TIRA, Daniel Edward, 1943-
AN INTRODUCTION TO THE THEORY AND APPLICATION
OF THE PRODUCT-MOMENT FAMILY OF CORRELATIONS
VIA A COMPUTER ASSISTED INSTRUCTIONAL SYSTEM.

The Ohio State University, Ph.D., 1970
Education, psychology

University Microfilms, A XEROX Company, Ann Arbor, Michigan
AN INTRODUCTION TO THE THEORY AND APPLICATION OF THE
PRODUCT-MOMENT FAMILY OF CORRELATIONS AND A
COMPUTER ASSISTED INSTRUCTIONAL SYSTEM

Dissertation

Presented in Partial Fulfillment of the Requirements
for the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

by

Daniel Edward Tira, a.s.

The Ohio State University
1970

Reviewed by

[Signature]
Advisor
College of Education
PLEASE NOTE:
Not original copy. Several pages have indistinct print. Filmed as received.

University Microfilms.
ACKNOWLEDGEMENTS

Dissertations, as well as most research studies, typically are not conceived, conducted and finally written without the guidance and/or assistance of a great number of people. Such is the case with this study. My sincere appreciation must be extended to Dr. Edwin Novak for the guidance and inspiration he has given me throughout this undertaking. Appreciation must be extended likewise to Drs. Willavene Wolf and James K. Duncan for their support and encouragement.

When dealing with computer systems, it is necessary to depend on numerous individuals. In this regard, I truly thank Dr. Gerald Gaston and the staff of the Ohio State Regional Medical Program - CAI Study for the use of their CAI system and for their varied assistance. Included in this group are Mrs. Julie Hushak, Mrs. Ruann Pengov, Mrs. Sandra Bricker and Mrs. Mary Logue. Also I would like to express my appreciation to Mrs. Nancy Gansneder for her time and aid in operationalizing the CAI program.

Finally, but by no means in the least, my extreme gratitude to my ever-patient wife, Mary.
VITA

August 5, 1943

1965

B.S., St. Procopius College, Lisle, Illinois

1965-1967

Teaching Assistant, Mathematics Department, The Ohio State University

1967-1968

Graduate Research Associate, The Ohio State University Evaluation Center

1968-1969

Graduate Research Associate, College of Education, Faculty of Educational Development, The Ohio State University

1969-1970

Assistant Supervisor, Ohio State Regional Medical Program-Computer Assisted Instruction Study
Instructor, College of Dentistry, The Ohio State University

FIELDS OF STUDY

Major Field: Educational Research

Minor Fields: Educational Psychology

Psychological Statistics
## LIST OF ILLUSTRATIONS

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td></td>
</tr>
</tbody>
</table>

- Scatterplot of Raw Scores for Two Variables Reflecting a Linear Relationship
- Computational Example of a Partial Variance Term as the Weighted Sum of Individual Variances
- Derivation of the Generalized Correlation Coefficients (Squared)
- Definition of the Generalized Correlation Coefficients
- Scatterplot of Raw Scores for Two Variables Illustrating Their Lines of Regression
- Relationship of Regression Lines to Differing Degrees of Correlations
- Definitions of Regression Equations and Their Respective Regression Coefficients
- The Relationship Between the Regression Coefficient and Specific Variance Terms
- Definition of the Generalized Correlation Coefficients in Terms of Their Respective Regression Coefficients
- The Equality of Regression Coefficients When Raw Scores are Transformed to Standard Scores
- Definition of the Pearson r and Its Relationship to the Regression Equations
- Derivation of the Standard Error of Estimate
- Derivation of the Point-Biserial r from the Pearson r
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.</td>
<td>Derivation of the Point-Biserial r from the Pearson r (cont'd.)</td>
<td>75</td>
</tr>
<tr>
<td>15.</td>
<td>Derivation of the Phi Coefficient from the Pearson r</td>
<td>78</td>
</tr>
<tr>
<td>16.</td>
<td>Derivation of Spearman's Rho from the Pearson r</td>
<td>80</td>
</tr>
<tr>
<td>17.</td>
<td>Derivation of Spearman's Rho from the Pearson r (cont'd.)</td>
<td>81</td>
</tr>
<tr>
<td>18.</td>
<td>Progressive Differentiation of the Product-Moment Family of Correlations from a Generalized Correlation Coefficient</td>
<td>87</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I. INTRODUCTION</strong></td>
<td>1</td>
</tr>
<tr>
<td>Rationale for the study</td>
<td>1</td>
</tr>
<tr>
<td>Review of Related Research</td>
<td>7</td>
</tr>
<tr>
<td>Objectives of the Study</td>
<td>23</td>
</tr>
<tr>
<td>Methodology</td>
<td>24</td>
</tr>
<tr>
<td>Evaluation</td>
<td>28</td>
</tr>
<tr>
<td><strong>II. DEVELOPMENT OF THE CAI PROGRAM</strong></td>
<td>30</td>
</tr>
<tr>
<td>Introduction</td>
<td>30</td>
</tr>
<tr>
<td>Requirements</td>
<td>30</td>
</tr>
<tr>
<td>Design</td>
<td>32</td>
</tr>
<tr>
<td>Production</td>
<td>33</td>
</tr>
<tr>
<td>Evaluation</td>
<td>34</td>
</tr>
<tr>
<td>Design of the Text</td>
<td>35</td>
</tr>
<tr>
<td>Definition of Correlation</td>
<td>35</td>
</tr>
<tr>
<td>Covariance as a Possible Measure of Correlation</td>
<td>36</td>
</tr>
<tr>
<td>Concept of Partial Variance</td>
<td>38</td>
</tr>
<tr>
<td>Comparison of Total Variance and Partial Variance</td>
<td>40</td>
</tr>
<tr>
<td>Derivation of a Generalized Correlation Coefficient Based on Proportion of Accountable Variance</td>
<td>42</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Homogeneity of Variance as It Relates to Correlation</td>
<td>47</td>
</tr>
<tr>
<td>Inadequacy of the Generalized Correlation Coefficient as a Unique Measure of Relationship</td>
<td>50</td>
</tr>
<tr>
<td>The Regression Coefficient as a Possible Measure of Relationship</td>
<td>50</td>
</tr>
<tr>
<td>The Regression Coefficient and the Problem of Uniqueness of Measurement</td>
<td>62</td>
</tr>
<tr>
<td>Standardization of Raw Scores as a Solution to the Uniqueness Problem</td>
<td>63</td>
</tr>
<tr>
<td>The Pearson Product-Moment Correlation</td>
<td>65</td>
</tr>
<tr>
<td>The Point-Biserial r</td>
<td>72</td>
</tr>
<tr>
<td>The Phi Coefficient</td>
<td>76</td>
</tr>
<tr>
<td>Spearman's Rho</td>
<td>77</td>
</tr>
<tr>
<td>Applications of the Coefficients</td>
<td>79</td>
</tr>
<tr>
<td>Conclusion</td>
<td>83</td>
</tr>
<tr>
<td>Illustrations of the Learning Tenets Applied in Program Sequences</td>
<td>84</td>
</tr>
<tr>
<td>III. EVALUATION OF THE CAI PROGRAM.</td>
<td>95</td>
</tr>
<tr>
<td>Introduction</td>
<td>95</td>
</tr>
<tr>
<td>Analysis of Responses</td>
<td>97</td>
</tr>
<tr>
<td>Discussion</td>
<td>104</td>
</tr>
<tr>
<td>Recommendations</td>
<td>109</td>
</tr>
<tr>
<td>Summary</td>
<td>114</td>
</tr>
<tr>
<td>APPENDIXES</td>
<td>117</td>
</tr>
<tr>
<td>A. Computer Assisted Instructional Program</td>
<td>118</td>
</tr>
<tr>
<td>B. Opinionnaire</td>
<td>205</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>210</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

Rationale for the Study

Approximately ten years ago, the educational community introduced to the use of the computer for instructional purposes. Since that time much enthusiasm and hope have been expressed on behalf of Computer-Assisted Instruction (CAI). Those who have recognized the potential of CAI have worked earnestly to convince the skeptics of its possibilities for the further development of the teaching-learning process. However, after ten years, CAI seems to have progressed relatively little toward general acceptance as a serious learning force in education.

There are two major reasons for this. The first is concerned with the economic aspects of computer-based instruction. The second reason for the lack of widespread implementation of CAI is due to the relatively limited range of subject matter that has been programmed for instructional use via the computer.

Unlike many innovations in education which have been adopted on a wholesale basis in the past, CAI does not possess the relatively negligible price tag associated with other of these new educational devices. Estimates of CAI cost per student-terminal hour have reflected this.
Annsto and Seidel (1967) estimated the cost of CAI at the elementary and high school levels at $3.75 per student-terminal hour. They compared this with the cost of traditional instruction estimated at the rate of $0.36 per student hour. Another estimate by Stolurow (1967) showed an even greater cost of CAI. He stated that assuming the system were to be used 200 hours per month, computer costs would approximate $7.50 per student-terminal hour for the elementary and high school student.

As with most innovations, however, the high costs associated with their development tend to decrease as they are implemented and used on a larger scale. Ironically, this leads to the second major problem associated with the development of CAI. In order to introduce CAI on a fairly wide basis, thus reducing the costs concomitant with it, a generous selection of course material programmed for the computer system is necessary. Aside from the costs related to the hardware (e.g., the computer) of a CAI system, a great deal of time, thus money, accompanies the development of the software (the programs). Although a good number of CAI programs have been authored to date, many of these are concerned with presentation of identical subject matter. Moreover, the programs that have been developed are quite apt to be incompatible (on a subject matter basis) with the specific needs of a particular population of individuals. Thus, they are found to be suitable to only the group or institution for which they were written.

Associated with the software problems of CAI is at least one hardware aspect. Given that a program has been developed and "perfected" at institution A on computer system B, it cannot be assumed that it can
be readily employed at institution X which has computer system Y. It is reasonable to understand, therefore, why Hickey (1968) has gone as far as to state that, "program preparation is undoubtedly the principal bottleneck in CII (p.87)." Needless to say, the future of the computer for instructional purposes is a direct function of how well and how many suitable programs can be developed.

With these two general problems in mind, support of a computer-based instructional system may be considered questionable. If one would examine the capabilities of CII, however, justification for support of such an effort is readily attainable. Gerard (1966) has listed a number of benefits which CII can realistically offer the student that the traditional classroom approach cannot. Some of these include the potential for 1) personalized tutoring; 2) automatic measurement of progress; 3) the opportunity to work with vastly richer materials and more sophisticated topics; 4) the highest quality materials being produced by master teachers; 5) these quality materials to be used widely; and 6) individual modification of the programs, if necessary, to suit specific needs. Stolnover (1967) adds another important capability of CII in that it can allow research on teaching to be conducted under controlled conditions that could not be achieved as easily in the classroom.

Additionally, comparative research on the effectiveness of CII, although by no means unequivocal, has tended to support the potential of CII over other instructional systems. Two well-conducted research studies which support the use of CII are that of Grubb and Selfridge (1965) and that of Schurdak (1967).
One of the capabilities that was mentioned seems to direct itself toward the heart of instruction more than do the others. This, of course, is the potential of CAT for personalized tutoring. Put in other terms, CAT presents the potential, more than most of already existing systems, for the individualization of instruction. Although seen by many in various lights, a representative definition of individualization of instruction has been given by Coulson (1966) as that form of instruction in which "...the mode, content, and sequence of instruction are tailored to the individual's needs at any moment in time... (p. 3)." It seems fairly clear that the individualization of instruction, as posited here, would necessarily demand the accounting capabilities of a system such as that manifested in the electronic digital computer. For a "typical" teacher to be responsive to the needs of any individual at any point in time would not seem feasible.

The recognized predecessor to CAT, Programmed Instruction (PI), claimed to offer the potential for individualization of instruction, especially when PI was taken as the "scrambled book" approach developed by Crowder (1959). However, with due consideration to this approach, by its very nature it is limited in its capability of being sensitive to the moment-to-moment needs of the individual student. That is, this method of instruction is only partially response-sensitive in that it can process only certain responses and, furthermore, can redirect the learner on a relatively small base of information. Because of the large storage capacity of the computer, individualized subject matter presentations can be better realized since more information could be available.
ble on which to make decisions as to where the student should be directed in meeting his needs at that point in time.

Although the computer, more than any existing instructional method, has the potential to individualize instruction through indepth and comprehensive programming, past attempts with varying approaches toward individualization of instruction via CAI have been, in general, inconclusive. It seems that the major problem remaining is the identification of those particular means by which individualization of instruction may be realized and, moreover, utilized in an effective manner through CAI.

There appear to be two encompassing approaches to this problem. One may be considered external to the CAI program, though highly related to it. The other is inherent within the program itself. The former might include general characteristics of the learner (e.g., level of competence with respect to a particular subject matter) and other specific characteristics of the student (e.g., ability to type) which would aid in placing him in an appropriate "track" within the program. The latter approach is concerned with the actual production of the program. It would include those "method" variables (e.g., principles of learning) which would better enhance the learning process both generally and individually.

Needless to say it is easier to speak in general terms about the individualization of instruction than to actually identify and employ those variables which are central to this goal. As yet, research in learning over the years has not produced a viable set of guidelines with which to accomplish this end. This is not due to a lack of information, because the literature is replete with investigations in this area. Since the majority of these studies come from the behavioristic camp of
learning theorists, it is understandable that most of the program approaches in PI and CAI have incorporated these notions on learning. Programmed instruction is founded entirely on these principles, especially as it was developed according to the reinforcement theory of B.F. Skinner (1954). To date there has been a great deal of carry over from PI to CAI in this regard. In terms of overall effectiveness as instructional approaches, wholesale application of behavioristic principles of learning has shown inconclusive, and at times contradictory results. Since CAI does possess tremendous potential, at least on the drawing board, another approach may be more promising. It is suggested that a cognitive learning framework applied to CAI should be examined.

Because the prime intent of CAI is the individualized instruction of the student, it seems only logical that the CAI program be effective in its application. Furthermore, this must be the case to justify the existence of CAI as an instructional system. Since a behavioristic route has been followed in the past with marginal and/or questionable success, it appears that alternative approaches should be explored.

Assuming there are various effective instructional programming approaches via CAI, there seems to be no real limit as to the application of CAI to any explicitly defined course of study. However, one of the more popular areas of CAI program concentration to date has been that of statistics. For the most part, though, these programs have dealt with the descriptive or the inferential aspects of this subject matter. In some instances, programs have been generated which contain instruction in both these areas of statistics. Examination of the literature has
revealed the absence of CAI programs devoted entirely to correlational
techniques. Not necessarily a fundamental subject area of study outside
of the social and behavioral sciences, correlational methods have become
some of the more important tools of research in those disciplines. For
this reason, it is felt that attention should be directed to the develop­
ment of a C^I course of study dealing with fundamental correlational
concepts, their relationship to some of the more popular correlational
techniques, and the application of these methods in various empirical
situations.

Review of Related Research

The inadequacies stated with reference to CAI in the introductory
paragraphs were either explicitly referred to in the literature or im­
plicitly dealt with because of their conspicuous absence. With reference
to the paucity of operational CAI programs, Rogers (1960) states that
"Since the earliest experiments with CAI systems, it has been recognized
that a major obstacle to the successful application of CAI is the lack
of quality course materials (p. 31)." He further admits that "...the
availability of CAI curriculum materials for use in formal education
is a critical matter...(p. 31)."

Kittel et al. (1967) have identified three reasons why CAI has not
been convincing to the majority of educators as a feasible instructional
tool. One of these was the shortage of good CAI course content. Within
the same frame of reference, Gentile (1967) has drawn the conclusion
that three residual problems still remain after reviewing the first
generation of CAI systems. One of his observations was directed toward
the problem of development of CAI programs as being a hindrance to the progress of CAI.

Undoubtedly the foremost reason for this state of affairs concerns the large number of man hours required in the preparation of these programs. Estimates of man hours spent in program development per hour of actual program instruction have ranged from 36 for 'simple' drill exercises (Rosenbaum, Feingold, Frye, and Bennik, 1967) to anywhere from 200 (Rosenbaum et al., 1967) to over 300 (Reynolds, 1968) for programs involved in the presentation of new content material to the students. It should be pointed out, though, that these estimates are quite variable depending on the experience, competency, and imagination of the program author, the type of program (drill and practice, tutorial, etc.) and, more importantly, the quality of the program developed.

Another reason for the scarcity of CAI programs lies in the fact that until recently authoring a CAI program required expertise in programming languages in addition to skills needed to design an instructionally sound teaching medium. Lately, however, a number of author languages have been developed which allow for CAI course authors to program content matter with a minimum of difficulty, at least in comparison to the situation that existed before these languages were introduced. Frye (1968) has given a listing of these CAI-oriented languages in use at the time of his writing.

Unlike conventional computer languages, use of those developed for CAI do not require typical programming skills. Moreover, as Frye (1968) states, "...these languages include capabilities for building and administering instructional sequences. They monitor the students' activities,
collect performance records, and then make the information available to
authorized persons (p. 25)." Expanding on the capabilities of these
languages, Frye (1968) further specifies that "...the languages provide
convenient methods for accepting answers so that many variations of an-
swers can be matched...(p. 35)." He also states that most of these lan-
guages allow for redirecting students to various sequences on the basis
of the student's response history while progressing through the program.

CAT programs in the area of statistics have been almost entirely
limited to the presentation of concepts with respect to the descriptive
and/or inferential ends of this subject matter. Frye (1968) has devel-
one a program solely concerned with the topic of inferential statistics.
Researchers (Rosenbaum, et al., 1967) at System Development Corporation
have authored a program in the application of descriptive and inferential
properties of statistics. Grubb and Selfridge (1963) were among the
first to have programmed a course in descriptive statistics. Lastly,
Hansen (1966) reported the production of a program package in descriptive
and inferential statistics. As far as can be determined, then, no atten-
tion as yet has been given to the development of a CAT program which
focuses entirely on correlational methods.

Educational and psychological literature is saturated with studies
dealing with various assumed learning principles and their impact on the
learning process. Many of these same learning principles have been in-
vestigated in the applied setting of programmed instruction. Since
CAI is considered to be an outgrowth of PI, it was and is assumed by
many that the implications derived from these PI studies are directly
applicable to CAI. It is felt, here, that such an assumption is not only unwarranted but methodologically unsound. With regard to this matter, Gentile (1967) has remarked that the generalizability of programmed instruction to CAI with respect to the research in learning principles employed in PI cannot be assumed without replication of studies.

The preceding discussion tends to elicit a most important question. Is there such a thing as a theory of programmed instruction? or computer assisted instruction? In answer to this question, one may respond with another question. Is there a learning theory? Contrary to much opinion, it appears that the response to this last question must be no. Granted, there are numerous positions on how man learns. But to claim that there are viable learning theories, in the explicit sense of the word "theory", is to delude oneself. Thus, since PI and CAI have incorporated a number of the principles associated with these various positions on learning, doubt is further cast on the existence of a theory of programmed instruction or one of computer assisted instruction.

Though "...advances made in programmed learning have been based very little upon a strict application of learning theory, regardless of what devotees of the different theories may assert..." (Hilgard, 1964, pp.136-137)," PI, in general, has followed the tenets of the learning approach primarily taken by the behavioral psychologists, especially Skinner's position of reinforcement "theory". Quite fundamentally, this position with regard to programmed instruction takes the form of a linear presentation of material, sequentially organized with each successive piece of information overlapping with its predecessor to some degree. After
each "frame" of information has been given, the student is required to respond to a question concerning this material. If a correct response is given, the student is reinforced by the confirmation of his correct answer or by some other form of praise. Because of the degree of overlap between frames, chance of resulting error in response by the student is relatively low.

In terms of a very general summary, a number of statements may be made with regard to this method of programmed instruction. Borrowing from Dick (1965) and from Espich and Williams (1967), they are as follows: 1) The subject matter is systematically presented in small bits to the student; 2) the student is required to become an active participant in the learning situation by constructing an answer to a question; 3) if needed, help is afforded the student to make the desired response to the stimulus by giving him clues, by leading him toward it, or by providing the response itself; 4) when a response is made, the student receives immediate information about the quality of that response; 5) he then continues at his own rate to the next frame.

Another programmed instructional approach which allowed branching to various remedial sequences on the basis of the student's response was introduced by Crowder (1959). This departed from the strict linear format of Skinner's programs. Instead of the small step approach of the linear programs the branching programs present material in larger "chunks". There is usually a multiple choice question following it which differs from Skinner's position that the learner must construct his own response to an item rather than choose an answer from a set of alternatives. Stu-
dents are then branched to the program sequence associated with the particular response he gives. However, the branching style of programmed instruction, like its linear counterpart, accepts an S-R model of learning as its base.

Most of the learning research efforts in PI have been concerned with the effects of singular aspects of Skinner’s viewpoints. A few of the more prominent concepts of this "theory" as applied to programmed instruction that have been reported in the literature have dealt with step size within the program (Evans, Glaser and Homme, 1960; Klaus, 1955; Skinner, 1958; Coulson and Silberman, 1960; Smith and Moore, 1962), cueing or prompting (Hershberger and Terry, 1965; Seidel, 1966; Briggs, 1961; Cook and Spitzer, 1960; Silberman, Melarrango and Coulson, 1961), and knowledge of results (reinforcement) (Keats, 1968; Karer and Clark, 1963; Major, 1968; Mitzel et al., 1967; Rothkopf, 1966; Pressey, 1960; Little, 1960; Stephens, 1960; Jones, 1950; Krumboltz and Weisman, 1962; Lambert et al., 1962; Gilman, 1968). Another consideration related to this approach is that of frame difficulty (Clausmeier and Check, 1962).

The four variables of step size, cueing, knowledge of results and frame difficulty are seen as basic, among others, to the rationale of programmed instruction. Step size refers to the "...relative magnitude of transition between task units...(Ausubel, 1966, p.326)." According to the linear programming approach advocated by Skinner, small step size between frames of the program is highly desirable. In this way, the possibility of error in response by the student is held to a minimum although this type of programming format is found to be more time con-
sumin. Related to step size in terms of attempting to minimize student response errors is the notion of cueing or prompting. "Prompting seems to make the occurrence of a response more probable and makes possible the design of an instructional sequence in which each learning step makes more likely the correct response in the next step (Taber, Glaser and Schaeffer, 1965, p.7)." However, "as instruction proceeds, prompts are gradually withdrawn so that the student learns to perform and discover new knowledge without artificial prodding... (Taber et al., 1965, p.7)."

Since programmed instruction has its roots in an S-R model of learning, much emphasis is placed on knowledge of results (feedback) as a form of reinforcement. Much research has been conducted on the completeness (Chansky, 1960; Bryan and Rigney, 1956; Trowbridge and Cason, 1932), immediacy (Meyer, 1960; Sax, 1960; Evans et al., 1960) and frequency of feedback (Bourne and Haggard, 1960; Chansky, 1960; Lambert, 1967). Though basic to this model of learning, the research results on knowledge of results tend to be equivocal. There appears to be a complex state of affairs involved in when to give feedback, how to give it and the kind of feedback to be utilized. In reference to when to give feedback, the issue seems to be concerned with what kind of learning is desired at the point of potential feedback (e.g. concept learning vs. a more mechanical type of learning). With regard to how to give feedback, the problem stems from whether to present feedback continuously (after each item) or intermittently. The kind of feedback to be used causes concern in the attempt to determine whether the feedback should be partial or entire, or just the remarks "right" or "wrong".
Frame difficulty undoubtedly affects learning in various ways. Some of these are "...learning time, rate of learning..., and the amount of material that is learned and retained (Ausubel, 1968, p. 125)." Other variables affected include the learner's self-confidence, his motivation, and his anxiety. In a Skinner (linear) program approach, frame difficulty is overcome, in a sense, in that each frame is arranged such that there is a high probability of success in responding to it. A problem of frame difficulty presents itself in derivatives of this strict linear programming technique. However, the problems that arise here are compensated for by the nature of that method employed, typically that of the branching program.

From the amount of research directed toward programmed instruction and, especially, the learning principles on which it was founded, one would tend to get the impression that a behavioristic position of learning is the only alternative in this regard. Since the behavioristic tenets of learning are, and have been, highly supported by a good number of educators and psychologists, it is reasonable to understand why programmed instruction, and later, CAL, have progressed in this manner. Yet others (Ausubel, 1968; and Pressey, 1962, 1963, 1964, for example) have questioned the validity of some of the fundamental properties of behavioral learning psychology as evidenced in programmed instruction and computer-assisted instruction. This is not to imply that they are not receptive to automated instruction. Pressey (1926) suggested a self-instructional testing device, the concept of which became the forerunner of many of the present "teaching machines". Likewise, Ausubel (1968) has stated that "...pro-
programmed instruction...is potentially the most effective method for transmitting the established content of most subject-matter fields (p. 346)."

The most criticized aspect of the linear programming technique has been the issue of the small step size which lies at the heart of this approach. The basis for the use of small step size in programmed instruction is derived from Skinner's view of the learning process. His global position of operant conditioning is reflected here in what he terms the shaping of behavior. Each sequential frame represents one step to the successive approximation of some global behavior -- in this case the learning of some material. With regard to this principle, Pressey (1962) has remarked:

The student is shown material...one bit or frame at a time in the window of a mechanism or space of a programmed textbook. We cannot readily look back at what he has been over or ahead to sense what is to come, or discover any outline or structure to the material... For effective reading, for general understanding of main ideas, and for adequate study and review, this procedure seems to be as clumsy as asking a person to apprehend a picture by letting him see, in a set order, only one square inch at a time (p. 31).

Ausubel (1968) has commented that the small step size presentation of subject matter "...tends to fragment ideas presented in the program so that their interrelationships are obscured and their logical structure is destroyed (p. 324)." He further states that:

...just because task size is small and error rate is low one cannot warrantedly assume that the learning of sequentially presented ideas is necessarily rendered easy and successful, and that consolidation of existing material is therefore assured before new material is presented. In fact, the very fragmentation of content may seem to ensure mastery of the component task units at the expense of understanding the logic of the larger segments of subject matter of which they are a part (p. 324).
Finally, Jone (196?) supports the statements made by Ausubel and Pressey in remarking:

Insuring that the student can take the next step successfully in a program by sufficiently granulating the material and then arranging it systematically, is no guarantee that he will understand the logical development involved. Also the student will not remember very well the facts he does learn if he fails to comprehend the logical structure and relationships of the concepts presented (p.85).

Since prompting and knowledge of results are inherent functions of the manner in which subject-matter is presented, discussion of these principles can be made according to that particular approach taken. That is, these principles are not limited to any single method of introducing material to the student, but are more general concepts applicable to many instructional techniques. They do, however, assume certain characteristics depending on the particular approach utilized. Thus, they will not be dealt with solely in relation to linear programming.

By its very nature, linear programming is quite inflexible. "In such a program every student pursues a straight course through the program responding to every frame, with no deviations or reversals (Taber et al., 1965, p.129)." In reference to this inflexibility, Taber, Glaser and Schaeffer (1965) have further stated that "...linear program is constructed for the average student and as a result may be dull and overly simple for bright students (p.130)." It seems fairly clear that one way to lessen the rigidity of the linear technique is to employ an approach posited in the branching format of programming. Aside from this aspect, the presentation of material as suggested in this method is flexible enough to allow various learning viewpoints.
to be applied, even though, as stated before, the most popular position in this regard has been that of an S-R nature (Thelen, 1963).

The criticism leveled at the principle of step size provides an indication of an alternative means of directing learning set in a programmed instruction framework. Each of the three remarks made against the Skinnerian approach included or implied the necessary objective of the student's achievement of an understanding of the logical whole of the material presented and the internal relationships of the concepts involved. This is in contrast to the learning of bits of information within a subject matter at the expense of learning the meaning generated by these bits when interrelated in a logical fashion. It is the logical-whole approach which seems to be the most reasonable of the various means of attaining learning achievement.

A supportive position with respect to the concept of meaningful learning has been advanced by Ausubel (1967). In his position a differentiation is made between meaningful learning and rote learning. Before this differentiation is clarified, a distinction must first be made between meaningful learning and the learning of meaningful material.

In meaningful learning, the materials are only potentially meaningful. If they were already meaningful, the goal of meaningful learning (that is, the acquisition of new meanings) would be accomplished in advance. It is true, of course, that in most potentially meaningful learning tasks, the component parts of the material are already meaningful; but in these instances the task as a whole is only potentially meaningful. For example in the learning of a new geometrical theorem, each of the component words is already meaningful, but the learning task as a whole (that is, learning the meaning of the theorem) is yet to be accomplished (p.218).

* Underlining is mine
Before this discussion proceeds further, a clarification of what is meant by "meaning" in this context should be made. About this Ausubel (1967) states that:

...material possesses potential or logical meaning if it can be related on a nonarbitrary substantive basis to a hypothetical human cognitive structure exhibiting the necessary ideational background and cognitive maturity...if the material manifests the characteristics of nonarbitrariness, lucidity, and plausibility, then it is, by definition also relatable to the aforementioned hypothetical cognitive structure (p. 216).

Expanding on this topic, Ausubel (1967) continues by proposing that New meanings are therefore acquired when potentially meaningful symbols, concepts, and propositions are related to and incorporated within cognitive structure on a nonarbitrary, substantive basis. Since cognitive structure itself tends to be hierarchically organized with respect to level of abstraction, generality, and inclusiveness, the emergence of most new meanings reflects the subsumption of potentially meaningful symbolic material under more inclusive ideas in existing cognitive structure. Sometimes the new material is merely illustrative of or derivable from an already established and more inclusive concept or proposition in cognitive structure. More typically, however, new meanings are learned by a process of correlative subsumption. The new learning material in this instance is an extension, elaboration, or qualification of previously learned concepts or propositions (p. 217).

