (i.e. increasing L), or pumping money into the economy (i.e. increasing K)? (or, <u>what combination</u> of both together?).

3. In CMU-DAP, there is an archive named GAS, with the following variables, one observation per state:

i) Registered autos, buses, and trucks in 1973.
"VEHICLES"

ii) State gasoline tax per gallon, in cents, in 1973.
"GASTAX"

iii) Motor Fuel consumption, in thousands of gallons, in 1973.
"CONSUMPTION"

iv) Population, 1970 census.
"POP"

v) Population density, 1970 census, land area only.
"POPDENS"

vi) Per capita income, in 1973.
"PCINCOME"

The data show motor fuel consumption in 1973 along with 5 other, possibly related, variables.

(a) Plot each variable against fuel consumption. Comment on these plots. Do all of these variables appear related to fuel consumption?

(b) Calculate the LS regression (i.e. fit a multiple regression to the data). Does the model make sense? Comment with respect to

(i) causality and the underlying theory of the situation being modeled

(ii) possible dependence among any of the "independent" variables

(c) Calculate and plot the residuals. What do they tell you about your fitted line?

(d) Calculate R^2. What does this tell you about your fitted line?

(e) From your discussion in (b), decide upon the <u>two</u> "independent" variables which you feel yield the most "reasonable" model from a causality/dependence point of view. Repeat parts (b), (c), and (d) for this "reduced" model. Compare these results to those found for the "complete" model and discuss.

(f) Does it bother you to have some data from 1970 and some from 1973? How do you think this affects the "validity" (such as it is) of the models ("complete" and "reduced")?

(g) If you could have any other "independent" variable(s) you desire to use to "predict" fuel consumption, which one(s) would you add? Which of the 5 provided would you retain? Write your model (but do not attempt to calculate any parameters).

4. Many people dream of "beating the stock market" by being able to predict accurately its behavior (and hence being able to buy low and sell high).

Coen, Gamma, and Kandall proposed the following model to predict the London Stock market:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 X_{1,t-6} + \beta_3 X_{2,t-7}$$

where t = time (in quarters, from 1952/3 to 1967/4, t = 1 for 1952/3, etc.)

Y_t = _Financial Times_ ordinary share index at time t

X_{1t} = United Kingdom car production at time t

X_{2t} = _Financial Times_ commodity index at time t

Note that two of the "independent" variables are lagged, that is, the values at some previous time are used to predict the next value of Y. (The data are stored however as Y_t, $X_{1,t-6}$, and $X_{2,t-7}$).

The data are stored in DAP under the archive STCKMKT, with variable names

SHAREIND

CARPROD-LAG6

COMMIND-LAG7

(a) Plot each "independent" variable against Y. Comment on these plots.

(b) Calculate the LS regression (i.e., fit a multiple regression line to the data). Does the model make sense? (Comment with respect to:

(i) Causality

(ii) Possible dependence among any of the "independent variables")

(c) Calculate and plot the residuals. What do they tell you about your fitted line?

(d) Calculate R^2. What does this tell you about your fitted line?

(e) Based upon your answers to (b), (c), and (d) above, would you be willing to use this model to "play the market"? Why or why not?

Homework Unit 4
Solutions

1. **Omaha Data**

(a) MFU vs. POP is plotted in Figure A. An approximately linear pattern is discernable (note the line).

MFU vs. NWHITE-PCT is plotted in Figure B. Transformation by log or negative reciprocals is clearly required.

MFU vs. the negative reciprocals of NWHITE-PCT is plotted in Figure C. The trend is approximately linear (note line). A plot of MFU vs. LOG of NWHTE!-PCT demonstrates less linearity. (Plot not shown.) MFU vs. YEARS-ED-C is plotted in Figure D. The pattern is roughly linear (note the line).

MFU vs. PCT-GT65Y is plotted in Figure E. Again, transformation by log or negative reciprocals seems necessary.

MFU vs. LOG of PCT-GT65Y is plotted in Figure F. A roughly linear pattern is discernable (note the line). A plot of MFU vs. the negative reciprocals of PCT-GT65Y shows little improvement over the LOG transformation, so we will stick with the LOG.

A common and useful transformation for proportions (% divided by 100) is the arcsin of the squareroot of the data. Plots using this transformation of NWHITE-PCT and PCT-GT65Y show little noticeable improvement over the above negative reciprocal and log transformations. Because the arcsin of the squareroot transformation is difficult to interpret, only a substantial improvement in linearity warrants its use.

(b) The MREG output for the transformed data is in Figure G. Note that in both cases the coefficient of years of education is large and positive (education seems to <u>increase</u> income), whereas the coefficients for percent nonwhite and percent over 65 are large and negative, although different in each case--as they should be since we transformed these two variables. The population effect seems negligible.

A unit analysis is not really appropriate to this problem. Unlike the simple addition of international currencies (in which all must be converted to common units), in this problem we are attempting to "predict" or "measure" median income (dollars) by several variables measured in differing units (# people per tract, # nonwhite/100 people in tract, years of education, and # people over 65/100 people in tract).

Dependence among the "independent" variables should be investigated further. For example:

- % of people older than 65 may be correlated with years of education, due to
  - current adult education trends
  - the fact that most of the elderly are women, and higher education of women was (allegedly) more prominent in the 1930's.
  - older people have had more years in which to be educated!
- % nonwhite may be highly (negatively) correlated with years of education
- If one believes nonwhite mortality rates to be higher than those for whites, % older than 65 may be highly negatively correlated with % nonwhite.

Causality is another question. It does make a certain amount of sense to suggest that changes in the age, race, and education levels of the population "cause" changes in the median income level. We certainly do not expect population to be causal.

(c) A stem-and-leaf display and a boxplot of the residuals are shown in Figure H. The residuals appear well-behaved--symmetric about median 0, small, and with few outliers.

(d) R^2 for the transformed data is .7863 (Figure G). The model fits rather well, from an R^2 point of view.

(e) Median income appears to be highly positively influenced by years of education, and negatively influenced by both % nonwhite and % over 65. This suggests that the CAUSES for lower nonwhite and elderly incomes should be investigated in more detail, and policies to benefit those groups (minority education, training programs, more efficient use of elderly resources, etc.) should be examined. Before any such policy is adopted, however, all of its implications should be thoroughly studied.

(f) Since population appears not to greatly affect the model, this X variable might be dropped. (If the resulting model is very different, however, this move must be reconsidered.)

(g) Some possibilities might be

- property tax rate (if this varies by tract)
- % of tract zoned for apartments (this however could be double-edged in the case of luxury apartments)
- % of population in professional fields (i.e. lawyers, doctors, etc.)
- population density (usually inversely related to median income)

Of course, there are many, many, other possibilities.

FIGURE A

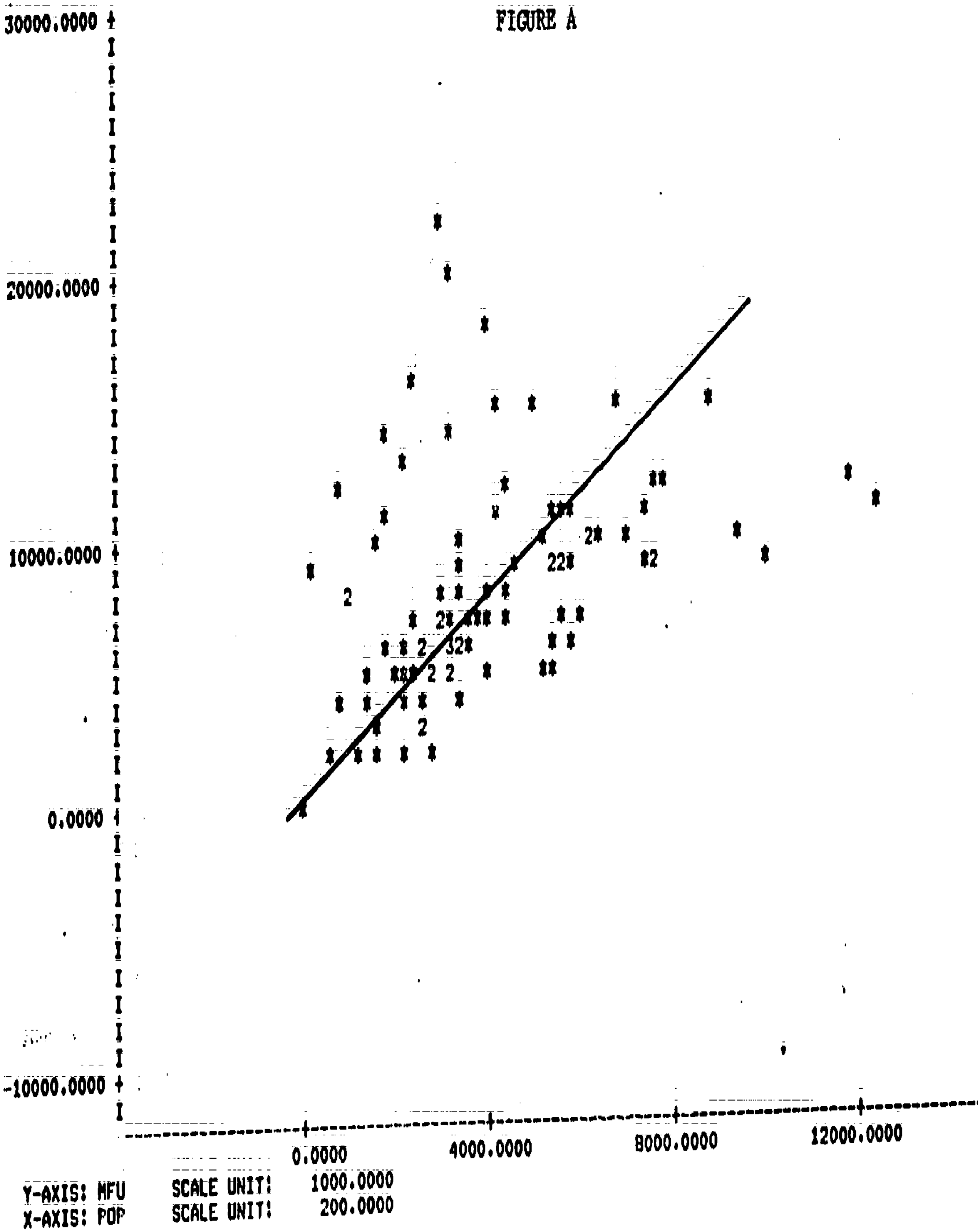
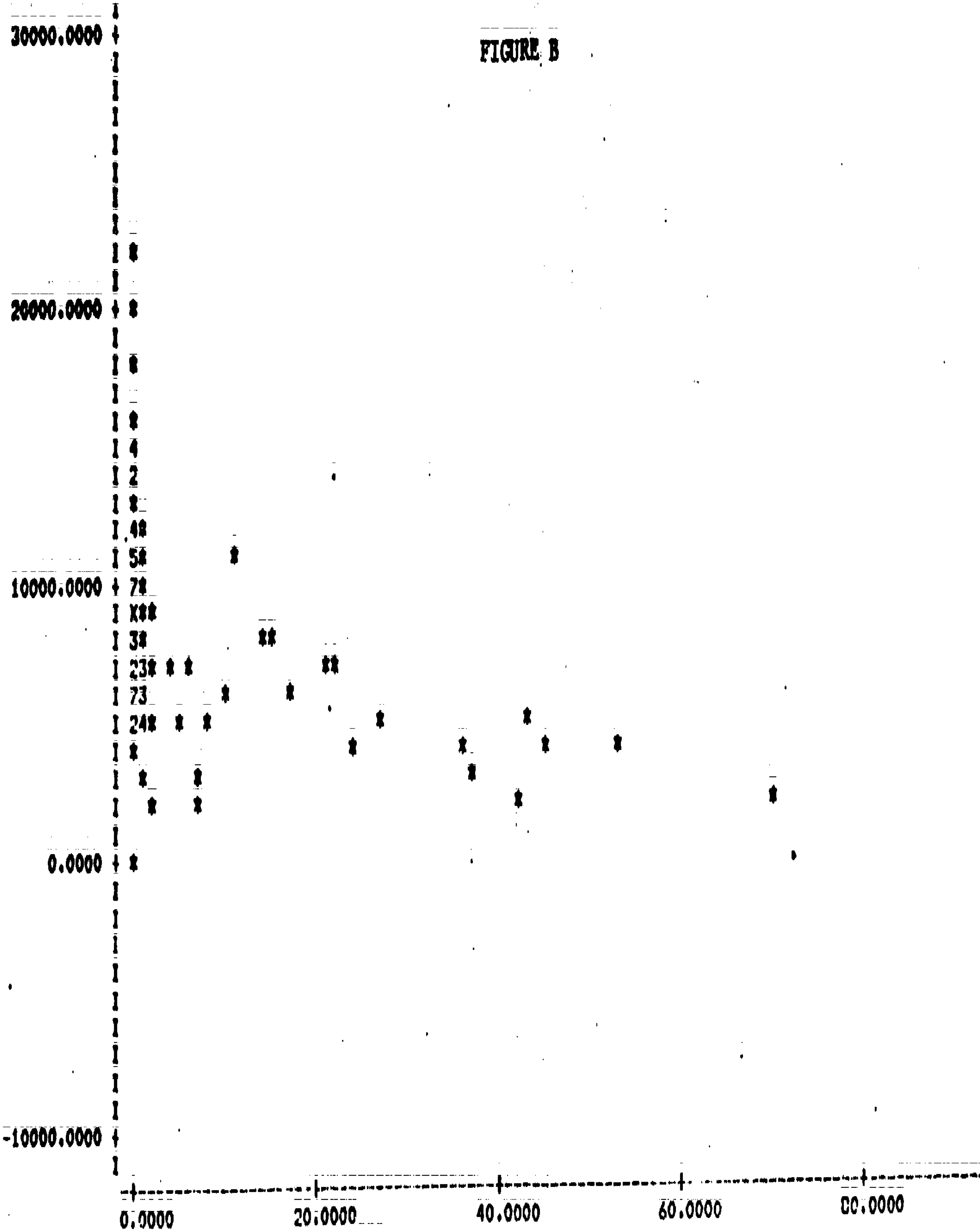

30000.0000
20000.0000
10000.0000
0.0000
-10000.0000
0.0000
4000.0000
8000.0000
12000.0000
Y-AXIS: MFU SCALE UNIT: 1000.0000
X-AXIS: POP SCALE UNIT: 200.0000

FIGURE B

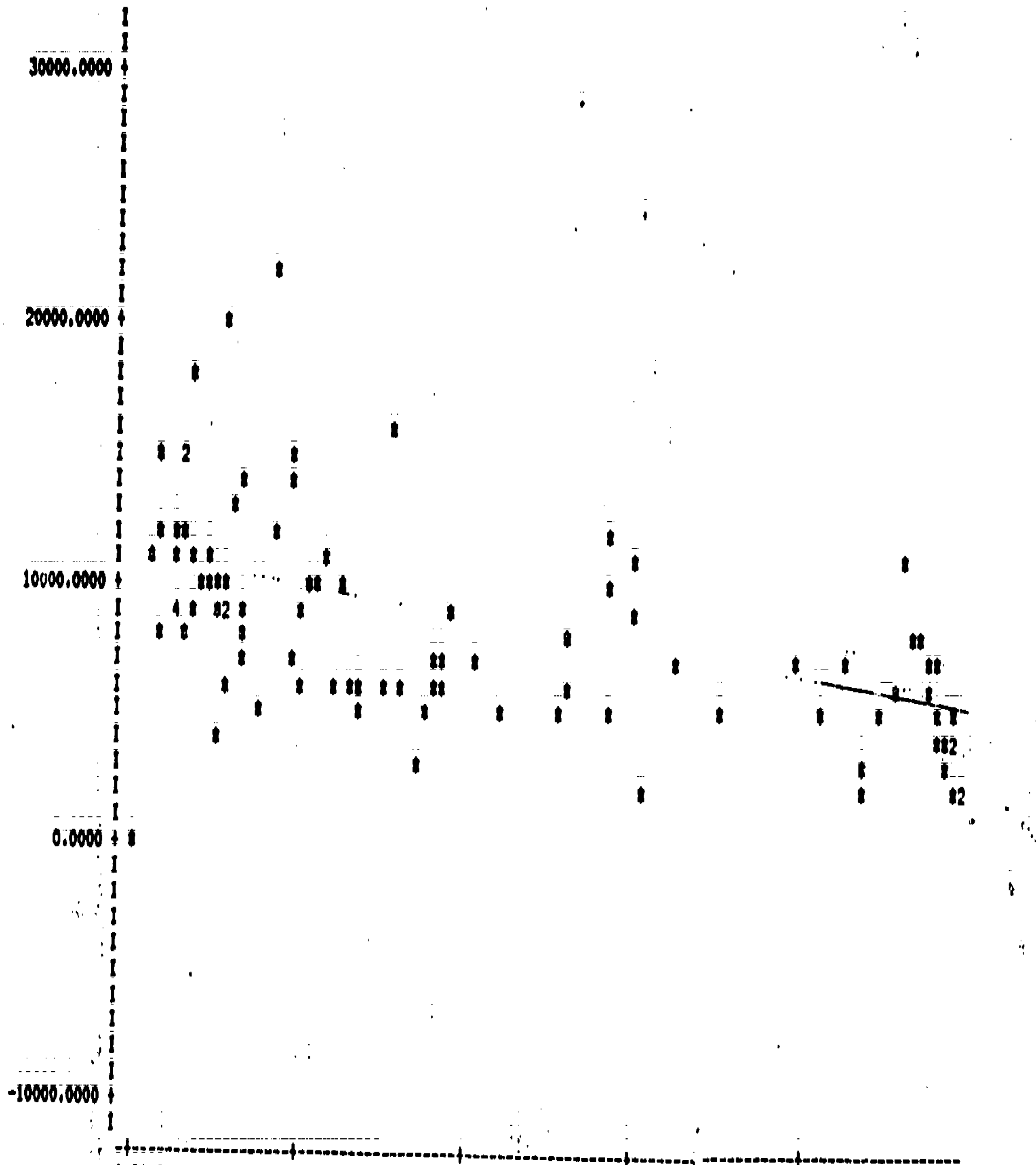


Y-AXIS: MFU SCALE UNIT: 1000.0000
X AXIS: NWHITE PCT SCALE UNIT: 1.0000

XVI.II.334

PLOT MFU VS INWHITE

FIGURE C

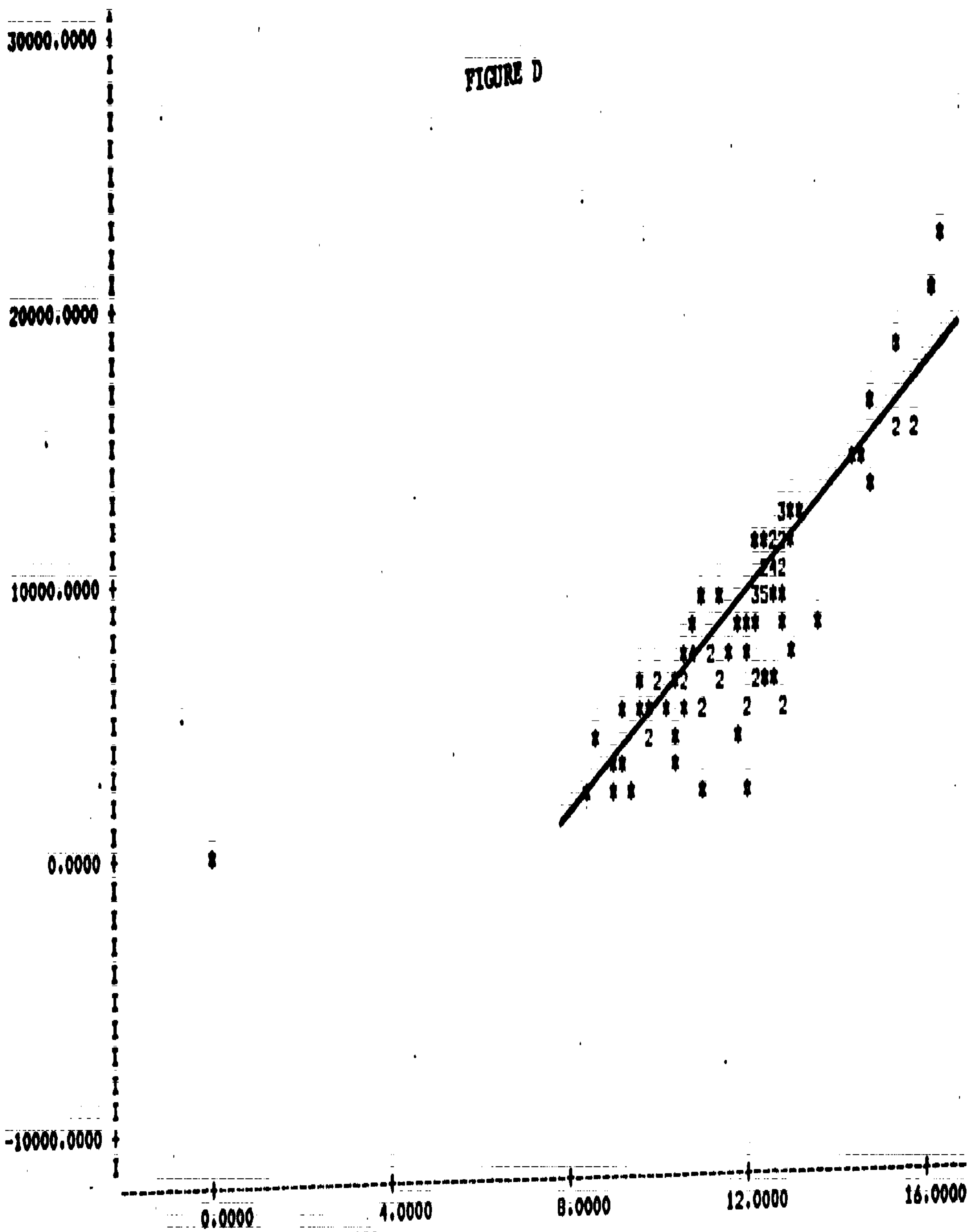


Y-AXIS: MFU SCALE UNIT: 1000.0000
X-AXIS: INWHITE SCALE UNIT: 0.0100

XVI.II.335

Module II

FIGURE D



Y-AXIS: MFU SCALE UNIT: 1000.0000
X-AXIS: YEARS-ED-C SCALE UNIT: 0.2000

XVI.II.336

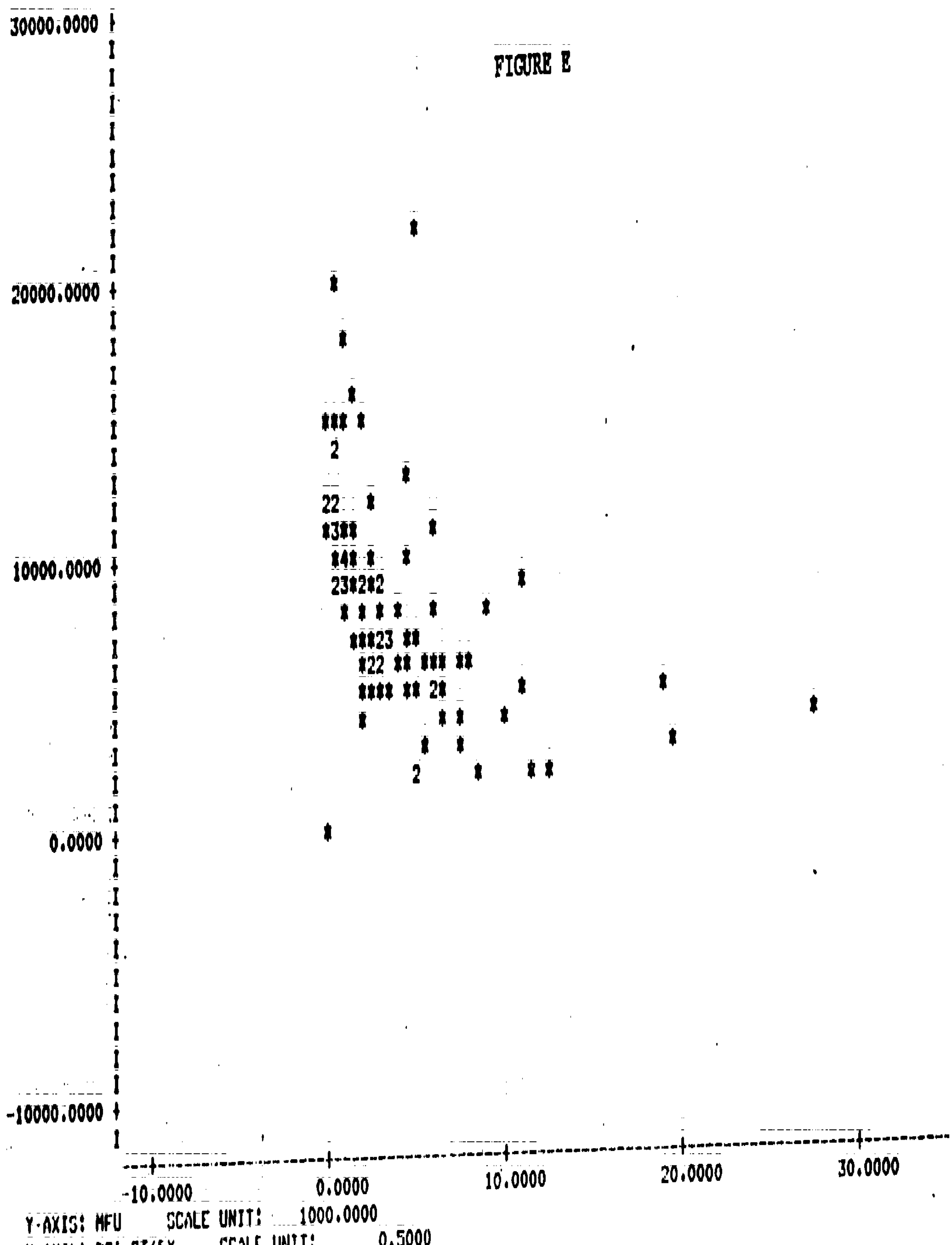
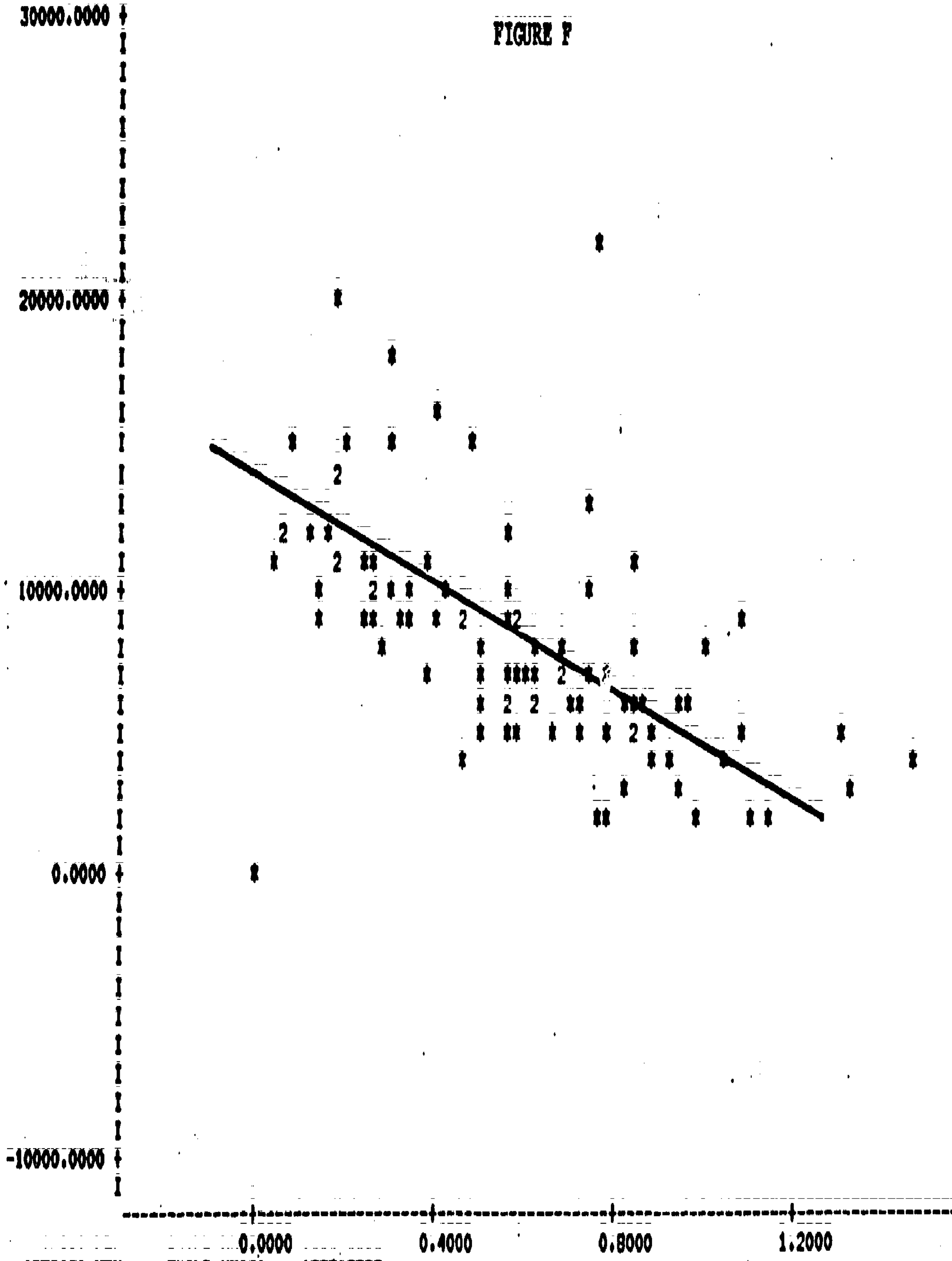


FIGURE E
30000.0000
20000.0000
10000.0000
0.0000
-10000.0000
-10.0000
0.0000
10.0000
20.0000
30.0000
Y-AXIS: MFU SCALE UNIT: 1000.0000
X-AXIS: PCT GT65Y SCALE UNIT: 0.5000

FIGURE F



Y-AXIS: MFU SCALE UNIT: 1000.0000
X-AXIS: LGT65Y SCALE UNIT: 0.0200

XVII.III.338

FIGURE G

MREG MFU VS POP INWHITE YEARS!-ED!-C LGT65Y SAVERES RES2

RESPONSE	MEAN	STD. DEV.
MFU	8244.7266	3972.7734

CARRIER:	CONSTANT	POP	INWHITE	YEARS-ED-C	LGT65Y
COEFFICIENT	-4056.4160	-0.2932	-2168.5540	1255.4548	-4721.1680
S.E. COEF.		0.1066	670.1370	102.9374	796.8584
MEAN		4056.6145	-0.6025	11.8695	0.5757
STD. DEV.		2415.7300	0.3430	2.1032	0.3179

MULTIPLE R SQUARED 0.7863

ANALYSIS OF VARIANCE TABLE

	SS	DF	MS	RMS
FIT	1.1790E+09	4	2.9474E+08	17167.9414
RESIDUAL	3.2043E+08	91	3.5212E+06	1876.4761
TOTAL	1.4994E+09	95		

	F	F PROB.
FIT	83.7048	1.0000

XVI.II.340

FIGURE H

```
STEM RES2

VARIABLE ' RES2 '.
 UNIT =     100.0000
            LO I   -5920.2266  -4235.4609  -3711.3889
  5         -3  I 30
  8         -2. I 855
 13         -2  I 44322
 18         -1. I 98755
 21         -1  I 410
 30         -0. I 998777555
 43         -0  I 4433322111000
( 16)        0  I 0001112333344444
 37          0. I 5566677888899999
 20          1  I 000012233
 11          1. I 5568
  7          2  I 13
  5          2. I 9
  4          3  I 44
            HI I     4632.9141    7673.5938
!
```

```
BOXPLOT RES2 THREE

  SCALE UNIT:    200.0000
        -4000.0000          0.0000        4000.0000        8000.0000       12000.0000
-----------+-------------------+-------------------+-------------------+-------------------+----
VARIABLE: RES2
                                 +----+----+
     *          * * <------------I    +    I------------>   *          *
                                 +----+----+
!
```

2. Cobbdouglas Problem

(a) We have $(\frac{V}{N}) = A(\frac{L}{N})^{\alpha} (\frac{K}{N})^{\beta}$

taking natural logs: $\ln(\frac{V}{N}) = \ln [A(\frac{L}{N})^{\alpha} (\frac{K}{N})^{\beta}]$

$$= \ln A + \ln(\frac{L}{N})^{\alpha} + \ln(\frac{K}{N})^{\beta}$$

$$= \ln A + \alpha \ln(\frac{L}{N}) + \beta \ln(\frac{K}{N})$$

$$V^* = \ln(\frac{V}{N}) = b_0 + b_1 L^* + b_2 k^*$$

the usual multiple regression form.

(We prefer natural log for theoretical reasons. The analysis could also have been done using base 10 logs.)

Plots of V^* vs. L^* and V^* vs. K^* are shown in Figures A and B. Both plots are remarkably linear for real data. (But note that since we are dealing with a model based on a specific and well-defined theoretical model we would not perform any transformations on V^*, L^*, or K^*.

(b) The MREG output is shown in Figure C. We are "predicting" or "modeling" value added (per establishment) in dollars by manhours (labor) per establishment and capital flow (dollars) per establishment. Note that all variables are measured in units PER ESTABLISHMENT. The underlying economic theory considers labor and capital services flow as causing value added, and not dependent on each other (although one certainly tends to increase both in order to increase value added).

(c) A stem-and-leaf plot of the residuals is shown in Figure D. Except for the one HI value, they appear fairly well behaved. The residuals are then plotted against each independent variable; against K^* in Figure E and L^* in Figure F. Both plots appear "random".

(d) From the MREG output, $R^2 = .9597$, hence our model "explains" the data very well.

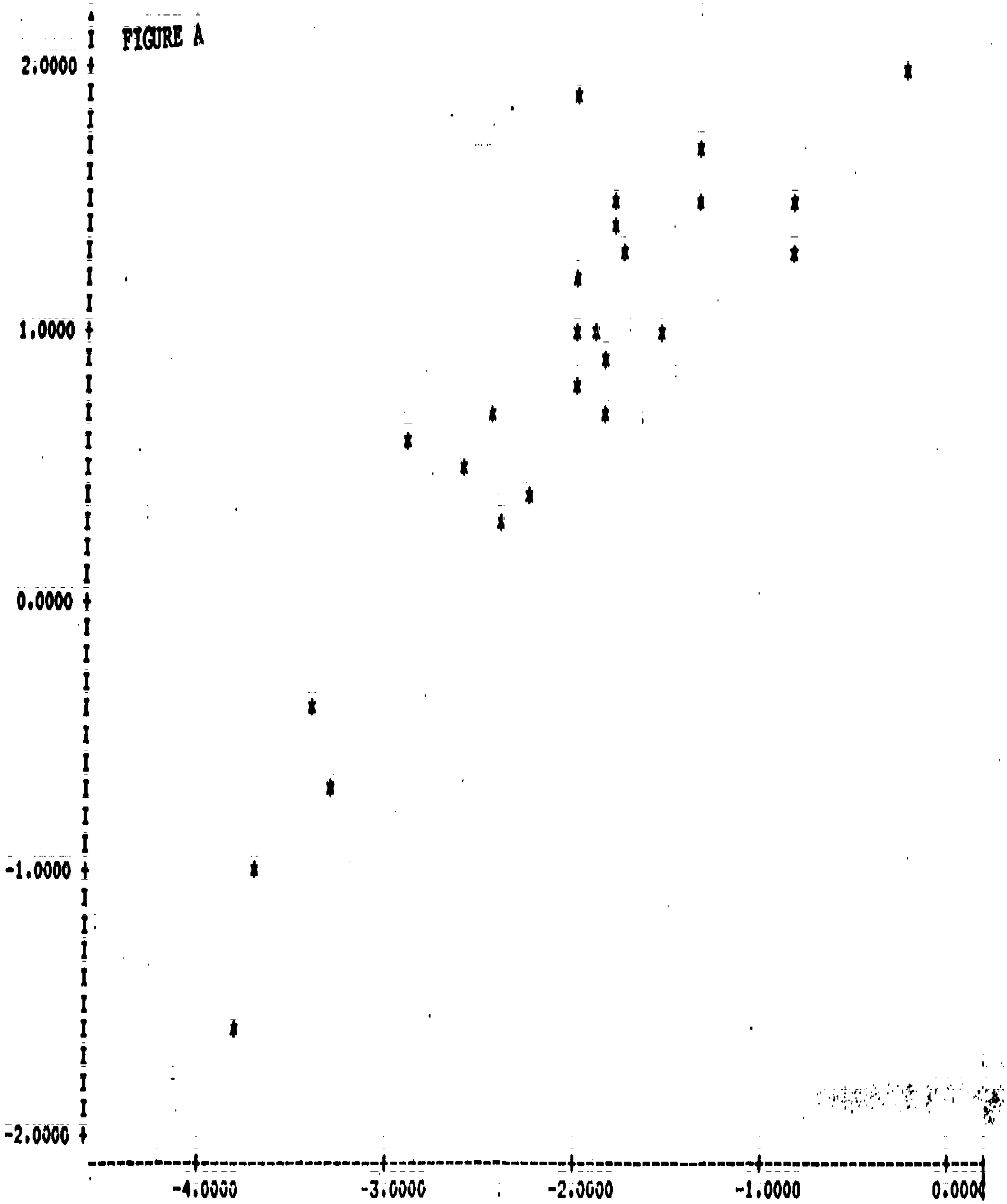
(e) The elasticity of value added with respect to labor/establishment is $\alpha = .9276$. The elasticity of value added with respect to capital/establishment is $\beta = .2788$.

Recall from Tufte page 114 that the elasticity of Y with respect to X measures the percentage change in Y with respect to the percentage change in X. Hence, if labor/establishment is doubled (i.e. 100% increase), value added increases by 93%. Similarly, if capital flow is doubled, value added increases by 28%.

(f) While labor/establishment appears to provide the larger proportional increase in value added (since $\alpha < \beta$), the key to this question is the unknown <u>costs</u>. Given a fixed amount of resources, what is the <u>unit</u> cost (in common units) for a given proportional change in labor/establishment compared to the <u>unit cost</u> for the same proportional change in capital/establishment.

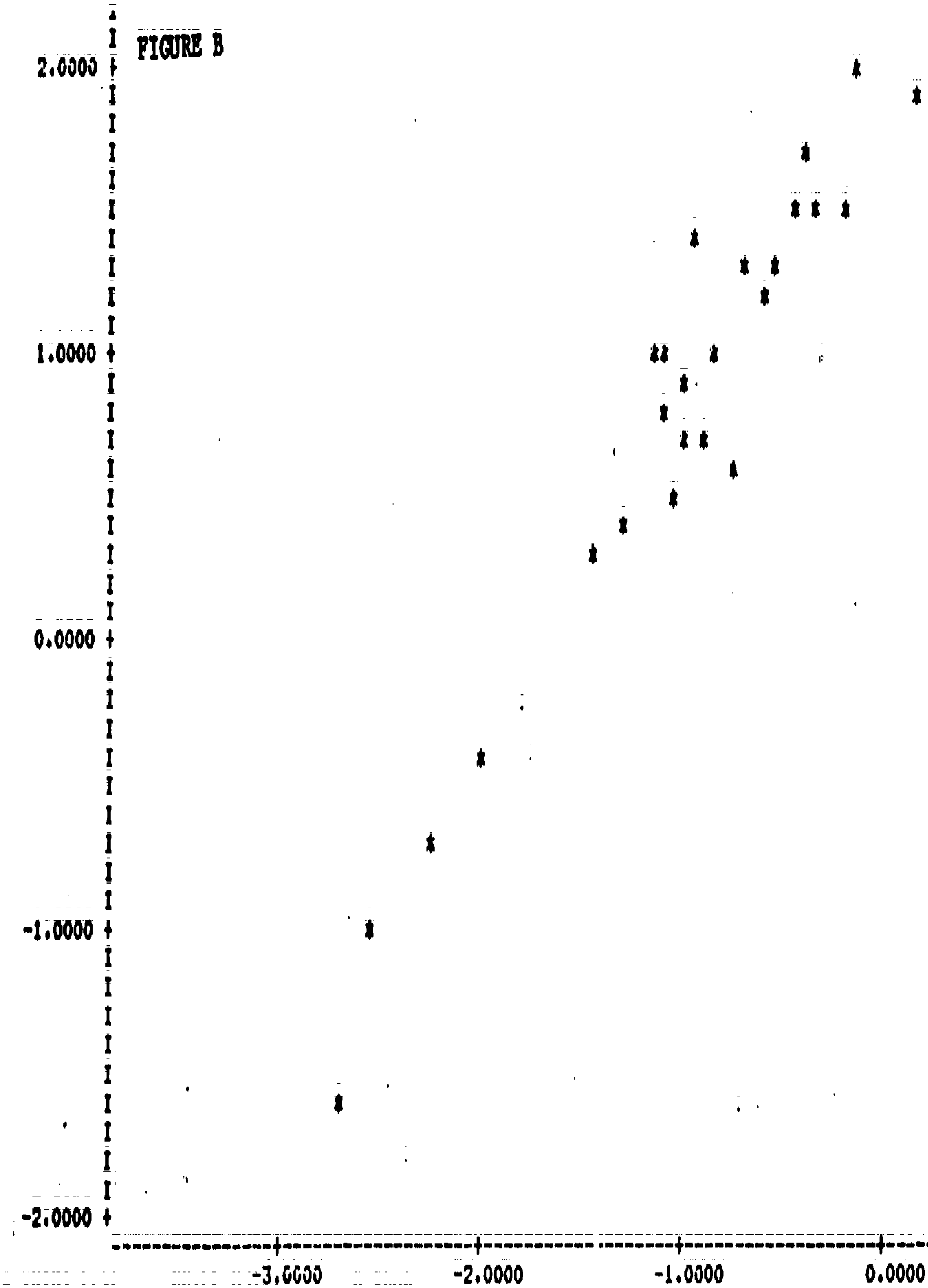
For example, if each % change in capital/estab costs \$X, and each % change in labor/est. costs \$Y = \$2X, then for approximately the same amount of resources we could (approximate) double value added/establish by either quadrupling capital services/estab or doubling labor/estab., or by a combination of increased labor <u>and</u> capital flow.

Hence the optimal policy is not obvious, depending heavily upon these unknown costs. Research is required to determine these costs.



Y-AXIS: LVAL SCALE UNIT: 0.1000
X-AXIS: LCAP SCALE UNIT: 0.0500

FIGURE B



Y-AXIS: LVAL SCALE UNIT: 0.1000
X-AXIS: LLAR SCALE UNIT: 0.0500

XVI.II.344

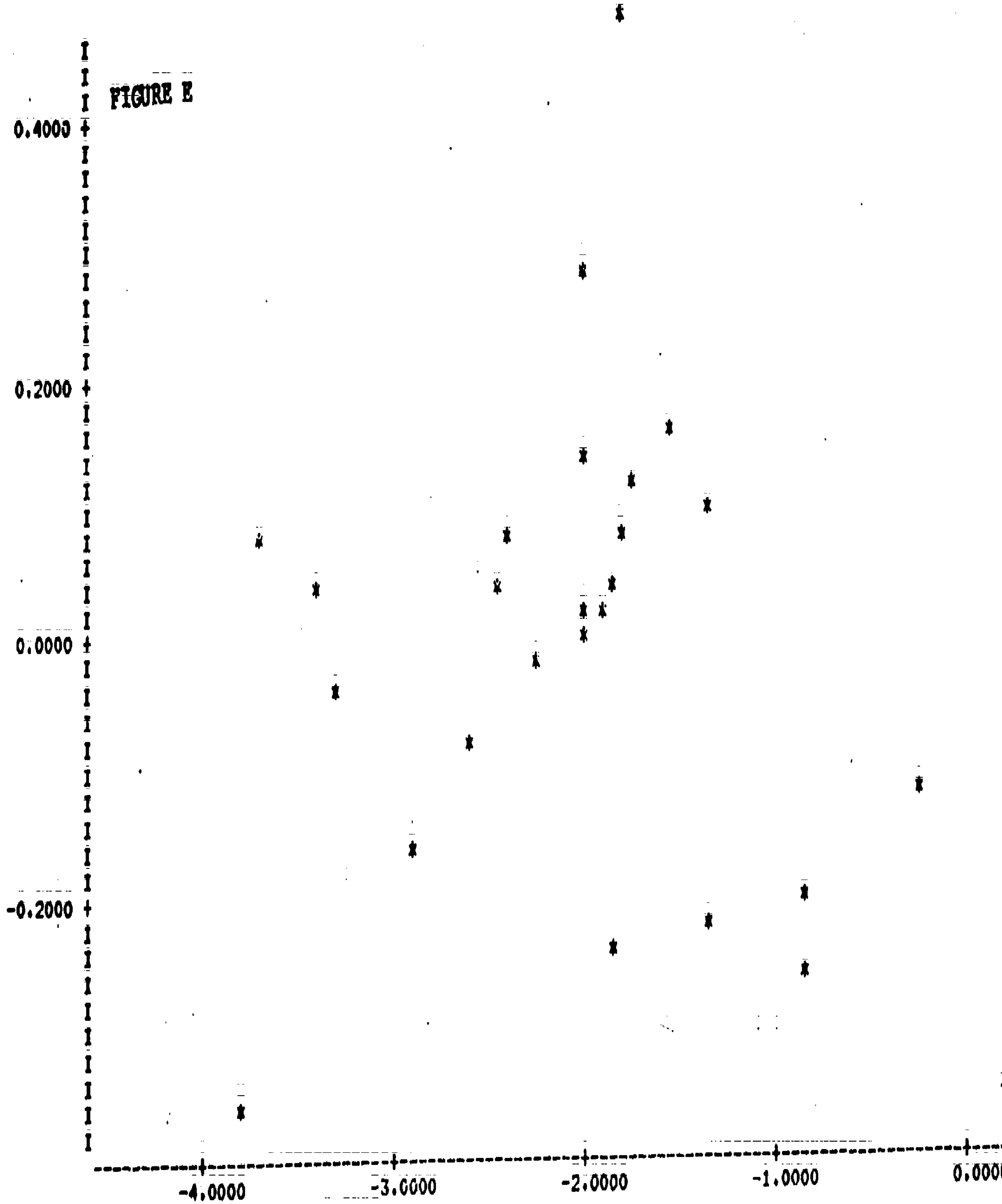
FIGURE C

```
MREG LVAL VS LCAF LLAB SAVERES RES3
   RESPONSE            MEAN   STD. DEV.
       LVAL          0.7717      0.8993

   CARRIER:        CONSTANT        LCAF        LLAB
COEFFICIENT          2.2932      0.2788      0.9276
 S.E. COEF.                      0.0807      0.0983
       MEAN                     -2.0820     -1.0146
  STD. DEV.                      0.8750      0.7164

MULTIPLE R SQUARED         0.9597

ANALYSIS OF VARIANCE TABLE
                         SS      DF           MS          RMS
FIT                 18.6281       2       9.3140       3.0519
RESIDUAL             0.7819      22       0.0355       0.1885
TOTAL               19.4100      24

                          F      F PROB.
FIT                262.0535       1.0000
:
```

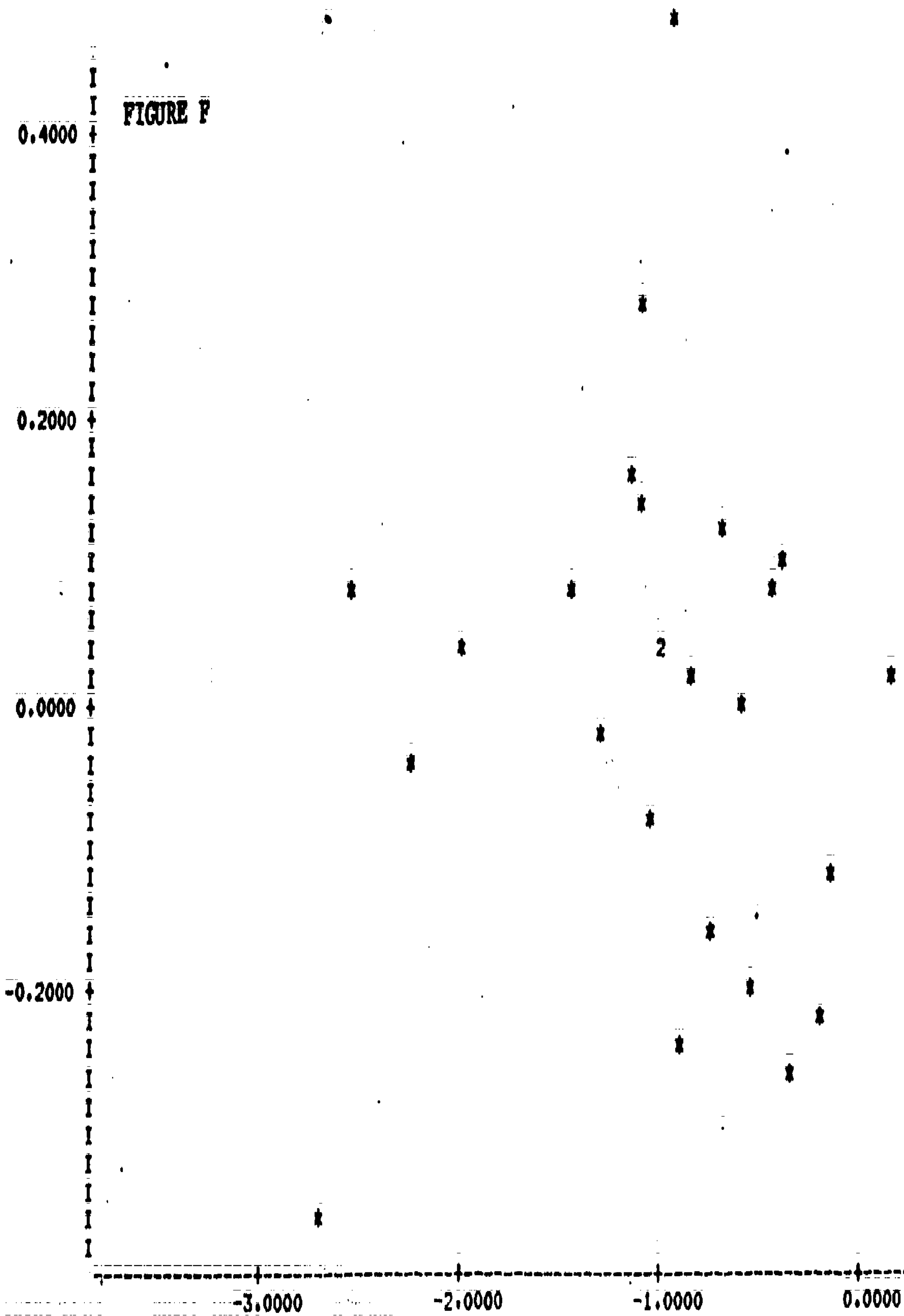
FIGURE D

```
 STEM RES3

VARIABLE ' RES3 '.
 UNIT =        0.0100
    1          -3  I 5
    5          -2  I 6410
    7          -1  I 51
   10          -0  I 732
 ( 10)          0  I 0114447689
    5           1  I 245
    2           2  I 7
               HI  I       0.4780
:
```

FIGURE E



Y-AXIS: RES3 SCALE UNIT: 0.0200
X-AXIS: LCAP SCALE UNIT: 0.0500



Y-AXIS: RES3 SCALE UNIT: 0.0200
X-AXIS: LLAR SCALE UNIT: 0.0500

3. Gas Problem

In this problem the first thing we must do is consider the units and magnitudes of the data we are working with. CONSUMPTION, VEHICLES, and POP are each in units of 1 with magnitudes of 6 or 7. PCINCOME is in units of 1 with a magnitude of 3. GASTAX and POPDENS are in units of 1 with magnitudes of 0 and 1, respectively. The very large differences in these magnitudes will be reflected in large differences in the magnitudes of the variances of the variables. We will also be faced with both very large and very small coefficients in the regression equation. To avoid these scaling problems, we need to equalize the magnitudes by increasing the units of those variables with magnitudes greater than 0 or 1. (There may also be situations in which units should be decreased to achieve equality of magnitudes.) The appropriate DAP commands are:

```
LET  CONSUMP = CONSUMPTION/1000000
LET  VEHIC = VEHICLES/1000000
LET  P = POP/1000000
LET  PCI = PCINCOME/1000
```

(NOTE: The variable names on the left hand sides of the equations can be any name of your choice.)

The resulting variables CONSUMP, VEHIC, and P are in units of 1 million with magnitudes of 0 and 1. PCI is in units of 1 thousand with magnitude 0.

(a) A plot of CONSUMP vs. VEHIC is shown in Figure A. The data look remarkably linear.

A plot of CONSUMP vs. GASTAX is shown in Figure B. The pattern (if any) appears linear.

A plot of CONSUMP vs. P is shown in Figure C. The data appear reasonably linear.

A plot of CONSUMP vs. POPDENS is shown in Figure D. While a transformation appears to be in order, none seems to improve the pattern. LOG, SQRT, and Negative Reciprocal were tried with no noticeable improvement.

A plot of CONSUMP vs. PCI is shown in Figure E. The plot suggests some linearity and no transformation seems applicable. The cone shaped spread is of some concern, but a Log transformation of CONSUMP does not improve the plot much. A transformation of the dependent variable would also necessitate transformations of the independent variables which exhibit linearity in the raw form. The slight improvement in spread in the one plot is not worth the sacrifice of simplicity of the model.

Examining these plots, VEHIC and P appear most, and GASTAX and POPDENS least, related to fuel consumption.

(b) The MREG output for CONSUMP vs. each of the five X variables is shown in Figure F.

While the model "makes sense" in that we are attempting to model or predict fuel consumption by five measurable qualitatively related X variables, it is a good example of one in which we have too many unnecessary "independent" variables, resulting in much complexity with little or no offsetting improvement in accuracy (as we shall see in (e) below)

First, we might expect VEHIC to be related to P. Certainly we expect more vehicles when there are more people. In cities, however, we might expect considerably fewer vehicles (many city dwellers do not own cars, although the number of cabs and buses would be larger). In suburban or rural areas, we might expect a larger number of vehicles for that population. The number of vehicles per person is not constant across geographic area, although some relationship between these two variables obviously exists.

Economic theory tells us that GASTAX should not have a great effect on CONSUMP. Our plots in part (a) reinforce this. We expect people to continue to utilize the available modes of transportation in order to commute to and from work, regardless of the marginal differences in the price of gasoline due to taxes (as opposed to dramatic price increases caused by other factors).

PCI should probably have a "threshold" effect on CONSUMP. When median income is below some threshold value, people cannot afford a private car and must rely on public transportation. Above this threshold, CONSUMP would then rise to some limiting value, after which increases in income do not have any effect.

POPDENS serves as a sort of index of urbanization. As discussed above, we expect fewer autos per capita (but more buses, cabs, and possibly truck traffic) in highly urban regions, while we expect the reverse in rural areas.

(c) A stem-and-leaf and boxplot of the residuals is shown in Figure G. The plots should make us suspicious of our model, since the residuals are not particularly well behaved. Note the large number of HI values, and the overall location of the batch (i.e. nonzero median). A plot of the residuals vs. each independent variable would also be helpful, perhaps indicating a hidden relationship (e.g. consumption vs. a quadratic in POPDENS). The residuals are all small, however, relative to the size of the original CONSUMP units.

(d) R^2 for this model is .9875 (Figure F). This indicates that the model "explains" the data very well. (But the answer in (c) above should caution us against evaluating the adequacy of a model solely on its R^2 value).

(e) From our discussion in (b) and the plots in (a), we know that one of the two must be VEHIC. Since we expect P to be functionally related to VEHIC, PCI is the best choice for a second variable, as is indicated by the t statistics.

The MREG output for this model is shown in Figure H. Note that the coefficients for these two variables are almost exactly the same as those in the complete model (Figure F).

A stem-and-leaf, and a boxplot of the residuals is shown in Figure I. There is some improvement (fewer HI values) over the full model, but the plots are similarly suspicious.

R^2 for this (reduced) model is .9864, almost the same as that for the "complete" model.

Now, note that in the reduced model:

(i) The coefficients of the remaining variables are virtually identical to their respective coefficients in the "complete" model.

(ii) The residual patterns are similar to those of the "complete" model.

(iii) The R^2 value is essentially the same as that of the "complete" model.

These are three very strong indications that the three independent variables dropped from the "complete" model were in fact unnecessary.

(f) There is really no reason even to consider using the "complete" model, since the reduced model does just as "well" (e above). Note that in the reduced model all data are from 1973.

For policy, decision making, or detailed study purposes, however, all data should be from the same time frame.

(g) One possibility is to use AUTOS, BUSES, and TRUCKS instead of VEHICLES.

Price per gallon at the pump (PPGAL) might be a good predictor, since we expect radically higher prices to discourage leisure and "unnecessary" driving.

Miles of highway (MILES!—ROAD) might be used as another indication of mobility in each state.

Indicator variables for minimum driving ages (IND!-AGE) and for mandatory state auto inspections (IND!-INSPECT) might also be used to detect possible influences of driving habits of automobile efficiency on gas consumption

Other possibilities no doubt exist. Clearly however consumption depends primarily on the number of vehicles and the degree of use.

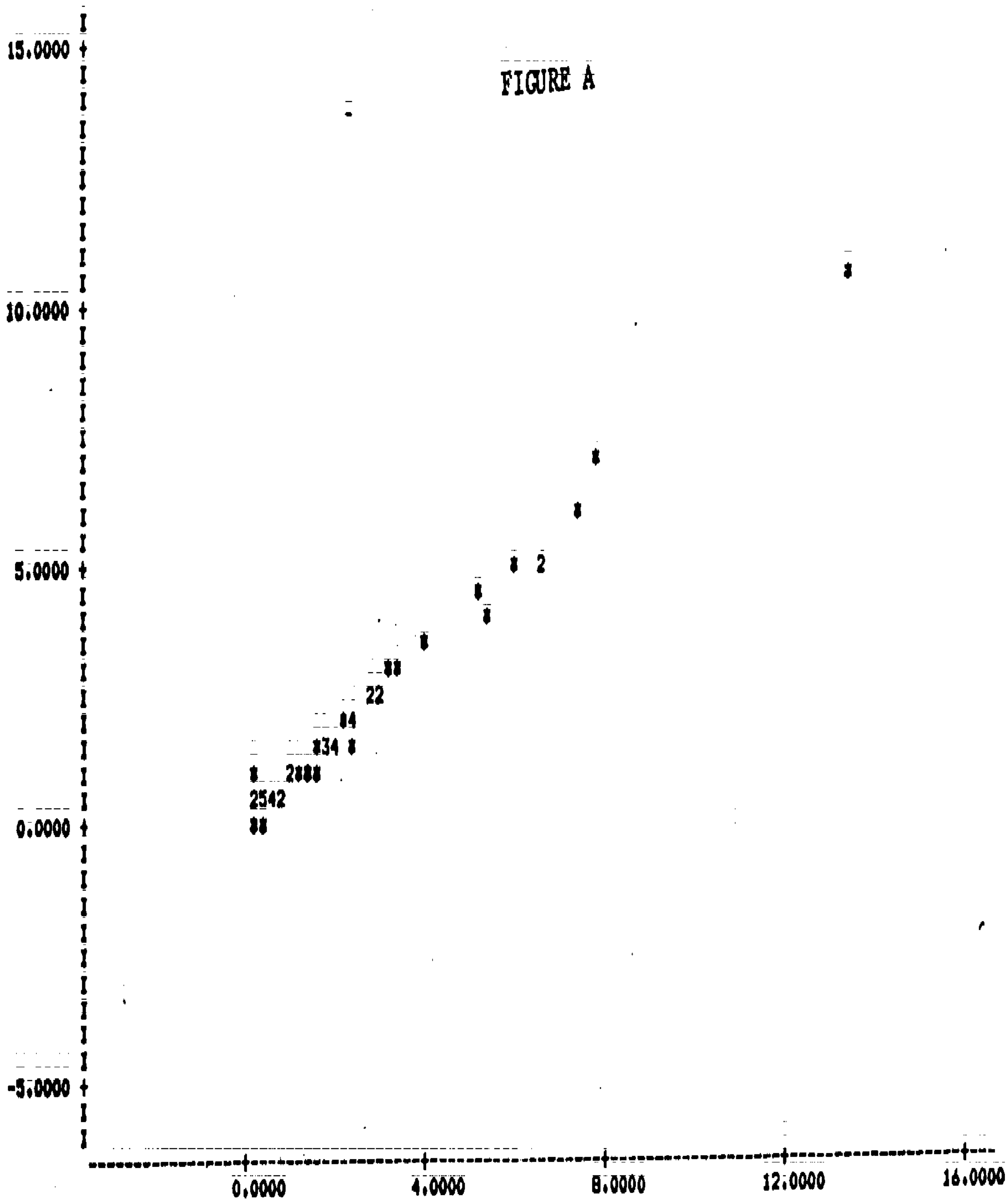
The model might then be

$$\text{CONSUMPTION} = b_0 + b_1\ \text{AUTOS} + b_2\ \text{BUSES} + b_3\ \text{TRUCKS} + b_4\ \text{PPGAL}$$
$$+ b_5\ \text{MILES!-ROAD} + b_6\ \text{AGE} + b_7\ \text{INSPECT}$$

A baffling outcome of the regressions in this problem is the negative coefficient for PCI. This is not what we would expect in theory nor what we would have guessed from the plot of CONSUMP vs PCI (Figure E). It seems likely that the large variation in the CONSUMP coordinates for the large PCI values permits a great deal of latitude in fitting a least squares line. Given this unexpected and difficult to explain outcome, we might try a univariate regression of CONSUMP vs VEHIC. The MREG result is shown in Figure J. Note that the R^2 is virtually unchanged. The residual structure is even somewhat improved (more symmetric, smaller values) as is shown by a stem-and-leaf display and boxplot (Figure K).

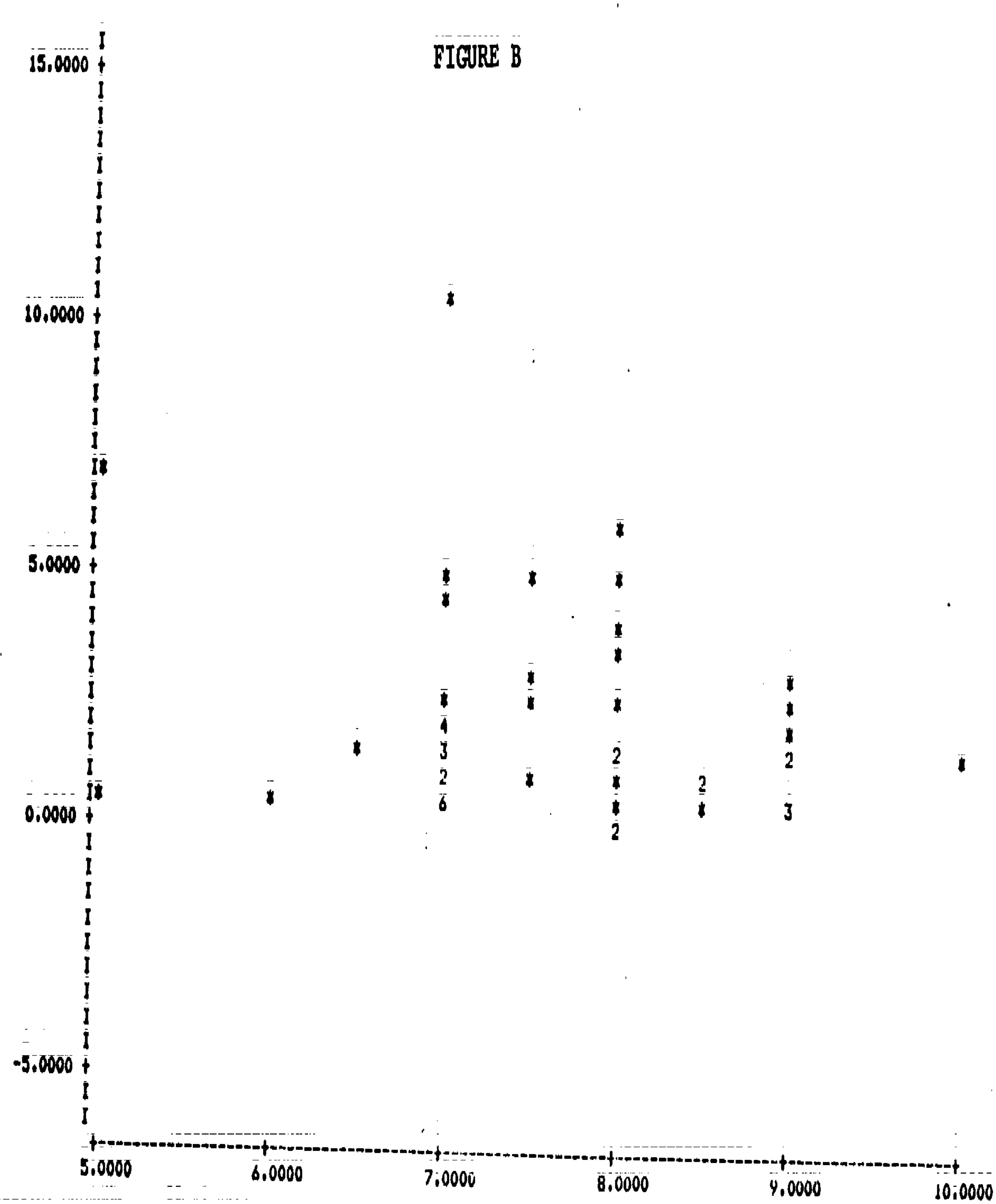
Anytime we are able to achieve a very high R^2 with a good residual structure using univariate regression, we should be very reluctant to add other variables for small "improvements" in the explanatory power of the model. This problem is a clear case of where the univariate model is the "best".

PLOT CONSUMP VS VEHIC



Y-AXIS: CONSUMP   SCALE UNIT: 0.5000
X-AXIS: VEHIC   SCALE UNIT: 0.2000

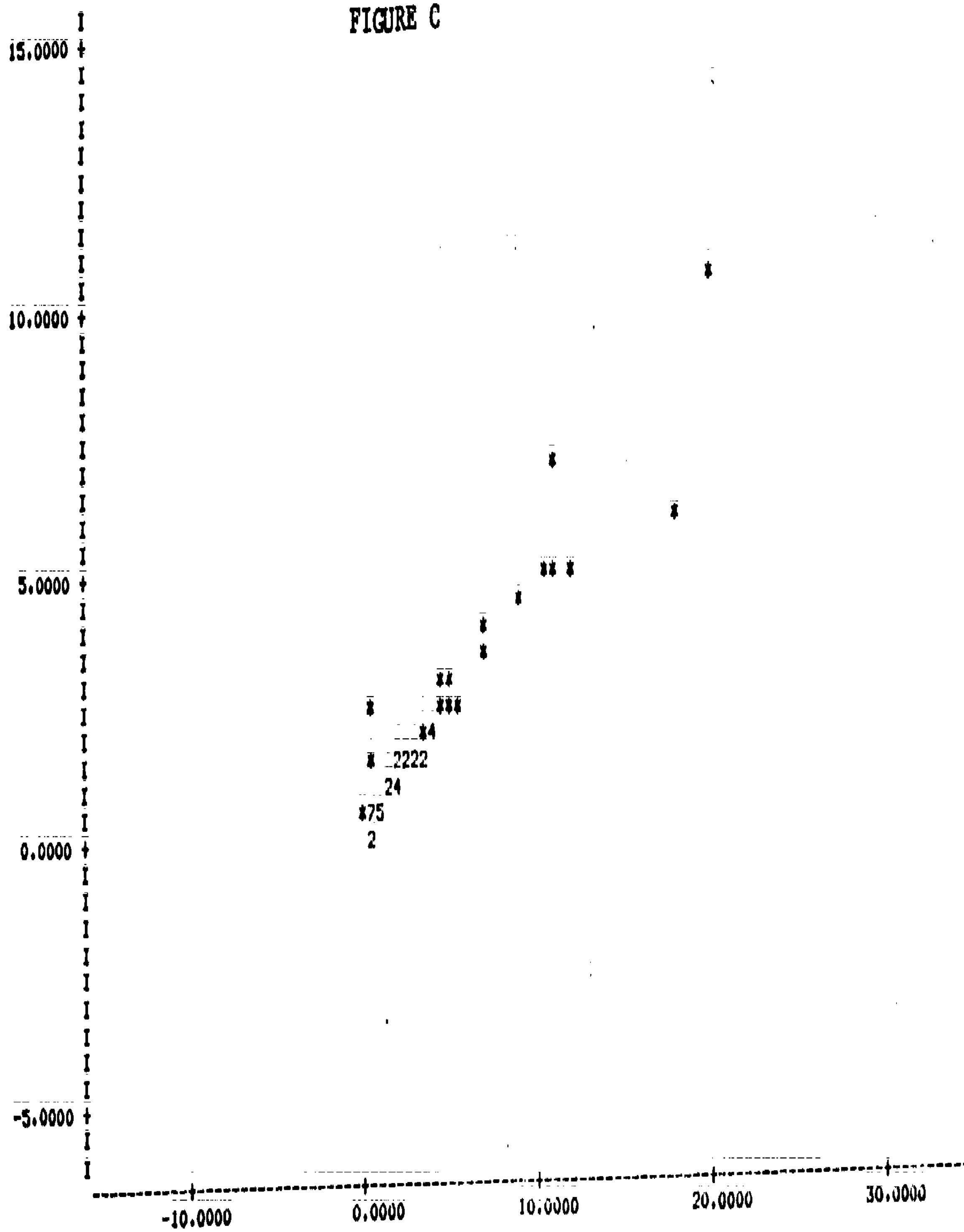
PLOT CONSUMP VS GASTAX



Y-AXIS: CONSUMP SCALE UNIT: 0.5000
X-AXIS: GASTAX SCALE UNIT: 0.0500

PLOT CONSUMP VS P

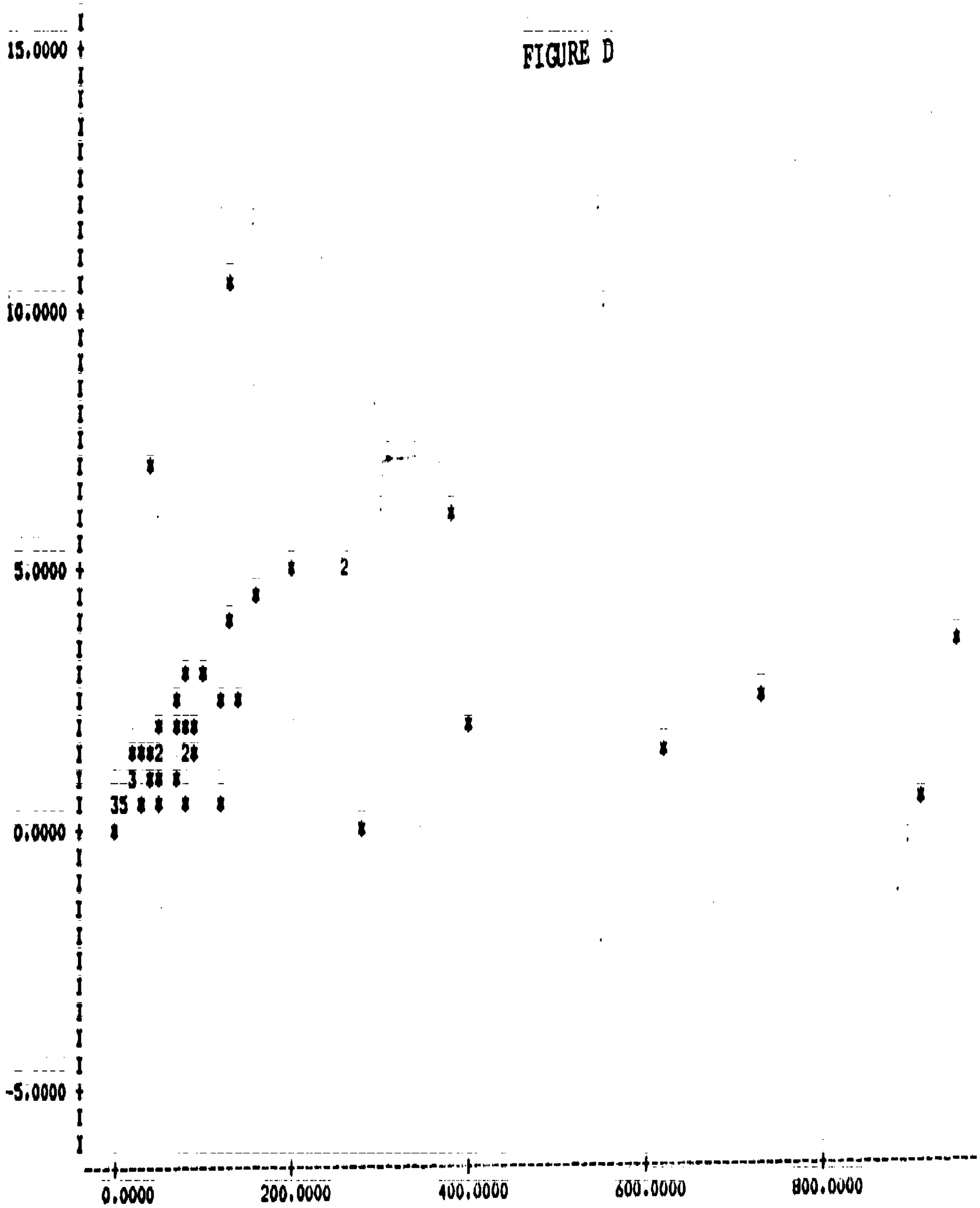
FIGURE C



Y-AXIS: CONSUMP SCALE UNIT: 0.5000
X-AXIS: P SCALE UNIT: 0.5000

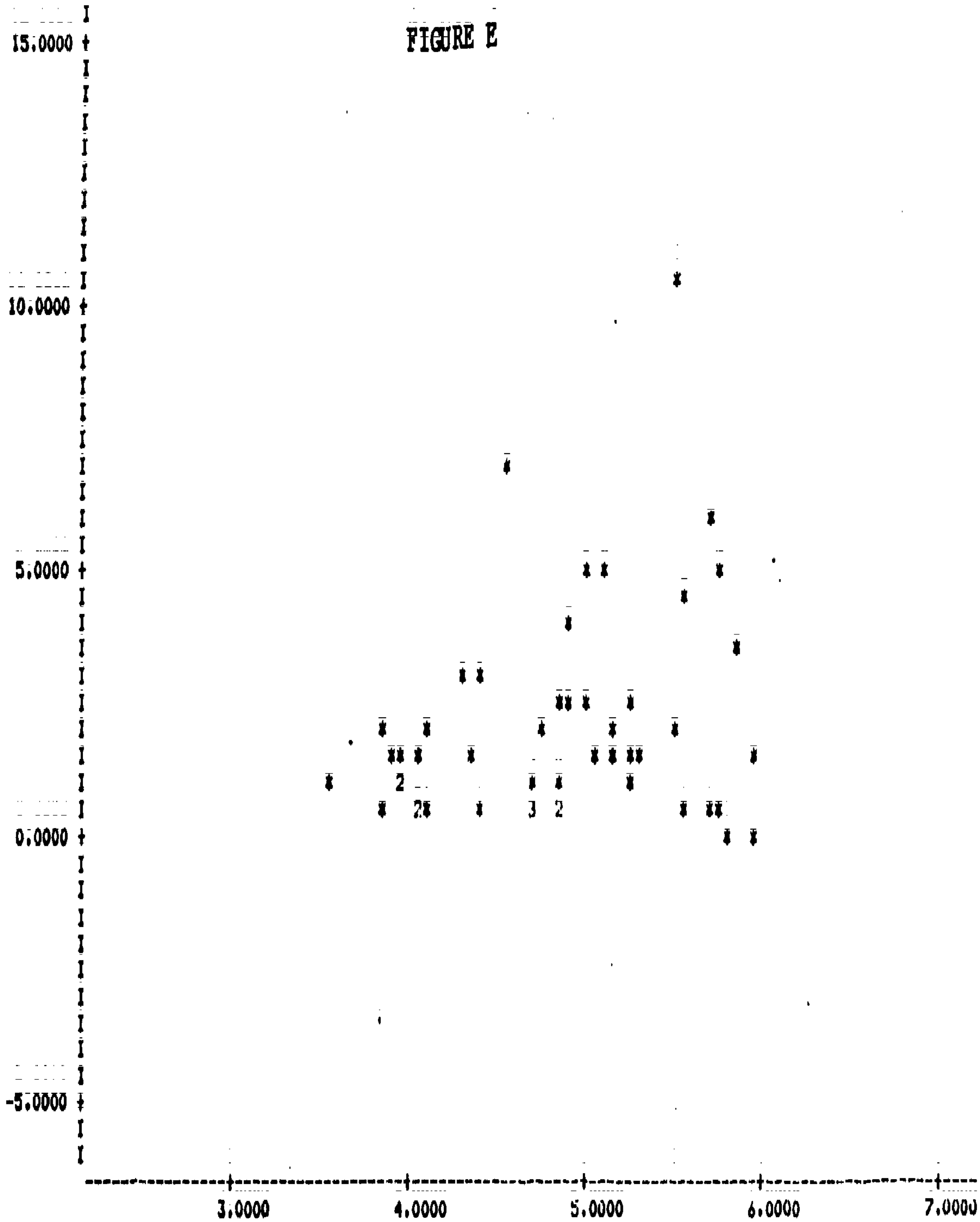
XVI.II.354

PLOT CONSUMP VS POPDENS



Y-AXIS: CONSUMP SCALE UNIT: 0.5000
X-AXIS: POPDENS SCALE UNIT: 10.0000

PLOT CONSUMP VS PCI



Y-AXIS: CONSUMP SCALE UNIT: 0.5000
X-AXIS: PCI SCALE UNIT: 0.0500

XVI.II.356

FIGURE F

```
MREG CONSUMP VS VEHIC GASTAX P PCI POPDENS SAVERES RESALL

  RESPONSE          MEAN    STD. DEV.
   CONSUMP        2.0340       2.0240

  CARRIER:      CONSTANT        VEHIC       GASTAX           P         PCI      POPDENS
COEFFICIENT       1.1935       0.7337      -0.0461      0.0396     -0.1614  2.2892E-05
 S.E. COEF.                    0.0459       0.0364      0.0268      0.0581       0.0002
T STATISTIC                   15.9777      -1.2687      1.4757     -2.7758       0.1237
       MEAN                    2.4743       7.6200      3.9025      4.8412     144.6380
  STD. DEV.                    2.5391       1.0079      4.3923      0.6606     219.2286

MULTIPLE R SQUARED        0.9875
```

FIGURE G

```
 STEM RESALL;BOXPLOT RESALL THREE

VARIABLE ' RESALL '.  UNIT =        0.0100

              LO I       -0.4151
    4         -2. I 966
              -2  I
    9         -1. I 88755
   17         -1  I 44333100
 ( 9)         -0. I 999986666
   24         -0  I 21110
   19          0  I 0111234
   12          0. I 6
   11          1  I 014
    8          1. I 59
               2  I
               2. I
    6          3  I 0

              HI I       0.3359      0.3831      0.4479      0.7011      0.7807

   SCALE UNIT:        0.0200

   -0.4000               0.0000              0.4000              0.8000
   -+-------------------+--------------------+-------------------+----------------

VARIABLE: RESALL

                       +---+---+
   *      <-------I    +   I------------> * * *            *   *
                       +---+---+
```

FIGURE H

```
MREG CONSUMP VS VEHIC PCI SAVERES RES2D

   RESPONSE          MEAN    STD. DEV.
    CONSUMP        2.0340      2.0240

   CARRIER:      CONSTANT       VEHIC        PCI
COEFFICIENT        0.7758      0.8013    -0.1496
  S.E. COEF.                   0.0141     0.0541
T STATISTIC                   56.9363    -2.7662
       MEAN                    2.4743     4.8412
  STD. DEV.                    2.5391     0.6606

MULTIPLE R SQUARED        0.9864
```

FIGURE I

```
STEM RES2D;BOXPLOT RES2D THREE

VARIABLE ' RES2D '.  UNIT =       0.0100

    1          -4  I 0
    2          -3  I 8
    6          -2  I 5532
   18          -1  I 644443322100
 ( 13)         -0  I 9976655543310
   19           0  I 0003468
   12           1  I 015557
    6           2  I 2
    5           3  I 07
    3           4  I 1
               HI  I        0.7505      0.8676

   SCALE UNIT:        0.0200

   -0.4000              0.0000              0.4000              0.8000              1.2000
     +--------------------+--------------------+--------------------+--------------------+-----

VARIABLE: RES2D

                         +---+------+
     <-----------------I    +      I----------------->                  *      *
                         +---+------+
```

```
MREG CONSUMP VS VEHIC SAVERES RESID
```

FIGURE J

```
  RESPONSE          MEAN    STD. DEV.
   CONSUMP        2.0340      2.0240

  CARRIER:      CONSTANT       VEHIC
COEFFICIENT       0.0774      0.7908
 S.E. COEF.                   0.0145
T STATISTIC                  54.6873
       MEAN                   2.4743
  STD. DEV.                   2.5391

MULTIPLE R SQUARED       0.9842
```

```
STEM RESID!BOXPLOT RESID THREE

VARIABLE ' RESID '.  UNIT =        0.0100

   2          -3. I 96
   3          -3  I 1
   4          -2. I 7
   7          -2  I 330
  10          -1. I 986
  13          -1  I 443
  22          -0. I 999988755
( 10)         -0  I 4444433311
  18           0  I 02233
  13           0. I 578
  10           1  I 134
               1. I
   7           2  I 2
   6           2. I 58
   4           3  I 3

              HI I       0.3780      0.8470      1.0353

   SCALE UNIT:        0.0200

  -0.4000            0.0000            0.4000            0.8000            1.2000
    +------------------+-----------------+------------------+------------------+-----
VARIABLE: RESID
                       +----+----+
    <------------------I    +    I-------------> *                    *        *
                       +----+----+
```

FIGURE K

734

4. Stockmarket Problem

(a) SHAREIND vs. CARPROD-LAG6 is plotted in Figure A and looks somewhat curved.

SHAREIND vs. COMMIND-LAG7 is plotted in Figure B. A somewhat linear pattern (with negative slope) is discernable.

SHAREIND vs. TIME (DM) is plotted in Figure C. A strong linear (although jagged) pattern is clear.

Note that since we are dealing with a model based on a specific and well defined theoretical model we would NOT perform any transformations on Y or the X's.

(b) The MREG output is shown in Figure D.

(i) Causality is not really an appropriate question here, since we are not inferring that Y is "caused" by the independent variables, but rather that Y can be "predicted" via these variables. Using car production and the commodity index (with suitable time lag for effect) to "measure" the health of an economy may indeed be a reasonable economic theory.

(ii) The independent variables may indeed be related (for example, we might expect car production to increase with population and hence with time). Other interdependencies should be investigated via economic theory.

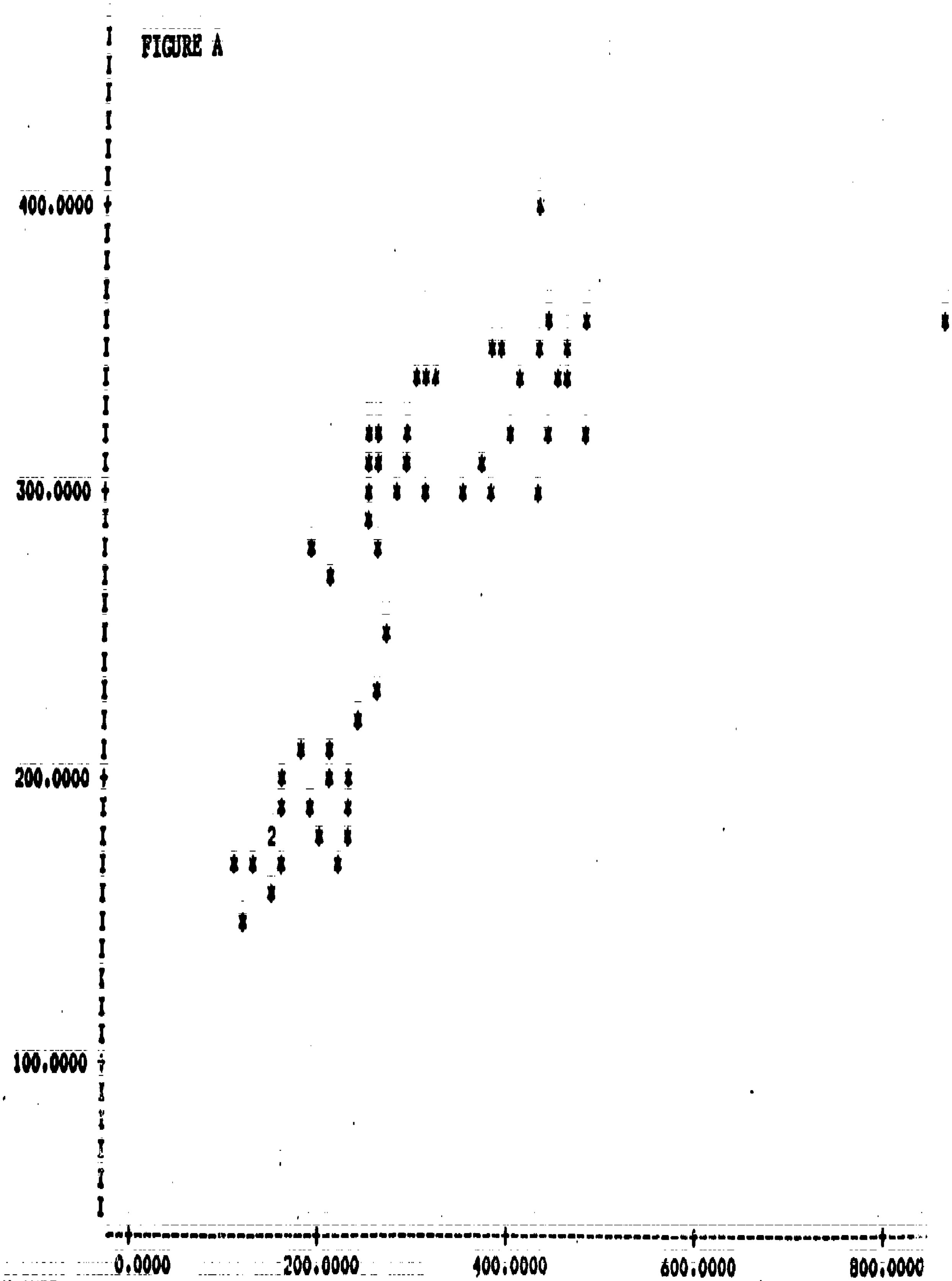
(c) A stem-and-leaf plot of the residuals is shown in Figure E. They are fairly well behaved (although we should note the large number of leaves on the -2 stem, and on stems ≥ 4). A plot of the residuals vs. time suggests a cycle pattern (Figure F). Plots of residuals vs. CARPROD-LAG6 (Figure G) and COMMIND-LAG7 (Figure H) are more "random".

(d) The MREG output shows $R^2 = .8295$. This suggests that the model provides a "good" explanation of the data.

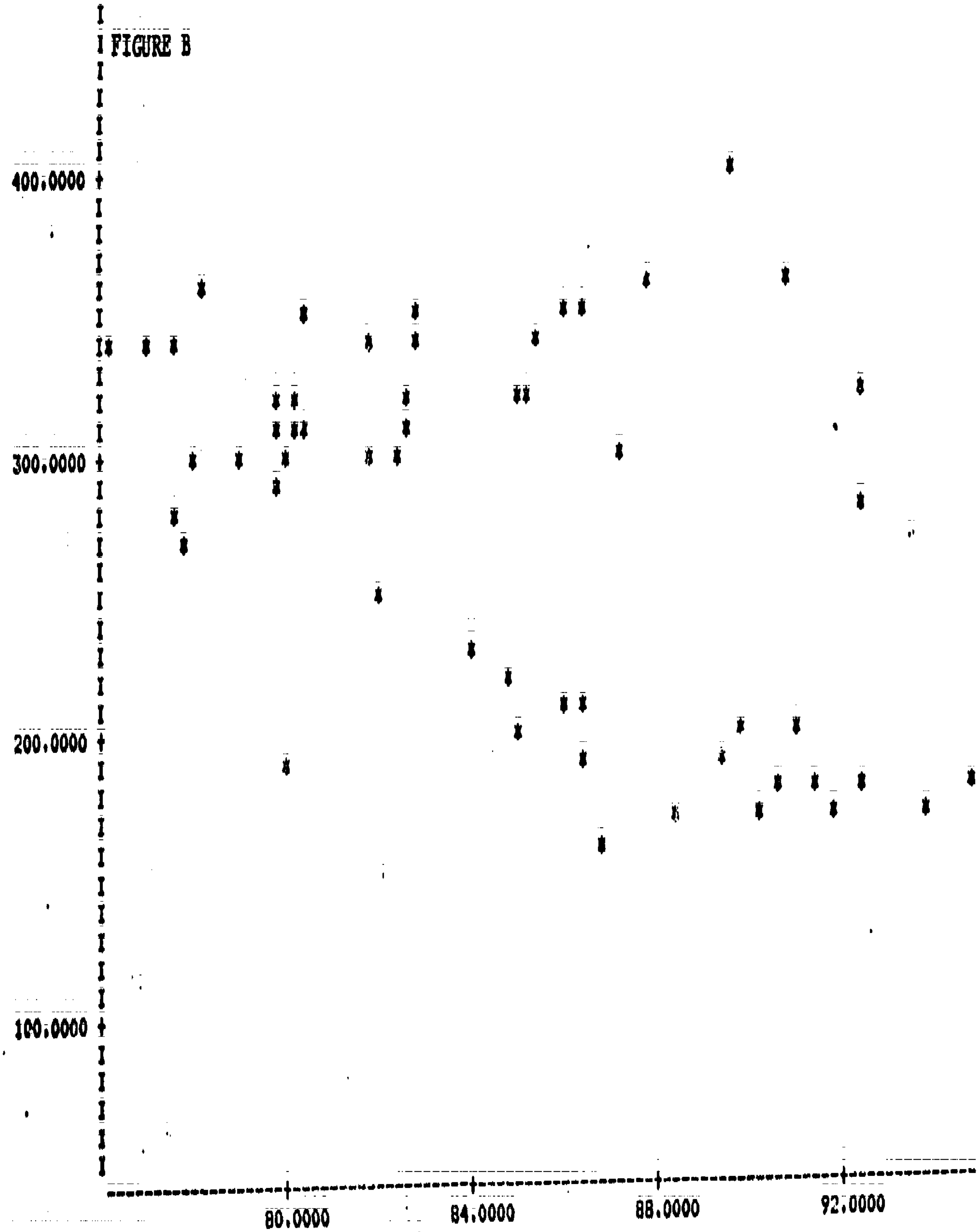
(e) Your answer to this part will depend in part on your degree of risk aversion. For most people, the model is not sufficiently accurate, nor does it inspire enough confidence, to be used as a basis for "playing the market". Those who are risk-seeking (i.e. "gamblers" or just "sportive") might be willing to give this model a go in an attempt to "beat the system".

735

FIGURE A



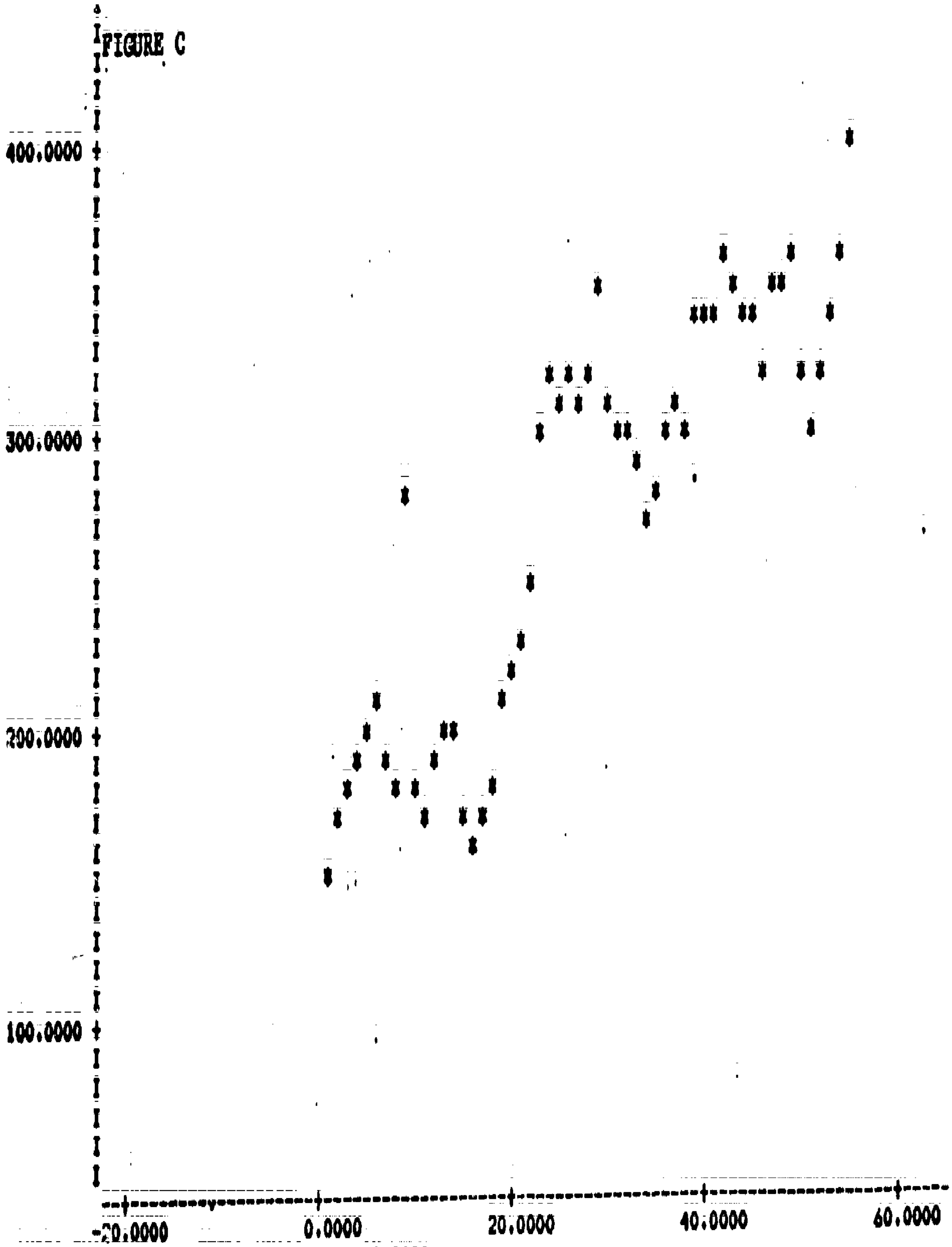
Y-AXIS: SHAREIND SCALE UNIT: 10.0000
X-AXIS: CARPROD-LAG6 SCALE UNIT: 10.0000



Y-AXIS: SHAREIND SCALE UNIT: 10.0000
X-AXIS: COMMIND-LAG7 SCALE UNIT: 0.2000

XVI.II.362

FIGURE C



Y-AXIS: SHAREIND SCALE UNIT: 10.0000
X-AXIS: TIM SCALE UNIT: 1.0000

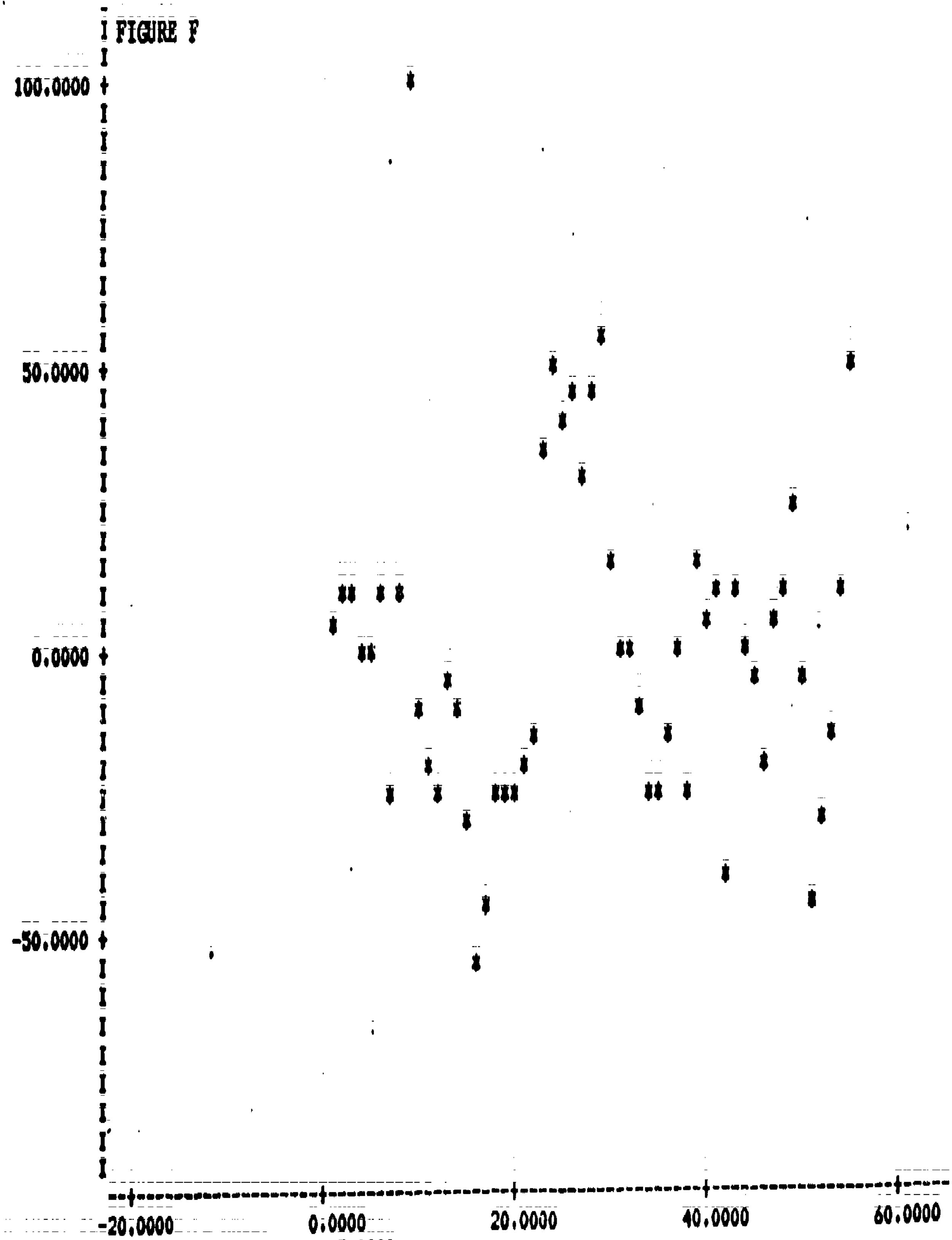
FIGURE D

```
MREG SHAREIND VS TIM CARPROD!-LAG6 COMMIND!-LAG7 SAVERES RES4
   RESPONSE              MEAN     STD. DEV.
   SHAREIND          272.6689       70.5904

   CARRIER:          CONSTANT           TIMCARPROD-LAG6COMMIND-LAG7
COEFFICIENT          425.8179        2.6257        0.1268       -3.1126
  S.E. COEF.                         0.4340        0.0504        0.8503
       MEAN                         28.0000      295.8901       84.8747
  STD. DEV.                         16.0208      131.7769        5.2503

MULTIPLE R SQUARED          0.8295

ANALYSIS OF VARIANCE TABLE
                          SS      DF            MS           RMS
FIT              223190.6250       3    74396.8750      272.7578
RESIDUAL          45891.4766      51      899.8328       29.9972
TOTAL            269082.1250      54

                           F       F PROB.
FIT                  82.6786        0.9966
:
```

FIGURE E

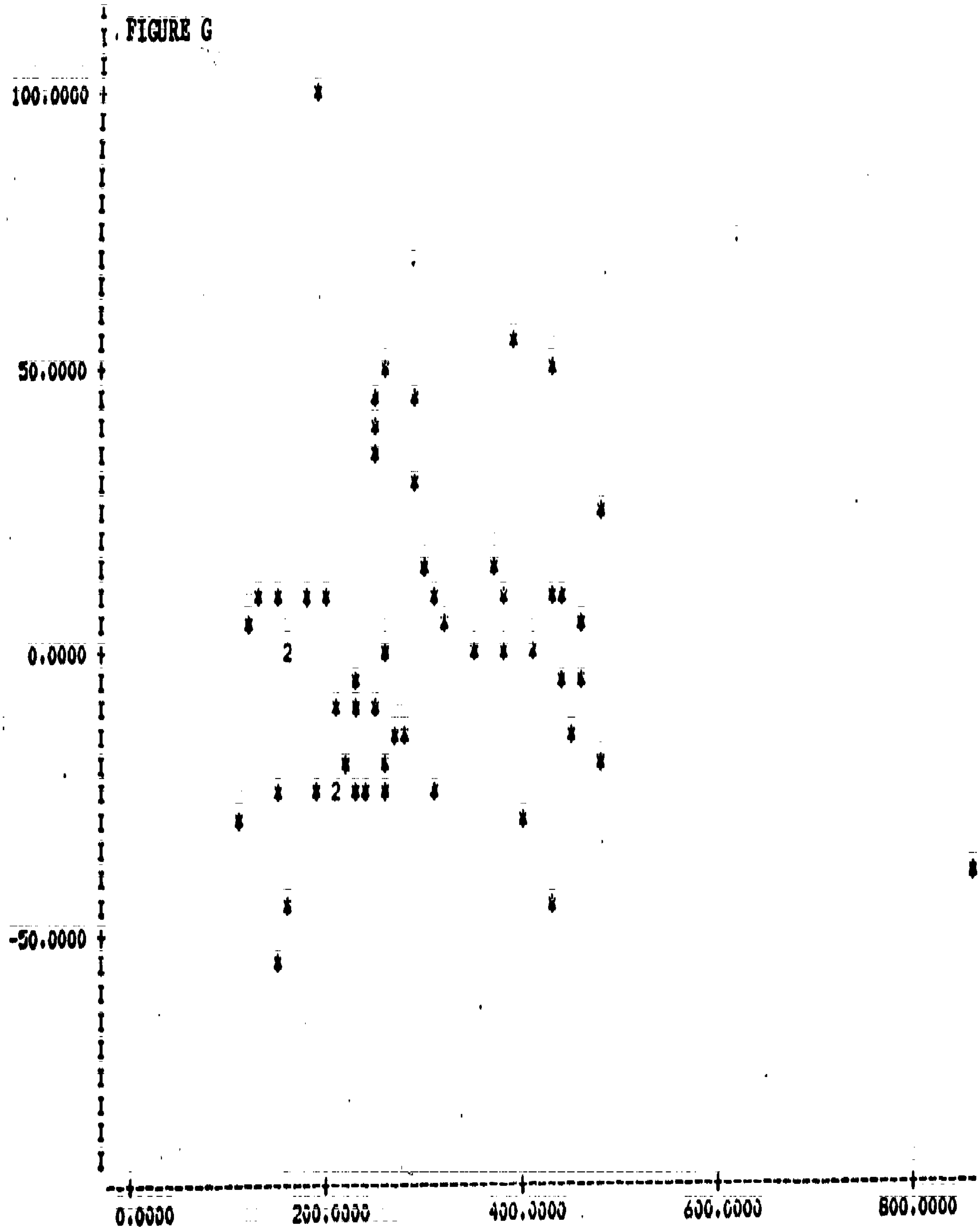
```
 STEM RES4

VARIABLE ' RES4 '.
 UNIT =          1.0000
     1          -5   I 5
     4          -4   I 431
     5          -3   I 0
    16          -2   I 97766554320
    23          -1   I 9442210
(    5)         -0   I 54420
    27           0   I 0012246889
    17           1   I 0000056
    10           2   I 79
     8           3   I 2
     7           4   I 055
     4           5   I 016
                HI   I       98.2473
:
```

742



Y-AXIS: RES4 SCALE UNIT: 5.0000
X-AXIS: TIM SCALE UNIT: 1.0000

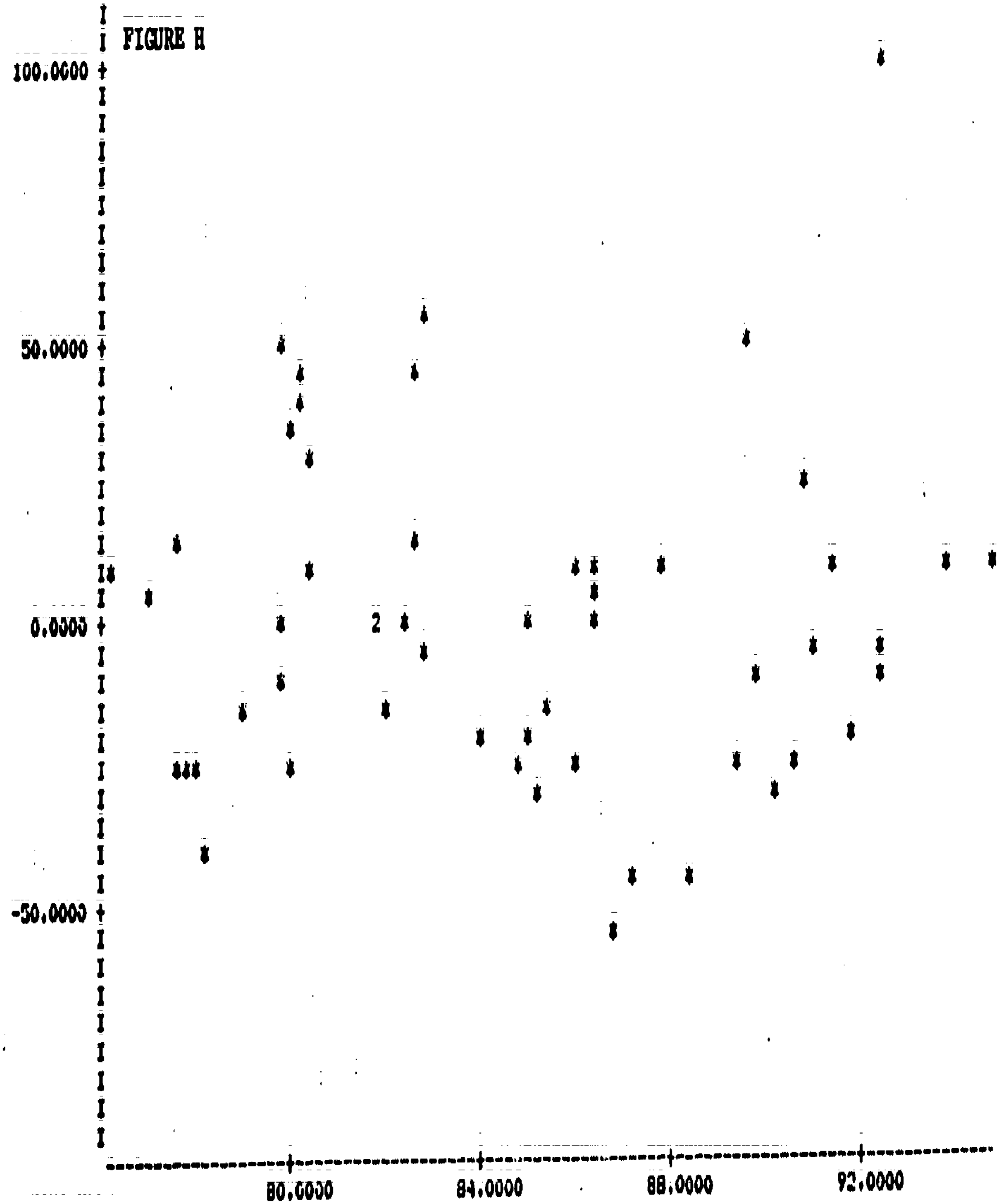
FIGURE G



Y-AXIS: RES4   SCALE UNIT: 5.0000
X-AXIS: CARPROD-LAG6   SCALE UNIT: 10.0000

XVI.II.366

FIGURE H



Y-AXIS: RES4 SCALE UNIT: 5.0000
X-AXIS: COMMIND-LAG7 SCALE UNIT: 0.2000

Quiz, Unit 4

You have 30 minutes to complete the quiz. Answer all questions, but **briefly**.

Background

You are working for the federal Department of Housing and Urban Development (HUD). The department has just completed a project in which \$1.3 million was allocated to developers for the purpose of constructing new graduate level public management curricula in three "need" areas. Solicited proposals were scored by experts. The approximately 200 proposals received were divided equally in the three need areas. The maximum obtainable score was 81 and the lowest was 0. To determine how "fair" the granting procedure was HUD has asked you to perform an evaluation. You gathered data on various characteristics of the proposals, constructed a model and used least squares techniques to estimate parameters. The results appear below.

Dependent Variable: Total Score for Proposal

Independent Variable	Coefficient Estimate	t-statistic
Length (in pages)	.74	5.55
Budget Request (in dollars)	.03	.21
Need area 1	-7.11	3.10
Need area 2	1.80	.76
Constant	47.87	

$R^2 = .21$

Explanation of Variables

Length - number of pages; Min = 6 Max = 58.

Budget Request - how much money was requested to do the task.

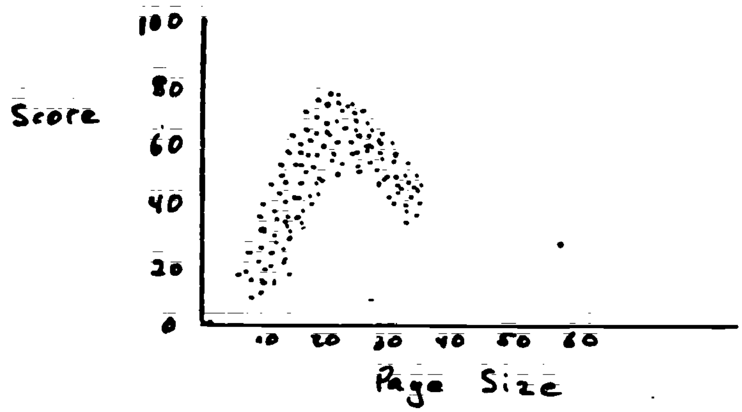
Need areas 1, 2, 3--areas of training in which HUD believes new curricula are required. Introduced as 0/1 indicator variable with following structure:

Need Area:	1	2	3
Variable Values:	1	0	0
	0	1	0

Problems

1. Why do the regression results imply that quality was not the only factor influencing the selection process? 20 points

2. The following is a plot of page size against total score.



a. Why does this plot suggest that the regression results are suspect? 10 points

b. Give two alternative forms for the page variable that would help summarize its behavior better. 10 points

3. HUD has been under attack for favoring schools in the Northeast over other areas of the country. How would you test this proposition using these data, assuming you know which institutions submitted the proposals? 20 points

4. a. What is the "effect" of a proposal's being in need area 3? 20 points

b. What is the estimated standard error of the length coefficient? 20 points

Quiz, Unit 4
Solutions

1. Both page length and curriculum need area one had significant coefficients indicating that these had an effect on the selection process.

2a. There is curvature evident in the variable for length of proposal; the relationship is not strictly linear. Linear regression assumes a linear relationship between the dependent and each independent variable.

b. A parabolic fit of the form $p + p^2$ or using splines would have been appropriate alternative forms.

3. The proposition could be analyzed through the use of dummy variables indicating the region of the institutions submitting the proposals.

4a. The "effect" of proposal being classed in curriculum need area 3 was the constant, i.e., 47.87. The coefficients for need areas 1 and 2 make adjustments to the constant.

b. $\frac{\text{coefficient}}{\text{standard error}} = \text{t-statistic}$

$\frac{\text{coefficient}}{\text{t-statistic}} = \text{standard error}$

$\frac{.74}{5.55} = .133$

751

Final Examination
First Term

Name ____________________________________

All answers should be written on th. test. Total point score is 100. Point value for each question are indicated. Do not spend all your time on questions with low point values! You have 2.5 hours to complete this examination. Your answers should be concise and to the point.

For the following problems assume that you are a public manager in a government agency.

(1) When you were first interviewed for this job, your supervisor could not understand why a public manager needed to take QMPM. You answered her by making four points. Circle them: 4 pt.

a. Knowledge of statistics improves rational th.king.

b. A good knowledge of statistics makes you a better liar.

c. Public managers are essentially decision makers.

d. You need a graduate education in statistics to read the New Yo Times.

e. M public policy analyses are unplanned and post hoc.

f. Public managers need to be able to perform, interpret and present quantit. ive analyses.

g. Knowledge of skills in data analysis assures a GS-11 rating.

h. More often then not, data relevant to public policy decisions are quantitative and "messy".

(2) The first task you encounter on the job involves an analysis of physician offices by census tract in urban areas. The policy issue concerns the equity of access by urban residents to physicians.

a. To begin the analysis you ask to see a batch of physician office data for a single city. Your supervisor asks you what a batch is. You reply.... 2 pt.

752

b. To get a feel for the "average" number of physicians per tract in this batch she suggests that you calculate the batch mean. You counter by suggesting that you calculate the median or even the mode. She asked what advantage these have over the mean and you replied.... 3 pt.

c. Her next request concerns variation in the data. "Calculate the variance", she says. But you hestiate and calculate the H-spread. How are they related, and when would you not hesitate to calculate the variance? 4 pt.

d. Impressed with your knowledge she asks you to draw a histogram. You construct a stem-and-leaf display instead. What is one essential difference between these two kinds of displays? 2 pt.

e. The stem-and-leaf shows that physician offices are frequent in only a few census tracts, and that, by and large, most tracts have only a few offices and many have none. Sketch an outline of what the stem-and-leaf of this data probably looks like. 2 pt.

f. When you see the stem-and-leaf of the physician data, you immediately suggest re-expressing. What might a re-expression achieve? 2 pt.

g. When you talk about this problem, you casually mention the simple ladder of powers. What is the simple ladder of powers, and why is it relevant to this problem? 4 pt.

h. She then asks how you decided which transformation to use, and you tell her about a general technique. What is it? 2 pt.

i. Since physician offices are counted data, you had in mind a specific transformation before you even try this method for finding a transformation. What transformation was it that you had in mind? 2 pt.

j. A colleague suggests looking at the data differently. Instead of counts he suggests dividing the number of offices in each tract by the total number of offices in the city. This would yield a variable that was the proportion of a city's physician offices in each tract. What transformation would you be likely to try on the data when they are in this form? 2 pt.

k. Elated with your thorough analysis of the single batch of physician data, another colleague suggested that you look at data from several cities simultaneously and compare them. Knowing that cities vary quite a bit in size you thought that re-expression would be needed. Why might you need to transform the data for this comparison, and how would you find the appropriate transformation? 4 pt.

(3) Pursuing the analysis of office location, you decide to see if you could construct a model relating the number of offices in a tract to other public policy relevant features of the census tract.

a. In constructing the model you are told to include median income and socioeconomic status as independent variables, but you suspect that these variables are highly correlated. This means that 1 pt.

(a) There is a linear relationship between them.

(b) Their covariance is positive.

(c) They are related in a curvilinear fashion.

(d) The covariance equals the product of the standard deviations.

b. You expect to use least squares techniques to estimate your model. Consequently, you suspect that two problems may arise because of these correlated variables. 2 pt.

(a) R^2 will be near unity.

(b) Their coefficient estimates will be unreliable.

(c) The residuals will be random.

(d) $\underset{\sim}{y}'$ will not invert.

(e) The computer may have problems with $(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$.

(f) The coefficient of determination will be indeterminant.

c. To get around these problems you suggest two possible solutions. 2 pt.

(a) Use only one of the pair.

(b) Add one to each variable and take logs.

(c) Use weighted least squares.

(d) Use ridge regression.

(e) Use the arc-sin square root transformation.

755

d. When you tell your supervisor about your intentions to use least squares to estimate the model she asks what this means. Your reply is that it uses one specific minimization criterion which is: minimize ..... 1 pt.

(a) $\Sigma(Y_i - \hat{Y}_i)$

(b) $\Sigma|Y_i - \hat{Y}_i|$

(c) $\Sigma(Y_i - \hat{Y}_i)^2$

(d) $\Sigma(Y^2 - \hat{Y}_i^2)$

(e) None of the above.

e. You continue your explanation by saying that, "If the assumptions underlying least squares hold, then this procedure yields optimal estimates of the coefficients". What are the assumptions? 3 pt.

f. "But in what sense are least squares regression lines optimal? she asks. You reply... 3 pt.

g. "OK", your colleague pipes up, "so they are optimal when the assumptions hold. But suppose that for our data the assumptions don't hold. What does that imply?" In your response you touch on the consequences of failure in each assumption. What do you say? 4 pt.

(4) At this point you have established your credentials as a competent policy researcher. Word passes quickly through offices in an agency, and when you turn around you meet an economist from down the hall. He explains that he, too, is working on a physician study but he has been investigating the growth of the physician supply in different countries. He says he is going to fit a straight line to the data from five countries and contrast slopes. You are aghast and demand to see plots of supply versus time. He shows you the five plots below and you know that you are right again. Instead of simply cranking the data through you counsel fitting different models to summarize each batch of paired values. Which alternatives below do you suggest for each of the plots? 10 pt.

i) logged dependent variable

ii) logged independent variable

iii) logged dependent and independent variable

iv) continuous spline

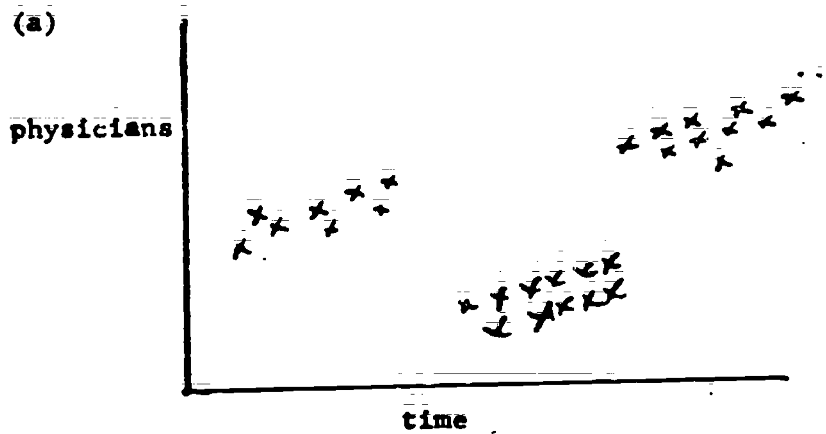
v) discontinuous spline

vi) straight line linear model

vii) linear model after removal of outliers

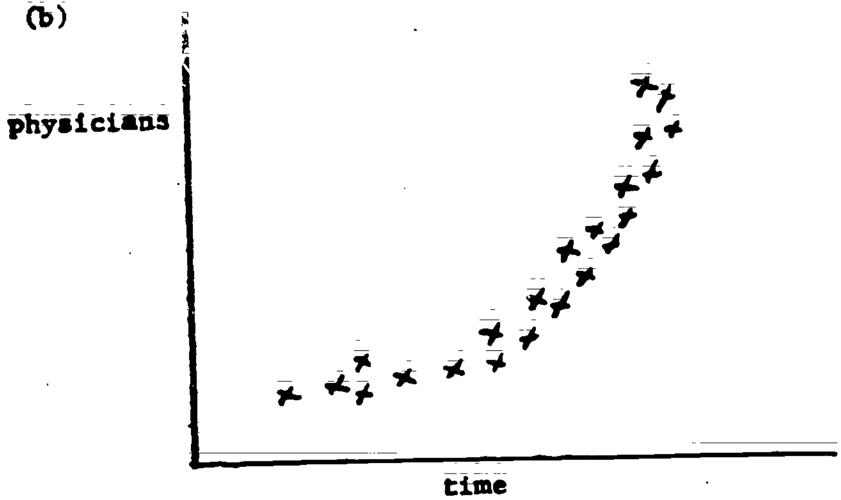
viii) linear model with dummy variable(s)

(a)

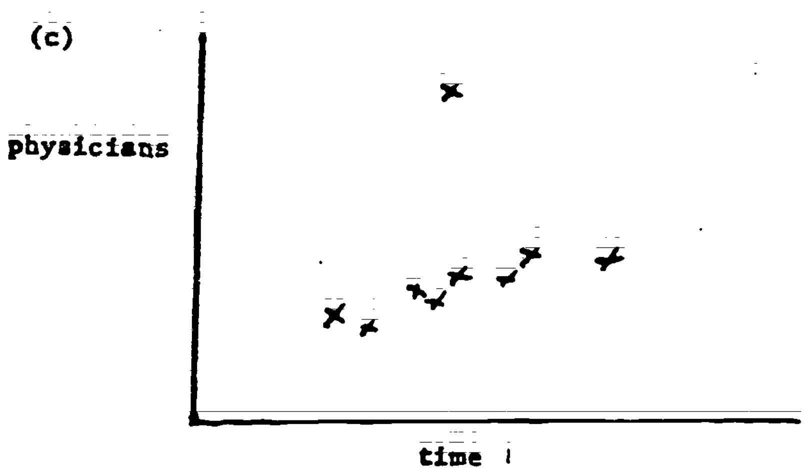


757

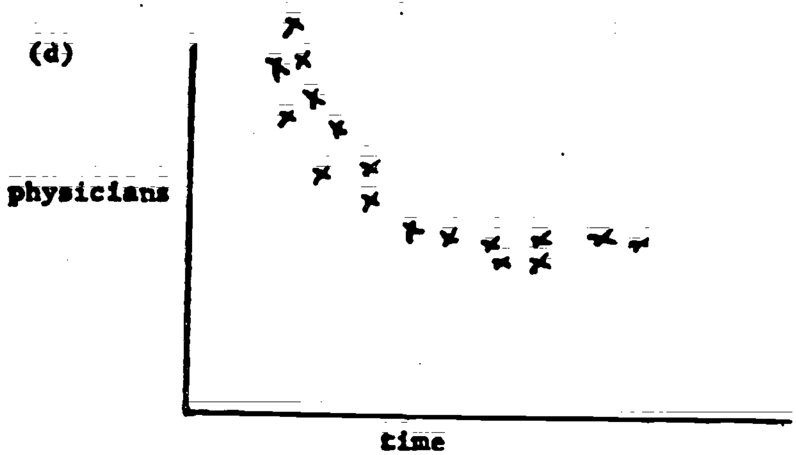
(b)



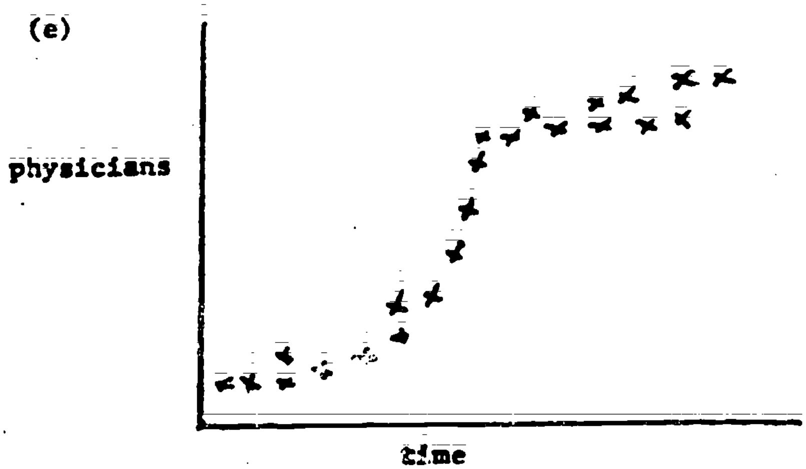
(c)



(d)



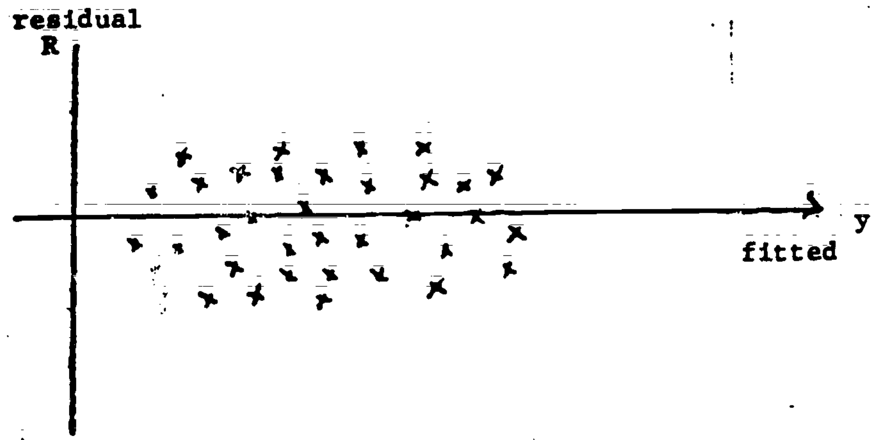
(e)

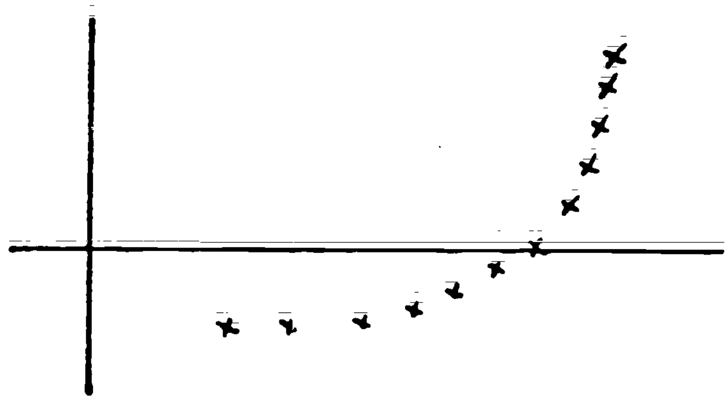


(5) Very concerned about this affront to his knowledge in the presence of his peers, the economist explains that it doesn't make any difference. Retaining your cool, you politely disagree and for each of the five cases you describe how a simple linear fit and the fits you've suggested would differ. Your explanations focus on what the residuals from the simple fits would look like. <u>10 pt</u>.

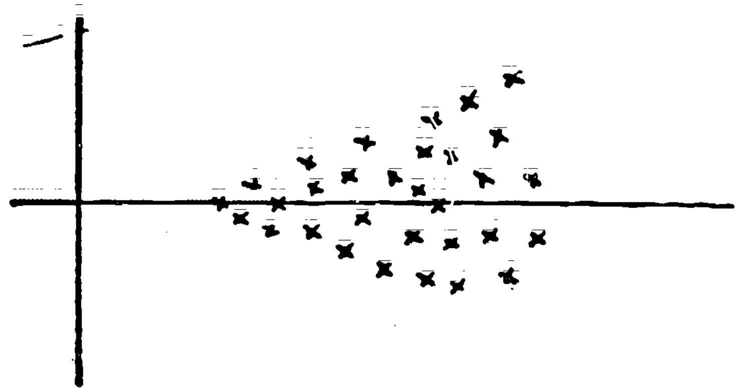
759

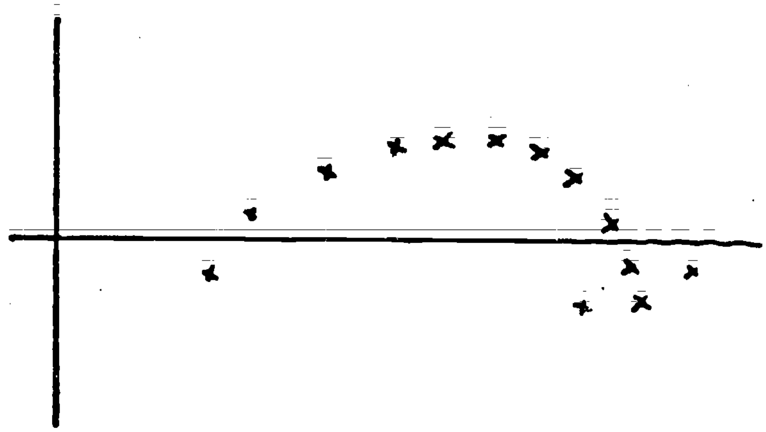
(6) Impressed by your knowledge your colleague admits to having performed regressions automatically before. He shows you the following residual plots and asks what he should do next each. 10 pt.

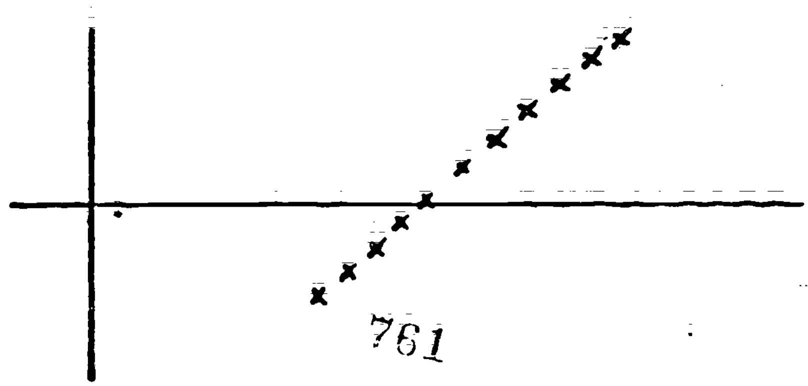




760



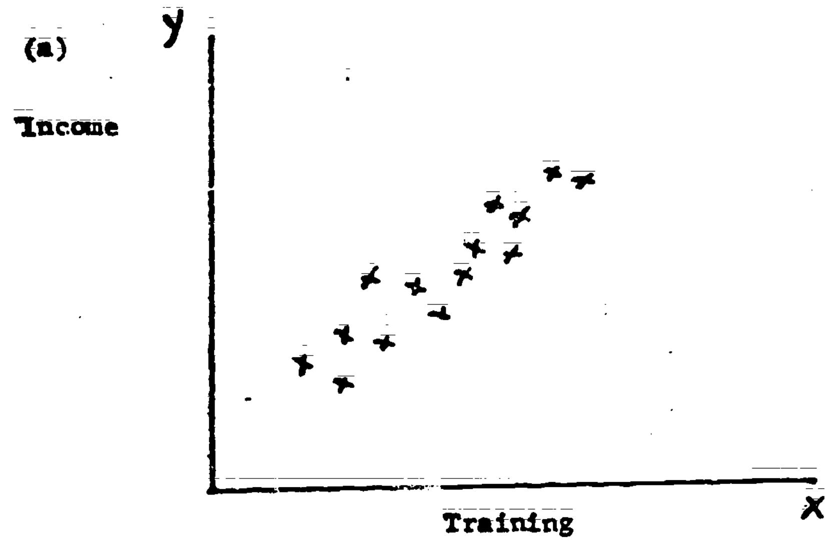




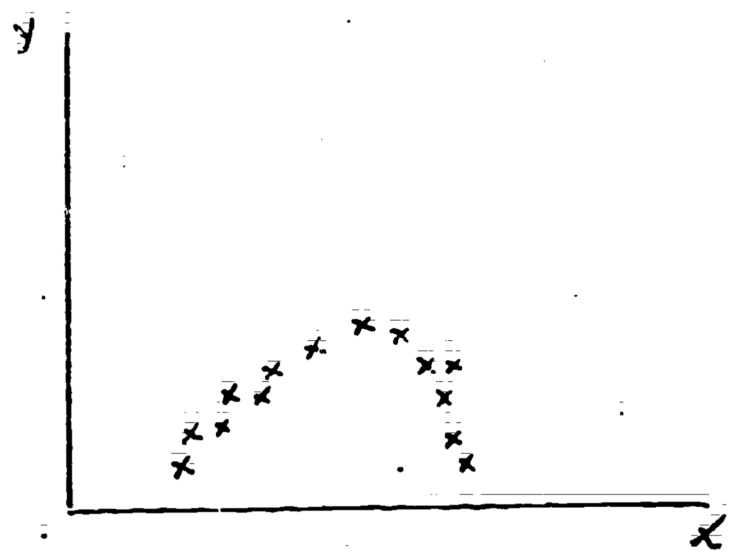
761

(7) The same colleague tells you of a study of income and years of physician training in different countries. He shows you the following plots. Since each is a univariate situation you suggest resistant lines. But he doesn't understand and you sketch (on the plots) how least squares and resistant lines would look in each situation. 10 pt.

(a)

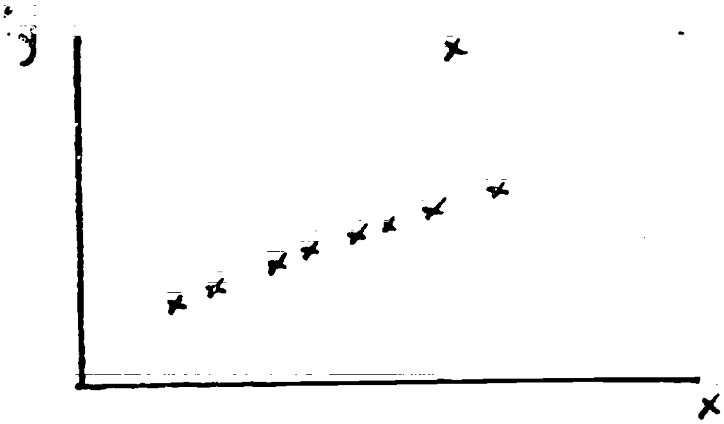


(b)

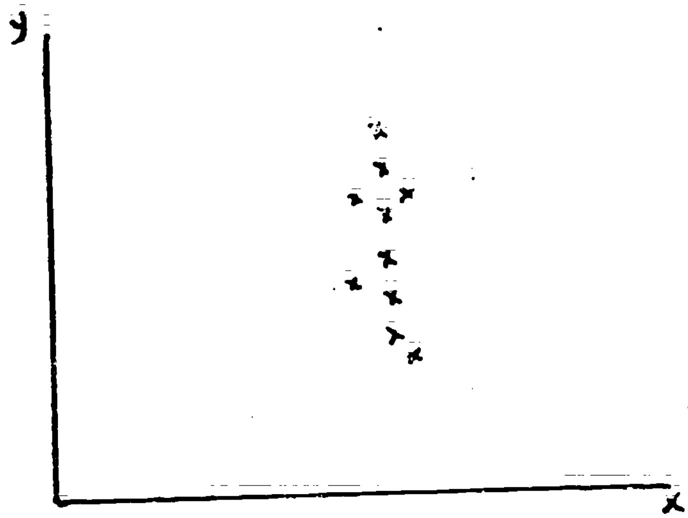


762

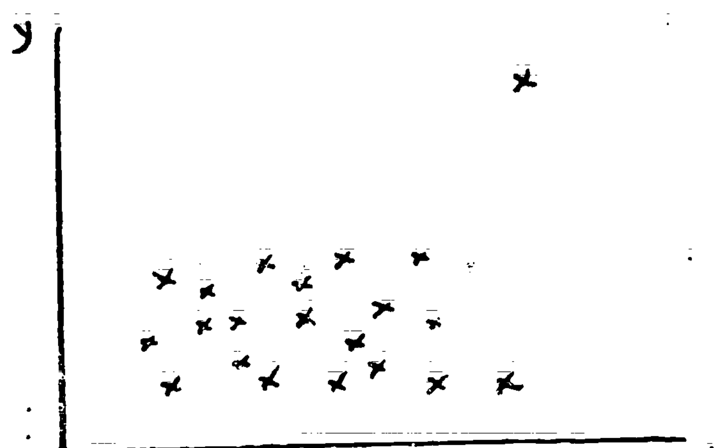
(c)



(d)



(e)



(8) Once you got rid of your now better informed colleague you are once again able to concentrate on your own problem, physician location in cities. Someone draws your attention to an article on physician location by the famous scientific research team of Kaplan and Leinhardt. You know this paper and recall its presentation and findings. In particular, you recall that they found little or no effect for an area's income and racial characteristics. In the paper they resent the results of their regression runs and imply that this should convince. Were you convinced? If so, why? If not, what else would you want to see before you were convinced? 6 pt.

(9) You also recall that they included variables that had policy relevance from two points of view. What are these points of view and what are examples of variables in each category? 4 pt.

Final Examination Solutions
First Term

(1) Correct answers are (c), (e), (f), and (h), although Huff emphasizes that (b) may be true.

(2)a. A batch of data is a set of similar numbers, obtained in some consistent fashion.

b. Each average, the mean, median, and mode, is a typical value for a batch. The mean is a perfectly good average if the batch is well-behaved, or even just symmetric. However, unlike the median, if there are discrepant or outlying observations in the batch, the mean is very sensitive to these departures from "well-behavedness", since we must sum all the observations to find the mean. In the batch given to you by your supervisor, the shape will not be symmetric because of many tracts with zero physicians; rather, the shape will be skewed to the right. Hence, the median, the middle observation of the batch, or perhaps the mode, the data value with the largest frequency, will be more typical of the batch than the mean.

c. The H-spread, the difference between the hinges, and the variance are both measures of spread. In a well-behaved batch,

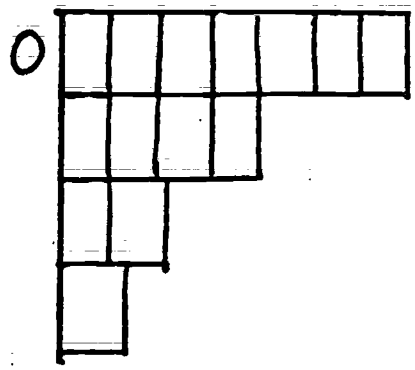
$$3/4 \cdot \text{H-spread} = \sqrt{\text{variance}}.$$

The H-spread is a more resistant measure of spread than the variance. Only with a well-behaved batch is the variance an acceptable measure.

d. A stem-and-leaf display retains additional information on the data values by using the digits immediately to the right of the stems to indicate "heights" or frequencies of each stem. With a histogram, only the heights of the bars are indicated, and the leaf digits are discarded.

765

e. Unit = 1%



or values with this shape.

f. A reexpression of these data will promote symmetry within the batch and perhaps decrease the number of outlying values.

g. The ladder of powers is a graphical representation of reexpressions of the form $X \rightarrow X^R$ for values of R such as -1, -1/2, 0(=log), 1/2, 1, 2. We can use the ladder to determine the effects that various reexpressions will have on the original batch, e.g., a batch skewed to the right has a long right tail, and a transformation down the ladder, $X \rightarrow X^{1/2}$ or log X, will force more of the observations into the right tail of the data, while making the skewness less noticeable.

h. Examine the mean, median, and midhinge = 1/2 (UH + LH), and perhaps the mid extreme = 1/2 (E + E). In a symmetric batch, these quantities are equal. By reexpressing the 5-number summary of the batch, you can compute the 4 quantities and examine their equality/inequality and hence, find the best transformation. Hint:

If med < midsp < midext., go down the ladder.
If med > midsp > midext., go up the ladder.

i. Square root, or perhaps logarithms, moving down the ladder.

j. Square root of the arcsine of proportion physician offices in each tract.

k. One suggestion is merely to examine the proportion of physician offices in each tract in each city, since these data are independent of the size of the city. If you desire to work with the actual numbers and discover that the spread of the batches increases or decreases as the location of each batch changes, then a log median vs log midspread plot will reveal a reexpression that stabilizes the spreads.

(3)a. Correct response is (a).

b. Correct responses are (b) and (e).

c. Correct responses are (a) and (d).

d. Correct response is (c).

e. Assumptions are

(1) The model is correct, i.e., y is a linear function of the x's.
(2) Residuals are independent.
(3) Residuals are homoscedastic.
(4) Residuals are well-behaved.

f. Out of all linear unbiased estimates, the least squares line is the one with minimum variance. Least squares lines are optimal only if the 4 assumptions hold.

g. (1) If the model is not correct the regression coefficients do not estimate the true population coefficient values. Moreover, none of the computed regression statistics are believable.

(2) Nonindependence of the residuals indicates that the observations are related. The R^2 statistic will not measure the goodness of fit, and the standard errors of the regression coefficients will not be accurately computed.

(3) Heteroscedastic errors have the same effect on the computed regression statistics as nonindependence of errors.

(4) Non-well-behavedness of residuals invalidates various distributional assumptions, e.g. batches of regression coefficients will not be well behaved.

(4)a. Dummy variable.

b. Logged dependent variable.

c. Straight line linear model after removal of outlier.

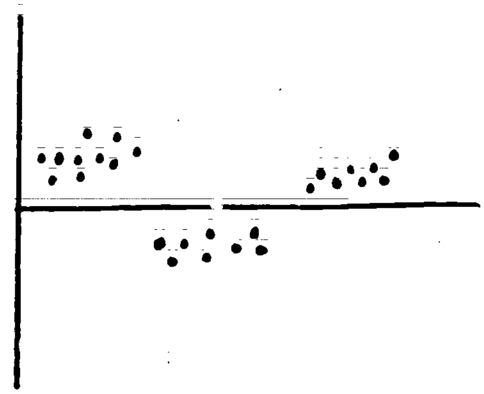
d. Logged dependent or logged independent variable, or both if necessary.

e. Continuous spline.

(5) For each of the suggested fits, the residuals, plotted against time, should appear as a random swarm of points, centered on the time axis, and as a well-behaved batch when displayed in a stem-and-leaf.

The residuals of the fits using the original data will exhibit various patterns, as shown below:

(a)



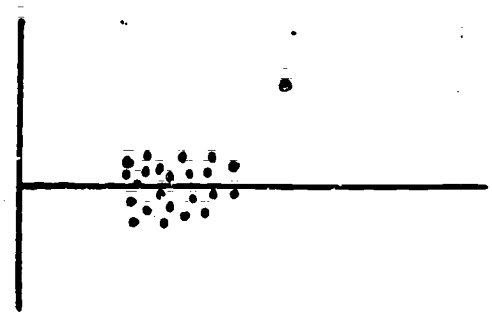
3 clusters of residuals

(b)

quadratic pattern

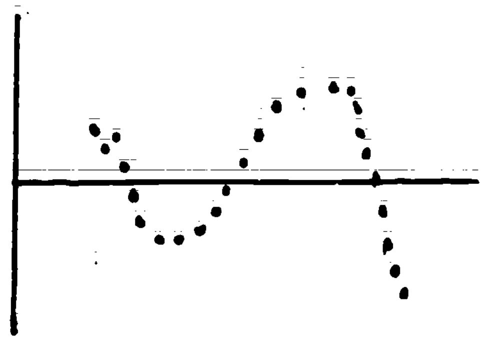
768

(c)



(d) same as b.