With the brief explanations of what is to be taken as meaning and meaningful learning in mind, a clarification of the distinction between meaningful learning and rote learning is possible. In contrast to meaningfully learned material, rote-learned tasks "...are discrete and relatively isolated entities that are relatable to cognitive structure only in an arbitrary, verbatim fashion...(Ausubel, 1968, p. 108)." This form of learning does not permit the establishment of the conceptual
relationships which are seen as fundamental to meaningful learning. Furthermore, "since rotely-learned materials do not interact with cognitive structure in a substantive, organic fashion, they are learned and retained in conformity with the laws of association...(Ausubel, 1968, p.106).

It is possible, on the other hand, that already meaningful material (or the component parts of this material) can be rotely-learned. Thus, the rationale for the distinction between meaningful learning and the learning of meaningful material. But, unless a meaningful learning approach is taken, as opposed to one that is rote in nature, this material (or its meaningful component parts) will be learned only as a set or series of symbols (e.g. words) that are arbitrarily related to each other and also to cognitive structure.

The description of meaningful learning as held by Ausubel is by no means exhaustive. The tenets that were presented however, do represent the salient features of his viewpoint on learning. This viewpoint is likewise amenable to most of the cognitive learning positions generally speaking. The issue which must be expounded upon now is the application of this learning approach in an instructional system. The expansion of the basics of this cognitive learning point of view will be directed toward the manner in which material may be presented via a computer.

Within this framework learning may be affected from two standpoints. According to Ausubel (1967), cognitive structure, the "processor" of incoming information,

...can be influenced substantively by the generality and integrative properties of the particular organizing
and explanatory principles used in a given (discipline)*, and programmatically by methods of presenting and sequentially ordering units of...knowledge that affect the clarity, stability, and cohesiveness of that structure (pp. 650-651).

As far as the first of these is concerned, it is primarily a matter of the identification of the particular organizing and explanatory principles of a specific content area which serve as a foundation for further enunciation of that subject matter or aspects of it. With regard to the programmatic factor, Ausubel (1967, 1968) contends that there are a few general notions that are applicable for efficient programming: 1) the use of organizers; 2) progressive differentiation of learning tasks; 3) consolidation; 4) integrative reconciliation; and 5) the internal logic of the learning material. Of these, the principles of progressive differentiation and integrative reconciliation are considered to be of central importance in the programming of meaningful subject matter. The others assume ancillary roles in coordination with these two. A brief description of all of these follows.

"The function of the organizer is to provide ideational scaffolding for the stable incorporation and retention of the more detailed and differentiated material....in the learning passage, as well as to increase discriminability between the latter and related interfering concepts in cognitive structure (Ausubel, 1967, p.236)." Organizers precede the learning material itself and represent a higher level of abstraction, generality and inclusiveness that the learning material itself. With regard to progressive differentiation, Ausubel (1962) states that:

When subject matter is programmed in accordance with the principle of progressive differentiation, the most general
and inclusive ideas of the discipline are presented first and are then progressively differentiated in terms of detail and specificity. This order of presentation presumably corresponds to the natural sequence of acquiring cognitive awareness and sophistication when human beings are exposed either to an entirely unfamiliar field of knowledge or to an unfamiliar body of knowledge....The assumption we are making here, in other words, is that an individual's organization of the content of a particular subject matter discipline in his own mind consists of a hierarchical structure in which the most inclusive concepts occupy a position at the apex of the structure and subsume progressively less inclusive and more highly differentiated subconcepts and factual data (p.651).

By consolidation, Ausubel (1967) implies "...the mastery of ongoing lessons before new material is introduced....(p.238)." In this respect, ...

the instructor makes sure of continued readiness and success in sequentially organized learning. This kind of learning presupposes of course, that the preceding step is always clear, stable and well organized. If it is not, the learning of all subsequent steps is jeopardized (Ausubel, 1967, pp.236-9).

Where the learning of a certain task presupposes an understanding of a previous one, this notion of consolidation of material was seen to have significant impact on the learning and retention of the later material (Ausubel, and Fitzgerald, 1961). In speaking of this principle relative to PI or GAI, Ausubel (1967) states:

By deferring the instruction of new material until prior material in the learning sequence is consolidated, automated instruction maximizes the effect of stability of cognitive structure on new learning. Moreover, by supplying immediate feedback, it rules out and corrects alternative wrong meanings, misinterpretations, ambiguities, and misconceptions before they have an opportunity to impair the clarity of cognitive structure and thereby inhibit the learning of new material (p.240).

Very basically, integrative reconciliation can be described as that principle by which relationships between presented ideas are explored,
where significant similarities are identified, and where real and apparent inconsistencies are reconciled. Absence of the application of this principle is reflected in a number of undesirable learning consequences. One situation that often arises is the use of multiple terms to represent the concepts which are actually identical in nature or are highly related. If this process is not manifested in the learning sequence, much confusion is apt to result. Moreover, slight differences are not clearly discriminated.

The last of the five principles, the internal logic of instructional material, is

...relevant for meaningful learning and retention outcomes, since the existence of logical meaning with the material (its relatability to corresponding relevant ideas that human beings generally can learn) is a prerequisite for the potential meaningfulness (its relatability to a learner's cognitive structure), and hence for the emergence of psychological...meaning. Logical meaning...is a function of the plausibility, lucidity, and nonarbitrariness of the material rather than of its logical or substantive validity. Hence 'internal logic' is used somewhat idiosyncratically here to designate those properties of the material that enhance these latter criteria of logical meaning (Ausubel, 1968, pp. 328-329).

Ausubel (1967, 1968) lists a number of aspects of the internal logic of material which tend to affect its potential meaningfulness to the student. Some of these are:

1) Adequacy of definition and diction, including precise, consistent, and unambiguous use of terms; the definition of all new terms prior to use; and the use of the simplest and least technical language that is compatible with conveying precise meaning.

2) Use of concrete-empirical data and relevant analysis when
developmentally warranted or otherwise helpful in the acquisition, clarification, or dramatization of meaning.

3) Stimulation of an active, critical, reflective, and analytic approach on the part of the learner, by encouraging him to reformulate presented ideas in terms of his own vocabulary, experiential background, and structure of ideas.

4) Explicit conformity with the distinctive logic and philosophy of each subject-matter discipline.

5) The selection and organization of subject-matter content around principles that have the widest and most general explanatory and integrative power.

6) Systematic sequential organization of material with careful attention to gradation of difficulty level.

7) Consistency with the principles of progressive differentiation and integrative reconciliation.

8) The use of appropriate organizers.

Application of Wubel's five general principles to the applied setting of CMI seems most reasonable. Because of the structured nature of CMI, the sequential and relational implications explicated by these principles lend themselves readily to this form of material presentation. Furthermore, the increased capability of a computer system to process information on which to branch better insures the potential meaningfulness of the material to any individual student. For individualization of instruction, as seen here, is nothing more than the insurance that each student internalizes the meanings within a given body of subject matter.

Objectives of the Study

The objectives of the study are viewed as being twofold. The first
is the development of a CMI program devoted to the presentation of the theory and application of correlational methods fundamental to the social and behavioral sciences. Explicitly, the Product-Moment family of correlational techniques will be treated. These include the Pearson "r", the point-biserial "r", Spearman's rank-difference correlation (rho), and the phi coefficient. The second intent of the study is the attempt to increase the potential for individualized instruction and thus to increase the potential for greater learning by employing a specifically defined cognitive learning theory within the development of the CMI program.

It must be stated, however, that this study is not experimental in nature but developmental. As defined here, the development of the CMI program will encompass four stages or phases. They are: 1) requirements of the program; 2) design of the program; 3) production of the program; and 4) evaluation of the program. Thus, the study is solely an initial attempt to develop a CMI program in a particular subject area within the framework of a specific learning theory. Furthermore, future improvement of the developed program will be made on the basis of the outcome of the evaluation phase of this developmental process.

Methodology:

In addition to the hardware used in a CMI system, the structure of CMI rests wholly on the program on which the system operates. Very basically, the worth of CMI as an instructional system is largely dependent on the effectiveness of the program which directs it. In this respect,
there are numerous ways of developing a CAI program. The major variables involved include the determination of the subject matter, the mode of instruction desired, the manner of progression through the program, the learning principles to be utilized within the program, and the CAI language and computer system to be used.

Suppes (1966) has outlined three broad modes of instruction which are used in CAI. The first and simplest is that of individualized drill and practice systems. These are meant as supplements to materials presented in the classroom by the teacher. The second and more complex mode is that of the tutorial systems. This mode assumes the responsibility for the presentation of concepts and for developing skills in their use. The last and most sophisticated mode is concerned with dialogue systems wherein the actual dialogue interaction would be possible between system and student. Other authors (Rogers, 1968; Stolov, 1968; Zinn, 1967) define other, more specific instructional strategies via CAI. Many of the strategies not taken into consideration by Suppes' classification are for particular "effects" desired by their use such as simulation and gaming and information retrieval. For the most part, however, Suppes' categorization is more relevant for purposes of this study. Thus, for this study, the tutorial mode will be employed since it logically represents and manifests the potential for the learning and application of new materials through the computer system.

The question of the manner of progression through the program has been dealt with in the discussion of learning approaches. Because it is felt that the potential capabilities of the computer to "instruct" would
be severely restricted using a fixed-sequence (linear) approach, the branching technique will be utilized in this program. Moreover, as implied previously, the use of a linear sequencing would significantly restrict the attempt at the individualization of instruction which is more feasible with the branching method. At a point of possible branch within the program, a number of alternative branches will be available. The specific branch will depend on the specific response given by the student while actually maintaining him on the same track of the program if his response is not totally deviant from the desired response.

A necessary adjunct to the utilization of the branched-sequenced type of program is the establishment of the criteria upon which branching takes place. Rodgers (1967) has listed a number of criteria for branching currently used by CAI program authors. They are wrong responses, confidence ratings, latencies, self-evaluation, experimental presentation, and test scores. However, Briggs (1968) has indicated that the criteria for branching decisions that have been employed to date may be a reason for the less than encouraging results which fail to show branching unequivocally superior to linear programs. It is felt here, though, that if much consideration is given to overcome learning inadequacies meaningfully by means of the branch, the use of relevant criteria for the branch will indirectly manifest themselves. That is, if the program is logically and meaningfully sequenced, use of the particular student response to a given item and/or the response history of the student up to the point of possible branch will be sufficient criteria on which to base relearning. This should be so because of the logical structural properties of
this form of program.

The learning principle of reinforcement in the form of immediate corrective feedback is assumed to play a standard role in CAI and will do likewise in this study. However, the program as a whole will be developed within the framework of the cognitive learning approach posited by Ausubel. Application of this learning approach will be reflected by the employment of the following five principles: 1) the use of organizers; 2) progressive differentiation of learning tasks; 3) consolidation; 4) interactive reconciliation; and 5) the internal logic of the learning material.

The language to be used in writing this program will be that developed by IBM and entitled Coursewriter III. Although more flexible in desirable factors than many of the existing CAI languages, there are others which offer more capabilities. However, the use of any of these is not feasible at the present time at this institution because of the computer systems available. The computer system to be used will be the IBM 360, Model 40 with an IBM 1050 input-output teletypewriter terminal. Graphs, mathematical derivations, and illustrations will be presented via an Anscomna Model 900 slide projector with a random access slide seeker attachment. These slides will be viewed on a Caritol front surface mirror, rear projection screen.

The content of the CAT program to be developed will consist of the Product-Moment family of correlational methods. Basically, the program is to have three primary emphases: 1) the presentation of general correlational concepts; 2) the development of the Product-Moment family
seldom goes further than a research report and has little impact on classroom practices (p.631).”

Results from studies using the comparative style of evaluation have generated, for the most part, a body of seemingly contradictory and/or equivocal statements as to the effectiveness of CAI. It is proposed that this is so mainly because the programs used in these studies have not been "internally" appraised and revised in a continual process until the CAI program was worth utilizing for such comparative purposes. Thus, the choice of the third method of evaluation to be used in this study.

For the purpose of program evaluation, then, a carefully selected sample of approximately ten graduate students with varying degrees of statistical background and understanding of basic statistical concepts (as perceived by the investigator) will take the CAI course and will also complete an opinionnaire directed toward the critical examination of the program. Items in this instrument will reflect choice of specific concepts for presentation, clarity of their presentation, difficulty of the concepts chosen, the general effectiveness of the course as an instructional device, and the degree to which the instructional objectives of the program were met. In the analysis of the opinionnaire, attention will be directed especially toward the differential responses made by the evaluators. In this way, an indication of the relative value of the program may at least be partially deduced. Finally, as stated above, this evaluation will serve as the basis for the future revisions of the program.
CHAPTER II
DEVELOPMENT OF THE CAI PROGRAM

Introduction

The purpose of this chapter is to describe the development of the CAI program. The title of this chapter may tend to be misleading in that the term "development" has traditionally been used in various ways. The term development, as used in this study, refers to the total process of creating the instructional program. This process begins with the initial determination of requirements and ends in the production and evaluation of the program. For this study, then, the developmental process is seen to pass through a sequence of four phases: 1) requirements (criteria); 2) design; 3) production; and 4) evaluation. Each of these phases will be treated separately as it related specifically to this study.

Requirements

The requirements phase of the developmental process served to establish the basic framework upon which the instructional program to be developed was designed, produced, and finally evaluated. This phase consisted of determining the scope, objectives, and user of the instructional program, its methods and strategies to be used in its development, and the identification of the equipment, configuration and the CAI programming language available for use.

The guidelines defined for the design, production, and evaluation of the CAI program were that it be a program in the theory and application
of the Product-Moment family of correlations presented according to Ausubel's cognitive learning approach, and that it be applicable to those students who have had experience with basic descriptive and inferential statistics. More specifically, the requirements were outlined as follows:

a) The program is to be developed in an attempt to achieve the following instructional objectives:

1) The student will be capable of identifying some possible but inadequate definitions of a correlation coefficient

2) The student will be capable of manifesting a comprehension of the theoretical development of an appropriate measure of correlation and that this coefficient represents a general statement of correlation

3) The student will be capable of demonstrating knowledge of some fundamental concepts of linear regression as they relate to the theory of correlation

4) The student will be capable of recognizing that, under certain conditions, the developed generalized correlation coefficients reduces to the Pearson Product-Moment correlation

5) The student will be capable of recognizing that, under other specific conditions, the Pearson Product-Moment correlation reduces to the Point-Biserial r, the Phi Coefficient, or Spearman's Rho, depending on the particular set of conditions stated

6) The student will exhibit a sufficient knowledge of the appropriate empirical applications of the presented correlation coefficients

b) The program is to be developed as being applicable to those users who have a basic knowledge of descriptive and inferential statistics

c) The program is to be developed within the framework of Ausubel's cognitive learning approach. Specific adaptations of the learning position to be applied:

1) The use of organizers

2) Progressive differentiation
One final requirement was defined. It concerned the constraints imposed on the developmental process by the computer system utilized and its CAI programming language. Strategy of material presentation via the system was limited to the capabilities which the language and the system afforded.

**Design**

The design phase of this developmental study constituted the generation of the textual material and accompanying visuals which served as the basis for the CAI program. The text was written as one would compose a series of lectures on the respective correlational topics. It was found that preparing textual material in this way for programming in CAI was beneficial for two primary reasons. Taking this approach in the formulation of the text served to facilitate the application of Ausubel's learning principles, keeping in mind the development of various particular topics into meaningful, integrated wholes. Secondly, this approach also facilitated the actual CAI programming (production phase) of the designed text in the attempt to preserve these meaningful, integrated wholes.

The constraints of the CAI language and/or the computer system dictated the design of some material in visual form (slides, figures, etc.) rather than for programming and presentation via the IBM 1050 teletype-writer terminal. In most instances this was reflected in the fact that mathematical derivations of statistical expressions were not amenable to
the latter form of presentation on this system. The lack of system functions (routines) executing statistical computations for application purposes in the program further dictated the design of material to overcome this limitation.

It should be noted that portions of the designed text and visuals were adapted from a presentation of correlational theory given by Edwin Ghiselli (1964). Specifically, this is in reference to the subject matter concerning the development of a generalized correlation coefficient.

Production

The translation of the textual material generated in the design phase to a CAI interactive program defined this phase of the developmental process. Narrative designed in the previous stage was coded in the IBM Coursewriter III language for input into the IBM 360/90 computer system. The textual material as it was programmed for interactive use appears in Appendix A.

It should be mentioned that there was not a continual one-to-one correspondence between the design of the instructional course and the translation of its contents to the interactive material as it appears in the CAI program. There were primarily two related reasons for this. The first concerns CAI programming effectiveness. It was felt that some concepts embodied in relatively lengthy narrative were not best presented as effectively or efficiently as such in a CAI setting. Moreover, presenting concepts in such a fashion via a CAI system does not take full advantage of the interactive capabilities of this system. Thus, where it was assumed pertinent, the narrative material was modified in an attempt to
optimize the use of the interactive mode. This certainly raised the question as to whether such modifications did actually prove to be more effective in an instructional sense. Basically, the more general inquiry contained within this question is whether CAI is more effective as a teaching instrument than are other forms of instruction. The answer to this is yet to be determined. Furthermore, the attempt to contribute to this answer was not seen as an objective of this particular study.

The second reason for the occasional disparity between the designed material and the CAI program rests in the learning approach taken. Portions of the material required modification or expansion in an attempt to increase their relevancy between adjacent sections of the program and, further, in an attempt to increase the overall meaningfulness of various sections when introduced in an interactive fashion. In general, this could be viewed also as a problem of transforming narrative material for interactive usage, attempting, in a global sense, to preserve the continuity of the presentation of concepts.

**Evaluation**

Evaluation of the CAI program was based on an opinionnaire completed by a carefully selected sample of persons with varying degrees of both statistical background and understanding of fundamental statistical concepts. Items in the opinionnaire were chosen in general, to reflect the acceptability and clarity of concepts as presented in the program in terms of their instructional value (their meaningfulness) to the student and, more specifically, to reflect the degree to which the instructional ob-
jectives were met by the program. An analysis of the responses to the opinionnaire is given in Chapter III of this study.

Design of the Text

The remainder of this chapter is based upon the design of the textual material from which the CAI program was derived. It will be presented as was designed. Selected portions of the material will then be given as they were coded in the Coursewriter III language. These instructional sequences were chosen to illustrate the application of certain tenets of Ausubel's learning approach which was described in the previous chapter.

Definition of Correlation

Very primitively, a correlation coefficient is another type of summarizing measure as that of the mean and/or variance of a set of data. It indicates, in a general sense, the degree to which data on different variables relate to each other, regardless of the specific type of correlational coefficient in question. If data on two variables are related, this implies that these data on the respective measures have something in common.

The study of correlations can be put into two separate though overlapping classes. The categories, themselves, in which the correlations can be placed are determined directly as a result of the intent of the user. If the purpose of the user is concerned with accuracy with which data on one variable can be predicted from data on another variable, then
the investigator will make use of correlation to aid in forecasting future behavior from behavior in the past. The other purpose for which application of a correlational method is appropriate is to determine the extent to which individual differences on the measures of the variables of interest are due to the same underlying factors.

Whether used for prediction or for attempting to assess the extent to which common factors are at play in determining individual differences on a pair of variables, the statistical composition of the correlation is fundamentally identical. Thus, the concept of correlation is built heavily around the idea that variance with respect to one variable is to an extent due to the same "reasons" as the variance with reference to the other variable. In other terms, what "causes" individual differences on one variable also "causes" to a degree, individual differences on another variable.

Let \( (X_1, X_2, \ldots, X_n) \) and \( (Y_1, Y_2, \ldots, Y_n) \) be sets of scores obtained by individuals 1 thru n on two measures, X and Y. Recall that the variance of X may be obtained by summing the squared deviations from the mean (\(X\)) and dividing by the number of deviation scores. The variance of Y can be computed in the same manner. Knowing that the variance for each set of data can be gotten, it is possible to obtain a measure of how the persons' scores on X vary in relation to how they vary on Y.

**Covariance as a Possible Measure of Relationship**

Following the example of the variance of any scores, if one were to multiply the deviation (from the mean) score of a subject with respect
to the Y variable, sum these over all persons and divide by the number of products, the result would be what is defined to be the COVARIANCE of X and Y. The covariance indicates, in a sense, how scores on one variable relate to their counterparts on another variable.

A first impression might be to terminate our discussion of correlation here and use the covariance as a satisfactory measure of correlation between variables. If done, however, it would be most unfortunate for at least two reasons. The covariance of two variables is a relatively "unstable" measure of relation. That is, it can take on any value on the real number line, limited only by the ranges of values defined by the scales associated with the variables from which they were computed. Thus, this term would fluctuate or could be made to fluctuate by mere choice of scale of measurement. Furthermore, since there would be no standard upper or lower limits to the covariance, how would it be known if there were a great deal of relation between variables or not. For instance, what would a covariance of 9.33 mean to me in terms of relation between two variables from which it was computed? Thus, we should not deal solely with the covariance as a measure of relationship.

From a logical point of view, however, it might be unwise to merely abandon the idea of covariance as a measure of relationship since it does speak to this issue of relationship although in an incomplete fashion. Let us, then, attempt to construct another logical frame of reference with regard to the relationship between variables and attempt to integrate it with what we have discovered about the covariance term in this respect.
Concept of Partial Variance

To begin with, please refer to Figure 1. This is simply a scatterplot of scores achieved by individuals on two measures, X and Y. Each point on the graph represents the scores obtained by a person on both of the measures. For example, the triangle indicates the fact that some person got a score of 3 on measure X and a score of 11 on measure Y.

Consider, now, the information contained at the bottom of this figure. This data can be taken to be a summary of Y scores at each X score individually. Furthermore, on the near right-hand side of the diagram, the distribution of Y scores in terms of their deviation from the overall mean Y score is given. The dots at the various points along this line represent the frequency of deviation scores for those deviation values.

The extent of the total amount of variation of Y scores is implicit within this distribution of deviation scores. The exact amount of total Y variance is given below the distribution in Figure 1, which is 40.04 in this case. This total variance may be due to many factors, some of which might be the same as those which produce variation among individuals on measure X. If we could determine statistically that portion of the total variation of Y that can be attributed to the same factors which underlie variation in X, then we would be taking a positive step in attempting to find a measure of relationship. This is saying nothing more than the fact that people tend to vary on two measures because of
Figure 1
Scatterplot of Raw Scores for Two Variables Reflecting a Linear Relationship

Adapted from Ghiselli (1968, p. 109)
similar reasons. The degree to which they vary "jointly" is, to an extent, the degree to which they are related.

Let us attempt to determine this amount of variance which may be attributable to both variables X and Y. To do this, let us take a back door approach in that we will try to pare down the total variance of Y by defining that part of this variance which is not associated with X. What we are basically doing is somewhat similar to what you may have experienced in your first statistics course with respect to the partitioning of the total variance into identifiable variance components in the analysis of variance.

Refer once again to the summary data at the bottom of Figure 1. Under each column value of X you will find a variance. This represents the variance of Y scores relative to that particular value of X. It is, as usual, based on the averaged sum of squared deviations from the mean. However, in this case, the means used for these particular variances are those respective means of the Y scores taken at each value of X. For example, the variance of Y at X equal 2 is based on a mean Y value of 15. On the other hand, the mean Y value used at X equal 6 to compute the variance of that column of Y scores is 20. In general, looking at all of the variances of Y at their particular values of X has the effect of viewing the variance of Y with the influence of X held constant. In essence, what is being done is obtaining the variances of Y without allowing X to vary with respect to (or influence) any one of them.

Comparison of Total Variance and Partial Variance

The distribution of deviation scores for values of Y computed on the basis of mean Y scores at each point of X is given at the far right-
hand side of Figure 1. This differs from the previous distribution of deviation scores in that scores in this distribution represent deviations from the particular means of the various columns of Y scores instead of deviations from the overall mean of Y. You will notice that the spread (variation) of these scores around the mean deviation score (0 as usual) is considerably smaller than the spread of the distribution of deviation scores to the left. Recall that this distribution of deviation scores was based on the overall mean Y score across, or without particular regard to, the X scores. If holding X constant had little or no effect on the variance of Y scores, then these two overall distributions would be similar. What this implies is that if there is some sort of relationship between X and Y scores, then removing the effect of variation among X scores (holding X constant) from the variation of Y scores should manifest itself in a smaller variance of Y. In our illustration this seems to be the case.

Although we have diagrammatically shown that in the case where there is a suspected relationship between scores on two variables the variation of one of them (in our illustration, Y) can be reduced substantially by controlling for the other. We next must determine a way to do this mathematically for practical purposes.

Once again consider the variance terms at the bottom of Figure 1. They vary, if you will, from the values of 2.67 to 17.75 depending on the specific value of X. Again each represents variation in Y not attributed to X since X is constant at each point. It is obvious that these variance terms are much smaller than the overall variance of Y. However,
what is needed is an overall summary of all these terms. This would define the variance of $Y$ with the influence of $X$ held constant across our particular range of $X$ values. Although a full derivation will not be attempted here, it can be shown that, statistically, this variance is nothing more than a weighted average of the $Y$ variances particular to specific values of $X$. The mathematical description of this statement is given in Figure 2. Moreover, you can see from the example that this holds true empirically in that overall variance of $Y$ with $X$ held constant is equal to the weighted sum of the individual variances of $Y$ at the particular values of $X$. This term is known as the partial variance of a variable, in this case $Y$. It is partial in that the "effects" of another variable have been eliminated.

It may be important to note here that we could have easily applied this procedure to the overall variation of $X$ with $Y$ held constant. It was just a matter of descriptive convenience that it was done as it was.

**Derivation of a Generalized Correlation Coefficient Based on Proportion of Accountable Variance**

Let us now re-examine what we have done. We have gotten a measure of variance of $Y$ wherein the "effects" of $X$ relative to $Y$ have been nullified. The question to be raised now is what exact bearing this has on the definition of a measure of relationship.

Consider for a moment the meaning of the statement "the variance of $Y$ with the effects of $X$ held constant" in terms of the underlying factors which tend to produce individual differences (variance) on measures $X$ and $Y$. Would this statement not imply that this resultant term would
Figure 2

Computational Example of a Partial Variance Term
as the Weighted Sum of Individual Variances

\[ \sigma^2_{y \cdot x} = \frac{1}{n} \sum_{i=1}^{n} \sigma^2_i \]

where \( n_i \) represents the number of scores in column "i" and \( n \)
is the total number of scores in the sample.

From our example,

\[ \sigma^2_{y \cdot x} = \frac{3(2.67) + 4(7.25) + 5(6.80) + 8(17.75) + 10(17.60)}{50} \]
\[ + \frac{8(17.75) + 5(6.80) + 4(7.25) + 3(2.67)}{50} \]

\[ \sigma^2_{y \cdot x} = \frac{602.02}{50} \]

\[ \sigma^2_{y \cdot x} = 12.04 \]

that variance which was determined previously on the basis of deviation scores.
be an indication of the extent of variation in \( Y \) due to factors other than those which produce variation in \( X \)?

Since the effects of \( X \) have been held constant in this term, it is only reasonable to assume that the variance in \( Y \) which remains must be due to reasons which cannot be similarly attributed to individual differences on measure \( X \). Granting this, then, the algebraic difference between the overall variance of \( Y \) and the partial variance of \( Y \) with respect to \( X \) must be the variance due to factors common to \( X \) and \( Y \) to some degree. In other words this difference marks the extent to which variation among people on measure \( Y \) is due to the same factors which elicit differences among individuals on measure \( X \). Refer to the first portion of Figure 3 (equation 3.1) for the algebraic representation of this.

We now find ourselves in somewhat the same situation as before when we were discussing the concept of covariance as a possible measure of relationship. We could view the difference designated as \( \sigma^2_{y'} \) in Figure 3 as being a measure of relationship also since it does define differences on subjects (variance) of one variable which is produced by factors which tend to produce differences on subjects (variance) in another variable. But, once again this index is incomplete as a measure of relationship for the same reasons as was the covariance term. That is, the size of this index varies to the extent that choice of different scales of measurement will result in diverse magnitudes of this measure. Implicit in this statement is the fact that it lacks a standard to which one may make a judgement as to the degree of relationship existing between two variables. Thus, one further step must be taken.
Figure 3

Derivation of the Generalized Correlation Coefficients (Squared)

\[ C_y^2 = \text{total variation among individuals with respect to their} \]
\[ \text{Y scores (in deviation form)} \]

\[ C_{y,x}^2 = \text{variation among individuals with respect to their} \]
\[ \text{Y scores after variation among them with respect to their} \]
\[ \text{X scores has been removed} \]

\[ C_y^2 - C_{y,x}^2 = \text{variation among individuals with respect to both} \]
\[ \text{their X scores and their Y scores} \]

Let \( C_{y,y}^2 = C_y^2 - C_{y,x}^2 \)

\[ \frac{C_{y,y}^2}{C_y^2} = \text{the proportion of total variation among individuals with} \]
\[ \text{respect to their Y scores which can be accounted for by} \]
\[ \text{variation among individuals with respect to their X scores;} \]
\[ \text{in other terms, this ratio represents the proportion of} \]
\[ \text{common factor variance between the X and Y variables} \]

Likewise,

\[ \frac{C_{x,y}^2}{C_x^2} = \text{the proportion of total variation among individuals with} \]
\[ \text{respect to their X scores which can be accounted for by} \]
\[ \text{variation among individuals with respect to their Y scores} \]

\[ \eta_{y,x}^2 = \frac{C_{y,y}^2}{C_y^2} \]

\[ \eta_{x,y}^2 = \frac{C_{x,y}^2}{C_x^2} \]
The remainder of Figure 3 delineates this step. What is done is to form the ratio of $\sigma^2 y'$ to $\sigma^2 y$, where $\sigma^2 y'$ is defined as given in Figure 3. The logical description of what this ratio represents is also given in Figure 3. As may be apparent to you, the ratio between these variances is identical to the ratio of their respective standard deviations. Since it is more customary to deal with standard deviations, the proportion is given in this respect in Figure 4. Basically, this may be thought to be our measure of correlation.