(e)



trigonometric pattern

(6)a. Do nothing.

b. Transform: X up the ladder, or Y down.

c. Transform Y to remove heteroscedasticity, or used weighted least squares.

d. Fit a quadratic to the data:

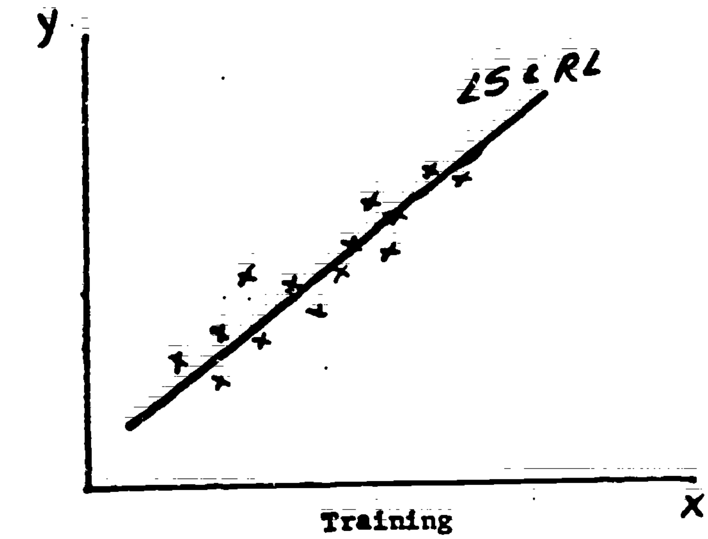
$$y = a + bX + cX^2$$

e. Computations probably incorrect. Make sure that X has been included in the model, where X = variable of x-axis.

769

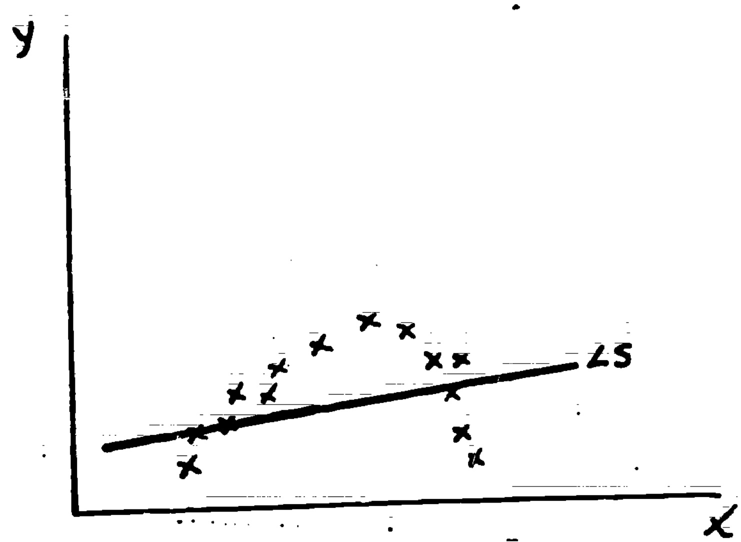
(7)

(a)



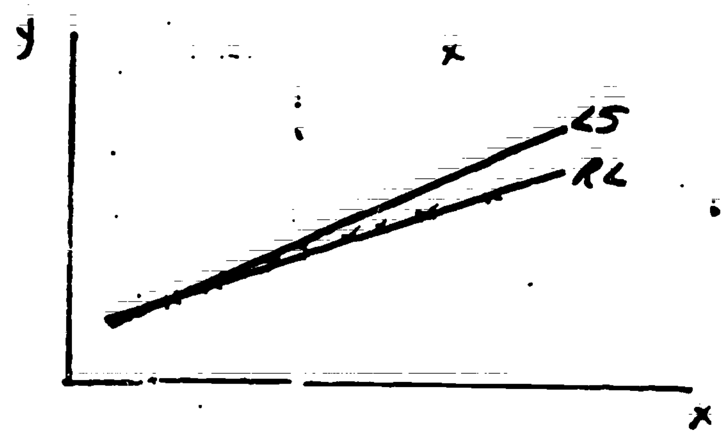
Income

(b)

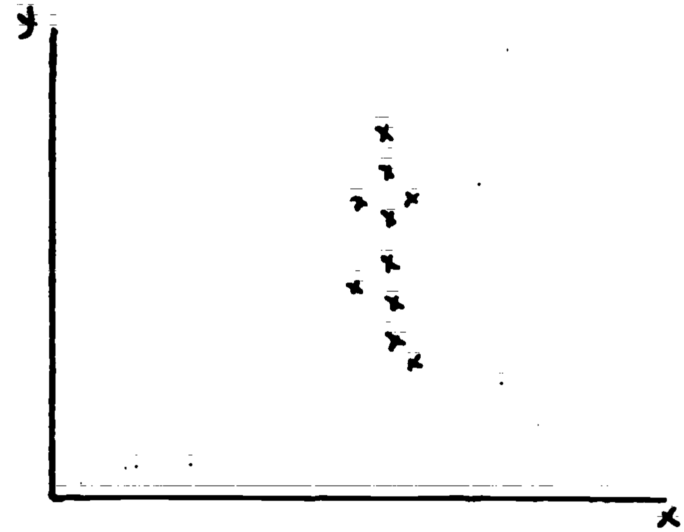


We would not fit a linear resistant line but would fit a parabola instead.

(c)

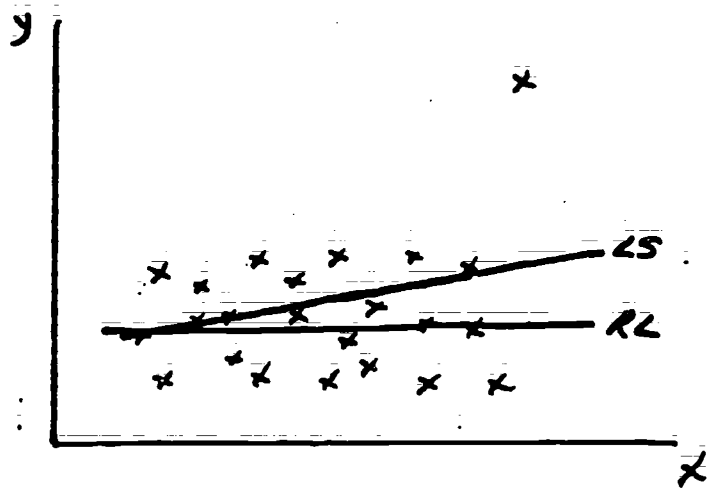


(d)



Both lines are undefined since there is not enough variation in the X variable. However, different computer programs for least squares regression will fit different lines.

(e)



771

(8) Kaplan and Leinhardt studied a rather controversial issue and their results, as indicated in the exam questions, contradicted most expectations. As the exam questions also pointed out, however, they reported only the regression results. That is, they gave coefficient values and t-statistics and expected the reader to be convinced by their results. In any analysis and especially one involving a highly controversial issue, we would want more information on the data analysis procedure. The issue of belief here is one of an evaluation of the effectiveness of the analysis. Thus, we would want to see such things as a correlation matrix to see if collinearity occurred, stem-and-leaf and plots of residuals to detect heteroscedasticity, non linearity and non-well behavedness, discussion of possible interactions and plots of independent variables against the dependent variable to explore for needed transformations. A general discussion of the exploratory phase should always appear but actual exploratory results are especially important in this case because policy may be influenced by the analytic results and if the results are due to poor analysis the policy may do more harm then good.

(9) Policy variables may be of two types. On the one hand, they may be "policy manipulable", i.e., we may be able to construct public policies which actually change these variables. On the other hand, they may be "policy directive", i.e., while not actually amenable to manipulation by policy they may guide policy development by focusing attention on special situations or target groups. Examples of the former variables are zoning regulations, hospital beds, physician offices, education, income. Examples of the latter are race, age, population. It should also be pointed out that if the policy issue is one involving the location of individuals even policy directive variables may become policy manipulable variables. For example, we can change the age distribution of a census tract by erecting an apartment house solely for senior citizens.

Handout
Covariances and Independence in the Bivariate Multiple Regression Model

Consider the situation of two "independent" x variables:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

The least squares solution for $\underset{\sim}{b}$ is

$$\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}$$

To make this result easier to comprehend, center the variables by subtracting their means:

$$X_{i1} - \bar{X}_1, \; X_{i2} - \bar{X}_2, \; Y_i - \bar{Y}$$

This shift of location forces the line to pass through the origin and, therefore, to have a y intercept of 0. Thus, $b_0 = 0$ in this "new" model and we do not need a column of ones in the $\underset{\sim}{X}$ matrix.

$\underset{\sim}{X}$ becomes

$$\begin{pmatrix} X_{11} - \bar{X}_1 & X_{12} - \bar{X}_2 \\ X_{21} - \bar{X}_1 & X_{22} - \bar{X}_2 \\ \vdots & \vdots \\ X_{N1} - \bar{X}_1 & X_{N2} - \bar{X}_2 \end{pmatrix}$$

and $\underset{\sim}{X}'$ is

$$\begin{pmatrix} X_{11} - \bar{X}_1 & X_{21} - \bar{X}_1 & \cdots\cdots & X_{N1} - \bar{X}_1 \\ X_{12} - \bar{X}_2 & X_{22} - \bar{X}_2 & \cdots\cdots & X_{N2} - \bar{X}_2 \end{pmatrix}$$

773

$\underset{\sim}{X}'\underset{\sim}{X}$ is the product of these two matrices:

$$\underset{\sim}{X}'\underset{\sim}{X} = \begin{pmatrix} \Sigma(X_{i1} - \bar{X}_1)^2 & \Sigma(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) \\ \Sigma(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) & \Sigma(X_{i2} - \bar{X}_2)^2 \end{pmatrix}.$$

Recall that the variance of X is defined as

$$\text{Var } X = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

and the covariance of X_k and X_p is

$$\text{Cov}(X_k, X_p) = \frac{1}{N} \sum_{i=1}^{N} (X_{ik} - \bar{X}_k)(X_{ip} - \bar{X}_p) .$$

(When the covariance of X_1 and X_2 is 0 then the variables are not linearly related; when it is > 0 or < 0 they are linearly related.) Therefore, $(\underset{\sim}{X}'\underset{\sim}{X})$ can be written as

$$\underset{\sim}{X}'\underset{\sim}{X} = N \cdot \begin{pmatrix} \text{Var } X_1 & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_1, X_2) & \text{Var } X_2 \end{pmatrix}$$

The inverse of this symmetric matrix is simply

$$(\underset{\sim}{X}'\underset{\sim}{X})^{-1} = \frac{1}{N(\text{Var } X_1 \text{ Var } X_2 - (\text{Cov}(X_1, X_2))^2)} \cdot \begin{pmatrix} \text{Var } X_2 & -\text{Cov}(X_1, X_2) \\ -\text{Cov}(X_1, X_2) & \text{Var } X_1 \end{pmatrix}$$

To evaluate $\underset{\sim}{X}'\underset{\sim}{y}$ we simply multiply

$$\underset{\sim}{y} = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_N - \bar{y} \end{pmatrix} \qquad \text{by } \underset{\sim}{X}' .$$

771

to get

$$\begin{pmatrix} \Sigma(Y_i - \bar{Y})(X_{i1} - \bar{X}_1) \\ \Sigma(Y_i - \bar{Y})(X_{i2} - \bar{X}_2) \end{pmatrix}$$

But this is

$$\begin{pmatrix} m\mathrm{Cov}(Y,X_1) \\ m\mathrm{COV}(Y,X_2) \end{pmatrix} = m \cdot \begin{pmatrix} \mathrm{Cov}(Y,X_1) \\ \mathrm{Cov}(Y,X_2) \end{pmatrix} = \underset{\sim}{X}'\underset{\sim}{Y}$$

Now, $(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\, \underset{\sim}{X}\underset{\sim}{Y}$ can be written

$$\frac{1}{(\mathrm{Var}X_1\mathrm{Var}X_2 - (\mathrm{Cov}(X_1,X_2))^2)} \begin{pmatrix} \mathrm{Var}X_2 & -\mathrm{Cov}(X_1,X_2) \\ -\mathrm{Cov}(X_1,X_2) & \mathrm{Var}X_1 \end{pmatrix} \begin{pmatrix} \mathrm{Cov}(Y,X_1) \\ \mathrm{Cov}(Y,X_2) \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

Multiplying out and setting up equations for b_1 and b_2 gives

$$\frac{1}{\mathrm{Var}X_1\ \mathrm{Var}X_2 - (\mathrm{Cov}(X,X_2))^2} (\mathrm{Var}X_2\ \mathrm{Cov}(Y,X_1) - \mathrm{Cov}(X_1,X_2)\mathrm{Cov}(Y,X_2)) = b_1$$

and

$$\frac{1}{\mathrm{Var}X_1\ \mathrm{Var}X_2 - (\mathrm{Cov}(X_1,X_2))^2} (\mathrm{Var}X_1\ \mathrm{Cov}(Y,X_2) - \mathrm{Cov}(X_1,X_2)\ \mathrm{Cov}(Y,X_1)) = b_2$$

By examining these equations we can see two important aspects of least squares estimation. First, if the two variables X_1 and X_2 are identical then $\mathrm{Cov}(X_1,X_2)$ will be $\mathrm{Cov}(X_1,X_1)$, and this equals Var X_1. Consequently, the difference $\mathrm{Var}X_1\ \mathrm{Var}X_2 - (\mathrm{Cov}(X_1,X_2))^2$ will reduce to $(\mathrm{Var}X_1)^2 - (\mathrm{Var}X_1)^2 =$ and there will be no solution (or an infinite number of solutions) to the equations for the b coefficient. Obviously, when the denominator is close to 0 (when X_1 and X_2 are very similar) computers will begin to have problems giving precise answers.

On the other hand, when the two X variables are not linearly related at all then $Cov(X_1,X_2) = 0$. Consequently, the equation for b_1 (and similarly for b_2) reduces to:

$$\frac{1}{VarX_1\ VarX_2}\left(VarX_2\ Cov(Y,X_1)\right) = b_1$$

Or

$$\frac{Cov(Y,X_1)}{VarX_1} = b_1 \quad .$$

Writing this out yields

$$\frac{\Sigma(X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\Sigma(X_{i1} - \bar{X}_1)^2} = b_1$$

which is the equation we obtained for the univariate situation. The equation for b_2 would be

$$\frac{\Sigma(X_{i2} - \bar{X}_2)(Y_i - \bar{Y})}{\Sigma(X_{i2} - \bar{X}_2)^2} = b_2 \quad .$$

These equations yield the same value for $\underset{\sim}{b}$ as would be obtained if individual univariate regressions were run. In general we observe that the least squares solution of the p variable multiple regression situation will yield the b_i values as p individual univariate regressions when all variables have zero covariances. (This will not be true for b_0 -- why not?) When the X variables are not strictly statistically independent (i.e., when the covariances of the X variables do not equal zero) the multiple regression solution and the univariate regression solutions will differ.

IPM

Handout

What to look for in reading technical reports

eneral

hat is the problem being addressed?
s it a substantive or methodological issue?
s it a basic or secondary issue?
s it an applied issue, a theoretical issue, or a combination?
s it part of an established research tradition or does it stand alone?
an you see any relevance for your concerns?

ata

hat are the data?
here do they come from?
ow were they gathered?
re there "data problems" (missing values, poor sampling, poorly defined measures, etc.)?
re the data relevant to the problem addressed or are there better sources of information on the topic?
re the data available if you wanted to pursue the analysis?

ethod

hat procedure(s) was used?
s the analytic procedure appropriate to the data?
ill it speak to the problem addressed?
o you understand it?
ave you used it before or is it new?
s its application here novel and innovative or typical and expected?
re there other procedures which could have been used?
id exploration precede confirmation?
f not, do you believe the analysis?
an the method be applied to other areas?

esults

hat are they?
ave you learned anything?
s it important?
o you believe it?
s it relevant to your own concerns?
re there any policy implications?
ow do they relate to other things you know about this or related problems?
o you believe the results are robust and will hold outside of the limited context of this study, i.e., can they be generalized?

Miscellaneous

Was it worth the effort?
Is it a landmark study useful for citation in other contexts or better forgotten?
Has the author(s) published anything else that you might follow-up?
Are the references useful?
Was the presentation adequate and convincing?
Can the data be mined for other issues?
Are there any outstanding problems?
Were all the questions raised dealt with?
What are the directions for future research?

Some Principles of Graphics for Scatterplots

This handout is a continuation of the Module I handout concerning some standards for graphics. The earlier handout was concerned with tables and charts; this handout focuses on scatterplots, or graphic displays of (X,Y) paired observational data sets. As before, some of these principles are due to Edward R. Tufte.

The principles discussed here are:

(1) Less is more;

(2) Suppressing the frame and grid;

(3) Pay attention to details;

(4) Friendly lettering and other aesthetic considerations:

(5) Using parallel plots.

We intersperse our text with many examples, both good and bad, taken from the pages of The New York Times, Business Week, and several scientific journals. As mentioned in the earlier handout, we feel that paying attention to the details expounded upon in these pages will help you produce good displays.

Principle 1: Less is More

"Less is more" is the first principle in both this handout and the earlier graphics handout. It is also the most important. Why should we waste pages and pages of text when a simple and explicit graphical display suffices?

Scatterplots, especially of time series data, are used in reports and articles to a greater extent today than ever before. They increase in value as more authors recognize their usefulness in presenting and summarizing the relationship between two quantifiable variables.

As an illustrative example, consider the scatterplot shown in Figure 1.

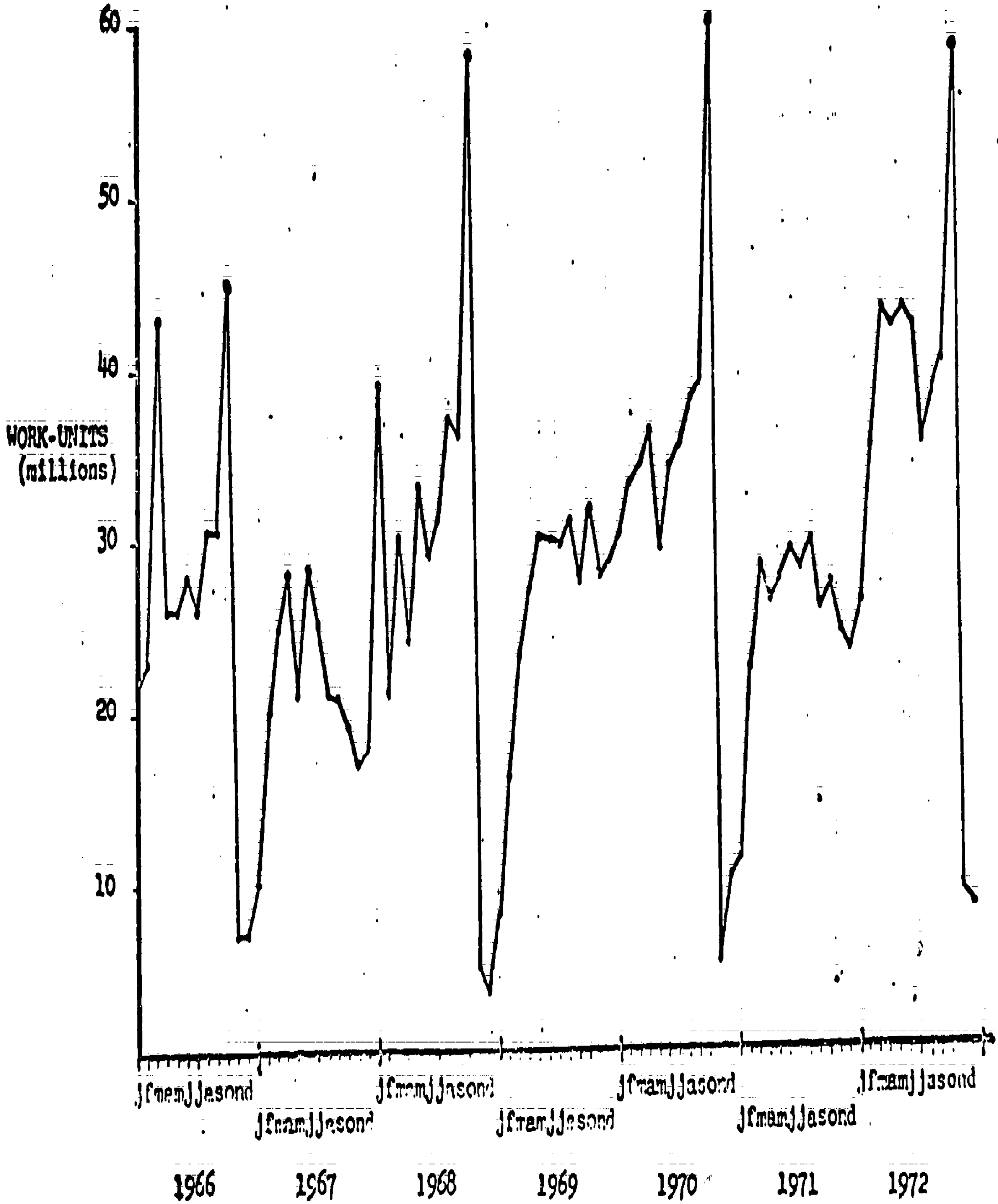
Tufte plotted the work load of the Publications Distribution Service for the U.S. House of Representatives as a time series. This diplay has an immediate impact on the reader. There are dramatic peaks in the data every second October -- right before Election Day!

The New York Times published a very involved, 700 word article to describe this phenomenon (N.Y. Times, June, 1975, p. 28). A display, such as that prepared by Tufte, could certainly have shortened the article and enhanced reader enjoyment.

FIGURE 1

U.S. HOUSE OF REPRESENTATIVES: PUBLICATIONS DISTRIBUTION SERVICE,

MILLIONS OF WORK-UNITS PER MONTH, 1966-1972



QMPM

XVI.II.400

Principle 2: Suppressing the Frame and Grid

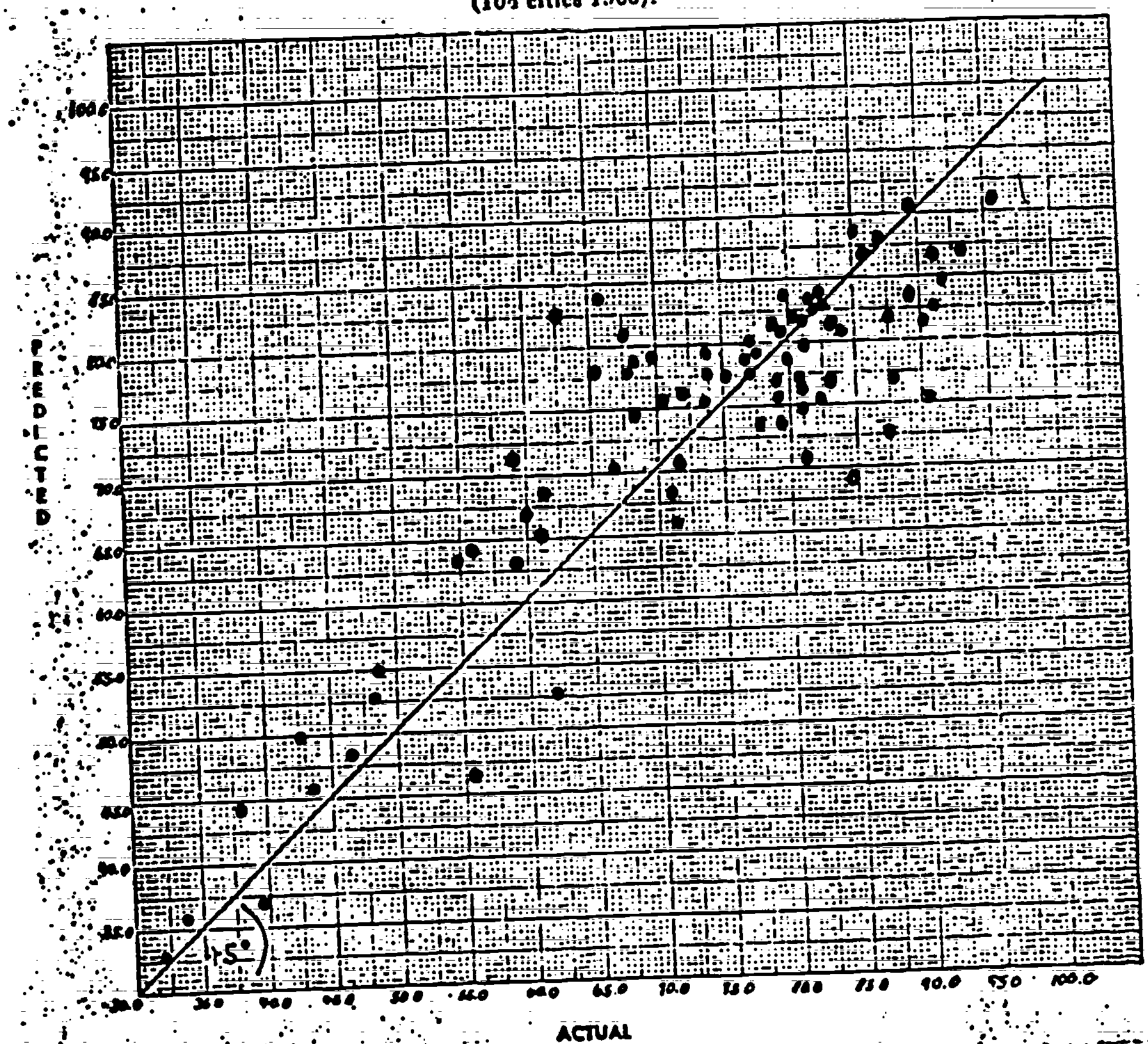
Considerations when making a scatterplot were discussed in the Prerequisite Inventory, Module I, and Chapter 4 of Hoaglin's forthcoming text A First Course in Data Analysis. We want to emphasize that final versions of scatterplots should not contain the graph paper grid, nor should the frame of the paper be included.

Consider Figure 2, a graphic that violates this principle. It is difficult to find the points in the grid. Figure 3 shows the same plot, first with grid (and quite a few of the points) suppressed, and then second, underneath the first, with grid and frame erased. The first plot greatly simplifies the relationship between X and Y by summarizing it with a line (where is the equation?) and a few token points (perhaps too few). The second plot is slightly better. Comments are Tufte's.

A final note: These displays show the relationship between actual registration rates and predicted rates. Isn't this relationship one of "actual data values" to "fitted data values"? How can we better analyze these data, using techniques introduced in class?

FIGURE 2

Relationship of Actual Rates of Registration to Predicted Rates (104 cities 1960).



Source: Stanley Kelley, Jr., Richard E. Ayres, and William G. Bowen, "Registration and Voting: Putting First Things First," American Political Science Review, 61 (June, 1967), p. 371.

784

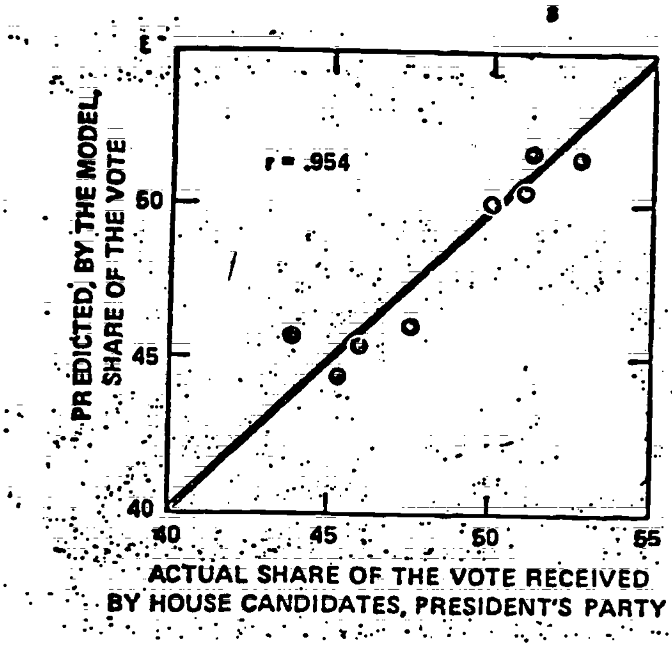
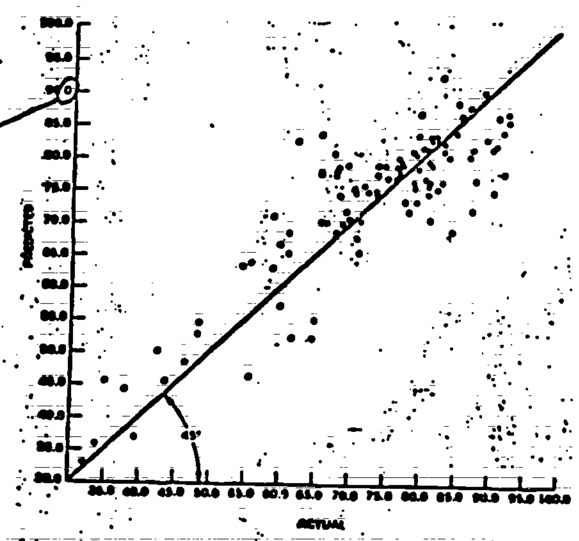


FIGURE 3

ACTUAL AND PREDICTED SHARE OF THE TWO-PARTY VOTE RECEIVED BY CONGRESSIONAL CANDIDATES OF PRESIDENT'S PARTY

Source: Edward R. Tufte, "Determinants of the Outcome of Midterm Congressional Elections," American Political Science Review, 69 (September, 1975), p. 818.

extra digits not needed; ".0" should be deleted from each number



Relationship of Actual Rates of Registration to Predicted Rates (104 cities 1960).

Source: Stanley Kelley, Jr., Richard E. Ayres, and William G. Bowen, "Registration and Voting: Putting First Things First," American Political Science Review, 61 (June, 1967); figure from reprinted version in Edward R. Tufte, ed., The Quantitative Analysis of Social Problems (Reading, Massachusetts: Addison-Wesley, 1970), p. 267.

Principle 3: Pay attention to Details

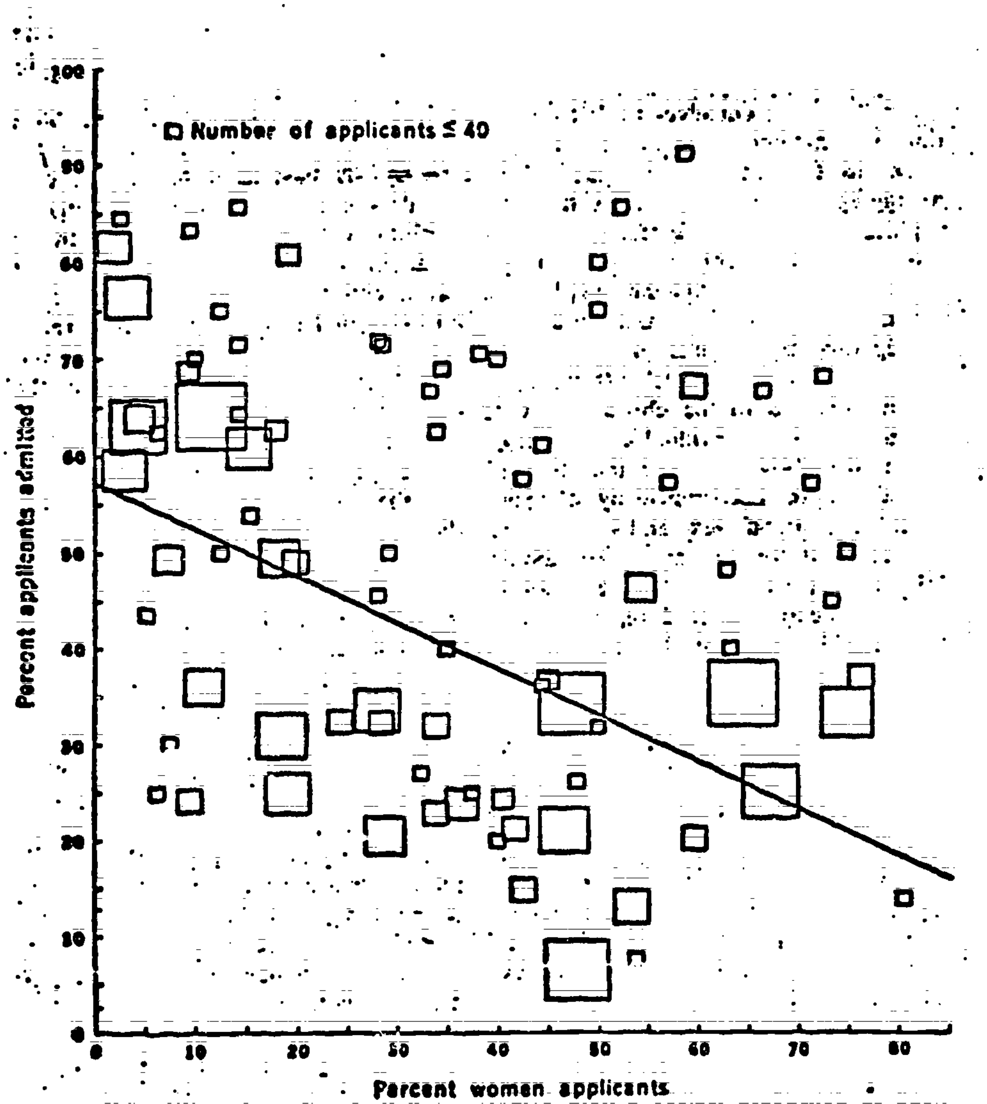
By giving the details of a scatterplot proper attention, a good display can be made even better. If the graphic can be reproduced in color, then by all means, make the points black, the scale green, and the fitted line red. Choose a good symbol to use for plotting points. We prefex X over • because the former symbol is larger. If you have different types of points, use different symbols to highlight the differences.

Consider Figure 5 from an article published in Science. The authors are analyzing sex discrimination in graduate school admissions at Berkeley. They have plotted the percent women admitted (Y), versus percent women accepted (X), one "box" per department. Note that the size of the boxes is related to the total number of applicants to the department. Here are some critical comments:

1. No detailed scale for size of box given. Why not label the largest boxes, and give total number of applicants?
2. Why is the minimum box size ≤ 40?
3. Fitted line does not fit the data!
4. Lettering is of poor quality.
5. Could we improve on the use of the boxes?

786

FIGURE 5



Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

Source: P. J. Bickel, E. A. Hammel, and J. W. O'Connell, "Sex Bias in Graduate Admissions: Data From Berkeley," Science, 187 (February 7, 1975), p. 400.

Principle 4: Friendly Lettering, etc.

This principle is briefly stated: pay attention to the "aesthetic" details. "Friendly lettering" is a good example. Instead of typing comments on the display, letter them by hand. If necessary, let a professional do it. Graphics by Roger Hayward are good examples of friendly plots, as shown in the plots of Figure 6 taken from a Chemistry text. Hayward is best known for his graphics work found in the "Amateur Scientist" section of Scientific American.

788

FIGURE 6

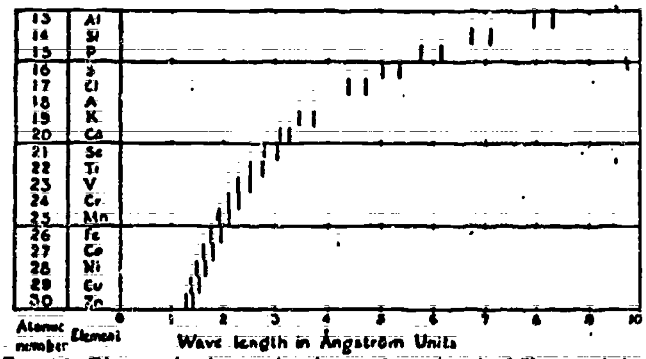


FIG. 4-1. Diagram showing regular change of wavelength of X-ray emission lines for a series of elements.

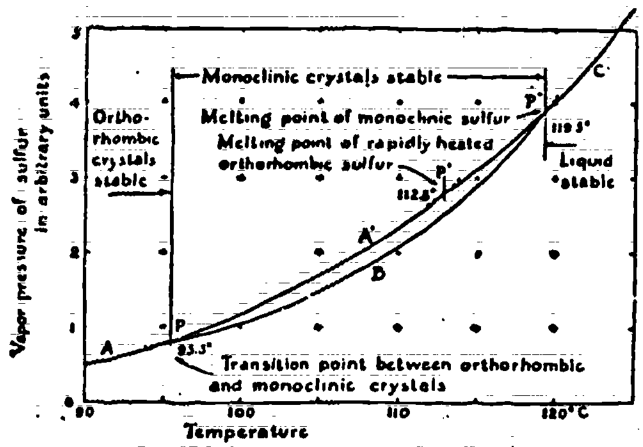


FIG. 17-2. Vapor pressure curves for sulfur.

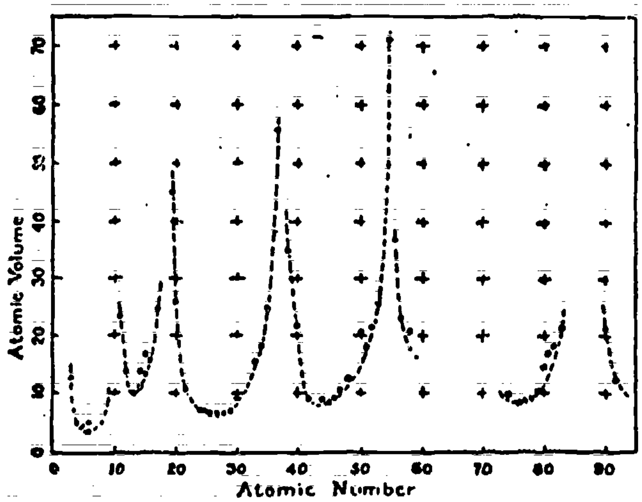


FIG. 5-1. Curve of atomic volume (volume containing 1 gram-atom) of elements as function of atomic number, illustrating periodicity of properties.

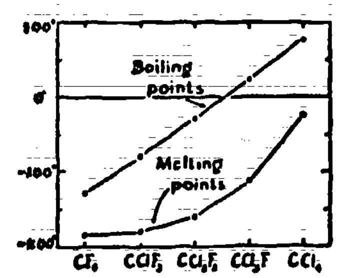
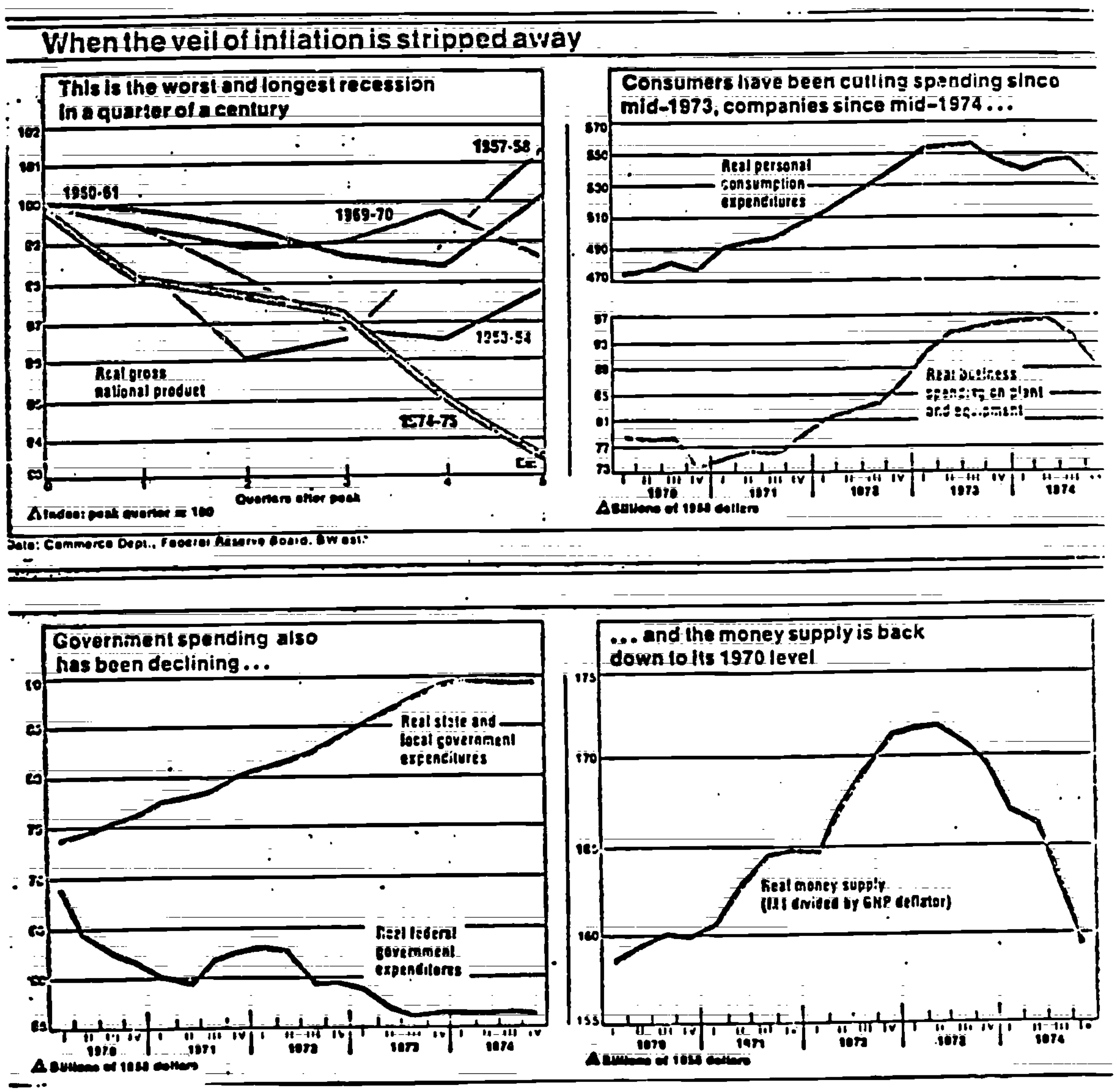


FIG. 15-5. The effect of molecular symmetry on melting point.

Principle 5: Using Parallel Plots

Parallel plots are useful in summarizing complex data sets. The graphs in Figure 7 are from Business Week and are quite attractive. Business Week usually produces very good displays of economic time series data.

FIGURE 7



790

QUANTITATIVE METHODS FOR PUBLIC MANAGEMENT

MODULE III, REVISED

Developed by

SCHOOL OF URBAN AND PUBLIC AFFAIRS
CARNEGIE-MELLON UNIVERSITY

SAMUEL LEINHARDT, PRINCIPAL INVESTIGATOR
and
STANLEY S. WASSERMAN

Under Contract to

THE URBAN MANAGEMENT CURRICULUM DEVELOPMENT PROGRAM
THE NATIONAL TRAINING AND DEVELOPMENT SERVICE
5028 Wisconsin Avenue, N.W.
Washington, D.C. 20016

Funded by

The Office of the Assistant Secretary
for Policy Development and Research
U.S. Department of Housing and Urban Development

Package XVI

Acknowledgements

Assistance in the preparation of this package was provided by Blaine Aikin, Larry Albert, Joseph Chmill, Steve Clark, Marjorie Farinelli, Janice Greene, Gretchen Hemmingsen, Paul W. Holland, J. Michael Hopkins, Gaea Leinhardt, Richard Sandusky, Christine Visminas, Diane Warriner, and Tammar Zeheb.

Acknowledgements

Assistance in the preparation of this package was provided by Blaine Atkin; Larry Albert; Joseph Chmill; Steve Clark; Marjorie Farinelli; Janice Greene; Gretchen Hemmingsen, Paul W. Holland; J. Michael Hopkins, Gaea Leinhardt; Richard Sandusky, Christine Visminas; Diane Warriner, and Tammar Zeheb.

TABLE OF CONTENTS

Material intended solely for the instructor is denoted by an (I). Material that should also be distributed to the students is denoted by an (S).

Introduction to Module III

Overview

Module III of the Quantitative Methods for Public Management package contains three units, numbers 5, 6 and 7. Unit 5, Probability and Sampling, introduces the student to the notions of probability and random variables. The relative frequency approach to probability is emphasized and a sampling experiment is used to provide a concrete, empirical feel for this fundamental idea. Following the introduction of probability notions, the concept of a random variable, a variable which takes on values with associated probabilities is introduced. Distributions for selected random variables are then discussed with special emphasis placed on those random variables traditionally associated with linear models. The shape of distributions is illustrated using the graphics tools of Module I.

Distribution is felt to be important to future public managers because many policy relevant analytic situations require knowledge of the characteristic shape and moments of a variety of well known random variables. Simple computing probabilities of events, such as the probability of finding a physician in a census tract when the physician distribution is known to be Poisson, requires knowledge of how the random variable behaves. The distributions introduced in Unit 5 cover those most likely to be encountered in the field. Furthermore, exposure to these selected distributions prepares the student to use others as the occasion arises in the computation of moments and the performance of such inferential procedures as constructing confidence intervals and hypothosis tests. Numerous examples in presentation material, homework, and exams in Module III illustrate the everyday utility of applied distribution theory.

Unit 6, Inference, introduces the student to the use of probability notions in determining the precision of parameter estimates and testing hypotheses. The material covered here is traditional statistical inference. However, students are cautioned against blind use of these procedures through careful consideration of the stringent assumptions implicit in the approach.

Unit 7, Sample Surveys, departs from the usual material taught in QMPM in that it concerns data collection rather than data analysis. The unit introduces the student to the use of surveys, the design of questions, questionnaire layout, fielding procedures and sampling designs. Emphasis is placed on the utility of surveys in policy analysis and their shortcomings. The objective of the unit is to create intelligent consumers of survey reports rather than skilled survey researchers. Since this is a vast area covered in only three lectures, it is implicitly assumed that when the student becomes a practitioner and has need for a survey a professional organization with experience in conducting survey research will be retained.

Specific Objectives

Unit 5

Upon successful completion of Unit 5 a student will acquire an understanding of what is meant by the mathematical notion of probability and will be able to use this notion in the study of random variables and their applications. The student will be able to specify the distributions of selected continuous and discrete random variables including the Normal, rectangular, exponential, uniform, binomial and Poisson. The student will also be able to compute first and second moments for random variables with these distributions and to recognize when empirical data are likely to be observations on random variables with these distributions. The student will also be familiar with the t, X^2 and F distributions, how they are related to each other and to the Normal, and how they arise in a linear model estimated from sample data.

Unit 6

Upon successful completion of Unit 6 a student will be able to apply probability notions in the performance of statistical inferential proceedures. The student will be able to apply knowledge of probability distributions and moments to compute confidence intervals and confidence levels using known random variables. The student will also be able to construct hypotheses tests and specify significance levels for tests. The student will have learned to perform these operations on single parameters such as the mean of a sample and on coefficients estimated in a multiple regression equation.

Unit 7

Upon successful completion of Unit 7 a student will be able to recognize a situation requiring the use of a sample survey and to design and field a simple survey instrument. The student will also have developed a critical capacity permitting effective review of survey instruments and results and will be able to compute elementary statistics to estimate precision in the case of simple random sampling. The student will also be able to identify features of more complicated probability sampling proceedures such as cluster, stratified, systematic, or multistage sampling and will be able to assess their advantages and disadvantages in particular situations. The student will also be able to assess the benefits and disadvantages of various fielding methods such as face-to-face, telephone, and mailed response interviews.

Unit 5
Reading Assignments

Lecture	Assignment
Lecture 5-0	Mosteller, Rourke, and Thomas Chapter 3 Mueller, Schuessler, and Costner, Chapter 11
Lecture 5-1	Tufte , Chapter 2
Lecture 5-2	Mosteller, Rourke, and Thomas, Chapters 5 and 7
Lecture 5-3	Draper and Smith, Chapter 2

In addition, read the following articles in Tanur, et.al.:

pages 102-11
164-75
212-19
244-52
372-84
407-15

Texts:

Draper, N. and H. Smith, Applied Regression Analysis, New York: John Wiley & Sons, 1966.

Mosteller, F., R. Rourke, and G. Thomas, Probability with Statistical Applications, Second Edition, Reading, Massachusetts: Addison-Wesley, 1970.

Mueller, J., K. Schuessler, and H. Costner, Statistical Reasoning in Sociology, Third Edition, Boston: Houghton Mifflin, 1977.

Tanur, J., et.al., Statistics: A Guide to the Unknown, San Francisco: Holden Day, 1972.

Tufte, E. R., Data Analysis for Politics and Policy, Englewood Cliffs, New Jersey: Prentice-Hall, 1974.

Prerequisite Inventory
Module III

This prerequisite inventory contains a brief introduction to the vocabulary and notation of elementary probability and set theory. If you are still uncertain about any of these concepts after reading the inventory, please consult a member of the teaching staff.

Probability is a measure of chance. Discussions of chance are by no means limited to the classroom. The weatherman often states that there is a 60% chance of rain, or a friend might remark that "chances are I'll be home late again tonight." Or, someone else may state that it is likely that federal income tax will rise this year.

In spite of our familiarity with chance, it is difficult to come up with a rigourous definition of probability. Mathematicians and philosophers cannot agree on a single definition. One group of statisticians believes in objective probability. Objective probabilities are derived from repeated observations of the happening that is in question. The long run relative frequency with which the happening occurs is taken to be its probability. For example, suppose you work for a health agency which is testing the effects of red dye #2, a suspected carcinogen, on rats. You have tested red dye #2 on 3000 identical rats. Five hundred of those rats have developed cancer. You conclude that the probability that the next rat to receive red dye #2 will develop cancer is 500/3000, or 1/6.

Probabilities may be based on equally likely chances. The traditional examples of this type of probability are the flip of a coin and the roll of a die. When a coin is tossed.heads and tails are equally likely; each has probability 1/2. The sides of a die are equally likely, so the probability that a particular side will appear is 1/6. Suppose that we want to meet with some physicians in Buffalo to discuss their opinions on malpractice insurance. We know that physicians are distributed unevenly across census tracts. Let's assume that there are 1000 physicians, of which only 1 is in the first census tract. If we select the first physician that we meet with at random, then the probability of selecting each particular physician is equal, namely a 1 in 1000 chance, a probability of .001. With only one physician in the first census tract, we can also say that the probability that our first physician is from that tract is .001. If there are 18 physicians in census tract two, then the probability that our first physician is from that tract is .018.

Another definition of probability is <u>subjective</u> or <u>personal</u> probability. Subjective probabilities are based on personal belief or professional judgment. Not all statisticians accept the idea of subjective probability, but some are willing to assign measures of chance to happenings that cannot be repeated to obtain a long-run frequency. In 1977, the secession of Nantucket and Martha's Vineyard from Massachusetts seemed possible. We cannot repeatedly put Martha's Vineyard and Nantucket in a position to secede and count the number of times that they actually do secede. Nor is secession one of a number of equally likely events. It may be possible, however, for us to analyze the existing conditions which affect the secession decision and to state our opinion that there is a 20% chance that the islands will secede within the next ten years. This 20% is an example of a subjective probability.

Statisticians do have a precise language for describing the happenings to which probabilities are assigned. An experiment is an act that can be repeated under given conditions. In the examples above, testing red dye #2 on a rat, flipping a coin, rolling a die, and selecting a physician to interview are all experiments. Secession from Massachusetts is difficult to repeat and is therefore not an experiment.

An experiment has one or more possible outcomes. We will rarely be interested in experiments with one outcome, since that outcome occurs with probability 1 (certainty). The outcome of flipping a coin must be either heads or tails. The outcome of giving red dye #2 to a single rat must be either cancer or not cancer. The outcome of randomly selecting a physician from 1000 must be one of the 1000 physicians.

An elementary event is the outcome of an experiment. We will often be interested in more complex events which incorporate more than one outcome. These are called compound events. In the red dye #2 example, we do not care whether or not a particular rat develops cancer. Rather, we are interested in whether or not a sufficient number of rats develop cancer so that we can conclude that red dye #2 is harmful. Suppose we have decided to conclude it is harmful if more than half of the rats get cancer. This event, more than 1/2 of the rats developing cancer, is a combination of the elementary events that 1501 or 1502 or 1503, and so on up to 3000, rats develop cancer.

A set is a collection of numbers or objects that can be grouped together in some context. We may talk about the set of all black city managers or the set of lifetimes of General Motors automobiles. The elements of a set are listed between brackets { } and are separated by commas. A set of lifetimes of automobiles in months may look like

{68, 73, 79, 80, 83, 83, 91}.

The possible outcomes of an experiment are often listed in set notation. When this is done, the set S is called the sample space of the experiment. The sample space of the experiment of testing red dye #2 on one rat is

{cancer, not cancer}.

The sample space of the experiment of testing the effects of red dye #2 on 3000 rats is

{no rats get cancer, 1 rat gets cancer, 2 rats get cancer, 3 rats get cancer, ..., 3000 rats get cancer}

In large sets containing an obvious progression, we use ... to stand for "and so forth up to" followed by the final element in the set. The set

{1, 2, 3, ...}

is the set of positive integers and is of infinite length. (There is no "final" element.)

A subset of a set S is made up only of elements in S. Subsets may contain all of the elements in S or any number of elements less than the total number. The set which contains no elements is called the empty set and is symbolized by { } or Ø. The empty set is a subset of every set. An event (see above) is a subset of a sample space.

We may want to describe the union or the intersection of two or more sets. The intersection of sets consists of the elements which the sets have in common. The symbol for intersection is $\cap$. For example, consider A and B where A is the set of grades on a test in a classroom with computer-aided instruction and B is the set of grades on a similar test in a traditional classroom:

A: {80, 83, 86, 86, 89, 93}
B: {62, 70, 79, 83, 89, 90}

Then $A \cap B =$ {83, 89}

The union of sets, symbolized $\cup$, enumerates all of the elements that appear in one or more of the sets. Referring to A and B above,

$A \cup B =$ {62, 70, 79, 80, 83, 83, 86, 86, 89, 89, 90, 93}

We say that two sets, C and D, are mutually exclusive if they have no elements in common. This is equivalent to the mathematical statement $C \cap D = \{\ \}$. The set of U.S. Senators and the set of U.S. Representatives at one point in time are mutually exclusive sets.

Suppose that E is some event which is a subset of the sample space S. Then the set $\overline{E}$, consisting of all elements in S which are not in E, is called the complement of E. Complements have the properties that

$$\{E \cap \overline{E}\} = \{\ \}$$

and $$\{E \cup \overline{E}\} = S, \text{ the sample space.}$$

We will be discussing the probability that a particular event occurs. The probability of an elementary event is the likelihood that the experiment will result in that outcome. The probability associated with an entire sample space is 1.

You will see different notations used to denote the probability of an event. The most common are P{event}, Pr{event}, P(event), and Pr(event). When we roll one die, P{5} = 1/6 and P{an even number} = 1/2.

A population is a group of people (or things) specified by some characteristic. A population may be all people in the United States, 35-year-old congresswomen from Poughkeepsie, or men with income greater than $50,000. A sample is a subset of an entire population.

In Module III, we will use our knowledge of probability to generalize from samples to populations. Probability will enable us to quantify the uncertainty about the population which is due to our sampling, i.e., our observation of only part of the set consisting of all members of the population. Unless we analyze the entire population, there will always be uncertainty.

A sample is a batch of data, and we continue to use measures of location and scale that were discussed in Module I. In particular, the sample mean, $\bar{X}$, and the sample standard deviation, s, will be used. It is a useful property of sampling that as the size of a sample increases, both in absolute number and relative to the population size, the sample mean approaches the mean of the population.

A final concept with which to be familiar before proceeding to Module III is the difference between continuous and discrete. A discrete variable may take on one of a finite or countably infinite set of values. The number of black city managers and the set of positive integers are examples of discrete variables. A continuous variable may take on values from a set consisting of an interval of the real number line. Length of a particular road, average height of European men, and all numbers between 0 and 1 are examples of continuous variables.

Homework
Prerequisite Inventory, Module III

Questions 1-10 refer to sets A through D.

A: {0, 1, 2}
B: {1, 3, 5, 7, 9}
C: {2, 4, 6, 8, 10}
D: {1, 2, 3, 4, 5}

List the elements of the following sets (in set notation):

1.	$A \cup D$	6.	$(A \cup D) \cap (A \cup C)$
2.	$A \cap D$	7.	$(A \cap B) \cup (A \cap C)$
3.	$B \cap D$	8.	$A \cup B \cup C \cup D$
4.	$B \cap C$	9.	$A \cap B \cap D$
5.	$B \cup C$	10.	$(B \cap D) \cup (C \cup A)$

In questions 11-15, give the sample space of the described experiment.

11. Annual family income for a family in Detroit, given that the head of the household is chairman of the board of a major automobile manufacturing company.

12.. Lifetimes (in years) of individuals in Washington D.C. who died between 1960 and 1970 from a heart attack.

13. The number of cars passing a building in a one hour period.

14. Percentage of black students in your master's class.

15. The number of women in a room of 10 people

Homework Solutions
Prerequisite Inventory, Module III

1. {0, 1, 2, 3, 4, 5}
2. {1, 2}
3. {1, 3, 5}
4. { }
5. {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
6. {0, 1, 2, 4}
7. {1, 2}
8. {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
9. {1}
10. {0, 1, 2, 3, 4, 5, 6, 7, 8, 10}
11. Set contains 3 incomes (one for GM, Chrysler, Ford), each element in excess of $200,000
12. Set contains positive integers between 20 and 90 (approximately) with most elements between 40 and 60.
13. [0, 1, 2, ...]
14. If class has N students, set is [0%, $\frac{1}{N}$ x 100%, $\frac{2}{N}$ x 100%, ..., $\frac{N-1}{N}$ x 100%, 100%]
15. [0, 1, 2, 3, ..., 10]

Lecture 5-0. Introduction to Unit 5

Introduction to Unit 5, Probability and Sampling

Lecture Content:

1. Introduction to objectives, problem, and notation for Unit 5
2. Discussion of a quantification of the notion of uncertainty

Main Topics:

1. Specific Introduction to the Objectives of Unit 5
2. Presentation of general problem of Unit 5
3. Notation for Unit 5
4. Definition of Probability

Topic 1. Specific Introduction to the Objectives of Unit 5

I. Questions to be answered in Unit 5

1. What is probability?

   a. In a frequency context, the probability of an event is the proportion of time that the event occurs on a large number of trials; e.g., P{head on a coin toss}

   b. However, probability is also a subjective notion and may vary from person to person -- chance, uncertainty, probable, likelihood of occurrence

2. How do we calculate and manipulate probabilities?

   a. We define simple rules for finding probabilities of equally likely outcomes

   b. Intersections and unions of events are easily visualized via Venn diagrams, and the corresponding probabilities found

3. What is a random variable, and how do we determine and utilize its probability distribution?

   a. A random variable is a variable whose specific value is not known with certainty

   b. A random variable is characterized by its probability distribution, which gives probabilities that the variable will have certain values.

   c. The probability distribution allows us to make certain statements about the random variable. For example, we may calculate the average value of the random variable.

4. What are some examples of random variables? [Random variables (usually) are either discrete or continuous.]

   a. Discrete Random Variables

      i. Discrete random variables take on a finite or countably infinite number of values; e.g., number of customers arriving at a supermarket between noon and 1 PM on a given day (Poisson)

      ii. Some discrete random variables include
          Binomial
          Poisson
          Uniform

809

b. Continuous Random Variables

i. Continuous random variables take on any value in some interval of the real line; e.g., heights of individuals (~Gaussian)

ii. Some continuous random variables include
Exponential
Rectangular
Gaussian
χ^2

5. How do we apply probability theory in data analysis?

We discuss regression analysis with probabilistic assumptions for errors

II. Skills to be mastered in Unit 5 (1)

1. Quantification of uncertainty through the concept of probability

2. Identification of the mathematical formulae of various sampling or probability distributions

3. Recognition of several types of random variables, and the ability to compute their expectations

4. Assessment of goodness of fit in multiple regression by utilization of probability.

Topic 2. Introduction to the Problems of Unit 5

I. What is probability?

1. Experiment: An activity or procedure involving alternative outcomes. Each has an associated probability. Repeatable under given conditions.
   Example: Flip a fair coin 10 times and let X = number of heads.

2. Each outcome is an event, either elementary or a combination of elementary events. All possible outcomes for an experiment is the sample

3. State that based on these 10 trials, Pr{head} = x/10

4. Note that this definition of probability is based on an infinite number of trials; only as $n \to \infty$, will $x/n \to 1/2$

II. How do we calculate probabilities?

1. Let A be an event
   Pr{A} must be between 0 and 1, inclusive.

   Intuitive: 0 means event doesn't occur; 1 means event does.

2. If the universe = set of all possible outcomes contains the events $A_1, A_2, \ldots, A_n$, then $\Sigma Pr\{A_i\} = 1$.

3. By using Venn diagrams we can compute probabilities of intersections, unions, etc. Intersection is "and", Union is "or" (2)

III. What are examples of random variables? (3)

1. Number of heads in n tosses of a coin = X
   X is a binomial random variable

2. Number of arrivals of airplanes at an airport in a specific hour = X
   X is a Poisson random variable

3. Number obtained on a single roll of a die = X (either 1, 2, 3, 4, 5, or 6)
   X is a Uniform random variable

4. Waiting time between arrivals of customers in a supermarket = Y
   Y is an exponential random variable

5. Random variables occurring in nature, e.g. intelligence scores, lengths of rose petals are (usually) assumed to be Gaussian.

IV. Conclusion

1. Need methods to recognize which distribution a batch of data came from

2. Need to be well versed in mathematical probability

Topic 3. Introduction to the Notation of Unit 5

I. We let capital letters such as, X, Y, Z,... denote random variables

II. Small letters x, y, z, ... denote realizations of these variables; e.g., we write $Pr\{X=x\}$, where x is an element of the sample space of X.

1. Read as "the probability that X, the random variable, will take the value x."

2. "Realization" indicates a value actually taken by a random variable

813

Topic 4. Probability

I. In this section, we define some basic notions of probability.

1. Probability is defined for <u>events</u>, or occurrence of a certain phenomenon. Events are notated A, B, ...

   a. A head on a single coin toss is an event

   b. 3 heads in 10 coin tosses is an event

   c. A leaf density of 6.93 is an event. Note P{A} between 0 and 1

2. The collection of all possible events, relative to a specific experiment, is called the <u>sample space</u> or <u>universe</u>, $S = \{A_1, A_2, \ldots, A_n\}$ It depends on the definition of the experiment.

   a. Events are subsets of S

   b. If we toss a coin 20 times and record the number of heads, then S = {0, 1, 2, ..., 19, 20} heads

   c. $\Sigma P\{A_i\} = 1$

3. Probability of an event A is the number of times the event occurs (or the number of successes) divided by the total number of trials, for a large number of trials $_{Pr}\{A\} = \frac{\text{number of "successes" of A}}{\text{total number of outcomes}}$

   i.e., it is a relative frequency.

4. Suppose we have 2 events A and B. The event $C = A \cup B$ is the <u>union</u> of A and B, and is the occurrence of either A or B or both A and B

5. Suppose we have 2 events D and E. The event $F = A \cap E$ is the intersection of D and E, and is the occurrence of both D and E.

6. $P\{C\} = P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}$; if A and B are <u>disjoint</u>, then $P\{A \cup B\} = P\{A\} + P\{B\}$

7. $\bar{A}$ is the complement of A
   $P\{\bar{A}\} = 1 - P\{A\}$

8. If 2 events are <u>independent</u>, or have nothing to do with each other, then $P\{A \cap B\} = P\{A\}P\{B\}$

9. In general, when 2 events are not independent,
$$P\{A \cap B\} = P\{A|B\}P\{B\}$$
where $P\{A|B\}$ is the conditional probability, read A given B.

10. Note that $P\{A|B\} = P\{A \cap B\}/P\{B\}$

II. Discuss example (4-5)

815

Lecture 5-0
Transparency Presentation Guide

Lecture Location Outline	Transparency Number	Transparency Description
Topic 1. Section II. 1.	1	Skills to be mastered
Topic 2. Section II. 3.	2	Venn Diagrams
Section III. 1.	3	Examples of Random Variables
Topic 4. Section II.	4	Experiment to Introduce Probability
Section II.	5	Probability Calculations

810

(1)

Skills to be mastered in Unit 5

1. Quantification of uncertainty

2. Identification of various sampling distributions

3. Recognition of several random variables

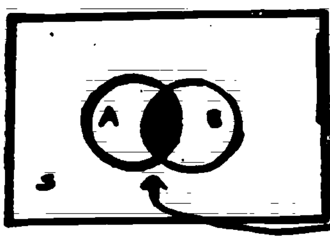
4. Goodness of fit in multiple regression via probability theory

[2]

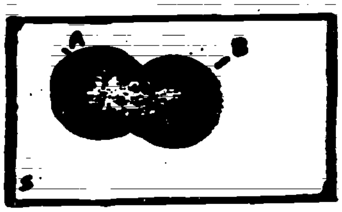
Venn Diagrams

Sample Space S

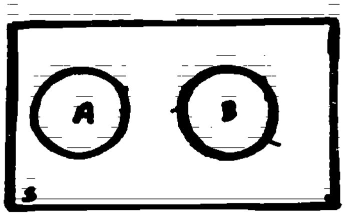
A and B are sets



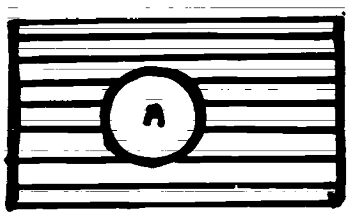
$A \cap B$
intersection



$A \cup B$
union



A and B
disjoint



complement of A

5-0

818

(3)

Examples of Random Variables

1. Number of heads in n tosses of a coin = X
   X is a <u>binomial</u> random variable

2. Number of arrivals of airplanes at Logan International Airport in a specific hour = X
   X is a <u>Poisson</u> random variable

3. Number obtained on a single roll of a die = X (either 1, 2, 3, 4, 5, 6)
   X is a <u>uniform</u> random variable

4. Waiting time between arrivals of customers at Giant Eagle, Center & Craig, = Y
   Y is an <u>exponential</u> random variable

5. Random variables occurring in nature, e.g. intelligence measures, lengths of rose petals, are (usually) assumed to be <u>Gaussian</u>

819

5-0

Experiment to Introduce Probability (4)

There are 1000 individuals in town Statecity. You interview each and record whether they are poor (income < poverty level) and whether they have a high school education.

The results:

	HS grad	Not HS grad	
Poor	50	250	300
Not poor	500	200	700
	550	450	1000

entries are numbers of people with specified characteristics

Events:

A = Poor, high school grad
B = Poor, not high school grad
C = Not poor, high school grad
D = Not poor, not high school grad

$P\{A\} = {}^{50}/_{1000} = .05$
$P\{B\} = {}^{250}/_{1000} = .25$
$P\{C\} = {}^{500}/_{1000} = .50$
$P\{D\} = {}^{200}/_{1000} = .20$

5-0

Experiment, continued (5)

Suppose we draw one person at random.

P{Poor and Not High School Grad} = .25
[intersection]

Note that

P{Poor} P{Not High School Grad} = .30 × .45 = .135 ≠ .25

so events are not independent

Unions

P{Poor ∪ HS grad} = .30 + .55 − .05 = .80

P{Not Poor ∪ HS grad} = .70 + .55 − .5 = .75

etc.

Conditional Probability

P{Poor / Not High School grad} = $\frac{.25}{.45}$ = .56

P{Not Poor / High School grad} = $\frac{.50}{.55}$ = .91

Lecture 5-1. Sampling Distributions

Sampling Distributions: Notion of random variables introduced by means of a sampling experiment. (1)

Lecture Content:

1. Discuss discrete and continuous random variables, numbers determined by the outcome of an experiment
2. Simple random sampling experiment from three probability distributions

Main Topics:

1. Random variables
2. Sampling Experiment
3. Sampling Distributions

Topic 1. Random variables

I. Basic Issue: Experimentally determined ... bers

1. Random variation arises in nearly al social and physical science experiments

   a. We may wish to measure the numb... of racial disorders occurring in inner city high schools in Boston

   b. Or we may wish to determine the boiling point of a certain chemical compound

   c. In both instances, the computed quantities will vary from school to school, or replication to replication

2. It is important to quantitatively define the nature of variation in our experiments

3. We describe this variation in probabilistic terms, to indicate our lack of certainty in the outcomes of the experiments

II. Problem: How do we characterize this variation?

1. Probabilities are defined only for long-run frequencies of events, where the experiment has been replicated many, many times

2. It is usually not profitable for us to conduct our experiments for policy decisions the required number of times

   a. Generally, we are lucky to have more than 100 replications of an experiment, because of either lack of time and money, or the small size of the sampling space

   b. What limited inferences can we make on fewer than 100 numbers?

   c. Stem-and-leaf displays indicate the nature of the variation, but rarely do they mimic the appearance of a known distribution.

3. Example illustrates this problem (2)

III. Solution: Attempt to describe the random nature of the variable by one of the well-known probability models

823

1. Essentially, we borrow strength from statistics and assume that the variable in question follows a known probability model

2. This approximation is reasonable in many instances; however, we must remember that it is just an approximation

3. The stem-and-leaf display is our most powerful analytical tool in determining which model to assume

IV. Definitions (3)

1. A random variable is a variable whose value is a number determined by the outcome of an experiment

2. If X is a random variable, with possible values x_1, x_2, ..., x_n, and associated probabilities $f(x_1)$, $f(x_2)$, ... $f(x_n)$, then f is called the probability function of X.

3. A random variable is like any other variable except that we know more about the random variable, namely the probabilities associated with its realizations

4. Random variables are either discrete or continuous (or sometimes a combination of these two)

   a. Discrete random variables take one of a finite or countably infinite set of values

   b. Continuous random variables take any value from an interval of real numbers

5. Random variables may also be vector-valued as in multiple regression

6. In the next lecture we discuss some special random variables at length

824

Topic 2. Sampling Experiment

I. Basic Issue: How can we best learn about random variation

1. We can collect many data sets, all of which have a random nature

2. However, it is more expedient to construct random numbers in a controlled "statistical laboratory"

3. We sample from 3 distributions--Gaussian, Rectangular, Exponential--and study the nature of the variability of several familiar statistics

II. Problem: How do we perform this controll ' experiment

1. Let: X be a random variable with a Gaussian distribution

   Y be a random variable with a rectangular distribution

   Z be a random variable with an exponential distribution

2. X, Y, and Z are continuous random variables, with probability functions as shown (4)

3. We shall draw 100 "samples" from each of these distributions, with sample size 20, by using a pseudo-random number generator

III. Solution: Study variation in our favorite statistics

1. For each sample, from each distribution, we compute $\bar{X}$, S^2, M, and ΔH

2. Thus we have 100 sample means from

   a. Gaussian distribution

   b. Rectangular distribution

   c. Exponential distribution

   Similarly for sample variances, medians, midspreads

3. We make a stem-and-leaf of each set of numbers and study the variation

4. Questions to be answered:

825

a. How much variability?

b. Is variability of a statistic constant over distributions?

c. Can we characterize the variability mathematically?

Note: Discuss how this sampling experiment might arise outside of our "statistical laboratory"

IV. Experiment

1. Sampling distribution of $\overline{X}$, $\overline{Y}$, $\overline{Z}$ (5a) (5b)

   a. Note symmetry

   b. Batches appear quite well behaved, especially Gaussian

2. Sampling distribution of Medians (6a) (6b)

   a. Also symmetric, and, except for Exponential, well behaved

   b. Spread is larger

3. Sampling distribution of S_x^2, S_y^2, S_z^2 (7a) (7b)

   a. Note skewness, to the larger values

   b. Rectangular, not very varied

4. Sampling distribution of midspreads (8a) (8b)

   a. Also skewed, but not as much as S^2

   b. Less varied, except for rectangular

5. This accords well with theory (9)

   1. Sample Mean ~ Gaussian by important Central Limit Theorem as $N \to \infty$

   2. Median ~ Gaussian, but with larger variance

   3. Sample Variance ~ χ^2, a skewed distribution

   4. Midspread ~ Gaussian (!)

Lecture 5-1
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 5-1 Outline
Topic 1		
Section II. 3.	2	Policy Question
Section IV 1.	3	Definitions
Topic 2		
Section II. 2.	4	Some Continuous Random Variables
Section IV. 1.	5a 5b	Sampling Distributions of $\overline{X}$, $\overline{Y}$, $\overline{Z}$
2.	6a 6b	Sampling Distributions of Median
3.	7a 7b	Sampling Distribution of s_x^2, s_y^2, s_z^2
4.	8a 8b	Sampling Distribution of Midspread
5.	9	Theoretical formulae for Sampling Distributions

827

Lecture 5-1

Sampling Distributions: Notion of random variation introduced by means of a sampling experiment

Lecture Content:

1. Discuss discrete and continuous random variables: numbers determined by the outcome of an experiment

2. Simple random sampling experiment

3. Several important sampling distributions

Main topics:

1. Random variables

2. Sampling experiment

3. Sampling distributions

Policy Question: Should instruction in foreign languages be dropped from high school curriculum, in a certain city, due to lack of interest by the students?

There are 19 high schools in the city.
In 1975-6, the percentages of students taking Spanish, French, German, and Latin, by school:

29.6	32.1	16.7	30.2	30.8	17.9	19.1
27.7	27.1	20.6	21.3	25.8	27.2	25.1
19.2	31.5	34.0	26.9	19.2		

1**	6 7 9 9 9
2	9 7 7 0 1 5 7 5 6
3**	2 0 0 1 4

unit = 1%

1**	6 7 9 9 9
2·	0 1
2**	5 5 6 7 7 9 9
3·	0 0 1 2 4

unit = 1%

How do we characterize this variation?

What is the probability function of the random variable X = % students studying foreign language?

Definitions

<u>Random variable</u> - a variable whose value is a number determined by the outcome of an experiment

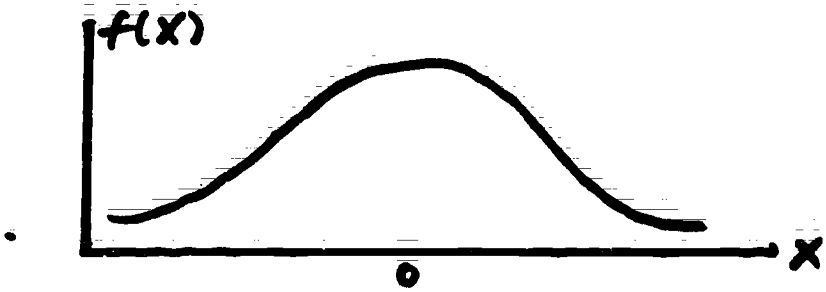
If X is a random variable, with possible outcomes $x_1, x_2, \ldots, x_n$, and associated probabilities $f(x_1), f(x_2), \ldots, f(x_n)$, then f is called the <u>probability function</u> of X

<u>Discrete</u> random variables - takes on one of a finite or countably infinite set of values

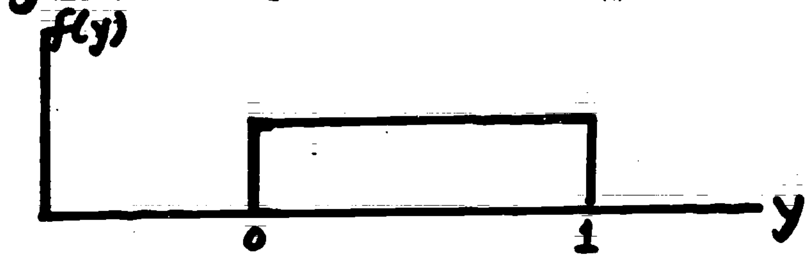
<u>Continuous</u> random variable - takes a value from a set of infinite size

Some Continuous Random Variables

1. X Gaussian $\quad X \in \mathbb{R}$



2. y Rectangular $\quad 0 \le y \le 1$



3. Z Exponential $\quad z \ge 0$

[5a.]

SAMPLING DISTRIBUTIONS of $\bar{X}, \bar{Y}, \bar{Z}$

Stem and Leaf Displays

GAUSSIAN

UNIT = 10^{-2}

-4	988641
-3	6666559
-2	77443110
-1	987666432210
-0	88887766665555333322222100
0	00122223444679
1	012355679999
2	0012556899
3	2366
4	9
5	12

RECTANGULAR

UNIT = 10^{-2}

3.	8999999
4	11111
T	22223
F	44555555
S	6666666777
4.	888888899999999
5	00000000111111
T	22222222223333333
F	4444555
S	6677
5.	8899
6	
T	23
F	44

HI	0.66

EXPONENTIAL

UNIT = 10^{-2}

5	1
5.	7
6	44
6.	5699
7	0122334
7.	56699
8	11122334
8.	5778888899
9	0111122224
9.	5566788888999
10	00112344
10.	7777777888
11	004
11.	5889
12	1
12.	779
13	00
13.	57799
14	1112

[5b]

Number Summaries

	Gaussian	Rectangular	Exponential
Min	-0.50	0.39	0.51
LH	-0.17	0.46	0.83
Med	-0.03	0.50	0.97
UH	0.14	0.54	1.08
Max	0.52	0.66	1.42
Mean	-0.02	0.50	0.98
Std. Dev	0. 23	0.06	0.21
Med	-0. 03	0.50	0.97
Δ H	0.31	0.08	0.25

Sampling Distributions of Median
Stem-and-Leaf Display

GAUSSIAN

UNIT = 10^{-2}

LO		-0.76	-0.76
5	90		
4	50		
3	554310		
2	88865933222220		
1	98765432211110		
0	99876654311111110000		
0	22235778		
1	0122233346688		
2	012237788		
3	00378		
4	223		
5	2		
6	2		
HI		0.69	0.76

RECTANGULAR

UNIT = 10^{-2}

2	4
2.	
3	4444
3.	5566778999
4	0011122222334
4.	55555556666666677777788999
5	001111111233333333344
5.	5555666777789
6	0234
6.	556689
7	012

EXPONENTIAL

UNIT = 10^{-2}

3	34
3.	5566999
4	134
4.	666789
5	11123344
5.	55566777889
6	011233
6.	5556666779
7	00012233344
7.	55567788889
8	0011114
8.	566
9	01
9.	6899
10	4
10.	7
11	9
HI	1.18 1.19 1.20 1.28 1.36

5-1

Module III

[6 b]

Number Summaries

	Gaussian	Rectangular	Exponential
Min	-0.76	0.25	0.33
LH	-0.21	0.43	0.54
Med	-0.02	0.49	0.67
UH	0.15	0.55	0.79
Max	0.76	0.73	1.36
Mean	-0.02	0.50	0.69
Std. Dev.	0.28	0.10	0.22
Med	-0.02	0.49	0.67
ΔH	0.36	0.13	0.25

Sampling Distributions of S_x^2, S_y^2, S_z^2

Stem-and-Leaf-Displays

Gaussian (Unit $= 10^{-2}$)

3	8
4	1222789
5	5579
6	14456668
7	022345558
8	0111112334445566666667789
9	00012245 6679
10	02269
11	61446
12	001 33667
13	2234577
14	025
15	003
16	8
17	1
HI	· 1.95 2.00

Rectangular (Unit $= 10^{-3}$)

3.	78
4	
4.	9
5	114
5.	56678
6	1123
6.	5555666778889
7	2444
7.	56678899
8	11122234
8.	555666666677888999
9	012444
9.	5556677888
10	0012234
10.	569
11	001224
11.	57

Exponential (Unit $= 10^{-1}$)

0	1
T	23333333
F	444444444448
S	666666666666677777777
0.	888888888899999999999
1	000000111
T	223333333
F	445555
S	67
1.	88899
2	01
HI	2.45 2.62 2.65 2.70

Number Summaries

	Gaussian	Rectangular	Exponential
Min	0.38	0.04	0.19
LH	0.75	0.07	0.62
Med	0.88	0.09	0.88
UH	1.21	0.10	1.32
Max	2.00	0.12	2.70
Mean	0.96	0.08	0.99
Std. Dev	0.33	0.02	0.56
Med	0.88	0.09	0.88
Δ H	0.46	0.03	0.70

[8a]

Sampling Distribution of Midspread, ΔH

Stem-and-Leaf Displays

GAUSSIAN

LO 0.51

UNIT = 10^{-2}

7	269
8	355
9	0122345689
10	1678
11	0013445566799
12	023446666789
13	001112334445577788
14	0122256789
15	01146689
16	139
17	0134688
18	0
19	245
HI	2.11 2.15 2.15 2.29

RECTANGULAR

UNIT = 10^{-2}

2.	5699
3	0001234
3.	55666889999
4	0011112233344444
4.	555555555556688888899
5	000001111223444
5.	5555666678889
6	0001112244
6.	5568

EXPONENTIAL

UNIT = 10^{-2}

4	5
5	00289
6	02345567
7	00112233678
8	1112345568888
9	000012355789
10	0123566779
11	013444456
12	123334469
13	122335
14	11235
15	338
16	1
17	5
18	15
19	4
HI	1.99 2.16 2.28

Module III

5-1

Number Summaries

	Gaussian	Rectangular	Exponential
Min	0.51	0.25	0.46
LH	1.14	0.41	0.78
Med	1.31	0.47	1.00
UH	1.52	0.56	1.25
Max	2.29	0.69	2.28
Mean	1.34	0.48	1.06
Std.Dev	0.33	0.10	0.37
Med	1.31	0.47	1.00
ΔH	0.38	0.14	0.47

Theoretical formulae for Sampling Distributions

Sample Mean $\bar{X}$ { Gaussian, Rectangular, Exponential } ~ Gaussian in large samples (exact if data are Gaussian)

Median M (Order statistic) { Gaussian, Rectangular, Exponential } ~ Gaussian in even larger samples (exact if data are Gaussian)

Variance s^2 { Gaussian, Rectangular, Exponential } ~ χ^2 Chi-Square in large samples (exact if data are Gaussian)

Midspread ΔH (difference in order statistics) { Gaussian, Rectangular, Exponential } ~ Gaussian in large samples (exact if data are Gaussian)

Lecture 5-2. Expectations of Random Variables

Expectations of random variables: Mathematical formulae for random variables, and means and for means and variances of these variables

(1)

Lecture Content:

1. Define Binomial, Poisson, and Uniform random variables and the contexts in which they occur
2. Define Exponential, Rectangular, and Gaussian random variables and the contexts in which they occur
3. Discuss mathematical expectations, variances, and independence of random variables

Main Topics:

1. Discrete random variables
2. Continuous random variables
3. Moments and other properties of these variables

Topic 1. Discrete Random Variables

I. Basic Issue: When can a batch of data be characterized by one of the special types of discrete random variables?

1. Recall that discrete random variables take on only a finite or countably infinite number of values (one-to-one association with the integers)

   a. For example X = number of Orientals in a census tract, has sample space $S = \{0, 1, 2, \ldots, N\}$, where N = total population of tract

   b. Or Y = number of rides taken on PAT busses in 1976, has $S = \{0, 1, 2, \ldots\}$, a countably infinite set of possible values

2. We deal here with discrete random variables in general

   a. X = discrete random variable

   b. S sample space of $X = \{x_1, x_2, \ldots, x_n\}$ where n may be infinitely large

   c. $f(x_i) = P\{X = x_i\}$ is the probability function of X

   d. The x_i are called "mass points", and f a "probability mass function", since f gives positive probability or weight (= mass) to only the x_i. (f is also simply called a probability function)

3. We shall discuss when X can be described with one of our special mass functions; i.e., when is X Binomial, Poisson, or Uniform

II. Problem: With only a few realizations of X, what can we say about the discrete random variable?

1. Occasionally we are able to take several samples (record several observations) of X

   a. A stem-and-leaf display should be made of this batch, and the shape studied quite closely and compared to the shapes shown later in this lecture

   b. If we suspect that X is either Binomial, Poisson, or Uniform, compute $\bar{X}$ and S^2 to compare with the "theoretical" values of these quantities

2. But if we have no observations on X, we must use whatever knowledge we have available about X to characterize its random nature

III. Solution: Three special random variables to use when appropriate

1. Binomial random variable (2)

   a. Assume an experiment involves N trials or observations, each trial being "independent", i.e. distinct from the other N-1 trials

   b. Assume each of the N trials has only 2 outcomes, a "0" or "1" (which could stand for any pair of mutually exclusive outcomes)

   c. Let p = Probability of an occurrence of a 1 on a single trial (this does not vary from trial to trial)

   d. Then X = number of 1's on N trials is a Binomial random variable

   e. Mass function $f(x) = \binom{N}{x} p^x (1-p)^{N-x}$ (3)

2. Poisson random variable (4)

   a. (Rar· ·ents $p \approx 0$, approximation of binomial) Assume a fix·· interval of time or space

   b. Consider a specific type of event or occurrence in the interval

   c. Let λ (lambda) be the <u>average</u> (mean) number of events that occur in the interval

   d. Let X = number of events that occur in the interval is a Poisson Random Variable

   e. Mass function $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ (5)

3. Uniform random variable (6)

   a. Assume an experiment with a sample space $S = \{x_1, x_2, \ldots, x_n\}$, n finite.

   b. If each x_i is equally likely to occur, then X is a uniform random variable

   c. Mass function $f(x) = \frac{1}{n}$

   d. ··dom numbers are realizations of uniform random variables. (7)

817

Topic 2. Continuous Random Variables

I. Basic Issue: When can a random variable be continuous?

1. In general, all measurements are discrete--there is a smallest possible fractio hat we can measure.

2. However, the thing measured is theo etically continuous

3. Continuous random variables may have symmetric, skewed, or even flat probability functions

   a. Y = continuous random variable

   b. S = sample space of Y = $\{y|a \leq y \leq b\}$, where a and b are any Real numbers

   c. f(y) is called the density function of Y, since the probability has been smeared over an interval (a,b), and every smaller interval has a chunk of probability

II. Problem: When can we assume that a specific continuous random variable can be characterized by one of our special density functions

1. We must use our intuition about the range of values of Y and the shape of empirical realizations Foreknowledge--and ch sensitivity

2. I can prove that Y is either Gaussian, Rectangular, c exp ential, then we have found a very important result

III. Solution: Specific continuous random variables

1. Gaussian (Normal)

   a. The "well-behaved" distribution of single batches

   b. Random variable is symmetric, and takes on values between $-\infty$, and ∞

   c. Important to notice the tails, and make sure that they are not too fat.

   d. There are many bell-shaped curves! e.g., Cauchy-- $f(y) = [\pi(1+y^2)]^{-1}$--thick tailed. $\mu = \infty$, no mean (expectation), integral doesn't converge

818

e. Density function of Gaussian

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-1/2\left(\frac{y-\mu}{\sigma}\right)^2}, \; Y \in R$$

μ(mean) $\in R$, σ(std.dev.) > 0.

2. Rectangular (9)

a. Flat over an interval (a,b)

b. Density function

$$f(y) = \frac{1}{b-a}, \; a \leq y \leq b$$

3. Exponential (Waiting time)

a. Waiting times are invariably exponentially distributed

b. Density function

$$f(y) = \theta e^{-\theta y}, \; y \geq 0.$$

θ^{-1} = average (mean) waiting time $\theta > 0$

819

Topic 3. Moments and other properties of these variables

I. Basic Issue: How can we <u>summarize</u> the random quality of these variables?

1. What is the average, or typical, value of a random variable (in terms of long term results of repeated experiments)

2. What is the variance of a random variable?

3. The mean (μ) and variance (σ^2) of a random variable X are useful and easily computed summarizations of X

II. Problem: How do $\overline{X}$ and S^2 compare to μ and σ^2

1. Suppose we have N sample observations on X and compute $\overline{X}$ and S^2

2. Then $\overline{X}$ estimates μ and s^2 estimates σ^2

3. Only as $N \to \infty$ does $\overline{X} \to \mu$ and $s^2 \to \sigma^2$; i.e. only in very large samples are the sample estimates identical to the population values

4. Remember our sampling experiment and the way the sample quantities varied about the true values

III. Solution: How to compute μ and σ^2 (10)

1. $\mu = E(X)$ -- read "expected value of X"--is the first moment of the random variable

$\mu = \sum_{i=1}^{n} f(x_i)x_i$, if X is discrete $S = \{x_1, x_2, \ldots, x_n\}$

$\mu = \int xf(x)dx$, if X is continuous

2. $\sigma^2 = E((X - \mu)^2)$ - read "Expected Value of $(X - \mu)^2$ -- is the second moment, about the mean, of the random variable

$\sigma^2 = \sum_{i=1}^{n} f(x_i)(x_i - \mu)^2$, if X is discrete

$\sigma^2 = \int (x - \mu)^2 f(x)dx$, if X is continuous

3. Examples

a. Discrete

850

i. Binomial

$\mu = Np$ $\sigma^2 = Np(1-p)$

ii. Poisson

$\mu = \lambda$ $\sigma^2 = \lambda$

iii. Uniform

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

Note that these are <u>not</u> sample quantities

b. Continuous

i. Gaussian

$\mu = \mu$ $\sigma^2 = \sigma^2$

ii. Rectangular

$\mu = \frac{1}{2}(b-a)$ $\sigma^2 = \frac{1}{12}(b-a)^2$

iii. Exponential

$\mu = \Theta^{-1}$ $\sigma^2 = \Theta^{-2}$

851

Lecture 5-2
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 5-2 Outline
Topic 1		
Section III		
1.	2	Binomial Random Variable
1.e	3	Binomial Mass Functions
2.	4	Poisson Random Variable
2.e	5	Poisson Mass Functions
3.	6	Uniform Random Variable
3.d	7	Sample Page of Random Numbers
Topic 2		
Section III		
1.	8	Gaussian Random Variable
2.	9	Rectangular Random Variable
Topic 3		
Section III		
1.	10	Mathematical Expectations

[1]

Lecture 5-2

Expectations of Random Variables:
Probability functions, means, and variances of several random variables

Lecture Content:

1. Define Binomial, Poisson, and Uniform random variables
2. Define Exponential, Rectangular, and Gaussian random variables
3. Discuss mathematical expectations and variances

Main topics:

1. Discrete Random variables
2. Continuous Random variables
3. Moments of Random variables

[2]

Binomial Random Variable

. Experiment involves N "independent" Bernoulli trials.

- Each of the N trials has 2 outcomes, either a "0" or a "1".
- Let $p = P\{1\}$ on each and every trial
- X = number of 1's on the N trials is a <u>Binomial</u> Random Variable

Mass function

$$f(x) = \binom{N}{x} p^x (1-p)^{N-x}; \; x = 0, 1, \ldots, N$$

$$\binom{N}{x} = \frac{N!}{x!(N-x)!} = \frac{N(N-1)\cdots(N-x+1)}{x(x-1)\cdots 2\cdot 1}$$

Examples

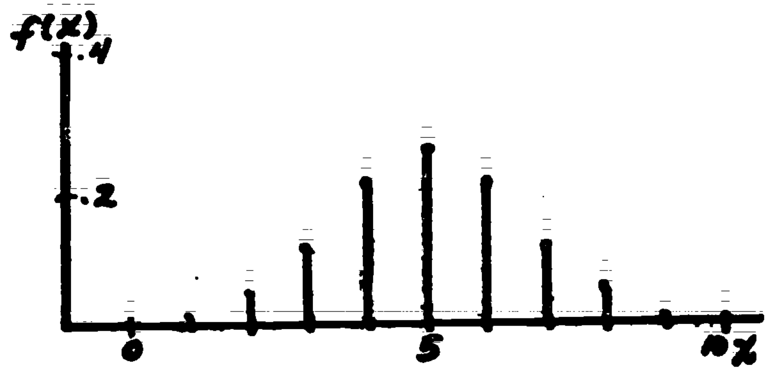
1) # heads in N coin tosses (classic)

2) # people in a group of size N with birthdays on January 1.

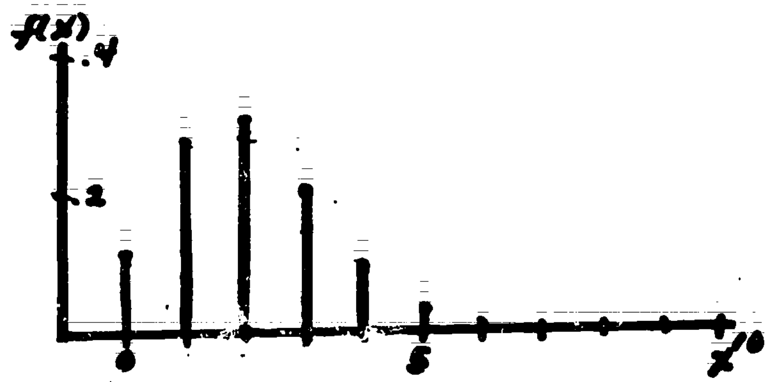
3) # defective parts in a batch of size N

5-2

Binomial Mass functions



N = 10
p = 1/2



N = 10
p = 4/5

N = 10
p = 1/10

[4]

Poisson Random Variable

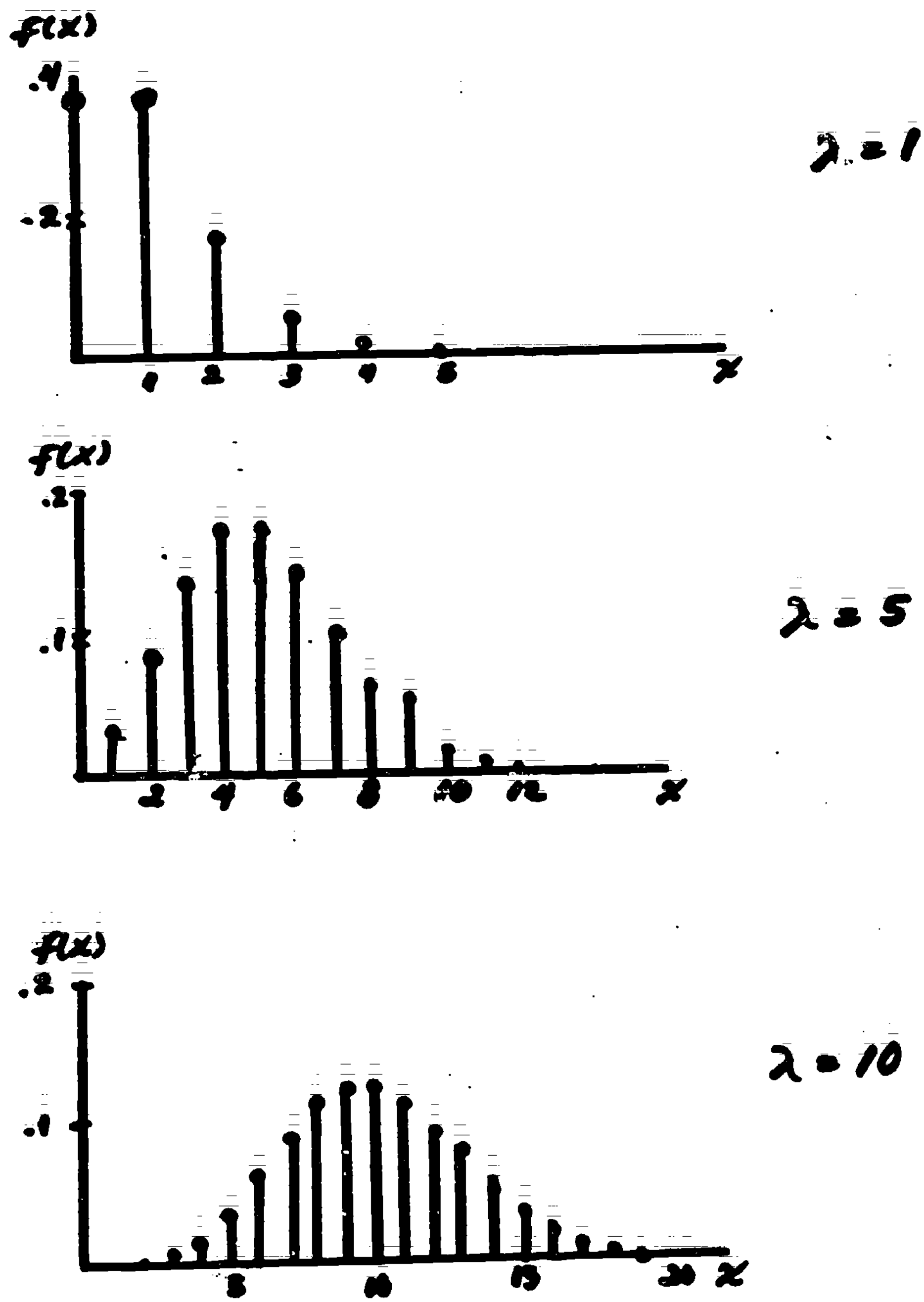
- Assume a fixed interval of time or space
- Consider a specific type of event or occurrence in the interval (rare).
- Let λ be the mean number of events that occur in the interval
- X = number of events that occur in the interval is a <u>Poisson</u> Random Variable

Mass function
$$f(x) = \lambda^x e^{-\lambda}/x! \; ; \; x = 0, 1, 2, \ldots$$

Examples

1) Flying bomb hits on London in WW II
2) Airplane arrivals
3) Prussian foot Soldiers (classic)
4) Physician offices ..
5) Discoveries.

Poisson Mass functions

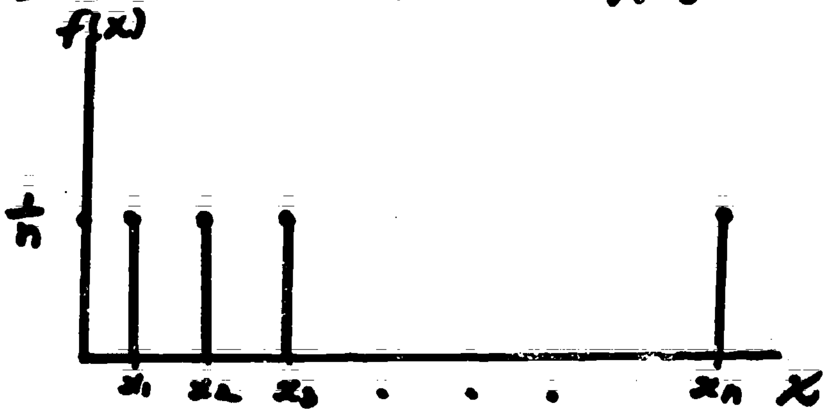


5-2

Uniform Random Variable

- Experiment with a sample space
  $S = \{x_1, x_2, \ldots, x_n\}$, n finite.
- Each x_i is equally likely to occur
- X is a Uniform Random Variable

Mass Function $f(x) = \frac{1}{n}$; $x = x_1, x_2, \ldots, x_n$.



Examples

1) Roll of a single die
2) Birthday of an individual
3) Random numbers

5-8

QMPM

[7]

Random Numbers

17623	47441	27821	91845	01654	50375	23941	44848
45054	58110	92081	97624	73750	68343	40727	81203
73700	58730	06111	64486	61163	22132	22896	14305
58374	03905	06865	95353	88445	85514	23627	79176
01981	17531	97372	39558	94180	71108	19121	11958
45639	02485	43905	01823	11433	12220	36719	35435
98832	38188	24080	24519	61838	68801	49856	21739
66638	03619	90006	95370	29908	68103	36855	19127
77580	87772	86877	57085	41018	69556	06402	03436
67125	98175	40912	11245	85095	01581	92299	06166
83808	98907	71829	99475	59705	78103	66710	41743
56402	58166	97902	88528	75904	55208	77905	20705
52728	15101	[illegible]	37706	78125	68672	79455	94534
44873	35302	[illegible]	35088	89088	89918	20787	05691
67352	41526	[illegible]	75440	83345	95889	39333	86027
35370	14915	16569	51945	62896	00342	66647	57096
00081	74957	87787	68849	96498	38270	80532	54307
58478	99297	43519	62410	30974	47335	04918	42974
21211	77299	74967	99038	57901	06163	99162	53285
98964	64425	33536	15079	88494	80633	47785	53996
81496	23006	56872	71401	34883	00045	89682	86664
71361	41989	92589	69788	21373	46458	28935	63903
36341	20326	37489	34626	16828	79262	23678	05589
34183	22856	18724	60122	33723	27616	92335	57136
98272	13969	12429	03093	01542	75066	73921	97188
75689	70722	88553	83900	00100	12787	74100	95536
64204	95212	31320	03783	82697	03389	19303	21646
16574	42305	56300	84227	28137	17549	22698	72955
93552	74363	30051	41367	02218	21570	33798	83789
48907	79840	34607	62668	56175	82515	23318	42207
42569	82391	20135	79906	80020	21622	67659	07878
81618	15125	48987	01250	20271	23004	48372	77621
31413	23756	45218	81976	38731	98944	02658	90698
77600	15175	67415	88801	48183	24263	49297	32923
42909	78616	45210	73186	48163	34158	03177	51696
02955	84348	46416	77911	45658	15024	66664	18730
93611	93346	71212	24405	71128	15524	55966	14763
11509	95853	02747	61889	19041	42899	49464	93965
83899	30932	90572	98071	32672	67506	93010	94527
73581	76740	03842	64009	15823	48310	04351	15521
49604	14809	12317	78162	82810	18981	62581	31642
81825	76822	87170	77235	74772	80840	05819	29023
60534	44842	16934	99406	52931	38199	85632	23761
41788	74409	76177	55519	95395	87644	09722	99231
73818	74434	02371	94293	76695	33451	57139	90612
81583	53439	99095	55578	83564	[illegible]	[illegible]410	57272
21858	30489	39251	70150	28355	[illegible]	[illegible]594	35807
09538	68184	62119	20229	84684	[illegible]	[illegible]350	66808
03249	74135	43003	53132	21135	[illegible]	[illegible]36	55887
59838	31226	89460	45191	24226	[illegible]	[illegible]97	41204

5-2

[8]

Gaussian Random Variable

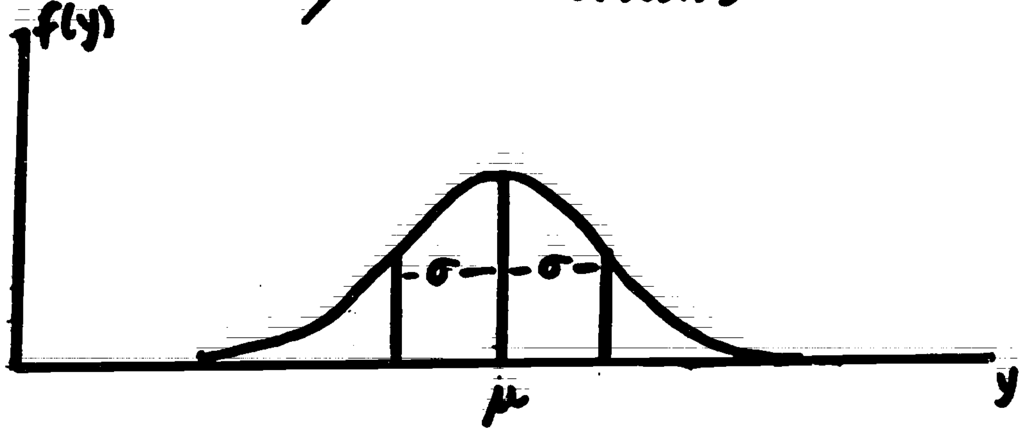
- Our "well-behaved" distribution
- Symmetric, $-\infty < y < \infty$.
- Note the tails - make sure that they are not too "fat"
- There are many distributions for "bell-shaped" curves

Density function

$$f(y) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} ; \quad -\infty < y < \infty$$

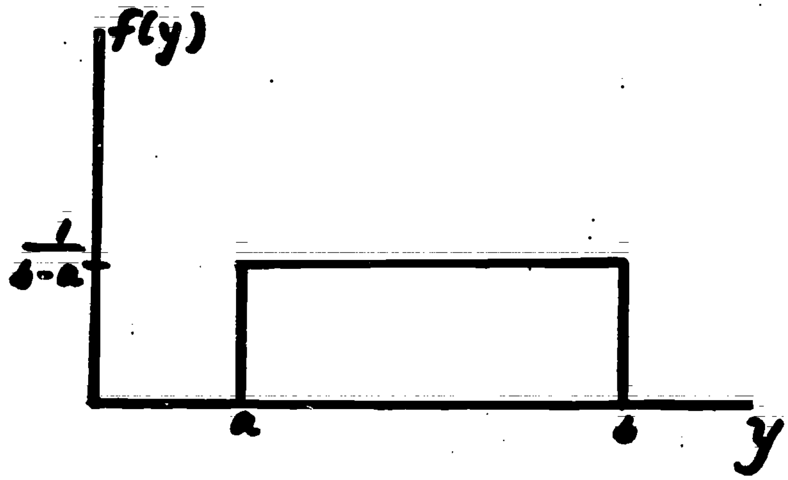
$\sigma > 0$ (std. dev.)

$-\infty < \mu < \infty$ (mean)

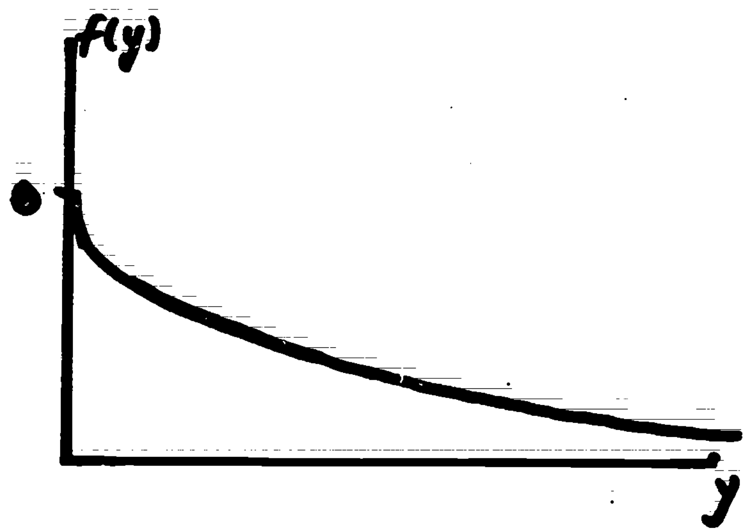


5-2

Rectangular Random Variable



Exponential Random Variable



5-2

Expectations

$\mu = E(X)$ $\qquad$ $\sigma^2 = E((X-\mu)^2)$

$\mu = \sum_{i=1}^{n} f(x_i)\, x_i$ $\qquad$ $\sigma^2 = \sum_{i=1}^{n} f(x_i)(x_i - \mu)^2$

$\mu = \int y f(y)\, dy$ $\qquad$ $\sigma^2 = \int (y-\mu)^2 f(y)\, dy$

	μ	σ^2
Binomial	Np	$Np(1-p)$
Poisson	λ	λ
Uniform	$\frac{1}{n} \sum x_i$	$\frac{1}{n} \sum (x_i - \mu)^2$
Gaussian	μ	σ^2
Rectangular	$\frac{1}{2}(b-a)$	$\frac{1}{12}(b-a)^2$
Exponential	θ^{-1}	θ^{-2}

Lecture 5-3. Probability and the Linear Model

Probability and the Linear Model: Probabilistic Assumptions Regarding the Errors of the Linear Model

(1)

Lecture Content:

1. Discuss probability distribution of the error terms in a linear model
2. Introduce several continuous probability distributions important to the analysis of a linear model

Main Topics:

1. The Mathematical Form of the Linear Model
2. Some Additional Distribution Theory

Topic 1. Mathematical Form of the Linear Model

I. Basic Issue: Review our notions of regression and introduce probability into our analysis

1. Unit 4 presented multiple regression as model fitting

2. No formalized goodness-of-fit measures were discussed, since such analyses depend on probability theory, that at the time, we had not discussed

3. In this lecture, and in Unit 6, we reintroduce the linear model, presenting the relevant probabilistic assumptions, and discuss how to evaluate a fitted model with probabilistic assessments

II. Problem: What do we assume about the random variation of the various components in the model

1. We have a vector of responses $\underset{\sim}{y}$, a matrix of vectors of carriers $\underset{\sim}{X}$, a vector of regression coefficients $\underset{\sim}{\beta}$, and of course, residuals.

2. One approach is to assume that $\underset{\sim}{y}$ is multivariate Gaussian distributed, with mean $\underset{\sim}{X}\beta$; we essentially work with the conditional distribution of $\underset{\sim}{y}$ given $\underset{\sim}{X}$.

   a. Regression may be approached strictly via conditional expectations

   b. We always assume that the rows of $\underset{\sim}{X}$ are known, fixed constants

   c. Hence it seems logical to say that given $\underset{\sim}{X}$, $\underset{\sim}{y}$ is a random variable

3. A simpler method of introducing probability focuses on the residuals of the model

   a. This approach exclusively will be utilized by us

   b. We assume that each residual $\hat{y}_i - y_i$ is a univariate Gaussian random variable

III. Solution: The linear model with Gaussian errors

1. First we review the linear model (2)

a. Assume y_i is a linear function of $x_{i1}, x_{i2}, \ldots, x_{ip}$, $i = 1,2,\ldots, N$

b. We write

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i$$

β_0 is a constant term in the model; e_i is the ith residual.

c. In matrix form

$\underset{\sim}{y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{e}$ (add a column of 1's to $\underset{\sim}{X}$)

$\underset{\sim}{y}$ (Nx1) $\underset{\sim}{X}$ (N x (p+1)

$\underset{\sim}{\beta}$ ((px1)x1) $\underset{\sim}{e}$ (Nx1)

d. Assumptions on $\underset{\sim}{e}$

i. $E(\underset{\sim}{e}) = \underset{\sim}{0}$ each residual has mean (expected value) = 0

ii. Residuals are independent, with constant variance σ^2

iii. Hence $Cov(\underset{\sim}{e}) = \sigma^2 \underset{\sim}{I}$ (NxN) matrix

iv. Since $\underset{\sim}{y} = \underset{\sim}{X}\,\beta + e$; i.e. $\underset{\sim}{y}$ is the sum of $\underset{\sim}{e}$ and $\underset{\sim}{X}\underset{\sim}{\beta}$, $\underset{\sim}{y}$ has the following moments: (3)

A. $\underset{\sim}{y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{e}$

B. $E(\underset{\sim}{y}) = E(\underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{e})$

$= \underset{\sim}{X}\underset{\sim}{\beta} + E(\underset{\sim}{e}) = \underset{\sim}{X}\underset{\sim}{\beta}$

C. $Cov(\underset{\sim}{y}) = Cov(\underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{e})$

$= Cov(\underset{\sim}{e}) = \sigma^2 \underset{\sim}{I}$

v. The new assumption is $e_i \sim$ Gaussian $(0,\sigma^2)$
Gaussian residuals

865

2. This new probabilistic assumption also affects the distribution of $\underset{\sim}{b}$, the Least Squares Regression Coefficients (4)

i. Remember,

$$\underset{\sim}{b} = \min_{\underset{\sim}{b}^*} (\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{b}^*)'(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{b}^*)$$

minimize sum of squares of residuals

ii. By differentiating, we find,

$$\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1} \underset{\sim}{X}'\underset{\sim}{y}$$

$= \underset{\sim}{M}\underset{\sim}{y}$, where $\underset{\sim}{M} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'$ a linear combination of $\underset{\sim}{y}$

iii. Hence, since a linear combination of Gaussian random variables is also Gaussian, b is Gaussian, with

A. $E(\underset{\sim}{b}) = E(\underset{\sim}{M}\underset{\sim}{y}) = \underset{\sim}{M}E(\underset{\sim}{y}) = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{\beta} = \underset{\sim}{\beta}$ (unbiased)

B. $\mathrm{Cov}(\underset{\sim}{b}) = M\,\mathrm{Cov}(\underset{\sim}{y})M$

$$= \underset{\sim}{M}\sigma^2\underset{\sim}{I}\underset{\sim}{M}'$$

$$= \sigma^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$$

$$= \sigma^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$$

Topic 2. Some Additional Distribution Theory

I. Basic Issue: Introduce other probability functions important in regression analysis

1. When fitting a model, we examined t-statistics, R^2, and pairwise correlation coefficients

2. Each of these quantities is a random variable with a specific distribution function

II. Distributions

1. $t = \dfrac{b_i}{S.E.(b_i)}$ follows a t distribution on N-p degrees of freedom, if $\beta_i = 0$ (5)

a. $$f(t) = \frac{\left(\frac{N-p-1}{2}\right)!}{\sqrt{\pi(N-p)}\ \left(\frac{N-p-2}{2}\right)!} \cdot \frac{1}{\left(1 + \frac{t^2}{N-p}\right)^{\frac{(N-p+1)}{2}}}$$

$-\infty < t < \infty$. (6a)

b. As $N \to \infty$, $f(t) \to$ Gaussian (0,1) (6b)

2. $$r_{ij} = \frac{\sum_k (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{\sqrt{\sum_k (X_{ik} - \bar{X}_i)^2 \ \sum (X_{jk} - \bar{X}_j)^2}} \qquad (7)$$

Sample Correlation Coefficient of X_i and X_j

a. If $\rho_{ij} = 0$ (population value is zero)

$r_{ij} \sim$ Gaussian $(0, \frac{1}{N-3})$ approximately

b. This approximation only holds for large N.

3. $\frac{1}{N-p} \ \Sigma(y_i - \hat{y}_i)^2 / \sigma^2$ follows a χ^2 distribution on (N-p) degrees of freedom. Sum of squares of Gaussian random variables are ~ Chi-square. (8)

867

4. If $SS_k \sim \sigma^2 \chi^2_k$ and $SS_\ell \sim \sigma^2 \chi^2_\ell$,

then $\dfrac{SS_k/k}{SS_\ell/\ell} \sim F_{k,\ell}$ F distribution, ratios of variances.

5. Since R^2 is a ratio of sums of squares,

$$\frac{R^2/(p-1)}{(1-R^2)/(N-p)} \sim F_{p-1,\, N-p}$$

if no linear relationship exists between $\underset{\sim}{Y}$ and $\underset{\sim}{X}$. (9)

Lecture 5-3
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 5-3 Outline
Topic 1 Section III		
1.	2	Review of the Linear Model
1.d.iv	3	New assumption
2.	4	Distribution of $\underset{\sim}{b}_{LS}$
Topic 2 Section II.		
1.	5	Student's t distribution
1.b	6a 6b	t versus Gaussian, for df = 1, 5, 10, 30
2.	7	Sample Correlation Coefficient
3.	8	χ^2 and F distributions
5.	9	Some characteristics of Important Distributions

889

[1]

Lecture 5-3

Probability and the Linear Model:
Probabilistic Assumptions Regarding the Errors of the Linear Model

Lecture Content:

1. Discuss the probability distribution of the error terms

2. Introduce several important probability distributions

Main Topics:

1. Mathematical Form of the Linear Model

2. Some Additional Distribution Theory

[2]

Review of the Linear Model

a. Y_i is a linear function of $X_{i1}, X_{i2}, \ldots X_{ip}$
$i = 1, 2, \ldots, N$

b. $Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} \ldots + \ldots + b_p X_{ip} + e_i$
e_i = ith residual or error

c. In matrix form $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{e}$

d. Assumptions on $\underset{\sim}{e}$

i Each e_i has zero expectation. $E(e_i) = 0$

ii Residuals are independent
$Cov(e_i, e_j) = 0$, all $i \neq j$

iii Residuals have constant variance
$Var(e_i) = \sigma^2$

Hence, $Cov(\underset{\sim}{e}) = E(\underset{\sim}{e}\underset{\sim}{e}') =$

$$\begin{pmatrix} Var(e_1) & Cov(e_1 e_2) & \cdots & Cov(e_1 e_N) \\ Cov(e_1 e_2) & Var(e_2) & & Cov(e_2 e_N) \\ \vdots & & \ddots & \\ Cov(e_1 e_N) & & & Var(e_N) \end{pmatrix} = \sigma^2 I$$

871

5-3

[3]

Regression Assumptions

Since $\underset{\sim}{y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{e}$, i.e. $\underset{\sim}{y}$ is the sum of the constant $\underset{\sim}{X}\underset{\sim}{\beta}$ and random variable $\underset{\sim}{e}$, we have

1. $E(\underset{\sim}{y}) = E(\underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{e}) = \underset{\sim}{X}\underset{\sim}{\beta} + E(\underset{\sim}{e}) = \underset{\sim}{X}\underset{\sim}{\beta}$
2. $Cov(\underset{\sim}{y}) = Cov(\underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{e}) = Cov(\underset{\sim}{e}) = \sigma^2 \underset{\sim}{I}$

New assumption:

$e_i \sim$ Gaussian, mean 0, variance σ^2

written $e_i \sim Gau(0, \sigma^2)$

Moreover,

$$\underset{\sim}{e} \sim Gau_n(\underset{\sim}{0}, \sigma^2 \underset{\sim}{I})$$

↑

n dimensions

and

$$\underset{\sim}{y} \sim Gau_n(\underset{\sim}{X}\underset{\sim}{\beta}, \sigma^2 \underset{\sim}{I})$$

5-3

[4]

Distribution of $\underset{\sim}{b}$, Least Squares Regression Coef.

Remember $\underset{\sim}{b} = \min_{\underset{\sim}{b}^*} (\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{b}^*)'(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{b}^*)$

$\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{y} = \underset{\sim}{M}\underset{\sim}{y}$, linear combination

Hence $\underset{\sim}{b}$ is Gaussian, since a linear combination of Gaussian random variables ($\underset{\sim}{y}$) is also Gaussian.

$$E(\underset{\sim}{b}) = E(\underset{\sim}{M}\underset{\sim}{y}) = \underset{\sim}{M}E(\underset{\sim}{y}) = \underset{\sim}{M}\underset{\sim}{X}\underset{\sim}{\beta}$$
$$= (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{\beta} = \underset{\sim}{\beta} \quad \text{(unbiased)}$$

$$\begin{aligned} Cov(\underset{\sim}{b}) &= \underset{\sim}{M}\,Cov(\underset{\sim}{y})\,\underset{\sim}{M}' \\ &= \underset{\sim}{M}\sigma^2\underset{\sim}{I}\,\underset{\sim}{M}' \\ &= \sigma^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1} \\ &= \sigma^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1} \end{aligned}$$

873

(5-3)

t distribution   Student's t

$$t = \frac{b_i}{S.E.(b_i)}$$

$S.E.(b_i)$ = (i,i)th element of $s^2_{y.x}(\underline{X}'\underline{X})^{-1}$

t follows a t distribution on $(N-p)$ degrees of freedom <u>when $\beta_i = 0$</u>

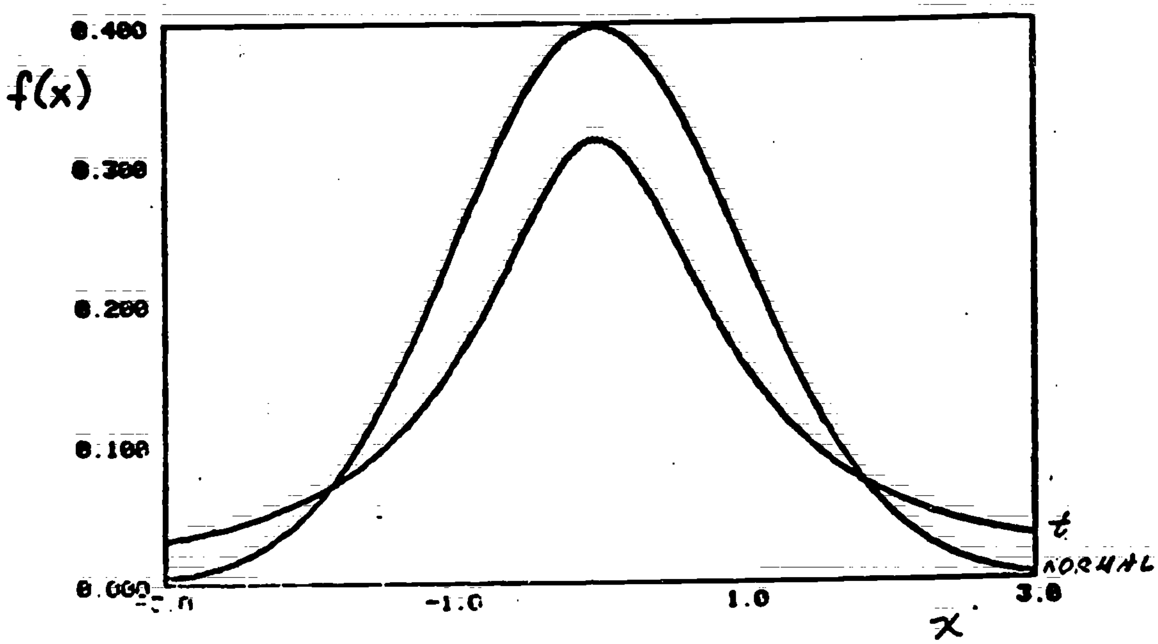
$$f(t) = \frac{\left(\frac{N-p-1}{2}\right)!}{\sqrt{\pi(N-p)}\left(\frac{N-p-2}{2}\right)!} \frac{1}{\left(1 + \frac{t^2}{N-p}\right)^{\left(\frac{N-p+1}{2}\right)}}$$

$$-\infty < t < \infty$$

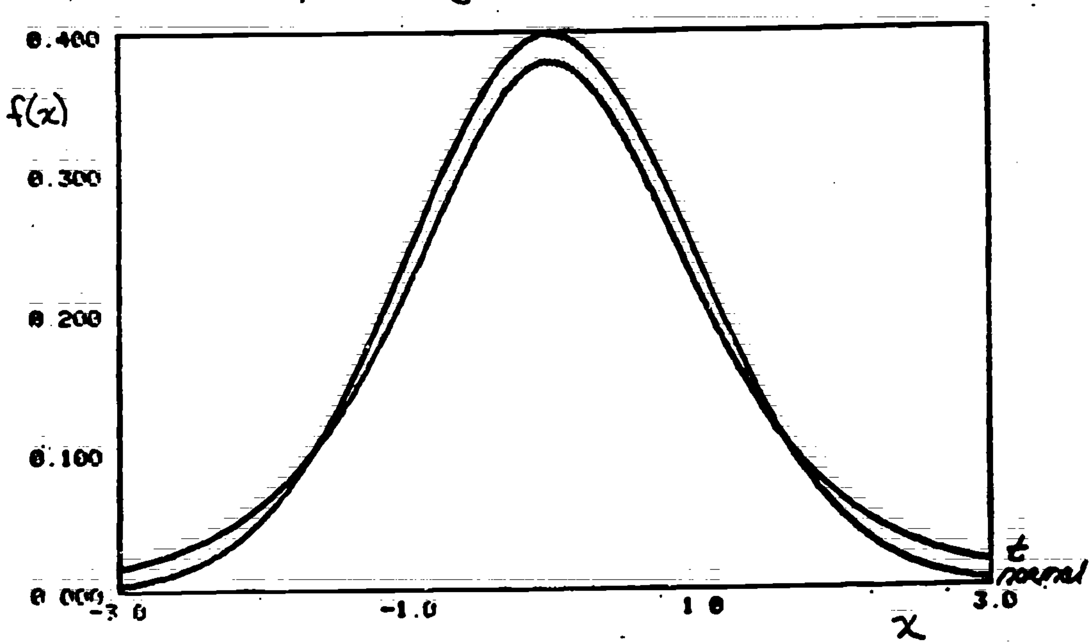
As $N \rightarrow \infty$, $f(t) \rightarrow$ Gaussian $(0,1)$

Comparison of Gaussian with t, 1 degree of freedom

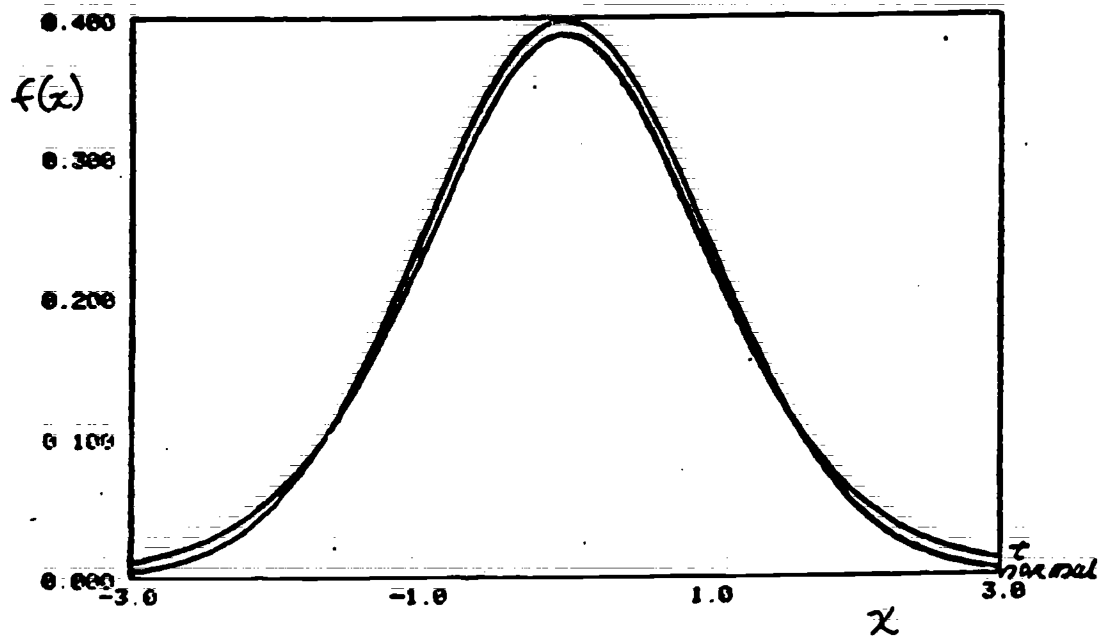
[6a]



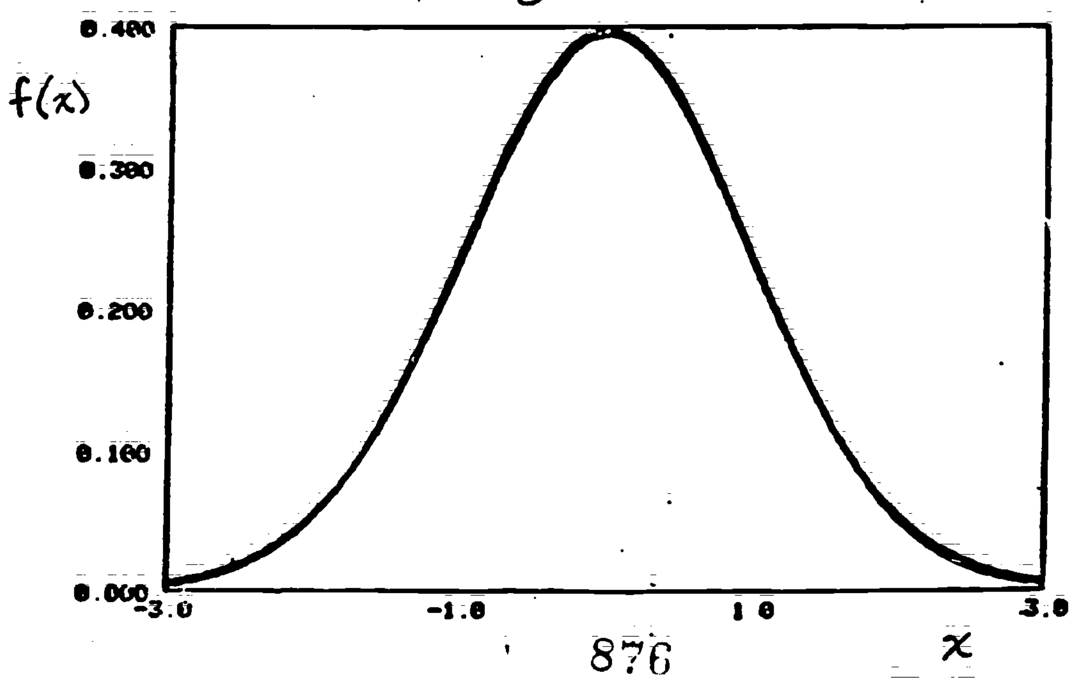
Comparison of Gaussian with t, 5 degrees of freedom



Comparison of Gaussian
with t, 10 degrees of freedom



Comparison of Gaussian
with t, 30 degrees of freedom



[7]

Sample Correlation Coefficient

$$r_{ij} = \frac{\sum_k (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{\left(\sum_k (X_{ik} - \bar{X}_i)^2 \sum_k (X_{jk} - \bar{X}_j)^2\right)^{1/2}}$$

If $\rho_{ij} = 0$ (population value is zero)

Then $r_{ij} \sim Gau\left(0, \frac{1}{N-3}\right)$

This approximation holds only for large N.

[8]

χ^2 and $\mathcal{F}$ distributions

$\frac{1}{N-p} \sum (y_i - \hat{y}_i)^2 / \sigma^2$ follows a χ^2 distribution on $(N-p)$ d.f. (χ^2_{N-p})

(In general, variances (sums of squares of Gaussian rand. vars) follow χ^2 distributions

If $SS_k \sim \sigma^2 \chi^2_k$ and $SS_l \sim \sigma^2 \chi^2_l$,

then $\dfrac{SS_k / k}{SS_l / l}$ follows an $\mathcal{F}$ distrib. with (k, l) d.f.

$\mathcal{F}$ distributions are obtained from ratios of variances

Since R^2 is a ratio of sums of squares, if there is no linear relationship between $\underline{x}$ and $\underline{y}$

$$\frac{R^2/(p-1)}{(1-R^2)/(N-p)} \sim \mathcal{F}_{p-1,\, N-p}.$$

5-3

[9]

Summary Characteristics of Important Distributions

χ^2_n = sum of n squared standard Gaussians

(chi square with n degrees of freedom)
as $n \to \infty$, $\chi^2 \to$ Gaussian

$E(\chi^2_n) = n$, $V(\chi^2_n) = 2n$

$t_f = \dfrac{z}{\left(\frac{\chi^2}{f}\right)^{1/2}}$ (as $f \to \infty$, $\chi^2/f \to 1$ and $t \to N$)
(if $f > 30$ t and z are very close)

[f = degrees of freedom]

$F_{m,n}$ = ratio of two χ^2 divided by their degrees of freedom = $\dfrac{\chi^2/m}{\chi^2/n}$

as $n \to \infty$, $F \to \chi^2/m$
as $m, n \to \infty$ $F \to 1$

5-3

Homework Problems
Unit 5

1. Let N represent a non-response to a mailed questionnaire and R represent a response. We mail questions to four people on a specific day.

   a) How many elements are in the sample space if we are interested in the number of responses to the questionnaire mailed on the given day? List them. Are these outcomes equally likely?

   b) How many elements are in the sample space if we are interested in the response to each questionnaire mailed on the given day? (Each questionnaire is distinctly identified by a code number.) List them. Depict the set of outcomes representing responses to three out of the four questionnaires. Depict the set of outcomes representing at most one response.

   c) What is the probability that the non-responses represent questionnaires lost in the mail?

2. An econometric model predicts whether the GNP will increase (i), decrease (d), or remain the same (s) in the following year. The GNP will then be observed to increase (I), decrease (D), or remain the same(S).

   a) List all the elements of the sample space.

   b) Depict the events that the model predicts correctly.

   c) Depict the events that the model predicts correctly using a Venn Diagram (Hint: first consider the possible outcomes (part a) as a matrix.)

3. According to accident reports, 25% of all accidents which occurred while equipment was being used were caused by faulty equipment and 75% by improper use of the equipment. The probability that on a given day an accident will occur while equipment is being used is .05. Use set notation to define the following events as unions, intersections complements, etc. Then calculate the probability that each event will occur on a given day.

   a) An accident occurs caused by faulty equipment.

   b) An accident occurs caused by improper use of equipment.

   c) No accident occurs.

880

d) An accident was caused by faulty equipment, given that an accident has occurred.

e) An accident occurs caused by either faulty equipment or improper use.

f) An accident occurs, but the equipment was found not to be faulty.

4. A jail has 490 inmates. It is known from the records that

300 committed armed robbery
200 committee larceny
50 committed homicide
20 committed armed robbery and homicide
30 committed larceny and homicide
20 committed larceny and armed robbery
10 committed all three crimes

a) Draw the Venn Diagram illustrating this problem.

If we draw an inmate's file at random, what is the probability that the inmate committed:

a) Two, but only two, of the three types of crimes.

b) At least one of the three types of crimes.

c) Homicide, given the inmate committed armed robbery.

d) Homicide, given the inmate did not commit armed robbery

e) Armed robbery or larceny

f) Only one of the three types of crimes

g) Arson

5.

Age at Marriage, Husband and Wife, New Haven, Conn.

Age of Husband	Age of Wife								Total
	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50 +	
15-19	42	10	3						55
20-24	153	504	51	10	1				719
25-29	52	271	184	22	7	2			538
30-34	5	52	87	69	13	5			231
35-39	1	12	27	29	21	2	3		95
40-44		1	9	18	17	8	2	1	56
45-49	1		3	6	16	16	7	1	50
50 and over			1	4	11	15	21	43	95
Total	254	850	365	158	86	48	33	45	1,839

Source: A. B. Hollingshead, "Cultural Factors in the Selection of Marriage Mates," American Sociological Review 15, 1950, p. 622.

a) Are the ages of husband and wife independent? What striking fact about these data immediately answers this question?

b) What is the probability that one partner was between the ages 30-34?

c) What is the probability that one partner was between the ages of 30-34 and the other was between the ages of 20-24?

d) What is the probability that the husband was between the ages of 30-34 and the wife was between the ages of 20-24?

e) What is the probability that the wife was between the ages of 20-24 given that the husband was between the ages of 30-34?

f) What is the probability that both partners were at least 45 years old?

g) What is the probability that at least one partner was at least 45 years old?

h) What is the probability that neither partner was at least 45 years old?

6. You are working on your annual report for the mayor of a small city. The fire department reports that last year they responded to the following number of false alarms per week:

# weeks	# false alarms per week
1	0
4	1
7	2
10	3
10	4
8	5
7	6
3	7
1	8
1	9
52	

a) Use a stem and leaf display to identify the distribution which the data follow.

b) What is the average number of false alarms per week?

c) What is the standard deviation of the number of false alarms per week?

d) Use the answers to (b) and (c) in the probability function you specified in (a) to verify your choice of distribution (just calculate two of the theoretical number of occurrences, i.e. $P\{X = 0\}$, $P\{X = 1\}$).

e) If each false alarm costs the city $1600, what should you budget for false alarms for the first three months for next year?

883

7. As chairperson of a public service organization's trust fund you are considering buying one of two stocks, both currently priced at $46 per share, for a one month trading venture. You have estimated the probability distribution for the closing prices of the two stocks (rounded to the nearest dollars) one month hence as follows:

Stock A		Stock B	
Closing price	f(price)	Closing price	f(price)
44	.1	44	.005
45	.1	45	.015
46	.1	46	.030
47	.1	47	.100
48	.1	48	.350
49	.1	49	.350
50	.1	50	.100
51	.1	51	.030
52	.1	52	.015
53	.1	53	.005

a) Find the expected value of one share of each stock.

b) Find the variance of the price of one share of each stock (Financial analysts often refer to variance as "risk").

c) Which stock would you purchase and why?

8. a) As a city manager, you stop off to visit the 7th Precinct Police station to look at the crime statistics (crimes per day) of the past 3 months. Assuming a "typical" period (not a crime wave), and discounting the "full moon" theory, to what distribution to you expect the data to conform?

b) Later you enter the comptroller's office to pick up some financial data. After fitting a regression to this data, you plot the residuals and determine that the regression fits remarkably well. To what distribution do the residuals conform?

c) Stopping for lunch at a hamburger joint (since city managers can't afford real food) you pass the time waiting in line by noting how long each customer takes to be served. The statistician in you immediately recognizes that these data fit a distribution, which you rush off to plot. What distribution caused you to miss lunch?

9. a) When you return to your office you resume work on the financial data. Looking at the daily expense account sheets for the past year, you can't help wondering about the distribution of the last digit (the "unit" digit denoting single dollars), so you make a plot. What distribution do you expect this data to follow?

b) Finally getting to work, you correct the errors in the financial reports. What distribution do you expect the number of errors per report to follow?

c) Late in the afternoon you visit a new housing project. The construction supervisor tells you that he has found 9 faulty valves in the lot of 96 acquired for the site. Since you expect to need another 200 lots (of 96 each) over the next 6 months, you need a rough estimate of how many additional valves to order to replace the faulty ones. What distribution do you expect the number of faulty valves to follow?

d) Finally, after a hard day running around and using your profound quantitative skills, you retire to your favorite nightspot. However, your acute mind does not fail to notice the number of mugs of draught beer ordered by the customers. You repeat this exercise every night for a month, except for Sundays, when you stay home to view "Masterpiece Theater" and "Evening at Symphony" with the BSO, to catch up on your cultural events. What distribution do you expect the beer consumption data to follow?

10. The following are the earnings for the city hall staff for the week of February 4, 1977:

earnings for week ended February 4, 1977	#employees with given earnings
187.50 to 194.99	2
195.00 to 202.49	7
202.50 to 209.99	9
210.00 to 217.49	14
217.50 to 224.99	10
225.00 to 232.49	6
232.50 to 240.00	2

a) Is the underlying distribution (of weekly earnings) discrete or continuous? Why?

b) Compute the mean of the above distribution.

c) Is the answer in (a) the same as you would have obtained had you calculated the ratio:

$$\frac{\text{total earnings for all city hall staff during the week ended 4 Feb. 77}}{\text{total \# city hall staff during the week ended 4 Feb. 77}}$$

Why or why not?

d) Compute the median of the above distribution.

e) In which direction is the data skewed?

f) The city comptroller stated that the total payroll for the week ended 4 Feb. 77 was $10675.18. Do you have any reason to doubt this statement? Support your position very briefly.

g) Is the mean computed in (a) a satisfactory description of the "average" or typical earnings of these 50 employees in the week of 4 Feb. 77? Why or why not?

888

QMPM

Homework Unit 5
Solutions

1(a) There are 5 possible outcomes {0, 1, 2, 3, 4}. They are (probably) not equally likely, but we do not know for certain.

(b) There are $2^4 = 16$ outcomes (4 questionnaires each with 2 possible outcomes; N = no response, R = response):

{NNNN, NNNR, NNRN, NNRR, NRNN, NRNR, NRRN, NRRR, RNNN, RNNR, RNRN, RNRR, RRNN, RRNR, RRRN, RRRR}

{3 responses out of 4} = {RRRN, RRNR, RNRR, NRRR}

{at most one response} = {NNNN, NNNR, NNRN, NRNN, RNNN}

(c) Without data from a carefully planned and correctly implemented experiment, this question cannot be objectively answered. Remeber that there are many reasons for a non-response.

2(a) {(i,I), (i,S), (i,D), (s,I), (s,S), (s,D), (d,I), (d,S), (d,D)}

(b) {model predicts accurately} = {(i,I), (s,S), (d,D)}

(c)

	d	s	i
I	(d,I)	(s,I)	(i,I)
S	(d,S)	(s,S)	(i,S)
D	(d,D)	(s,D)	(i,D)

shaded area represents event {model predicts accurately} Venn diagram in this instance includes a grid.

3a) P(accident caused by faulty equipment) = .25 (.05) = .0125

b) P(accident caused by improper use) = .75 (.05) = .0375

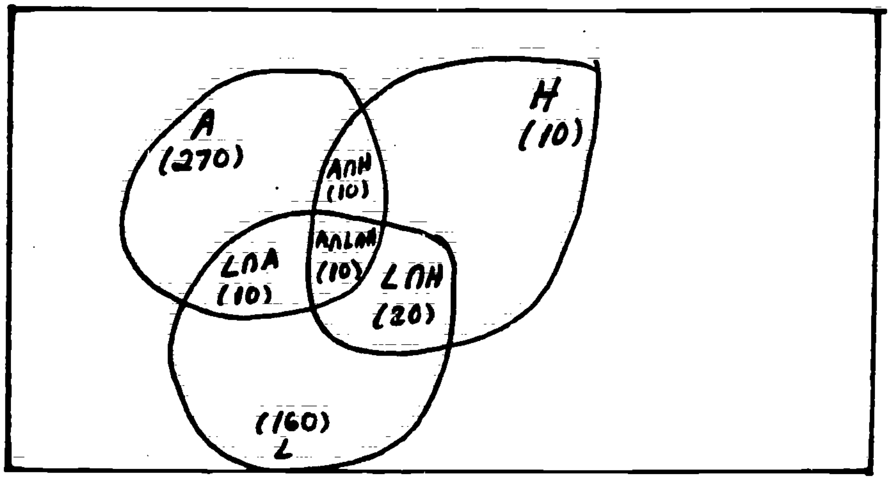
c) P(no accident) = 1 - P(accident) = .95

d) P(faulty equipment/accident) = .25

e) P (accident faulty equipment) (accident improper use) = P(accident) = .05

f) P(not faulty equipment/accident) = P(improper use/accident) = .75

4.



where A = armed robbery
L = larceny
H = homicide

a) $p(\text{2 of the 3}) = p(L\cap A \text{ or } L\cap H \text{ or } A\cap H)$

$= p(L\cap A + p(L\cap H) + p(A\cap H)$, since disjoint.

$= \frac{10}{490} + \frac{20}{490} + \frac{10}{490}$

$= \frac{49}{490} = .08$

b) p(at least one of the 3 types) = 1, since everyone in the jail committed at least one of the three crimes.

c) $p(H|A) = \frac{p(H\cap A)}{p(A)}$

$= \frac{10/490}{300/490} = \frac{10}{300} = .033$

d) $p(H|\bar{A}) = \frac{p(H\cap \bar{A})}{p(\bar{A})} = \frac{30/490}{190/490} = \frac{30}{190} = .16$

e) $p(A\cup L) = p(A) + p(L) - p(A\cap L) = \frac{300+200-20}{490}$

$= \frac{480}{490} = .98$

888

f) p(only one of the three types) = p(A only) + p(L only) = p(H only)

$$= \frac{270}{490} + \frac{160}{490} + \frac{10}{490}$$

$$= \frac{440}{490} = .90$$

g) p(arson) = 0, no individual was assumed to have committed arson.