As before, this index could have been derived using as its base the variance of $X$ just as well. It was done here for the variance of $Y$ solely out of convenience.

Refer to Figure 3 and equations 3.1, 3.2, and 3.3 once again. Since the variance of $Y$ with $X$ held constant must be less than the overall variance of $Y$, the ratio of the former to the latter must be less than 1. If these two variables are identical, then the ratio equals 0 and the coefficient is 0. This implies that holding $X$ constant has no effect on the variance of $Y$ (the factors which produce variation in $X$ are not the same as those which produce variation in $Y$) and, thus, the relationship between the two measures is non-existent (an index of zero). On the other hand, if the variance of $Y$ with $X$ held constant is zero, the ratio equals one and thus the coefficient is equal to 1. This indicates that holding the factors underlying $X$ constant accounts for all the variance in $Y$.

Thus, there is a perfect relationship between $X$ and $Y$ and it is shown in the fact that the coefficient is 1. Of course, there are necessarily intermediate degrees of relationship between the two and the index will
manifest this in terms of a coefficient which ranges between zero (no relationship) and one (perfect relationship). Unlike the other two attempts at defining a measure of relationship, this index is, in a sense, standardized. Since we know the limits of it (in terms of no relationship and perfect relationship), we have a basis on which to evaluate the extent of the association.

Consider for the moment that in deriving this measure of relationship between two variables, we did not impose any explicit constraints on our procedure. That is, we did not make obvious statistical assumptions in producing the resultant index. Thus, equations 3.2 and 3.3 of Figure 3 or their counterparts 4.1 and 4.2 of Figure 4 can be said to be definitions of a generalized correlation coefficient.

This last statement is a very important one. It is so because in using this general measure of variable relationship, the underlying form of the relationship (linear or curvilinear) between the variables does not affect its application. For our illustration, the relationship apparently tended toward linearity. We could have used an example which manifested a curvilinear relationship just as easily. As a matter of fact, Figure 6 is an example of such a case.

**Homogeneity of Variance as it Relates to Correlation**

There is one detail that must be dealt with in terms of the meaningfulness of our coefficient, however. As you will recall, we used a partial variance term in our derivation. This turned out to be an average weighted sum of the variances of Y at specific points of X. Recall also that
Figure 4
Definition of the Generalized Correlation Coefficients

\[ \eta_{y,x} = \frac{C_{y'}}{C_y} \]

\[ \eta_{x,y} = \frac{C_{x'}}{C_x} \]
this was interpreted to be the extent of variation among individuals on measure Y remaining after the variation due to differences among individuals on measure X had been eliminated. But, consider for the moment the case wherein these variances of Y at specific values of X differ to a large degree. This implies that particular values of X influence Y scores differentially and thus the nature of the overall relationship between X and Y, in terms of the factors which underlie this relationship, is a very complex situation. Moreover, the average of these variance terms would not give a clear picture of this situation and initial interpretations of the resulting index of relationship may be misleading especially if careful analysis of the scatterplot has not been attempted.

Looking back to Figure 1, you may notice that there is not a "constant" relationship between X and Y as you proceed from X equals zero to X equals 8. At these end points of X, the relationship between X and Y is greater than it is at X equals 5. Another way of saying this is that the variances of Y at some values of X are smaller than those at other points of X. Thus, a well-comprehended relationship between the two variables might be difficult to achieve. As a point of consideration, however, notice that the variance of Y at various values of X is calculated on the basis of a relatively small number of cases. Thus, in our example, some of the more divergent variances may likely be due to measurement or sampling error.

In order to be able to interpret the relationship of the variables with at least minimal understanding, it should be taken that the variances of variable one at the particular values of variable two are similar in
magnitude. This should also be the case for variable two at particular values of variable one. This notion is known as homogeneity of variance.

Inadequacy of the Generalized Correlation Coefficient as a Unique Measure of Relationship

As you recall, an overall or generalized coefficient of correlation was developed which seemed to meet our initial need of an index that gave us the degree to which variables were related and which was "standardized" such that it would be interpretable from one situation to the next. Recall also, however, that this generalized coefficient took two forms, one based primarily on the variance of the Y variable and the other on the variance of the X variable. Unfortunate as it may be, these two measures of the relationship between the same two variables will, for the most part, not be the same. An illustration will exemplify this. Notice in Figure 5 that the coefficient for Y given X is .830 while that for X given Y is .913. Needless to say this easily presents a problem since we have here two different measures of relationship for the same two variables. Let us attempt a solution to this.

The Regression Coefficient as a Possible Measure of Relationship

Turn to Figure 5. This is our original graphical example. But, now notice that there are two lines drawn through the scatter of points. These lines represent lines of "best fit", one with respect to the distribution of points with X values as the base, and the other with Y values as the base. The points could just as easily have been described by curves of "best fit" in our example, but for purposes here we will utilize linear
Figure 5
Scatterplot of Raw Scores for Two Variables
Illustrating Their Lines of Regression

\[ \eta_{y,x} = 0.830 \]
\[ \eta_{x,y} = 0.913 \]

Adapted from Ghiselli (1964, p. 109)
representations. Aside from this, the trend of the scatter of points seems to be linear in nature. A clarification of what is meant by "best fit" will follow shortly.

From your mathematical experiences, recall that any straight line may be written in the general form of $Y = mX + a$, where "m" represents the slope of the line and "a" gives that point where the line crosses the Y (or vertical) axis of the coordinate axis system. When an increase in X values (going to the right along the horizontal axis) results in an increase in Y values on the line, then the slope of the line will be positive. Conversely, when an increase in X is met with a decrease in Y along the line, the slope will be negative. This fact has meaningful implication with respect to the further definition of a correlation coefficient as we shall see soon.

When we are dealing with the plot of scores on two variables and are concerned with their degree of relationship, the two lines drawn through the scattering of points are known as regression lines. In this case, the slope (m) is known as the regression coefficient. For convenience, let us denote this term as "b". Of course, the "b's" of our two regression lines will generally be different.

The importance of the linear regression coefficient in terms of correlation can be illustrated as follows. Please turn to Figure 6. In Box 1, you will notice a great deal of variability of the points around the two regression lines drawn. In this example there exists little or no correlation since the variability of Y scores at each point of X is almost identical to the total variability of Y scores. The same holds
Figure 6
Relationship of Regression Lines
to Differing Degrees of Correlation
true for the variability of X scores. In Box 2, the variability around these lines is somewhat less pronounced. Notice also that the regression lines (again, lines of best fit) seem to be approaching one another. In this example the relationship between X and Y is greater than in the first since it appears that the variance of scores around the lines is less than the total variance of either X or Y, depending on the particular line to which you are referring. Boxes 3 and 4 show even greater relationship between these variables. The variability around the lines is diminished and the two lines further approach themselves.

Notice in Box 1 that the slope of line (b) is very great and the slope of line (a) is zero. In Box 2 the slope of (b) is less than in Box 1 but the slope of (a) is greater. Likewise in Boxes 3 and 4 respectively. From what was said previously about the degrees of relationship exhibited in the various boxes, it may be said that the greater the magnitudes of the linear regression coefficients relative to each other, the greater the degree of linear relationship between the variables under consideration. Thus, it might be wise to develop a working definition of the regression coefficient as a possible measure of linear relationship.

For this purpose, a deviation score form of the regression equation will be used. Here a deviation score represents a score's deviation from the overall mean and not a column or row mean. A complete derivation of the regression coefficients may be found on pages 127-128 in the Ghiselli reference listed at the bottom of Figure 7. The results of these derivations is given in Figure 7, equations 7.1 and 7.2. Please refer to
Figure 7
Definitions of Regression Equations and Their Respective Regression Coefficients

\[ b_{y,x} = \frac{\sum xy}{\sum x^2} \]  \hspace{1cm} 7.1

\[ b_{x,y} = \frac{\sum xy}{\sum y^2} \]  \hspace{1cm} 7.2

\[ \bar{y}_i = b_{y,x} \bar{x}_i \]  \hspace{1cm} 7.3

\[ x_i' = b_{x,y} y_i \]  \hspace{1cm} 7.4

\[ y_i' = b_{y,x} \bar{x}_i \]  \hspace{1cm} 7.5

Chiselli, Edwin E.  \textit{Theory of Psychological Measurement}
them now. Again we can see that the two coefficients will differ unless it is the case that the variance of X and the variance of Y are identical. Notice that in place of the "m's" in equations 7.3 and 7.4 there are the working definitions of the slopes. In these cases they are the respective regression coefficients. It should also be noted that in deviation score form, the Y intercept ("a") is mathematically forced to be zero. Thus the regression equations (in deviations) appear as those shown in Figure 7 as equations 7.3 and 7.4.

For possible expedience in the continued development of this topic, it is now necessary to deal further with the concept of the line of regression. In the example (Figure 5) notice that the regression line fitted according to the Y values (line(a)) passed through the mean of each column of scores. Notice also that this is not the case for the regression line fitted according to the X values (line(b)). Here the mean values of X at values of Y seem to fluctuate randomly around the line. For the most part, this is the typical occurrence in real situations. However, this ostensibly may tend to cause a problem with respect to the continuation of our discussion. It is so because of the manner in which the generalized correlation coefficient was derived and because of the manner in which the regression coefficients, as given in Figure 7, were determined. Recall that in defining the partial variance term, deviations were expressed as those from the mean Y values at each value of X. With this defined the general coefficient followed. In determining the "operational" form of the regression coefficients, it was assumed that the regression lines passed through the means of the column scores and
the means of the row scores respectively. It is for this reason that the $y$ and $x$ scores of equations 7.3 and 7.4 of Figure 7 are expressed as mean deviation scores. Thus, to this point, a strong emphasis has been placed on the fact that the lines representing the trend in the data (the lines of best fit) pass through the row and column means of these data. However, as was expressed above, real situations do not usually comply with this theoretical standard. How then do we resolve this issue?

As was seen in Figure 5, one of the regression lines does not pass through the row means. The $x$ means tend to scatter randomly about this regression line. Notice that the number of scores on which each mean is based is relatively small. An assumption which we have to make now is that under the condition that we had all possible scores of this nature instead of only a limited sample of them, the means would change slightly and would fall on the regression line. Viewed in another way, if we took samples of scores from a population of bivariate scores with respect to these two variables of interest and plotted their row and column means, these means would tend to fluctuate randomly about the lines of best fit for each sample. If we put all of these samples together (ie., took the population of scores) the marked tendency would be that the means would lie on their respective lines of best fit. With this notion accepted, we can continue with our development and, moreover, accept what has been derived to this point.

While on the topic of regression lines, it was noted previously that these were lines of best fit. To clarify what is implied by a line of
best fit through a scatter of points, it will be defined to be that line wherein the horizontal or vertical distance (depending on whether you are considering one or the other of the regression lines) between points in the plot and the line is at a minimum. Put in other terms, the mean of the deviations of the points from the line would be less than that of any other line drawn through these points. Note that a line of best fit can be found whether or not it passed through the means of either column scores or row scores. It should be apparent, however, that if you have a rare case where a line may be drawn through all of the means of, say, the column scores, this would necessarily represent the line of best fit for these points without any further substantiation. This is so because of a principle you learned in your basic statistics course. It states that minimum score deviations occur when these deviations are from the mean of those scores.

Since our regression coefficients were determined on the basis that the regression line passed through the means of the scores, and that these coefficients were the slopes of their respective regression lines, a question can now be raised as to whether this is the case for the regression lines when they do not pass through the means. It can be shown that the regression coefficients as originally derived likewise represent the slopes of these regression lines and are identical in nature (Ghiselli, pages 134-136). In this explanation it is also shown that in using the original regression coefficients, the "new" regression lines satisfy the criterion of best fit lines by employing the definition of best fit.

We are now at a point where we have developed equations which can be
used to linearly describe a set of points in the best possible way. Included in these equations (as the respective slopes of the lines) are the regression coefficients. Let us now see if we can relate these to our generalized correlation coefficient since we have determined that these coefficients might be used as measures of relationships. Turn to Figures 8 and 9.

Here we have developed and expressed the generalized correlation coefficients in terms of the respective regression coefficients (9.3 and 9.4). We still are in somewhat the same shape as before, however, when our regression coefficients were found assuming that the regression lines passed through all of the means. Since these last derivations centered on this assumption, can we view this result the same as when the regression line does not pass through the means? It is possible to justify equations 9.3 and 9.4 for this more general case if we take into consideration what was said about the line of best fit as it would be if we had all possible scores. That is, we assumed that under this ideal condition these lines of best fit would theoretically pass through the means. In this situation the deviation "d" in Figure 8 would be the score deviation from the line of best fit and would also be the score deviation from the mean scores.

Although derived under more rigid assumption that the regression line passed through the means, equations 9.3 and 9.4 may be used to obtain estimates of measures of relation between two variables when it is known that the regression lines do not pass through the means of column and/or row scores. Since these equations were derived under the ideal case and since the typical regression line still represents the line of best fit, accep-
Let us call the deviation of a score from the regression line as "d" where $\bar{y}_i = b_{y,x}x$ is the equation for the line in deviation score form.

Thus $d = y_i - \bar{y}_i$ where $y_i$ is any score in column "i" (or for a fixed $X$).

Notice, that we are assuming once again that the regression line passes through the column means (thus $\bar{y}_i$).

Now $\frac{\sum d^2}{n} = \frac{\sum (y_i - \bar{y}_i)^2}{n} = \sigma_{y,x}^2$ from before.

Since $\sigma_{y,x}^2 = \sigma_y^2 - \sigma_{y_i}^2$ from our deviation of the generalized correlation coefficient

$$\frac{\sum d^2}{n} = \sigma_{y,x}^2 - \sigma_y^2 - \sigma_{y_i}^2$$

From the determination of the "new" regression line as lines of best fit, it can be shown (Chiselli, page 135) that

$\sigma_{y,x}^2 = \sigma_y^2 - b_{y,x} \sigma_x^2$

Thus $\sigma_y^2 - \sigma_{y_i}^2 = \sigma_y^2 - b_{y,x} \sigma_x^2$

or $b_{y,x} \sigma_x^2 = \sigma_{y_i}^2$

and $\sigma_{y_i}^2 = b_{y,x} \sigma_x$
Similarly it could be shown that

$$C_{x'} = b_{x,y} C_y$$  \hspace{1cm} 9.1

Dividing by $C_y$ in equation 8.1 and $C_x$ in equation 9.1 we have:

$$\eta_{y,x} = \frac{C_{y'}}{C_y} = b_{y,x} \frac{C_x}{C_y}$$  \hspace{1cm} 9.2

and

$$\eta_{x,y} = \frac{C_{x'}}{C_x} = b_{x,y} \frac{C_y}{C_x}$$  \hspace{1cm} 9.3
tance of the use of these equations as best estimates of relationships should be reasonable.

The Regression Coefficient and the Problem of Uniqueness of Measurement

From a generalized statement of relationship between two variables which is based on the proportion of accountable variation in one of the variables by the other, we have restated this index in terms of the regression coefficient relative to that particular generalized correlation coefficient. The advantage in restating the correlation in these terms was seen in the fact that it now is capable of defining both positive and negative relationships between variables since the slope (b) can be either positive or negative. However, once again we face the unfortunate situation wherein we have two correlation indices for the same two variables. That is, depending on whether we use one or the other of the two regression coefficients, we will get two distinct measures of correlation since the two regression coefficients will not usually be the same. Furthermore, you will recall in equation 9.3 and 9.4 that the correlations were given not only in terms of the regression coefficients but also in terms of the ratio of the standard deviations of X and Y. These ratios will not necessarily be identical either.

The most feasible way of getting out of this bind would be to do something that would result in the equalizing of the two regression coefficients. If this is possible, the next step would find another way of equalizing the standard deviation of Y to that of X. If both of these "transformations" are possible, then we would have a unique measure of correla-
Standardization of Raw Scores as a Solution to the Uniqueness Problem

Probably the best way to approach the solution of our first problem would be an attempt to determine why we get two regression lines and thus two regression coefficients. A relatively simplistic answer lies in the fact that the magnitude of the regression coefficients are inherently dependent on the underlying scales of measurement of the raw scores on which they were based. Thus, if we could equate the scales of measurement of the two variables such that they were comparable, our first problem would be solved in that the two regression coefficients would be identical in magnitude.

A basic technique in measurement to accomplish this task is to standardize the underlying raw scores. In other terms, this transforms the raw scores into a scale of measurement whose units of measurement are comparable to any other set of raw scores which have been transformed similarly. A very fortunate by-product of this transformation lies in the fact that the variance of such "standard scores" is one. Thus, by transforming the raw scores of both variables into standard scores, we not only equate the regression coefficients but also equate the variances of both variables. A mathematical proof of this is given in Figure 10 along with the form the regression coefficients and thus the generalized correlation coefficients take on after the transformation of the underlying raw scores to standard scores.
The Equality of Regression Coefficients When Raw Scores are Transformed to Standard Scores

From our derivation of the regression coefficients, it was found that

$$b_{y,x} = \frac{\sum xy}{N \sigma_x^2}$$
and $$b_{x,y} = \frac{\sum xy}{N \sigma_y^2}$$

When the raw scores are transformed into standard scores in the usual manner (i.e., $$z_x = \frac{x - \bar{x}}{\sigma_x}$$), the regression coefficients are now expressed as:

$$b_{y,x} = \frac{\sum z_x z_y}{N \sigma_x^2}$$
$$b_{x,y} = \frac{\sum z_x z_y}{N \sigma_y^2}$$

Since the variance of a set of standard scores is equal to unity, we have the following:

$$b_{y,x} = \frac{\sum z_x z_y}{N \cdot 1}$$
and $$b_{x,y} = \frac{\sum z_x z_y}{N \cdot 1}$$

Since the right hand side of both of these equations is identical, we have our desired result, namely

$$b_{y,x} = b_{x,y}$$

Furthermore, since $$\sigma_{z_x}^2 = \sigma_{z_y}^2 = 1$$, our expression for the correlation coefficients are

$$\frac{\sigma_{z_y}^2}{\sigma_{z_y}^2} = b_{y,x} \frac{\sigma_{z_x}^2}{\sigma_{z_y}^2}$$ and $$\frac{\sigma_{z_x}^2}{\sigma_{z_x}^2} = b_{x,y} \frac{\sigma_{z_y}^2}{\sigma_{z_x}^2}$$

or

$$\frac{\sigma_{z_y}^2}{\sigma_{z_y}^2} = b_{y,x}$$ and $$\frac{\sigma_{z_x}^2}{\sigma_{z_x}^2} = b_{x,y}$$
The Pearson Product-Moment Correlation

A small summary of what we have done to this point may be beneficial. Through a series of derivations we have developed a generalised correlation coefficient in two forms, have expressed these coefficients in terms of their respective linear regression coefficients, and lastly have shown that under a certain condition they are identical. Thus we now have a single, unique measure of linear relationship between two variables which has the desired capability of expressing both positive and negative relationships between variables. The condition, of course, under which this finally occurred was that the underlying raw scores of both variables were to be given as standard scores. Under these circumstances, the correlation found is more popularly known as the Pearson Product-Moment Correlation. Referring to Figure 11, we have this defined in equation form.

Because of their relative importance, let us review the major conditions under which the Pearson r, as it is known, was developed. First it is assumed in the use of this correlation coefficient that the underlying relationship between the variables under consideration is linear in nature. Use of this coefficient when the relationship is not linear will give a false representation of the magnitude of the true relationship. Secondly, it is assumed that a homoscedastic relationship exist between the variables. That is, it is assumed that the spread (variance) of scores about the best fitting straight line is approximately the same at all levels of both variables. If the spread is not roughly the same, the relationship is more complex than "assumed" to be and the Product-Moment Correlation would fail to reveal this information. Furthermore,
Because it was determined that $b_{y.x} = b_{x.y}$ it follows that

$$\frac{C_{z_y'}}{C_{z_y}} = \frac{C_{z_x'}}{C_{z_x}}$$

Thus,

$$\frac{C_{z_y'}}{C_{z_y}} - \frac{C_{z_x'}}{C_{z_x}} = b_{y.x} \quad b_{x.y} = \frac{\Sigma z_y z_x}{N} = r_{xy}$$

Or, just

$$r_{xy} = \frac{\Sigma z_y z_x}{N} \quad \text{(Pearson Product-Moment Correlation)}$$

In terms of standard scores and the Pearson coefficient, the regression equations become:

$$\bar{z}_{y_1} = r_{xy} z_{x_1}$$

and

$$\bar{z}_{x_1} = r_{xy} z_{y_1}$$

Or, more generally:

$$z'_{y_1} = r_{xy} z_{x_1}$$

and

$$z'_{x_1} = r_{xy} z_{y_1}$$

In raw score form, these equations become:

$$\bar{y}_1 = r_{xy} \frac{C_y}{C_x} (x_1 - \bar{x}) + \bar{y} \quad 11.1$$

and

$$\bar{y}_1 = r_{xy} \frac{C_y}{C_y} (y_1 - \bar{y}) + \bar{y} \quad 11.2$$

or

$$y'_1 = r_{xy} \frac{C_y}{C_x} (x_1 - \bar{x}) + \bar{y} \quad 11.3$$

and

$$y'_1 = r_{xy} \frac{C_y}{C_y} (y_1 - \bar{y}) + \bar{y} \quad 11.4$$
interpretation of this particular coefficient would be difficult under this circumstance.

A final assumption or condition with respect to the Pearson r which has not been considered yet is that of the underlying distribution of the variables. This correlation assumes that scores on one variable are distributed normally at each respective value of the other variable and vice versa. This assumption (as with the other two) is central to the mathematical foundations of this correlation. As a result, the assumption of normality plays a major role in the development of the inferential statistical properties of this coefficient of correlation. Thus, this assumption has greater impact with respect to the interpretation of the coefficient than with respect to its use.

Although normality is used as the model for the underlying distributions of the variables, departure from the normal assumption does not seriously affect the use of the Pearson r. An important consideration, however, does relate to the distributions of the variables. If the distributions of these variables differ from one another (in terms of their shapes—skewness and kurtosis), it can be shown that the maximum value of r is not 1.00 in this situation but some value less than this depending on the degree of disparity between the distributions. However, this may only be an academic point since it is rare that a perfect relationship between typical variables studied in the behavioral sciences is found.

To get away from the flow of the course for a moment, let us talk about a more practical value of the correlation coefficient and its lim-
itations. If you will recall, it was stated near the outset of this course that there were two major uses of the correlation coefficient. One of these was indirectly made reference to in the development of our generalized correlation coefficient. This was to attempt to determine the extent to which individual differences on two measures were due to the same underlying factors. The other was the use of the correlation coefficient for predictive purposes. The question arises as to how it might be utilized for this latter purpose. Let us now take an indirect way of demonstrating this.

Suppose you had a distribution of scores for a known sample of persons on a variable, call it variable Y. Suppose also that you were asked to guess a person's score from a number of scores you had at your disposal. Since it was most likely shown in your first statistics course that the mean of a distribution is that score from which the sum of squared deviations is at a minimum, the logical or the best estimate of any person's Y score in the distribution would be the mean. However, as you know, even though the mean would be the best guess in such a situation, it would not be all that accurate. The degree of accuracy would depend on the variance of that distribution.

Suppose now we had more information at our disposal. Say we had another distribution of scores for the same sample of people on some variable (variable X). Let us assume for our purposes that there exists a relationship between the scores on both variables. Let us further assume that this relationship is linear in nature. With this added information, it is now possible to better "guess" or predict a person's score on variable Y.
knowing his score on variable X.

From our development of the generalized correlation coefficient, recall that the scatter of Y scores around the regression line at any point of X was less than the scatter of Y scores not taking X into consideration. Practically speaking the range and variance of Y scores at any value of X is somewhat diminished in comparison to that if a value of X were not taken into account. Thus, if you were to guess a person's Y score knowing his score on X it would most likely be the mean Y value at that specific value of X.

Let us back up one step and add on a reservation to the last statement. The answer of the mean Y value at a specific value of X is probably the best answer. In an ideal sense, these mean values would fall on a straight line, that line being our regression line for Y values or otherwise known as the line of best fit for these points. But, as was noted in detail in the previous discussion, most often the means do not lie on the regression line. When one takes samples of scores from a hypothetical population of scores and plots these, the various Y means (at points of X) will fluctuate differently, though randomly, about the regression line. Furthermore, to repeat, when one has the total population of scores at hand, these Y means should fall quite near the regression line if not on it. Thus, in the long run (to get back to our illustration), the most accurate prediction of a Y score would be the Y value on the regression line which corresponds to that person's X score knowing that person's X score.

It should be realized that the Y score will not be predicted precisely from the X score even though this is the best estimate. Thus, there
is a measure of error associated with each prediction. Fortunately, it is possible to get some indication of the magnitude of this error. To do so, let us essentially return to the concept of score deviation from the regression line since this is fundamentally what this error is — the difference of the actual score from the predicted score which happens to be on the regression line. For this development turn to Figure 12.

Equations 12.1 and 12.2 represent what is known as the Standard Error of Prediction or the Standard Error of Estimate, one for the estimation of Y from X and the other for prediction of X from Y. They describe the degree of error included in predicting one score from another for this particular sample of scores. Equation 12.3 is also the Standard Error of Estimate, but only given in terms of standard scores. Note that in this case, both forms of this standard error are identical.

Returning to Figure 12, notice that when there is no relationship between the variables (r=0), the standard error of the estimate is equal to the standard deviation of the predicted variable (in the case of equations 12.1 and 12.2). Similarly, when this relationship is perfect (r=1), the standard error is zero. These developments should be familiar since they are really the same ones as were discussed when talking about the partial variance term in the development of the generalized correlation coefficient. Further appraisal of the standard error of estimate should indicate to you that it is nothing more than an average partial standard deviation expressed in terms of the measure of linear relationship between two variables and standard deviation of one of them.

The assumption of homogeneity of variance has rather important im-
Derivation of the Standard Error of Estimate

Let us represent the score deviation from the regression line as before: $y_i - \overline{y}_i$ where $\overline{y}_i$ is that $y$ value which falls on the line.

Then $\Sigma (y_i - \overline{y}_i) = \sigma_y.x = \sigma_y - b_{y.x} \sigma_x$ from before.

Since $b_{y.x} = \frac{\sum xy}{N \sigma_x^2}$ we have

$$\sigma_y.x = \sigma_y^2 - \left(\frac{\sum xy}{N}\right) \frac{\sigma_x^2}{\sigma_y}$$

Moreover, since $r_{xy} = \frac{\sum xy}{\sigma_x \sigma_y}$, $\frac{\sum xy}{N \sigma_x \sigma_y} = \sigma_x \sigma_y r_{xy}$

Substituting this into the equation, we have:

$$\sigma_y.x = \sigma_y^2 - \frac{\sigma_x \sigma_y \sigma_{r_{xy}}}{\sigma_y}$$

or

$$\sigma_y.x = \sigma_y^2 (1 - r_{xy}^2)$$

or

$$\sigma_y.x = \sigma_y \sqrt{1 - r_{xy}^2}$$  \hspace{1cm} (12.1)

Similarly

$$\sigma_{x.y} = \sigma_x \sqrt{1 - r_{xy}^2}$$  \hspace{1cm} (12.2)

If we would have used standard scores instead of deviation scores,

$\sigma_x = \sigma_y = 1$ and

$$\sigma_y.x = \sigma_{x.y} = \sqrt{1 - r_{xy}^2}$$  \hspace{1cm} (12.3)
plication with respect to the interpretation of the standard error of esti-
mate. The index as given in equations 12.1 and 12.2 is more-or-less an
overall (or average) index of variability around the regression line. To
be sure, in a real sense it is reasonable to assume that there will be
differing degrees of spread about the regression line as one goes from one
end of it to the other. Thus, the standard error of estimate is just an
estimate in itself -- but remember it is the best estimate we can make as
to the spread of scores around the line of best fit. If we can assume
that the variability of scores around the line is fairly standard at points
along the line, then our "estimate" of the standard error of estimate pre-
sents us with a pretty good overall picture of the real situation.

One other assumption has implication to the standard error of esti-
mate in terms of more detailed use and interpretation of it. If the
scores in the rows and columns are normal and assuming, again, homogeneity
of variance, then we know that 68% of the scores fall between plus and
minus one standard error of estimate in this case. Thus, if a person's
score is predicted to be 100 and the standard error of estimate is found
to be 10, the probability that person's true score falls in the range 90-
110 is .68.

The Point-Biserial r

Most of the measuring instruments with which we deal in the behavior-
al sciences are considered to be based on a continuous scale of measure-
ment. That is, they assume that there is a multitude of scores that could
be recorded between the lowest and highest scores possible on the instru-
ment. The Pearson Product-Moment Correlation, which we have just discus-
sed, presumes that the variables to which it is applied possess this underly-
ing form of measurement. Many times, however, a measure of relation-
ship is required between a variable with underlying continuous scale and
another which allows for only two possible scores, e.g., right-wrong, yes-
no, Democrat-Republican, etc.. The question arises as to what to do when
we have a continuous variable and one dichotomous variable. Such is the
case, for example, when we would like to correlate a total test score
with an item score when the item score is given as either correct or in-
correct. The solution to this is relatively simple and it flows from the
Pearson r.

Since a dichotomous scale of measurement is only a special case of a
continuous scale of measurement, it seems reasonable to apply the Pearson-
ian coefficient to a problem of this sort. In this case, one of the vari-
ables would have only scores of 0 or 1 to represent the two possibilities
on its scale. As a matter of information, 0 and 1 are used only for the
sake of convenience. Although direct application of the Pearson r is ap-
propriate here, a more expedient form of this correlation coefficient is
derived and given in Figures 13 and 14 for a special case of this nature.
Given in this form, the Pearson r is known as the point-biserial coeffi-
cient of correlation.