5. a) It is clear that the ages of husband and wife are <u>not</u> independent, since the counts cluster along the diagonal from upper left to lower right, with many empty cells in the other corners.

b) P(one partner between 30-34) = P(wife between 30-34) + P(husband between 30-34) - P(both between 30-34) = $\frac{158 + 231 - 69}{1839}$ = $\frac{320}{1839}$ .174

c) P(husband between 30-34 wife between 20-24) + P(wife between 30-34 husband between 30-34) = $\frac{52}{1839} + \frac{10}{1839} = \frac{62}{1839}$ .03

d) P(husband between 30-34 wife between 20-24) = $\frac{52}{1839}$ .028

e) P(wife between 20-24/husband between 30-34) = $\frac{52}{231}$ .225

f) P(husband 45-49 50+) (wife 45-49 50+) = $\frac{7 + 1 + 21 + 43}{1839}$ = $\frac{72}{1839}$

g) P (wife 45-49 50+) (husband 45-49 50+) = P(wife 45-49) + P(wife 50+) + P(husband 45-49) + P(husband 50+) - P(wife & h P(wife and husband 45-49) - P(wife 45-49 and husband 50+) - P(wife and husband 50+) = $\frac{33 + 45 + 50 + 95 - 7 - 1 - 21 - 43}{1839}$ = $\frac{151}{1839}$ .082

h) P(neither partner was 45 or older) = 1 - P(at least one partner was 45 or older) = 1 - .082 = .918

6. (a) the data arepoisson, $\lambda = 4$

   (b) 4 false alarms per week (λ)

   (c) variance of a poisson = λ; standard deviation is therefore $\sqrt{\lambda}$ or 2

   (d) 3 months x 4-1/3 weeks/month = 13 weeks x 4 false alarms/week = 52 expected false alarms

   52 x 1600 = \$83,200 is the minimum which should be budgeted. To insure that the department does not run short, more should be budgeted (probably enough for another 2 false alarms (one standard deviation) per week

7. (a) E(A) = E(B) = \$48.50

   (b) Var(A) = 8.25

   Var (B) = 1.57

   (c) Assuming other factors equal, since the "risk" of B is less than that of A with the same expected value, B is preferred.

8. (a) Poisson (or possibly uniform)

   (b) Gaussian

   (c) exponential

9. (a) uniform

   (b) poisson

   (c) binomial

   (d) Gaussian (or possibly rectangular, maybe even Poisson)

10. (a) While the underlying distribution is technically discrete (since fractional cents are not permitted) the measurement \$.01 is so small that we usually consider such distributions to be essentially continuous.

(b) $\bar{X} = \frac{10,680}{50} = 213.60$

(c) No. The ratio gives the exact mean while part (b) gives only a close approximation, since a frequency table was used.

(d) Median class is 210.00 to 217.49. Median observation is $\frac{50+1}{2} = 25.5$ the observation. Median = $\$210 + \frac{25-18}{14}\ \$7.50$ = \$213.75 (interpolation)

(e) By comparing the mean and median, we note that the data are slightly skewed to left, but for practical purpose it is symmetrical.

(f) No, since $50(\bar{X}) = \$10,680$. This figure is an <u>estimate</u> of the total payroll which is close to the figure of \$10,475.18

(g) Yes, the mean here seems typical since there is little skewness to distort it.

891

Quiz
Unit 5

Name ______________________________

Point values are given in parentheses, preceding every question. You have sixty (60) minutes to complete this quiz.

Please write all your answers on these pages, in the space provided. Answers should be brief and succinct: clearly expressed.

When appropriate, answers may be left in fractional form, e.g., 693/721.

(20) 1. You have constructed a univariate linear regression model relating the response variable, miles per gallon for 1977 Volkswagon Rabbits, and the carrier variable, tire pressure per square inch. You have only 15 paired observations to estimate the parameters a and b in the following model:

$$y_i = a + b\,x_i; \quad i = 1, 2, \ldots 15.$$

The least squares estimate of b is -0.16, very nearly zero.

a) What distribution does the quantity

$$\frac{-0.16}{s\left(\sqrt{\Sigma(X_i-\bar{X})^2}\right)^{-1}}$$

follow, where $s^2 = 1/13\,\Sigma(y_i-\hat{y}_i)^2$? Sketch the shape of its probability function.

b) Is it correct to assume that this quantity is Gaussian?

c) When can this assumption accurately be made?

(10) 2. As an employee of Pennsylvania Department of Transportation, you are concerned with structural faults in steel I-beams used to construct bridges in the Pittsburgh metropolitan area. You inspect a shipment of I-beams from the International Steel Company. You are told that the probability of a fault in any given I-beam is 0.0005. What distribution do you expect the number of defective beams to follow? If there are $N = 4000$ I-beams in the shipment, what are the mean and variance of X = number of defective I-beams in the shipment?

(10) 3. You are interested in the traffic flow off the 6th, 7th, and 9th Street Bridges into the North Side of Pittsburgh. On a specific Friday afternoon between 4 and 6 p.m., you record the time in seconds between cars as the cars drive off the 9th Street Bridge into the North Side.

You compute the average waiting time between automobiles to be 10 seconds.

Sketch the most likely probability function for the waiting times. With what random variable is this function associated?

(15) 4. In studying the records of Aggravation Airlines, it has been found that the actual arrival time of the scheduled 5:00 p.m. flight from Philadelphia to Pittsburgh, due in at 6:00 p.m., is a uniformly distributed random variable in the range of 6:00 p.m. to 7:30 p.m. Let $X=1$ represent 6:00 p.m., $X=2$ represent 6:01 p.m., etc.

a) Write out the mathematical expression for $f(X)$.

b) What is the probability that the plane will be late?

c) What is the probability that it will be more than 1 hour late?

(5) 5. The number of potholes along a 100 yard stretch of the Parkway East is a Poisson random variable with a mean of 40. What is the probability that along a specific 100 yard length there are <u>no</u> potholes? Leave answer in terms of a power and multiple of e.

(20) 6. Consider the following data, giving the number of people arrested, by race and age, in 1976 in a small town in Ohio:

	White		Black	
Age	Arrested	Population	Arrested	Population
15-24	378	27,000	65	5,000
25-44	324	36,000	32	4,000
45-74	108	27,000	3	1,000
Totals	810	90,000	100	10,000

Assuming that an individual is drawn at random, find the probability of the following events:

a) A person is arrested.

b) A person who is arrested is white.

c) A black person is between the ages of 25-74.

d) A person is arrested, given that he or she is 25 or over.

e) A person who is arrested, is black and aged 45-74.

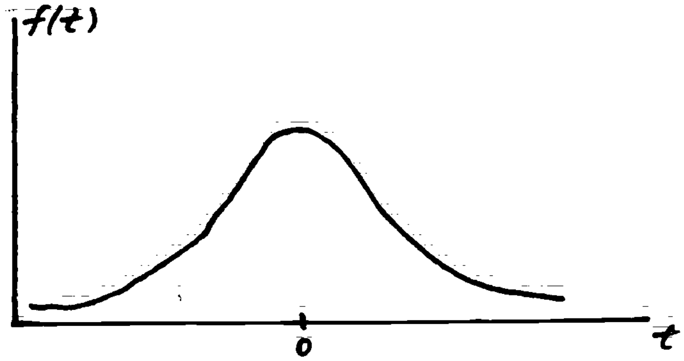
f) A person, who is arrested, is aged 25-44.

(10) 7. You have a large population of individuals and have recorded the 1976 Federal Income Tax paid by each individual. You break the population into 200 separate batches, and compute the mean and variance of the tax payments for each batch. Theoretically, what probability distribution should the variance of the tax payments follow?

(10) 8. In the article "The Use of Subjective Probability Methods in Estimating Demand", by Hanns Schwarz, what is subjective probability, and how does it differ from probability defined as relative frequencies ? How does Schwarz use subjective probability to get reasonable estimates of demand ?

Quiz, Unit 5
Solutions

1. a) t distribution, on 13 degrees of freedom.

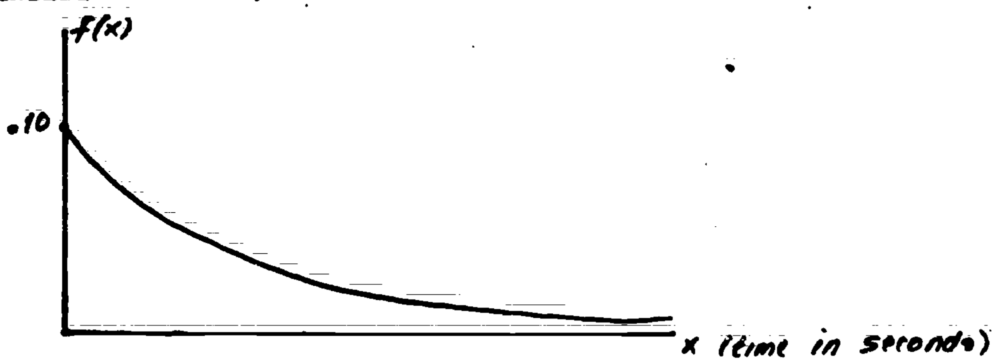


b) No.

c) Never can this assumption be exact. However, when $n > 30$, approximating t by the Standard Gaussian distribution is generally acceptable.

2. Binomial, N=4000, p=.0005
$\mu = Np = 4000\ (.0005) = 2$
$\sigma^2 = Np(1-p) = 2(.9995) = 1.9990$

3. Exponential 1/θ = 10



4. a) f(X) = 1/91; X = 1,2,...,91.

b) 90/91

c) 30/91

5. e^{-40}

896

6. a) 910/100,000 = 91/10,000

   b) 810/910 = 81/91

   c) 5000/10,000 = 1/2

   d) 467/68,000

   e) 3/910

   f) 356/910

7. Multiple of a X^2 random variable

8. Subjective probability is derived from personal <u>opinions</u> about events that have occurred or will occur, rather than strictly from direct observations of past events.

   Schwarz chooses arbitrary weights for opinions about demand (will buy, may buy, won't buy) according to such factors as purchaser and expected date of purchase.

Reading Assignments
Unit 6

Lecture	Assignment
Lecture 6-0	Mosteller, Rourke, and Thomas, Sections 9-1 - 9-4
Lecture 6-1	Mueller, Schuessler and Costner Chapter 13 Mosteller, Rourke, and Thomas, Sections 10-1 - 10-4
Lecture 6-2	Mueller, Schuessler and Costner Chapter 14 Mosteller, Rourke, and Thomas, Chapter 12

In addition please read the following articles:

Tanur, pages 220-8

Tufte, pages 285-351, 391-406

Texts:

Mosteller, F., et. al., Probability with Statistical Applications, Second Edition. Reading. Massachusetts: Addison-Wesley, 1970.

Mueller, J. H., et.al., Statistical Reasoning in Sociology, Third Edition, Boston: Houghton Mifflin, 1977.

Tanur, J., et.al., editors, Statistics: A Guide to the Unknown, San Francisco: Holden Day, 1972.

Tufte, E. R., editor, The Quantitative Analysis of Social Problems, Reading, Massachusetts: Addison-Wesley, 1970.

Lecture 6-0. Introduction to Unit 6

Introduction to Unit 6. Statistical Inference

<u>Lecture Content</u>:

1. Introduction to the objectives, problem, and notation of Unit 6

<u>Main Topics</u>:

1. Specific Introduction to the Objectives of Unit 6
2. Presentation of General Problem of Unit 6
3. Notation for Unit 6

Topic 1. Specific Introduction to the Objectives of Unit 6

I. Questions to be answered in Unit 6

1. Is it ever possible to study an entire population?

   a. Such a complete study is called a <u>census</u>; every individual in the population is sampled

   b. Usually, the researcher only has the opportunity and ability to study a fraction of the population, called a <u>sample</u>

   c. The findings of the study apply only to the sample; the sample statistics <u>estimate</u> the true population value

2. How much can we infer from these findings, in our effort to study the entire population?

   a. This process of "extending" our analytical results is called <u>statistical inference</u>

   b. The example illustrates the problem (1)

3. Can we quantitatively assess how good an estimate is?

II. Skills to be mastered in Unit 6 (2)

1. Calculation of probabilities using the Gaussian probability function

2. Making estimations about the values of parameters in the population

3. Placing intervals around these estimates to give a range of possible parameter values

4. Testing relationships concerning the variables in the population

900

Topic 2. Introduction to the Problems of Unit 6

I. What is statistical inference?

1. For a specific batch of data, obtained as a sample from a larger population, we compute various statistics:

   a. p = proportion

   b. $\overline{X}$ = sample mean

   c. M = sample median

   d. r = sample correlation coefficient

2. These quantities are estimates of true population values, called <u>parameters</u>

   a. $\overline{X}$ estimates μ

   b. r estimates ρ

   c. p estimates P

3. Statistical inference is concerned with how well these statistics estimate population values

   a. How much "faith" can we have in any given estimate?

   b. Our notion of "faith" will be quantified through the use of probability, especially probabilities from the Gaussian distribution

4. Statistical inference makes the risk associated with the use of a specific statistic explicit and known

II. Calculation of Gaussian probabilities

1. Suppose that X is a Gaussian random variable, with mean μ and variance σ^2

   a. Transparency shows f(X), and probabilities associated with # standard deviations from μ. (3

   b. We desire to compute $P\{a \leq X \leq b\}$ for some $a < b$.

2. We standardize X to a standard Gaussian random variable, and then use tabulated values of this standard distribution

3. For example: (4)

$$P\{a \leq X \leq b\} = P\{\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\}$$

$$= P\{Z \leq \frac{b-\mu}{\sigma}\} - P\{Z \leq \frac{a-\mu}{\sigma}\}$$

where $Z \sim$ Gau $(0,1)$

4. Transparency shows Tables of Standard Normal Random Variable. (5)

$P\{0 \leq X \leq a\}$ for $a > 0$.

5. Remember:

a. $P\{X \leq a\}$ where $a < 0$

$= P\{X \geq -a\}$ by symmetry

b. $P\{X \geq a\} = 1 - P\{X \leq a\}$

c. $P\{-\infty < X \leq a\}$, $a \leq 0$

$= .5 + P\{0 \leq X \leq a\}$

(Work several examples using the Tables)

Topic 3. Notation for Unit 7

I. Conventions

1. Population values denoted by Greek letters

   μ, σ^2, ρ, β

2. Sample estimates denoted by Latin letters

   $\bar{X}$, s^2, r, b

II. Variables

1. X, Y denote variables; x, y are realizations

903

Lecture 6-0
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Topic 1		
Section I.		
2.b	1	Inference Problem
Section II.		
1.	2	Skills to be Mastered
Topic 2		
Section II.		
1.a	3	Gaussian Probability Distribution
3.	4	Calculating Gaussian Probabilities
4.	5	Normal Curve Areas

[1]

Inference Problem

As an employee of Datatown International Airport Planning Committee, you are studying the location of the coffee shops within the airport complex.

The question to be answered is whether one or more additional coffee shops should be opened to reduce the waiting lines at the sole restaurant now operating.

You begin your analysis by going through restaurant records to record the prices of meals, ordered by the patrons.

Of course it is not feasible to do this for all customers (records are not computerized)

Therefore you must sample.

You compute $\overline{X}$ = average price per meal.

How well does $\overline{X}$ estimate μ, the population value?

[2]

Skills to be mastered in Unit 7

1. Calculation of probabilities using the Gaussian probability function

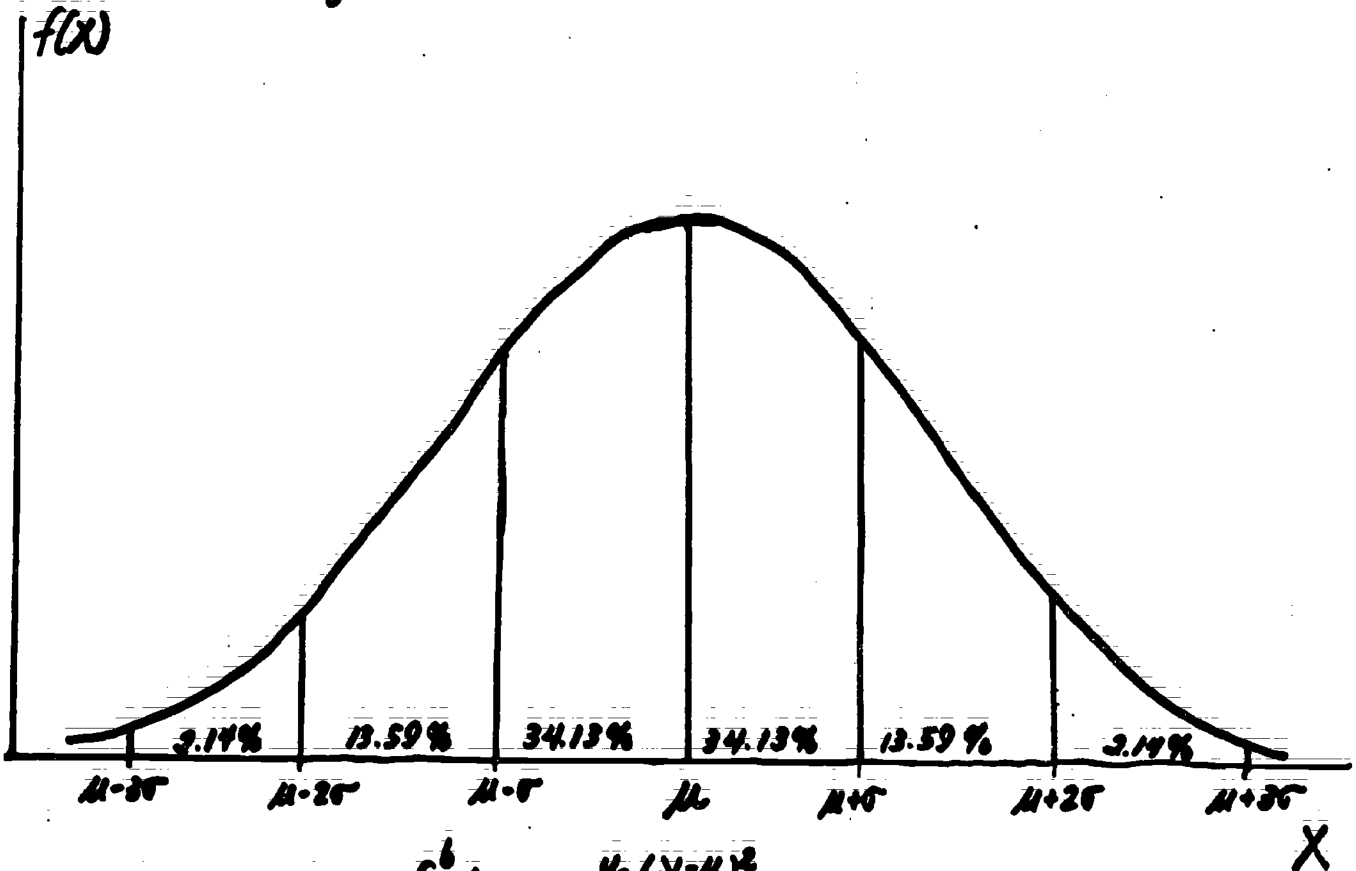
2. Making estimations about the values of the parameters in the population

3. Placing intervals around these estimates to give a range of possible parameter values

4. Testing relationships concerning the variables in the population

Gaussian Probability Distribution [3]

QMPM



$$P\{a \le X \le b\} = \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy$$

906

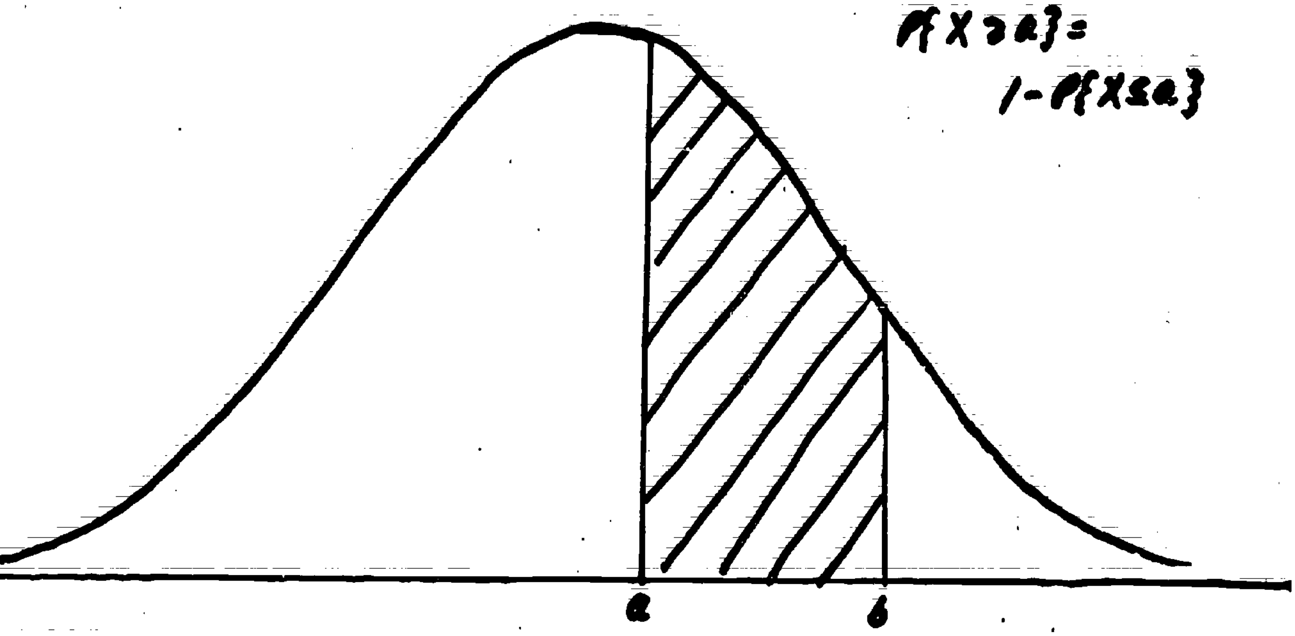
07

6-0

alculating Gaussian Probabilities

$\{a \le X \le b\} = P\{X \le b\} - P\{X \le a\}$

Note:
$P\{X > a\} =$
$1 - P\{X \le a\}$



$\{X \le b\} = P\left\{\frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right\}$

$\frac{X-\mu}{\sigma}$ is a Gaussian (0,1) R.V.

Probabilities for Gau(0,1) found in Standard Normal Tables

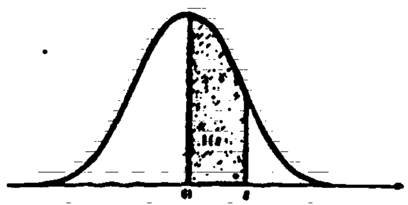
Standard Normal Table

Normal Curve Areas

Area under the standard normal curve from 0 to z, shown shaded, is $A(z)$.

Examples. If Z is the standard normal random variable and $z = 1.54$, then

$A(z) = P(0 < Z < z) = .4382,$
$P(Z > z) = .0618$
$P(Z < z) = .9382,$
$P(|Z| < z) = .8764$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Lecture 6-1. Quantifying Uncertainty of Estimates

Quantifying Uncertainty of Estimates with Confidence Intervals: Interval **Bounds Between** which the Population Parameter will fall with a specified frequency.

Lecture Content:

1. Sampled data and parameter estimates
2. Quantifying certainty (or uncertainty) in our estimates

Main Topics:

1. Notion of Confidence Intervals
2. Calculating Confidence Intervals

(There are no transparencies for this lecture.)

Topic 1. Notion of Confidence Intervals

I. General Problem: Sampled data yield estimates of population parameters which will not be equal to the parameter. How can we quantify the certainty (or uncertainty) that we have in our estimate?

II. Solution: Confidence intervals--interval bounds between which the population parameter will fall with specifiable probability.

III. Skills to master: Confidence intervals for mean, regression coefficient, correlation coefficient, proportion.

IV. Specific Notions

1. Inference goes from sample to population. We measure a feature of the sample data and infer the value of the population parameter from this. Thus, we call the sample statistic an estimate of a parameter.

2. Parameter is constant. But from sample to sample measured estimate can vary.

3. Each estimate is a value of a random variable whose distribution may be known from theory or assumption.

   Ex: means of samples are Normal in large samples.

4. Point estimate: single value. But this may be in error. In fact, we don't expect it to equal the parameter. Simply reporting the number gives no indication of how close we believe the estimate is to the parameter.

5. Interval estimate: bounds for an interval containing the point estimate (not necessarily symmetric) which we know contains the parameter with certain probability

6. The probability that the interval covers the parameter is the <u>confidence level</u>. The interval is called a <u>confidence interval</u>.

7. Note that since the parameter falls in the interval with certain probability < 1, it may not actually be in the interval. (95% confidence means 1 in 20 chance of being wrong).

Topic 2. Calculation of Confidence Intervals

I. Specific Methods

1. The normal approximation for $\overline{X}$.

   The population mean, μ is estimated by the sample mean, $\overline{X}$, an unbiased estimate of μ since $E(\overline{X}) = \mu$.

2. What is its "sampling" distribution?

   From statistical theory (central limit theorem), regardless of distribution of original x's, frequency distribution of $\overline{X}$ in repeated random samples of size n tends to the Normal as $n \to \infty$

   (Note that the closer to normal the original distribution, the smaller n can be for using normal approximation. Others may require $n >> 100$. The more skewed x, the larger n should be)

3. What is its standard deviation? (Often called standard error because it indicates the amount of error in using $\overline{X}$ as a measure of μ).

   $$\sigma_{\overline{X}} = \sigma/\sqrt{n}$$

4. Since $\overline{X}$ is normally distributed (for repeated random samples and large n) and we know its expected value and standard deviation, we can construct a standardized normal deviate:

   $$z_{\overline{X}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

5. Now we can specify the probability that $\overline{X}$ lies between two limiting values L_1 and L_2 by determining the probability that $Z_{\overline{X}}$ lies between L_1 and L_2 and this we do by examining a table of the percentage points of the cumulative normal.

II. Confidence intervals for μ: σ known.
(Example: IQ scores)

1. Random sample of size n, $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$

If Y is drawn from $N(\mu,\sigma)$

$$P\{\mu - 1.96\sigma \leq y \leq \mu + 1.96\sigma\} = .95$$

Thus, if $\bar{X}$ is drawn from $N(\mu,\sigma/\sqrt{n})$

$$P\{\mu - 1.96\sigma/\sqrt{n} \leq \bar{X} \leq \mu + 1.96\sigma/\sqrt{n}\} = .95$$

or

$(\bar{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1.96\sigma/\sqrt{n})$ is our interval.

Thus the 95% confidence interval for μ is

$$\bar{X} \pm 1.96\sigma/\sqrt{n}$$

Since ± 2.58 contains 99% of the standard normal, 99% conf. limits are

$$\bar{X} \pm 2.58\sigma/\sqrt{n}$$

In general, p% confidence levels are

$$\bar{X} \pm Z_p\sigma/\sqrt{n}$$

where Z_p is a value in the cumulative normal table such that the area .

(One sided tests use Z_p such that area is P.)

2. Sample size

We want estimate accurate to $\pm L$ set probability of $\bar{X}$ lying between $\pm$ L = .95, say.

Then

$$P\{\mu - L \leq \bar{X} \leq \mu + L\} = .95$$

But $P\{\mu - 1.96\sigma/\sqrt{n} \leq \bar{X} \leq \mu + 1.96\sigma/\sqrt{n}\} = .95$

Thus $L = 1.96\sigma/\sqrt{n}$.

Making 1.96 ≈ 2 we have

$n = 4\sigma^2/L^2$ for 95% prob.

$n = 6.6\sigma^2/L^2$ for 99% prob.

3. σ unknown

a. Use s as estimate of σ. It is based on (n-1) degrees of freedom. Now

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

t differs from N(0,1) only when $n \leq 30$.

b. P% confidence intervals

$= \bar{X} \pm t_{p/100 \cdot 1/2}\, s/\sqrt{n}$ where t has n-1 d.f.

4. Regression coefficient

$$t = \frac{b_i - \beta_i}{S_{b_i}} \qquad \text{on } n\text{-}k \text{ df}$$

How is S_{b_i} computed?

$$\text{var}\,(\hat{\beta}) = \frac{\sigma^2}{\Sigma(X_i - \bar{X}_i)^2}$$

with univariate regression, through the origin.

usc residual variance

$$S_{b_i} = \frac{\sqrt{\dfrac{\Sigma(y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\Sigma(X_i - \bar{X})^2}}$$

(numerator is standard error of regression
denominator is (n-1) standard error of X_i)

916

$$= \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{(n-2)\ \Sigma(X_i-\bar{X})^2}}$$

Then C% confidence intervals are

$$\beta_i = \hat{\beta}_i \pm t_{c/100\cdot 1/2}\ S_{b_i}$$

5. Correlation coefficient

Fisher's $r \rightarrow Z$ transformation

$$Z = \frac{1}{2}\ [\log_e\ (1+r)\ - \log_e(1-r)]\ \sim N$$

$$\sigma_z = \frac{1}{\sqrt{(n-3)}}$$

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2\Sigma(Y-\bar{Y})^2}} = \frac{Cov(X,Y)}{\sigma_x\sigma_y}$$

Then p% confidence interval:

$$\rho = Z \pm Z_{p/100\cdot 1/2}\ \sqrt{(n-3)}$$

6. Proportion

If r of n have attribute proportion in population, $\hat{p} = r/n$. When $n >> 30$, $\hat{p}$ is $N(p, PQ/n)$; use $\hat{p}$ and $\hat{q}(=1-\hat{p})$ to estimate p,q and c% confidence interval

$$p = \hat{p} + Z_{c/100\cdot 1/2}\ \sqrt{\hat{p}\hat{q}/n}$$

when n is small use

$$P = \hat{p} \pm (Z_{c/100}\cdot 1/2\ \sqrt{\hat{p}\hat{q}/n} + 1/2n)$$

(Careful when p and q are not near .5 and n is < 75)

Lecture 6-2. Significance Testing

Significance Testing: Determining the reasonableness of a hypothesis

Lecture Content:

1. Null hypotheses: H_0
2. Determining whether to reject or not reject H_0
3. Significance Levels

Main Topics:

1. Significance Tests: Concepts
2. Significance Tests: Techniques

Topic 1. Significance Tests: Concepts

I. Problem--Sometimes a point value for a population parameter is assumed or hypothesized. How can sampl[illegible] results be used to test the reasonableness of the hypothe[illegible]

II. Solution--Significance test--calculate a test criterion from the sample data; if it falls into a region of rejection, the null hypothesis, i.e., the hypothesized value for the parameter, is rejected and the departure is called statistically significant. If the null hypothesis is true the test has a known probability of obtaining a significant result which is called the significance level of the test.

III. Notion of a null hypothesis

A. Considerations

1. This is a statistical hypothesis, an assertion that the population parameter has a certain value. It is called null because the assertion is that there is no difference between the hypothetical value and the parameter's actual value. This is nonetheless hypothetical because we have no evidence (yet) that the hypothesized and true values are equal.

2. This leads to a decision making situation. We want to construct a procedure with which we contrast the null hypothesis with evidence drawn from sample data.

3. If the value computed from the data is "very different" from the null hypothesis we reject it. If it is "similar" we do not reject it.

4. The null hypothesis can describe a single parameter, such as a regression coefficient or a difference in parameters, such as the difference in means.

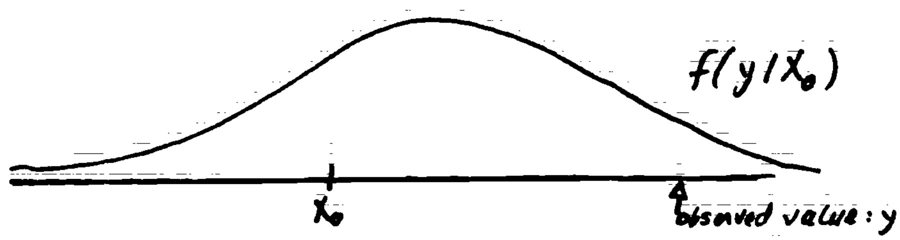
B. Notion of a rejection region

1. How can we specify what is a "very different" value lending to rejection or a "similar" value leading to non-rejection?

2. Use probability. If we know the sampling distribution of the estimate under the null hypothesis then we can compute the probability of observing a value like that computed from the sample data. (Exactly, in fact.)

3. When this probability is quite small, we can argue that that the observed value is unlikely to be an observation on a random variable with parameter equal to the null hypothesized value. Thus the data lead us to reject the null hypothesis or, we say, they fail to confirm it.

Example:



Y may have arisen from $f(Y/X_0)$ but such a value is almost a rare event in this distribution. We must decide whether the rare event occurred, or whether X_0 seems reasonable as the parameter of the distribution.

4. We can establish levels which bound small areas of probability such that if the observed value falls beyond the inner bounds of the levels we agree to reject the null hypothesis. These are the rejection regions. Between the limits lies the acceptance region.

5. If we have no idea on which side of the null hypothesis the sample value will fall (disjoint alternative hypothesis) we need two levels. If we have a prior idea (one alternative hypothesis) we need only one level. In the first case we have a two tailed test, two rejection regions in the tails of the assumed distribution. In the latter case we have a one tailed test.

C. Notion of power of a test

1. The estimate of the parameter calculated from the sample is an observation on a random variable whose sampling distribution is known when the null hypothesis is true.

2. Since the probability of observing a rare value is small but nonetheless positive sometimes a value falling in the rejection region will be <u>falsely rejected</u>.

3. The probability of rejecting a true null hypothesis is the <u>significance level</u> of the test. It is equal to the area in the rejection regions and is 1 minus the <u>confidence level.</u>

4. The probability of rejecting a false null hypothesis is the <u>power</u> of the test. (1) (2)

D. Types of errors in decision making

		H_0: T	H_0: F
H_0	R	I α	Power $1-\beta$
	not R	$1-\alpha$	II β

To reduce type I error (α) increase confidence, i.e., min. α

To reduce type II error (β) increase power, i.e. min. β

There is a trade off between type I and type II errors. In general when one decreases, the other increases.

Topic 2. Significance Tests: Techniques

I. General procedure:

1. Determine test statistic
2. Establish null hypothesis
3. Determine sampling distribution of test statistic under null hypothesis.
4. Set levels for rejection (or simply report p value) usually .1, .05, .01. (Discuss looking up critical values in appropriate table of the distribution)
5. Perform test

II. Examples

1. For μ; σ known: $H_0 = \mu_0$

   Compute $Z_{\bar{X}} = \dfrac{\bar{X}-\mu_0}{\sigma_{\bar{X}}} = \dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$

   Is $|Z_{\bar{X}}| > Z_\alpha$? If yes, reject H_0 at α .

2. For μ; σ unknown, $n \geq 30$: $H_0 = \mu_0$

   use $s/\sqrt{n}$ to estimate $\sigma/\sqrt{n}$ and proceed as above.

3. For μ;σ unknown, $n < 30$: $H_0 = \mu_0$

   $$t = \frac{\bar{X} - \mu_0}{s_{\bar{X}}} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$ where t has (n-1) df

4. Two means from independent samples

Note:

a. Difference between two normally distributed random variables is normal

b. $\sigma^2_{x_1-x_2} = \sigma^2_{x_1} + \sigma^2_{x_2}$ where x_1 & x_2 are i.i.d.

   Then $\dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}$ is test statistic

c. If the X_i are drawn from some population $n_1 = n_2$ with variance σ^2 known then

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = 2\sigma^2/n$$

With known σ use Z
With unknown σ use

pooled $s^2 = (s_1^2 + s_2^2)/2$; t has $2(n-1)$ df

d. $n_1 \neq n_2$ then $\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sigma^2 \dfrac{n_1 + n_2}{n_1 n_2} = \dfrac{\sigma^2}{n_1} + \dfrac{\sigma^2}{n_2}$

and when σ is unknown

σ unknown use

$$s^2_{\bar{x}_1 - \bar{x}_2} = \frac{s^2}{n_1} + \frac{s^2}{n_2} \text{ where t has } n_1 + n_2 - 2 \quad \text{d.f.}$$

5. Correlation coefficient

Use Fisher's transformation

$$z = \frac{1}{2} [\log_e(1 + r) - \log_e(1-r)]$$

which has $\sigma_z = \dfrac{1}{\sqrt{(n-3)}}$

and proceed as with normally distributed test statistic.

6. Regression coefficient

$$t = \frac{b_i - \beta_0}{s_{b_i}}$$

where t has $(n-k)$ d.f. and k is tot[illegible] number of parameters being estimated.

7. Regression

$$\frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F_{k-1,\, n-k}$$

if <u>no</u> linear relationship exists between $\tilde{y}$ and $\tilde{X}$.

where k is number of parameters estimated and n is sample size.

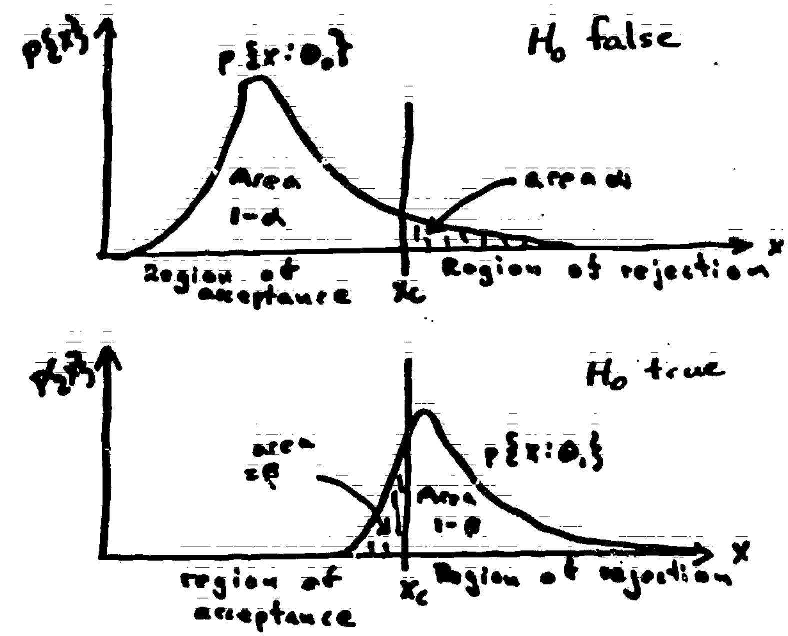
Lecture 6-2
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Topic 1		
Section III		
C.4	1	One Sided Test
C.4	2	Two Sided Test

(1)

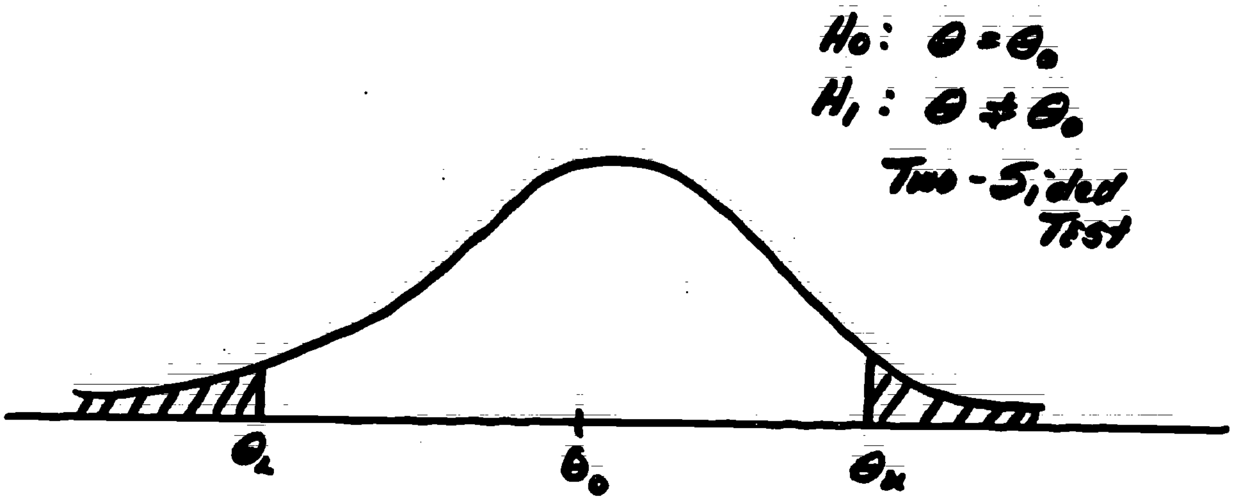
One Sided Test

$H_0: \theta = \theta_0$
$H_1: \theta > \theta_0$



Relation between significance level, power and types of errors in hypothesis testing.

(2)



Reject θ_0 if $\hat{\theta}_0 > \theta_u$ or $\hat{\theta}_0 < \theta_L$

Shaded area = α ($\alpha/2$ in each tail, if symm.)

Pr{Rejecting H_0 when H_0 is true}

θ_u and θ_L are "critical values"

We specify α in advance --

control for Type I error rate

We do not control for Type 2 error rate

β = Pr{Accepting H_0 when H_0 is false}

Want β as small as possible

or $(1-\beta)$ = "Power" as large as possible

926

6-2

Homework
Unit 6

1) Of the 3,017 families in Ellwood City, PA in 1970, a random sample of 300 families was taken to determine the mean family income. A 95% confidence interval ($8,812 to $9,116) was established on the basis of the sample.

Using only the above information comment on the truthfulness of the following statements:

a) Of all possible samples of size 300 drawn from this population, 95% will have sample means between $8,812 and $9,116.

b) Of all possible samples of size 300 drawn from this population, 95% will have population means between $8,812 and $9,116.

c) Of all possible samples of size 300 drawn from this population, 95% of the confidence intervals will contain the true population mean.

d) 95% of the families in Ellwood City have incomes in the range $8,812 to $9,116.

2) One can always decrease the width of a confidence interval by increasing the sample size. Why then does one not always determine the desired width and sample accordingly?

3) Suppose you are interested in the proportion of families in the United States that have 4 or more children. Let the true population proportion be P. Since your office does not have a copy of the current Statistical Abstracts, you are instructed to estimate P based on a very small sample of 1,000 families.

a) Let p be the estimate of P from your sample. What is the (large) sampling distribution of p?

b) Suppose we found a p of .125. Construct a 90% confidence interval about p based on these sampling results. What do you report to your supervisor concerning the true population proportion? In policy context, when would the point estimate p be preferred to the 90% confidence interval?

c) Suppose that n is quite small, and P = .90. Explain why the sampling distribution of p will be asymmetric, and tell your supervisor why the Gaussian approximation is inaccurate in this instance.

4) Suppose that as an employee of HEW, you are studying the effect of the apparent decline in intelligence exhibited by high school upperclassmen on the allocation of federal funds to public schools. You have SAT scores for high school seniors throughout the country for 1966-1975.

It is a well known fact that an individual's score on the Mathematics Scholastic Aptitude Test, administered by the College Entrance Examination Board, is a random variable with mean 500, standard deviation 100; moreover, for all but very specific purposes, it is Gaussian.

a) What is the probability that an individual's score is greater than 700?

b) What is the probability that a score is between 400 and 650?

c) Two sisters each have scores between 500 and 550. What is the probability of the simultaneous occurrence of these two seemingly independent events?

d) Your supervisor states that the simple assumption you used to calculate the probability in (c) (independence) is not at all correct. Why?

5) You are conducting a study for a dean of a highly regarded school of public policy into the ages of incoming master of public administration students. Your data consist of 22 students.

a) You find that $\bar{X}$ = 24.5 years, and S^2 = 2.1 years^2, with n = 22. Construct a 92% confidence interval and interpret your results.

b) In what way(s) is your confidence interval similar to a hypothesis test?

6) The National Training and Development Service has kindly given you data on the evaluation of 195 proposals for curriculum development. Each proposal is submitted to one of 8 need areas.

Your data analysis reveals that page length and the indicator variable for need area 3 are important determinants of the final score awarded to a proposal.

The regression model of total summed score on page length, (page length)^2, and need area 3 was constructed. Results are given below:

	Pages	(Pages)^2	Indicator Need Area 3	Constant
Coefficient	1.540	-.024	-6.686	41.884
Standard error	.271	.0058	2.014	

R^2 = 0.19

p = 195 y = total summed score (Max = 90)

929

a) Test the hypothesis that the model relating y to the 4 carriers is not additive, and hence that no linear regression exists.

b) Place confidence intervals about the least squares coefficients for pages and (pages)^2.

c) Comment on the results of (a) and (b). How can you explain the rather contradictory finding of such a small R^2? Would a stem-and-leaf display of the residuals be useful?

7) In a random sample of 1,000 individuals, 600 were in favor of capital punishment. Test the hypothesis (α = .10) that individual attitudes in the population are equally divided for capital punishment and against it.

8) You have access to the grade reports of 9 students in the class. You find that the sample correlation between undergraduate GPA and fall term OMPM grade is only 0.15. Can you conclude that there is no relationship between these 2 variables?

9) A linear regression model relates the response Emigration from 33 SMSA's with populations greater than 500,000 to 3 carriers: welfare payments per capita, immigration into the SMSA, and average annual temperature.

The results:

	Coefficient	t-statistic	standard error
Constant	-0.1978		
Welfare	0.3324	17.94	0.0185
Immigration	0.0046	2.14	0.00215
Temperature	0.0026	1.34	0.00192

$\sigma^2 = 0.00482$

$R^2 = 0.9394$

Comment on these results by constructing hypothesis tests, with $\alpha = .05$.

10) In a sample of 400 professors, you find that the average annual salary is $23,200, with a standard deviation of $4,000. Test the hypothesis that the population value is $25,000. Let the probability of a Type 1 error be .10.

929

Homework Solutions
Unit 6

1. The only valid statement is (c). This is precisely what we mean by a 95% confidence interval. Remember, we are examining a confidence interval for the one (only one) population mean. We calculate the interval using a random sample.

2. The cost of taking a larger sample may be uneconomical in terms of return on the sample info -- or the sample size may be limited by other factors such as physical, time, moral/ethical, etc. constraints. We usually predetermine n as the largest sample size possible within time, cost, availability, etc. constraints.

3. (a) A normal distribution may be used to approximate the sampling distribution of p. Although the ratio of p:q (q being the percentage of families with fewer than 4 children) is likely to be considerably less than .5:.5, the sample size is sufficiently large to counter any resulting skewedness. (Note that although 1,000 families is a small sample of the total number of families in the country, it is a large sample from the standpoint of developing sampling distributions. The distribution will have an estimated mean (μ_p) and standard deviation (σ_p) of P and $\sqrt{\frac{PQ}{1000}}$ respectively.

(b) We have been given the sample size (1000) and the sample mean (p = .125). We are asked to determine the critical value boundaries (limits of the estimates of the mean) within which we can be 90% sure that the true population mean will fall. The calculations of these values are as follows:

$$Pr\left(p - Z\sqrt{\frac{pq}{n}} \leq P \leq p + Z\sqrt{\frac{pq}{n}}\right) = \qquad Z_{.05} = 1.65$$

$$.125 - 1.65\sqrt{.00011} \leq P \leq .125 + 1.65\sqrt{.00011}$$

$$.108 \leq P \leq .142$$

We prefer a point estimate of P when we need to make decisions based on a particular value of P (e.g. how much should be budgeted to provide a good to every 4 child family.)

(c) The shape of the sampling distribution depends on sample size and the relationship of P to Q. As the ratio of P to Q departs from 1, the distribution becomes increasingly skewed. The greater the skewedness, the more likely it is that the samples means will be distorted. Large samples "smooth out" this distortion so that the sample distribution of the percentage approaches the normal.

930

4. (a) Let X = the individual's score
We want to find Pr(X > 700) which is equivalent to 1-Pr(X ≤ 700)
We can subtract the population mean and divide by the standard deviation on either side of the inequality. Since Z is of the form $\frac{X - \mu}{\sigma}$, the probability determination can readily be made. The calculations are:

$$\Pr(X > 700) = 1 - \Pr(X \leq 700)$$
$$= 1 - \Pr\left(\frac{X - 500}{100} \leq \frac{700 - 500}{100}\right)$$
$$= 1 - .9772$$
$$= .0228$$

(b) By the same logic as above

$$\Pr(400 \leq X \leq 650) = \Pr\left(\frac{400 - 500}{100} \leq \frac{X - 500}{100} \leq \frac{650 - 500}{100}\right)$$
$$= \Pr(-1 \leq Z \leq 1.$$
$$= .3413 + .4332$$
$$= .7745$$

(c) First we find the of occurrence of a score between 500 and 550

$$\Pr(500 \leq X \leq 550) = \Pr\left(\frac{500 - 500}{100} \leq \frac{X - 500}{100} \leq \frac{550 - 500}{100}\right)$$
$$= \Pr(0 \leq Z \leq .5$$
$$= .1915$$

If the events are truly independent then the probability of their simultaneous occurrence is the product of their probabilities. Since both events have probabilities of .1915, the joint probability is: $(.1915)^2$ or .0367.

(d) He is right because the sisters share similar genetic make-up and environmental experience. Both factors can influence intelligence. The events are therefore not independent and the probability of both scores being between 500 and 550 is probably greater than .0367.

931

5. (a) The t-distribution is required here since the sample size is small. The confidence interval is constructed as follows:

$$\bar{X} - t_{.04}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + t_{.04}\left(\frac{s}{\sqrt{n}}\right)$$

$$24.5 - 2.1\left(\frac{1.45}{\sqrt{22}}\right) \leq \mu \leq 24.5 + 2.1\left(\frac{1.45}{\sqrt{22}}\right)$$

$$23.35 \leq \mu \leq 24.65$$

There is a 92% chance that this interval will contain the true mean of ages.

(b) To test whether a given population mean is the same as another a confidence interval may be established. This interval corresponds to the region of "non-rejection" of the null hypothesis.

6. (a) The null hypothesis to be tested is H_0: $R^2 = 0$
The F-distribution is appropriate here.

$$\frac{R^2}{1 - R^2} \cdot \frac{N-p}{p-1} = \frac{.19}{.81} \cdot \frac{191}{3}$$

$$= 14.93$$

$$F_{3,191;.05} = 8.5$$

Reject H_0. There is a linear relationship.

(b) Pages

$$(1.54 - t_{.05}(.271)) \leq \beta_1 \leq (1.54 + t_{.05}(.271))$$

$$(1.54 - 1.96(.271)) \leq \beta_1 \leq (1.54 + 1.96(.271))$$

$$1.00 \leq \beta_1 \leq 2.07$$

Pages ^{2}

$$(-.024 - 1.96(.0058)) \leq \beta_2 \leq (-.024 + 1.96(.0058))$$

$$-.035 \leq \beta_2 \leq -.013$$

(c) Neither confidence interval contains 0. Hence both carriers explain a portion of the total variation. Also, the additive model is consistent with the data since N is large, even though R^2 is small. An examination of residual plots would be informative.

7. Large sample, testing for $\frac{P}{Q} = 1$.

H_0: $P = .5$

H_1: $P \neq .5$

$$Z = \frac{.6-.5}{\sqrt{\frac{(.6)(.4)}{1000}}} = \frac{.1}{.015}$$

$$= 6.67 \qquad Z_{.05} = 1.65$$

Since calculated Z exceeds table $Z_{.05}$, reject H_0

8. H_0: $\rho = 0$

H_1: $\rho \neq 0$

$$Z = \frac{.15-0}{\sqrt{\frac{1}{6}}}$$

$$= \frac{.15}{.41} \qquad Z_{.05} = 1.96$$

$$= .37$$

Since calculated $Z < Z_{.05}$, H_0 cannot be rejected

9. $t_{31;.05} = 2.04$ Since the computed t-statistics for Welfare and Immigration are both greater than 2.04, they are "significant." Temperature is not significant at the 5% level

$$\frac{R^2}{1-R^2} \cdot \frac{N-p}{p-1} = \frac{.9394}{.0606} \cdot \frac{31}{2}$$

$$= 15.5 \cdot 15.5 \qquad F_{2,31;.05} = 3.32$$

$$= 240.25$$

There is clearly a linear relationship

Emigration = Constant + β_1 Welfare + β_2 Immigration

10. H_0: $\mu = 25{,}000$

H_1: $\mu \neq 25{,}000$

$$Z = \frac{23{,}200 - 25{,}000}{\frac{4000}{\sqrt{400}}}$$

$$= \frac{-1800}{200} \qquad Z_{.05} = 1.65$$

$$= -9$$

$$= -9 < -1.65 \quad \therefore \text{ reject } H_0$$

Unit 6
Quiz

Name: ____________________

Write all your answers on these pages. Point totals are given in parentheses prior to each question. You have forty (40) minutes for this quiz.

(45) 1. You have estimated a linear model relating the response single family housing starts in Pittsburgh (Y) to carriers Median Price of a new unit in thousands (X_1), % Unemployment (X_2), and Population in thousands (X_3). You collect data for 1960-1972. The model is

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e.$$

Least squares estimates of the parameters are

$$\underset{\sim}{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 57363 \\ -182.5 \\ 222 \\ -19.1 \end{pmatrix}$$

and diagonal terms of the Variance-Covariance Matrix are

$$s^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1} = \begin{pmatrix} 3.38 \times 10^8 & & & \\ -- & 4.48 \times 10^3 & & \\ -- & -- & 2.70 \times 10^3 & \\ -- & -- & -- & 5.04 \times 10^1 \end{pmatrix}$$

and $R^2 = .902$.

(a) Last term we stated that in testing the importance of a coefficient in a linear model, you should consider the carrier "important" if the t-statistic was greater than 2 or 3 in absolute value. Why?

(b) Does the carrier X_3 differ from zero?

(c) What hypothesis do you test to determine whether or not the response is linearly related to the set of carriers as a whole? Under the null hypothesis, the test statistic is distributed as a specific random variable. Which distribution is it, and why is this the correct one?

(25) 2. Your supervisor states that 5% of the census tracts in Pittsburgh have median family size greater than 6 individuals/family. In disbelief, you gather data on the 86 census tracts and find that median family size per tract is remarkably well behaved, with $\mu = 4.5$ and $\sigma^2 = .20$. Is your supervisor correct? Why or why not?

(30) 3. The computer center at Robber Baron University claims a 95% availability for their HAL-250 computer. You are somewhat skeptical of this statement, so you gather data for the 30 days that you used the system for your latest paper. You calculate the average availability to be 85% with associated standard deviation $\frac{s}{\sqrt{n}}$ of 5%.

(a) Construct a 95% confidence interval for the true percentage.

(b) Based on this interval, state and test a hypothesis (α = .05) to determine the truth of the computer center's assertion.

(c) Are the distributional assumptions that you made to test the hypothesis in (b) appropriate? Why or why not?

TABLE III

The Normal Distribution

$$\Pr(X \le x) = N(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw$$

$$[N(-x) = 1 - N(x)]$$

x	$N(x)$	x	$N(x)$	x	$N(x)$
0.00	0.500	1.10	0.864	2.05	0.980
0.05	0.520	1.15	0.875	2.10	0.982
0.10	0.540	1.20	0.885	2.15	0.984
0.15	0.560	1.25	0.894	2.20	0.986
0.20	0.579	1.282	0.900	2.25	0.988
0.25	0.599	1.30	0.903	2.30	0.989
0.30	0.618	1.35	0.911	2.326	0.990
0.35	0.637	1.40	0.919	2.35	0.991
0.40	0.655	1.45	0.926	2.40	0.992
0.45	0.674	1.50	0.933	2.45	0.993
0.50	0.691	1.55	0.939	2.50	0.994
0.55	0.709	1.60	0.945	2.55	0.995
0.60	0.726	1.645	0.950	2.576	0.995
0.65	0.742	1.65	0.951	2.60	0.995
0.70	0.758	1.70	0.955	2.65	0.996
0.75	0.773	1.75	0.960	2.70	0.997
0.80	0.788	1.80	0.964	2.75	0.997
0.85	0.802	1.85	0.968	2.80	0.997
0.90	0.816	1.90	0.971	2.85	0.998
0.95	0.829	1.95	0.974	2.90	0.998
1.00	0.841	1.960	0.975	2.95	0.998
1.05	0.853	2.00	0.977	3.00	0.999

from: Hogg, R. V. and A. T. Craig, Introduction to Mathematical Statistics, Third Edition, New York: Macmillan, 1970.

Quiz Unit 6
Solutions

1. a) We have been implicitly testing the hypotheses that our β coefficients are zero. We know that $(\beta-\hat{\beta})/s_{\hat{\beta}}$ is distributed as a t random variable with N-p degrees of freedom (N observations, p variables). For large N-p and $\beta=0$, 95% of the $\hat{\beta}$'s will fall in the interval (-2,2). A t-statistic outside of that interval allows to reject the hypothesis that $\beta=0$. For smaller degrees of freedom, we use the larger interval (-3,3).

   b) H_0: $\beta_3 = 0$

   H_1: $\beta_3 \neq 0$

   $$t = \frac{b_3-0}{s_{b_3}} = \frac{-19.1-0}{\sqrt{50.4}} \approx -2.7$$

   Since we have only 13 observations--the years 1960-1972--we use the confidence interval (-3,3). The t-statistic is within this range, so we <u>cannot</u> say that X_3 is significantly different from zero.

   c) H_0: $R^2 = 0$

   H_1: $R^2 > 0$

   The test statistic $\dfrac{R^2/(p-1)}{(1-R^2)/(N-p)} \sim F_{p-1,\ N-p}$

   Again, there are N observations and p variables.

   R^2 and $(1-R^2)$ are ratios of sums of squares. Each has the same denominator. Hence their ratio is a ratio of two χ^2 random variables. The ratio of two χ^2's, divided by their degrees of freedom, is distributed F.

2. Median family size ~ Gau (4.5, .20)

   $$z = \frac{\text{median family size} - 4.5}{\sqrt{.20}} \sim \text{Gau }(0,1)$$

   Pr(Median family size > 6)

   $$= \Pr\left(\frac{\text{Median family size} - 4.5}{\sqrt{.20}} > \frac{6 - 4.5}{\sqrt{.20}}\right)$$

$= Pr(Z > 3.33)$

$< .001$ (from the normal probability table)

Our supervisor, w..o claims that P:(Median family size > 6)

= .05, is wrong.

Alternatively, if we note that 6 is more than 3 standard deviations from the mean (median = mean in a well-behaved batch), we know that our supervisor has overestimated the frequency of median family sizes greater than 6.

3. a) With a large number of observations, a 95% confidence interval is described by

$$p \pm Z_{.025}\left(\frac{s}{\sqrt{n}}\right)$$

$$.85 \pm 1.96\ (.05)$$

(.752, .948) is the 95% confidence interval.

b) H_0: $P = .95$

H_1: $P \neq .95$

Since the confidence interval that we con ..cted in part (a) is our acceptance region when $\alpha = .05$, we reject H_0. We disagree with the computer center.

c) We have relied on the assumption that our data are approximately normal. However, the distribution is very skewed with a p of .85 or .95. (p is bounded by 1.) In light of the skewness, 30 is not a large enough sample size to justify our assumption.

911

Unit 7
Reading Assignments

Lecture	Reading
7-0	Warwick and Lininger, chapters 1, 2, 3 Davis, "Are Surveys Any Good..."
7-1	Warwick and Lininger, chapters 6, 7, 8; Sudman, "Sample Surveys".
7-2	Warwick and Lininger, chapters 4, 5;

Optional:

Coleman, et.al. "Relation of School Factors...";
Duncan, "Measuring Social Change...";
Featherman and Hauser, "Design for a Replicate Study...";
Stokes, "Some Dynamic Elements..."
Winsborough, "Age, Period and Cohort..."

References:

Warwick, D. P. and C. A. Lininger, The Sample Survey: Theory and Practice, McGraw-Hill, 1975.

Davis, J. A., "Are Surveys Any Good, and if so, for What?" in Perspectives on Attitudes Assessment: Surveys and Their Alternatives: Proceedings of a Conference, Smithsonian Institution Technical Report #2, August 1975, National Technical Information Service # AD-A014321; pp. 41-47.

Sudman, S. "Sample Surveys" in Annual Review of Sociology, Volume 2, Edited by A. Inkeles, J. Coleman, and N. Smelser. Palo Alto: Annual Reviews Inc., 1976. pp. 107-120.

Optional Readings:

The following three articles appear in Social Indicator Models, edited by K. C. Land and S. Spilerman, Russell Sage Foundation, New York, 1975.

Duncan, O. D., "Measuring Social Change via Replication of Surveys", pp. 105-128.

Featherman, D. L. and R. M. Hauser, "Design of a Replicate Study of Social Mobility in the United States", pp. 219-252.

Winsborough, H. H., "Age, Period, Cohort, and Education Effects on Earnings by Race--An Experiment with a Sequence of Cross-Sectional Surveys," pp. 201-218.

The following two articles appear in The Quantitative Analysis of Social Problems, edited by Edward R. Tufte, Addison-Wesley Publishing Company, Reading, Massachusetts, 1970.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York, "Relation of School Factors to Achievement" and Integration and Achievement", from Equality of Educational Opportunity.

Stokes, Donald E. "Some Dynamic Elements of Contests for the Presidency".

Lecture 7-0. Introduction to Unit 7

Introduction to Unit 7--Sample Surveys

<u>Lecture Content</u>:

1. Definition of a Sample Sur
2. Examples of Sample Surveys

<u>Main Topics</u>:

1. What is a Survey
2. Components of a Survey
3. Motivation for Conducting a Sample Survey

(There are no transparencies for this lecture.)

Reference: Warwick and Lininger, Chapters 1-3

Topic 1. What is a Survey?

I. A Data collection procedure (to e distinguished from a data analysis procedure).

II. A detailed investigation, mapping, or inspection to enumerate or observe characteristics of a population.

III. Examples

1. Survey of wildlife in a region
2. Survey of objects on a desk
3. Survey of rocks in a soil sample
4. Survey of opinions held by residents of a city

IV. Major forms

1. Census--complete survey: every object in the relevant population is involved.
2. Sample survey--partial survey: members of the population are selected and the entire population's characteristics are inferred from the sample.

915

Topic 2. Components of a Survey

I. Instrument

1. Observation rule
2. Interview topics
3. Questionnaire
4. Continuous record

II. Fielding procedure

1. Interview structure
    a. Open-ended
    b. Structured
    c. Item response
2. Data collector
    a. Interviewer
        i. Face-to-Face
        ii. Telephone
    b. Self-administered
        i. Questionnaire
        ii. Diary
    c. Unobtrusive observer
        i. Participant
        ii. Mechanical recorder

III. Data recording and reduction

1. Items on schedule
2. Coders
3. Direct to computer
4. Machine readable forms

IV. Analysis plan

V. Sampling procedure (for sample surveys)

1. Ad hoc
2. Arbitrary
3. Probability
4. Oversampling

VI. Overall design

1. Cross-section
2. Panel, successive samples
3. Snowball
4. Multiple questionnaires (for different categories of respondents)
5. Multiple linked items
6. Timing

VII. Staff requirements

1. Administrative
2. Clerical
3. Field
4. Scientific
   a. Questionnaire design
   b. Sampling procedure

VIII. Sequence of activities

Issue defined → population defined → instrument designed → instrument tested → sample designed → sample selected → instrument fielded → data returned → data coded → data cleaned and reduced → analysis commences

(Potential biases and errors occur at each stage.)

Topic 3. Motivation for conducting a sample survey

I. Nature of data--must interact with people

1. Opinions, attitudes, experiences
2. Past unrecorded actions
3. Enumeration
4. Behavioral intentions
5. Legislative requirements (U.S. census)

II. Why sample?

1. Cost
2. Efficiency--not all observations needed
3. Necessity--not all population members available

III. What purposes can a survey serve?

1. Describe a population
2. Test hypotheses and theories about behavior
3. Deduce goals, interests or desires
4. Evaluate programs
5. Forecast outcomes

IV. Problems with surveys

1. Cost
2. Interaction required--obtrusive
3. Time consuming--for respondent and collection
4. Error prone

Lecture 7-1. Survey Design

Survey Design: Designing instruments and fielding procedures for administering sample surveys

Lecture Content:

1. Concerns of the survey designer
2. Examples of surveys

Main Topics:

1. Respondents
2. Questionnaire
3. Interview
4. Examples

(There are no transparencies for this lecture.)

Reference: Warwick and Lininger, Chapters 6-8

Topic 1. Respondents

I. Who is to be surveyed--who is the survey about?

1. An individual

2. Heads of households

3. Income earners

4. Parents

5. Participants in a particular program

6. etc.

II. Where are the respondents located?

1. Geographically

2. Socioeconomically

3. Behaviorally

III. Where will they be interviewed?

1. Residence

2. At program site

3. On the street

4. In store

IV. What impact does nature of respondents have on survey?

1. Language

2. Types of questions that can be asked

3. Timing

4. Access

5. Security

6. Response rate--cooperativeness

Topic 2. Questionnaire

I. What controls are required?

1. Age
2. Sex
3. Race
4. Ethnicity
5. Family type and size
6. Marital status
7. Income
8. Occupation
9. Education
10. Others...

II. What indicators can be used?

1. Duncan scale of occupational prestige
2. U.S. Bureau of Labor Statistics or Census Bureau definitions
3. Review other used measures (may be able to contrast results)

III. Interview situation

1. Problems
   a. Phone--selective, short, unknown respondent
   b. Self administered--who really did it?
   c. Face-to-face--interviewer training
2. Advantages
   a. Mail--cheap
   b. Phone--cheap, fast

IV. Open versus closed response

1. Closed--prompts meaning, limited, category coverage
2. Open--difficult to quantify irrelevant responses--lack of verbal ability but gets spontaneous and unexpected information
3. Combination of open and closed

V. Question writing--objectives

1. Simplicity--for interviewer and interviewee includes structure, vocabulary, and responses
2. Spec[illegible]y--single issue focus
3. Avoi[illegible]crations--biases and prompts
4. Permit catch-all category
5. Construct appropriate context
6. Depersonalize answers
7. Make relevant to respondent
8. Voice in respondent's style
9. Balance questions positively and negatively
10. Avoid overly consistent response categories and sequence
11. Avoid extreme statements
12. Build in consistency checks
13. Construct effective flow and branching
14. Construct simple layout
15. Keep size to minimum
16. Provide handouts for complicated answers

Topic 3. Interview

I. Use trained interviewers (Reference: Warwick and Lininger, Chapter 8)

II. Presentation of interviewer should be natural and unobtrusive

III. Perform random checks on interviews (by calling, etc.)

IV. Examples

1. Choose two surveys, one of poor quality (such as a magazine self-report questionnaire) and one of professional quality (such as one administered by the National Opinion Research Center or the instrument appearing on pages 172-181 of Warwick and Lininger.)

2. Make certain that students have copies of these survey instruments

3. Have students administer the instruments (or parts) to one another

4. Discuss positive and negative features of the questionnaires

Lecture 7-2.. Sample Design

Sample Design for Surveys--The use of statistical procedures for selecting respondents and estimating errors.

Lecture Content:

1. Types of sampling procedures
2. Statistics for simple random sampling

Main Topics:

1. Review motivation for sampling
2. Sampling procedures
3. Probability sampling

(There are no transparencies in this lecture.)

Reference: Warwick and Lininger, Chapters 4-5

Topic 1. Review motivation for sampling

I. Why sample?

1. Can all population members be interrogated?

2. Need all population members be interrogated?

3. Cost of complete census may be too high

4. Adequate level of precision may be reached with sample

5. Sample may be better than census--ask more questions of fewer people

II. What is role of sampling procedure?

1. Select individuals to interrogate

2. Provide [illegible]anism for estimating error

Topic 2. Sampling procedures

I. Probability--individuals selected by chance mechanism with certain known probabilities of inclusion

1. Simple random sampling--equally likely and independent inclusion probabilities

2. Many variations (non-equal probability)

a. Stratified

b. Clustered

c. Multistage

II. Non-probability

1. Haphazard

2. Judgmental--interviewer determined

3. Quota--categories outside

4. Experts--(paid by interviewer)

5. Purposive

(Note: Non-probability methods do not permit estimating errors in inferring features of the population from characteristics of the sample--thus it cannot be known what size sample is required to obtain some specified level of precision.)

Topic 3. Probability Sampling

I. Structure

1. List: Inventory of population units

2. Sampling units: Actual sampling basis

3. Frame: Operational procedure to account for population

II. Types of errors

1. Instrument measurement error

2. Interviewer bias

3. Sampling error--this we can quantify in terms of a confidence interval around a mean response:

   a. Once we sample, response is a random variable

   b. When variance and distribution of random variable are known, confidence interval for the mean can be obtained

   c. Stating a confidence level, we can obtain an estimate of the needed sample size

III. Simple random samples--equally likely and independent

1. All units chosen individually

2. All units have same chance of being chosen

3. Selection of one unit does not prejudice selection of any other

4. Various mechanisms from list

   a. Random number table

   b. Computer pseudo-random numbers

   c. Mechanical devices

IV. Introduction to statistics for simple random sampling (SRS)

Note: For a more extensive treatment of sampling the instructor should consult Kish, L., The Sample Survey, New York: Wiley, 1965.

(Notation: upper case letters refer to population lower case letters refer to sample)

When a survey is to sample attributes (numerical) in a population we are interested in several issues (use binomial if attribute is dichotomous or Normal approximation to the binomial):

1. We will examine average opinion in the sample, $\bar{x}$.

2. From the sample average we will infer the population average, $\bar{X}$.

3. Given a level of precision for this inference, we will specify the sample size, n.

4. To select n responses randomly we will need an inventory or list of the population, N.

Since we are sampling the population, the sample mean obtained from one sample is only one of many possible means of similar samples, i.e., it is a random variable with expected value $E(\bar{x}) = \bar{X}$ and, in SRS, is distributed $N(\bar{X},\sigma^2/n)$. That is, it is Normally distributed and an unbiased estimate of the population mean.

Note that $\bar{x}$ is Normally distributed even if the distribution of the attribute being sampled is not Normal in the population (except in cases where $n < 30$).

V. Confidence intervals for the population mean

1. Variance of population and sample.

$$\sigma^2 = \frac{\Sigma(X_i-\bar{X})^2}{N} \qquad \text{population variance}$$

$$s^2 = \frac{\Sigma(x_i-\bar{x})^2}{n} \qquad \text{sample variance}$$

Using sample variance as estimate of population variance we use $\frac{N-n}{N} s^2$ and note that as N becomes large this correction become very small.

2. Variance of sample means from similarly sized and drawn samples (referred to as the square of the standard deviation of the sample means or the standard error).

3. Confidence intervals

Theoretical Value: $\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N}\,\frac{\sigma^2}{n}}$

Estimate using sample variance:

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N}\,\frac{s^2}{n}}$$

95% confidence interval for $\bar{X}$ is $\approx$

$\bar{x} \pm 2.0 s_{\bar{x}}$

99% confidence interval for $\bar{X}$ is $\approx$

$\bar{x} + 2.6 s_{\bar{x}}$, etc.

4. Example

Perform an experiment with the class by asking all to record their age to the tenth of a year. Sample the group randomly and draw a sample of ten. Compute statistics for constructing a confidence interval for average class member's age from actual average age.

VI. Sample size

Examine the equation for the estimate of the standard error,

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N}\,\frac{s^2}{n}} \quad = \sqrt{\frac{N-n}{N}}\sqrt{\frac{1}{n}}\ (s)$$

and discuss the relative impact on error reduction that occurs by increasing proportionate sample size, $\frac{N-n}{n}$, and absolute sample size, $\frac{1}{n}$.

Note the relative efficiency of increasing absolute size.

Since absolute size increases reduce $s_{\bar{x}}$ by $1/\sqrt{n}$ note that there are decreasing percentage improvements as the ratio of sample size to population increases. The typical national sample survey uses n between 1500 and 2500. 959

VII. Modifications of SRS

In advanced classes discuss

1. Stratification--divide population into strata
2. Clustering--elements chosen as groups
3. Systematic selection--use selection interval
4. Unequal probability--weight selection probabilities
5. Multistage sampling--selection involving two or more successive stages

Reference: Warwick and Lininger, pp. 95-110 and Chapter 5.

Homework
Unit 7

1. Write a short essay discussing the merits and disadvantages of using sample surveys to collect data for policy analysis.

2. Discuss the merits and disadvantages of using archival survey data for policy analysis.

3. Design an interview schedule to be administered in 30 minutes to obtain information relevant to one of the following urban policy issues:

    1. Public transportation
    2. Hospital care of the elderly
    3. Prenatal care of welfare mothers
    4. Satisfaction with garbage and sanitation services
    5. Air pollution

   Be certain to include relevant control variables and discuss your planned analysis in terms of policy amenable independent variables.

4. Using the questionnaire designed in the prior reading assignment design a sampling procedure in which a 1% SRS sample would be drawn from an urban population. Discuss the stages through which the sample is actually drawn. Choose five questions and estimate their standard errors as: .01, .05, .1, .2, .7. Compute confidence intervals for population attributes sampled by each question using sampling fractions of 90%, 75%, 50%, 10%, 1%, .01%. Assume the city has a population of 500,000. Assuming a cost of $50/ completed interview, discuss the comparative merits of each sampling design.

Quiz
Unit 7

Name: ____

Please read this quiz thoroughly before writing answers to any of the questions. Make your answers brief and to the point. Excessive wordiness and rambling responses will detract from your total score. You have thirty (30) minutes to answer this quiz.

Examine the following questionnaire that appeared in the national magazine Ms. and was meant to be filled in and returned by Ms. readers. (Do not answer these questions. Quiz questions begin on the third page.)

First National Television Test on Sexual Attitudes (from Ms. Magazine)

Fact Questions

1. Men are more aggressive than women.

   True ☐ False ☐

2. Single men are psychologically better off than married men.

   True ☐ False ☐

3. Most women are supported by men and therefore work for luxuries not necessities.

   True ☐ False ☐

4. The average full-time male worker earns:

   ☐ 70% More ☐ 38% More ☐ 18% More

   than his female counterpart.

5. Of all girls born in 1977, what percentage will work outside the home during their lifetimes?

   ☐ 33% ☐ 55% ☐ 90%

Opinion Questions

1. If you could send only one child to college, would you send:

   your son ☐ your daughter ☐ your oldest child ☐

2. The more assertive and independent a woman is, the less sexually attractive she is to men.

   agree ☐ disagree ☐

3. A woman who decides not to have children is:

   ☐ Missing one of life's greatest satisfactions

   ☐ Unfeminine

   ☐ Fulfilled in other equally valuable ways

   ☐ Probably physically unable to have them

4. What gives <u>you</u> the most satisfaction?

   ☐ Family ☐ Running a home

   ☐ Love Life ☐ Career

5. Who gets the better deal in this society?

   Men ☐ Women ☐

---- END OF QUESTIONS THAT APPEARED IN MS. MAGAZINE ----

983

QMPM quiz questions:

1. Comment on the split of these 10 questions by Ms. magazine into fact and opinion. Are any of the questions "double-barreled" or applicable to only a fraction of the respondents? Are the questions "loaded"? Are the answers allowed for each question sufficient and accurate? What other comments can you make about the nature of the questions?

2. Given a large response rate, what conclusions can be drawn by an investigator from individual responses to these 10 questions? What can one say about national "sexual attitudes" from this survey. What qualifications must be made when generalizing from this survey and why?

3. Assume you have $150,000 to spend on a national survey of sexual attitudes and that the questionnaire has been designed and field tested. Describe a workable, reliable and efficient sampling and implementation strategy. Be sure to discuss whether clustered or stratified sampling should be employed and the nature of the field work.

Quiz Solutions
Unit 7

1. There were many aspects of the survey which could be criticized, including the following:

   (a) The distinction between fact and opinion is arbitrary.

   (b) There was no response alternative for "don't know" or "no opinion."

   (c) Many of the words are loaded, such as "aggressive" and "assertive".

   (d) Many of the words are subject to interpretation, such as "psychologically better off."

   (e) "Most women are supported by men and therefore work for luxuries not necessities" is a double-barreled question. (It asks two different questions.)

   (f) There is no indication on the questionnaire of the respondent's age or sex or marital status; yet these will probably greatly influence the responses.

   (g) In many cases, the answers are not exhaustive. For instance, there are many possible reasons for a women to decide not to have children other than the alternatives listed.

2. Even if there is a large response rate, the respondents will be that group of <u>Ms.</u> readers who would answer the questionnaire. The characteristics of that group are certainly different from those of the national population. The survey asks for no demographic or biographical data, but such factors as age and sex and location affect "sexual attitudes." Further, we have just finished discussing flaws in the survey itself. Therefore, even with a large response rate, we do not want to say anything about national sexual attitudes on the basis of this survey.

3. There are many correct answers to this question; errors are apt to be of omission, not commission. In designing your survey strategy, did you consider...

   (a) who the population is that you wish to generalize about? (the whole country? adults only? sexually active adults only?)

   (b) how large a sample to have, in either absolute numbers or percentages?

   (c) what sampling strategy to use (cluster? stratified? simple random sample?) and the relative advantages of the strategy you chose? Remember the large scale of a national survey.

   (d) how to administer the survey (in person? by phone? by mail?) and the relative advantages of your choice?

   (e) the potential embarrassment to respondents, particularly in a face-to-face interview given by someone of the opposite sex?

   (f) the cost of your strategy?

QUANTITATIVE METHODS FOR PUBLIC MANAGEMENT

MODULE IV, REVISED

Developed by

SCHOOL OF URBAN AND PUBLIC AFFAIRS
CARNEGIE-MELLON UNIVERSITY

SAMUEL LEINHARDT, PRINCIPAL INVESTIGATOR
and
STANLEY S. WASSERMAN

Under Contract to

THE URBAN MANAGEMENT CURRICULUM DEVELOPMENT PROGRAM
THE NATIONAL TRAINING AND DEVELOPMENT SERIVCE
5028 Wisconsin Avenue, N.W.
Washington, D.C. 20016

Funded by

The Office of the Assistant Secretary
for Policy Development and Research
U.S. Department of Housing and Urban Development

Package XVI

Acknowledgements

Assistance in the preparation of this package was provided by Blaine Aikin, Larry Albert, Joseph Chmill, Steve Clark, Marjorie Farinelli, Janice Greene, Gretchen Hemmingsen, Paul W. Holland, J. Michael Hopkins, Gaea Leinhardt, Richard Sandusky, Christine Visminas, Diane Warriner, and Tammar Zeheb.

TABLE OF CONTENTS

Material intended solely for the instructor is denoted by a (I). Material that should also be distributed to the students is denoted by a (S).

Introduction to Module IV

Overview

Module IV of the Quantitative Methods for Public Management package contains two units, numbers 8 and 9. Unit 8, Two-way classifications for continuous data, introduces the student to the construction of models for summarizing continuous data arrayed in a two-way table, a table-type data structure quite common in public policy studies. Three variables are involved, two factors and a response. The general strategy is to fit a simple additive model to the table, compute fitted values and residuals and examine the quality of the model. The fitting procedure employed involves iterated decomposition of the table using repeated removal of medians (i.e., median polish) or means (i.e., mean polish). A procedure is introduced for determining whether the data need to be transformed to improve the appropriateness of an additive model. Techniques are also discussed for handling ordinal levels in the factors and for constructing a model with an interaction term.

Unit 9, Discrete Multivariate Anaylsis, introduces the student to the analysis of contingency tables, another table-type data structure common in policy studies. The data in this case are discrete frequencies, counts of the simultaneous occurrence of two or more conditions. The question posed by analysis is whether or not the table provides evidence of independence in the variables. The strategy is to introduce students to contingency tables via traditional test for goodness of fit in one dimensional tables and then to develop log-linear models in the analysis of higher dimensional tables.

Specific Objectives

Unit 8

Upon successful completion of Unit 8 a student will be able to recognize continuous data that can be arrayed in a two-way layout and will be able to analyze the data. Analysis could include the construction of elementary additive models using median or mean polish, computation of comparison values and construction of diagnostic plots, identification of data requiring transformations, selection of appropriate transformations, construction of displays of coded residuals, evaluation of the fit of the model, plots of effects for ordinal factors, development of extended summaries for ordinal factors, and development of extended models incorporating an interaction term.

Unit 9

Upon successful completion of Unit 9 a student will be able to identify data which can be analyzed as a one, two, or more dimensional contingency table. Analysis will include determination of appropriate probability models, construction of cross-product ratios, computation of Pearson's χ^2 test for goodness of fit in the case of a one dimensional table, construction of log-linear models in higher dimensional tables, tests for independence of variable and for interactions. Students will have obtained experience in constructing log-linear models for frequency data arising in commonly reported tables such as opinion surveys and censuses.

Prerequisite Inventory
Units 8 and 9

In this module we analyze data which come in two forms. The data in Unit 8 are two-way tables, which relate one Y and two X variables. In Unit 9 we look at contingency tables, which list the number of observations in the different categories of one or more variables.

Comprehension of Module I is assumed. Stem-and-leaf displays and medians are topics covered in Module I that are also used in this module. The topics in this inventory are:

1. Review of Numbers: Amounts and Counts
2. Review of Resistant Lines
3. Review of Hypothesis Testing and χ^2
4. Data Structures

If you are uncertain about any of these topics after reading this inventory, please consult a member of the teaching staff. Mastery of this material is essential before proceeding to Module IV.

Section 1. Review of Types of Numbers

In Unit 1 four types of numbers were discussed: amounts, counts, bounded numbers, and differences. In this module it will be necessary to distinguish counts from amounts. Two-way tables contain amounts. Contingency tables contain counts.

Amounts are levels of a variable. Amounts may either be either discrete or continuous, but for our purposes we usually think of them as continuous. When we discuss thousands of dollars of income, income can take on so many values that the variable is essentially continuous even though the smallest unit it can be expressed in is .01 dollars.

As another example, distance is a continuous variable and 56.34 miles is an amount.

A count is the number of observations in a category. Counts take on only non-negative integer values. The number of people in the U.S. with income greater than $20,000 is an example of a count.

Compare these 2x2 tables:

Average income (in $) of Transylvania residents, by race and sex

	Male	Female
Black	8,400	8,000
White	9,000	8,200

Number of Transylvania residents, by race and sex

	Male	Female
Black	87,508	88,981
White	195,067	198,216

The table on the left introduces a new variable (average income) but tells nothing about the number of people whose incomes contributed to the averages on each of the four cells. The table on the right tells the number of observations in each of the four categories but introduces no new variable. The table on the left is called a "Two-Way" table of amounts; the one on the right, a "Contingency Table" of counts.

Section 2. Review of Resistant Lines

A clear understanding of resistant lines is important for two reasons: many of the concepts used in describing two-way tables are analogous to techniques used in fitting resistant lines, and there are relationships in two-way table analysis that are best described by fitting resistant lines.

A resistant line is a fit which describes the relationship of paired (X,Y) data. If X and Y are linearly related (in raw or transformed units), a resistant line summarizes that relationship with a single equation. Unlike least squares regression lines, resistant lines are not much affected by a couple of points which deviate from the linear trend.

To fit a resistant line, break the ordered X's into thirds, carrying along with each X its paired Y value. Calculate a conditional typical value of X and of Y for each of the three minibatches of paired values. The conditional typicals will be the (median X, median Y) of each third, although these pairs may not have been paired among the original N ordered pairs.

Before fitting a line, check to see if the data need to be transformed. To do this, proceed to list the conditional typicals:

(X_L, Y_L) (X_M, Y_M) (X_H, Y_H)

Calculate the two slopes:

$$m_1 = \frac{Y_H - Y_M}{X_H - X_M} \quad \text{and} \quad m_2 = \frac{Y_M - Y_L}{X_M - X_L}.$$

If the data are linear, then m_1/m_2 will equal 1. If the ratio is <1, transform X down the ladder of powers; if > 1, transform up the ladder of powers. You need transform only the three summary points to see if the transformation is successful. After you decide on the appropriate transformation, then transform all of the data.

Once the data are linear, the next step is to remove the tilt (or slope) from the line. The slope is determined by

$$m = \frac{Y_H - Y_L}{X_H - X_L}$$

where X may now represent transformed data.

Remove the slope by rewriting each Y_i as $Y_i - mX_i$. The new conditional typicals are

$$(X_L, Y_L - mX_L) \quad (X_M, Y_M - mX_M) \quad (X_H, Y_H - mX_H).$$

The level (or intercept) of the line is the median of $Y_L - mX_L$, $Y_M - mX_m$, $Y_H - mX_H$. Subtract the level from each Y value. Now we're left with residual = Y - mX - level

We may choose to polish the line by treating the residuals (i.e., Y - mX - level) as a new batch of Y's, repeating the fitting procedure described above, and adding the polished fit to the original fit. We calculate a new batch of residuals from the polished fit and may polish again if we'd like to. The decision to polish is usually based on the appearance of the residuals (in a stem-and-leaf or as plotted against X).

Section 3. Review of Hypothesis Testing and X^2

Many aspects of probability and inference are utilized in Module IV. In particular, you should feel comfortable with hypothesis testing, levels of confidence (α), and χ^2 distributions.

In hypothesis testing, we establish a null hypothesis, called H_0, which we express in quantitative terms. The null hypothesis is generally a supposition about a population parameter. We then do whatever analysis is appropriate to the hypothesis, based on a sample from the population in question and on the assumption that H_0 is true. If our analysis leads to conclusions that are "unlikely", we reject H_0, i.e., conclude that it cannot be true. Otherwise, we do not reject H_0, i.e., conclude that based on our analysis H_0 could be true.

The decision as to whether or not a result is "likely" is not a subjective decision but is based on probability. We cannot be correct all the time, but we can decide how much "being wrong" we are willing to tolerate. The proportion of times we expect to be wrong is α, and $1-\alpha$ (times 100%) is our level of confidence. Commonly used levels of confidence are 90%, 95%, and 99%.

We are able to quantify our confidence in this precise manner because of our knowledge about underlying probability distributions. For example, in least squares regression we tested the hypothesis that a true β-coefficient was zero. We made use of our knowledge of t-distributions to determine whether the sample coefficient was likely to be non-zero when the true $\beta = 0$.

In Unit 9 we will use our knowledge of the χ^2 distribution in hypothesis tests. A χ^2 random variable is defined as the sum of squared normal random variables. It is characterized by one parameter, its degrees of freedom. In theory, degrees of freedom are determined by the number of normal random variables which are squared to form the χ^2; in practice, we will figure out the degrees of freedom from the number of variables and number of observations in our data. Just as we used t-tables and Z-tables, there are χ^2 tables which tell the probability with which a χ^2 random variable takes on values within specified regions.

Section 4. Data Structures

Most of the data that we have looked at so far have been either one-dimensional or two-dimensional. One-dimensional data are typically single batches of data, written as a list of numbers, a vector, or an nx1 or 1xn table. Examples:

Census Tract	Number of Physicians	Quiz Grades
1	80	99
2	43	97
3	5	86
⋮		72

Paired (X,Y) data for multiple regression analysis may be thought of as having two dimensions, one dimension for each variable. Along one dimension are the variables (e.g., income, age) and along the other is whatever characterizes the observations (e.g., census tract, city). For example, the hospital insurance data:

	old premium	new premium
Children's	866	646
Beth Israel	833	635
McLean	255	218
Mt. Auburn	162	148
Deaconess	435	348

illustrate a set of 5 paired observations arranged in a 5x2 table.

Now consider the following tables.

		Male		Female	
		Black	White	Black	White
Age	<40				
	≥40				

Number of college deans
by age, race, and sex

978

Each of the eight cell entries represents an observation across the three dimensions of age, race, and sex. If we could present such tables in three dimensions, we would have done so, placing one of the tables on top of the other. Since we have to present the data on two-dimensional paper, we placed the tables next to each other. The decision to split into separate tables on the basis of sex was arbitrary; we could as easily have written

Age

	< 40	
	Male	Female
Black		
White		

	≥ 40	
	Male	Female
Black		
White		

or any one of four other combinations.

With an understanding of three-dimensional data, we can easily extend our knowledge to larger dimensions. Suppose we want to add region to the college dean data. Below is one way to represent the four-dimensional data.

		Male			
		Northeast	South	Midwest	West
< 40	Black White				
≥ 40	Black White				

		Female			
< 40	Black White				
≥ 40	Black White				

Homework
Prerequisite Inventory, Units 8 and 9

1. Identify the dimensions of the following tables as 1, 2, 3, or more dimensions and state whether the cells of the table contain counts (number of observations) or amounts (variable).

   a. median age of college students, by class and college

   b. enrollment in each of the elementary schools in the city of Pittsburgh

   c. number of blue-collar and white-collar workers in major U.S. cities

   d. number of demolitions in 1976 by building type and census tract

   e. number of patients in Philadelphia hospitals, by hospital, illness, and age

Answer questions 2-10 as briefly as possible.

2. When you examine a batch of residuals, what are you looking for?

3. How might you want to examine residuals from a fitted line?

4. What values can a count take on?

5. What type of data do you fit resistant lines to?

6. Identify (median X, median Y) in the following (X,Y) batch:

   (3, 13)
   (5, 11)
   (6, 18)
   (6, 10)

7. How many steps of polish are necessary when fitting a resistant line?

8. When do we conclude that a null hypothesis is true?

9. In hypothesis testing, why are we willing to be wrong some of the times that we reject the null hypothesis?

Homework
Prerequisite Inventory, Units 8 and 9
Solutions

1. a) two-dimensional table of amounts

   b) one-dimensional table of counts

   c) two-dimensional table of counts

   d) two-dimensional table of amounts

   e) three-dimensional table of counts

2. Gaussian shape, centered and clustered at zero, very few outliers

3. Plot the residuals against X or Y or $\hat{Y}$

4. 0, 1, 2, 3, ...

5. Paired (X,Y) data which exhibit a linear relationship in either raw or transformed units

6. (5.5, 12)

7. It depends on the shape and size of the residuals after the original fit (and each step of polish), and on whether you are fitting the data by hand or computer.

8. We never conclude that a null hypothesis _is_ true; we conclude that it _could be_ true if for a specified level of confidence the truth of the null hypothesis could lead to the observed sample statistic(s).

9. If we weren't willing to be wrong some of the time, we would never reject the null hypothesis. We can never know a true population parameter; but we can decide what percentage of the time we are willing to be wrong.

Unit 8
Reading Assignments

Lecture	Reading
8-0	Tukey, Chapter 10, pages 331-348
8-1	Tukey, Chapter 10, pages 348-363
8-2	Tukey, Chapter 11
8-3	Singer, "Exploratory Strategies and Graphical Displays" Journal of Interdisciplinary History, volume 7, pages 57-70

In addition, please read the following article:

Fairley & Mosteller, pp. 23-50

Texts:

Fairley, W.B. and F. Mosteller, Statistics and Public Policy, Reading, Massachusetts, Addison-Wesley, 1977.

Tukey, J.W., Exploratory Data Analysis, Reading, Massachusetts, Addison-Wesley, 1977.

Lecture 8-0. Introducation to Unit 8

Introduction to Unit 8--Two-Way Tables

Lecture Content:

1. Definition of Two-Way Tables

2. Examples of Two-Way Tables

Main Topics:

1. What is a Two-Way Table

2. Examples of this common data form

3. What does the analysis mean

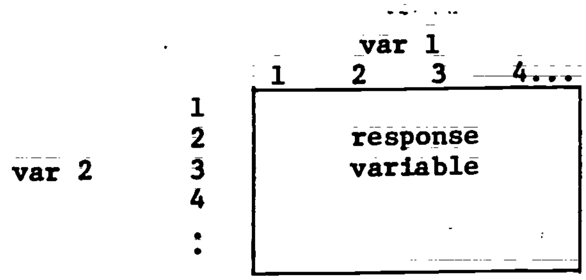
(There are no transparencies for this lecture.)

Reference: Tukey, Chapter 10

Topic 1. Introduction to Unit 8--Two-Way Tables

I. What is a two-way table?

1. A rectangular array of responses laid out in rows and columns



2. Data comes as triples

3. Variables (factors) 1 and/or 2 may be ordinal

4. Response is numeric

II. Examples--common data form

1. Pittsburgh food data

2. Infant mortality be region and year

3. Others? Unemployment by year, reg.

III. What does <u>analysis</u> mean?

1. Question: what effect does each factor have on the response.

Data = row effect + column effect + common

Decomposition into effects
Use residuals for evaluation

2. Question: possible role for tsansformations?

Analytic procedure--Median Polish

Lecture 8-1. Analyzing Two-Way Tables of Responses

Analyzing two-way tables using median polish (Simple Fits): The use of median polish to construct simple summaries of two way tables. (1)

Lecture Content:

1. Discuss simple model for two-way table
2. Discuss median polish

Main Topics:

1. Two-way tables
2. Simple additive summary
3. Median polish

Topic 1. Structure of a two-way table

		Factor 1		
		Level 1	Level 2	... Level n
Factor 2	Level 1			
	Level 2		Response	
	⋮			
	Level m			

I. Simple additive "model" (2) (3)

1. Data = Fit (Response) + Residual

   Response = Contribution (F_1) + Contribution (F_2) + Common

   Common = Typical for entire table

2. Row fit = conditional typical on row

   Column fit = conditional typical on column

3. Row effect = row fit - common

   Column effect = column fit - common

4. Thus,

   Response = Fit = row effect + column effect + common

   or = row fit + column fit - common

986

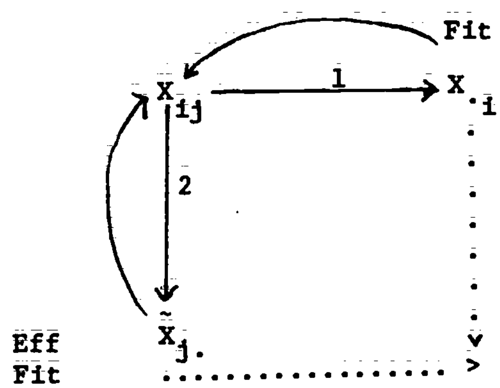
II. Elementary Analysis (Because we have only an additive model) (4)

	F1 levels	row eff	row fit
F2 levels	residuals		
col. eff. col. fit		com	(-com)

1. Technique (using Means or Medians)

   Median Polish: Decomposition of a two-way table into row and column effects by repeated (iterated) removal of medians. (5)

2. Procedure



3. Details of Method

   a. Get row medians and grand median

   b. Subtract from X_{ij} and then get column medians

   c. Check row medians: all zero?

   d. If no--subtract row medians from X_{ij} and get column medians

   e. Check column medians: all zero?

   f. If no--repeat

   g. Yes--add parts (eff.) and common to get fits

4. Fit: Row Effs + Col Effs + common

   or Row Fit + Col Eff

   or Row Eff + Col Fit

   Then, Residual = Data - Fit

5. Construct--stem & leaf--check for symmetry

   Construct--Elementary analysis table--check for non-additivity by examining for opposite corner sign pattern

   Examine effects and fits (6) (7) (8)

   Example 1. High School Grades and GPA

   Example 2. Infant Mortality by region

III. Problems

1. Code residuals

Symbol	Residuals
X	
	uh + step
x	
	uh
.	
	lh
o	
	lh + step
0	

2. Repeated values: take cell medians

3. Holes: Skip--after more polish, get fitted values

989

Lecture 8-1
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 8-1 Outline
Topic 1 Section I		
1.	2	Effects and Common in Multiple Batches
1.	3	Two-way table of Responses
Section II		
1.	4	Two-way table. Elementary Analysis
1.	5	Median Polish: Procedure
5.	6	Predicting Freshman College Grades
5.	7	Median Polish: College Grades 1
5.	8	Median Polish: College Grades 2

[1]

Lecture 8-1

Analyzing Two-way Tables of Responses

Objective: Find simple summary of two-way table.

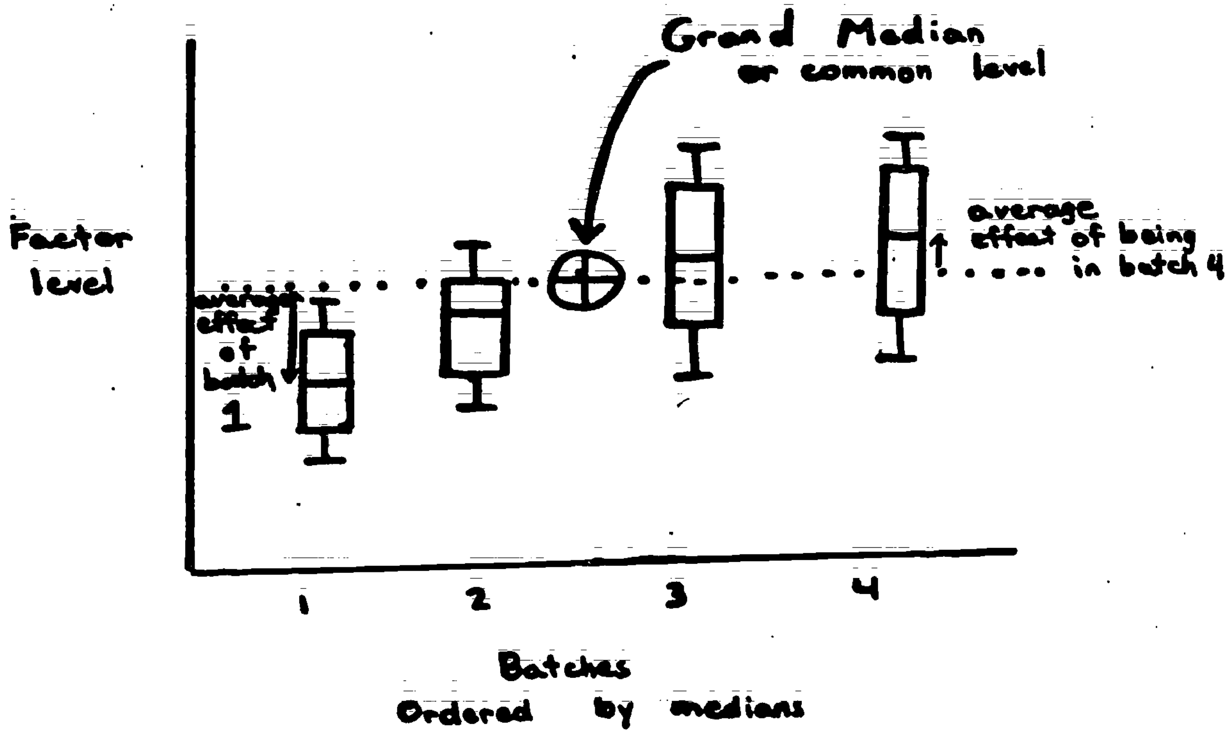
Simple means: additive contributions from both factors.

Summary means: decomposing table into effects and conditional typicals or fits.

Technique: Median polish,

the decomposition of a two way table by repeated (iterative) removal of medians.

The Notion of "Effects" and "Common" in Multiple Batches [2]



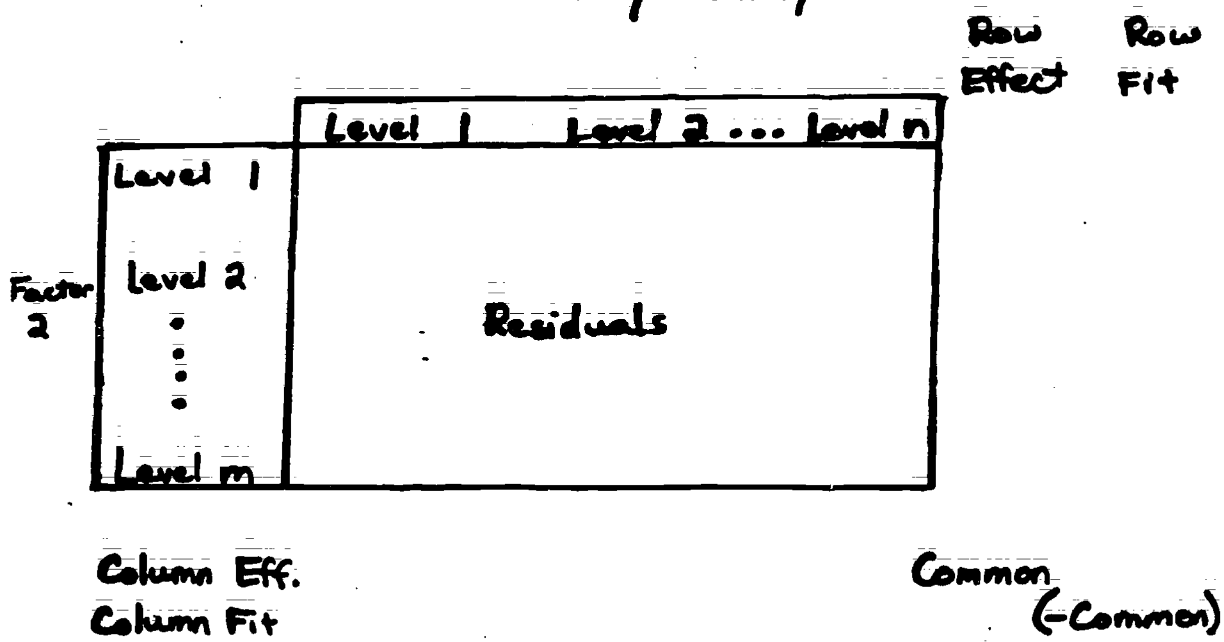
Two-Way Table of Responses (Data) [3]

		Factor 1	
	Level 1	Level 2	.... Level n
Factor 2	Level 1		
	Level 2	Responses	
	⋮		
	Level m		

[4]

Two-way Table

Elementary Analysis



Data = Fit (Response) + Residual

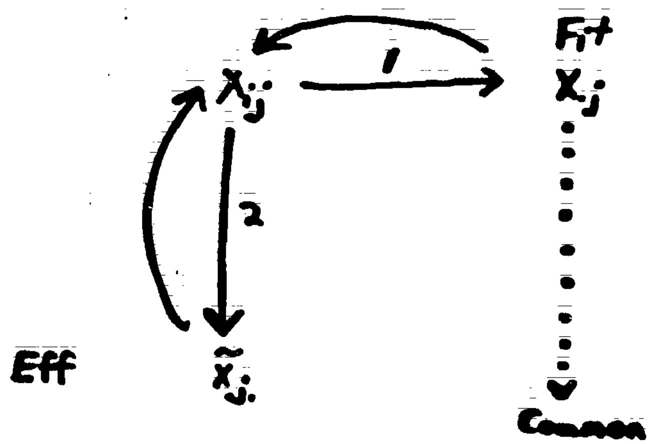
Response = Contribution (F1) + Contribution (F2) + Common

Common = Typical for entire table
Fit = Common + Effect
Row Fit = conditional typical for row
Column Fit = conditional typical for column

8-1

[5]

Median Polish: Procedure



1. Get row medians and grand median.
2. Subtract from X_{ij} and then get column medians.
3. Check row medians: all zero?
4. No - subtract row medians from $\tilde{X}_{ij}$ and get column medians
5. Check column medians: all zero?
6. No - repeat.
7. Yes - add parts and common to get fits.

[6]

Predicting Freshman College Grades from High School Grades

Ave. Grade in H.S.	# M	# F	$\overline{GPA}$ M	$\overline{GPA}$ F
A or A+	1262	1686	2.94	3.08
A-	2035	2732	2.69	[illegible]
B+	2324	3893	2.41	2.59
B	4247	4174	2.18	2.34
B-	3121	1982	2.07	2.15
C+	3014	[illegible]	1.92	[illegible]
C	2312	927	1.77	1.83
D	128	19	1.61	1.73

From : *Predicting Academic Performance in College* by A.W. Astin, Free Press, 1971, p.5.

8-1

Median Polish: College Grades ①

	M	F	Row Med	I→ M	I→ F
A/A+	294	308	301	-7	7
A-	267	283	275	-8	8
B+	241	257	250	-9	9
B	218	234	226	-8	8
B-	207	215	211	-4	4
C+	192	202	197	-5	5
C	177	183	180	-3	3
D	161	173	167	-6	6
			218.5		
Col Med				-6.5	6.5

↓II M	↓II F		III→ M	III→ F
-.5	.5	0	-.5	.5
-1.5	1.5	0	-1.5	1.5
-2.5	2.5	0	-2.5	2.5
-1.5	1.5	0	-1.5	1.5
2.5	-2.5	0	2.5	2.5
1.5	-1.5	0	1.5	1.5
3.5	-3.5	0	3.5	3.5
.5	-.5	0	.5	.5
			0	

8-1

Median Polish: College Grades (2) [8]

Fits

	M	F	Eff.	Fit.	Residuals	
A+/A	294.5	307.5	82.5	301	-.5	.5
A-	268.5	281.5	56.5	275	-1.5	1.5
B+	243.5	256.5	31.5	250	-2.5	2.5
B	219.5	232.5	7.5	226	-1.5	1.5
B-	204.5	217.5	-7.5	211	2.5	-2.5
C+	190.5	203.5	-21.5	197	1.5	-1.5
C	173.5	186.5	-38.5	180	2.5	-2.5
D	160.5	173.5	-51.5	167	.5	-.5
Eff.	-6.5	6.5	218.5			
Fit	212.0	225.0		-218.5		

Elementary Analysis (Original Scale)

	M	F	Eff.	Fit
A/A+	-.005	.005	.825	3.01
A-	-.015	.015	.565	2.75
B+	-.025	.025	.315	2.50
B	-.015	.015	.075	2.26
B-	.025	-.025	-.075	2.11
C+	.015	-.015	-.215	1.97
C	.025	-.025	-.385	1.80
D	.005	-.005	-.515	1.67
Eff.	-.065	.065	2.185	
Fit	2.12	2.25		-2.185

3	5
2	55
1	555
0	5555
1	555
2	55
3	5

8-1

Lecture 8-2. Evaluating Additivity

Diagnostic Plots: Evaluating the adequacy of an additive model as a summary for a two-way table. (1)

Lecture Content:

1. Detecting nonadditivity
2. Computing comparison values
3. Diagnostic plots
4. Transformations

Main Topics:

1. Review additive model
2. Discussion of comparison values and diagnostic plots

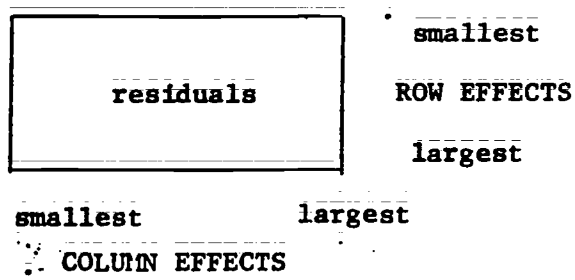
Topic 1. Review additive model

I. **Summary has:**

Data = Row effect + Column effect + common + residual

II. **Departures from additivity**

1. Once elementary analysis is completed, arrange residuals in "effect order"



(2)

2. Examine effect ordered residuals for evidence of opposite corners sign patterns.

+	-
-	+

or

-	+
+	-

Topic 2. Comparison Values and Diagnostic Plots

I. Comparison Values and Disgnostic Plots:

1. Comparison value = $\dfrac{\text{row effect} \cdot \text{column effect}}{\text{common}}$ for each cell (3)

2. Plot residual x comparison value

3. Note that:

   a. If residuals are all around zero this plot will be flat

   b. If residuals equal comparison values or equal comparison values times some constant, there is some non-additive (multiplicative) component in residuals

4. Explore for non-additivity by putting resistant line through plot

5. Flatten diagnostic plot by re-expressing data X_{ij}^{1-m} using ladder of powers

6. Redo entire procedure to determine if re-expression was effective

7. Note:

   a. Weak patterns in the diagnostic plot will not re-express well

   b. Non-monotone patterns require more complicated fits

   c. 1-m must be interpreted loosely--it is a guide, an approximation

8. <u>Example</u>: HS grades and Freshman GPA (4-8)

Lecture 8-2
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 8-2 Outline
Topic 1 Section II 1.	2	Residuals in "Effect" Order
Topic 2 Section I	3	Comparison Values and Diagnostic Plot-
8	4	High School Grades and Freshman Grade Point Average
8	5	High School Grade Data: Residuals and Comparison Values
8	6	Diagnostic Plot of Grade Data
8	7	Diagnostic Plot of Log (Grade Data)
8	8	Elementary Analysis of Log (Grade Data)

1000

[1]

Lecture 8-2

Evaluating additivity in a two-way table using diagnostic plots.

Lecture Content:

1.) Detecting non additivity.

2.) Computing comparison values.

3.) Constructing diagnostic plots.

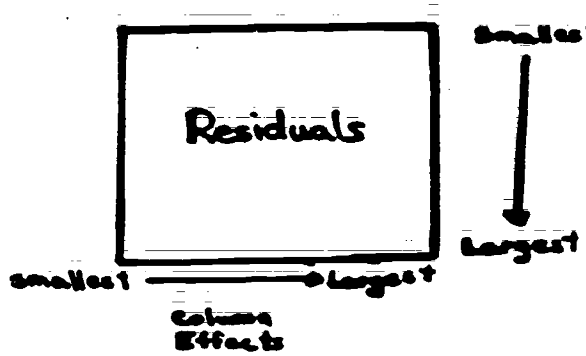
4.) Performing transformations on two-way tables.

8-2

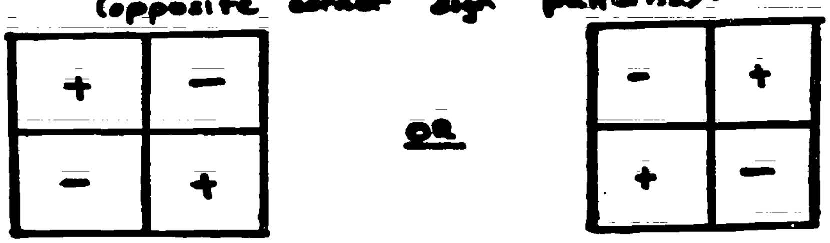
Additive Model:

Data = Row effect + Column effect + Common + Residual

Residuals in "effect order":



Departures from additivity
(opposite corner sign patterns):

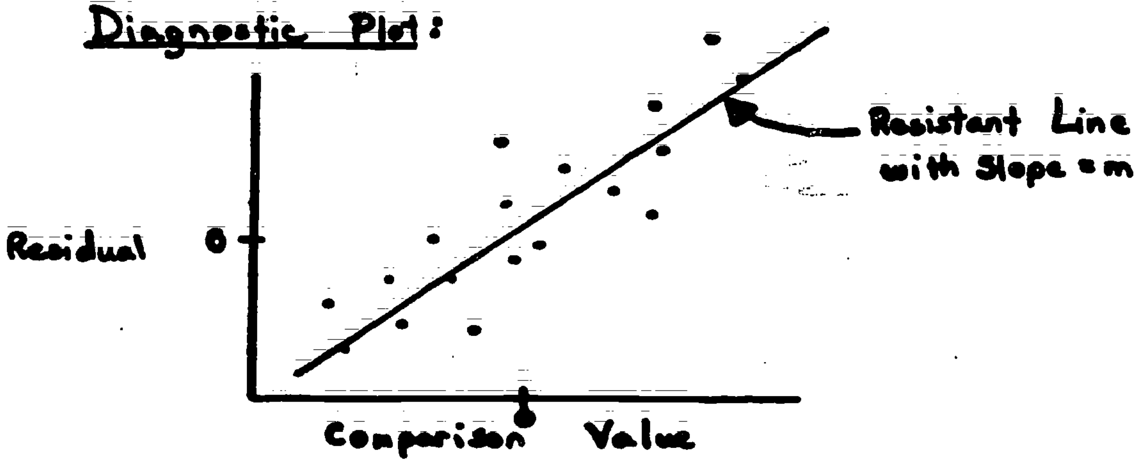


Comparison Value:

$$\text{Comparison Value} = \frac{\text{Row effect} \cdot \text{Column effect}}{\text{common}}$$

N.B.: There is one comparison value for every cell in the table.

Diagnostic Plot:



Transformation:

Transform X_{ij}^{1-m} for all X_{ij} and repeat entire process.

[4]

Example: HS Grades and Freshman GPA by Sex

Original Data

	M	F
A/A+	2.94	3.08
A-	2.67	2.83
B+	2.41	2.59
B	2.18	2.34
B-	2.07	2.15
C+	1.92	2.02
C	1.77	1.83
D	1.61	1.73

Elementary Analysis

	M	F	Row Effs.	Row Fits
A/A+	−.005	.005	.825	3.01
A-	−.015	.015	.565	2.75
B+	−.025	.025	.315	2.50
B	−.015	.015	.075	2.26
B-	.025	−.025	−.075	2.11
C+	.015	−.015	−.215	1.97
C	.035	−.035	−.385	1.80
D	.005	−.005	−.515	1.67
Col. Effs.	−.065	.065		
Col. Fits	2.12	2.25		2.185

Notice opposite corners sign pattern in residuals

8-2

[5]

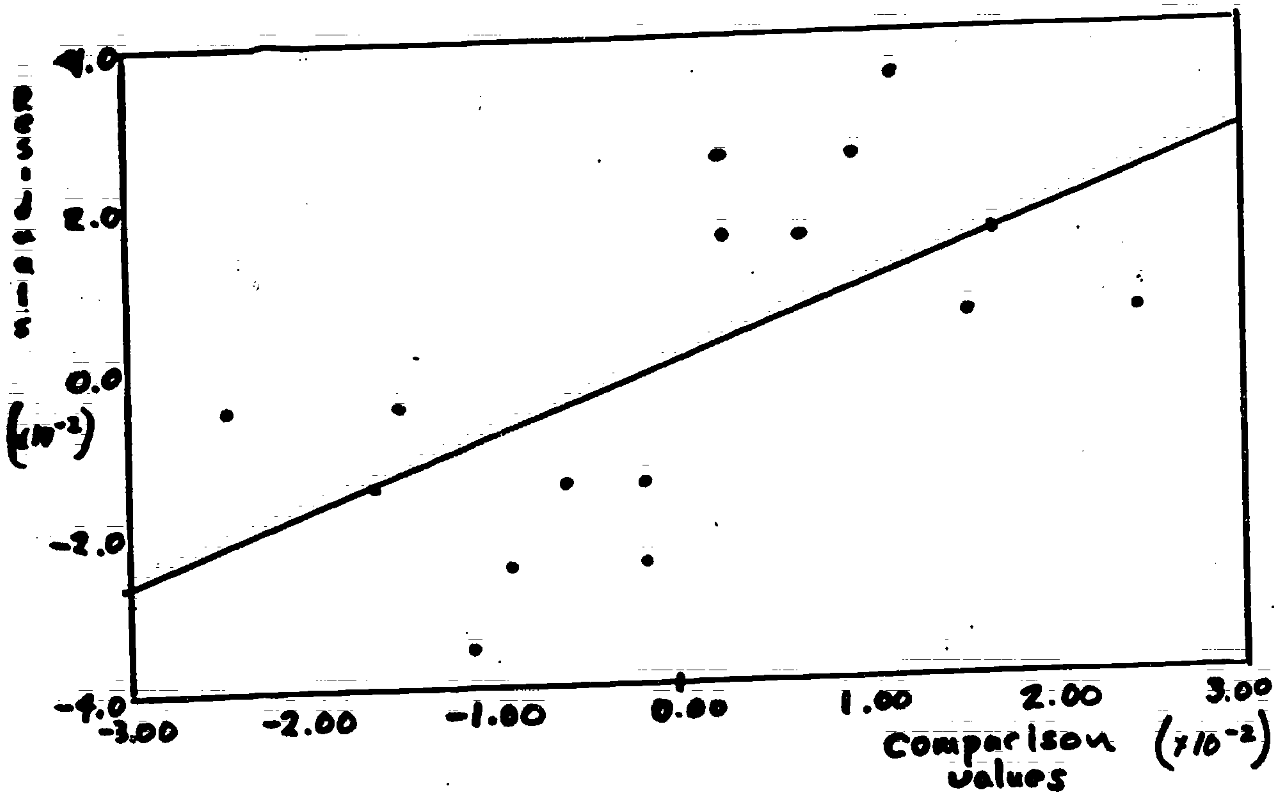
H. S. Grade Data Residuals and Comparison Values

M		F	
Residual	Comparison Value	Residual	Comparison Value
-.005	-.025	.005	.025
-.015	-.017	.015	.017
-.025	-.009	.025	.009
-.015	-.002	.015	.002
.025	.002	-.025	-.002
.015	.006	-.015	-.006
.035	.011	-.035	-.011
.005	.015	-.005	-.015

$$\text{Comparison Value} = \frac{\text{Row Effect} \times \text{Column Effect}}{\text{Common}}$$

[6]

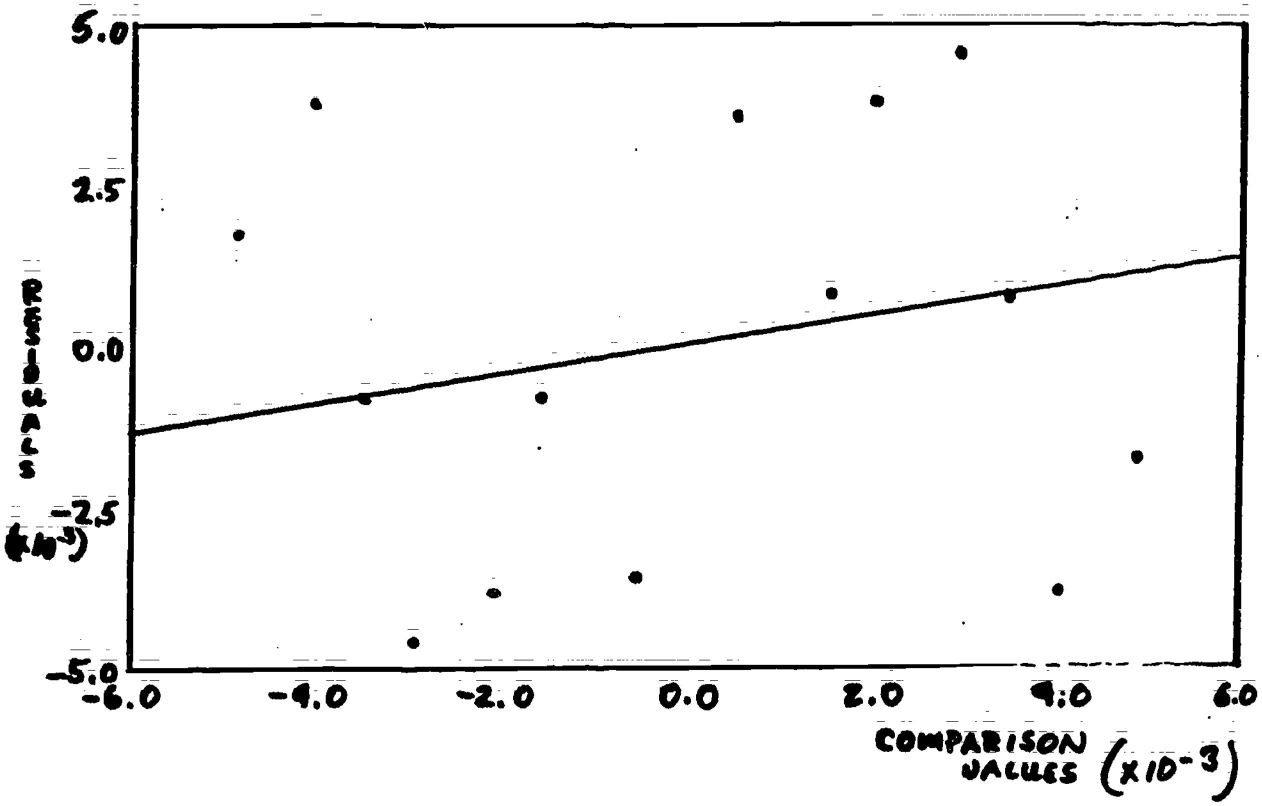
Diagnostic Plot of Grade Data



m = .89

[7]

Diagnostic Plot of Log (Grade Data)



$m = .23$

8-2

[8]

Elementary Analyses of Log (Grade Data)

	M	F	Row Eff.	Row Fit
A/A+	0.00	−0.00	.14	.48
A-	−0	0	.10	.44
B+	−0	0	.06	.40
B	−0	0	.01	.35
B-	0	−0	−.01	.33
C+	0	−0	−.04	.30
C	0	−0	−.08	.26
D	−0	0	−.12	.22
Col Eff.	−0.01	0.01		.34
Col Fit.	.33	.35		

<u>Notice</u>: the sign pattern has been reduced but not eliminated.

Lecture 8-3. Extending the model

Extending the Model: Summarizing effects in ordinal data using fitted lines and developing extended fits for interactions.

(1)

Lecture Content:

1. Discuss summaries for effects

2. Discuss interactions

Main Topics:

1. Plotting effects to construct simple summaries

2. Extended fits

808

Topic 1. Plotting effects to construct simple summaries

I. Simple additive model for categorical data (Review)

	Column Factor	
Row Factor	Response	Row Effects
	Column Effects	common

1. Data = Fit + Residual

   becomes:

   Response = column effect + row effect + common

2. Transformation of the response variable may be required to improve additivity.

   Then model is:

   $(\text{Response})^r$ = column effect + row effect + common

   and the data structure is conceived of as:

	Column Factor	
Row Factor	$(\text{Response})^r$	Row Effects
	Column Effects	common

3. When the factors are categories of categorical variables then a summary formula for the effects is <u>not</u> possible and each effect, one for each category of each factor, must be represented in the model.

1010

II. Additive model with ordinal data

1. When factors have quantitative levels, i.e., ordinal data, then it is possible to consider fitting a model with summaries for effects.

2. Generally, we can try to fit

$$(\text{Response})^{r} = f(\text{Factor } 1) + f(\text{Factor } 2) + \text{common}$$

where the right hand functions are linear or linear through a transformation.

(Note that this is the most general representation. It is not necessary for the response to be transformed or for both factors to be ordinal.)

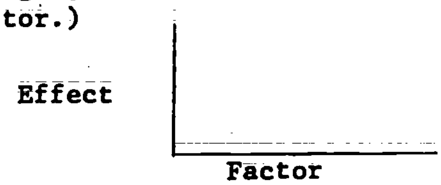
3. An alternative representation:

$$(\text{Response})^{r_0} = (a_1 + b_1 F_1^{\,r_1}) + (a_2 + b_2 F_2^{\,r_2}) + \text{common}$$

(Note that this assumes that the effects for each factor can be summarized as a linear function of the factor's levels.)

III. Finding summaries for effects of ordinal variables.

1. Plot effect against level using x-axis for level. (One plot for each factor.)



2. If it appears to be reasonable, fit a line to the scatterplot (transform factor levels if this is required.)

3. Use the equation obtained as the summary of effects for the factor.

IV. <u>Example</u>: College grade point average as a response to sex of student and high school average grade (3)

1. Simple additive and logged fits (effects only are shown)

2. Plot of row effect against row factor (high school grade.)

line: Row effect = −.56 + .23 (HSG) (3)

Model:

GPA = (−.56 + .23 (HSG)) + (Sex Effects) + common

3. Plot or row effect against row factor (logged response). (3)

line: row effect = −.12 + .04 (HSG)

Model:

log GPS = (−.12 + .04 (HSG)) + Sex Effect + common

V. <u>Example</u>: Moody bonds--net interest as a response to year and grade.

1. Original data and effect analysis (4)

2. Plot of row effects against row factor level (year).

   Line: row effect = −.50 + .30 (year − 1964) (5)

3. Plot of column effects against column factor level (bond grade).

   Line: column effect = −.33 + .26 (grade) (5)

4. Model

Net interest = (−.50 + .30 (year − 1964)) + (−.33 + .26 (grade))

or

Net interest = .30 (year − 1964) + 2.6 (grade) + 3.70

Topic 2. Extended Fits

I. Purpose

1. To include an interaction effect
2. To improve additivity where transformations do not make sense

II. Model

$$\text{Response} = \text{Row Effect} + \text{Column Effect} + \text{common} + k\,\frac{(\text{Row effect})\,(\text{Column effect})}{\text{common}}$$

where k is slope of a line through the diagnostic plot

III. Procedure

1. Perform elementary analysis
2. Construct a diagnostic plot
3. Fit a line and find k
4. Compute difference between residuals from elementary analysis and $k\,\frac{ce.re}{com}$. These are new residuals
5. Contrast improvement by computing sum of absolute residuals
6. Compute fitted values from basic model

IV. Construct example using college grade point average data.

Lecture 8-3
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Beginning	1	Lecture 8-3 Outline
Topic 1		
Section IV.		
1.	2	Grade point average by high school grade and sex
2,3	3	Row effects plotted for raw and logged data
Section V.		
1.	4	Average net interest costs for bonds
2,3	5	Row and column effects plotted

1014

[1]

Extending the Model

Summarizing effects in ordinal data using fitted lines and developing extended fits for interactions

Lecture Content:

1. Discuss summaries for effects
2. Discuss interactions

Main Topics:

1. Plotting effects to construct simple summaries
2. Extended Fits

Example 1 - GPA = HS grade = sex

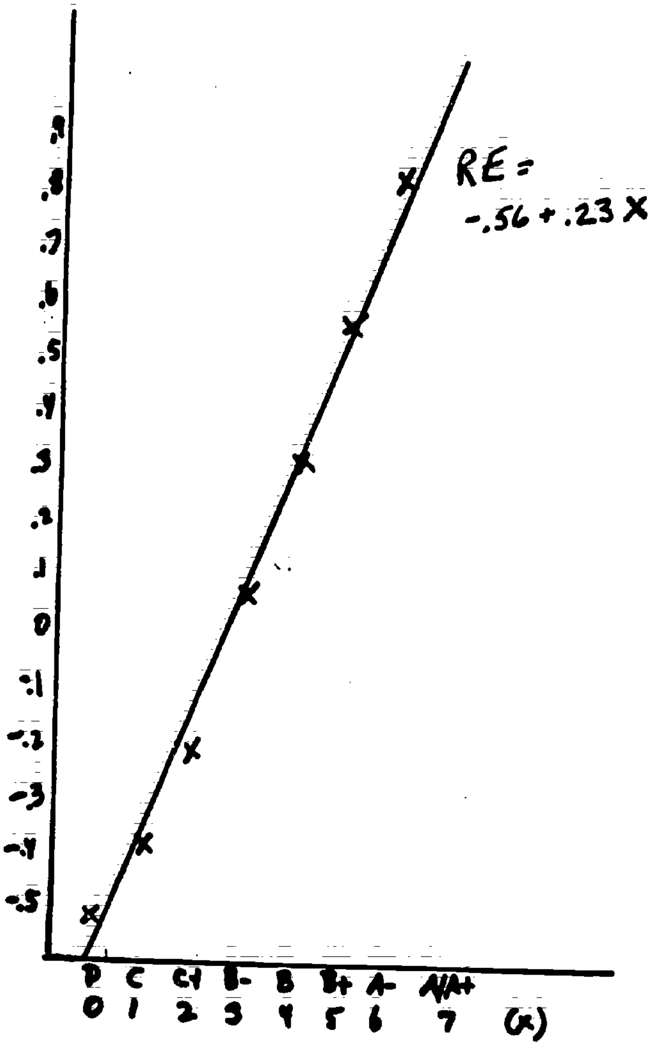
Additive Model

HSG	Sex: M	F	Eff
A/A+			.825
A-			.565
B+			.315
B	GPA		.035
B-			-.075
C+			-.215
C			-.385
D			-.515
Eff	-.065	.065	

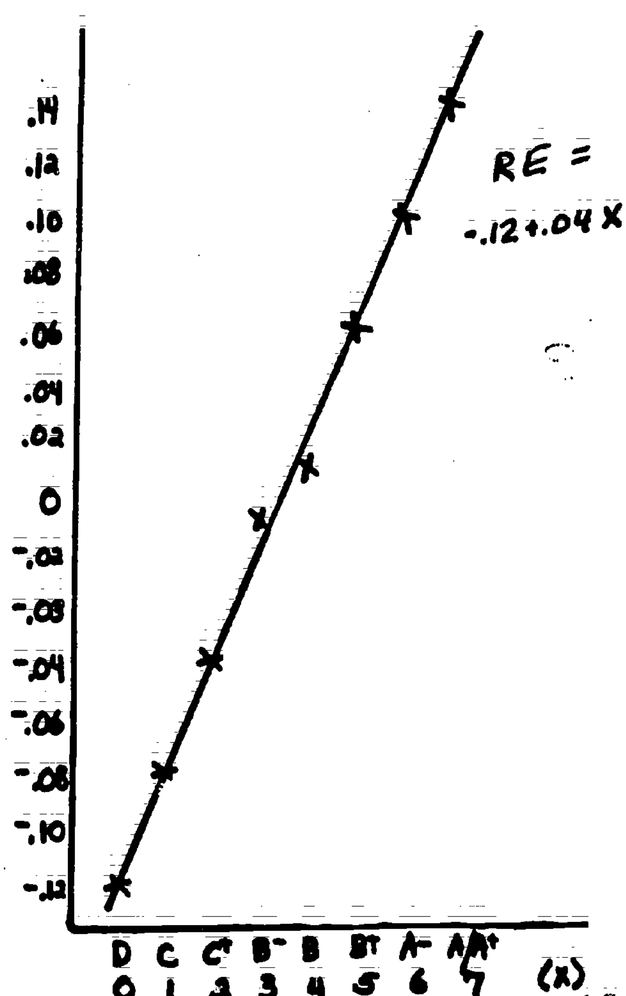
Logged Model

HSG	Sex: M	F	Eff
			.14
			.10
			.06
	GPA		.01
			-.01
			-.04
			-.08
			-.12
Eff	-.01	.01	

GPA × Sex × HS Grade
Raw Data
Row Effects



GPA × Sex × HS Grade
Logged Data
Row Effects



8-3

XVI.IV.47

[4]

Ave. Net Int. Cost (in %)

	Aaa	Aa	A	Baa	Ba
1964	2.88	3.07	3.17	3.43	3.80
1965	2.93	3.11	3.16	3.44	3.76
1966	3.26	3.48	3.56	3.86	4.01
1967	3.56	3.79	3.86	4.17	4.68
1968	3.96	4.23	4.40	4.74	5.05
1969	5.05	4.41	4.73	5.07	5.53
1970	6.04	5.90	6.28	6.71	7.09
1971	5.10	5.02	5.14	5.93	6.60
1972	4.54	4.60	4.92	5.48	5.84
1973	4.53	4.74	4.77	5.18	5.17

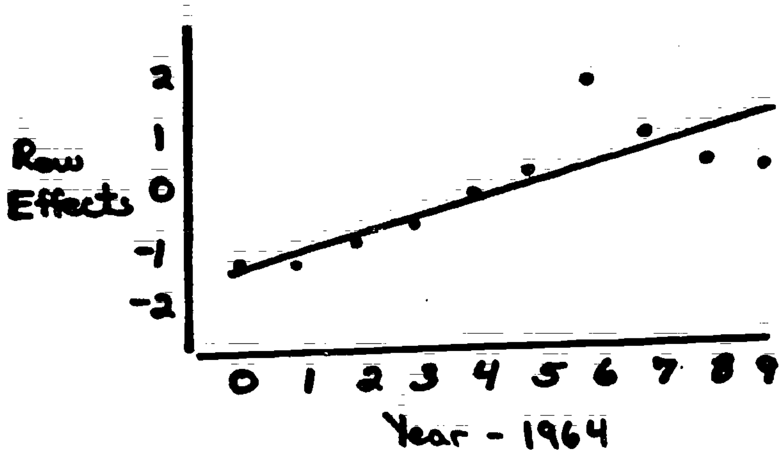
Original Data

						RE	RF
64	0.0	.011	.015	-.083	-.099	-1.379	3.155
65	.045	.046	0.0	-.078	-.144	-1.374	3.160
66	0.0	.041	.025	-.033	-.269	-.999	3.535
67	-.025	.026	0.0	-.048	-.076	-.674	3.86
68	-.103	-.011	.063	.045	-.031	-.196	4.338
69	.595	-.224	0.0	-.018	.056	.196	4.730
70	0.0	-.319	-.035	.038	.031	1.781	6.315
71	0.0	-.259	-.235	.198	.481	.841	5.375
72	-.165	-.224	0.0	.202	.176	.386	4.920
73	0.0	[illegible]	-.015	.018	-.379	.271	4.805
CE	[illegible]	-.096	0.0	.358	.744	4.533	
CF	[illegible]	4.938	[illegible]	4.891	5.278		

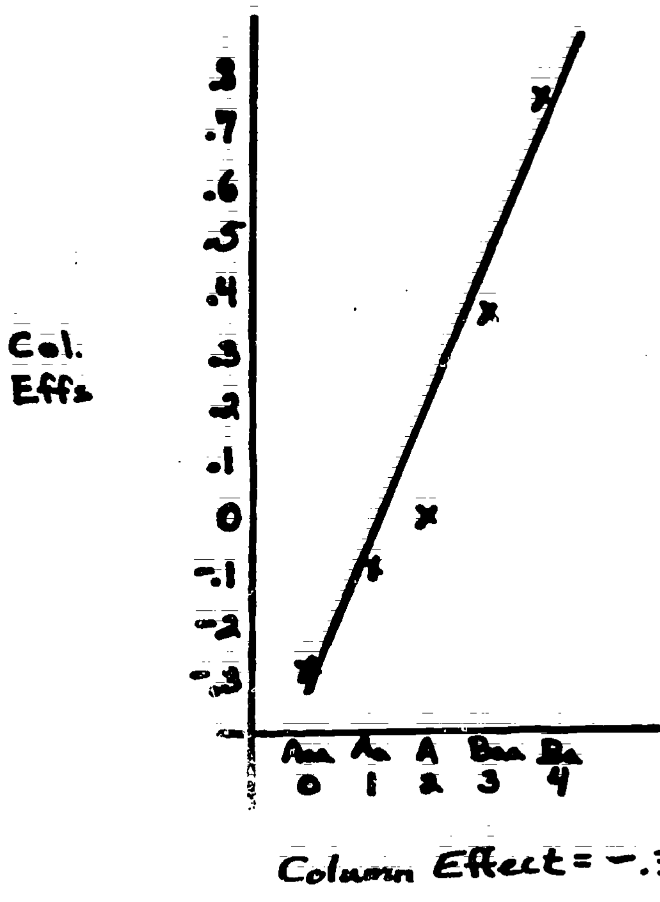
1019

8-3

[5]



Row Effect = −.50 + .30 (Year − 1964)



Column Effect = −.33 + .26X

MOODY BOND DATA (RAW) 8-3

Homework, Unit 8

1. Average interest rates by loan size and geographical region appear below. (Units are % per year).

location	size of loan $1000	$7000	$10000	$30000	$1000000
New York City	10.0	9.0	8.5	8.3	7.8
South and West	10.8	10.4	9.8	9.1	8.6
North and East (Excluding NYC)	10.9	10.3	10.2	9.1	8.6

(a) Median polish this table, and comment on residuals.

(b) Compare the three locations by examining the location effects.

(c) How do the interest rates vary with loan size?

(d) Suggest some reasons fo[illegible]ked effects across loan size and region--assume loan mo[illegible]ommodity sold in a market.

(e) Assume that you are the [illegible] manager of a large, non-profit rehabilitation orga[illegible]anchester, New Hampshire. The organization has de[illegible]nstruct a new building to provide offices and recreational [illegible]ilities. A loan of $100,000 is required to finance construction. Your board of directors suggests that many small loans be made in the Manchester area so that local financial institutions will benefit. But you are concerned with the organization's growing indebtedness and suggest a different strategy to minimize cost. What is the strategy? What arguments would you use in support of your position?

2. Pursuing a study of the equity of basing school support on local property assessments you gather data on assessed values for single family dwellings by age of dwelling and metropolitan area. The data appear below.

(a) Analyze the table.

(b) What substantive explanations can you provide for any consistencies in effects across age and region?

(c) If school support is based on assessed property values, what can you say about the distribution of the burden of school support across these metropolitan areas?

	City				
Age of dwelling	Atlanta	L.A.	D.C.	Chicago	New York
< 5 years	222	238	229	224	243
6-10 years	227	239	225	231	240
11-30 years	222	221	224	212	249
21-30 years	195	216	230	NA	NA
31-40 years	199	214	213	198	192
> 40 years	195	206	205	221	251

Entries are in $.

3. Consider the two two-way tables shown below.

The first gives labor participation rates for women with children in 4 age classifications for 1950-70 in 5 year intervals. The second table gives labor participation rates for married women in 4 age classes for the same years. Entries are % of women with the specified row/column characteristics that are employed; e.g., 11.9% of women with children under 6 were employed in 1955.

(a) Analyze these tables using median polish.

(b) How do children affect the labor participation rates of women? Have these "children effects" been constant over time as evidenced by the columns of the first table? Your supervisor is particularly interested in the "children 6-17" effect. Why is this effect so much higher than the "no children under 18" effect? Why doesn't the rate increase as children become older?

(c) In general, is the participation rate higher or lower for married women than for women with children? Prove to your supervisor that this question is easily answered by examining only one fitted parameter from each table.

(d) Present to your supervisor the two relationships between the 6 years in the tables and the labor participation rates for women with children and the years and married women. Is there a linear relationship in either table? How do the fitted lines compare?

(e) Check to see if the row and column effects are additive in the raw unit of measurement. Check the residuals for any sign patterns and the comparison values for evidence of need for a transformation.

(f) If a transformation of either table is called for, re-express and analyze the transformed data.

LABOR FORCE PARTICIPATION RATES (in %)
MARRIED WOMEN (HUSBAND PRESENT)

AGE	YEARS 1950	1955	1960	1965	1970
20-24	28.5	29.4	30.0	35.6	47.4
25-34	23.8	26.0	27.7	32.1	39.3
35-44	28.5	33.7	36.2	40.6	47.2
45-54	26.8	33.9	40.5	44.0	49.5

Source: Department of Labor, Manpower Report of the President, 1973.

LABOR FORCE PARTICIPATION RATES (in %)
WOMEN WITH CHILDREN (HUSBAND PRESENT)

	YEARS 1950	1955	1960	1965	1976
With Children Under 6	11.9	16.2	18.6	23.3	30.3
Children 0-17	12.6	17.3	18.9	22.8	30.5
Children 6-17 Only	28.3	34.7	39.0	42.7	49.2
No Children Under 18	30.3	32.7	34.7	38.3	42.2

Source: Department of Labor, Manpower Report of the President, 1973.

4. This table shows average infant mortality rates over 1964-1966, whites and blacks, legitimate and illegitimate births, for 4 regions of the United States.

   (a) Analyze this table using both median polish and mean polish. How do the two fitted models compare? If there is a difference in fits, explain why.

   (b) Estimate the infant mortality rate for black illegitimate births in the western United States.

   (c) Suppose you work for an agency in HEW and have a $20 million 1977-78 appropriation for educating expectant mothers in pre- and postnatal care. How should this money be spent? Discuss how the funds should be allocated to regions of the United States. To whom should the educational campaign be directed; specifically, which age groups, which races, etc. The majority of your inferences should be based on this table.

AVERAGE INFANT MORTALITY RATES, 1964-1966
(average annual rates per 1000 live births)

	REGION OF U.S.			
	NORTHEAST	NORTH CENTRAL	SOUTH	WEST
White Legitimate	19.1	22.7	21.7	20.0
White Illegitimate	35.5	33.3	36.5	31.4
Black Legitimate	33.9	44.0	40.4	35.5
Black Illegitimate	43.6	39.9	45.1	NA

Source: Socioeconomic Issues of Health, 1974.

1024

Homework Solutions
Unit 8

Step 1. Find row medians

						med
1a)	10.0	9.0	8.5	8.3	7.8	8.5
	10.8	10.4	9.8	9.1	8.6	9.8
	10.9	10.3	10.2	9.1	8.6	10.2

Step 2. Subtract out row medians; find column medians

						part
	1.5	.5	0	-.2	-.7	8.5
	1.0	.6	0	-.7	-1.2	9.8
	.7	.1	0	-1.1	-1.6	10.2
med	1.0	.5	0	-.7	-1.2	9.8

Step 3. Subtract out column medians and new row medians then new column medians

						med	part
	.5	0	0	.5	.5	.5	-1.3
	0	.1	0	0	0	0	0
	-.3	-.4	0	-.4	-.4	-.4	.4
med	0	0	0	0	0		
part	1.0	.5	.0	-.7	-1.2	common	9.8

Step 4. Subtract out new row medians. All medians now = 0

						med	part
	0	-.5	-.5	0	0	0	-.8
	0	.1	0	0	0	0	0
	1	0	.4	0	0	0	0
med	0	0	0	0	0		
part	1.0	.5	0	-.7	-1.2	common	9.8

Resultant Table

	\$1000	\$7000	\$10000	\$30000	\$100000	eff	fit
NYC	0	-.5	-.5	0	0	-.8	9.0
SW	0	.1	0	0	0	0	9.8
NE	.1	0	.4	0	0	0	9.8
eff	1.0	.5	0	-.7	-1.2	common	9.8
fit	10.8	10.3	9.8	9.1	8.6		

1025

1a) (continued)

```
     Residuals
      Unit = .1

-1 |
-0 | 55
-0 | 00000
 0 | 0000011
 0 | 4
 1 |
```

The residuals tend to be small (0) or large (.4, -.5) as we expect from a resistant procedure. That two of the three large residuals are from NYC suggests further analysis of this location.

1b) Examining the location effects, we immediately note that there is really only one effect--NYC. The other two regions have zero location effects. Further, the NYC effect is large (almost a full percent) and negative--i.e., interest rates in NYC tend to be almost a full percent <u>lower</u> than the NE and SW regions (for the conditions under which the data were collected--time, term, loan size, etc.).

1c) It is quite clear that interest rates <u>decrease</u> monotonically as the loan size increases (for the conditions under which the data were collected).

1d) If we consider loan money as a commodity sold in a market, then much of the size and location effects might be explained by supply and demand. It is likely that money is more available (larger supply) in NYC than elsewhere due to the high density of financial institutions there. We do not, however, expect demand to be correspondingly higher in NYC since money consumers (individual or commercial) are at least as numerous in each of the other two regions. NYC, a "financial capital", thus exhibits lower interest rates.

Similarly, one would expect interest rates to decrease with loan size since (a) there is probably less demand for loans of $100,000 than of $1000, (b) the paperwork for any single loan is probably equivalent, so lender costs for one $100,000 loan would be significantly less than for a hundred $1000 loans, (c) there is probably less risk involved with the larger loans (would <u>you</u> loan $100,000 as readily as $1000?).

Note that these ideas might also help explain the two large residuals for NYC. Suppose large loans ($30,000 or $100,000) are available only from the larger banks while small loans ($1000) are available from all (but mostly the smaller) banks. The lower interest rates for the $7000--$10000 loans might be caused by a relatively smaller demand for these loans offered by the large NYC banking establishments. (A similar trend would not be expected for the $1000 loans since they may not be quite as readily available from those larger banks, and hence not experience quite the same degree of oversupply.) We assume all other considerations (time, term, etc.) are equal.