Before the days of the high speed computer, the point-biserial was
sometimes used in place of "r" by dichotomizing the scale of one of the
variables, taking, for example, the scores above the mean as 1 and those
below as 0. This was done because of the computational ease of the point
biserial relative to that of the Pearson r. However, the time and effort
Derivation of the Point-Biserial r from the Pearson r

From the definition of the Pearson r, we have the computational form as follows:

\[ r_{xy} = \frac{N \sum xy - (\sum x)(\sum y)}{\sqrt{N \sum x^2 - (\sum x)^2} \sqrt{N \sum y^2 - (\sum y)^2}} \]

Let us assume that variable \( X \) is dichotomous in nature, i.e. takes on possible values of 0 and 1. Then,

\[ \sum x = N_1 \quad \text{(number of } X \text{ scores of value 1)} \]
\[ \sum x^2 = N_1 \]
\[ \sum y = \sum y_0 + \sum y_1 \quad \text{(sums of } Y \text{ values with corresponding } X \text{ values of 0 and 1)} \]
\[ \sum xy = \sum xy_0 + \sum xy_1 = \sum y_1 \quad \text{(Since all } X_0 \text{'s are 0)} \]

Substituting,

\[ r_{xy} = \frac{N \sum y_1 - N_1 \sum y}{\sqrt{N N_1 - N_1^2} \sqrt{N \sum y^2 - (\sum y)^2}} \]
\[ = \frac{(N_0 + N_1) \sum y_1 - N_1 (\sum y_0 + \sum y_1)}{\sqrt{(N_0 + N_1)(N_1) - N_1^2} \sqrt{N \sum y^2 - (\sum y)^2}} \]
\[ = \frac{N_0 \sum y_1 + N_1 \sum y_1 - N_1 \sum y_0 - N_1 \sum y_1}{\sqrt{N_0 N_1 + N_1^2 - N_1^2} \sqrt{N \sum y^2 - (\sum y)^2}} \]
\[ = \frac{N_0 \sum y_1 - N_1 \sum y_0}{\sqrt{N_0 N_1} \sqrt{N \sum y^2 - (\sum y)^2}} \]
\[ = \frac{N_0 N_1 y_1 - N_0 N_1 y_0}{N \sum y_1 \sqrt{N_0 N_1}} \quad \text{Since } M_1 = \frac{\sum y_1}{N_1} \]
\[ = \frac{N_0 N_1 y_1 - N_0 N_1 y_0}{N \sum y_1 \sqrt{N_0 N_1}} \quad \text{and } M_0 = \frac{\sum y_0}{N_0} \]
Derivation of the Point-Biserial $r$ from the Pearson $r$ (cont'd)

\[ r_{\text{Pbis}} = \frac{(M_y - M_0) \sqrt{P_0P_1}}{\sigma_y} \]

Where $P_0$ = proportion of scores with value of 0

and $P_1$ = proportion of scores with value of 1.

\[ = \frac{N_0N_1 (M_y - M_0)}{N \sigma_y \sqrt{N_0N_1}} \]

\[ = \frac{(M_y - M_0) \sqrt{N_0N_1}}{\sigma_y N} \]

\[ = \frac{(M_y - M_0) \sqrt{N_0N_1}}{\sigma_y} / \sqrt{N_0N_1} / \sqrt{N^2} \]
saved in calculating the measure of relationship in this way is of no value because of information lost in doing so.

Because the point-biserial does not make the assumption of normality on the dichotomous variable, it can be said that this coefficient is more generally applicable than is the Pearson r. Again, however, it must be stressed that when r can be used as an appropriate coefficient of correlation, the point-biserial should be dismissed as an alternative. Even though the point-biserial is a Product-Moment correlation, numerically the point-biserial is not comparable to the Pearson r in value due to the previously mentioned loss of information incurred in artificially dichotomizing continuous data.

Recall that in the discussion of the Pearson r it was stated that the shapes of the distributions of the two variables affected the maximum size of the coefficient. The argument is much the same in the case of the point-biserial but to a greater extent. Because of the quite different nature of the variables being correlated, the shapes of these two distributions are necessarily quite unlike. Thus, the maximum value that can be attained by the point-biserial approaches .80.

Let us now proceed with the strategy begun with the introduction of the point-biserial correlation coefficient. Here we presented the case where one of the variables under consideration was dichotomous. Now let us take the case where both of the variables are of a dichotomous sort. The measure of relationship in this instance is known as the Phi Coefficient.

The Phi Coefficient
The discussion of phi follows along the same line as that of the point-biserial with the added notion that now both variables are dichotomous. As with the point-biserial, the derivation of phi follows directly from the Pearson r. Thus, it is fundamentally a product moment coefficient. The derivation of phi into its operational form from the Pearson r is given in Figure 15.

Unlike the point-biserial, the phi coefficient exhibits a possible range of values from -1.00 to +1.00. This is so since both variables are of dichotomous nature and therefore it is more feasible (in comparison to the point-biserial case) for their distributions to be alike. However, the maximum values of the coefficient, as with the other two, are still dependent on the relative shapes of the variable distributions.

As with the point-biserial, use of the phi coefficient on data which is suitable for application of the Pearson r results in loss of information from the user's viewpoint since there is a curtailment of information by dichotomizing otherwise continuous variables.

**Spearman's Rho**

The last correlation coefficient to be discussed in this course is the "rho" coefficient, known also as Spearman's rho or the Rank-Difference correlational method. It, like the two preceding correlation coefficients, is derived from the Pearson product-moment correlation. The possible range of values which rho can assume runs from -1.00 to +1.00.

As one of its titles implies, the rho coefficient is applied to data which are in the form of ranks on both variables. It describes the degree of relationship between two sets of ranked data on N individuals.
Let the table for X and Y values of 0 and 1 be given as follows:

```
+---+---+
|   | c | d |
+---+---+
| 0 | a | b |
+---+---+
| 1 | c | d |
+---+---+
```

Let a, b, c, d represent the number of scores falling into the four possible combinations of X and Y scores:

\[
\begin{align*}
\xi X &= a\cdot 0 + b\cdot 1 + d\cdot 1 = b + d \\
\xi X &= a\cdot 0 + c\cdot 0 + b\cdot 1 + d\cdot 1 = b + d \\
\xi Y &= a\cdot 0 + b\cdot 0 + c\cdot 1 + d\cdot 1 = c + d \\
\xi Y &= a\cdot 0 + b\cdot 0 + c\cdot 1 + d\cdot 1 = c + d \\
\xi XY &= a(0\cdot 0) + b(1\cdot 0) + c(0\cdot 1) + d(1\cdot 1) = d \\
N &= a + b + c + d
\end{align*}
\]

\[
\begin{align*}
N \xi X &= (a + b + c + d)d - (b + d)(c + d) \\
N \xi Y &= (a + b + c + d)d - (b + d)(c + d)
\end{align*}
\]

\[
\begin{align*}
\xi xy &= \frac{N \xi XY - \xi X \xi Y}{\sqrt{N \xi X - (\xi X)^2} \sqrt{N \xi Y - (\xi Y)^2}} \\
\xi xy &= \frac{(a + b + c + d)d - (b + d)(c + d)}{\sqrt{(a + b + c + d)(b + d) - (b + d)^2} \sqrt{(a + b + c + d)(c + d) - (c + d)^2}} \\
\xi xy &= \frac{ad + bd + cd + d^2 - bc - dc - bd - d^2}{\sqrt{(b + d)(a + b + c + d - b - d)} \sqrt{(c + d)(a + b + c + d - c - d)}} \\
\xi xy &= \frac{ad - bc}{\sqrt{(b + d)(a + c)} \sqrt{(a + b)(c + d)}}
\end{align*}
\]
or objects. Thus, it assumes an underlying scale of measurement for variables which would permit ranking. A derivation of rho from r under ranked data conditions is presented in Figures 16 and 17.

As with the point-biserial and phi, rho is sometimes used in place of the Pearson r as a measure of relationship. This implies the rearrangement of the data to comply with the bases of the coefficient in use. Like the other two, rho should not be employed where the application of r can be justified. Whereas the Pearson r deals with both the magnitude of scores and their respective relative positions in a series, rho is concerned only with the relative position of the data. In spite of this, rho does produce a value which is quite similar to r in magnitude if both were to be employed on the same continuous data (ranked for rho, of course) and if these continuous data were approximately normally distributed. Applying rho to the example data on which the previous three coefficients may have been applied, we can get a comparison of these measures and how the specific transformation of the data for use with the respective correlations affect its size in relation to the Pearson r.

Applications of the Coefficients

The use of any one of the presented correlation coefficients for analysis purposes is of course dependent upon a particular data situation. Assuming the ideal conditions of data collection and the applicability of one of the four coefficients presented in this program, determine which of these correlations is best applicable in the hypothetical situations which follow.
Figure 16

Derivation of Spearman's Rho from the Pearson r

Organisation of data when in terms of ranks for two variables might look as follows:

<table>
<thead>
<tr>
<th>PERSONS</th>
<th>1</th>
<th>R</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>or</td>
<td></td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>OBJECTS</td>
<td>1</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>K</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>E</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>O</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>R</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>S</td>
<td></td>
</tr>
</tbody>
</table>

X = \sum_{i=1}^{N} i \quad (Sum of ranks 1-N)

Y = \sum_{i=1}^{N} i \quad (Sum of ranks 1-N)

Since (X-Y)^2 = X^2 + Y^2 - 2XY

and \sum(X-Y)^2 = \sum X^2 + \sum Y^2 - 2 \sum XY

\sum XY = \frac{\sum X^2 \cdot \sum Y^2}{2}

Let (X-Y), the difference of a pair of ranks, be signified as D.

\[
\rho_{xy} = \frac{\sum XY - \sum X \sum Y}{\sqrt{\sum X^2 - (\sum X)^2} \sqrt{\sum Y^2 - (\sum Y)^2}}
\]

\[
\rho_{xy} = \frac{N \left[ \sum X^2 + \sum Y^2 - \sum D^2 \right] - \sum X \sum Y}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}}
\]

\[
= \frac{N \left[ \frac{N(N+1)(2N+1)}{6} \cdot \frac{N(N+1)(2N+1)}{6} - \sum D^2 \right] - \left[ \frac{N(N+1)}{2} \right] \left[ \frac{N(N+1)}{2} \right]}{\sqrt{\frac{N(N+1)(2N+1)}{6} - \frac{N^2(N+1)^2}{4} \sqrt{\frac{N(N+1)(2N+1)}{6} - \frac{N^2(N+1)^2}{4}}}}
\]
Derivation of Spearman’s Rho from the Pearson r (cont’d)

\[
\begin{align*}
\rho &= \frac{N \left( 2N(N+1)(2N+1) - \frac{\sum D^2}{2} \right)}{N^2(N+1)(2N+1) - \frac{N^2(N+1)^2}{4}} - \frac{N^2(N+1)}{4} \\
&= \frac{2N^2(N+1)(2N+1) - 6N \sum D^2 - 3N^2(N+1)}{2N^2(N+1)(2N+1) - 3N^2(N+1)} \\
&= \frac{\frac{4}{12} \left( N^3 + N^2 - 6 \sum D^2 \right)}{\frac{2}{12} \left( N^3 + N^2 - 6 \sum D^2 \right)} - \frac{N^3}{3} \sum D^2 \frac{6}{3N^3} - \frac{6N}{3N^2} - \frac{3N}{3N^1} \\
&= \frac{N \left( N^3 - N - 6 \sum D^2 \right)}{N(N^3 - N)} \\
&= \left( \frac{N^3 - N}{N(N^3 - N)} \right) - \frac{6 \sum D^2}{N(N^3 - N)} \\
&= 1 - \frac{6 \sum D^2}{N(N^3 - N)}
\end{align*}
\]

17.1
NOTE: For each case, type the number corresponding to the correlation that you would apply:
1. Pearson Product-Moment r
2. Point-Biserial r
3. Phi Coefficient
4. Spearman Rho
5. None of the above

Situation 1:
A sociologist asks you to analyze some data he has collected on persons' political affiliation (assumed Democrat or Republican). He also knows the racial background (assumed Negro or Caucasian) of these persons. He would like to determine whether there is a statistical relationship between political affiliation and race for the persons in his sample. What coefficient would you recommend?

Situation 2:
You, as a teacher-researcher, are interested in determining how well a number of subject-matter units were presented to a class (in terms of clarity, meaningfulness, etc.). Thus, you select two evaluators to rate units, relative to each other, in terms of effectiveness of presentation. Now you would like to see if there is a relationship between the ratings of both evaluators. Which correlation would you use?

Situation 3:
An admissions officer at a small liberal arts college would like to see if there is a statistical relationship between high school Grade
Point Averages (GPA) of incoming freshmen with their GPA's at the end of their first year in college. In the future he would also like to determine the statistical relationship between this group's H.S. GPA's and their four year college GPA. What correlation coefficient should he use in either instance?

 Situation 4:

A test constructor wishes to determine the appropriateness of items in his test at measuring mechanical skills. Thus, he determines five criteria which he feels define the concept of mechanical skill and measures a number of subjects on these criteria. These same subjects are given his test. What measure of relationship might he use to correlate each item with his criteria?

 Situation 5:

In the previous case, suppose the test constructor uses total scores of another measure of mechanical skill as the criterian measure of mechanical skill. What measure of correlation might he use now to determine the appropriateness of each of the items of his test at measuring mechanical skill?

 Conclusion

This concludes this very brief introduction to the study of correlation. By no means has the material presented exhausted the domain of correlations. What was presented, however, gives you a fundamental knowledge of the theory of correlations and a few methods for applications in various data situations. Attention was focused on the Product-Moment
family of correlations of which the Pearson r is the most popular. Although limited in some respects because of its assumptions, it and its three derivations prove to be quite useful where the data analysis demands some measure of variable relationship. Although not applicable to all such cases, they do find application in the majority of typical empirical analyses.

Illustrations of Learning Tenets Applied in Program Sequences

Ausubel speaks out forcibly against the popular use of small step size in many of the PI or CAI courses available today. In these cases, the result is material fragmentation and a relatively artificial presentation of the subject matter. In this respect, then, material was programmed for this course with the intention of preserving its underlying internal logic and meaningfulness. This was done by viewing the material in terms of naturally logical units. For purposes here, a logical unit is to be considered a body of information relative to a specific subject matter concept (or concepts) which explicates all or a portion of the concept (or concepts) in a meaningful fashion (i.e., meaningful as previously described). Although this strategy was utilized throughout the entire program, a brief example of the preservation of the logical unit, as was programmed, is as follows:

This last notion is a very important one. It is so because in using this general measure of variable relationship, the underlying form of the relationship between the variables does not affect its application. Which of the following would you believe best illustrates
what is meant by form of relationship. Answer A, B, or C.

A) leptokurtic vs. platykurtic
B) linear vs. curvilinear
C) negatively skewed vs. positively skewed

could B
ty Very Good! The relationship between two variables takes the form of either one of linearity or curvilinearity.

wa A
ty Kurtosis (lepto- or platy-) is an indication of shape of a distribution of scores but is not an indication of form of relationship. A relationship between variables is either linear or curvilinear.

wa C
ty Skewness (negative or positive) is an indication of shape of a distribution of scores but is not an indication of form of relationship. A relationship between variables is either linear or curvilinear.

Here, the concept of a recently developed generalized correlation coefficient is being highlighted as such -- a generally applicable coefficient. Indication of why it is generally applicable further seems to enhance the comprehension of the notion. This is done through questioning the student on what is implied by form of relationship. Through this procedure, the student will not only know why this coefficient is generally applicable (viz., no assumption of form of relationship) but will also develop an understanding of what is meant by "form of relationship" if he does not already possess this. Thus, the concept of general application
of this correlation coefficient (the logical unit) is explained in light of why it is so, a process of meaningfully relating the concept with that of form of relationship.

The principle of progressive differentiation might be considered as manifested in the actual structure of the entire program. Once again, this principle maintains that "...the most general and inclusive ideas of the discipline are presented first and are then progressively differentiated in terms of detail and specificity (Ausubel, 1967, p. 236). According to the format of the program, a correlation coefficient was developed from various attempts at determining a measure of relationship. The coefficient that was finally derived to satisfy the designated criteria of a useful correlation measure was found to be general in nature (that is, in application). From this point, conditions were placed on this coefficient until the Pearson Product-Moment Correlation was defined from it. Reducing the Pearson r under another condition produced the Point-Biserial r. One more condition was imposed on "r" and it resulted in the Phi Coefficient. Finally, Spearman's Rho was shown to be derived from the Pearson coefficient under other assumptions. Thus, all of these more specific coefficients were reduced from the Pearson r which, itself, was a reduction of the generalized correlation coefficient. In other words, all of these indices were, in turn, progressively differentiated from a general statement of statistical relationship. Because it would be impractical to illustrate the specific application of this principle here, a flow chart representing the major topical progression (relative to this principle) through the program is given in Figure 18.
Figure 18
Progressive Differentiation of the Product-Moment Family of Correlations From a Generalized Correlation Coefficient
The fundamental notions of Computer Assisted Instruction relate to Ausubel's principle of consolidation. As you recall, this learning tenet assumes very basically that new subject matter is presented only after material sequentially preceding this has been mastered. Through the interactive capabilities of the CAI system, an immediate indication of the student's achievement with respect to the subject matter just presented is readily available. Thus, decision can be made at this point whether or not to present the next topic depending, of course, on the degree of achievement exhibited by the student. One programming attempt at both determining the level of competence of the student with regard to a specific topic and also reinstructing him in the event that he is deficient in this respect so that a higher level of competence can be attained before proceeding is shown as follows:

Now, if we could equate the scales of measurement of the two variables such that they were comparable, our first problem would be solved in that the two regression coefficients would be identical in magnitude. From your statistical experience or possibly your background in measurement theory, what is one way of equating scales of measurement?

Good thinking. A basic technique in measurement to accomplish the task is to standardize the underlying raw scores. In other terms, this transforms the raw scores into another scale of measurement whose units of measurement are comparable to any other set of raw
scores which have been transformed similarly.

Good try. Let's try to get this answered in another way. Which of the following might be a way of equating scales of measurement? Answer A, B, or C.

A) By some linear transformation of the scales
B) By standardizing the underlying raw scores of the scales
C) By equating the possible ranges of scores on both scales of measurement

Correct. A basic technique in measurement to accomplish the task is to standardize the underlying raw scores. In other terms, this transforms the raw scores into another scale of measurement whose units of measurement are comparable to any other set of raw scores which have been transformed similarly.

To a point you are correct. However, more specifically, the desired way of equating scales of measurement would be to standardize the underlying raw scores of these scales, which is basically a linear transformation. In other terms, this transforms the raw scores into another scale of measurement whose units of measurement are comparable to any other set of raw scores which have been transformed similarly.

It is doubtful that this would help at all. The desired way of
equating scales of measurement would be to standardize the under-
lying raw scores of these scales, which is basically a linear tran-
formation. In other terms, this transforms the raw scores into a-
other scale of measurement whose units of measurement are comparable
to any other set of raw scores which have been transformed similarly.

Answer either A, B, or C.

In this example, the student is asked a question which requires a
specific constructed response by him. This is in contrast to a question
which requests that the student choose a response from among alternatives.
If he is incorrect, he is shown a few feasible alternative answers to the
question. Whether he chooses the proper alternative or not, he is in-
structed as to what the correct response is to the question. The fact
that a student might be allowed to proceed after either guessing the
correct answer or by getting it through a process of elimination (thus,
not really understanding this particular concept) can be dispelled by
noting that if the student answers incorrectly he is given corrective
feedback which indicates why his response was not the suitable one. Fur-
thermore, feedback is given also on correct responses noting why it is
correct. Thus, the student should be reasonably prepared to proceed to
the next topic.

The principle of integrative reconciliation basically indicates im-
portant similarities and likewise reconciles inconsistencies. In the
development of this course, an inconsistency between theoretical math-
ematical expressions and real manifestations of these (from which the ex-
pressions were developed) presented itself. This program sequence illustrating the application of the principle of integrative reconciliation is as follows:

For possible expedience in the continued development of this topic, it is now necessary to deal further with the notion of the line of regression. Recall from a previous discussion that mention was made of the line of "best fit". For purpose of the present discussion a small distinction will be made between line of "best fit" and regression line. Notice in Slide 8 that line "a" (the line based on mean Y values) passes precisely through the means of each column of Y scores. Referring to this slide again, can it be said of line "b" (that line based on the means of each row of X scores) that it passes through each row mean of X scores? Answer yes or no.

No
Correct
Yes
No it can't be said. This line does not pass through each X mean.
Answer yes or no please.
Up to this point, the manner in which the generalized correlation coefficient was determined, the manner in which the regression coefficients were determined, and the manner in which the equations 11.3 and 11.4 were expressed depended highly on the mean values of Y and the mean values of X. In fact, a strong emphasis has been placed on the fact that the lines representing the trend in the data (the regression lines) pass through the row and column means of these data.
This can be seen in equations 11.3 and 11.4 in that the left sides of these equations are given in terms of mean scores. But, as was just pointed out above, we have a regression line which does not pass through all X mean scores. Thus, it would seem that there is a discrepancy between what we have developed and real situations. Fortunately there is a solution to this. It can be shown (Ghiselli reference) and should be emphasized that the regression equations may be expressed in more general terms without affecting the validity of the generalized correlation coefficient or the validity of the regression coefficients as were mathematically developed as slopes of these equations. That is, these equations would not require that the lines pass through either column or row means respectively. These more general regression equations are given in Slide 11 as equations 11.5 and 11.6.

The basis for expressing the regression equations in this more general form is a mathematical principle known as "least squares". Lines developed under this principle requires that the average of the squared horizontal or vertical algebraic differences of points in a plot of scores to the line of best fit drawn through them be less than for any other lines that would be drawn through these points. In keeping with the initial regression line that was defined, if you have the rare case where a line may be drawn through all of the means of, say, the column scores, this would necessarily represent the line of best fit for these points without any further substantiation. This is so because of a principle you learned in your basic stat
courses. It states that minimum score deviations occur when these deviations are from the mean of those scores. Thus it may be said that the regression lines as first developed are "special cases" of the more general notion of lines of best fit.

In this illustration, the concept of regression line as had been defined and described had to be redefined and described in terms of a line of best fit. This was required to reconcile an obvious discrepancy of what was implicitly assumed by mathematical expressions and what was actually the case in their graphical manifestations.

Capitalizing on the internal logic of the subject matter is an important concept in Ausubel's position on learning. An attempt was made in this program to follow the logical sequence of topics in presenting that material necessary to fulfill the objectives of the course. As was seen in the first chapter, there are a number of avenues through which one may make use of the internal logic of the material in an attempt to meaningfully present concepts to the student. Some of these are reflected in the learning principles already demonstrated. Another was the use of concrete empirical information for illustrative intent. An example of this is contained in the following program sequence:

Notice the magnitude (.734) of the Phi coefficient relative to the Pearson r calculated on the same data (.82). By artificially dichotomizing both variables and applying Phi, we lost information. Thus, the phi coefficient, in this case, would not give a valid measure of relationship of X and Y as originally defined.

Note also, the size of Phi relative to the point biserial (.68).
The remainder of the aspects of internal logic given in Chapter 1 were neither readily nor necessarily applicable in the program or were assumed as standard instructional techniques. What is implied by the latter statement are definition of terms prior to use and use of the simplest and least technical language as possible. Thus, illustrations of the application of these aspects will not be presented.
CHAPTER III
EVALUATION OF THE CAI PROGRAM

Introduction

Content of this chapter is based on the results of the evaluation of the CAI program which was developed. Evaluation of the instructional program dealing with the product-moment family of correlational techniques is centered primarily on responses to an opinionnaire (Appendix B) completed by a carefully selected group of ten individuals after having taken the course. The responses were also discussed in detail with the evaluators. The evaluative opinionnaire was primarily focused around the content of the program in terms of choice of concepts presented, the manner in which they were presented, their instructional value to the student and the degree to which the general instructional objectives of the course were met by the CAI program. The purpose of this focus was to determine whether the program was valuable as an instructional presentation, whether the application of Ausubel's learning approach (manifested through the sequencing of the content matter had merit, and finally to identify those sections of the program which could be instructionally improved.

The criteria used in the selection of those individuals were the exposure they had to the field of statistics (e.g., number of courses, work experience with statistics, etc.) and the investigators perception as to
their comprehension of statistical concepts and their capabilities in correctly applying them. The purpose for this choice of evaluators was to determine, through their respective evaluative responses to the items in the opinionnaire, whether or not the program was sensitive to persons with diverse statistical backgrounds and more importantly, to persons with perceived differences in the level of understanding of statistical theory and application. It seemed reasonable that this should be examined since the program was designed for the student with only an initial course in descriptive and inferential statistics.

The sample of evaluators was comprised of individuals with widely differing experiences in basic statistical concepts. Some had only the introductory course in descriptive and inferential statistics while others had experienced a range of courses up to and including factor analysis. Specific courses included in this range were analysis of variance, experimental design, advanced experimental design, and correlational analysis. However, in keeping with the intent of a differential analysis of the responses of the opinionnaire based on the evaluator's level of understanding of fundamental statistical concepts, prime emphasis in the analysis was placed on the investigator's perception of their level of comprehension of statistical concepts rather than on the number of courses in statistics they had completed. Thus, founded on the investigator's familiarity with the statistical knowledge and abilities of these individuals, they were placed into three groups: 1) those with a low level of understanding; 2) those with a medium level of understanding; and 3) those with a high level of understanding. Four evaluators were in each of the
low and medium r with the remaining two evaluators categorized as having a high level of understanding of statistical concepts.

Analysis of Responses

Results of the analysis of responses to items of the opinionnaire follow:

Item 2: How would you rate the choice of concepts presented in the program? That is, considering the objective of presenting the four product-moment family correlations, how would you rate their development in terms of the specific concepts used in the program?

Very Poor Poor Average Good Very Good

This item attempted to determine the extent to which the particular conceptual development utilized in the definition of the various correlations was of value. It was considered valuable to determine this since most presentations of correlational analysis do not include a theoretical development based on many of the concepts that appear in this program. The choice of concepts used in the program was rated a being Good by all but one of the evaluators. The rating of that person, a member of the medium level of understanding group, was Average. It is felt that the reason behind the rating of the majority of evaluators on this item lies in the fact that this conceptual approach to correlation was novel to them, and possibly because it held promise as a viable means of presenting the topics of correlation.

Item 3: How would you rate the logical sequencing of these concepts in this program?

Very Poor Poor Average Good Very Good
The intent of this item was to determine if the evaluator felt that each succeeding concept flowed from previously presented ones in a logical and meaningful fashion. In Ausubel's terms, the item attempted to assess whether the CAI program made efficient application of the principle of internal logic of the subject matter. The responses of the evaluators ranged from Poor to Very Good, the majority of the ratings being Very Good. Those who felt the conceptual sequencing was Very Good came from both the medium and high level of understanding categories of individuals, while the other ratings were made by persons representing each of the three groupings of evaluators. The mode of these responses was Very Good while only one of the ratings was Poor which was given by an evaluator from the low level of understanding category. The median response to this item fell between Good and Very Good.

It would appear that the sequencing of concepts as presented in the course had no appeal for any particular group of evaluators, but did appeal in a relative sense, to the person's unique experiences in and understanding of statistics.

Item 4: In general, how would you rate the concepts presented in terms of their difficulty?

Very Easy Easy Average Difficult Very Difficult

The purpose of this item was to assess the absolute difficulty of the concepts given in the program as judged by the evaluators. Responses to the item showed that the evaluators viewed the concepts as falling into the Difficult-Very Difficult range, the median and mode being Difficult. However, one individual (from the medium level of understanding group)
rated the concepts as being only of Average difficulty.

Ratings of this item did not seem to be related to the ratings of the previous item (the sequencing of concepts) in that the concepts were assessed as being Difficult or Very Difficult regardless of how the evaluators felt they were sequenced.

**Item 5:** As the concepts were presented in the program, how would you rate your difficulty in understanding them?

Very Much Much Average Little Very Little

This item attempted to determine, in part, the effectiveness of the presentation. Analysis of the responses found the median rating to be Average difficulty and the mode response to be Much difficulty. However, a fairly definite pattern appeared with regard to the ratings from the three groups of evaluators. All individuals considered as having a high level of understanding of basic statistical concepts felt that they had Little difficulty in comprehending the inherent meaning of the concepts as offered in the program. On the other hand, those individuals who were classified as having a low level of understanding of basic statistical principles had Much difficulty in understanding the concepts. Those evaluators in the medium understanding group split their ratings between Little and Much difficulty. Finally, it may be interesting to note that there appeared to be a tentative pattern between those persons who felt that the sequencing of concepts was Good-Very Good and their ease in understanding these concepts.

**Item 6:** How would you rate the clarity of presentation of the various concepts in the development of the correlations?

Very Poor Poor Average Good Very Good
The modal judgement by the evaluators was that the clarity of presentation of the concepts was Good. One evaluator (from the low level of understanding category) did feel that the clarity was Poor while another (from the medium level of understanding group) felt that the concepts were presented with Average clarity. Again, there appears to be a slight trend in the responses made to this item and other items in the opinionnaire, specifically those made to the previous two items. Those who thought the sequencing was Good to Very Good and had Little difficulty in understanding the concepts also felt that the clarity of the conceptual presentation was Good. The individual who rated the clarity of the concepts as being Poor likewise felt the sequencing to be Poor and his understanding of these as Little.

Item 8: As a result of taking this CAI course, how would you rate your understanding of those concepts with which the program dealt?

Very Poor Poor Average Good Very Good

The rationale underlying this item was the attempt to determine, in part, whether the program was effective in fulfilling its instructional purpose. The median rating made by the evaluators on this item was Good. The modal rating of this item was Good.

Once again, there seemed to be a pattern emerging with respect to the responses of the evaluators, both in relation to the respective classifications of the individuals and in relation to other items in the opinionnaire. Those who were judged to have a medium or high level of understanding of statistical concepts considered their understanding of correlational concepts as being Good to Very Good by reason of the CAI
program. Those evaluators were classified as having a low level of understanding of statistical concepts felt that their comprehension of those correlational concepts treated in the program was only Poor to Average. Furthermore, the total sample of evaluators had impressions consistent with those they had with reference to the sequencing of concepts through the course, their understanding of these concepts, and the clarity of the concepts.

In general, then, the course appeared to be somewhat effective in conveying correlational concepts to the "better" students, while rather less effective in this respect with those persons judged low in their comprehension of general statistical concepts.

Item 9: In general, how would you rate the instructional value of this CAI program?

Very Poor Poor Average Good Very Good

Responses to this item were almost identical with those of Item 8. The only discrepancies were found in the high level of understanding group of individuals. Their ratings fell from Good and Very Good to Average and Good. A possible explanation for this might lie in the fact that these persons saw the contents of the program as only a review of material and of no particular value in increasing their knowledge of this area of statistics. The other two classes of individuals rated the instructional value of the program as being anywhere from Poor to Very Good. Those who derived Little or Average understanding from the concepts presented thought the same of the program's instructional value, while those who derived a Good understanding of these concepts as a result of the program had a Good opinion of the value of the CAI course.
The remainder of the items, with the exception of Item 12, was open-ended and was directed to the determination of which of the concepts presented lacked clarity, how the subject matter might be best presented (traditional lecture fashion, via CAI, or some combination of both), and where and how the program should be revised if revision was felt necessary. Specifically, those concepts or sections of the course that were thought to lack clarity (by one or more evaluators) were as follows: the sequence on covariance, the presentation on partial variance, that section on the development of the generalized correlation coefficient, and the discussion of regression and regression lines. All of these concepts were included in the theoretical development of correlation. No comments were rendered with regard to the more practical or applied aspects of the four product-moment correlations.

In an attempt to assess the general attitude of the evaluators with respect to how the topics outlined in this CAI course might be best presented, the respondents were asked whether this material would be provided better by lecture, by CAI, or by some combinations of the two. All but one evaluator thought that a combination of the two methods would be best. The general rationale for their choice was that each, in its own right, has merit and that a combination of the two could theoretically capitalize upon these positive aspects. The only evaluator who did not feel that a combination of the two would be best thought that this material is best presented via CAI. His reason for the choice was that the calculation capabilities of the computer could be used in an efficient way to aid the student's understanding of the concepts. Implicit in his comment is the
desire to have the student perform calculations pertaining to the practical application of the presented concepts based on random data generated by the computer. As an example, assuming the computer system capability to calculate a Pearson r based on supplied data, the student could envision immediately the degree to which the maximum value of r is limited by disparate data distributions generated by the computer.

All evaluators indicated the need for the revision of the CAI course. Aside from the areas of the program that were mentioned as being unclear, it was felt that revision was necessary for other various reasons. One of these was the "wordiness" of certain sections of the course. Others were the need for the presentation of more problems in the body of the program, the desire for the capability of optionally repeating program sequences, and the need to give students more practice in making more calculations.

Finally, Item 12 questioned the degree to which the evaluators felt the six broad instructional objectives of the CAI course were met. The objectives dealing with the theoretical development of the generalized correlation coefficient and the presentation of some fundamental concepts of linear regression as they relate to correlation received the lowest ratings in terms of the degree to which they were met. On the other hand, the evaluators felt that those objectives referring to the definition of the product-moment correlations and the relationships between them were met to a large extent. The objective concerning the applications of these correlations was considered met to a limited degree. Responses to this indicated both the need to include more problems throughout the program and also the need for more hypothetical situations to which the correla-
An analysis of the responses to items of the opinionnaire appears to indicate that the program was moderately effective in presenting the subject matter for those evaluators who were classified as having a high level of understanding of statistical concepts. Conversely, the program seemed to be less effective in this respect for those persons judged to have a low level of understanding of statistical theory. The medium group of evaluators seemed to fall into one or the other of these designations.

In designing and programming the course, the author assumed that the specific concepts presented and the manner in which they were sequenced would be meaningful to the individual who has had only a fundamental course in descriptive and inferential statistics. Allowing for a presumably basic understanding of the fundamental topics presented in such a course, it was further assumed that such a person would be capable of deriving the desired meaning of each of the concepts developed in the program and would likewise be capable of forming the desired meaningful relationships between these concepts. An analysis of individual differences according to the classifications of evaluators indicated that the CAI program did not achieve this purpose. At its present stage of development, the program would appear to be instructionally beneficial primarily for those persons with a relatively high level of understanding of a range of statistical concepts.

From the opinionnaire analysis, a number of broad but fundamental concepts were identified as being unclearly presented and, thus, assumed to
have been less-than-adequately understood by the majority of individuals in the low or medium level of understanding groups. The concepts which were identified as being unclear included those of covariance, partial variance, the generalized correlation coefficient, regression and regression lines. It is suggested, then, that the presentations of these concepts be subdivided with more of an emphasis on those specific bodies of sub-concepts which serve to define each of the larger concepts. It is hoped that this would make a greater contribution both to the absolute meaningfulness of the respective concepts and also to the relative meaningfulness of these concepts, especially for the low and medium levels of understanding individuals.

An example of what is implied in this process might be helpful. The concept of partial variance was developed in a sequential fashion based largely on the use of a scatterplot of scores given in one of the visuals accompanying the course (refer to pages 127 to 135 of Appendix A for this sequence of the program). As was programmed, one portion of the development of this concept required the student to make fairly large relationships between points in the plot of scores to their representations as deviation scores in given distributions of two types of deviation scores—one based on deviations from an overall mean value and the other based on deviation from respective column means of scores. It is proposed that this portion of the program be modified such that the students would develop the two distributions of deviation scores in a sequential fashion by actually computing a number of the respective deviation scores which comprise the distributions. In this way it is hoped that the student
would achieve a better understanding of the two types of deviation scores, two sub-concepts which are utilized in a fundamental way in the further development of the inclusive concept of partial variance.

From a theoretical standpoint, what is basically underlying the proposed procedure of "breaking down" the inclusive, major concepts of the program into smaller sub-concepts is the attempt to better relate the more general concepts to the person's existing cognitive structure by means of a more precise sequential development of the foundations of these concepts. That is, by establishing the components of a concept and further relating these in a logically sequential manner to already existing "anchor" concepts in the student's cognitive framework, a consequence should be a greater comprehension of the inclusive concept. It is felt that this proposed procedure is mandatory for those students having a low or medium level of understanding of basic statistical concepts since both the relationship between the components of a concept and the relationship to cognitive structure apparently were not made. In summary and in simple terms, major concepts must be redeveloped through the use of their defined sub-concepts, presenting these in a fashion which is more familiar and thus meaningful to the student.

With respect to the theoretical framework upon which the program was developed, it should be pointed out that it is the belief of the investigator that application of this learning approach is beneficial. Principles such as the use of the internal logic of the subject matter, that of progressive differentiation, and that of integrative reconciliation were, in principle, thought to have been applied properly. Illustrations of how
these were applied were given in Chapter II. However, what appears to have manifested itself from the analysis of responses to those items of the opinionnaire dealing with the conceptual development of the program (Items 2 through 6) is the implication for the probable need of a differential application of these principles. What is assumed in the differential application of the principles is the development of a concept according to various strategies, each exhibiting varying degrees of sophistication and complexity in arriving at an identical end—the meaningful presentation of the concept. The basis for this differential application of the principles obviously must be specific characteristics of the learner.

Several comments were made by the evaluators concerning the course in general. One was that portions of the program were too "wordy". Implicit in this comment are both the length of continuous narrative used in illustrating certain points (e.g., the discussion of regression lines and their relation to correlation, the discussion on the possible lack of generality of the correlation coefficient, etc.) and also the style of writing. Both of these inadequacies must be dealt with. The first actually reflects a restricted use of the interactive capabilities of a CAI system. At least in part, it also reflects the need to revise these sections on the basis of the suggestion made previously with regard to the subdivision of inclusive concepts, sequentially developing them in terms of their specific components. In doing so, moreover, it is felt that a change in the author's style of writing would be a consequence.

Another comment of the evaluators was that the student should be given the option to repeat sections of the program if so desired. Although, it
is felt that the option of repetition should be employed only after revision(s) of the program and further evaluation(s) of the revision(s) dictate its inclusion. A final general comment was offered that concerned the evaluators' desire that a greater degree of application should be made of the correlations presented. In evaluation of this comment, the writer would tend to extrapolate from it and suggest that the revision(s) include a greater degree of application of the major concepts of the course themselves. It is hoped that this would tend to further solidify their understanding.

Finally, the investigator is in agreement with the majority position that the subject matter of the program would be best presented by some combination of both lecture and CAI. The question of best method of material presentation is a very important one and also a very valid one. However, definite judgement as to which of the two (lecture vs. CAI) or some combination of the two would be most effective cannot be made at this point. It remains open for investigation, not only in a general sense, but also within the bounds of the best form of presentation of each of the specific concepts appearing in this particular CAI program.

It should be pointed out that the major limitations of the analysis of the evaluative responses were the absence of reliability and validity measures of the instrument (opinionnaire), the relatively small number of evaluators (10), and the method of classification of the evaluators into groups with perceived levels of understanding of fundamental statistical concepts. With regard to the first limitation, it is the opinion of the investigator that for the type of information required on which
to base the evaluation of this CAI course for the purpose of future revision, such measures of reliability and validity of the instrument were not necessary. Secondly, since the evaluators were carefully selected, their relatively small number would not seem to discredit the value of the information gathered through their evaluations. It is realized, though, that additional comments by evaluators other than those used in this study possibly could supply the investigator with information not revealed in the present evaluation. Lastly, in reference to the manner of categorization of the evaluators, it is the belief of the writer that other means of classifying these individuals would have resulted in similar groupings. Thus, consistent with the intent for a general evaluation of this CAI program to determine where its strengths and weaknesses lie, it was felt that this evaluation elicited suitably valid information upon which subsequent revision can be based despite these limitations.

Recommendations

From the evaluation and discussion of the results of the opinionnaire analysis, the direction of future revision and research on this CAI program is clearly dictated. It is assumed that revision of and research on this program will be complementary activities. As implied in the discussion section, recommendations for further research are found along two major dimensions. The first is concerned with the entering statistical characteristics (abilities) of the learner and the appropriate sequencing of course content responsive to these specific characteristics. The second area of needed research is concerned with the determination of those sections of the program which might be most effectively presented via
CAI and those which might be better presented using other instructional techniques or devices.

Based on the responses of evaluators, the necessity for course revision is indicated. This fact, moreover, is only keeping with the general scheme of the developmental process—continuous looping from evaluation back to the requirements, design, or production phases. Here, this process must be realized in the last two of these phases. The requirement that the student-user have a basic understanding of those concepts given in a course on descriptive and inferential statistics should be retained as is. However, this necessarily dictates the redesign and re-production of the program, again, based on information supplied from the evaluation phase. Where, why and how this is to be accomplished follows.

Because a theoretical introduction of the general concept of correlation is assumed to be a necessity in a course in the product-moment family of correlations and since the great majority of individuals reviewing this course had a great deal of difficulty with a relative lack of understanding of the specific concepts presented in this theoretical development, attention must be focused primarily on the revision of these particular concepts in order to satisfy the instructional objectives of the requirements phase related to this.

The first consideration for future revision must be directed to research on the learner variable. It is recommended that research be undertaken to identify those entering statistical characteristics of that population of potential users of the program which would be beneficial
in more specifically defining the level of understanding each student possesses with regard to fundamental statistical concepts which are perceived as being important for the future understanding of the concepts contained in the CAI course. Another characteristic of the learner that may be relevant for these purposes is the mathematical reasoning ability of the student--whether the student learns better from an inductive or a deductive approach to the subject matter. It is also proposed that this research be further directed to the development of an instrument which would validly and reliably measure the degree to which individuals possess these characteristics.

Complementing this research on the learner variable, an effort must be made to develop various sequential approaches to the development of those concepts of the program identified as being little understood. By varying sequential approaches (or multi-tracked approaches) the writer is implying the development of additional presentations of the same concepts which now appear in the program. These additional presentations would differ in degree of complexity. The conceptual development as it is currently presented is considered to be of a higher degree of complexity. Lower degrees of complexity of the program sequences will be defined by the degree to which the present conceptual sequences would be subdivided into the explicit development of the sub-concepts which comprise them. In keeping with the learning approach taken initially, the development of these additional sequences should be done within the framework of Ausubel's learning approach.

A last stage of research in this process must be concerned with the determination of that level of program sequence which is most appropriate
to produce maximum understanding for the learner with certain defined characteristics. Furthermore, a factor which would influence the choice of instructional strategies to be employed in this redevelopmental process would be the specific characteristics of the learner for whom a certain sequence would be most beneficial. Thus, without doubt, a great deal of interplay will be necessary between the three stages of this research process to arrive at the optimum "fit" between entering knowledge, specific sequence to be presented, and the instructional strategies involved in the specific sequences.

The process described above is not necessarily new to CAI or to instructional research in general. However, the "prediction" of the best sequencing of subject matter is not considered to be generalizable at this point by the writer. Therefore, investigation into this problem must be carried out within the bounds of a specific content area, in this case the theoretical approach to a particular body of correlational techniques.

The second domain of recommended research is concerned with the determination of those instructional means through which this subject matter of interest may be presented most effectively. The previously recommended schedule of research assumed presentation solely by means of CAI. However, once this research is completed, further research is proposed to determine which portions of the CAI presentation may be more effectively delivered through traditional lecture or through other forms of educational technology. However, at this point, it is hypothesized by the writer that the attempt at greater individualization of instruction implicit within the previously outlined research (and thus greater under-
standing of this material) would be better realized solely in the use of CAI. Again, this is subject to confirmation through this proposed research.

A final set of recommendations concerns the use of auxiliary equipment in the presentation of selected course material and the need for expansion of the capabilities of the particular CAI system utilized. Some of the material of the course could not be presented via the IBM 1050 terminal and thus had to be shown through the use of slides. Most of this material dealt with equational derivations. Thus, it is suggested that future utilization be made of a cathode ray tube (CRT) display facility to more effectively present this information in an actual sequential sense. With regard to the expansion of the capabilities of the CAI system utilized, it is recommended that system functions be written that would allow greater flexibility in the application of the specific concepts that appear in the program. This would include the programming of functions for the calculation of the four presented correlations, a function which would generate random data on which to base the simple calculations of these coefficients and other terms that are given in the course (e.g., covariance, partial variance, etc.), and also a function which would generate data according to certain specified conditions (e.g., data representing a positively or negatively skewed distribution of scores). With these additional capabilities, it is hoped that the student would derive a greater appreciation for the various ramifications that differing sets of data can have on the application of these correlations, aside from the experience of directly applying other statistical terms to varying
The purpose of this developmental study was to produce a Computer Assisted Instruction program in the product-moment family of correlations. Four phases were seen as comprising this developmental effort. The first was the definition of the requirements of the developmental process which serve as guidelines for this process. Here, these comprised the instructional objectives of the CAI course, the fact that it be one concerned with the four product-moment family correlations, and that the program be developed according to the cognitive learning approach described by Ausubel.

The second phase, design, dealt with the compilation of the specific topics to be treated in the program and the expression of these in narrative form. The production phase of the developmental process was concerned with the transformation of the narrative of the design stage into an interactive dialogue characteristic of CAI. It also included the coding (COURSEWRITER III language) and inputting of this material into an IBM 360/40 computer by means of an IBM 1050 teletypewriter terminal. The last developmental phase, evaluation, in general, was charged with the determination of the conceptual inadequacies of the program and its effectiveness in imparting the meaningfulness of the concepts presented. Redevelopment of the program will be founded on this basis.

Ten student-evaluators reviewed the CAI course and completed an evaluative opinionnaire consisting of items reflecting information needed for the future revision(s) of the program. These student-evaluators were purposefully selected because of their varying backgrounds in, and under-
standing of, statistical theory, as perceived by the writer. This sample was divided into three classes, depending on their adjudged level of understanding of fundamental statistical concepts. The responses of these persons to items in the opinionnaire were thus analyzed differentially, with the classification of the evaluator as the most important variable.

A synthesis of results of the opinionnaire analysis directed themselves to the implication that the CAI program was ineffective for the lesser two classes of individuals. Aside from this, certain of the concepts in the program were identified as being unclearly presented. Recommendations for further research inferred from these results included 1) need for the identification of specific statistical characteristics of the learner, and 2) the addition of several other presentations of the same concepts that appear in the course at present. The proposed approach in "redeveloping" these concepts is to focus more attention on the components or sub-concepts of the major concepts, developing them in a more precise fashion on this basis with each additional sequencing of the subject matter reflecting a different level of complexity of presentation. It is thus hypothesized that these concepts will become more meaningful to the student regardless of his entering understanding of basic statistical concepts (that which will serve, in part, to define the learner's statistical characteristics). Of course, this necessarily implies a third recommendation for further research, that being identification of the specific pairings of entering characteristics of the learner and the program sequence which would best facilitate the meaningful learning of the concepts on the CAI course.
Finally, it must be mentioned that it was felt that the learning approach taken was a valuable one, but only so for those student-evaluators classified as having a high level of understanding of statistical theory. Thus, the application of the principles of Ausubel's learning approach will be made likewise to the additional program sequences to be developed, assuming that this application to the redeveloped sequences would be as beneficial to the specific learners who would experience them as the initial application was to the "better" student-evaluators of the CAI program.
APPENDIX A

Computer Assisted Instructional Program
INTRODUCTION TO THE FAMILY OF GRETL TIPS

WELCOME TO COMPUTER ASSISTED INSTRUCTION (CAI). THE PRIMARY OBJECTIVE
OF THIS SIMULATION IS TO INCREASE YOUR UNDERSTANDING OF THE
BASIC CORRELATIONAL TECHNIQUES. WE WILL BE FOCUSING ON THE
APPLICATION OF THESE TOOLS.

THE GENERAL INSTRUCTIONAL OBJECTIVES THAT WILL SERVE AS THE FRAMEWORK ARE:

1. THE STUDENT WILL BE CAPABLE OF IDENTIFYING THE PURPOSE OF CAI.
   [IMPROVE PERCEPTIONS OF THE CRITICAL ROLE OF CAI]
2. THE STUDENT WILL BE CAPABLE OF APPLIQUING A COMPREHENSIVE VIEW OF THE
   DEPENDENT AND INDEPENDENT VARIABLES AND THE ROLE OF CAI.
   [DEVELOP A GLOBAL AND COMPLETE UNDERSTANDING OF THE BROAD ROLE OF CAI]
3. THE STUDENT WILL BE CAPABLE OF INTEGRATING THE CONCEPTS AS THEY RELATE TO THE
   THEORETICAL CORRELATION FACTOR.
4. THE STUDENT WILL BE CAPABLE OF IDENTIFYING THE OBLIGE TOWARDS THE
   DEVELOPMENT OF THE CRITICAL FACTORS OF CAI.
   [INCREASE UNDERSTANDING OF THE CRITICAL FACTORS]
5. THE STUDENT WILL BE CAPABLE OF IDENTIFYING THE OBLIGE TOWARDS THE
   DEVELOPMENT OF THE CRITICAL FACTORS OF CAI.
   [INCREASE UNDERSTANDING OF THE CRITICAL FACTORS]
6. THE STUDENT WILL BE CAPABLE OF EXPLORING THE EXPERIMENTAL APPLICATIONS OF CAI.
   [EXPLORE THE EXPERIMENTAL APPLICATIONS]
7. THE STUDENT WILL BE CAPABLE OF UNDERSTANDING THE CRITICAL FACTORS.
   [UNDERSTAND THE CRITICAL FACTORS]
CORRELATION COEFFICIENT INDICATES, IN A GENERAL SENSE, THE DEGREE TO WHICH DATA ON DIFFERENT VARIABLES RELATE TO EACH OTHER. IF DATA ON TWO VARIABLES ARE RELATED, THIS IMPLIES THAT WHAT THESE DATA MEASURE IN THE RESPECTIVE VARIABLES HAVE SOMETHING IN COMMON.

ON THE STUDY OF CORRELATION CAN BE PLACED INTO TWO SEPARATE DOMAINS ACCORDING TO THE ULTIMATE INTENT OF THE USES. THE STATISTICAL NATURE OF THE CORRELATION COEFFICIENT IS IDENTICAL IN BOTH CASES, HOWEVER, OF THE TWO WHICH TWO BEST REPRESENT THE USES TO WHICH A CORRELATION MAY BE PUT?

Choose two answers.

A. THE ATTEMPT TO DETERMINE THE AMOUNT OF COMMON VARIANCE BETWEEN TWO VARIABLES.
B. THE PREDICTION OF DATA ON ONE VARIABLE FROM DATA ON ANOTHER.
C. THE ATTEMPT TO DETERMINE THE EXTENT TO WHICH INDIVIDUAL DIFFERENCES ON THE MEASURES OF INTEREST ARE DUE TO SIMILAR UNDERLYING FACTORS.
D. THE ATTEMPT TO ESTABLISH A CAUSE AND EFFECT RELATIONSHIP BETWEEN FACTORS UNDERLYING TWO VARIABLES.

CA(2) CA(3) CA(4) CA(5)
TY. GOOD THINKING!!!
WA(2) AB AC CA(4) CA(5)

TY. YOU ARE PARTIALLY CORRECT. THE AMOUNT OF COMMON VARIANCE BETWEEN TWO VARIABLES IS IMPORTANT TO THE STUDY OF CORRELATIONS, AS WILL BE SEEN SHORTLY. HOWEVER, IT IS NOT ACTUALLY A MAJOR PURPOSE IN FINDING A CORRELATION COEFFICIENT. TRY AGAIN. CHOOSE THE (2) ANSWERS.

WA(2) AB OR CA(3) CA(4)

TY. TO DETERMINE A CAUSE-EFFECT RELATIONSHIP ON THE BASIS OF A CORRELATION IS TO TREAD ON TYPICALLY UNCERTAIN GROUND. TO REPEAT, A CORRELATION MEASURES ONLY A RELATIONSHIP BETWEEN MEASURES ON A PAIR OF VARIABLES. IN ORDER TO UN-equivocally DETERMINE A CAUSE-EFFECT SITUATION BY MEANS OF A CORRELATION, PERFECT EXPERIMENTAL CONTROL OF FACTORS AFFECTING OTHER VARIABLES IS NECESSARY. IN THIS WAY IT CAN ONLY BE KNOWN THAT THE EFFECT WAS DUE TO A CERTAIN CAUSE AND NOT DUE TO ANOTHER OR OTHERS. VERY RARELY IS THIS THE CASE.

HOWEVER, TRY AGAIN. CHOOSE THE (2) ANSWERS.

WA(2) AB OR CA(3)

TY. BOTH OF YOUR ANSWERS ARE WRONG.
VARIANCE WITHIN THE THREE VARIANCES. IF YOU A SAVVY, WHAT IS THE NAME GIVEN TO THIS CONCEPT OF COVARIA

VARIANCE _______________ COVARIANCE VARIABLES? ______________

CA: NO

TY > CORRECT? *<______________>

UN > PLEASE REREAD THE QUESTION, AND ANSWER AGAIN? ______________

UN > THE ANSWER IS COVARIANCE, ______________

TY > COVARIANCE IS A POSSIBLE MEASUREMENT RELATIONSHIPS... ______________

JUST AS WE CAN DETECT THE VARIANCE A SET OF SCORES ON ONE VARIABLE, ______________

WE CAN LIKEWISE FIND THE AMOUNT OF COMMON VARIANCE BETWEEN TWO SETS... ______________

SCORES REPRESENT THE TWO VARIABLES. THEREFORE, IF SUBJECTS HAVE SCORED, SAY, C... ______________

TWO TESTS, AND IT IS HOW THESE SCORES VARY ONE OF THE TESTS RELATIVE... ______________

TY HOW THEY VARY ON THE SECOND IS GIVEN THE MEASURE: COVARIANCE, ______________

PY 53 |

TY > REFER TO THIS TIME TO SLIDE 1., HERE TWO SETS OF DATA 18 STUDENTS ARE PRESENTED, ______________

PA 101 |

TY THE SUM MEAN (X), VARIANCE (VAR) THE DEVIATIONAL SQUARE OF EACH SET OF SCORES... ______________

MAY BE FOUND IN THE USUAL MANNER AND ARE GIVEN ELABORATE TO THE SIDE... ______________

EACH COLUMN OF DATA, WHERE IS IT POSSIBLE TO DETERMINE A MEASURE RELATIVE FUNCTION... ______________

SCORES ON SET 1 VARY IN RELATION TO ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ________________ ______________

CA YES

TY > RIGHT, IT SURE IS, ________________

PA 52 |

CA NO |

TY > WITH THE AVAILABLE DATA GIVEN, IF SLIDE 2., IT IS POSSIBLE TO COMPUTE A... ________________

MEASURE OF COVARIANCE, ________________

PA OK |

UN > TYPE: YES OR NO ________________

INTR |

TY > ON YOUR STAT BACKGROUND, NOTED, THE ACTUAL GIVEN IS SLIDE 1., WHICH IF TRUE... ________________

> FOLLOWING REPRESENT THE COVARIANCE IF THESE SETS DATA? USE THE "CALC... ________________

FUNCTION IF NECESSARY, TYPE A, B, C, D AS YOUR ANSWER. ________________

> 【A, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【B, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【C, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【D, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【E, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【F, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【G, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【H, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【I, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【J, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【K, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【L, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【M, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【N, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【O, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【P, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【Q, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【R, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【S, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【T, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【U, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【V, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【W, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【X, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【Y, 3, 4, 1 ________________

> 【0, 17, 23 ________________

> 【Z, 3, 4, 1 ________________

> 【0, 17, 23 ________________
Slide 1

Computation of the Covariance
For Variables 1 and 2

<table>
<thead>
<tr>
<th>SS</th>
<th>1</th>
<th>2</th>
<th>DEV1</th>
<th>DEV2</th>
<th>PRODDEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>70</td>
<td>75</td>
<td>3.67</td>
<td>5.00</td>
<td>18.35</td>
</tr>
<tr>
<td>2</td>
<td>67</td>
<td>72</td>
<td>0.67</td>
<td>2.00</td>
<td>1.34</td>
</tr>
<tr>
<td>3</td>
<td>70</td>
<td>75</td>
<td>3.67</td>
<td>5.00</td>
<td>18.35</td>
</tr>
<tr>
<td>4</td>
<td>71</td>
<td>71</td>
<td>4.67</td>
<td>1.00</td>
<td>4.67</td>
</tr>
<tr>
<td>5</td>
<td>67</td>
<td>70</td>
<td>0.67</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>64</td>
<td>68</td>
<td>-2.33</td>
<td>-2.00</td>
<td>4.06</td>
</tr>
<tr>
<td>7</td>
<td>71</td>
<td>72</td>
<td>4.67</td>
<td>2.00</td>
<td>9.34</td>
</tr>
<tr>
<td>8</td>
<td>63</td>
<td>67</td>
<td>-3.33</td>
<td>-3.00</td>
<td>9.99</td>
</tr>
<tr>
<td>9</td>
<td>65</td>
<td>67</td>
<td>-1.33</td>
<td>-3.00</td>
<td>3.99</td>
</tr>
<tr>
<td>10</td>
<td>65</td>
<td>68</td>
<td>-1.33</td>
<td>-2.00</td>
<td>2.66</td>
</tr>
<tr>
<td>11</td>
<td>65</td>
<td>68</td>
<td>-1.33</td>
<td>-2.00</td>
<td>2.66</td>
</tr>
<tr>
<td>12</td>
<td>65</td>
<td>71</td>
<td>-1.33</td>
<td>1.00</td>
<td>-1.33</td>
</tr>
<tr>
<td>13</td>
<td>66</td>
<td>68</td>
<td>-0.33</td>
<td>-2.00</td>
<td>0.66</td>
</tr>
<tr>
<td>14</td>
<td>65</td>
<td>71</td>
<td>-1.33</td>
<td>1.00</td>
<td>-1.33</td>
</tr>
<tr>
<td>15</td>
<td>61</td>
<td>62</td>
<td>-5.33</td>
<td>-8.00</td>
<td>42.64</td>
</tr>
</tbody>
</table>

SUM 995 1050

\( \bar{x} \) 66.33 70.00

VAR 11.58 12.93
I... 