1e) Should you follow the advice of your board of directors, you would expect to pay over 10% interest on the loans (since interest rates for the NE for loans of $10000 or less are 10.2% or greater). Moreover, should a single $100000 loan be taken, the interest rate would be only 8.6%, a savings of at least 1.6% or $1600.

Of course, an even more clever strategy would be to go to NYC and take the loan there. The resultant interest rate would be 7.8%. This strategy would save at least 2.4%, or $2400 over that suggested by your board of directors.

A reasonable compromise would be to suggest that membe[illegible] your board take a pay cu[illegible] o compensate the company for the larger cost of implementing their plan.

Note that since the organization is non-profit and public service, a consideration such as generating good-will (which often induces corporations to pursue more cosstly strategies) is not an issue. However, if by taking the more costly loans from other local banks, other benefits (such as fund raising aid from these institutions) accrue, a more complex cost-benefit analysis is required.

						med.
2a)	222	238	229	224	243	229
Step 1	227	239	225	231	240	231
	222	221	224	212	249	222
	195	216	230	---	---	216
	199	214	213	198	192	199
	195	206	205	221	251	206

---

						part	
Step 2	-7	9	0	-5	14	229	
	-4	8	-6	0	9	231	
	0	-1	2	-10	27	222	
	-21	0	14	---	---	216	
	0	15	14	-1	-7	199	
	-11	0	-1	15	45	206	
med	-6	4	1	-1	14	219	(common)

---

						med	part	
Step 3	-1	5	-1	-4	0	-1	10	
	2	4	-7	1	-5	+1	12	
	+6	-5	1	-9	13	+1	3	
	-15	-4	13	---	---	-4	-3	
	+6	9	13	0	-21	6	-20	
	-5	-4	-2	16	31	-2	-13	
part	-6	4	1	-1	14		219	common

---

						part	
Step 4	0	6	0	-3	1	9	
	1	3	-8	0	-6	13	
	5	-6	0	-10	12	4	
	-11	0	17	---	---	-7	
	0	3	7	-6	-27	-14	
	-3	-2	0	18	33	-15	
med	0	2	0	-3	1	219	common
part	-6	4	1	-1	14		

---

						med	part	
Step 5	0	4	0	0	0	0	9	
	1	1	-8	+3	-7	1	13	
	5	-8	0	-7	11	0	4	
	-11	-2	17	---	---	-2	-7	
	0	1	7	-3	-28	0	-14	
	-3	-4	0	21	32	0	-15	
part	-6	6	1	-4	15		219	common

Step 6

						part
	0	4	0	0	0	9
	0	0	-9	2	-7	14
	5	-8	0	-7	11	4
	-9	0	19	---	---	-9
	0	1	7	-3	-28	-14
	-3	-4	0	21	32	-15
med	0	0	0	0	0	
part	-6	6	1	-4	15	219 common

Step 7

						med	part
	0	4	0	0	0	0	9
	0	0	-9	2	-7	0	14
	5	-8	0	-7	11	0	4
	-9	0	19	---	---	0	-9
	0	1	7	-3	-28	0	-14
	-3	-4	0	21	32	0	-15
med	0	0	0	0	0	219 common	
part	-6	6	1	-4	15		

The resultant table is

	Atlanta	LA	DC	Chicago	NY	eff	fit
< 5 years	0	4	0	0	0	9	228
6-10	0	0	-9	2	-7	14	233
11-20	5	-8	0	-7	11	4	223
21-30	-9	0	19	---	---	-9	210
31-40	0	1	7	-3	-28	-14	205
> 40	-3	-4	0	21	32	-15	204
eff	-6	6	1	-4	15	common	= 219
fit	213	225	220	215	234		

Note the larger NY effect, and the large NY residuals.
Note that except for < 5 yrs., age effect decreases as age increases.

2b) One might expect assessments to correspond to a large degree to the cost of living of a given area. This certainly appears to be the case, with the greatest location effect for NYC (which has the highest cost of living of those cities examined) and the smallest location effect (greatest negative) for Atlanta. Similarly for LA, DC, and Chicago.

Similarly, we expect assessment to decrease with dwelling age. This certainly appears to be the case, except for dwellings constructed in the past 5 years. Since these are aggregate figures, this phenomenon might be explained by a recent wave of low-cost housing construction in the larger cities.

2c) If school support is based on assessed property values, we expect the burden of this support to fall most heavily upon newcomers to the city, i.e., those who move into the newer ($\leq$ 10 yrs. old) dwellings. This assumes a minimal mobility on the part of longtime residents. In an area of higher internal mobility, the burden of support falls more heavily on those who think they can afford to move into newer ($\leq$ 10 yrs. old) dwellings.

The impact of such an assessment structure may be to create a disincentive to the construction of new homes with a corresponding loss of jobs and a disincentive to mobility by young upwardly mobile families. This would yield reduced total tax revenues and make difficult the support of schooling in general. Such effects would be particularly hard felt in NYC which could probably use such potential tax revenues most. Thus, basing school support on property taxing is not likely to yield equitable education, at least on the basis of these data.

3.a) Table I is analyzed by median polish in Figures A through D, Table II in Figures E through H. In each case, the analysis proceeded as follows

i) The data were entered
ii) The tables were polished (Figures A and E)
iii) The residuals were analyzed (Figures B and F)
iv) A diagnostic plot was made (Figures C and G)
v) A resistant line was calculated for the diagnostic plot (Figures D and H)

By themselves these Figures do not constitute a complete analysis; we must interpret these Figures. This is done in parts (b), (c), (d), (e), and (f) below.

We should, however, be very cautious in our interpretation of these data since we do not know how participation rates were calculated, what the base group was (i.e., these are percents of what group?), what the eligibility requirements for this base group were, and whether in fact all of these considerations were even consistent for all of the years in question. Manipulating these factors (or just changing definitions from year to year), can create a table of "participation rates" which reflect anything we wish.

For this problem, however, we will assume that the above points have already been addressed and answered to our satisfaction.

b) Predictably, the presence of young children (< 6 years old) seems to lower the labor participation rate of women, as evidenced by the large negative effect of the first two rows of Table II (compared to the large positive effect for the fourth row--women with no children under 18). The presence of children in general also seems to lower the average rate (see (c) below).

More interesting, however, is that the rate for women with OLDER children (6 to 17 years of age) is HIGHER than that for women with no children under 18. We can hypothesize at least several reasons why this might be so (although this is an excellent question for further study):

--families with older children are more likely to need the additional income;

--women with children over 18 are likely to be older, and hence possess fewer, obsolescent, or just "rusty" skills. (but see the effects of age in Table I);

1031

--many women whose children are over 18 (note that this is an open ended age bracket; the children could be 37) may have reached voluntary--or mandatory--retirement age.

Note also that these effects are NOT constant over time. In 1950, the rate for women with children between 6 and 17 was lower than that for women with children over 18. Yet thereafter the situation is reversed. This situation might be due to the large utilization of women--especially women without children at home--in the work force during World War II (whose effects would continue for several years, perhaps even through 1950), and perhaps also Korea (1950-1952).

A similar, although temporary, reversal occurs between the rates for women with children under 6 and with children under 17 in 1965. Reasons for this situation are more difficult to propose.

c) Comparing the common values for each table as general indications of OVERALL level, we note that the common value for Table I (33.7) is greater than that for Table II (27.0), which suggests that the participation rate for married women in general is greater (by about 7%) than that for women with children. (This should not be too surprising).

d) First note that there are only FIVE years given in each table.

These plots are shown for Table I in Figure I and for Table II in Figure K. The resistant line for each is shown in Figures J and L respectively.

Both plots are nearly linear, the third and fourth point of each lying somewhat below the fitted line.

Note the similarity in slope between the two fitted lines. (.896 vs. .855). The two lines therefore differ only by a constant of about 8%. (This is calculated by comparing the ordinates at each of several years. We cannot simply compare the constant terms of the two resistant lines since the slopes are not precisely equal).

Note how this corresponds to our answer in (c) above. The participation rate for married women seems to have been consistently (over time) about 8% higher than that for women with children.

e) To check additivity, we examine three indicators:

i) the residual sign patterns (Figures A and E)
ii) the residual behavior (Figures B and F)
iii) the diagnostic plots (Figures C, D, and G, H)

A residual sign pattern for Table I (Figure A) does not seem particularly prominent. The poor behavior of the residuals (Figure B) points to a definite lack of additivity. The diagnostic plot (Figure C) whose slope (as calculated in Figure D) is 1.46, confirms this. Reexpression will be pursued in part (f).

Similarly, there is no residual sign pattern for Table II (Figure E). The residuals (Figure F), while not especially well behaved, have few outliers. The slope of the resistant line (calculated in Figure M) for the diagnostic plot (Figure G) is extremely close to 0, a decisive indication of additivity.

f) We noted in (e) above that the slope of the resistant line for the diagnostic plot for Table I (Figures C, D) was 1.46, confirming the other indications of nonadditivity. This value (approximately 1.5) suggests reexpression to reciprocal roots. Since reciprocals (let alone reciprocal roots) are difficult to interpret, a log reexpression was tried first.

Figures M through P show the analysis of the log data. The behavior of the residuals, and the slope of the resistant line for the diagnostic plot, both suggest the inadequacy of this transformation.

The (negative) inverse reexpression (still somewhat easier to interpret than inverse roots) was tried next. Figures Q through T show this analysis. The residuals are much better behaved, although the slope of the resistant line for the diagnostic plot suggests (predictably enough) a further reexpression by square roots.

An analysis of the (negative) inverse roots might therefore be done next, if the increased additivity is deemed worth the corresponding increase in difficulty of interpretation.

1033

TWO WAY TABLE OF DATA IN VARIABLE: MARWOM   FIGURE A.

	1	2	3	4	5
1:	28.3000	29.4000	30.0000	35.6000	47.4000
2:	[illegible]	26.0000	27.7000	32.1000	39.3000
3:	[illegible]	33.7000	36.2000	40.6000	47.2000
4:	[illegible]	33.9000	40.5000	44.0000	49.5000

ELEMENTARY ANALYSIS BY MEDIAN POLISH.

	1	2	3
1:	2.7125	0.0063	-1.5875
2:	1.6000	0.1938	-0.3000
3:	-1.6000	-0.0063	0.3000
4:	-5.6000	-2.1063	2.3000
EFF:	-5.8000	-2.1938	0.0000
FIT:	27.9437	31.5500	33.7437

	4	5	EFFECT	FIT
1:	-0.3875	4.5125	-2.1563	31.5875
2:	-0.3000	0.0000	-5.7437	28.0000
3:	0.3000	0.0000	2.1563	35.9000
4:	1.4000	0.0000	4.4563	38.2000
EFF:	4.4000	11.3000	33.7437	0.0000
FIT:	38.1437	45.0437	0.0000	-33.7437

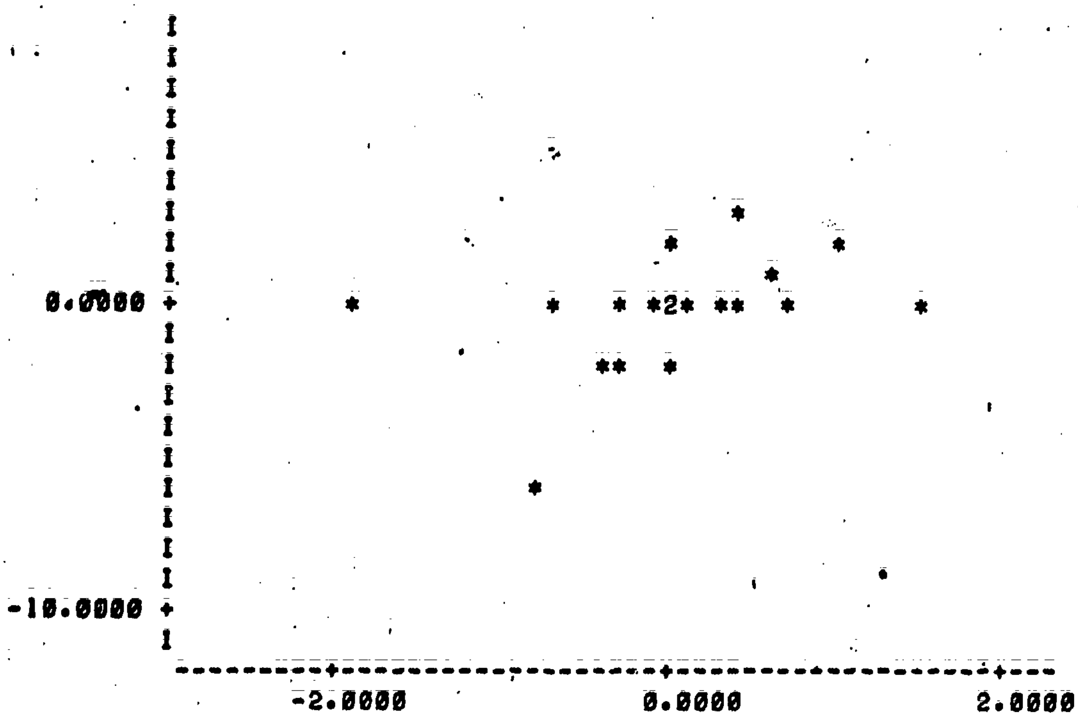
FIGURE B.

```
:STEM RES1
VARIABLE ' RES1 '.
 UNIT =       0.0100
               LO I     -5.6000  -2.1063  -1.6000  -1.5875
   5           -3 I 8
   7           -2 I 99
               -1 I
   8           -0 I 0
(  4)           0 I 0000
   8            1 I 9
   7            2 I 99
               HI I     1.4000   1.6000   2.3000   2.7125
               HI I     4.5125
```

```
:BOXPLOT RES1 THREE

  SCALE UNIT:         0.2000
      -4.0000                  0.0000                  4.0000
-------+-----------------------+-----------------------+----
VARIABLE: RES1
                     *         +++
   *               *          <I+I      **   *   *         *
                     *         +++
```

PLOT RES1 VS CVS1

FIGURE C



Y-AXIS: RES1 SCALE UNIT: 1.0000
X-AXIS: CVS1 SCALE UNIT: 0.1000

FIGURE D

LINE RES1 VS CVS1

AFTER 1 STEPS OF POLISH THE FITTED RESISTANT LINE IS:
RES1 = -0.0938 + 1.4649 * CVS1

QMPM

```
TWO WAY TABLE OF DATA IN VARIABLE: WOMCHILD                    FIGURE E

              1          2          3          4          5
   1:      11.9000    16.2000    18.6000    23.3000    30.3000
   2:      12.6000    17.3000    18.9000    22.8000    30.5000
   3:      28.3000    34.7000    39.0000    42.7000    49.2000
   4:      30.3000    32.7000    34.7000    38.3000    42.2000

ELEMENTARY ANALYSIS BY MEDIAN POLISH.
              1           2           3
   1:      -0.0187     -0.2000      0.0000
   2:       0.0187      0.2375     -0.3625
   3:      -2.9813     -1.0625      1.0375
   4:       2.2812      0.2000      0.0000
 EFF:      -6.6812     -2.2000      0.0000
 FIT:      20.3000     24.7812     26.9812
              4           5                   EFFECT         FIT
   1:       0.5500      0.4625              -8.3812     18.6000
   2:      -0.6125      0.0000              -7.7188     19.2625
   3:       0.5875      0.0000              10.9812     37.9624
   4:      -0.5500     -3.7375               7.7188     34.7000
 EFF:       4.1500     11.2375              26.9812      0.0000
 FIT:      31.1312     38.2187               0.0000    -26.9812
```

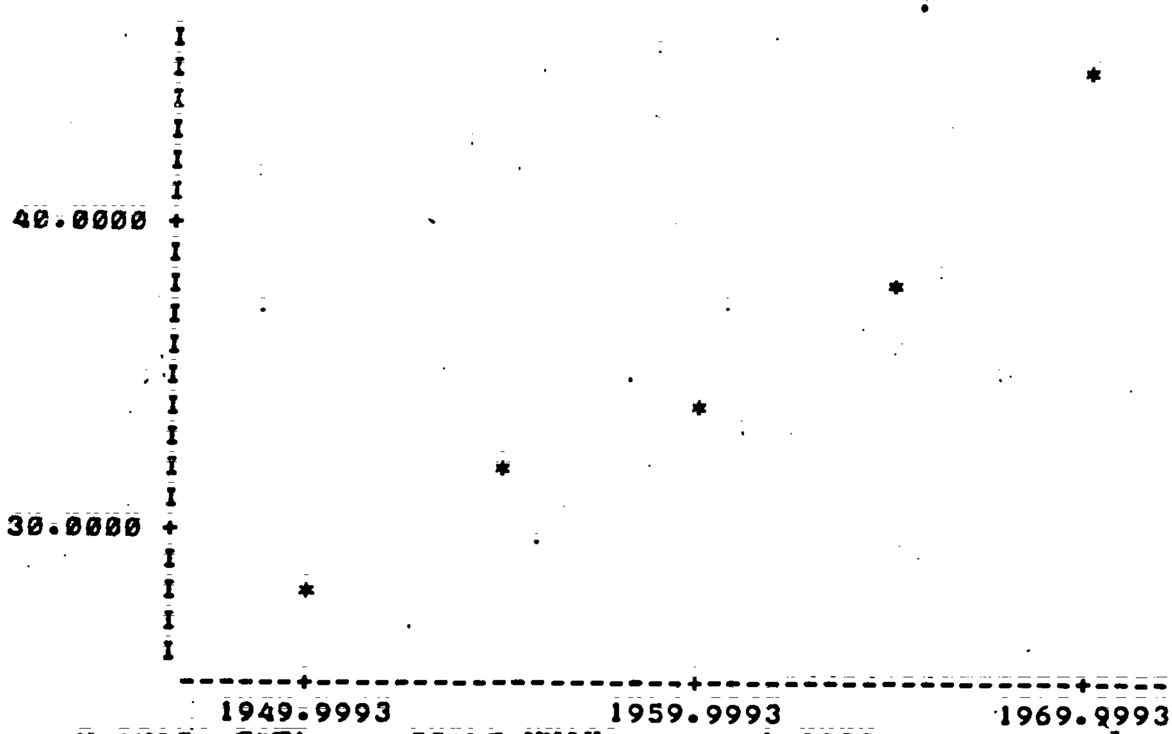
```
STEM RES2                                                FIGURE F

VARIABLE ' RES2 '.
 UNIT =        0.1000
                LO I       -3.7375      -2.9813
   3            -1 I 0
               -0. I
   4              S I 6
   5              F I 5
   6              T I 3
   8             -0 I 10
 ( 6)             0 I 000001
   6              T I 2
   5              F I 455
                  S I
                 0. I
   2              1 I 0
                 HI I        2.2812
BOXPLOT RES2 THREE

    SCALE UNIT:          0.2000
                            0.0000                   4.0000
       ------------------+-------------------+------------
VARIABLE: RES2
                             +-++
      *    *             <--I +I--->       *
                             +-++
```

```
PLOT RES2 VS CVS2                  FIGURE G
             +
             I
             I
             I              *
             I
             I
             I                        *
             I       *          *             *
      0.0000 +        *            *  2  2      2            *
             I                  *     *     *
             I                    *
             I
             I
             I
             I          *
             I                                        *
             I
             I
     -5.0000 +
             I
             I
             ----+-------------------+-------------------+---------
               -4.0000              0.0000              4.0000

    Y-AXIS: RES2      SCALE UNIT:        0.5000
    X-AXIS: CVS2      SCALE UNIT:        0.2000
```

FIGURE H

```
LINE RES2 VS CVS2

AFTER    1 STEPS OF POLISH THE FITTED RESISTANT LINE IS:
RES2 =       0.0802 +      -0.0402 * CVS2
```

```
:INPUT YEARS NOBS 5                  FIGURE I

ENTER DATA
81950 1955 1960 1965 1970

     5 VALUES READ.

:INPUT E/T1 NOBS 5

ENTER DATA
:27.94 31.55 33.74 38.14 45.04

     5 VALUES READ.

:

PLOT E/T1 VS YEARS
```

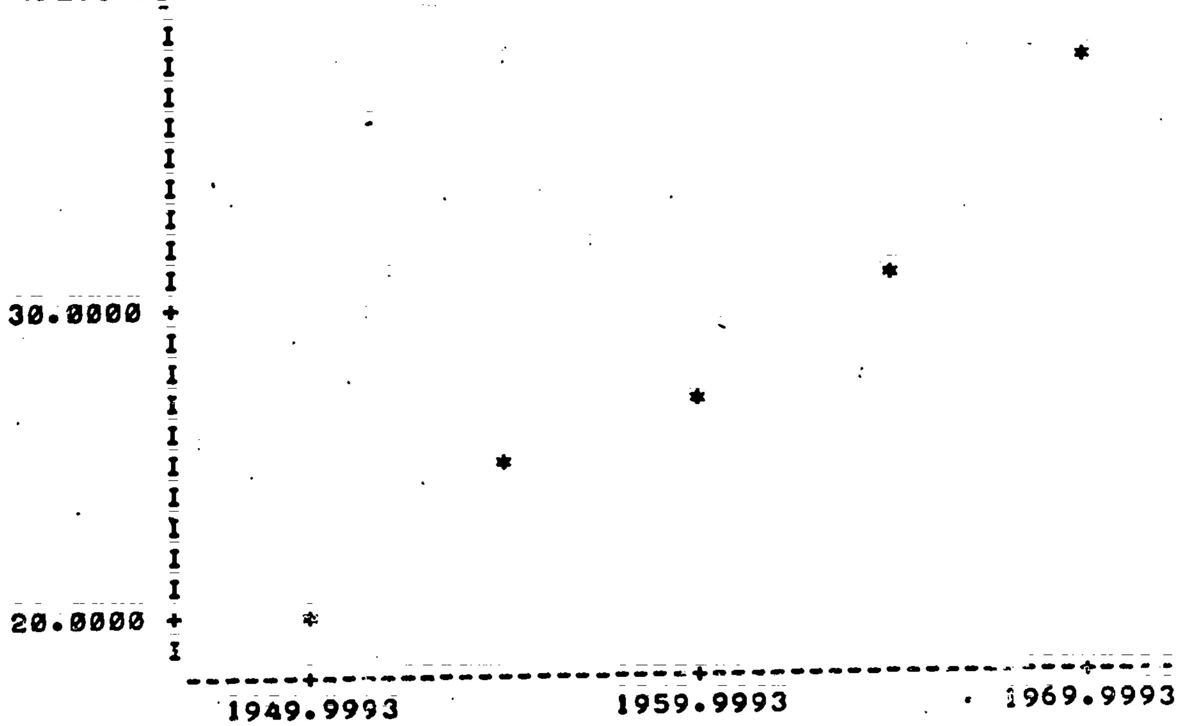



```
   Y-AXIS: E/T1       SCALE UNIT:        1.0000
   X-AXIS: YEARS       SCALE UNIT:         0.5000

:LINE E/T1 VS YEARS                 FIGURE J

AFTER    1 STEPS OF POLISH THE FITTED RESISTANT LINE IS:
E/T1 =   -1639.3252 +      0.8550 * YEARS
```

```
:INFUT FIT2 NOBS 5
```

FIGURE K

```
ENTER DATA
:20.30 24.78 26.98 31.13 38.22

     5 VALUES READ.

:PLOT FIT2 VS YEARS
```




```
    Y-AXIS: FIT2      SCALE UNIT:        1.0000
    X-AXIS: YEARS      SCALE UNIT:        0.5000

 :

  LINE FIT2N\N\  VS YEARS
```

FIGURE L

```
 AFTER    1 STEPS OF POLISH THE FITTED RESISTANT LINE IS:
 FIT2 =  -1726.9070 +      0.8960 * YEARS

 :
```

1041

```
TWO WAY TABLE OF DATA IN VARIABLE: LOGMAR
```

FIGURE M

```
              1          2          3          4          5
   1:      1.4548     1.4683     1.4771     1.5514     1.6758
   2:      1.3766     1.4150     1.4425     1.5065     1.5944
   3:      1.4548     1.5276     1.5587     1.6085     1.6739
   4:      1.4281     1.5302     1.6075     1.6435     1.6946
ELEMENTARY ANALYSIS BY MEDIAN POLISH.
              1            2            3
   1:      Ø.Ø413       Ø.ØØØØ      -Ø.Ø213
   2:      Ø.Ø119      -Ø.ØØ45      -Ø.ØØ71
   3:     -Ø.Ø119       Ø.ØØ[illegible]      Ø.ØØ71
   4:     -Ø.Ø472       Ø.ØØ[illegible]      Ø.Ø472
 EFF:     -Ø.Ø849      -Ø.Ø36        Ø.ØØØØ
 FIT:      1.4401       1.4949       1.525Ø
              4            5                    EFFECT        FIT
   1:     -Ø.ØØ39       Ø.Ø378              -Ø.Ø266      1.4984
   2:      Ø.ØØØØ       Ø.ØØ52              -Ø.Ø754      1.4496
   3:      Ø.ØØØØ      -Ø.Ø172               Ø.Ø266      1.5516
   4:      Ø.Ø263      -Ø.ØØ52               Ø.Ø353      1.56Ø3
 EFF:      Ø.Ø569       Ø.1396               1.525Ø      Ø.ØØØØ
 FIT:      1.5819       1.6646               Ø.ØØØØ     -1.525Ø

:
```

FIGURE N

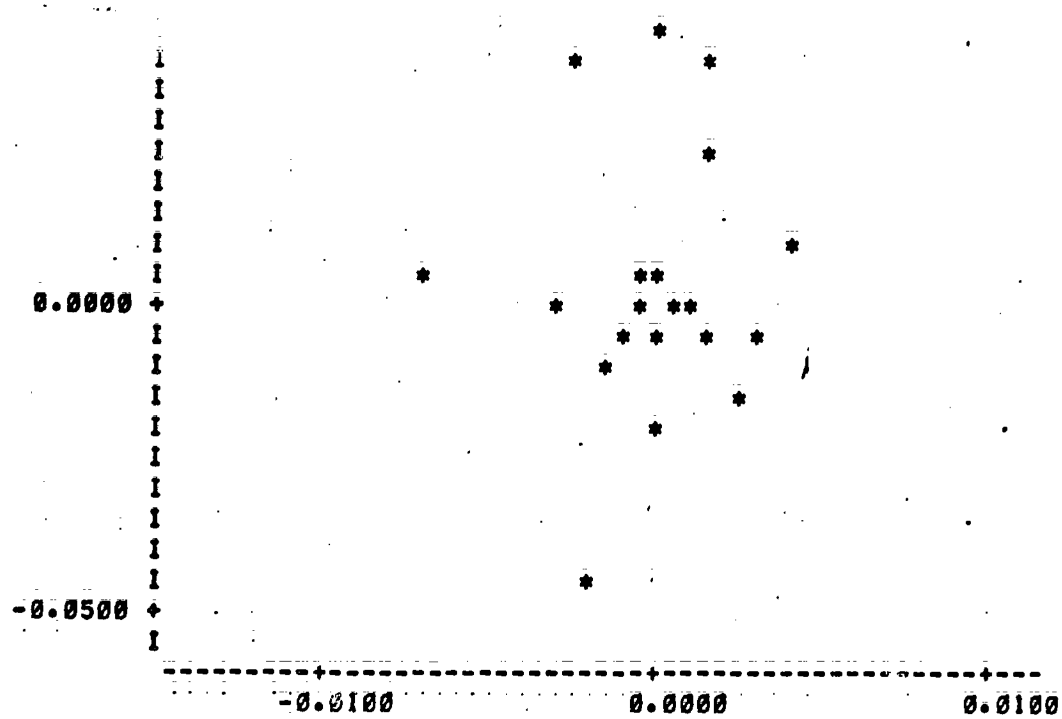
```
 STEM RES4

VARIABLE ' RES4 .
 UNIT =        Ø.ØØ1Ø
                LO I         -Ø.Ø472
   2           -2  I 1
   3           -1. I 7
   4           -1  I 1
   6           -Ø. I 75
   8           -Ø  I 43
(  4)           Ø  I ØØØØ
   8            Ø. I 567
   5            1  I 1
                HI I          Ø.[illegible]3   Ø.Ø378   Ø.Ø413   Ø.Ø472

:
```

```
BOXPLOT RES4 THREE

   SCALE UNIT:          Ø.ØØ2Ø
        -Ø.Ø4ØØ                   Ø.ØØØØ                   Ø.Ø4ØØ
     ----+--------------------+----------------------+--------
VARIABLE: RES4
                                 +---+---+
      *             <-------I    +    I->        *        * *  *
                                 +---+---+
```

1042

PLOT RES4 VS CVS4

FIGURE O



Y-AXIS: RES4 SCALE UNIT: 0.0050
X-AXIS: CVS4 SCALE UNIT8 0.0005

LINE RES4 VS CVS4

FIGURE P

AFTER 1 STEPS OF POLISH THE FITTED RESISTANT LINE IS:
RES4 = 0.00[illegible]2 + 1.1572 * CVS4

TWO WAY TABLE OF DATA IN VARIABLE: MARINV   FIGURE Q

	1	2	3	4	5
1:	-0.0351	-0.0340	-0.0333	-0.0281	-0.0211
2:	-0.0420	-0.0385	-0.0361	-0.0312	-0.0254
3:	-0.0351	-0.0297	-0.0276	-0.0246	-0.0212
4:	-0.0373	-0.0295	-0.0247	-0.0227	-0.0202

ELEMENTARY ANALYSIS BY MEDIAN POLISH.

	1	2	3
1:	0.0034	0.0000	-0.0015
2:	0.0000	-0.0009	-0.0008
3:	0.0000	0.0009	0.0008
4:	-0.0033	0.0000	0.0026
EFF:	-0.0067	-0.0022	0.0000
FIT:	-0.0368	-0.0323	-0.0301

	4	5	EFFECT	FIT
1:	-0.0003	0.0022	-0.0017	-0.0318
2:	0.0002	0.0013	-0.0052	-0.0353
3:	-0.0002	-0.0013	0.0017	-0.0284
4:	0.0006	-0.0015	0.0028	-0.0273
EFF:	0.0040	0.0085	-0.0301	0.0000
FIT:	-0.0261	-0.0216	0.0000	0.0301

:STEM RES4

FIGURE R

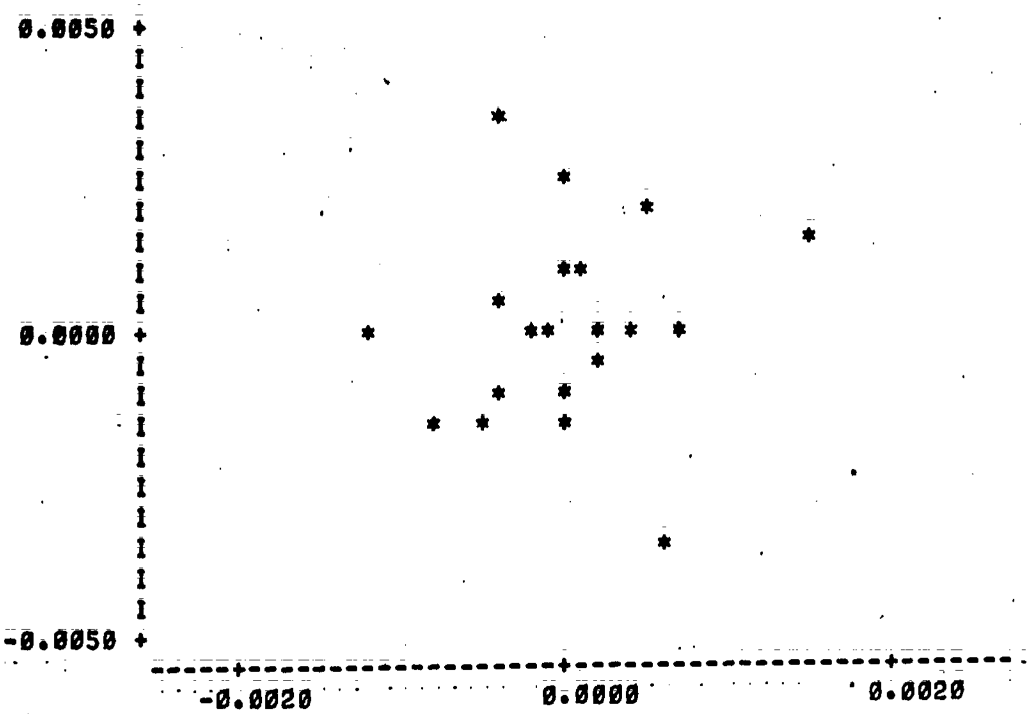
```
VARIABLE ' RES4 '.
  T =          0.0001
               LO  I       -0.0033
  2            -1.  I  5
  4            -1   I  43
  6            -0.  I  97
  8            -0   I  22
( 5)            0   I  00002
  7             0.  I  579
  4             1   I  3
                1.  I
  3             2   I  1
  2             2.  I  6
               HI   I       0.0034
```

:BOXPLOT RES4 THREE

```
    SCALE UNIT:        0.0002
                        0.0000                   0.0040
     ------------------+-------------------+---------------
VARIABLE: RES4
                   +---+---+
       *      <---I    +    I--------->      *
                   +---+---+
```

PLOT RES4 VS CVS4

FIGURE S



Y-AXIS: RES4 SCALE UNIT: 0.0005
X-AXIS: CVS4 SCALE UNIT: 0.0001

:LINE RES4 VS CVS4

FIGURE T

AFTER 1 STEPS OF POLISH THE FITTED RESISTANT LINE IS:
RES4 = -0.0001 + 0.3888 * CVS4

4.a) The analysis using median polish is shown in Figures U through X; the hand calculations are shown in Figure Y for comparison. (Note that hand calculations were done using mortality rates per 10,000 live births. Differences between computer and hand calculations. aside from the units difference, are due to rounding).

The mean polish is shown in Figure Z. Residual and diagnostic plots were not done for the mean polish. See also (d) below.

Comparing the fitted models from the mean and median polish (Figure Y vs. Z) shows the two to be similar, although there is no reason to expect the two to be the same. Indeed, just as we expect the mean and median to be the same only in very well-behaved batches, we expect a difference between the results of mean and median polish in real (and hence not likely to be well-behaved) data.

When performing the analysis on this data, we shall use the median polish, since it has the desirable quality of being resistant.

b). This is estimated by: common + west effect + B-I effect

From Figure U (median polish): 36.076 + 6.705 + (-2.576) = 40.25

From Figure Y (hand calculated median polish): 36.2 + 7.1 + (-2.9) = 40.4

From Figure Z (hand calculated <u>mean</u> polish): 34.0 + 8.3 + (-2.3) = 40.0

(Note the similarity among the three.)

c) One, equitable distribution of funds to each geographical region would be in proportion to its need, i.e., in proportion to the column <u>fits</u>. Hence, we would allocate (from Figure U).

$$\frac{35.8}{35.8 + 36.3 + 38.4 + 33.5} = 24.9\% \text{ of the total to the NE}$$

$$\frac{36.3}{35.8 + 36.3 + 38.4 + 33.5} = 25.2\% \text{ of the total to the NC}$$

$$\frac{38.4}{35.8 + 36.3 + 38.4 + 33.5} = 26.7\% \text{ of the total to the south}$$

$$\frac{33.5}{35.8 + 36.3 + 38.4 + 33.5} = 23.2\% \text{ of the total to the west}$$

The educational campaign in each region might then be directed to each of the four groups in proportion to the calculated fits (or actual observed values) for that region. (NOT the row fits, which "average" overall regions).

The above method however only responds to the data presented to us in Table III. A far better--although long term--solution would be to determine the (probably common) underlying causes of infant mortality and allocate the money to a centralized facility (for medical research or training of medical personnel for example), to regions in proportion to need (for more maternity care beds, or simply more ambulances), or perhaps even to national and regional mass media for educational broadcasting. In any case, UNDERSTAND the problem before pouring money into it. These data do NOT provide all the required information for UNDERSTANDING the problem. We don't even know if the observed patterns are consistent over time.

d) In part (a) we analyzed the raw data by median and mean polish. The slope of the resistant line (Figure X) of the diagnostic plot (Figure W) suggests reexpression. Although the value of the slope (-.39), or about -1/2) suggests reexpression by the 3/2 power, a more easily interpreted reexpression is to square the data (2 power). An analysis of the squared data is shown (by median polish) in Figures AA through DD. Note the slope of the resistant line (Figure DD) of the diagnostic plot for the reexpressed data. We might consider using the fits from THIS analysis (Figure AA) in part (c) above.

TWO WAY TABLE OF DATA IN VARIABLE: INFMORT

FIGURE U

	1	2	3	4
1:	19.1000	21.7000	21.7000	20.0000
2:	35.5000	33.3000	36.5000	31.4000
3:	33.9000	44.0000	40.4000	35.[illegible]000
4:	43.6000	39.9000	45.1000	N.A.

ELEMENTARY ANALYSIS BY MEDIAN POLISH.

	1	2	3
1:	-1.0701	1.0273	-1.0428
2:	1.6752	-1.0273	0.1025
3:	-3.9248	5.6727	0.0025
4:	1.0701	-3.1324	-0.0025
EFF:	-0.2513	0.2513	2.3214
FIT:	35.8248	36.3273	38.3974

	4	EFFECT	FIT
1:	2.1547	-15.6547	20.4214
2:	-0.1000	-2.0000	34.0760
3:	0.0000	2.0000	38.0760
4:	N.A.	6.7051	42.7811
EFF:	-2.5761	36.0760	0.0000
FIT:	33.5000	0.0000	-36.0760

1 MISSING VALUES IN SAVED RESIDUALS.

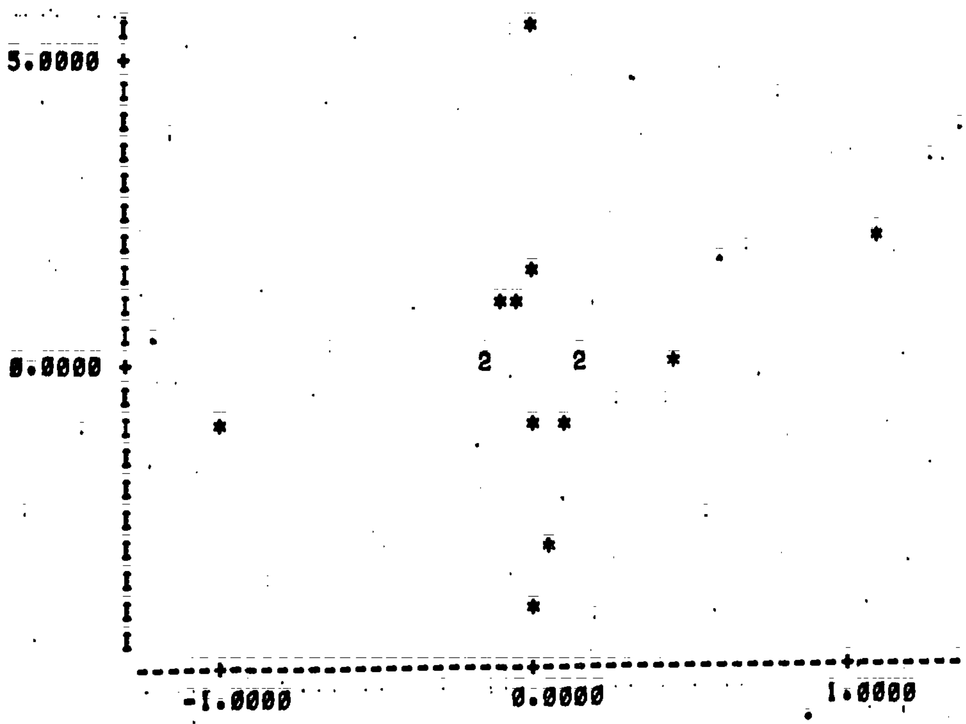
FIGURE V

```
:STEM RES3

VARIABLE ' RES3 '.
 UNIT =        0.1000
   2           -3  I 91
               -2  I
   5           -1  I 000
   7           -0  I 10
 ( 3)           0  I 001
   5            1  I 006
   2            2  I 1
               HI  I        5.6727

:

 BOXPLOT RES3 THREE

    SCALE UNIT:          0.2000
     -4.0000                  0.0000                   4.0000
      +-----------------------+-----------------------+-----------
VARIABLE: RES3
                         +-----+-----+
      <------------------I     +     I------>                  *
                         +-----+-----+
```

1050

PLOT RES3 VS CVS3

FIGURE W



Y-AXIS: RES3 SCALE UNIT: 0.5000
X-AXIS: CVS3 SCALE UNITS 0.0500
1 COORDINATE PAIRS CONTAINED MISSING VALUES AND WERE NOT PLOTTED.

:LINE RES3 VS CVS3

FIGURE X

1 PAIRS CONTAINED MISSING VALUES, WERE NOT ENTERED IN FIT.
AFTER 1 STEPS OF POLISH THE FITTED RESISTANT LINE IS:
RES3 = 0.0526 + -0.3885 * CVS3

Figure Y. Median Polish

					median	
Step 1	191	217	217	200	209	
	355	333	365	314	344	
	339	440	404	355	380	
	436	399	451	---	436	
					part	
Step 2	-18	8	8	-9	209	
	11	-11	21	-30	344	
	-41	60	24	-25	380	
	0	-37	17	---	436	
median	-9	-2	19	-25	362	
					median	part
Step 3	-9	10	-11	16	0	-153
	20	-9	+2	-5	-3	-18
	-32	62	5	0	3	18
	9	-35	-2	--	-2	74
part	-9	-2	19	-25		362
					part	
Step 4	-9	10	-11	16	-153	
	23	-6	5	-2	-21	
	-35	59	2	-3	21	
	11	-33	0	--	72	
median	1	2	1	-2	362	
part	-9	-2	19	-25		
					median	part
Step 5	-10	8	-12	18	-1	-153
	22	-8	4	0	2	-21
	-36	57	1	-1	0	21
	10	-35	-1	--	-1	72
part	-8	0	20	-27		362
					part	
Step 6	-9	9	-11	19	-154	
	20	-10	2	-2	-19	
	-36	57	1	-1	21	
	11	-34	0	--	71	
median	1	0	0	-1	362	
part	-8	0	20	-27		

Figure Y. continued

					median	part
Step 7	-10	9	-11	20	0	-154
	19	-10	2	-1	0	-19
	-37	57	1	C	0	21
	10	-34	0	--	0	71
median	0	0	1*	0		
part	-7	0	20	-29		362

*due to roundoff

median polish
(Unit = .1%)

	NE	NC	S	W	effect	fit
W-L	-10	9	-11	20	-154	108
W-I	19	-10	2	-1	-19	343
B-L	-37	57	1	0	21	383
B-I	10	34	0	--	71	433
effect	-7	0	21	-29	362	
fit	355	362	383	323		

1054

Figure Z. Mean Polish

					mean	
Step 1	191	217	217	200	206.25	206
	355	333	365	314	341.75	342
	339	440	404	355	384.50	384
	436	399	451	---	428.67	429
					part	
Step 2	-15	11	11	-6	206.25	206
	13	-9	23	-28	341.75	342
	-45	56	20	-29	384.50	384
	7	-30	22	---	428.67	429
mean	-10	7	19	-21	340.29	340
					mean	part
Step 3	-5	4	-8	15	2	-134
	23	-16	4	-7	1	2
	-35	49	1	-8	2	44
	17	-37	3	--	-4	89
part	-10	7	19	-21	common =	340
					part	
Step 4	-7	2	-10	13	-132	
	22	-17	3	-8	3	
	-37	47	-1	-10	46	
	+21	-33	7	---	85	
mean	0	0	0	-2		
part	-10	7	19	-21	common =	340
					mean	part
Step 5	-7	2	-10	15	0	-132
	22	-17	3	-6	0	3
	-37	47	-1	-8	0	46
	21	-33	7	--	-2	85
part	-10	7	19	-23	common =	340
					mean	part
Step 6	-7	2	-10	15	0	-132
	22	-17	3	-6	0	3
	-37	47	-1	-8	0	46
	23	-31	9	--	0	83
mean	0	0	0	0	common =	340
Step 7	NE	NC	S	W	effect	fit
W-L	-7	2	-10	15	-132	108
W-I	22	-17	3	-6	3	343
B-L	-37	47	-1	-8	46	386
B-I	23	-31	9	--	83	423
effect	-10	7	19	-23	common =	340
fit	330	347	359	317		

TWO WAY TABLE OF DATA IN VARIABLE: INFMOR2 FIGURE AA

	1	2	3	4
1:	364.8098	470.8894	470.8894	400.0000
2:	1260.2493	1108.8892	1332.2493	985.9600
3:	1149.2090	1935.9985	1632.1597	1260.2493
4:	1900.9587	1592.0078	2034.0081	N.A.

ELEMENTARY ANALYSIS BY MEDIAN POLISH.

	1	2	3
1:	-49.0110	49.1062	-125.0697
2:	110.2163	-49.1062	0.0780
3:	-287.9233	490.9023	12.8886
4:	49.0110	-267.9014	-0.0780
EFF:	-3.9812	3.9812	178.1571
FIT:	1293.5825	1301.5449	1475.7207

	4	EFFECT	FIT
1:	150.2522	-879.7622	417.8015
2:	0.0000	-143.5498	1154.0139
3:	-12.8105	143.5500	1441.1138
4:	N.A.	558.3650	1855.9287
EFF:	-168.0542	1297.5637	0.0000
FIT:	1129.5095	0.0000	-1297.5637

1 MISSING VALUES IN SAVED RESIDUALS.

:STEM RES5 FIGURE BB

```
VARIABLE ' RES5 '.
 UNIT =      10.0000
              LO I     -287.9233    -267.9014
   3          -1 I 2
              -0. I
   7          -0 I 4410
(  5)          0 I 00144
              0. I
   3           1 I 1
   2          1. I 5
              HI I      490.9023
```

:BOXPLOT RES5 THREE

```
    SCALE UNIT:       20.0000
                      0.0000                  400.0000
       -------------+--------------------+-----------
VARIABLE: RES5
                    +-+-+
       **      <---I + I----->                   *
                    +-+-+
```

```
PLOT RESS VS CVS5                                      FIGURE CC
                     I
                     I
         500.0000 +                         *
                     I
                     I
                     I
                     I
                     I
                     I
                     I
                     I
                     I                           *
                     I                          **
           0.0000 +                       2         2            *
                     I                           **
                     I
                     I   *
                     I
                     I                           *
                     I                           *
                     I
                     I
                      -------+-------------------+-------------------+-----
                        -100.0000              0.0000            100.0000

         Y-AXIS: RES5      SCALE UNIT:      50.0000
         X-AXIS: CVS5      SCALE UNIT:       5.0000
        1 COORDINATE PAIRS CONTAINED MISSING VALUES AND WERE NOT PLOTTED.
```

```
LINE RES5 VS CVS5                                      FIGURE DD

  1 PAIRS CONTAINED MISSING VALUES, WERE NOT ENTERED IN FIT.
AFTER   1 STEPS OF POLISH THE FITTED RESISTANT LINE IS:
RES5 =        0.0574 +      -0.0010 * CVS5
```

Unit 8
Quiz

Table 1 appeared in a recent issue of The Retired Officers' Journal. It presents the monthly pay received by members of the US armed forces by pay grade (job levels 4-10) and years of service (12-26) effective October, 1976.

Assume that you are a staff member of a congressional committee which is considering the unionization of the armed forces. Your supervisor wants to contrast pay in the military with pay received by professionals in unionized situations (such as at some universities). But first she wants to understand the table and has asked you to analyze it.

The analysis has been done for you by computer. Parts of the analysis and questions about these parts follow.

1.a Is it true that an individual in the armed forces gets a pay raise every year? Explain your answer by reference to Table 1.

1.b What is the monthly pay for someone pay grade 7 who has been in the service for 16 years?

Table 2 shows the pay data in median polished, "bordered table" form.

2. Based only on Table 2 and the stem-and-leaf display of the residuals from the fit in Figure 1, argue that to determine the monthly pay of an individual in the armed forces one needs more information than pay grade and years of service of the individual. Assume that the individual under consideration is in pay grade 4-10 and has been in the service either 12, 14, 16, 18, 20, 22, or 26 years.

Figures 2, 3, 4 show diagnostic plots of the untransformed data and two transformations, base 10 logarithms and square root.

3. How are "comparison values" defined? In simple layman's non-quantitative language, tell your supervisor (and us, of course) the purpose of the diagnostic plot and why a log or square root transformation might be required.

4. What is the preferable mode of analysis for this table, a transformation of the data, or an extended fit? Give the equation of the extended fit for these data.

Lastly consider the plots of the effects versus respective variable in Figure 5.

5. Construct a simple equation, a function of pay grade and years in service, that approximates the monthly military pay of an individual. What is the yearly pay of an individual, grade 9, with 24 years of service?

TABLE 1.--MONTHLY MILITARY BASIC PAY, OCTOBER, 1976 ("MILITPAY") ENTRIES ARE IN $

	Years of Service						
Pay Grade	12	14	16	18	20	22	26
10	3407	3407	3650	3650	3895	3895	4137
9	2920	2920	3164	3164	3407	3407	3650
8	2804	2804	2920	3047	3164	3291	3291
7	2318	2434	2678	2862	2862	2862	2862
6	1703	1761	2040	2145	2191	2318	2514
5	1586	1692	1820	1924	1982	2051	2051
4	1529	1599	1669	1715	1715	1715	1715

TABLE 2.--MEDIAN POLISH OF MILITPAY TABLE

Pay Grade	Years of Service 12	14	16	18	20	22	26	Effects
10	.	o	.	o	.	.	**+**	1033
9	.	o	.	o	.	.	**+**	546
8	+	.	.	.	.	+	.	358
7	.	.	+	**+**	.	.	o	0
6	o	o	.	.	.	.	**+**	-577
5	.	.	.	.	.	.	o	-797
4	+	+	.	.	o	**0**	**0**	-974
Effects	-315	-255	-72	0	117	173	244	2689

Key to Symbols

Negative		**0**
LH-ΔH		
		o
LH		
M		.
UH		
		+
UH+ΔH		
		+
Positive		

1061

FIGURE 1.--RESIDUALS FROM MEDIAN POLISH FIT.

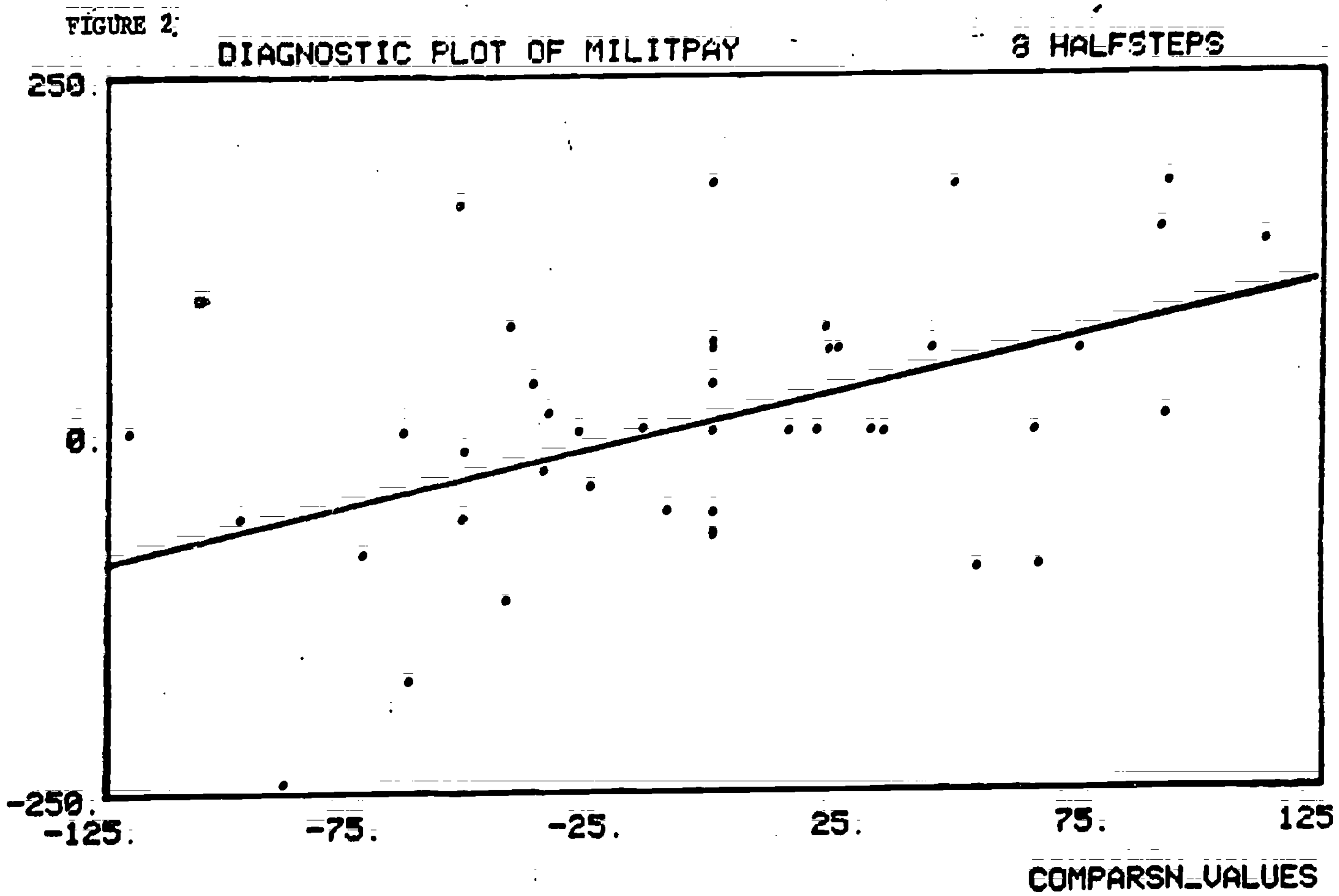
unit = 10^1

```
 LO  |-244, -173

 -1  |1
-0** |998
    s|77766
    f|55
    t|32
 -0  |10
  0  |ZZZZZZZZZZZZZ001
    t|333
    f|55555
    s|677
  0**|

  Hi |129, 139, 158, 171, 171, 173
```

Z = "hard" zero.

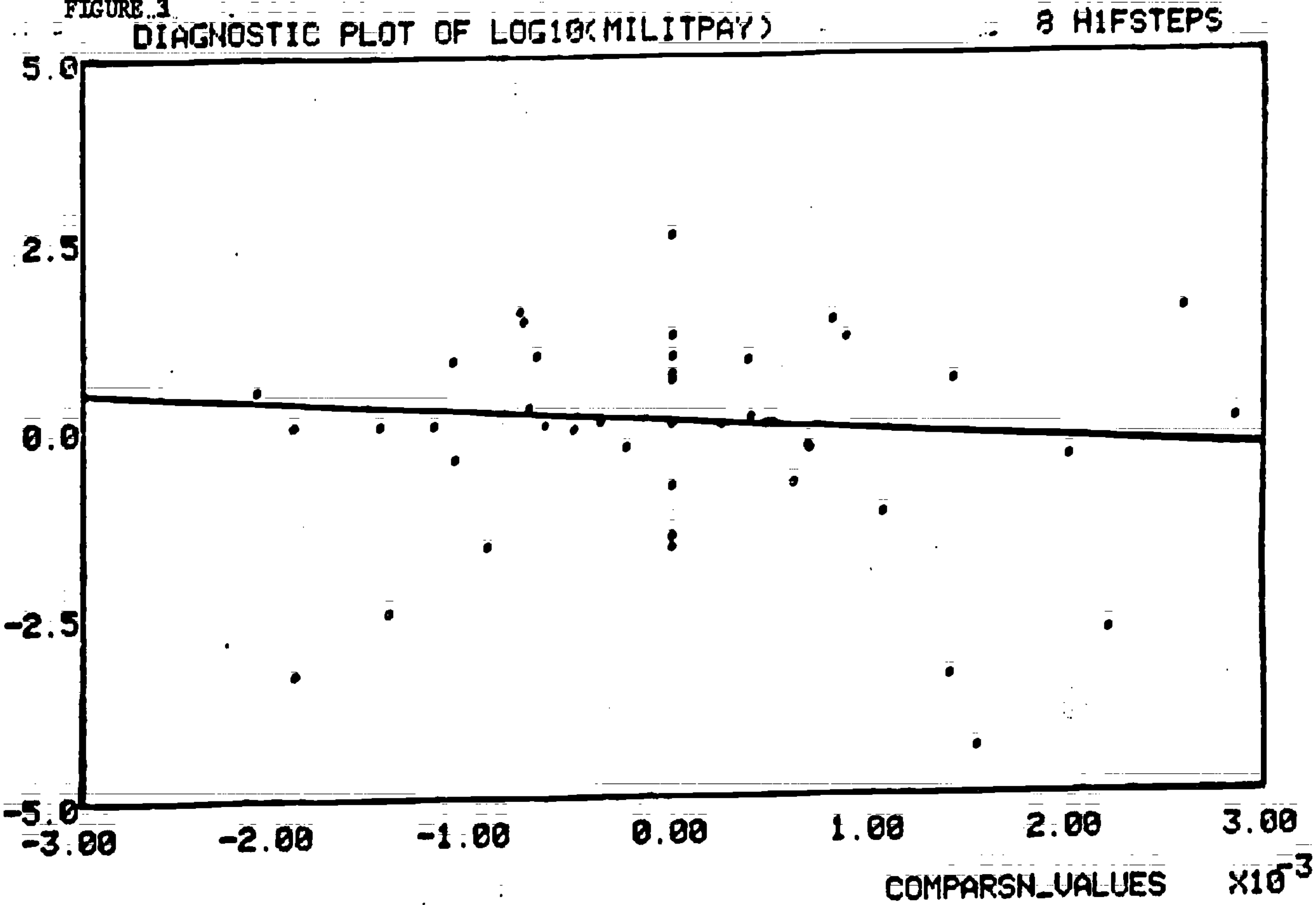
1062

FIGURE 2:



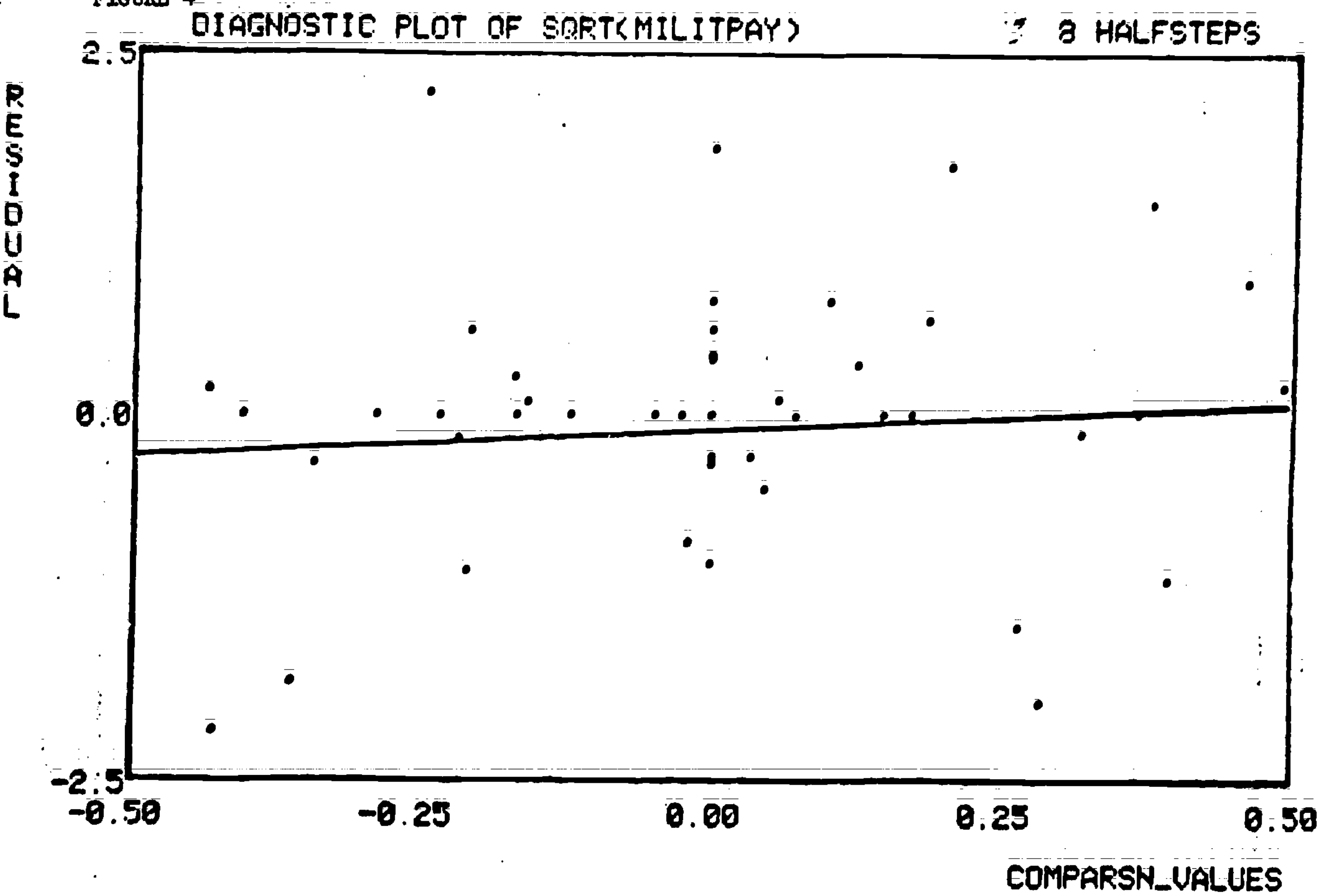
ROBUST EQUATION IS: Y=0.764703X+5.25783

FIGURE 3



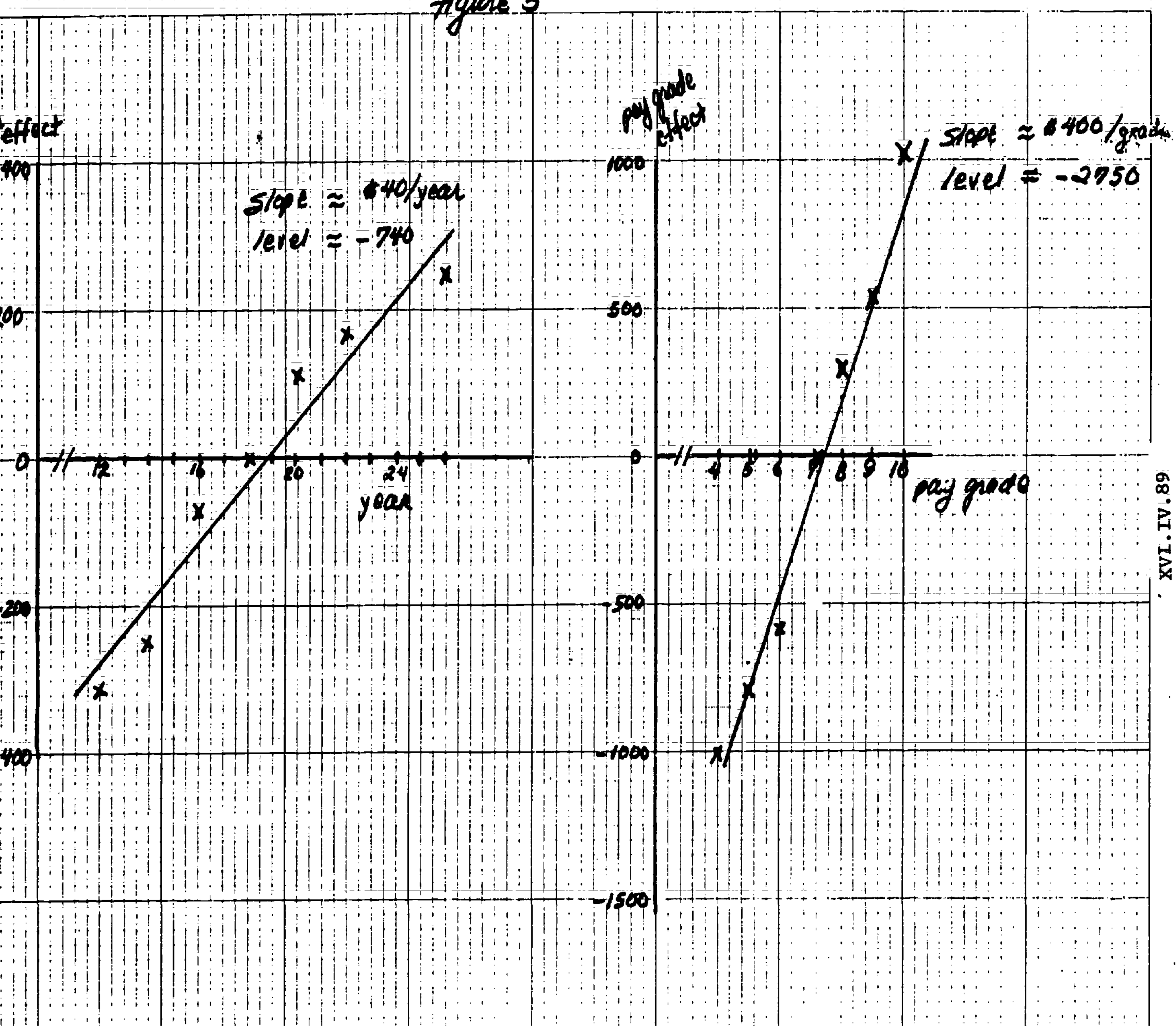
ROBUST EQUATION IS: Y=-1.38231 X+0.000554

FIGURE 4



ROBUST EQUATION IS: Y=0.354679X-0.106905

Figure 5


effect
400
200
0
-200
-400
slope ≈ $40/year
level ≈ -740
12
16
20
24
year
pay grade effect
1000
500
0
-500
-1000
-1500
slope ≈ $400/grade
level ≈ -2750
4
5
6
7
8
9
10
pay grade

Unit 8
Quiz Solutions

1a. Although the general tendency is for pay to increase with years of service, pay does not increase for every 2 years of service within a given pay grade. For example, in grade 4, there is no increase after 18 years of service.

b. Monthly pay for someone in pay grade 7 with 16 years of service is $2,678.

2. First of all, without a table illustrating the residuals the exact monthly pay can only be estimated. Second, the stem-and-leaf of the residuals shows that they are not a well-behaved batch due to the presence of outliers. This would indicate that the additive model is "missing part of the action". Third, the coded residuals also show that the largest and smallest residuals are along the edges of the table, indicating the need for a transformation or an extended fit.

All of these clues indicate that the effects and residuals of the linear model shown in Table 2 are inadequate for summarizing the data and a transformation or extended fit should be tried.

3a. $$\text{Comparison Values} = \frac{\text{(row effect)(column effect)}}{\text{common term}}$$

b. A diagnostic plot, which graphs the comparison values on the horizontal axis and the residuals from a median polish on the vertical axis, is an indication of the adequacy of the additive model. If the plot indicates a linear relationship between the residuals and the comparison values, then the additive model is inadequate. We should try a transformation of the original data or an extension of the additive model, via multiplicative interaction terms. The slope of a linear relationship found in the diagnostic plot should be subtracted from 1 to determine a transformation that might be appropriate. In this particular case, the plot has a slope of .76. Subtracting this from 1.0 gives us .24 (about 1/4). We want to keep the transformations simple so we try a square root instead of a quarter root (moving slightly up the ladder of powers) or a logarithm (moving slightly down on the ladder of powers).

4a. An extended fit of the data is the preferable mode of analysis for this table.

- it makes sense that there would be an interaction between level or grade achieved and the number of years spent in the service.

- the diagnostic plot indicates a quarter root and not one of the simpler transformations we like to work with.
- transformations using square roots and logarithms didn't completely work.
- the coded residuals show that the largest and smallest residuals are at the borders and corners of the table.

To make sure the extended fit was the best, compare the Σ |Residuals| from the extended fit and the two transformations for the smallest sum, (assuming that the residuals were all placed in the same units).

4b. The extended fit model is

$$\text{Data} = \text{Common} + \text{Row Effect} + \text{Col Effect} + K\left(\frac{\text{Row effect} \cdot \text{Col effect}}{\text{Common term}}\right)$$

$$\text{Data} = 2689 + RE + CE + (.76/2689)(RE \cdot CE)$$

$$= 2689 + RE + CE + .00028\ RE \cdot CE$$

5. $$\text{Data} = 2689 + [-740 + 40/\text{year}] + [-2750 + 400/\text{pg}]$$

$$= 2689 - 740 - 2750 + 40/\text{year} + 400/\text{pg}$$

$$= -801 + 40/\text{year} + 400/\text{pg}$$

for Grade = 9, Years = 24

$$\text{Monthly Pay} = -801 + 40(24) + 400(9)$$

$$= -801 + 960 + 3600$$

$$= 3759$$

$$\text{Yearly pay} = 3759 \cdot 12 = \$45{,}108$$

Unit 9
Reading Assignments

Lecture	Reading
9-0	Tanur, Pages 52-65 Bickel, Hammel, O'Connell article in Fairley and Mosteller, pages 113-30
9-1	Mueller, et.al. Pages 480-8 Fienberg, Chapters 1 and 2
9-2	Mueller, et.al. Pages 489-500 Fienberg, Chapter 3
9-3	Fienberg, Chapter 4
9-4	Fienberg, Chapter 5

In addition, please read any articles in Fairley and Mosteller that you have not already read.

Texts:

Fairley, W. and F. Mosteller, Statistics and Public Policy, Reading, Mass.: Addison-Wesley, 1977.

Fienberg, S.E., The Analysis of Cross-classified Categorical Data, M.I.T. Press, in press.

Mueller, J.H., et.al., Statistical Reasoning in Sociology, Third edition, Boston: Houghton-Mifflin, 1977.

Tanur, J., etal., editors, Statistics: A Guide to the Unknown, San Francisco: Holden-Day, 1972.

1073

Lecture 9-0. Introduction to Unit 9

Introduction to Unit 9, Discrete Multivariate Analysis

Lecture Content:

1. Discrete vs. Continuous Multivariate Data
2. Multinomial Distribution for Contingency Tables
3. Examples

Main Topics:

1. Discrete Multivariate Data
2. Multinomial Distribution
3. Examples of Contingency Tables

Topic 1. Discrete Multivariate Data

I. Basic Issue: New "type" of data

1. Everything we have discussed thus far, both response and carrier variables, has been continuous

2. This implies that within a specific range, the dependent variable could take on any possible value

3. For this unit, we change this assumption

II. Problem: How do we structure "discrete" data?

1. We now assume that we have a set of variables that take on only a finite number of discrete values

   a. Moreover, within this set we have no "independent/dependent" dichotomy

   b. Example: Alive/Dead variable;only two values or categories

2. We take all our variables and look at all combinations of the categories

   a. We examine all possible intersections

   b. Each intersection is called a cell

3. We then take a sample (perhaps exhaustive) from a population, sample size N, and record the number of observations falling within each cell

4. Number of observations in each cell is called the frequency count of the cell

III. Solution: Data structure is a Contingency Table

1. The set of all cells and the frequencies of the cells is called a contingency table

2. The set of all frequencies is known as a Discrete Multivariate Data Set

   a. The number of variables, n, is the dimensionality of the contingency table

   b. n may be 1, 2, 3, etc.