TY >-CORRECT*<<<<

HR COVAR 1

WAIT] A B N

TY >-YOUR ANSWER INCORRECTLY. WOULD YOU LIKE TO REVIEW THE COMPUTATIONAL DEFINITION?<<<<

OF COVARIANCE OR WOULD YOU LIKE TO RECOMPUTE IT FOR THIS EXAMPLE?<<<<

(TYPE YES FOR THE FIRST OR NO FOR THE LATTER)<<<<

BR QUEST 1

UN TYPE EITHER A, B, C, OR D<<<<

QUEST 2

QU 1

CA YES

BR COVAR 1

CA NO

BR INTR 1

UN >>TYPE: YES OR NO<<<<

COVAR 1

QU >>JUST AS THE VARIANCES OF THE TWO SETS WERE COMPUTED AS THE SUMS OF THE SQUARES OF THE DEVIATIONS FROM THEIR RESPECTIVE MEANS DIVIDED BY THE NUMBER OF DEVIATIONS,<<<<

IF ONE WERE TO MULTIPLY THE DEVIATION SCORES ON SET 1 BY THEIR COU TERS<<<<

PARTS ON SET 2, SUM THESE PRODUCTS AND DIVIDE BY THE NUMBER OF PRODUCTS,<<<<

THE RESULT WOULD BE WHAT IS DEFINED TO BE THE COVARIANCE OF SET 1 WITH <<<<<<

SET 2 (COV 1,2)<<<<

THE COVARIANCE CAN BE REPRESENTED SYMBOLICALLY AS SUMP-ODV/1 WHERE SUMP-ODV<<<<

REFERS TO THE SUM OF THE PRODUCTS OF THE RESPECTIVE DEVIATIONS FROM SET 1<<<<

AND SET 2 AND N IS THE NUMBER OF PRODUCTS SUMMED. FROM THE EXAMPLE, WHAT ARE THE <<<<<<

VALUES OF SUMP-ODV AND N, RESPECTIVELY?<<<<

CAW] 8160,158<<<<

TY >>GOOD*<<<<

CAW] 8158,1406<<<<

TY >>YOU'VE GIVEN THE CORRECT ANSWERS BUT IN REVERSE ORDER. SUMP-ODV IS <<<<<<

140 AND N IS 1<<<<

WALL] 81426}
TY >> YOU GOT PART OF IT. NOW WHAT IS THE VALUE OF S1?<-----
    THIS, TYPE BOTH THE VALUES OF S1 AND S2, RESPECTIVELY.<--------
TY >> (L) ELSE
TY >> YOU GOT PART OF IT. NOW WHAT IS THE VALUE OF S1?<-----
    YOU DETERMINE THIS TYPE WITH THE VALUES OF S1 AND S2, RESPECTIVELY.<--------
TY >> (L)
TY >> NOW USING THE "CALC" FUNCTION IN THE I/O PANEL, THE VALUES OF S1 AND S2, RESPECTIVELY.<--------
    AFTER YOU HAVE CALCULATED IT, TYPE THIS.<--------
TY >> (L) ?<------
TY >> RIGHT<--------
TY >> THE COVARIANCE HERE IS ?<------

COVARS<------
OU >> BECAUSE OF THE NATURE OF THE COVARIANCE TERM AS AS DESCRIBED, IT IS A PHYSICAL<------
    IMPRESSION THAT WE HAVE TO DETERMINE THE MEAN VALUE OF A CORRELATION COEFFICIENT<------
    HERE AND USE THE COVARIANCE AS A SATISFACTORY MEASURE OF CORRELATION.<------
    WOULD YOU LIKE TO TERMINATE THE DISCUSSION HERE AND USE THE COVARIANCE AS A<------
    CORRELATION COEFFICIENT?<--------
OU >> NO<------
TY >> RIGHT<--------
OU >> (L) YES<------
TY >> (L) CHOICE<------
OU >> IF YOU ACCEPT THE COVARIANCE TERM AS A MEASURE OF CORRELATION, IT WOULD BE INAPPROPRIATE.<--------
    TYPE YES OR NO.<------
OU >> COVEXP<------
OU >> THE COVARIANCE TERM IS NOT AN APPROPRIATE MEASURE OF CORRELATION. IT IS<------
    NOT FOR AT LEAST TWO REASONS. THE MEAN VALUE OF THE VARIABLES IS<------
    SHALL WE SAY, A RELATIVELY "INSTABLE" MEASURE OF RELATION. THAT IS, IT<------
    CAN TAKE ON ANY VALUE IN THE REAL NUMBER LINE (POSITIVE OR NEGATIVE)<------
    LIMITED ONLY BY THE RANGE OF VALUES MENTIONED BY THE SCALES.<------
    MEASUREMENT ASSOCIATED WITH THE VARIABLES IS IN WHICH THEY ARE COMPUTED. THUS,<------
    THE TERM WOULD FLUCTUATE IN VALUE DUE TO FLUCTUATION IN THE<------
    CHANGES IN SCALES OF MEASUREMENT.
Slide 2

Scatterplot of Raw Scores for Two Variables Reflecting a Linear Relationship

\[ \bar{X}_1, \bar{Y}_1 \]

10 12.5 15 17.5 20 22.5 25 27.5 30 \( \bar{Y} = 20.0 \)

\[ \bar{Y}_1 \]

2.67 7.25 6.80 17.75 17.60 17.75 6.80 7.25 2.67 \[ \bar{Y}_1 = 40.04 \]

\[ \alpha_{xx} = 12.04 \]

\[ \alpha_{yy} = 3.47 \]

Adapted from Ghiselli (1964, p. 109)
This data can be taken to be a summary of Y scores at each X score individually. Furthermore, on the right-hand side of the diagram, the distribution of Y scores (distribution II in terms of their deviation from the overall mean) is given. The dots at the various points along this line (distribution II) represent the frequency of deviation scores for those respective deviation values. How many deviation scores are there with a value of \( +4 \) and \( -2 \)?

There are four (4) of them.

Here: the distribution of Y is displayed in deviation scores. For purposes which will become apparent as this discussion progresses, as you may realize, the extent of the total amount of variation of Y.
The following best describes this implication?

A. The range of these deviation scores.
B. The spread of the deviation scores.
C. The symmetry of the distribution around mean X.

Good. The spread of scores (whether deviation or other) around a mean always connotes the notion of variance.

Although the range of scores is sometimes used as a rough measure of variation, the relative spread of scores (in this case deviation scores) around a mean connotes variation.

Although the symmetry of a distribution around a mean (here, X) may be considered as a direct by-product of variation around scores, the spread of scores (whether deviation or other) around a mean connotes the notion of variance.

The exact amount of total Y variance is given by the distribution in slide 2. In this case, 40% of this total variance may be due to many factors, some of which might be the same as those which produce variation among individuals.

Measure Y. If we could determine statistically the fraction of the total variance of Y that can be attributed to the same factors which underlie...
VARIATION IN X, THEN WE WOULD BE TAKING A POSITIVE STEP IN ATTEMPTING THE
FIND A MEASURE OF RELATIONSHIP, WHICH OF THE FOLLOWING RESOURCES HAS THIS
MIGHT BE SO:

A. THE DEGREE TO WHICH X AND Y VARY "SIMILARLY" IS THE DEGREE TO WHICH THEY
   ARE RELATED.

B. THIS IS NOTHING MORE THAN A ROUGH DEFINITION OF CORRELATION AS PRESENTED PREVIOUSLY.

C. LOGICALLY, THIS STATEMENT IS CORRECT, BUT IT IS IMPOSSIBLE TO
   FIND A MEASURE OF RELATIONSHIP STATISTICALLY TAKING THIS APPROACH.

CA 8]
TY > YOU ARE CORRECT. FOR THIS IS SAVING OTHER FACTORS ARE AT PLAY WHICH
   TEND TO PRODUCE VARIANCE AMONG SUBJECTS IN BOTH X AND Y.

CA A]
TY > CLOSE, BUT NOT QUITE RIGHT. STATEMENT A IS NOTHING MORE THAN A ROUGH
   DEFINITION OF COVARIANCE. STATEMENT X IS THE CORRECT RESPONSE SINCE THE
   DEFINITION OF CORRELATION IS DIRECTED TO THAT DATA ABOVE, IT MIGHT
   BE WISE FOR YOU TO REVIEW THE REASONS WHY THE COVARIANCE TOOK ITS
   FORM.

WA C]
TY SORRY, IT IS QUITE POSSIBLE TO DETERMINE STATISTICALLY A MEASURE OF CORRELATION.
   USING THIS APPROACH AS YOU WILL SEE SOON, TRY AGAIN.
UN > TYPE: A OR C

OU > LET US ATTEMPT TO DETERMINE THIS WITH A VARIANCE WHICH MAY BE ATTRIBUTABLE TO ACTH.
   VARIABLES X AND Y. TO DO THIS, LET US TAKE A "BACK UP" APPROACH IN THAT WE WILL TRY
   TO PARSE DOWN THE TOTAL VARIANCE (E Y) IN TERMS OF THAT PART OF THIS VARIANCE
   WHICH IS NOT ASSOCIATED WITH X. THAT IS, A BASICALLY POINT IS SIMILAR TO THE
   WHAT YOU MAY HAVE EXPERIENCED IN YOUR FIRST STAT COURSE WITH RESPECT TO THE PARTIAL
   VARIANCE, IN IDENTIFICATION VARIANCE COMPONENTS IN THE ANALYSIS.

RD > REFER AGAIN TO THE SUMMARY DATA AT THE BOTTOM OF SLIDE 2. PRESS THE RETURN KEY WHEN YOU ARE FINISHED.

OU > UNDER EACH MEAN VALUES A KEY WILL FILL VARIANCE. THIS PRESENTS THE VARIANCE
   OF X SCORES RELATIVE TO THAT PARTICULAR VALUE OF X. IT IS, AS USUAL, BASED ON THE
   AVERAGE SQUARES OF SIGNED DEVIATIONS FROM THE MEAN. HOWEVER, IN THIS CASE, THE
   MEAN SCORES GET USED FOR THESE PARTICULAR VARIANCES FOR THESE RESPECTIVE PARTS OF THE Y SCORES TAKEN.
AT EACH VALUE OF X, FOR EXAMPLE, THE VARIANCE OF Y AT X = 1 IS BASED ON A MEAN Y-value of 15. ON THE OTHER HAND, THE MEAN Y VALUE CALLED AT X = 1 TO COMPUTE THE VARIANCES OF THAT COLUMN, IF Y SCORES IS 15, COULD YOU SAY THAT??

TAKING THE VARIANCES OF Y ARE PARTIALLY VALUES OF X HAVE THE EFFECT OF LOOKING AT THE VARIANCE OF Y WITH THE INFLUENCE OF X HELD CONSTANT? ASSUME YES, LET'S LOOK AT IT.>

CA YES?

TV > ABSOLUTELY!! IN ESSENCE WHAT HAS BEEN DONE IS TO OBTAIN THE RESPECTIVE VARIANCES OF Y WITHOUT ALLOWING X TO VARY WITH RESPECT TO (X, INFLUENCE) ANY CASES ELSE. THIS IS THE VARIANCE OF Y WITH THE INFLUENCE OF X HELD CONSTANT OR DEVIATION. IN ESSENCE WHAT HAS BEEN DONE IS TO OBTAIN THE RESPECTIVE VARIANCES OF Y WITHOUT ALLOWING X TO VARY WITH RESPECT TO (L, INFLUENCE) ANY CASES ELSE. THIS IS THE VARIANCE OF Y WITH THE INFLUENCE OF X HELD CONSTANT OR DEVIATION.

CA NO!

TV > SORRY, BUT IT DOES HAVE THE EFFECT OF LOOKING AT THE VARIANCE OF Y WITH THE INFLUENCE OF X HELD CONSTANT OR DEVIATION. IN ESSENCE WHAT HAS BEEN DONE IS TO OBTAIN THE RESPECTIVE VARIANCES OF Y WITHOUT ALLOWING X TO VARY WITH RESPECT TO (X, INFLUENCE) ANY CASES ELSE. THIS IS THE VARIANCE OF Y WITH THE INFLUENCE OF X HELD CONSTANT OR DEVIATION.

UN > PLEASE TYPE YES OR NO.>>

>>= COMPARISON OF TOTAL VARIANCE AND PARTIAL VARIANCE.<

OU => THE DISTRIBUTION OF DEVIATION SCORES FOR VALUES OF Y COMPUTED ON THE BASIS OF MEAN Y SCORES AT EACH POINT OF X IS GIVEN AT THE FAR RIGHT HAND SIDE OF SLIDE 2. THIS IS DIFFERENT FROM THE PREVIOUS DISTRIBUTION. DEVIATION SCORES IN THAT SCORES IN THIS DISTRIBUTION REPRESENT DEVIATIONS FROM THE MEANS OF THE VARIOUS COLUMNS XM SORE, FOR EXAMPLE, DEVIATIONS FROM THE OVERALL MEAN Y VALUE. LOOKING AT SLIDE 2, NAME THE ONE Y SCORE THAT HAS A DEVIATION SCORING 11 IN THIS RESPECT OF +3.00 AND THE ONE Y SCORE THAT HAS A DEVIATION SCORING OF -2.10. SEPARATE YOUR ANSWERS WITH A LUMP.<

CA(N) 633613, 636328,N.

TV > VERY GOOD<<

WA(L) 2381,

TV > YOU ARE PARTIALLY CORRECT. 23 IS PART OF THE CORRECT ANSWER. TRY AGAIN.

WA(L) 1381.

TV > YOU ARE CORRECT. 23 IS PART OF THE CORRECT ANSWER. TRY AGAIN.

UN > YOU ARE CORRECT. 23 IS PART OF THE CORRECT ANSWER. TRY AGAIN.

UN > YOU ARE PARTIALLY CORRECT. 13 IS PART OF THE CORRECT ANSWER. TRY AGAIN.

UN > YOU ARE CORRECT. 13 IS PART OF THE CORRECT ANSWER. TRY AGAIN.

UN > YOU ARE CORRECT. 13 IS PART OF THE CORRECT ANSWER. TRY AGAIN.

UN > YOU ARE CORRECT. 13 IS PART OF THE CORRECT ANSWER. TRY AGAIN.

UN > YOU ARE CORRECT. 13 IS PART OF THE CORRECT ANSWER. TRY AGAIN.

UN > THE ANSWERS ARE 23 AND 13, RESPECTIVELY.
Q. Now, does the variation of deviation scores around the mean deviation score (\( \sigma_y \)) usual of distribution 2 seem considerably smaller than that of those of distribution 1? Answer yes or no.

CA. Yes

TY. You are 50% right.

CA. No.

TY. Look again. The spread of deviation scores for distribution 2 is \( \sigma_y \) FAI.LY OBVIOUSLY smaller than that of 1.

TY. Please answer yes or no.

Q. Recall that the deviation scores of distribution 1 were based on the overall y mean. (20 to be exact) across, or without particular regard to, the x scores. Now, if holding X constant had little or no effect on the variance of Y scores, as shown in distribution 2, which of the following would you expect with regard to the spread of scores in distribution 2? Answer either A, B, or C.

A. The spread of deviation scores of 2 would be identical to that of 1.

B. The spread of deviation scores of 2 would be greater than that of 1.

C. The spread of deviation scores of 2 would be less than that of 1.

CA(1) A.

TY. You are correct. If holding X constant had little or no effect on the variance of Y scores, then the variance of Y scores with X held constant would be identical to the simple variance of Y scores.

CA(1) B.

TY. If holding Y constant had little or no effect on the variance of Y scores, then the variance of Y scores with Y held constant would be identical to the simple variance of Y scores. Thus, the spread of deviation scores of 2 would not be greater than that of 1.

DD. DD.

CA(1) C.

TY. If holding Y constant had little or no effect on the variance of Y scores, then the variance of Y scores with Y held constant would be identical to the simple variance of Y scores. Thus, the spread of deviation scores of 2 would not be less than that of 1.

IN. Type A, B, or C.

Q. Turn to slide 3. Notice that holding X constant does not affect the total variance of Y since the variance of Y with the effects of X removed is identical with the overall variance of Y. This is directly given, again, in the comparison of distributions 1 and 2 of deviation scores.
Scatterplot of Raw Scores for Two Variables Reflecting No Relationship

\[
\begin{align*}
\bar{y}_1 &= 20 \\
\bar{y} &= 20 \\
\bar{y}_x &= 160 \\
\bar{y}_y &= 12.65 \\
\bar{y}_x &= 12.65 \\
\bar{y}_y &= 12.65 \\
\end{align*}
\]
Let us now relate this notion of holding X constant with respect to Y to the situation of a relationship between X and Y, once again. If these were a relationship between the two variables, by definition of correlation there would be no variance of Y about the mean.

Thus, if we find that holding one variable constant (or letting it vary), X implies which of the following? Select the most appropriate, if either A, B, or C, above.

A. The covariance between the variables is zero.
B. There is no apparent relationship between the variables.
C. The two distributions of deviation scores are identical.

Call (1). 61

Ty > exactly. If there is --- some sort of relationship between two variables, then ---

Removing the effect of variation of one of them from the other should manifest ---

itself in the reduction of variance of the latter. The illustration given in Slide 2 is an example of this, also note that alternatives A and C above are valid --- implications of this situation.

Call (1). 61

Ty > this is true. More importantly, though, there would be no apparent relationship ---

between the variables, if there is --- some sort of relationship between two variables ---

then removing the effect of variation of one of them from the other should manifest ---

itself in the reduction of variance of the latter. The illustration given in slide 2 is an example of this.

BR PR 

Call (1). 61

Ty > this is true. More importantly, though, there would be no apparent relationship ---

between the variables. (the) (the) (the) (the)

UN > type: A  B  O2. (the) (the)

OM3

QW > although we have diagrammatically shown that in the case where there is a suspected relationship between scores on two variables, the variation ---

of one of them is in our illustration, x, can be substantially reduced ---

controlling for the other. We next illustrate the way to do this ---

mathematically for practical purposes.

Once again, look at the variance terms at the bottom of slide 2. They vary, if you will, from the values of 2.67 to 2.73 depending on the specifics.

Value of X, again, each represents variation in Y that is attributed to X.

Since X is --- at each point.
PLEASE SUPPLY THE LUSING WORD.  

CA CONSTANT  

CALL CONTROL  

TV > YOU ARE INCORRECT****

UN > I CANT RECOGNIZE YOUR ANSWER. PLEASE TRY AGAIN****

UN > X IS CONSTANT OR CONTROLLED AT EACH POINT****

QU > IT IS OBVIOUS THAT THE VARIANCE TERMS AT THE BOTTOM OF SLIDE 2 ARE MUCH SMALLER THAN THE OVERALL VARIANCE OF Y. HOWEVER, WHAT IS QUERITED IS AN OVERALL SUMMARY OF ALL THESE TERMS. THIS WOULD DEFINE THE VARIANCE OF Y****

WITH THE INFLUENCE OF X HELD CONSTANT ACROSS OUR PARTICULAR RANGE. IF****

X VALUES.********

A DESCRIPTIVE DERIVATION OF THIS SUMMARY INDEX IS GIVEN IN THE GASTELLU REFERENCE***

SHOWN AT THE BOTTOM OF SLIDE 11, WE WILL NOT GO INTO IT HERE. FOR OUR PURPOSES THIS****

VARIANCE IS NOTHING THAN A WEIGHTED AVERAGE OF EACH OF THOSE Y VARIANCES****

AT PARTICULAR Y VALUES. AN EXAMPLE BASED ON OUR ILLUSTRATION IS****

GIVEN IN SLIDE 4.********

QU THIS VARIANCE TERM THAT HAS JUST BEEN PRESENTED IS KNOWN AS THE PARTIAL VARIANCE OF****

SOME VARIABLE (HERE Y). IT IS CONSIDERED PARTIAL IN THAT THE EFFECTS OF****

ANOTHER VARIABLE (HERE X) HAS BEEN REMOVED FROM IT.********

IT MAY BE IMPORTANT TO NOTE THAT WE COULD HAVE EASILY APPLIED THIS****

PROCEDURE TO THE OVERALL VARIATION OF X WITH Y HELD CONSTANT. IT WOULD****

JUST A MATTER OF DESCRIPTIVE CONVENIENCE THAT IT WAS DONE AS IT WAS.********

DERIVATION OF A GENERALIZED CORRELATION COEFFICIENT BASED ON PROPORTIONALITY****

OF ACCOUNTABLE VARIANCE********

QU NOW, LET US RE-EXAMINE WHAT WE HAVE DONE. WE HAVE OBTAINED A MEASURE OF VARIANCE OF****

Y WHERE IN THE "EFFECTS" OF X RELATIVE TO Y HAVE BEEN NULLIFIED. THE****

QUESTION TO BE RAISED NOW IS WHAT EXACT MEANING THIS HAS ON THE DEFINITION OF****

A MEASURE OF RELATION********

CONSIDER FOR A MOMENT THE MEANING OF THE STATEMENT "THE VARIANCE OF Y WITH****

THE EFFECTS OF X HELD CONSTANT" (PARTIAL VARIANCE OF Y RELATIVE TO X) IN TERMS OF THE****

UNDERLYING FACTORS WHICH TEND TO PRODUCE INDIVIDUAL DIFFERENCES (VARIANCE) OF****

MEASURES X AND Y. WOULD THIS STATEMENT NOT IMPLY THAT THIS RESULTANT TERM WOULD BE****

AN ILLUSTRATION OF THE EXTENT OF VARIATION IN Y DUE TO FACTORS OTHER THAN****

THOSE WHICH PRODUCE VARIATION IN X?********

TYPE : YES OR NO********

CA YES

TV > RIGHT. SINCE THE EFFECTS OF X HAVE BEEN HELD CONSTANT IN THIS TABLE, IT IS ONLY****
Computational example of a Partial Variance Term as the Weighted Sum of Individual Variances

\[ \sigma^2_{y|x} = \frac{1}{n} \sum_{i} n_i \sigma^2_{y_i} \]

where \( n_i \) represents the number of scores in column "i" and \( n \) is the total number of scores in the sample.

From our example,

\[ \sigma^2_{y|x} = \frac{3(2.67) + 4(7.25) + 5(6.80) + 8(17.75) + 10(17.60) + 8(17.75) + 5(6.80) + 4(7.25) + 3(2.67)}{50} \]

\[ \sigma^2_{y|x} = \frac{602.02}{50} \]

\[ \sigma^2_{y|x} = 12.04 \]

that variance which was determined previously on the basis of deviation scores.
REASONABLE TO ASSUME THAT THE VARIANCE IN Y WHICH REMAINS MUST BE DUE TO INDIVIDUAL DIFFERENCES IN MEASURE X.

CAN YOU RECONSIDER THIS POINT FOR A MOMENT. SINCE THE EFFECTS OF X HAVE BEEN HELD CONSTANT, IT IS ONLY REASONABLE TO ASSUME THAT THE VARIANCE IN Y WHICH REMAINS MUST BE DUE TO REASONS WHICH CANNOT BE SIMILARLY ATTRIBUTED TO INDIVIDUAL DIFFERENCES ON MEASURE X.

UNIQUE TYPE: YES OR NO.

OR.

QU > GRANTING THE ABOVE STATEMENT, THEN, WHICH OF THE FOLLOWING CAN BE SAID OF THE ALGEBRAIC DIFFERENCE BETWEEN THE OVERALL VARIANCE OF Y RELATIVE TO X AND THE PARTIAL VARIANCE OF Y WITH RESPECT TO X?

A. THE DIFFERENCE MARKS THE EXTENT TO WHICH VARIATION AMONG PEOPLE ON MEASURE Y IS DUE TO SIMILAR FACTORS WHICH ELICIT VARIATION AMONG PEOPLE ON MEASURE X.

B. THIS DIFFERENCE MARKS THE EXTENT TO WHICH VARIATION AMONG PEOPLE ON MEASURE Y IS NOT DUE TO SIMILAR FACTORS WHICH ELICIT VARIATION AMONG PEOPLE ARE RELATED.

CA AI.

TY > VERY GOOD, BECAUSE THE PARTIAL VARIANCE OF Y WITH RESPECT TO X IS THAT VARIANCE OF Y NOT ASSOCIATED WITH THE VARIABLE X. THE DIFFERENCE BETWEEN THE TOTAL VARIANCE OF Y RELATIVE TO X AND THE PARTIAL VARIANCE OF Y MUST BE THAT VARIANCE "CAUSED" BY SIMILAR FACTORS.

CA BI.

TY > INCORRECT, REMEMBER, THE PARTIAL VARIANCE OF Y WITH RESPECT TO X IS THE VARIANCE TERM OF Y WITH THE RELATIVE EFFECTS OF X REMOVED FROM IT. THUS, THE DIFFERENCE BETWEEN THE TOTAL VARIANCE OF Y RELATIVE TO X AND THE PARTIAL VARIANCE OF Y WITH RESPECT TO X MUST BE THAT VARIANCE WHICH IS DUE TO FACTORS COMMON TO BOTH.

WACI.

TY > YOU'RE JUMPING THE GUN A LITTLE. THIS DIFFERENCE IS NOT AS YET A MEASURE OF VARIABLE RELATIONSHIP. PLEASE SELECT ANOTHER أساء.
Now, refer to Slide 5 for the algebraic representation of this difference. (8)

We now find ourselves in somewhat the same situation as before, when we were discussing the concept of covariance as a possible measure of relationship. (8)

We could view the difference designated as a sigma squared y prime in.

In Slide 5, as being a measure of relationship also since it uses different differences in subjects (variance) of one variable which is produced by factors. Which tend to produce differences in subjects (variance).

In another variable. But, once again this index is incomplete as a measure of relationship for the same reasons as was the covariance term.

Which two of the following, if you recall, were the reasons why the covariance term was not a complete measure of correlation?

A. Different scales of measurement will result in different magnitudes of the term.

B. It is not consistent from one situation to the next.

C. The term lacks a standard by which to make a judgment as to the degree of relationship existing between two variables.

D. It is a biased estimate of correlation.

Cal: CA, CA

Try right, good thinking.

Un: Wrong, a change in scales of measurement will change the value of the term. Either covariance of the variance differences and there is no relative standard by which we could judge how much relationship there is between the variables of interest.

Br: BR

Qu: Because of the incompleteness of this variance difference as a measure of relationship, one further step must be taken.

The remainder of Slide 6 delineates this step. What is done is to form the ratio of sigma squared y prime to sigma squared y prime. Where sigma squared y prime is defined as given in Slide 5. The logical description of what this ratio represents is given in Slide 5.

De: As may be apparent to you, the ratio between these variances is identical to the ratio of the respective standard deviations. Since it is more customary to deal with standard deviations, and the proportion is given in this respect in Slide 6. Basically, this may be thought...
Slide 5
Derivation of the Generalized Correlation Coefficients (Squared)

\[ \gamma_{yx} = \text{total variation among individuals with respect to their Y scores (in deviation form)} \]

\[ \gamma_{yx}' = \text{variation among individuals with respect to their Y scores after variation among them with respect to their X scores has been removed} \]

\[ \gamma_{yx} - \gamma_{yx}' = \text{variation among individuals with respect to both their X scores and their Y scores} \]

Let \[ \gamma_{yx}' = \gamma_y - \gamma_{yx} \]

\[ \frac{\gamma_{yx}'}{\gamma_y} = \text{the proportion of total variation among individuals with respect to their Y scores which can be accounted for by variation among individuals with respect to their X scores; in other terms, this ratio represents the proportion of common factor variance between the X and Y variables} \]

Likewise,

\[ \frac{\gamma_{xy}'}{\gamma_x} = \text{the proportion of total variation among individuals with respect to their X scores which can be accounted for by variation among individuals with respect to their Y scores} \]

\[ \gamma_{yx}^2 = \frac{\gamma_{yx}'}{\gamma_y} \]

\[ \gamma_{xy}^2 = \frac{\gamma_{xy}'}{\gamma_x} \]
Slide 6

Definition of the Generalized Correlation Coefficients

\[ \eta_{y,x} = \frac{C_{y'}}{C_y} \]

\[ \eta_{x,y} = \frac{C_{x'}}{C_x} \]
TO BE OUR MEASURE OF CORRELATION.


Q2. FROM THE DEFINITION AND DEVELOPMENT OF SIGMA Y PRIME AND SIGMA Y (AS GIVEN IN SLIDES 5 AND 6) WHICH OF THE FOLLOWING IS CORRECT?

A. SIGMA Y PRIME WILL ALWAYS BE GREATER THAN OR EQUAL TO SIGMA Y.

B. SIGMA Y PRIME WILL TYPICALLY BE GREATER THAN SIGMA Y, BUT AT TIMES...

C. SIGMA Y PRIME WILL ALWAYS BE LESS THAN OR EQUAL TO SIGMA Y.

D. SIGMA Y PRIME WILL TYPICALLY BE LESS THAN SIGMA Y, BUT AT TIMES...

W A: C1

T Y: RIGHT. SINCE SIGMA SQUARED Y PRIME, BY DEFINITION, IS THE DIFFERENCE BETWEEN THE TOTAL VARIANCE OF Y (SIGMA SQUARED Y) AND THE PARTIAL VARIANCE OF Y.

SIGMA Y PRIME MUST BE LESS THAN OR EQUAL (AT MOST) TO SIGMA Y.

W A: C2

T Y: NOT SO. IN FACT IT IS JUST THE OPPOSITE OF THIS ANSWER. TRY AGAIN.

W A: C3

T Y: SIGMA Y PRIME WILL NEVER BE GREATER THAN SIGMA Y. TRY AGAIN.

W A: D1

T Y: SIGMA Y PRIME WILL NEVER BE GREATER THAN SIGMA Y.

T R Y A G A I N.

U N: TYPE: A B C D.

Q3. WITH THE ABOVE IN MIND, WHAT WILL THE RATIO OF SIGMA Y PRIME TO SIGMA Y BE?

A. GREATER THAN OR EQUAL TO ONE.

B. LESS THAN OR EQUAL TO ONE.

C. USUALLY EQUAL TO ONE.

W A: C1

T Y: RIGHT. THIS FOLLOWS PRECISELY FROM THE ABOVE DISCUSSION. SINCE SIGMA Y PRIME IS ALWAYS LESS THAN OR EQUAL TO SIGMA Y.

W A: C2

T Y: INCORRECT. AS WAS JUST DISCUSSED, SIGMA Y PRIME WILL ALWAYS BE LESS THAN OR EQUAL TO SIGMA Y. THUS, THE RATIO OF THE former TO THE latter.

P A G E 17
MUST BE LESS THAN OR EQUAL TO ONE.

CA C1

ENTITY TIMES THIS RATIO MAY EQUAL ONE [A RARITY, THOUGH]. MOST TYPICALLY, THE RATIO WILL BE LESS THAN ONE, BUT NEVER GREATER THAN ONE, THIS FOLLOWS DIRECTLY FROM THE FACT THAT SIGMA Y IN THE IS ALWAYS LESS THAN (OR EQUALS)

TO SIGMA Y.,

UN > TYPE: A, B OR C


ON THE OTHER HAND, WHICH OF THE FOLLOWING IS THE CASE WHEN THE PARTIAL VARIANCE OF Y.

Y (WITH RESPECT TO X) IS ZERO? ANSWER A, B, OR C.

A. THE RATIO OF SIGMA Y PRIME TO SIGMA Y. IS LESS THAN ONE AND THERE IS SOME

EXISTENT STATISTICAL RELATIONSHIP BETWEEN Y AND Y.

B. THE RATIO OF SIGMA Y PRIME TO SIGMA Y. IS ZERO AND THERE IS NO EXISTENT

STATISTICAL RELATIONSHIP BETWEEN X AND Y.

C. THE RATIO BETWEEN SIGMA Y PRIME AND SIGMA Y IS ONE (1) AND THERE IS A

PERFECT STATISTICAL RELATIONSHIP BETWEEN X AND Y.

CA C1

ENTITY YOU ARE CORRECT***. SINCE THE PARTIAL VARIANCE OF Y IS ZERO, THIS INDICATES THAT

HOLDING THE FACTORS UNDERLYING X CONSTANT ACCOUNTS FOR ALL THE VARIANCE IN Y.

THEN, THE DIFFERENCE BETWEEN THE VARIANCE OF Y AND THE PARTIAL VARIANCE OF Y IS

EQUAL TO JUST THE VARIANCE OF Y. THUS, THERE IS A PERFECT RELATIONSHIP BETWEEN X

AND Y AND IT IS SHOWN IN THE FACT THAT THE RATIO OF SIGMA Y PRIME TO SIGMA Y IS 1.

GIVEN IN SLIDE 6, EQUATION 6.1, IS ONE.
TY >> NOT IN THIS CASE, SINCE THE PARTIAL VARIANCE WAS GIVEN AS EQUAL TO ZERO, THE DIFFERENCE BETWEEN THE TOTAL VARIANCE OF Y AND ITS REGRESSIVE TOTAL VARIANCE OF Y, THE SQUARE ROOT OF THIS DIFFERENCE WAS DETERMINED TO BE SIGMA Y'.

THUS, HERE, THERE IS A PERFECT RELATIONSHIP BETWEEN X AND Y AND IT IS SHOWN IN THE FACT THAT THE RATIO OF SIGMA Y' TO SIGMA Y (THE CORRELATION COEFFICIENT) AS GIVEN IN SLIDE 6, EQUATION 6.1, IS ONE.

TY AND NOT LESS THAN ONE.

TY AND NOT ZERO.

DIFFERENT ILLUSTRATIONS OF RELATIONSHIP HAVE JUST BEEN PRESENTED, ONE OF A LACK OF RELATIONSHIP, AND ONE OF PERFECT RELATIONSHIP. OF COURSE, THERE ARE NECESSARILY INTERMEDIATE DEGREES OF RELATIONSHIP, AND THE INDEX OF RELATIONSHIP WILL MANIFEST THIS IN TERMS OF A COEFFICIENT WHICH RANGES BETWEEN THE VALUE ZERO (NO RELATIONSHIP) AND THE VALUE ONE (PERFECT RELATIONSHIP). UNLIKE THE OTHER TWO ATTEMPTS IN DEFINING A MEASURE OF RELATIONSHIP (THE COVARIANCE TIMES SIGMA SQUARE Y PRIME), THIS INDEX IS, IN A SENSE, STANDARDIZED, SINCE WE KNOW THE LIMITS OF IT (IN TERMS OF YC OR PERFECT RELATIONSHIP), WE NOW HAVE A BASIS ON WHICH TO EVALUATE THE EXTENT OF THE ASSOCIATION.

NOW, CONSIDER FOR THE MOMENT THE STATISTICAL CONDITIONS UNDER WHICH OUR MEASURE RESPONDS TO PERFECT RELATIONSHIP AS A MATTER OF FACT, ARE THERE ANY EXPLICIT STATISTICAL ASSUMPTIONS IMPLIED IN THE DEVIATION? ANSWER: YES OR NO.

CALL END

TY >> RIGHT*** WE DID NOT IMPOSE ANY STATISTICAL ASSUMPTIONS IN PRODUCING THE RESULTANT INDEX, THUS EQUATIONS 5.2 AND 5.3 OF SLIDE 5 (AND THEIR COUNTERPARTS, 6.1 AND 6.2) OF SLIDE 6, CAN BE SAID TO BE DEFINITIONS OF A GENERALIZED CORRELATION COEFFICIENT.

CALL END

TY >> NOT ALL MENTION WAS EVER MADE, IMPLICITLY OR EXPLICITLY, WITH RESPECT TO ANY STATISTICAL ASSUMPTIONS UNDER WHICH THE INDEX WAS DERIVED, THUS EQUATIONS 5.2 AND 5.3 OF SLIDE 6 (AND THEIR COUNTERPARTS, 6.1 AND 6.2) OF SLIDE 6, CAN BE SAID TO BE DEFINITIONS OF A GENERALIZED CORRELATION COEFFICIENT.
I. THIS LAST ACTION IS A VERY IMPORTANT ONE. IT IS SO BECAUSE IN USING THIS GENERAL MEASURE OF VARIABLE RELATIONSHIP, THE UNDERLYING FORM OF THE RELATIONSHIP BETWEEN THE VARIABLES DOES NOT AFFECT ITS APPLICATION, WHICH OF THE FOLLOWING WOULD YOU BELIEVE BEST ILLUSTRATES WHAT IS MEANT BY FORM OF RELATIONSHIP?

A. LEPTOKURTIC VS. PLATYKURTIC
B. LINEAR VS CURVILINEAR
C. NEGATIVELY SKEWED VS. POSITIVELY SKEWED

CA A1
LD 1/C3
TY > VERY GOOD** THE RELATIONSHIP BETWEEN THE VARIABLES TAKES THE FORM OF EITHER ONE OF LINEARITY OR CURVILINEARITY.
BR PR 1
CA A1
LD 1/C3
TY > KURTOSIS (LEPT- OR PLATY-) IS AN INDICATION OF SHAPE OF A DISTRIBUTION OF SCORES, BUT IS NOT AN INDICATION OF FORM OF RELATIONSHIP. A RELATIONSHIP BETWEEN VARIABLES IS EITHER LINEAR OR CURVILINEAR.
BR PR 1
CA C1
TY > SKEWNESS (NEGATIVE OR POSITIVE) IS AN INDICATION OF SHAPE OF A DISTRIBUTION OF SCORES, BUT IS NOT AN INDICATION OF FORM OF RELATIONSHIP. A RELATIONSHIP BETWEEN VARIABLES IS EITHER LINEAR OR CURVILINEAR.
UN > TYPE A, B, OR C
PR 1
TY FOR OUR ILLUSTRATION, THE RELATIONSHIP APPARENTLY TENDED TOWARD LINEARITY. WE COULD HAVE USED AN EXAMPLE WHICH MANIFESTED A CURVILINEAR RELATIONSHIP JUST AS EASILY. AS A MATTER OF FACT, SLIDE 7 IS AN EXAMPLE OF SUCH A CASE. WE WILL BE SAID LATER IN REFERENCE TO HOW THE FORM OF RELATIONSHIP (EITHER LINEAR OR CURVILINEAR) IS DETERMINED.
BR HOME 1

HOW

OU > HOMOGENEITY OF VARIANCE AS IT RELATES TO CORRELATION
> THERE IS ONE DETAIL THAT SHOULD BE CONSIDERED IN THE MEANINGFULNESS OF THE COEFFICIENT. HOWEVER, AS YOU RECALL, WE USED A PARTIAL VARIANCE OF Y IN OUR DENOMINATOR.
Slide 7

Scatterplot of Raw Scores for Two Variables Illustrating a Curvilinear Relationship

Adapted from Chiselli (1964, p. 113)
A. AN AVERAGE WEIGHTED SUM OF THE VARIANCES OF Y AT SPECIFIC POINTS OF X.

B. THE OVERALL VARIANCE OF Y RELATIVE TO VALUES OF X.

C. A VARIANCE OF Y BASED ON ONLY LIMITED VALUES OF Y.

CALL AG1
TY > GOOD********>-<------>
BR HOMO1]
UN > THINK IT OVER AGAIN AND GIVE ANOTHER ANSWER.>-<------>
UN > WRONG AGAIN. MAYBE A LITTLE REVIEW MIGHT CLARIFY MATTERS.>-<------>
BR PARTE]

PARTRE

QU >> THE NOTION OF THE PARTIAL VARIANCE IS CENTRAL TO THE DEVELOPMENT OF CUR<----->
CORRELATION COEFFICIENT. RECALL THAT, IN GENERAL, A PARTIAL VARIANCE IS A MEASURE<-----
OF VARIANCE WITH THE RELATIVE EFFECTS OF ANOTHER VARIABLE REMOVED FROM IT. IN CUR<----->
CASE, THIS WAS THE VARIANCE OF Y WITH THE EFFECTS OF X REMOVED. TURNING TO SLIDE<------>
2, RECALL THAT AT EACH VALUE OF X THERE IS A VARIANCE OF Y FOR THOSE Y SCORES WHICH<-----
WERE "PAIRED" WITH THAT PARTICULAR X SCORE. BECAUSE EACH OF THESE VARIANCES IS<-----
PARTICULAR TO A SINGLE X SCORE, X IS CONSTANT WITHIN EACH. PUT IN OTHER TERMS, X<-----
HAS NO REAL EFFECT ON EACH OF THESE SPECIFIC VARIANCES. NOW IN ORDER TO GET A<-----
SINGLE OVERALL MEASURE OF THE VARIANCES OF Y FOR THE PARTICULAR VALUES OF X, WE<-----
FOUND THAT A WEIGHTED AVERAGE OF THESE PARTICULAR VARIANCES WOULD DO IT. THIS<-----
OVERALL VARIANCE WAS THEN DEFINED TO BE THE PARTIAL VARIANCE OF Y WITH THE EFFECTS<-----
OF X HELD CONSTANT OR REMOVED.>-<------>
BR HOMO1]

HOMO1]

QU >> NOW, CONSIDER FOR THE MOMENT THE CASE WHEREIN THE VARIANCES OF Y AT SPECIFIC VALUES<-----
OF X DIFFER TO A FAIRLY LARGE DEGREE. LOOKING BACK TO SLIDE 2, YOU MAY NOTICE THAT<-----
THERE ARE NOT EQUAL AMOUNTS OF VARIANCE OF Y AT THE VALUES OF X AS YOU PROCEED FROM<-----
X EQUALS 0 TO X EQUALS 9. HOW DO THE VARIANCES OF Y AT THE EXTREME VALUES OF X<-----
COMPARE TO THOSE AT THE MIDDLE VALUES OF Y? ANSWER A, B, OR C.>-<------>
A. THEY ARE REALLY RELATIVELY ALIKE.>-<------>
B. THOSE AT THE EXTREMES ARE SMALLER THAN THOSE TOWARD THE MIDDLE VALUES OF X.>-<------>
C. THOSE AT THE EXTREMES ARE LARGER THAN THOSE TOWARD THE MIDDLE VALUES OF X.>-<------>
CALL AG1
TY > RIGHT YOU ARE********>-<------>
CALL AG1
THEY ARE NOT RELATIVELY ALIKE. THE FACT IS THAT THE VARIANCES OF \( Y \) AT THE EXTREME VALUES OF \( X \) ARE SMALLER THAN THOSE TOWARD THE MIDDLE VALUES OF \( X \).

JUST THE OPPOSITE. THE VARIANCES OF \( Y \) AT THE EXTREME VALUES OF \( X \) ARE SMALLER THAN THOSE TOWARD THE MIDDLE VALUES OF \( X \).

THE DIFFERENCES IN \( Y \) VAIRANCES AT THE \( X \) VALUES WOULD TEND TO IMPLY THAT PARTICULAR VALUES OF \( X \) INFLUENCE \( Y \) SCORES DIFFERENTIALLY. THESE DIFFERENCES WOULD FURTHER TEND TO IMPLY THAT THE NATURE OF THE OVERALL RELATIONSHIP BETWEEN \( X \) AND \( Y \) IN TERMS OF FACTORS WHICH UNDERLIE THIS RELATIONSHIP IS RATHER COMPLEX. HOWEVER, THE AVERAGE OF THESE PARTICULAR VARIANCE TERMS WOULD NOT GIVE A CLEAR PICTURE OF THIS SITUATION, AND INITIAL INTERPRETATIONS OF THE INDEX OF RELATIONSHIP BASED ON THIS PARTIAL VARIANCE TERM WERE BE MISTAKING ESPECIALLY If CAREFUL ANALYSIS OF THE SCATTERPLOT HAD NOT BEEN MADE.

AS A POINT OF CONSIDERATION BASED ON THE ILLUSTRATION IN SLIDE 7, HOWEVER, NOTICE THAT THE VARIANCES OF \( Y \) AT VARIOUS VALUES OF \( X \) ARE CALCULATED ON THE BASIS OF A RELATIVELY SMALL NUMBER OF CASES, WHICH OF THE FOLLOWING, THEN MIGHT BE A REASON FOR:

1. THE MORE DIVERGENT VARIANCES OF \( Y \)
2. ERROR IN CALCULATING THE VARIANCES
3. THE NATURE OF THE TWO VARIABLES, \( X \) AND \( Y \)
4. SAMPLING ERROR

VERY GOOD

THIS MAY BE THE REASON: ASSUMING CORRECT CALCULATIONS, HOWEVER, THE REASON FOR THIS IS MOST LIKELY SAMPLING ERROR.

TO AN EXTENT YOU ARE RIGHT. BUT, IN A SPECIFIC SENSE, THE MORE PROBABLE REASON ELSEWHERE IS THE MORE DIVERGENT VARIANCES OF \( Y \) IS SAMPLING ERROR.

Because of the few cases on which each variance was calculated (especially at the extreme values of \( Y \)), it must always remain a possibility that a certain amount of sampling error is at play. In other words, the scores on which these divergent variances were calculated may not be adequately representative of a hypothetical
Population of scores from which these few were "sample".

Q7. Thus, in order to be able to interpret a statistical relationship between two variables with at least minimal understanding, it should be taken that the variance of variable one (here Y) at the particular values of variable two (here X) are similar in magnitude. This should also be the case for variable two at a particular values of variable one. Can you tell me that this statistical notion is known as?

CALC: HOMOSCEDASTIC
TY: EXCELLENT THINKING.
TY R21
TY: **********
CALC: HOMOSCEDASTIC
TY: YOU ARE CORRECT. IT IS ALSO MORE POPULARLY KNOWN AS HOMOGENEITY OF VARIANCE.
UN: I DON'T RECOGNIZE YOUR ANSWER. PLEASE TRY AGAIN.
UN: THIS NOTION IS KNOWN AS HOMOGENEITY OF VARIANCE.
BR: PR:

>>> INADEQUACY OF THE GENERALIZED COEFFICIENT AS A UNIQUE MEASURE OF RELATIONSHIP

Q7: As you recall, an overall or generalized coefficient of correlation was developed which seemed to meet our initial need of an index that gave... the degree to which variables were related and which was "standardized" such that it would be interpretable from one situation to...


LOOKING BACK AT SLIDES 2 AND 5, USE THE APPROPRIATE DATA AND EQUATIONS TO CALCULATE THE SQUARE OF THE CORRELATION COEFFICIENT WHICH IS APPROPRIATE FOR THESE SPECIFIC DATA... (USE THE "CALC" FUNCTION IF NECESSARY).

CALC: .696, .696

TY: CORRECT. THE SQUARE ROOT OF THIS IS THE CORRELATION COEFFICIENT (BASED ON Y) BETWEEN VARIANCE AND IS EQUAL TO .830. <<<

UN: GIVE IT ONE MORE TRY.
UN: USING THE PARTIAL VARIANCE OF Y AND THE TOTAL VARIANCE OF Y TERMS, YOU...
SHOULD HAVE SUBTRACTED THE FORMER FROM THE LATTER, THEN YOU SHOULD

HAVE USED EQUATION 5.2 TO GET THE SQUARE OF THE CORRELATION (BASED ON Y-VARIANCE) WHICH IS .59. THE SQUARE ROOT OF THIS IS THE CORRELATION AND .

IT IS EQUAL TO "30. (---------------)

BU BU

QU > NOW USING THE APPROPRIATE DATA (WHICH YOU HAVE NOT SEEN AS YET) AND EQUATION 5.3, THE SQUARE OF THE CORRELATION OF X AND Y (BASED ON X-VARIANCE) WOULD BE

.913. (---------------)

IT IS OBVIOUS, HERE, THAT THE TWO CORRELATIONS BASED ON THE SAME VARIABLES ARE DIFFERENT.

NEEDLESS TO SAY, THIS EASILY PRESENTS A PROBLEM SINCE WE HAVE HERE TWO DIFFERENT MEASURES OF RELATIONSHIP FOR THE SAME TWO VARIABLES.

LET US ATTEMPT A SOLUTION TO THIS PROBLEM. (---------------)

THE REGRESSION COEFFICIENT AS A POSSIBLE MEASURE OF RELATIONSHIP

QU TURN TO SLIDE 9. (---------------)

DA 51

QU THIS IS OUR GRAPHICAL EXAMPLE BUT NOTICE THAT THERE ARE TWO LINES DRAWN THROUGH THE SCATTER OF POINTS. THESE LINES REPRESENT LINES OF "BEST FIT", ONE WITH RESPECT TO THE DISTRIBUTION OF POINTS WITH X VALUES AS THE BASE AND THE OTHER WITH Y VALUES AS THE BASE. THAT IS, THE LINE FOR X IS DETERMINED BY THE MEANS OF X SCORES AT EACH POINT OF Y AND THE LINE FOR Y IS DETERMINED BY THE MEANS OF Y SCORES CORRESPONDING TO EACH POINT OF X. COULD OUR SCATTER OF POINTS HAVE BEEN DESCRIBED BY CURVES RATHER THAN STRAIGHT LINES.

CA YES!

CB NO!

TY THEY COULD HAVE, BUT FOR PURPOSES HERE, WE WILL UTILIZE LINEAR REPRESENTATIONS.

ASIDE FROM THIS, THE TENDENCY OF THE SCATTER OF POINTS SEEMS TO BE LINEAR.

UN > TYPE: YES OR NO

DM > NOTE: A CLARIFICATION OF WHAT IS MEANT BY "BEST FIT" WILL FOLLOW SOONER OR LATER.

ALSO RECALL THAT WHETHER THE POINTS ARE DESCRIBED LINEARLY OR CURVILINEARLY,

OUR GENERALIZED CORRELATION COEFFICIENTS ARE APPLICABLE.

FROM YOUR MATHEMATICAL EXPERIENCES, RECALL THAT ANY STRAIGHT LINE MAY BE "Y = " X + A " HERE " A " REPRESENTS THE
Scatterplot of Raw Scores for Two Variables Illustrating Their Lines of Regression

Adapted from Ghiselli (1964, p. 109)
Slope of the line and \( a \) gives that point where the line crosses the y (or vertical) axis of the coordinate axis system. An example of this would be \( y = 0.5x \). What is the slope of this line? (\( \text{CA(w): } 0.5 \\text{ E51} \))

**CA 01**

**TY >** Right. (\( \text{CA(w): } 0.5 \\text{ E51} \))

**TY >** No. The slope of this line is \( 0.5 \). (\( \text{CA(w): } 0.5 \\text{ E51} \))

**UN >** The slope of this line is \( 0.5 \). (\( \text{CA(w): } 0.5 \\text{ E51} \))

**QR OR I**

**QU >** Now, what is the y-intercept? (the point at which the line crosses the y-axis) of this line? (\( \text{CA(w): } 0.5 \\text{ E51} \))

**CA 01**

**TY >** Correct. This line crosses the y-axis at the origin, the point at which the two axes cross. (\( \text{CA(w): } 0.5 \\text{ E51} \))

**CA(w): 0.5 E51**

**TY >** Wrong. This is the slope of the line. The y-intercept of this line is zero or the origin (the point at which the two axes cross). (\( \text{UN}\) the y-intercept is zero. (\( \text{CA(w): } 0.5 \text{ E51} \))

**BR PR I**

**RD >** When an increase in x values (going to the right along the horizontal axis) results in an increase in y values (going up along the vertical axis), the slope (\( m \)) will be positive. Conversely, when a decrease in x is met with a decrease in y along the line, the slope will be negative. An example of both cases is given in slide 5. Line "A" represents one with a positive slope and line "B" with a negative slope. (\( \text{CA(w): } 0.5 \text{ E51} \))

Press the return key when you are finished. (\( \text{CA(w): } 0.5 \text{ E51} \))

**QU >** The notion of positive and negative slopes has meaningful implication with respect to the further definition of a correlation coefficient as we shall soon see. (\( \text{UN}\) when we are dealing with the plot of scores on two variables (as we are), and (\( \text{CA(w): } 0.5 \text{ E51} \))

are concerned with their degree of relationship, the two lines drawn through the (\( \text{UN}\)) scattering of points are known as "regression lines", in this case, the slopes of these lines, known as the "regression coefficients". For convenience, let us (\( \text{CA(w): } 0.5 \text{ E51} \))

symbolize these terms as "\( a \)" and "\( b \)". From previous discussions will the two (\( \text{CA(w): } 0.5 \text{ E51} \))

"b's" pertaining to the two regression lines as different, generally speaking? (\( \text{CA(w): } 0.5 \text{ E51} \))

**CA YES!**

**TY >** Right. (\( \text{CA(w): } 0.5 \text{ E51} \))

**CA YES!**
Slide 9

Lines with Positive or Negative Slope

Line A

Line B
TY INCORRECT. RECALL THAT

IN GENERAL THE SLOPES OF THESE LINES WILL BE DIFFERENT.

THE IMPORTANCE OF THE REGRESSION COEFFICIENT IN TERMS OF CORRELATION CAN BE ILLUSTRATED AS FOLLOWS. PLEASE TURN TO SLIDE 10. IN BOX 1 YOU WILL NOTICE A GREAT DEAL OF VARIABILITY OF THE POINTS AROUND THE TWO REGRESSION LINES DRAWN.

IN THIS EXAMPLE THERE EXISTS LITTLE OR NO CORRELATION SINCE THE VARIABILITY OF Y SCORES AT EACH POINT OF X IS ALMOST IDENTICAL TO THE TOTAL VARIABILITY OF Y SCORES, THE SAME HOLDS TRUE FOR THE VARIABILITY OF X SCORES. THUS IT WOULD APPEAR THAT HOLDING X CONSTANT WOULD NOT AFFECT THE SIZE OF THE

VARIANCE OF Y IMPLYING LITTLE OR NO STATISTICAL RELATIONSHIP BETWEEN X AND Y.

NOW, IN BOX 2, THE VARIABILITY AROUND THESE LINES IS SOMEWHAT LESS PRONOUNCED.

NOTICE ALSO THAT THE REGRESSION LINES SEEM TO BE APPROACHING ONE ANOTHER. IN THIS EXAMPLE WOULD YOU SAY THAT THE RELATIONSHIP BETWEEN X AND Y ARE LESS THAN, GREATER THAN, OR EQUAL TO THOSE OF BOX 1? (TYPE EITHER)

CA GREATER.

TY GOOD.

CA LESS.

TY NO.

CA EQUAL.

TY NO.

UN > TYPE: LESS GREATER OR EQUAL

TY THESE RELATIONSHIPS (PLURAL SINCE THERE ARE TWO COEFFICIENTS, REMEMBER!) ARE GREATER SINCE IT APPEARS THAT VARIANCES OF SCORES AROUND THE LINE ARE LESS THAN THE TOTAL VARIANCES OF EITHER X OR Y, DEPENDING ON THE PARTICULAR LINE TO WHICH YOU ARE REFERRING.
Slide 10
Relationship of Regression Lines to Differing Degrees of Correlation
Referring to boxes 3 and 4, there appears to be even greater relationships between X and Y, based on the same rationale as that for box 2. Moreover, as you can see, the regression lines further approach each other.

Notice, now, that in box 1 the slope (regression coefficient) of line "B" is very great (infinitely to be exact) and the slope of line "A" is zero. In box 2, the slope of line "B" is less than it was in box 1, but the slope of line "A" is greater. A similar situation occurs in boxes 3 and 4, respectively.

What was said about the degrees of relationship exhibited in the various boxes, along with the relative sizes of the regression coefficients in each case, which of the following statements would be most appropriate?

A. The greater the magnitude of the regression coefficients, relative to each other, the greater the degree of relationship between the variables under consideration.

B. The magnitudes of the regression coefficients relative to each other have no bearing on the degree of relationship between the variables under consideration.

C. The smaller the magnitudes of the regression coefficients relative to each other, the greater the degree of relationship between the variables under consideration.

CA A1
TY > Very good. *<--------->
BR or 1

CA B1
TY > I'm afraid you are wrong. Based on the discussions of magnitude of the regression coefficients and the degree of relationship exhibited with respect to these magnitudes, it was hoped that you would make the association that the greater the magnitudes of the regression coefficients relative to each other, the greater the degree of relationship between the variables under consideration.

BR or 1

CA C1
TY > Not quite. From the above discussions, it was hoped that you would make the association that the greater the magnitudes of the regression coefficients relative to each other, the greater the degree of relationship between the variables under consideration.

BR or 1
THIS IS A PHENOMENON WHICH RELATES THE NATURE OF REGRESSION COEFFICIENT (THE SLOPE OF THE REGRESSION LINE) WITH DEGREE OF LINEAR RELATIONSHIP. THUS, IT MIGHT BE WISE TO DEVELOP A COMPUTATIONAL DEFINITION OF THE REGRESSION COEFFICIENT AS A POSSIBLE MEASURE OF LINEAR RELATIONSHIP SINCE THIS FORM OF RELATIONSHIP IS MOST COMMON IN APPLICATION.


THE RESULTS OF THESE DERIVATIONS ARE GIVEN IN SLIDE 11, EQUATIONS 11.1 AND 11.2. PLEASE REFER TO IT NOW.

THE DEVIATION SCORES, (X AND Y), USED IN EQUATIONS 11.1 AND 11.2 ARE DEVIATION SCORES FROM THE RESPECTIVE OVERALL X AND Y MEANS. AGAIN, WILL THE TWO COEFFICIENTS IN THESE EQUATIONS BE IDENTICAL?

NOT CORRECT.

YES.

THEY WOULD ONLY BE IDENTICAL IF THE VARIANCE OF X AND THE VARIANCE OF Y ARE IDENTICAL SINCE THE TWO COEFFICIENTS ARE THE SAME EXCEPT FOR THESE VARIANCE TERMS.

QU NOTICE IN PLACE OF THE M'S OF THE LINEAR EQUATIONS 11.3 AND 11.4 OF SLIDE 11 THERE ARE THE WORKING DEFINITIONS OF THE SLOPES. IN THESE TWO CASES THEY ARE THE RESPECTIVE REGRESSION COEFFICIENTS. IT SHOULD ALSO BE NOTED THAT IN DEVIATION SCORE FORM, THE Y INTERCEPTS ("A") OF EQUATIONS 11.3 AND 11.4 ARE MATHEMATICALLY FORCED TO BE ZERO. THUS THE REGRESSION EQUATIONS (IN DEVIATION FORM) APPEAR AS THOSE SHOWN IN SLIDE 11 FOR POSSIBLE EXPERIENCE IN THE CONTINUED DEVELOPMENT OF THIS TOPIC, IT IS NOW NECESSARY TO DEAL FURTHER WITH THE NOTION OF THE LINE OF REGRESSION. RECALL FROM A PREVIOUS DISCUSSION THAT MENTION WAS MADE OF THE LINE OF "BEST FIT" FOR PURPOSE OF THE PRESENT DISCUSSION A SMALL DISTINCTION WILL BE MADE BETWEEN LINE OF "BEST FIT" AND REGRESSION LINE. NOTICE IN SLIDE 11 THAT LINE "A" (THE LINE BASED ON Y MEAN VALUES) PASSES PRECISELY THROUGH THE MEANS OF
Slide 11
Definitions of Regression Equations and Their Respective Regression Coefficients

\[
b_{y,x} = \frac{\sum xy}{N \sigma_x^2} \tag{11.1}
\]

\[
b_{x,y} = \frac{\sum xy}{N \sigma_y^2} \tag{11.2}
\]

\[
\bar{y}_i = b_{y,x} \bar{x}_i = \frac{\sum xy}{N \sigma_x^2} \bar{x}_i \tag{11.3}
\]

\[
\bar{x}_i = b_{x,y} \bar{y}_i = \frac{\sum xy}{N \sigma_y^2} \bar{y}_i \tag{11.4}
\]

\[
y_i' = b_{y,x} \bar{x}_i = \frac{\sum xy}{N \sigma_x^2} \bar{x}_i \tag{11.5}
\]

\[
x_i' = b_{x,y} \bar{y}_i = \frac{\sum xy}{N \sigma_y^2} \bar{y}_i \tag{11.6}
\]

Ghiselli, Edwin E. *Theory of Psychological Measurement*
EACH COLUMN OF Y SCORES. REFERENCING TO THIS SLIDE AGAIN, CAN IT BE SAID
OF LINE "B" (THAT LINE BASED ON THE MEANS OF EACH ROW OF X SCORES) THAT IS PASSING
THROUGH EACH ROW MEAN OF X SCORES?

CA NO

TY > CORRECT

CA YES

TY B21

TY, NO, IT CAN'T BE SAID, THIS LINE DOES NOT PASS THROUGH EACH X MEAN.