IV. Methods: How do we analyze a contingency table?

1. Generally researchers have calculated a X^2 statistic for the table and stated whether the statistic was greater than the tabulated 5% X^2 value, and then called it quits

2. No one really knew what to do with a table of dimension ≥ 3 -- could only handle 1 or 2 dimensional tables

3. Lately, we have begun to understand higher dimensional tables and have developed a sophisticated new technology--the log-linear model--for the analysis

Topic 2. Multinomial Distribution

I. Basic Issue: Probability model for a Contingency Table

1. We have k cells

2. P{observation lands in the ith cell} = p_i; i ranges over all cells

3. We take a sample $\underset{\sim}{Y}$ of size N

   a. $\underset{\sim}{Y} = (y_1, y_2, \ldots, y_N)$

   b. y_j = appropriate cell for jth observation

4. Let x_i = number of observations falling in the ith cell

II. Solution: Multinomial Distribution

1. $P\{X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k\} =$

$$\frac{N!}{\prod_{i=1}^{k} x_i!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

where $\sum_i p_i = 1, \quad \sum_i x_i = N$

2. So the probability distribution for a k dimensional contingency table is the multinomial

1077

Topic 3. Examples of Tables

I. 1 dimensional table--simple multinomial (2)

1. Test fit to a known distribution (uniform)

2. X^2 goodness of fit test

II. 1 dimensional table--simple multinomial. Another goodness of fit test, but what distribution

III. 2 dimensional table--2x3 (3)

1. Test for independence between the variables

2. X^2 test for independence

IV. 3 and 4 dimensional tables (4)

1. What do we do?

2. Independence between which variables?

3. There are many different models to consider

Lecture 9-0
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Topic 2 Section I 1.	1	Multinomial Distribution
Topic 3 Section I	2	Examples
III	3	Examples
IV	4	Examples

1079

Multinomial Distribution

Have a table with K cells
$P\{\text{observation belongs in the ith cell}\} = P_i$
i ranges over all cells.

Take a sample $\underset{\sim}{Y}$ of size N individuals or observations.

$$\underset{\sim}{Y} = (y_1, y_2, \ldots, y_N)$$

y_j = appropriate cell for j^{th} observation

Let X_i = #observations falling in the ith cell.

$$P\{X_1 = x_1, X_2 = x_2, \ldots, X_K = x_K\} =$$

$$\prod_{i=1}^{K} \frac{N!}{X_i!} P_1^{x_1} P_2^{x_2} \cdots P_K^{x_K}$$

where $\sum_i P_i = 1$ and $\sum_i x_i = N$

So the probability distribution for a K dimensional contingency table is the Multinomial.

9-0

[2]

Examples of Discrete Multivariate Data Sets (Contingency Tables)

1. Simple Multinomial - One dimensional Test for fit to a specific distribution (Uniform).

Random numbers, sample of N = 250

Digit

	0	1	2	3	4	5	6	7	8	9	
frequency	22	24	28	23	18	33	29	17	31	25	250

H_0 : Observations are uniformly distributed.

Should have 25 in each cell.

2. Another Simple Multinomial

Number of men on base when a home run is hit. Data from National League for a particular year.

Number On Base

	0	1	2	3	
Frequency of home runs.	421	229	96	21	765

What distribution?

Truncated Poisson, or Binomial

9-0

③ Two Dimensional Table

Distribution of Soviet Population in 1959
Published data by Party and Age.

Party Membership	Age 16-29	30-49	50+	Totals
Yes	1,623,109	4,663,348	1,952,674	8,239,131
No	51,585,678	48,597,649	36,908,810	137,092,135
% Party	3.05	8.76	5.02	5.67

Test for Independence of party and age

④ Three Dimensional Table

Sex, Marital Status, and Happiness 1972 NORC

		Not too Happy	Pretty Happy	Very Happy	TOTALS
Male	Single	15	86	20	121
	Married	52	254	203	509
	Ex-married	20	36	12	68
					698
Female	Single	10	86	18	72
	Married	45	254	254	565
	Ex-married	54	36	31	165
					802
TOTALS		196	766	538	1500

9-0

[4]

5. Four Dimensional Contingency Table 2^4

"Two" panel Study (Two Interviews) where at each interval, the respondent was asked if he or she had seen an advertisement for a certain product and if he or she had bought the product.

First Interview		Second Interview: See Yes, Buy Yes	Second Interview: See Yes, Buy No	Second Interview: See Yes, Buy Yes	Second Interview: See Yes, Buy No
See	Buy				
Yes	Yes	83	8	35	7
	No	22	68	11	88
No	Yes	25	10	95	15
	No	8	32	6	493

1083

9-0

Lecture 9-1. Simple Multinomials

Simple Multinomials--Testing for Goodness of Fit

Lecture Content:

1. Determination of appropriate probability models
2. Pearson's X^2 test for goodness of fit
3. Discrete probability models
4. Continuous probability models

Main Topics:

1. Making direct inferences about distributions
2. Specific probability models to fit

Topic 1. Making direct inferences about distributions

I. Basic Issue: Does population distribution have a specific form?

1. Examine empirical (sample) distribution

   a. Group data into a set of qualitative classes, C of them

   b. Compute expected frequencies for each cell with specific hypothesized distributions

2. Test the "goodness" of various theoretical distributions for the data (1)

II. Solution: Goodness of Fit Test

1. Have some Null hypothesized expected frequencies $\{E_i\}, \sum_i E_i = N$

2. Data give you observed frequencies $\{O_i\}$

3. Compute

$$X^2 = \sum_{i=1}^{C} \frac{(O_i - E_i)^2}{E_i}$$

4. X^2: weight the squared difference of O_i and E_i inversely by E_i; cells with large departures get more weight if E_i is small (2)

5. The quantity X^2 is called Pearson's Chi-Square Statistic

III. Method: How do we determine whether to reject H_0?

1. X^2 for large N, is distributed as a χ^2 random variable with C-1 degrees of freedom, when H_0 is true (3)

2. We lose 1 d.f. since N is fixed

1085

3. If $X^2 > \chi^2_{C-1;\alpha}$, reject H_0

   a. Probability that the sample data accord with H_0 is quite small

   b. Doubtful that this observed X^2 could have origin by chance

4. When can we use this inferential procedure?

   a. Each and every sample observation falls into one and only one category or class interval

   b. The outcomes for the N observations in the sample are independent

   c. Sample size N must be large

      i. If $C-1=1$, $E_i>10$

      ii. If $C-1>1$, $E_i>5$

Topic 2. Specific Probability Models to fit

I. Discrete Models

1. Binomial (n,p)

   a. n or fewer cells

   b. If we estimate p, we lose an additional 1 df

2. Poisson (λ)

   a. ? cells

   b. If we estimate λ, we lose an additional 1 df

II. Gaussian probability model

1. Must take infinite range and break it up into a finite number of cells

2. Postulate Gaussianity-

   a. Convert every observation into a standard score

   b. Lose 1 df for each parameter (μ, σ) that we must estimate

3. How many cells? Suppose we desire C.

   a. Make intervals of equal width.

$$\text{Min},\ \text{Min} + \frac{\text{Max-Min}}{C},\ \text{Min} + \frac{2\,(\text{Max-Min})}{C}, \ldots,\ \text{Max}$$

   b. Or make each interval such that probability of an interval is 1/C

4. Fixed probability intervals are preferred to fixed width intervals.

Lecture 9-1
Transparency Presentation Guide

Lecture Outline Location	Transparency Number	Transparency Description
Topic 1 Section I 2.	1	Study of Educational Achievement
Section II 4.	2	Sample and Expected Results
Section III 1.	3	Pearson's Chi-Squared Statistic

1088

[1]

Study of Educational Achievement (1976)

Population = set of all American citizens at least 25 yrs. of age.

Each subject can be placed into 6 educational categories based on his/her maximum formal educational achievement.

We also know the distribution of each of the 6 categories in 1960:

	Frequency (P_i)	
1. College Grad.	.18	
2. Some College	.17	6 naturally exclusive and exhaustive categories.
3. High School Grad.	.32	
4) Some High School	.13	
5) Finished 8^{th} grade	.17	
6) Did not finish 8^{th}	.03	

Has the distribution changed in 10 years?

H_0 : No change in distribution.

Take a random sample of $N = 200$.

[2]

Sample and Expected Results

	Observed	Expected
1	35	36
2	40	34
3	83	64
4	16	26
5	26	34
6	0	6
	200	200

Expected Frequencies $(E_i) = N P_i$

How close is the empirical dist. to the theoretical dist.?

Compute: $\sum_i \frac{(O_i - E_i)^2}{E_i} = 18.30$

Squared difference, weighted inversely by expected frequencies.

9-1

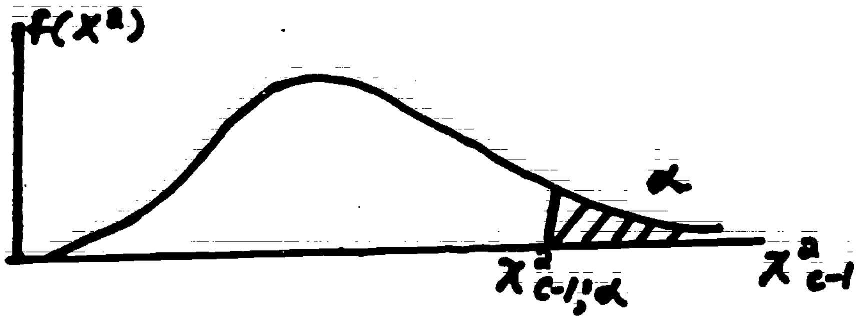
1090

[3]

The Quantity

$$\chi^2 = \sum_{i=1}^{c} \frac{(E_i - O_i)^2}{E_i}$$

for very large N, when H_0 is true, is distributed as an χ^2 random variable with $c-1$ degrees of freedom.



If $\chi^2 > \chi^2_{c-1;\alpha}$, reject H_0 with α type I error

"Goodness-of-fit" test to a Theoretical Distribution.

Lecture 9-2. 2x2 Contingency Tables

2x2 Contingency Tables: Examining Interactions

Lecture Content:

1. 2x2 array of counts
2. Measuring association
3. Log-linear models
4. Testing for independence

Main Topics:

1. Cross-Product Ratio for 2x2 tables
2. Log-linear model and presence of interaction

(There are no transparencies for this lecture)

Topic 1. Cross-Product Ratio for 2x2 tables

I. Basic Issue: Structure of Data

1. Variables A and B, both at levels 1 and 2

2. Two dimensional array of counts

		B: 1	B: 2	
A	1	x_{11}	x_{12}	x_{1+}
	2	x_{21}	x_{22}	x_{2+}
		x_{+1}	x_{+2}	$x_{++} = N$

3. x_{ij} are positive integers

4. We can convert the x_{ij} into probabilities

$$p_{ij} = x_{ij}/N$$

5. If we know N, x_{1+}, and x_{+1}, specifying any cell in the table allows us to fill in the other 3 cells

6. Hence, table itself has only 1 degree of freedom after specifying N, row margin, and column margin

II. Problem: How do data exhibit interaction?

1. If variables A and B are independent, then $x_{ij} = x_{i+}x_{+j}$ product of the marginal distributions

2. As Variables A and B exhibit more and more non-zero interaction, then x_{ij} differs more and more from $\frac{1}{N} x_{i+}x_{+j}$

3. How do we best measure the interaction present between A and B

III. Solution: Cross-Product Ratio

1. Natural "measure of association"

$$\alpha = \frac{x_{11}x_{22}}{x_{12}x_{21}}$$

2. **Properties of α**

   a. If A and B are independent, $\alpha=1$

   b. α is invariant under the simultaneous interchange of rows and columns

   c. α is invariant under row and column multiplications (not true for X^2)

3. α is also called the odds ratio

$$\alpha = \frac{p_{11}/p_{12}}{p_{21}/p_{22}}$$

   a. p_{11}/p_{12} = odds on being in the first level of B, given that you are in the first level of A

   b. p_{21}/p_{22} = odds on being in the first level of B, given that you are in the second level of B

Topic 2. Log-Linear model and presence of interaction

I. Basic Issue: Null model for data structure

1. If A and B are independent,
   $\log p_{ij} = U + U_{1(i)} + U_{2(j)}$; $i=1,2$, $j=1,2$
   a. $U = \frac{1}{4} \sum_{i,j} \log p_{ij}$
   b. $U + U_{1(i)} = \frac{1}{2} (\log p_{i1} + \log p_{i2})$, $i=1,2$.
   c. $U + U_{2(j)} = \frac{1}{2} (\log p_{1j} + \log p_{2j})$, $j=1,2$.
2. If A and B exhibit interaction, then
   $\log p_{ij} = U + U_{1(i)} + U_{2(j)} + U_{12(ij)}$
   a. $U_{12(11)} = -U_{12(12)} = -U_{12(21)} = U_{12(22)}$
   b. $U_{12(ij)}$ are interaction terms

II. Problem: How do we estimate the parameters, and determine whether $U_{12(ij)}$ is nonzero

1. Let $\ell_{ij} = \log p_{ij}$
2. Then
   a. $U = \frac{1}{4} \sum_{i,j} \ell_{ij}$
   b. $U_{1(i)} = \frac{1}{2} \sum_{j} \ell_{ij} - \frac{1}{4} \sum_{i,j} \ell_{ij}$
   c. $U_{2(j)} + \frac{1}{2} \sum_{i} \ell_{ij} - \frac{1}{4} \sum_{i,j} \ell_{ij}$
   d. $U_{12(ij)} = \ell_{ij} - \frac{1}{2} \sum_{j} \ell_{ij} - \frac{1}{2} \sum_{i} \ell_{ij} + \frac{1}{4} \sum_{i,j} \ell_{ij}$

III. Solution: Testing for independence or whether $U_{12(ij)} = 0$ for all i,j.

1. $X^2 = \sum_{i,j} \left[(x_{ij} - E(x_{ij}))^2 / E(x_{ij}) \right]$

   where $E(x_{ij}) = Np_{i+} p_{+j}$

   $= e^{U + U_{1(i)} + U_{2(j)}}$

2. **Definition: Goodman's Measure (likelihood ration statistic)**

$$G^2 = -2\,\Sigma_{ij}x_{ij}\ln(E(x_{ij})/x_{ij}) = 2\,\Sigma_{ij}x_{ij}\ln(x_{ij}/E(x_{ij}))$$

3. **Both distributed under**

H_0: $U_{12(ij)} = 0$, all i,j

as χ^2 random variables, 1 df

Lecture 9-3. Two dimensional contingency tables

Fitting Models to Two Dimensional Contingency Tables

Lecture Content:

1. Structure of Two Dimensional Contingency Tables
2. Log-linear models for two dimensional tables
3. Independence of the Variables

Main Topics:

1. Log-linear models
2. Testing the fit of the model

(There are no transparencies for this lecture)

Topic 1. Log-linear models

I. Basic Issue: Structure of the data

1. Variable A_1: I categories

2. Variable A_2: J categories

		A_2: 1	2	. . .	J	row margins
A_1	1	x_{11}	x_{12}	. . .	x_{1J}	x_{1+}
	2	x_{21}	x_{22}	. . .	x_{2J}	x_{2+}
	⋮	⋮				
	I	x_{I1}	x_{I2}	. . .	x_{IJ}	x_{I+}
column margins		x_{+1}	x_{+2}	. . .	x_{+J}	$x_{++} = N$ sample size

Table of observed frequencies

3. Let $p_{ij} = x_{ij}/N = P\{\text{observation falls in cell } (i,j)\}$

4. $m_{ij} = E(x_{ij})$ = Expected number of observations in (i,j)

$$\sum_{i,j} m_{ij} = N$$

5. $\ell_{ij} = \ln m_{ij}$

II. Method: Log-Linear model

1. Model:

$$\ell_{ij} = U + U_{1(i)} + U_{2(j)} + U_{12(ij)}$$

Saturated model

2. Note that model is for $\log_e m_{ij}$ not $\log_e (m_{ij}/N)$; however, they differ only by the U term

3. Using ANOVA/mean polish analogy, we define:

   a. Overall mean

$$U = \frac{1}{IJ} \sum_{i,j} \ell_{ij} = \frac{1}{IJ} \ell_{++}$$

1098

b. Main effect for variable A_1:

$$U_{1(i)} = \frac{1}{J} \sum_{j} \ell_{ij} - \frac{1}{IJ} \sum_{i,j} \ell_{ij}$$

$$= \frac{1}{J} \ell_{i+} - \frac{1}{IJ} \ell_{++};\ i = 1, 2, \ldots, I.$$

c. Main effect for variable A_2:

$$U_{2(j)} = \frac{1}{I} \sum_{i} \ell_{ij} - \frac{1}{IJ} \sum_{i,j} \ell_{ij}$$

$$= \frac{1}{I} \ell_{+j} - \frac{1}{IJ} \ell_{++};\ j=1,2,\ldots,J.$$

d. Two factors effect (interaction) between variables (may be zero):

$$U_{12(ij)} = \ell_{ij} - \frac{1}{J} \ell_{i+} - \frac{1}{I} \ell_{+j} + \frac{1}{IJ} \ell_{++}$$

4. Evaluating degrees of freedom

<u>U term</u>		
U	1	
U_1	I-1	(1 constraint $\Sigma U_{1(i)} = 0$)
U_2	J-1	(1 constraint $\Sigma U_{2(j)} = 0$)
U_{12}	(I-1)(J-1)	(I-1+J-1+1 constraints)

5. Other issues

a. We can also define cross-product ratios and express the U-terms as functions of them

b. Can also consider the effect of combining categories; for example

A.

Sex \ Age	10-30	30-50	50+
M			
F			

B. Becomes

Sex \ Age	10-30	30+
M		
F		

Topic 2. Testing the Fit of the Model

I. Basic Issue: Are A_1 and A_2 independent?

1. We fit the Model containing only the 2 one-dimensional margins

$$\ell_{ij} = U + U_{1(i)} + U_{2(j)}$$

2. This implies that

$$m_{ij} = \frac{1}{N} x_{i+} x_{+j}$$

3. We call this model 1/2, and compute

$$G^2 = 2 \sum_i \sum_j x_{ij} \log \left(\frac{x_{ij}}{m_{ij}}\right)$$

4. $G^2 \sim \chi^2_{(I-1)(J-1)}$

   to test H_0: A_1 & A_2 are independent

II. Secondary Issue: Evaluation of Fit itself

1. Compute Freeman - Tukey deviates

$$Z_{ij} = \sqrt{x_{ij}} + \sqrt{x_{ij}+1} - \sqrt{4m_{ij}+1} \simeq 2\sqrt{x_{ij}} - 2\sqrt{m_{ij}}$$

2. $Z_{ij} \sim N(0,1)$

3. Stem-and-Leaf display of the deviates should be Gaussian in shape; any Z_{ij} greater than 2 in absolute value is suspect.

Lecture 9-4. Three Dimensional Contingency Tables

Fitting Models to Three Dimensional Contingency Tables

Lecture Content:

1. Structure of Three Dimensional Contingency Tables
2. Log-linear Models for Three Dimensional Tables

Main Topics:

1. Log-linear models
2. Finding the "best" model

(There are no transparencies for this lecture)

1101

Topic 1. Log-linear models

I. **Basic Issue: Structure of the data** (1

1. Variable A_1: I categories
2. Variable A_2: J categories
3. Variable A_3: K categories

variable A_1

$A_1 = 1$ (rows: variable A_2; columns: variable A_3)

	1	2	...	K
1	x_{111}	x_{112}	. . .	x_{11K}
2	x_{121}	x_{122}	. . .	x_{12K}
.	.	.		.
.	.	.		.
.	.	.		.
J	x_{1J1}	x_{1J2}	. . .	x_{1JK}

$A_1 = 2$ (rows: variable A_2 = 1, 2, ..., J)

Entries are x_{2jk}

.
.
.

$A_1 = I$ (rows: variable A_2 = 1, 2, ..., J)

Entries are x_{Ijk}

II. Method: Log-linear model

1. Saturated Model:

$$\ell_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)} + U_{13(ik)} + U_{23(jk)} + U_{123(ijk)}$$

2. Constraints

   a. $\sum_i U_{1(i)} = \sum_j U_{2(j)} = \sum_k U_{3(k)} = 0$

   b. $\sum_i U_{12(ij)} = \sum_i U_{13(ik)} = \sum_j U_{23(jk)} = 0$

   c. $\sum_j U_{12(ij)} = \sum_k U_{13(ik)} = \sum_k U_{23(jk)} = 0$

   d. $\sum_i U_{123(ijk)} = \sum_j U_{123(ijk)} = \sum_k U_{123(ijk)} = 0$

3. We rarely compute these U-terms. We merely calculate G^2 to find best fitting model

4. Evaluating degrees of freedom (2)

U term	df	
U	1	
U_1	I-1	(1 constraint)
U_2	J-1	(1 constraint)
U_3	K-1	(1 constraint)
U_{12}	(I-1)(J-1)	(I+J-1 constraints)
U_{13}	(I-1)(K-1)	(I+K-1 constraints)
U_{23}	(J-1)(K-1)	(J+K-1 constraints)
U_{123}	(I-1)(J-1)(K-1)	(IJ+IK+JK-I-J-K+1 constraints)

5. To find the correct df for a G^2 of one of the <u>8</u> possible models, we merely subtract from IJK the degrees of freedom for every term in the model

1103

Topic 2. Finding the "Best" Model

I. Basic Issue: Descriptions of the Models

1. There are 8 relevant log-linear models, of 4 different types

2. The models, by type are

   a. Complete Independence, Model 1/2/3.

   $$\log m_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)}$$

   all interactions are zero

   b. Single association models.

   Models 12/3, 13/2, 23/1 all but one 2 factor interaction is zero.

   i. 12/3: $\log m_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)}$

   ii. 13/2: $\log m_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{13(ik)}$

   iii. 23/1: $\log m_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{23(jk)}$

   c. Conditional independence models.

   Models 12/13, 12/23, 13/23
   Conditional on the level of the variable included in the two interactions, the other two variables are independent

   i. 12/13: $\log m_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)} + U_{13(ik)}$

   ii. 12/23: $\log m_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)} + U_{23(jk)}$

   iii. 13/23: $\log m_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{13(ik)} + U_{23(jk)}$

1104

d. No three factor interaction model

Model 12/13/23

$$\log m_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)} + U_{13(ik)} + U_{23(jk)}$$

II. Problem: How do we compute cell estimates

1. All but the 12/13/23 model have "closed form" cell expected values

2. The cell estimates for 12/13/23 model must be found by Iterative Proportional Fitting (3)

3. Cell estimates, m_{ijk} are

   a. 1/2/3: $m_{ijk} = \frac{1}{N^2} x_{i++} x_{+j+} x_{++k}$

   b. 12/3 : $m_{ijk} = \frac{1}{N} x_{ij+} x_{++k}$

   c. 13/2 : $m_{ijk} = \frac{1}{N} x_{i+k} x_{+j+}$

   d. 23/1 : $m_{ijk} = \frac{1}{N} x_{+jk} x_{i++}$

   e. 12/13: $m_{ijk} = \frac{1}{x_{i++}} x_{ij+} x_{i+k}$

   f. 12/23: $m_{ijk} = \frac{1}{x_{+j+}} x_{ij+} x_{+jk}$

   g. 23/13: $m_{ijk} = \frac{1}{x_{++k}} x_{i+k} x_{+jk}$

   h. 12/13/23: m_{ijk} found by iterative proportional fitting

4. Main task is to determine which model fits

III. Solution: Hypothesis tests for each model

1. For each of the 8 models, we have a null hypothesis that the model is an accurate description of the data

1105

2. We compute a G^2 for each model, and determine whether

   $G^2 > \chi^2_{df;\alpha}$ where df is as follows

   a. 1/2/3: G^2 has $IJK-(I+J+K)+2$ df

   b. 12/3: G^2 has $(IJ-1)(K-1)$ df

   c. 13/2: G^2 has $(IK-1)(J-1)$ df

   d. 23/1: G^2 has $(JK-1)(I-1)$ df

   e. 12/13: G^2 has $(J-1)(K-1)I$ df

   f. 12/23: G^2 has $(I-1)(K-1)J$ df

   g. 23/13: G^2 has $(I-1)(J-1)K$ df

   h. 12/13/23: G^2 has $(I-1)(J-1)(K-1)$ df

3. Strive for simplicity: if 2 models fit, choose the less saturated of the two

4. Calculate Freeman-Tukey deviates for the best fitting model and examine them

5. Rearrange table to emphasize fit

6. Examine relevant 2 dimensional margins

(Thoroughly discuss example) (4)-(9)

Lecture 9-4

Transparency Presentation Guide

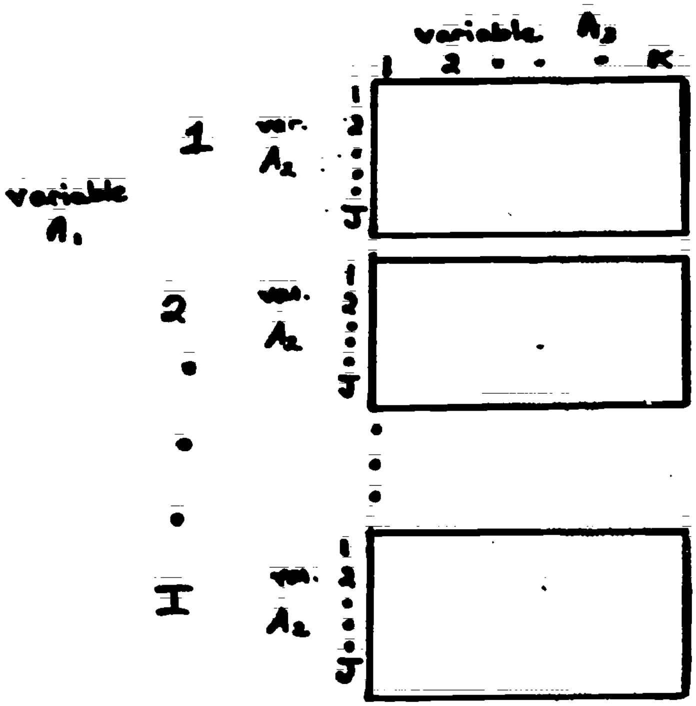
Lecture Outline Location	Transparency Number	Transparency Description
Topic 1		
Section I.		
1.	1	Log-linear model and data structure
Section II.		
4.	2	Degrees of freedom and model types
Topic 2		
Section II.		
3.	3	Cell estimates
Section III.		
6.	4	NBER-Thorndike, Hagen Study
6.	5	NBER-T,H data
6.	6	NBER-T,H Two Dimensional Margins
6.	7	Model fitting
6.	8	NBER-T,H Data Rearranged
6.	9	NBER-T,H Freeman Tukey Residuals

1107

[1]

Three Dimensional Contingency Tables

table: variables A_1, A_2, A_3 at levels I, J, K



entries are X_{ijk}

Fully saturated log linear model

$$\log M_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)} + U_{13(ik)} + U_{23(jk)} + U_{123(ijk)}$$

$$\sum_i U_{1(i)} = \sum_j U_{2(j)} = \sum_k U_{3(k)} = 0$$

$$\sum U_{12}(ij) = \sum U_{12}(ij) = \sum U_{13}(ik) = \sum U_{13}(ik)$$

$$\sum_i U_{12}(ij) = \sum_j U_{12}(ij) = \sum_i U_{13}(ik) = \sum_k U_{13}(ik) = \sum_j U_{23}(jk) = \sum_k U_{23}(jk) = 0$$

$$\sum_i U_{123}(ijk) = \sum_j U_{123}(ijk) = \sum_k U_{123}(ijk) = 0$$

We very rarely compute these U-terms . . . merely calculate G^2 to find best fitting model. 9-4

Degrees of Freedom

[2]

U-term	df
U	1
U_1	$I-1$
U_2	$J-1$
U_3	$K-1$
U_{12}	$(I-1)(J-1)$
U_{13}	$(I-1)(K-1)$
U_{23}	$(J-1)(K-1)$
U_{123}	$(I-1)(J-1)(K-1)$

To find df for G^2: subtract from IJK the degrees of freedom for every term in model.

e.g. 1/2/3 G^2 has $IJK-(I-1)-(J-1)-(K-1)-1 = IJK-(I+J+K)+2$ df.

There are 8 log-linears models of 4 different types.

Type A. 1/2/3 complete independence
all interactions are zero
$$\log \hat{M}_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)}$$

B. 12/3, 13/2, 23/1, Single association models.
all but one 2 factor interaction is zero.

12/3 $\log M_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)}$

13/2 $\log M_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{13(ik)}$

23/1 $\log M_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{23(jk)}$

C. 12/23, 13/23, 12/13 Conditional independence models
conditional on level of one variable, remaining two independent.

12/23 $\log M_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)} + U_{23(jk)}$

13/23 $\log M_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{13(ik)} + U_{23(jk)}$

12/13 $\log M_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)} + U_{13(ik)}$

9-4

D. 12/13/23 No three factor interaction model

$\log M_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)} + U_{13(ik)} + U_{23(jk)}$

Model	Degrees of Freedom	Cell estimates, $\hat{M}_{ijk}$ =
1/2/3	$IJK-(I+J+K)+2$	$\frac{1}{N^2} X_{i++} X_{+j+} X_{++k}$
12/3	$(IJ-1)(K-1)$	$\frac{1}{N} X_{ij+} X_{++k}$
13/2	$(IK-1)(J-1)$	$\frac{1}{N} X_{i+k} X_{+j+}$
23/1	$(JK-1)(I-1)$	$\frac{1}{N} X_{+jk} X_{i++}$
12/13	$(J-1)(K-1)I$	$\frac{1}{X_{i++}} X_{ij+} X_{i+k}$
12/23	$(I-1)(K-1)J$	$\frac{1}{X_{+j+}} X_{ij+} X_{+jk}$
13/23	$(I-1)(J-1)K$	$\frac{1}{X_{++k}} X_{i+k} X_{+jk}$
12/13/23	$(I-1)(J-1)(K-1)$	Found by Iterative Prop Fitting

1. We seek G^2 values slightly greater than the corresponding degrees of freedom.

2. Strive for simplicity -- if 2 models fit, choose the less saturated of the 2 if possible.

Calculate Freeman-Tukey deviates for the best fitting model, and stem-and-leaf them to find large deviations.
Rearrange table to emphasize fit - Examine relevant 2 dim. margins.

[4]

NBER - Thorndike, Hagen Study

Subjects all took a series of U.S. Air Force aptitude tests in 1943. Relatively homogenous in age, all had high school education or the equivalent and had been accepted by the Air Force for Aircrew Training Program.

Aptitude data gathered in 1943, Education and Occupation data obtained in following studies in 1955 and 1969.

Occupation classes — Variable 3

- O1 = Self-employed (business)
- O2 = Self-employed (professional)
- O3 = teacher
- O4 = salary (employed)

Education classes — Variable 1

- E1 = High School
- E2 = Some College
- E3 = College
- E4 = College +

Aptitude Classes — Variable 2

- A1 Lowest
- ⋮
- A5 Highest

9-4

[5]

O1 *Self Employed Business*

	E1	E2	E3	E4	
A1	42	55	22	3	122
A2	72	82	60	12	226
A3	90	106	85	25	306
A4	27	48	47	8	130
A5	8	18	19	5	50
	239	309	233	53	834

O2 *Self Employed Professional*

	E1	E2	E3	E4	
A1	1	2	8	19	30
A2	1	2	15	33	51
A3	2	5	25	83	115
A4	2	2	10	45	59
A5	0	0	12	19	31
	6	11	70	199	286

O3 *Teacher*

	E1	E2	E3	E4	
A1	0	0	1	19	20
A2	0	3	3	60	66
A3	1	4	5	86	96
A4	0	0	2	36	38
A5	0	0	1	14	15
	1	7	12	215	235

O4 *Salary, Employed*

	E1	E2	E3	E4	
A1	172	151	107	42	472
A2	208	198	206	92	704
A3	279	271	231	191	1072
A4	99	126	179	97	501
A5	36	35	99	79	249
	794	781	922	501	2998

9-4

[6]

Two - Dimensional Margins

	E1		E2		E3		E4		
A1	215	33%	208	32%	138	21%	83	13%	644
A2	281	27%	282	27%	284	27%	197	19%	1044
A3	371	23%	382	24%	446	28%	385	24%	1584
A4	128	18%	176	24%	238	33%	186	26%	728
A5	44	13%	53	15%	131	38%	117	34%	345

	O1		O2		O3		O4		
E1	239	23%	6	1%	0	0%	794	76%	1039
E2	309	28%	11	1%	0	0%	781	71%	1101
E3	233	19%	70	6%	12	1%	922	75%	1237
E4	53	5%	199	21%	215	22%	501	52%	968

	O1		O2		O3		O4		
A1	122	19%	30	5%	20	3%	472	73%	644
A2	226	22%	51	5%	63	6%	704	67%	1044
A3	306	19%	115	7%	91	6%	1072	68%	1584
A4	130	18%	59	8%	38	[illegible]	[illegible]	69%	728
A5	50	14%	31	9%	15	[illegible]	[illegible]	72%	345

9-4

1114

[7]

Model	df	χ^2	G^2	$\sum_i (F-T)^2$	
1/2/3	59	1094	1081	1154	
1/23	47	1012	1046	1135	
3/12	47	902	904	974	
12/23	38	860	867	937	
2/13	52	219	219	217	
13/23	40	179	183	184	
12/13	40	41	41	39	**
12/13/23	28	18.1	17.3	15.7	

df = 80 − (structural zeros) − (estimated parameters)

structural zeroes = 10

Model	estimated parameters	
1/2/3	$1+3+4+3$	= 11
1/23	$1+3+4+3+4\cdot3$	= 23
3/12	$1+3+4+3+3\cdot4$	= 23
12/23	$1+3+4+3+4\cdot3+3\cdot4$	= 35
2/13	$1+3+4+3+(3\cdot3-2)$	= 18
13/23	$1+3+4+3+(3\cdot3-2)+3\cdot4$	= 30
12/13	$1+3+4+3+(3\cdot3-2)+4\cdot3$	= 30
2/13/23	$1+3+4+3+(3\cdot3-2)+3\cdot4+4\cdot3$	= 42

[8]

Rearranged data to Reflect Conditional Independence of O and A, given E

E1	A1	A2	A3	A4	A5
01	42	72	90	27	8
02	1	1	2	2	0
03	0 NA	0 NA	1 NA	0 NA	0 NA
04	172	208	279	99	36

E2	A1	A2	A3	A4	A5
01	55	82	106	48	18
02	2	2	5	2	0
03	0 NA	3 NA	4 NA	0 NA	0 NA
04	151	198	291	126	35

E3	A1	A2	A3	A4	A5
01	22	60	85	47	19
02	8	15	25	10	12
03	1	3	5	2	1
04	107	206	331	179	99

E4	A1	A2	A3	A4	A5
01	3	12	25	8	5
02	19	33	83	45	19
03	19	60	86	36	14
04	42	92	191	97	99

9-4

[9]

$$\sqrt{O} + \sqrt{O+1} - \sqrt{4E+1}$$

Freeman-Tukey Residuals for Model 12/13

E1	A1	A2	A3	A4	A5
01	-1.06	.92	.52	-.41	-.61
02	-.03	-.32	.05	1.16	-.42
03	—	—	—	—	—
04	.61	-.45	-.25	.14	.44

E2	A1	A2	A3	A4	A5
01	-.41	.35	-.09	-.16	.82
02	.10	-.36	.65	.31	-.77
03	—	—	—	—	—
04	.30	-.13	.02	.13	-.39

E3	A1	A2	A3	A4	A5
01	-.76	.89	.14	.36	-1.15
02	.15	-.21	0	-.93	1.53
03	-.11	.27	.41	-.05	-.05
04	.43	-.38	-.07	.14	.16

E4	A1	A2	A3	A4	A5
01	-.65	.43	.86	-.63	-.47
02	.51	-1.19	.45	1.08	-1.03
03	.19	2.29	.08	-.81	-2.63
04	-.11	-.98	-.57	.10	2.29

9-4

Homework, Unit 9

1. The number of animal bites reported in three successive weeks in 1967 to the Chicago Board of Health were as follows:

Week	1	2	3
Number of Bites	268	189	199

   a. Test the null hypothesis that the weeks are identical.

   b. Explain why the uull hypothesis in part(a) was or was not rejected. How could differences or similarities in the three weeks produce this finding?

2. We have taken a random sample of 148 retarded children and recorded their IQ score (dichotomous: 55-69 and 40-54) and season of birth.

	Birth			
IQ	Summer	Autumn	Winter	Spring
55-69	29	19	12	18
40-54	13	17	20	20

   In this sample, are IQ and Birth Season independent? Why or why not?

3. Consider the incidence of leukemia among survivors of the atomic bombings of Hiroshima and Nagasaki. These cases were recorded from 1950-1958.

Dose in rads	% Population Exposed	Cases
>81	11.03	34
21-80	13.41	5
<20	75.56	12
		51

   a. If the number of reported cases were independent of the amount of radiation, what would be the expected number of cases for each of the three dosage categories?

   b. Test the null hypothesis that leukemia incidence is independent of the amount of radiation exposure.

1118

4. The data set to be analyzed for this problem concerns sex bias in Graduate admissions at Berkeley. The article was published in Science, volume 187, page 398, and is in the Fairley and Mosteller collection.

After careful perusal of the article, you should feel that the authors have not "done justice" to this table. What we need is a log-linear model, and, fortunately, the raw data is given to you on the next 2 pages.

Your assignment is to find the best fitting model for this 2x2x100 table, and to interpret it. Also determine whether there is a 3 factor interaction.

I suggest you fit all the possible models... there are 8 of them.

QMPM

Berkeley Graduate Admissions Data

Fall Quarter 1973

dept.	men admit	men deny	women admit	women deny
7	20	21	6	10
10	87	54	4	3
14	91	207	48	68
16	6	19	4	8
38	11	3	1	1
42	138	279	133	242
51	8	9	4	3
67	12	3	3	3
69	1	0	0	0
81	5	0	2	0
85	39	19	5	1
95	0	2	0	0
126	1	0	0	0
128	10	5	0	1
146	4	6	9	8
168	352	208	17	8
179	6	6	0	11
213	53	138	93	300
215	47	12	12	2
225	28	95	5	8
240	47	188	21	74
248	1	0	3	0
254	196	135	4	6
270	18	3	6	7
279	5	14	7	9
283	31	7	4	0
316	9	1	9	2
319	24	367	25	334
328	21	47	12	23
335	93	169	13	19
342	1	2	0	2
347	32	61	10	27
353	14	0	17	3
369	22	14	5	1
371	7	7	1	1
386	5	6	7	12
390	1	1	3	2
395	9	3	5	3
400	16	21	1	1
405	4	3	10	4
421	162	36	3	1
429	12	40	14	63
440	12	46	3	40

dept.	men admit	men deny	women admit	women deny
442	2	9	5	6
448	12	60	13	24
471	21	16	14	5
477	4	2	11	5
497	1	1	0	0
518	14	35	2	2
519	16	7	7	2
522	0	1	2	1
539	1	9	3	3
542	6	4	0	0
548	174	131	45	10
552	7	6	15	21
560	0	0	1	0
571	15	18	7	20
573	0	1	0	0
623	3	2	5	0
634	5	1	1	0
644	36	32	33	48
649	2	0	1	0
650	6	4	5	3
662	69	206	16	49
681	97	263	40	42
701	34	21	8	4
709	1	1	2	0
716	32	6	1	0
737	1	3	C	0
743	58	73	116	276
752	2	0	0	0
776	72	80	9	3
780	271	86	11	1
789	115	118	26	26
795	24	60	12	15
796	32	16	8	0
812	13	105	20	115
813	48	10	0	0
822	1	2	3	1
841	2	6	14	6
848	44	41	8	13
860	43	183	46	151
862	1	6	4	25
866	25	100	22	50
873	17	27	1	29
877	3	12	1	0

dept.	men admit	men deny	women admit	women deny
884	509	316	89	19
886	8	18	8	6
895	5	6	2	0
901	120	205	204	389
907	3	5	0	0
923	0	12	2	12
932	32	22	58	22

dept.	men admit	men deny	women admit	women deny
933	1	1	1	0
942	0	4	0	0
957	11	8	8	6
965	12	7	8	2
984	1	0	0	0
986	14	60	13	40
995	14	16	33	64

1120

Homework Unit 9
Solutions

1. a) Animal Bites reported in 3 successive weeks in 1967 to Chicago Board of Health

Week	1	2	3	Total
Bites	268	189	199	656

H_0: Weeks are identical $\qquad$ H_1: Weeks are not similar

E_i	218.67	218.67	218.67
O_i	268	189	199
$O_i - E_i$	49.33	-29.67	-19.67

$$X^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} =$$

$$(49.33)^2/218.67 + (29.67)^2/218.67 + (19.67)^2/218.67$$

$$= 16.9$$

$X^2_{2;.05} = 5.99 \qquad X^2_{2;.01} = 9.21$

Hence we reject H_0; the weeks are not identical.

b) We reject the null hypothesis because the probability that the data results in such an extreme (large) value of X^2 under H_0 is (much) less than .01. The differences among the three weeks could be due to weather (more bites in warm or sunny weather than cool or rainy weather), the lunar cycle, or, given data for many other weeks. "Week 1" may just be an outlier, or the high point of an "animal bite cycle". (It is difficult to make inferences about the reasonableness of one data value given a sample of only 3).

2. Season of Birth and IQ scores

H_0: Season of Birth and IQ Scores are independent.

IQ	Summer	Autumn	Winter	Spring	total (x_{i+})
55-69	29	19	12	18	78
40-54	13	17	20	20	70
total (x_{+j})	42	36	32	38	148 (x_{++})

$$\text{Expected Values} = \frac{x_{i+}x_{+j}}{x_{++}}$$

IQ	Summer	Autumn	Winter	Spring
55-69	22.1	19.0	16.9	20.0
40-54	19.9	17.0	15.1	18.0

$X^2 = 7.98$

$\chi^2_{3;.05} = 7.82$ $\chi^2_{3;.01} = 11.34$

The observed value of X^2 is just significant at the .05 level, hence we reject H_0 at the 5% level. Note that we <u>accept</u> H_0 (independence) at the 1% level.

3. Incidence of Leukemia among survivors of the atomic bombings of Hiroshima & Nagasaki (1950-1958)

Dose in Rads	% Population Exposed	Cases
81+	11.03	34
21-80	13.41	5
0-20	75.56	12
		51

H_0: Leukemia incidence independent of amount of exposure

Under H_0, 11.03% of the reported cases would be in 81+ rad group, 13.41% of the cases in 21-80 group, and 75.56% in 0-20 group.

E_i	$.1103 \times 51 = 5.63$	$.1341 \times 51 = 6.84$	$.7556 \times 51 = 38.54$
O_i	34	5	12

$\chi^2 = 161.73$ huge

$\chi^2_{2;.05} = 5.99$ $\chi^2_{2;.01} = 9.21$

Hence we reject H_0; the incidence of leukemia is <u>not</u> independent of exposure.

1123

4. The saturated model 123 is

$$l_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

where: i (= 1, 2, ..., 100) refers to department,

j (= 1,2) refers to sex,

k (= 1,2) refers to admit/deny.

We fit all 8 models

1/2/3
12/3
13/2
23/1
12/13
12/23
13/23
12/13/23

a) (1/2/3) complete independence

Under this model there is no association between any pair of variables, nor among all three together.

Department, Sex, and Admission decision are all independent.

The observed X^2 statistic was 5703

The observed G^2 statistic was 5688

The model has (IJK - I - J - K + 2) = 298 degrees of freedom. $\chi^2_{.05} \doteq \frac{1}{2}(\sqrt{595} + 1.65)^2 = 399.11$ Hence we reject this model.

1124

b) (1,2/3)

Under this model admission is independent of sex and department. Sex and department however are associated indicating a tendency for each sex to be more strongly preferred by some departments at the expense of other departments.

The observed X^2 statistic was 2262

The observed G^2 statistic was 2420

The model has (IJ-1)(K-1) = 199 degrees of freedom

$X^2_{.05} \doteq \frac{1}{2}(\sqrt{397} + 1.65)^2 = 232.74$

Hence we reject this model.

c) (1,3/2)

Under this model sex is independent of admission and department. Admission and department are however associated, indicating a tendency for some departments to be easier to enter than expected, while other departments are harder to enter than expected.

The observed X^2 statistic was 3118

The observed G^2 statistic was 3428

The model has (IK-1)(J-1) = 199 degrees of freedom

$X^2_{.05} \doteq \frac{1}{2}(\sqrt{397} + 1.65)^2 = 232.74$

Hence we reject this model.

d) (2,3/1)

Under this model department is independent of admission and sex. Admission and sex are associated, indicating an overall pattern of sex discrimination. Closer examination will suggest whether this discrimination favors males or females; the <u>mere presence</u> of the u_{23} term does <u>not</u> indicate which sex is favored.

The observed X^2 statistic was 5407

The observed G^2 statistic was 5585

The model has (JK-1)(I-1) = 297 degrees of freedom

$X^2_{.05} \doteq \frac{1}{2}(\sqrt{593} + 1.65)^2 = 338.04$

Hence we reject this model

e) (1,2/2,3) (conditional independence of variables 1 and 3)

Under this model department and sex, and admission and sex are conditionally independent (for a given sex, department and admission are independent). Hence for males (and also for females), individuals have equal chances for admission to each department.

The observed X^2 statistic was 2173

The observed G^2 statistic was 2316

The model has (J)(I-1)(K-1) = 198 degrees of freedom

$X^2_{.05} \doteq \frac{1}{2}(\sqrt{393} + 1.65)^2 = 231.65$

Hence we reject this model.

f) (1,3/2,3) (conditional independence of variables 1 and 2)

Under this model department and admission, and admission and sex are associated. Department and sex are conditionally independent (for a given admission decision, department and sex are independent).

The observed X^2 statistic was 3044

The observed G^2 statistic was 3324

The model has (K)(I-1)(J-1) = 198 degrees of freedom

$X^2_{.05} \doteq \frac{1}{2}(\sqrt{393} + 1.65)^2 = 231.65$

Hence we reject this model

g) (1,2/2,3/1,3) (No 3 factor interaction)

Under this model each pair of variables is associated, but together independent of the third.

The observed X^2 statistic was 151

The observed G^2 statistic was 155

The model has (I-1)(J-1)(K-1) = 99 degrees of freedom

$X^2_{.05} \doteq \frac{1}{2}(\sqrt{197} + 1.65)^2 = 123.02$

Hence we reject this model.

h) (1,2/1,3) (conditional independence of variables 2 and 3)

Under this model department and sex, and department and admission are associated. Admission and sex are conditionally independent (for a given department, admission and sex are independent). There is a tendency for some departments to attract more male (or female) applicants than otherwise expected and a tendency for some departments to be harder to enter than otherwise expected. However, the decision in each department, to admit or deny an applicant is independent of the applicant's sex. This is an extremely important conclusion.

The observed X^2 statistic was 156

The observed G^2 statistic was 159

The model has (I)(J-1)(K-1) = 100 degrees of freedom

$\chi^2_{.05} \doteq \frac{1}{2}(\sqrt{199} + 1.65)^2 = 124.14$

$\chi^2_{.01} \doteq 135.81$

$\chi^2_{.005} \doteq 140.17$

Hence we reject this model

However, this model fits better than the other 7 models, with the possible exception of 12/13/23. This model is preferred because it is more parsimonious than 12/13/23.

Conclusion:

Of all the models, the "best fit" was achieved by model (1,2/1,3), (h), conditional independence of variables 2 and 3. The fit and the Freeman-Tukey residuals for this model should now be computed. We rearrange the data into 100 2x2 small tables. Each has the structure:

		Admit	Deny
Department i	Male		
	Female		

for all departments i = 1,2,..., 100.

Within each of these tables, the sex and admit/deny variables are independent;

$$\text{i.e.} \qquad \alpha_i = \frac{x_{i11}\ x_{i22}}{x_{i12}\ x_{i22}} \approx 1.$$

We next give the stem-and-leaf and boxplot of the F-T residuals. We suppress displaying the 100x2x2 array of the rearranged data and residuals because of lack of space.

The stem-and-leaf display of the residuals is symmetric, and indicates that there are very few deviant cells. In fact, we see only 7 cells > 1.96 in absolute value, a number much smaller than the .05(400) = 20 expected by chance. Perhaps the variance of the residuals is smaller than 1.

```
 RES
BLE ' RES '.
 =       0.0100
        LO I      -3.5647    -2.7362    -2.6693    -2.0501    [illegible].6364    -1.5876    -1.4176    -1.2842
        LO I      -1.2542    -1.2460    -1.1913    -1.1800    [illegible]219    -1.1156
      -10  I 641
       -9  I 0
       -8  I 87444210
       -7  I 754100
       -6  I 88655444332200
       -5  I 8733320
       -4  I 964444333
       -3  I 9987766555444442110000
       -2  I 98865443333333321111111000
       -1  I 999987766654443333200000000
       -0  I 87776666655[illegible]4333333221111000
        0  I 0000000000000[illegible]2222222334455555678888999
        1  I 00112233334444[illegible]44556666777788
        2  I 00001123333455556788888899
        3  I 00000011111122233333444555666666667778899
        4  I 11111112334456677789999
        5  I 00111222335566679
        6  I 0013445688999
        7  I 233336789
        8  I 0001388
        9  I 0034489
       10  I 03379
       11  I 2247
       12  I 3
        HI I       1.2798     1.3622     1.4135     1.4280     1.6604     1.8300     2.0604     2.2110
        HI I       2.6020


 RES THREE
SCALE UNIT:        0.1000
                 -2.0000                 0.0000                 2.0000                 4.0000                 6.0000
  ---------------+-----------------------+-----------------------+-----------------------+-----------------------+----
ABLE: RES
            *          *  ***        +--+--+          *
  *              *      * *<--------I  +  I-------->**   **   **   *
            *          *  ***        +--+--+          *
```

Quiz, Unit 9

Name ______________________________

Write all your answers on these pages. Point totals are given in parentheses prior to each question. You have sixty (60) minutes for this quiz. Good luck!

(40) 1. You are interested in the relationship between family income and the number of children per family in a small township of 1000 families.

Unfortunately, you do not have available the family income and number of children for each family. Your research assistant has been able to gather from each family only answers to the two questions:

Is your family income (annual) greater than $8000 or less than $8000?

Does your family have 3 or more children, or 2 or fewer children?

a. Explain briefly to your township supervisor what you have implied by the statement:

"Based on this two dimensional contingency table, the number of children per family is independent of the annual family income."

1131

b. Your research assistant has misplaced the actual cell counts for the table, but has managed to keep the one-dimensional margins of the table in his head.

The margins are

		family income		
		$<$ \$8000	$\geq$ \$8000	
#	$<$ 3			550
children	$\geq$ 3			450
		300	700	1000

Amaze your research assistant by computing a two dimensional table from these one dimensional margins that exhibits no interaction.

c. Fortunately, your summer intern has filed away the actual cell counts. She says that the observed frequencies are:

		family income	
		$<$ \$8000	$\geq$ \$8000
#	$<$ 3	200	350
children	$\geq$ 3	100	350

Please construct and test a hypothesis for your supervisor for no interaction in the observed table. Use $\alpha = .025$. Is there any interaction present?

(50) 2. You are interested in studying what patient characteristics influence the length of stay of hospital patients after surgical procedures have been performed on them.

You have data on 10000 surgical patients from Massachusetts General Hospital in 1976.

The patients are placed into cells of a 4 dimensional contingency table with variables:

Var. 1: Length of Stay, 4 categories:
$\leq$ 1 day, 1 day - 1 week,
1 week - 1 month, > 1 month.

Var. 2: Age, 3 categories:
$\leq$ 30 years, 30-50 years, > 50 years

Var. 3: Sex, 2 categories:
Male, Female

Var. 4: Preoperative status, 4 categories:
1 = excellent, 2 = good, 3 = fair, 4 = poor

You use the stepwise procedure to fit loglinear models to this 4x3x2x4 table.

The first stage of the fitting process yields the following results:

Model	df	G^2
1/2/3/4	86	375.4
12/13/14/23/24/34	57	41.6
123/124/134/234	18	9.3

a. What can you conclude from these results with regards to choosing the "best-fitting" log linear model?

1133

b. You find the following test statistics for each of the 2 factor interactions:

Interaction	Conditional G^2	df
12	193.1	
13	4.6	
14	125.3	
23	3.2	
24	5.7	
34	1.9	

Fill in the df column, and determine (roughly) which 2 factor interactions are non-zero.

Note: Conditional G^2 statistics are differences of G^2 statistics for specific models. For example $G^2_{1/2/3} - G^2_{12/3} = G^2_{[12]} = 193.1$

c. Based on these results, write down the "best" loglinear model for this table in terms of the appropriate U-terms, <u>and</u>, in the context of this example, interpret it.

1134

(10) 3. You are interviewing individuals at random in the community to determine preferences for home energy consumption.

You ask each individual which of the following 4 energy alternatives he/she prefers:

Natural Gas
Oil
Coal
Solar Power

The 1000 individuals sampled have the following preferences:

	# individuals
Natural Gas	270
Oil	260
Coal	280
Solar	190
	1000

Test whether individual preferences are uniformly distributed among these 4 alternatives.

1135

Quiz, Unit 9
Solutions

1\. a. Since this answer is directed to a layman of both exploratory data analysis and statistics, perhaps the simplest and most effective response would be to say:

> "Independence implies that the variables do not influence one another in a consistent, measurable way. It is not possible to predict with accuracy the number of children in a family based upon their income."

A more technical response which would assume a background in the subject would be to say that:

> "The contribution of one category of a factor does not help define the contribution of any category of the other factor. In other words, the probability of a given observation falling in particular cell of the table is equal to the product of the marginal probabilities."

The form of the table of raw data can be described by the form:

		children ≥ 2	children ≤ 3
Income	< 8,000	$\frac{a}{b}x$	$\frac{b}{a}y$
Income	$\geq$ 8,000	ax	by

where x and y represent values (predicted or observed) of the table and a and b are positive constants

b\. The model of independence in the table is given by

$$\log m_{ij} = U + U_1 + U_2$$

where m_{ij} is a predicted cell value and U is the grand mean

U_1 is the additional contribution to the grand mean associated with the first variable

U_2 the additional contribution of the second variable

m_{ij} is also given by the equation $m_{ij} = \frac{x_{i+}x_{+j}}{x_{++}}$

where each x_{i+} is a row sum

each x_{+j} is a column sum

x_{++} is the sum of all cell values (all x_{ij}'s)

The solution (given the marginal sums) is:

# children		<$8,000	>$8,000	
	< 3	$\frac{(550)(300)}{1,000} = 165$	$\frac{(500)(700)}{1,000} = 385$	550
	≥ 3	$\frac{(450)(300)}{1,000} = 135$	$\frac{(450)(700)}{1,000} = 315$	450
		300	700	1,000

c. The "null" hypothesis to be tested is that there is no interaction between family income and the number of children (i.e. $U_{12} = 0$).

To test this hypothesis we use the X^2 formula:

$$X^2 = \Sigma \frac{(\text{Observed value} - \text{Expected Value})^2}{\text{Expected Value}}$$

We calculated the expected values in b above under the assumption of independence (i.e. $U_{12} = 0$). The observed values are given in this question.

The respective $\frac{(\text{Obs.} - \text{Exp.})^2}{\text{Exp.}}$ values for the table are

7.42	3.18
9.07	3.89

$$23.58 = \Sigma \frac{(\text{Obs-Exp})^2}{\text{Exp.}} = X^2$$

1137

Using $\alpha = .05$ for a one-tail test, we find $z = 1.65$.

Since $\sqrt{23.58} = 4.86 > 1.65$ the null hypothesis is rejected.

Our conclusion is that an interaction between the variables probably does exist and that our assumption of independence was probably incorrect.

2. a. By comparing the number of degrees of freedom (d.f.) with the G^2 value for each model we can discover which model fits best.

The model of independence (1/2/3/4) is apparently much too simple; the G^2 is almost four times the number of degrees of freedom.

The model of all two factor interactions displays a G^2 to d.f. ratio of .73 which is closest to 1 of any of the three models. However, as shown by the ratio of less than one, the model is "overfit". A simpler model might be preferred.

The model of three factor interactions is much too complex with a G^2/d.f. ratio of about .52.

The 12/13/14/23/24/34 is therefore the "best-fitting" model of the three given. Partitioning of this model would be advised to get a simpler model which is not overfit.

b. To calculate the degrees of freedom for each interaction we must consider the number of categories for each variable

Variable 1	has	4	categories	Let $I = 4$
2	"	3	"	$J = 3$
3	"	2	"	$K = 2$
4	"	4	"	$L = 4$

Since the question asks for the number of degrees of freedom for each interaction the calculations are straightforward multiplications of the degrees of freedom of the involved variables. Therefore the result is:

Interaction	G^2	d.f.
12	193.1	$(I-1)(J-1) = 6$
13	4.6	$(I-1)(K-1) = 3$
14	125.3	$(I-1)(L-1) = 9$
23	3.2	$(J-1)(K-1) = 2$
24	5.7	$(J-1)(L-1) = 6$
34	1.9	$(K-1)(L-1) = 3$

The question also asks what two factor interactions are non-zero. The sheer magnitude of the G^2 makes it apparent that 12 and 14 are non-zero. They are so overwhelming that inclusion of the others would probably contribute little to the goodness of fit of the model.

c. The best log-linear model is given by:

$$\log m_{ijkl} = U + U_1 + U_2 + U_3 + U_4 + U_{12} + U_{14}$$

Length of stay, age, sex, and preoperative status each contribute in describing particular types of patients and the frequency which they can be expected to be observed. In addition, there is a relationship (an interaction) between length of stay and age, and between length of stay and preoperative status. Although we do not know for sure what these relationships are, we can postulate that perhaps the very young or very old normally require extra care in their treatment and therefore generally stay longer. Similarly, it can be reasoned that the worse the preoperative status of the patient the longer he or she will have to stay in the hospital.

There are no other significant two factor interactions nor are there any three factor relationships which help much in describing the patient population.

3. If individual preferences of energy alternatives are uniformly distributed, each category would have the same number of observations (neglecting sampling error). This means that the expected values for each of the four alternatives is 250 (i.e. $\frac{1000}{4}$). The null hypothesis is H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$.

Using the X^2 formula $X^2 = \Sigma\left(\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}\right)$ we get a value $X^2 = 20$.

With 3 degrees of freedom, at a 95% level of significance, a table value of 7.815 is found. Since 20 > 7.815, the null hypothesis is rejected. We cannot assume that the preferences are uniformly distributed.

1139

Final Examination
Second Term

Name: ______________________________

All answers should be written on this test. Total point score is 100. You need not answer every question. Please read the examination before you begin writing. You have 2.5 hours to complete this examination.

Part I. This section is worth 40 points. Answer all 3 questions (#1, 2 and 3)

(1) You are interested in constructing a linear model relating the number of pharmacists per census tract to other policy relevant features of the tract.

a. You expect to use least squares techniques to estimate your model. Consequently, you suspect that two problems may arise because of these correlated variables.

(a) The residuals will not be independent.

(b) Their coefficient estimates will not be reliable.

(c) The F-test will produce a ~ ult.

(d) Both coefficient estimate ve large t-statistics.

(e) The computer may have problem with $(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$.

(f) The coefficient of determination will be indeterminant.

b. When you tell your supervisor about your intentions to use least squares to estimate the model she as s what this means. Your reply is that it uses one specific minimization criterion which is: minimize .....

(a) $\Sigma(Y_i - \hat{Y}_i)$

(b) $\Sigma|Y_i - \hat{Y}_i|$

(c) $\Sigma(Y_i - \hat{Y}_i)^2$

(d) $\Sigma(Y_i^2 - \hat{Y}_i^2)$

(e) None of the above.

c. You continue your explanation by saying that, "If the assumptions underlying least squares hold, then this procedure yields optimal estimates of the coefficients". What are the assumptions?

d. "But in what sense are least squares regression lines optimal?" she asks. You reply...

e. "Ok", your colleague says, "so they are optimal when the assumptions hold. But suppose for our data the assumptions do not hold. What does this imply with regards to the distributions of the standard set of test statistics that we always compute?

1141

2. a. Your supervisor states that 5% of the census tracts in Pittsburgh have median family size greater than 6 individuals/family. In disbelief, you gather data on the 86 census tracts and find that median family size per tract is remarkably well behaved, with $\mu = 4.5$ and $\sigma^2 = .20$. Is your supervisor correct? Why or why not?

b. The computer center at Robber Baron University claims a 95% availability for their HAL-250 computer. You are somewhat skeptical of this statement, so you gather data for the 30 days that you used the system for your latest paper. You calculate the average availability to be 85% with associated standard deviation $\frac{s}{\sqrt{n}}$ of 5%.

(i) Construct a 95% confidence interval for the true percentage.

(ii) Based on this interval, state and test a hypothesis ($\alpha = .05$) to determine the truth of the computer center's assertion.

(iii) Are the distributional assumptions that you made to test the hypothesis in (ii) appropriate? Why or why not?

1113

3. You are interested in studying what patient characteristics influence the length of stay of hospital patients after surgical procedures have been performed on them.

You have data on 10000 surgical patients from Massachusetts General Hospital in 1976.

The patients are placed into cells of a 4 dimensional contingency table with variables:

Var. 1: Length of Stay, 4 categories:
$\leq$ 1 day, 1 day - 1 week,
1 week - 1 month, > 1 month.

Var. 2: Age, 3 categories:
$\leq$ 30 years, 30-50 years, > 50 years

Var. 3: Sex, 2 categories:
Male, Female

Var. 4: Preoperative status, 4 categories:
1 = excellent, 2 = good, 3 = fair, 4 = poor

You use the stepwise procedure to fit loglinear models to this 4x3x2x4 table.

The first stage of the fitting process yields the following results:

Model	df	G^2
1/2/3/4	86	375.4
12/13/14/23/24/34	57	41.6
123/124/134/234	18	9.3

a. What can you conclude from these results with regards to choosing the "best-fitting" log linear model?

b. You find the following test statistics for each of the 2 factor interactions:

Interaction	Conditional G^2	df
12	193.1	
13	4.6	
14	125.3	
23	3.2	
24	5.7	
34	1.9	

Fill in the df column, and determine (roughly) which 2 factor interactions are non-zero.

Note: Conditional G^2 statistics are differences of G^2 statistics for specific models. For example $G^2_{1/2/3} - G^2_{12/3} = G^2_{[12]} = 193.1$

c. Based on these results, write down the "best" loglinear model for this table in terms of the appropriate U-terms, and, in the context of this example, interpret it.

1145

Part II. This section is worth 40 points. Answer 5 of the 7 questions.

1. If someone asked you to describe how exploratory and confirmatory techniques differed and what they were used for, how would you respond?

2. What feature distinguishes a table on which you would perform a two-way analysis from a table to which you would fit a log-linear model?

3. What does an extended fit incorporate that a simple two-way analysis does not?

   Choose one:

   (a) additive differences

   (b) multiplicative effects

   (c) column medians of zero

   (d) a $U_{12(ij)}$ interaction term

4. Suppose that a sample of voters in a certain district were selected by choosing every hundredth person from the list of registered voters and including that person and his/her spouse in the sample. Would this be a random sample? Why or why not?

5. After performing 4 half-steps of median polish, your two-way table of bordered and ordered residuals shows large positive values in the upper left and lower right corners and large negative values in the upper right and lower left corners. You should:

   Choose one:

   (a) Go to an extended fit

   (b) Perform more half-steps of median polish

   (c) Return to the original data and perform mean polish

   (d) Perform on X^2 test with the observed and expected frequencies

6. What does a 95% confidence interval about ρ mean?

7. What is the definition of the expectation of a continuous random variable if $f(x)$ is its probability density function?

Part III

The following questions are worth 20 points. All refer to chapters in in the book by Fairley and Mosteller. There are 10 questions each with five possible answers. Circle the one best answer. If you haven't done the readings and you guess at the answers how many points would you gain on average?

The following questions refer to Fairley's paper, "Accidents on Route 2".

1. In his initial exploration of the data on accidents Fairly used a stem-and-leaf display and concluded that:

   a. quarterly totals of accidents could not be predicted accurately by simply using an average value.

   b. the count of accidents by quarter should be transformed by taking its log and then predictions would be straightforward

   c. the data were remarkably symmetric

   d. missing values precluded any classical analysis

   e. a regression using least squares estimation would yield unacceptably low t-statistics for the time variables

2. When exploring year and quarter effects simultaneously Fairley tried an additive model and the following procedure to fit it:

   a. Log-linear contingency table analysis

   b. Linear regression

   c. Median polish

   d. Extended fit

   e. None of the above

1149

3. He also tried a multiplicative model. This involved

   a. testing for interactions

   b. multiplying marginals

   c. adding an extended fit

   d. multiplying by the conditional typical

   e. none of the above

The following questions refer to the chapter "A Statistical Search for Unusually Effective Schools" by Kiltgaard and Hall.

4. They used regression primarily as

   a. a confirmatory procedure

   b. an exploratory procedure

   c. an inferential precedure

   d. an experimental procedure

   e. an effective procedure

5. The policy implication that they derived from their study was:

   a. We need to build more effective schools

   b. Unusually effective schools cannot be produced

   c. Studies of educational effectiveness should focus on classrooms and programs

   d. Studies of educational effectiveness are doomed to failure because of colinearity problems

   e. Rural schools are more effective then urban schools.

The following question refers to the chapter by Shepard on "The Wait to See the Doctor".

6. Using a two-way analysis he concluded that,

   a. Doctor workload was significant at the 5% level

   b. Late startup of the clinic could not be represented in a linear model

   c. Late startup was a more important factor than doctor workload

   d. Doctor workload was a more important factor than late startup

   e. Race of doctor interacted with race of patient

The following questions refer to the chapter by Lave and Seskin, "Does Air Pollution Shorten Lives?"

7. This chapter

   a. proves that air pollution shortens life

   b. shows that nothing can be proven using regression

   c. proves that exploratory data analytic procedures are superior to confirmatory procedures

   d. could be improved by an extended study of the sensitivity of the results to the assumptions of least squares

   e. could be improved by an extended study of the sensitivity of the elasticities to the transformations performed

8. Another implication they draw is:

   a. in modern American, reducing air pollution is the only way to lengthen life expectancy

   b. the elasticity of poverty indicates that a reduction in poverty will result from a reduction in air pollution

   c. regression procedures should not be used to estimate models for large SMSAs.

   d. the most useful decision variable is the minimum level of a pollutant

   e. the most useful decision variable is the maximum level of a pollutant

1151

The next questions refer to the chapter by Gilbert, Light and Mosteller, "Assessing Social Innovations".

9. They distinguish between the following types of field trials:

   a. Expensive and inexpensive

   b. Purposeful and integrative

   c. Continuous and discrete

   d. Survey based and experimental

   e. Randomized and nonrandomized

10. This chapter described the application of which procedure in a policy context:

   a. Exploratory analysis

   b. Hypothesis testing

   c. Experimental meteorology

   d. Normal deviates

   e. Ridge regression

Final Examination, Second Term
Solutions

Part I.

1) a. parts b. and e.

   b. part c.

   c. Assumptions are

      i. The model is correct, i.e., y is a linear function of the x's.

      ii. Residuals are independent

      iii. Residuals are homoscedastic

      iv. Residuals are ~ Gaussian $(0,\sigma^2)$

   d. Of all linear unbiased estimates, the least squares regression line yields residuals with minimum variance. The line is "optimal" in this sense only if the four assumptions are true.

   e. i. If the model is not correct, the regression coefficients do not estimate the true population values. The coefficient estimates will be biased, although still normally distributed.

      ii. If the residuals are not independent, then one must consider the covariances of y_i and y_j when calculating sample distributions. The sums of squares will be χ^2, or mixtures of χ^2, but the degrees of freedom are indeterminate. This fact influences the distribution of t statistics, R^2, and the F statistic.

      iii. If the errors are heteroscedastic, then the residuals are not identically distributed. The sums of squares will be mixtures of χ^2 with varying degrees of freedom. The regression coefficient estimates will be linear combinations of Gaussian random variables. We will not know the linear combinations or mixtures unless we know the variance structure of the errors.

      iv. Invalidation of the assumption of Gaussianity is the most severe. None of the null hypothesized distributions will obtain; moreover, it may be quite difficult to compute the true distributions.

2) a. Median Family size ~ Gau (4.5, .20)

$$Z = \frac{\text{median family size} - 4.5}{\sqrt{.20}} \sim \text{Gau } (0,1)$$

Pr {median family size >6} =

$$\Pr \left\{ \frac{\text{median family size} - 4.5}{\sqrt{.20}} > \frac{6 - 4.5}{\sqrt{.20}} \right\} =$$

$$P\ \{Z > 3.33\} < .001$$

Our supervisor, who claims that

P{median family size >6} = .05 is incorrect

b. i. With a large number of observations, a 95% confidence interval is

$$P \pm Z_{.025} \left(\frac{s}{\sqrt{n}}\right) =$$

$$.85 \pm 1.96\ (.05) =$$

$$(.752,\ .948)$$

ii. H_0: $P = .95$

H_1: $P \neq .95$

Since our confidence interval that we constructed in part i, (.752, .948), does not contain .95, we reject H_0. We do not agree with the computer center.

iii. We have relied on the assumption that our data are approximately Gaussian. However, the true distribution is quite skewed with such a large P. In light of this skewness, a sample size of 30 is not large enough to justify our assumption.

3) a. The "best-fitting" log-linear model will have several two factor interactions, but should not have any three factor interactions.

b.

Interaction	df	
12	6	nonzero
13	3	
14	9	nonzero
23	2	
24	6	
34	3	

c. Model:

$$\ell_{ijk\ell} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{4(\ell)} + U_{12(i\ )} + U_{14(i\ell)}$$

Conditional on a patient's length of stay, his or her age and preoperative status are independent.

Part II.

1. Confirmatory techniques make distributional assumptions about the data. Based on these assumptions, inferences are made concerning the probabilities of various outcomes, and most likely parameter values. Of course, if the distributions are not accurate, then the inferences are invalid.

   Exploratory techniques do not make a priori assumptions about the data. Instead, they examine the relationships among the data and attempt to "describe" the data based on these relationships.

   Thus, even if distributional assumptions are not true, exploratory techniques may still be able to describe, summarize, and fit models to data.

2. The feature that distinguishes these tables is the nature of the cell entries. A two-way table has cell entries that are values of a third variable, a <u>response</u> variable. We in fact use the row and column variables to "explain" this response variable. In a contingency table, cell (i,j) is merely a count of the number of occurences of category i of variable A_1 <u>and</u> category j of Variable A_2.

3. Part b

4. This procedure does not yield a representative sample. The starting point in the list must be randomly chosen. A better procedure would be to use a table of random numbers to choose all individuals. Of course, spouses are not selected randomly; a spouse has a probability of unity of being in the sample if his/her spouse is included.

5. Part a.

6. A 95% confidence interval about ρ implies that if we obtained N samples and estimated ρ in each sample, and construct a 95% confidence interval about each $\hat{\rho}$, 95% of the intervals will contain the true ρ.

7. $\int xf(x)dx$

Part III.

1. Part a

2. Part e

3. Part b

4. Part b

5. Part c or e

6. Part c

7. Part d

8. Part d

9. Part e

10. Part b

1157