UN > TYPE: YES OR NO

QU > UP TO THIS POINT, THE MANNER IN WHICH THE GENERALIZED CORRELATION COEFFICIENT
WAS DERIVED, THE MANNER IN WHICH THE REGRESSION COEFFICIENTS WERE DETERMINED, AND THE
MANNER IN WHICH EQUATIONS 11.3 AND 11.4 WERE EXPRESSED DEPENDED HIGHLY ON THE
MEAN VALUES OF Y AND THE MEAN VALUES OF X. IN FACT, A STRONG EMPHASIS
HAS BEEN PLACED ON THE FACT THAT THE LINES REPRESENTING THE TREND IN THE DATA
(CALLED REGRESSION LINES) PASS THROUGH THE ROW AND COLUMN MEANS OF THESE DATA. THIS
CAN BE SEEN IN EQUATIONS 11.3 AND 11.4 IN THAT THE LEFT SIDES OF THESE
EQUATIONS ARE GIVEN IN TERMS OF MEAN SCORES, BUT, AS WAS JUST POINTED OUT
ABOVE, WE HAVE A REGRESSION LINE WHICH DOES NOT PASS THROUGH ALL X MEANS.
SCORES, THUS, IT WOULD SEEM THAT THERE IS A DISCREPANCY BETWEEN WHAT WE HAVE
DEVELOPED AND REAL SITUATIONS. Fortunately, THERE IS A SOLUTION TO THIS.

IT CAN BE SHOWN (SHIRLEY REFERENCE, PP. 134-136) AND SHOULD BE EMPHASIZED
THAT THE REGRESSION EQUATIONS MAY BE EXPRESSED IN MORE GENERAL TERMS WITHOUT
AFFECTING THE VALIDITY OF THE GENERALIZED CORRELATION COEFFICIENTS OR THE
VALIDITY OF THE REGRESSION COEFFICIENTS AS HELD MATHEMATICALLY DEVELOPED AS THE SLOPES
OF THESE EQUATIONS. THAT IS, THESE EQUATIONS WOULD NOT REQUIRE THAT
THE LINES PASS THROUGH EITHER COLUMN OR ROW MEANS RESPECTIVELY.

THESE MORE GENERAL REGRESSION EQUATIONS ARE GIVEN IN SLIDE 11 AS EQUATIONS 11.5
AND 11.6.

THE BASIS FOR EXPRESSING THE REGRESSION EQUATIONS IN THIS MORE GENERAL FORM
IS A MATHEMATICAL PRINCIPLE KNOWN AS "LEAST SQUARES." LINES DEVELOPED UNDER
THIS PRINCIPLE ARE KNOWN AS LINES OF BEST FIT. VERY FUNDAMENTALLY, THIS
PRINCIPLE REQUIRES THAT THE AVERAGE OF THE SQUARE OF THE VERTICAL
ALGEBRAIC DIFFERENCE OF POINTS IN A PLOT OF SCORES TO THE LINE OF BEST FIT
DRAWN THROUGH THESE POINTS IS LESS THAN ANY OTHER LINES THAT COULD BE
DRAWN THROUGH THESE POINTS. IN KEEPING WITH THE INITIAL REGRESSION
LINES THAT WERE DEFINED, IF YOU HAVE THE CASE WHERE A LINE MAY BE
DRAWN THROUGH ALL OF THE MEANS OF, SAY THE COLUMN SCORES, THIS WOULD NECESSARILY
REPRESENT THE LINE OF BEST FIT FOR THESE POINTS, WITHOUT ANY FURTHER
SUBSTANTIATION. THIS IS SO BECAUSE OF A PRINCIPLE YOU LEARNED IN YOUR
BASIC STAT COURSE. IT STATES THAT MINIMUM SCORE DEVIATIONS OCCUR WHEN
DEVIATIONS ARE FROM THE MEAN OF THOSE SCORES. THUS IT MAY BE SAID THAT
THE REGRESSION LINES AS FIRST DEVELOPED ARE "SPECIAL CASES" OF THE MORE
GENERAL NOTION OF LINES OF BEST FIT.

QU: LET'S SUMMARIZE WHAT WE HAVE JUST DISCUSSED.

IS THERE A SLIGHT DISTINCTION BETWEEN A REGRESSION LINE AND A LINE OF BEST FIT AS
Presented?

CA: YES!
TY: RIGHT.
BR: PR

CA: NO!
TY: THERE IS A SLIGHT DIFFERENCE AS WE HAVE DEFINED THEM.
BR: PR

UN: TYPE: YES OR NO
PR: TY

WE USED THE TERM REGRESSION LINE WHEN WE WERE DEVELOPING THE NOTION OF
REGRESSION COEFFICIENT AS A POSSIBLE MEASURE OF CORRELATION. IN THIS
RESPECT, REGRESSION LINES WERE EXHIBITED AS THOSE LINES WHICH PASS
THROUGH THE RESPECTIVE X AND Y MEANS OF A SCATTER OF POINTS, NOTING
THAT THERE ARE CASES (THE GREAT MAJORITY, TO BE EXACT) WHEREIN THE
THE REGRESSION LINE DID NOT PASS THROUGH A SERIES OF COLUMN OR ROW MEANS,
WE THEN DEFINED THE NOTION OF LINE OF BEST FIT. THIS WAS
DETERMINED TO BE A MORE GENERAL FORM OF THE REGRESSION LINE SINCE IT
DID NOT ASSUME THAT IT NECESSARILY PASS THROUGH COLUMN OR ROW MEANS.

QU: WHICH OF THE FOLLOWING IS A DESCRIPTION OF A LINE OF BEST FIT AS DEFINED?

A. THAT LINE WHEREIN THE SUM OF THE DEVIATIONS OF POINTS IN THE PLOT FROM THE LINE
   IS A MINIMUM.
B. THAT LINE WHEREIN THE SUM OF THE AVERAGED SQUARED HORIZONTAL OR
   VERTICAL DISTANCES (DEPENDING ON WHETHER YOU ARE CONSIDERING ONE OR
   THE OTHER OF THE TWO VARIABLES) BETWEEN POINTS IN THE PLOT AND THE LINE IS
   A MINIMUM.
C. That line wherein the sum of the averaged squared deviations of points in the plot from score means on the line is a minimum.

CA B)
TY >-You are correct.

CA A)
TY >-Not quite. It is important to remember that, according to the principle of least squares, the line of best fit is that line wherein the sum of the averaged squared horizontal or vertical deviations of points in the plot from the line is at a minimum.

CA C)
TY >-Not quite. Remember, we did not require the line of best fit to pass through score means. Further more, the line of best fit does require that the averaged squared distances from the points to the line be measured in a horizontal or vertical fashion, depending on the variable under consideration.

UN >-Type: A B or C

QU >-Does this "redefinition" of the regression line as the line of best fit alter the validity and/or applicability of our generalized correlation coefficient or that of the regression coefficients?

CA NO!
TY >-Right. More information on this matter is given in Ghiselli, pp. 134-136, as was mentioned.

CA YES!
TY >-No it doesn't. For a more in-depth discussion on this matter, refer to Ghiselli, pp. 134-135.

UN ANSWER: YES OR NO

DM III
QU >-The regression coefficient and the problem of uniqueness of measurement.

>After this rather lengthy aside into regression lines, lines of best fit, and regression coefficients, let us get back to the main tract. As you may recall, we reached a point wherein it was felt that the linear regression coefficients might be useful as measures of relationship. However, since we have already developed generalized correlational coefficients, it would expedient to try to relate these general coefficients to the computational definitions of our linear regression coefficients. This is done in slides 12 and 13. Equations 13.2 and 13.3 express the generalized correlation coefficients in terms of their respective regression coefficients. Note, however, the
Let us call the deviation of a score from the regression line as "d" where \( \bar{y}_i = b_{y,x} x \) is the equation for the line in deviation score form.

Thus \( d = y_i - \bar{y}_i \) where \( y_i \) is any score in column "i" (or for a fixed \( X \)).

Notice, that we are assuming once again that the regression line passes through the column means (thus \( \bar{y}_i \)).

Now \( \frac{1}{n} \sum d^2 = \frac{1}{n} \sum (y_i - \bar{y}_i)^2 = \gamma_{y,x} \) from before.

Since \( \gamma_{y,x} = \gamma_y - \gamma_y' \) from our deviation of the generalized correlation coefficient

\[
\frac{1}{n} \sum d^2 = \gamma_{y,x} = \gamma_y - \gamma_y'
\]

From the determination of the "new" regression line as lines of best fit, it can be shown (Ghiselli, page 135) that

\[
\gamma_{y,x}' = \gamma_y - b_{y,x} \gamma_x
\]

Thus

\[
\gamma_y - \gamma_y' = \gamma_y - b_{y,x} \gamma_x
\]

or

\[
b_{y,x} \gamma_x = \gamma_y'
\]

and

\[
\gamma_y' = b_{y,x} \gamma_x
\]
Similarly it could be shown that

\[ \gamma_{x'} = b_{x',y} \gamma_y \]  \hspace{1cm} (13.1)

Dividing by \( \gamma_y \) in equation 12.1 and \( \gamma_x \) in equation 13.1

we have:

\[ \gamma_{y,x} = \frac{\gamma_{x',y}}{\gamma_y} = b_{y,x} \frac{\gamma_x}{\gamma_y} \]  \hspace{1cm} (13.2)

and

\[ \gamma_{x,y} = \frac{\gamma_{x',y}}{\gamma_x} = b_{x',y} \frac{\gamma_y}{\gamma_x} \]  \hspace{1cm} (13.3)
MANNER IN WHICH "D" IS DEFINED IN SLIDE 12. IT IS THE DIFFERENCE BETWEEN AS:<-----
POPULATION OF RIVARIATE SCORES WITH RESPECT TO THESE TWO VARIABLES.<-----
THAT THIS Y SCORE IN THE LINE IS TAKEN TO BE THE MEAN Y SCORE OF C.<-----
THAT COLUMN OF Y SCORES, THUS, IN A SENSE, WE HAVE REGRESSED TO SOMEWHAT THE:<-----
SAME SITUATION AS WE HAD DISCUSSED BEFORE WITH RESPECT TO THE:<-----
REGRESSION LINES. THEREFORE, IT MAY BE QUESTIONED, ONCE AGAIN,<-----
WHETHER EQUATIONS 13.2 AND 13.3 BASED THE "D" ARE GENERALLY:<-----
APPLICABLE TO SITUATIONS WHERE THE REGRESSION LINES DO NOT PASS THROUGH THE COLUMN:<-----
AND/OR ROW MEANS. BUT, ALSO AS BEFORE, THIS WILL PROVE TO BE NO PROBLEM IN:<-----
THE APPLICABILITY OF 13.2 AND 13.3. THAT IS, THESE ARE VALID COEFFICIENTS:<-----
WHETHER THE LINES OF BEST FIT PASS THROUGH COLUMN OR ROW MEANS OR NOT.<-----
WOULD YOU, AT THIS POINT, LIKE TO KNOW THE RATIONALE BEHIND THIS?<-------->
CA YES  
BR RATION|
CA NO I
BR CONT I
UN > TYPE:  YES OR NO<-----
> <------
RATION <-----
RD THE JUSTIFICATION FOR OUR USE OF THE CORRELATION COEFFICIENT IN A MORE<-----
GENERAL WAY THAN UNDER THOSE CONDITIONS UPON WHICH THEY WERE DEVELOPED<-----
RESTS ON A RELATIVELY BASIC, BUT IMPORTANT ISSUE IN STATISTICS, AS AS<-----
WHOLE, AND HERE IN PARTICULAR, IF YOU WILL REFER TO SLIDE 8.<-----
AND LINE "B", THIS POINT MIGHT BE ILLUSTRATED. NOTICE THAT THE X MEANS TEND TO<-----
SCATTER RANDOMLY ABOUT THIS REGRESSION LINE. NOTICE ALSO THAT THE NUMBER<-----
OF SCORES ON WHICH EACH MEAN IS BASED IS RELATIVELY SMALL. AN ASSUMPTION<-----
WHICH WE HAVE TO MAKE IS THAT UNDER THE CONDITIONS THAT WE HAD ALL<-----
Possible SCORES OF THIS NATURE, IF ONLY A LIMIT ED SAMPLE OF THEM,<-----
The MEANS WOULD CHANGE SLIGHTLY AND WOULD FALL ON THE REGRESSION LINE.<-----
VIEWED IN ANOTHER WAY, IF WE TOOK SAMPLES OF SCORES FROM A HYPOTHETICAL<-----
POPULATION OF RIVARIATE SCORES WITH RESPECT TO THESE TWO VARIABLES<-----
OF INTEREST AND PLOTTED THEIR MEAN AND COLUMN MEANS. THESE MEANS WOULD TEND TO<-----
FLUCTUATE RANDOMLY ABOUT THE LINES OF BEST FIT FOR EACH SAMPLE. IF WE<-----
TOOK THE MEANS OF ALL OF THE RESPECTIVE SAMPLE MEANS, THE MARKED TENDENCY<-----
WOULD BE THAT THESE MEANS WOULD LIE ON THEIR RESPECTIVE LINES OF BEST FIT.<-----
THUS, IT IS ASSUMED THAT IN A THEORETICAL SENSE, THE LINES OF BEST FIT<-----
WOULD PASS THROUGH THE MEANS OF THE COLUMN OF RAW SCORES. IN THIS SITUATION.

THE DEVIATION "D" IN SLIDE 12 WOULD BE THE SCORE DEVIATION FROM THE Y VALUE

ON LINE OF BEST FIT AND WOULD ALSO BE THE SCORE DEVIATION FROM THE MEAN OF THE

Y SCORES FOR THAT COLUMN. THE SAME IS TRUE WITH RESPECT TO THE X SCORES.

PLEASE PRESS THE RETURN KEY WHEN YOU ARE READY.

CONT.

Q2. "A GENERALIZED STATEMENT OF RELATIONSHIP BETWEEN THE VARIABLES WHICH IS

BASED ON THE PROPENSITY OF ACCOUNTABLE VARIATION IN ONE OF THE VARIABLES.

BY THE OTHER, WE HAVE RESTATED THIS INDEX IN TERMS OF THE REGRESSION

COEFFICIENT RELATIVE TO THAT PARTICULAR GENERALIZED CORRELATION.

COEFFICIENT. WHICH OF THE FOLLOWING IS AN ADVANTAGE IN RESTATING THE?

CORRELATION IN THESE TERMS?

1. THE CORRELATION COEFFICIENT IS NOW CAPABLE OF DEFINING BOTH POSITIVE

AND NEGATIVE STATISTICAL RELATIONSHIPS.

2. WE NOW HAVE OUR GENERALIZED CORRELATION COEFFICIENTS EXPRESSED IN TERMS

OF A LINEAR FORM OF RELATIONSHIP WHICH IS THE MOST COMMON TYPE.

3. THERE IS REALISTICALLY NO ADVANTAGE IN RESTATING OUR GENERALIZED

COEFFICIENT IN TERMS OF THE REGRESSION COEFFICIENTS.

CA 1

TY > YOU ARE CORRECT. SINCE THE REGRESSION COEFFICIENT CAN BE EITHER POSITIVE OR

NEGATIVE (BECAUSE IT IS THE SLOPE OF THE REGRESSION LINE), OUR CORRELATION

IS NOT NOW RESTRICTED TO ONLY POSITIVE STATISTICAL RELATIONSHIPS. FURTHERMORE.

NUMBER 2 ABOVE IS ALSO AN ADVANTAGE.

CA 2

TY > YOU ARE CORRECT. MOREOVER, NUMB 2 ABOVE IS LIKEWISE AN ADVANTAGE.

SINCE THE REGRESSION COEFFICIENT CAN BE EITHER POSITIVE OR NEGATIVE

(BECAUSE IT IS THE SLOPE OF THE REGRESSION LINE), OUR CORRELATION IS NOT NOW

RESTRICTED TO ONLY POSITIVE STATISTICAL RELATIONSHIPS.

CA 12

CA 13

CA 31

TY > NOT AT ALL. BOTH NUMBERS 1 AND 2 ABOVE ARE ADVANTAGES. WITH RESPECT TO NUMBER 1.

SINCE THE REGRESSION CAN BE EITHER POSITIVE OR NEGATIVE (BECAUSE IT IS THE<
Slope of the regression line. Our correlation is not now restricted to only positive statistical relationships.

Now that we have indicated two advantages for restating the generalized correlation coefficients in terms of their respective regression coefficients, which of the following is a fairly apparent disadvantage?

A. The correlation coefficients now assume only linear relationships.
B. Compared to the generalized correlation coefficients as first defined, the new expressions of relationships are more difficult to compute.
C. There are still two correlation coefficients for the same two variables.

This is a true statement but it cannot be considered a disadvantage. Remember that if a more general form of statistical relationship is desired, one can always apply the original correlation coefficients developed. The disadvantage of our present situation is that we still have two correlation coefficients for the same two variables.

This is not necessarily true and therefore not a relative disadvantage. The disadvantage of our present situation is that we still have two correlation coefficients for the same two variables.

Depending on whether we use one or the other of the two coefficients, we will get two distinct measures of correlation since each of the two regression coefficients used will not usually be the same. Furthermore, as you can see in equations 13.2 and 13.3 of slide 13, the correlations were given not only in terms of the regression coefficients but also in terms of the ratio of the standard deviations of X and Y. The ratios will not necessarily be identical either.
**Transformation of Raw Scores as a Solution to the Uniqueness Problem**

The most feasible way of getting out of this first problem would be to do something that would result in the equalizing of the two regression coefficients. If this is possible, the next step would be to find another way of equalizing the standard deviations of X and Y. In this way, the ratio of the standard deviation of Y to that of X, if both these "transformations" are possible, we would then have a unique measure of correlation.

Probably the best way to approach the solution of our first problem would be to attempt initially to determine why we get two regression coefficients. Of the following, which one might you suggest as being a possible reason?

- Why we get two regression coefficients from one set of bivariate data?
- The choice of the coordinate system on which the points are plotted affects the magnitude of the regression coefficients.
- The fact that there are two regression coefficients is merely an unavoidable function of the data on which they are computed.
- The magnitude of the regression coefficients are inherently dependent on the underlying scales of measurement of the raw scores on which they were based. Thus, if the two variables are based on different scales of measurement, it is almost certain that there will be distinct regression lines and regression coefficients.

**Possible Reasons**

- The choice of the coordinate system on which the points are plotted affects the magnitude of the regression coefficients.
- The fact that there are two regression coefficients is merely an unavoidable function of the data on which they are computed.
- The magnitude of the regression coefficients are inherently dependent on the underlying scales of measurement of the raw scores on which they were based. Thus, if the two variables are based on different scales of measurement, it is almost certain that there will be distinct regression lines and regression coefficients.

**Possible Outcomes**

- Very good
- Contia
- CA AI
- CA BI
- UN Type: A B C

Indirectly, and incompletely this way be correct. However, the primary reason for distinct regression lines and coefficients is due to the underlying scales of measurement of the two variables. If they differ for each variable, there will most likely be two distinct lines and coefficients.

**Contia**

Now, if we could equate the scales of measurement to the two variables, such that they were comparable, our first problem would be solved in that.
THE TWO REGRESSION COEFFICIENTS WOULD BE IDENTICAL IN MAGNITUDE.
FROM YOUR STAT EXPERIENCE OR POSSIBLY YOUR BACKGROUND IN MEASUREMENT,
THEORY, WHAT IS ONE WAY OF EQUATING SCALES OF MEASUREMENT?

CA(l) & STANDS;SORES
TY > GOOD THINKING
BR NG1
UN > GOOD TRY, LET'S TRY TO GET THIS ANSWERED IN ANOTHER WAY.

MIGHT BE A WAY OF EQUATING SCALES OF MEASUREMENT?
A. BY SOME LINEAR TRANSFORMATION OF THE SCALES
B. BY STANDARDIZING THE UNDERLYING RAW SCORES OF THE SCALES
C. BY EQUATING THE POSSIBLE RANGES OF SCORES ON BOTH SCALES OF MEASUREMENT

BR PR
QU
CA A1
TY > CORRECT
BR NG1
CA A1
TY > TO A POINT YOU ARE CORRECT, HOWEVER, MORE SPECIFICALLY, THE DESIRED WAY
OF EQUATING SCALES OF MEASUREMENT WOULD BE TO STANDARDIZE THE UNDERLYING
RAW SCORES OF THESE SCALES, WHICH IS BASICALLY A LINEAR TRANSFORMATION.
IN OTHER TERMS, THIS TRANSFORMS THE RAW SCORES INTO ANOTHER SCALE OF
MEASUREMENT WHOSE UNITS OF MEASUREMENT ARE COMPARABLE TO ANY OTHER SET OF
RAW SCORES WHICH HAVE BEEN TRANSFORMED SIMILARLY.

BR CONT2
CA C1
TY > IT IS DOUBTFUL THAT THIS WOULD HELP AT ALL.
THE DESIRED WAY OF EQUATING
SCALES OF MEASUREMENT WOULD BE TO STANDARDIZE THE UNDERLYING RAW SCORES OF
THESE SCALES, WHICH IS BASICALLY A LINEAR TRANSFORMATION. IN OTHER TERMS,
THIS TRANSFORMS THE RAW SCORES INTO ANOTHER SCALE OF MEASUREMENT WHOSE
UNITS OF MEASUREMENT ARE COMPARABLE TO ANY OTHER SET OF RAW SCORES WHICH
HAVE BEEN TRANSFORMED SIMILARLY.

BR CONT2
UN > TYPE: A B CP
NG1
PR
TY > A BASIC TECHNIQUE IN MEASUREMENT TO ACCOMPLISH THE TASK IS TO STANDARDIZE THE
UNDERLYING RAW SCORES. IN OTHER TERMS, THIS TRANSFORMS THE RAW SCORES INTO
ANOTHER SCALE OF MEASUREMENT WHOSE UNITS OF MEASUREMENT ARE COMPARABLE TO
ANY OTHER SET OF RAW SCORES WHICH HAVE BEEN TRANSFORMED SIMILARLY.

BR CONT2 1

CONT2 1

RE > NOTE THAT A VERY FORTUNATE BY-PRODUCT OF THIS FORM OF A TRANSFORMATION LIES IN THE
FACT THAT VARIANCE OF SUCH "STANDARD SCORES" IS ONE. THUS, BY TRANSFORMING THE
RAW SCORES OF BOTH VARIABLES INTO STANDARD SCORES, WE NOT ONLY EQUATE THE
REGRESSION COEFFICIENTS BUT ALSO EQUATE THE VARIANCES OF BOTH VARIABLES.
A MATHEMATICAL PROOF OF THIS IS IN SLIDE 14. PRESS THE RETURN KEY WHEN YOU ARE READY TO CONTINUE.

QU > THE PEARSON PRODUCT-MOMENT CORRELATION TO THIS POINT, THROUGH A SERIES OF DERIVATIONS, WE HAVE DEVELOPED
A GENERALIZED CORRELATION COEFFICIENT IN TWO FORMS, HAVE EXPRESSED THESE COEFFICIENTS IN TERMS OF THEIR RESPECTIVE LINEAR REGRESSION COEFFICIENTS.
AND LASTLY HAVE SHOWN THAT UNDER A CERTAIN CONDITION (STANDARDIZATION OF OF BOTH UNDERLYING SCALES) ARE IDENTICAL, THUS WE NOW HAVE A SINGLE, UNIQUE
MEASURE OF LINEAR RELATIONSHIP BETWEEN TWO VARIABLES WHICH HAS THE DESIRED CAPABILITY OF EXPRESSING BOTH POSITIVE AND NEGATIVE RELATIONSHIPS BETWEEN
VARIABLES. BECAUSE OF THIS CAPABILITY, IT IS NOW POSSIBLE TO OBTAIN A PERFECT
NEGATIVE RELATIONSHIP OF -1.00. AS BEFORE, A PERFECT POSITIVE RELATIONSHIP IS MEASURED AS + 1.00 WITH NO RELATIONSHIP VALUED AT 0.00.
UNDER THESE CIRCUMSTANCES, THE CORRELATION FOUND IS MORE POPULARLY KNOWN AS THE PEARSON PRODUCT-MOMENT CORRELATION. REFERRING TO SLIDE 13, WE HAVE THIS DEFINED IN EQUATION FORM.

IN THE DERIVATION OF THE PEARSON R, AS IT IS KNOWN FROM THE GENERALIZED CORRELATION COEFFICIENT TWO ASSUMPTIONS EITHER HAVE BEEN IMPLIED OR STATED, WHICH OF THE FOLLOWING PAIRS OF ASSUMPTIONS BEST DESCRIBES THOSE OF THE PEARSON R?

A. - A LINEAR RELATIONSHIP BETWEEN VARIABLES
   - A HOMOSCEDASTIC RELATIONSHIP BETWEEN VARIABLES
B. - A REPRESENTATIVE SAMPLE (FROM A POPULATION) OF SCORES IS USED
   - THESE SCORES HAVE AN UNDERLYING LINEAR RELATIONSHIP
C. - A REPRESENTATIVE SAMPLE (FROM A POPULATION) OF SCORES IS USED
   - A HOMOSCEDASTIC RELATIONSHIP EXISTS BETWEEN THESE SCORES

EXCELLENT. IT IS ASSUMED IN THE USE OF THIS CORRELATION COEFFICIENT THAT
The equality of Regression Coefficients When
Raw Scores are Transformed to Standard Scores

From our derivation of the regression coefficients, it was found that

\[ b_{y,x} = \frac{\Sigma x'y}{N\sigma_x^2} \]

and

\[ b_{x,y} = \frac{\Sigma x'y}{N\sigma_y^2} \]

When the raw scores are transformed into standard scores in the usual manner (i.e., \( z_x = \frac{x - \bar{x}}{\sigma_x} \)), the regression coefficients are now expressed as:

\[ b_{y,x} = \frac{\Sigma z_xz_y}{N\sigma_x^2} \]

\[ b_{x,y} = \frac{\Sigma z_xz_y}{N\sigma_y^2} \]

Since the variance of a set of standard scores is equal to unity, we have the following:

\[ b_{y,x} = \frac{\Sigma z_xz_y}{N} \quad \text{and} \quad b_{x,y} = \frac{\Sigma z_xz_y}{N} \]

Since the right hand side of both of these equations is identical, we have our desired result, namely

\[ b_{y,x} = b_{x,y} \]

Furthermore, since \( \sigma_{z_x}^2 = \sigma_{z_y}^2 = 1 \), our expression for the correlation coefficients are

\[ \frac{\sigma_{z_y'}}{\sigma_{z_y}} = \frac{b_{y,x}\sigma_{z_x}}{\sigma_{z_y}} \quad \text{and} \quad \frac{\sigma_{z_x'}}{\sigma_{z_x}} = \frac{b_{x,y}\sigma_{z_y}}{\sigma_{z_x}} \]

\[ \text{or} \quad \frac{\sigma_{z_y'}}{\sigma_{z_y}} = b_{y,x} \quad \text{and} \quad \frac{\sigma_{z_x'}}{\sigma_{z_x}} = b_{x,y} \]
Definition of the Pearson r and Its Relationship to the Regression Equations

Because it was determined that \( b_{y.x} = b_{x.y} \) it follows that

\[
\frac{C_{z_y}}{C_{z_y}} = \frac{C_{z_x}}{C_{z_x}}
\]

Thus,

\[
\frac{C_{z_y}}{C_{z_y}} = \frac{C_{z_x}}{C_{z_x}} = b_{y.x} \quad b_{x.y} = \frac{\sum z_x z_y}{N} = r_{xy}
\]

Or, just

\[
r_{xy} = \frac{\sum z_x z_y}{N}
\]

(Pearson Product-Moment Correlation)

In terms of standard scores and the Pearson coefficient, the regression equations become:

\[
\bar{Z}_{y_1} = r_{xy} \bar{Z}_{x_i}
\]
and
\[
\bar{Z}_{x_i} = r_{xy} \bar{Z}_{y_1}
\]

Or, more generally:

\[
Z'_{y_1} = r_{xy} Z_{x_i}
\]
and
\[
Z'_{x_i} = r_{xy} Z_{y_1}
\]

In raw score form, these equations become:

\[
\bar{Y}_i = r_{xy} \frac{C_y}{C_x} (x_i - \bar{x}) + \bar{Y}
\]
and
\[
\bar{X}_i = r_{xy} \frac{C_x}{C_y} (y_i - \bar{y}) + \bar{x}
\]
or
\[
Y'_i = r_{xy} \frac{C_x}{C_y} (x_i - \bar{x}) + \bar{Y}
\]
and
\[
X'_i = r_{xy} \frac{C_y}{C_x} (y_i - \bar{y}) + \bar{x}
\]
THE UNDERLYING RELATIONSHIP BETWEEN THE VARIABLES UNDER CONSIDERATION IS LINEAR IN NATURE. USE OF THIS COEFFICIENT WHEN THE RELATIONSHIP IS NOT LINEAR WILL GIVE A FALSE REPRESENTATION OF THE MAGNITUDE OF THE TRUE RELATIONSHIP. IT ALSO IS ASSUMED THAT A HOMOSCEDASTIC RELATIONSHIP EXISTS BETWEEN THE VARIABLES. THAT IS, IT IS ASSUMED THAT THE SPREAD (VARIANCE) OF SCORES ABOUT THE BEST FITTING STRAIGHT LINE IS APproximately THE SAME AT ALL LEVELS OF BOTH VARIABLES. AS YOU MAY RECALL, IF THE SPREAD IS NOT ROUGHLY THE SAME, THE RELATIONSHIP IS MORE COMPLEX THAN "ASSUMED" TO BE AND THE PRODUCT-MOMENT CORRELATION WOULD FAIL TO REVEAL THIS INFORMATION. FURTHERMORE, INTERPRETATION OF THIS PARTICULAR COEFFICIENT WOULD BE DIFFICULT UNDER THESE CIRCUMSTANCES.

CA B]

TY > PARtLY CORRECT. FOR GENERALIZATION OR FOR INFERENTIAL PURPOSES, THE FACT THAT THE SAMPLE OF SCORES ON WHICH THE CORRELATION IS COMPUTED BE REPRESENTATIVE OF SOME POPULATION OF SCORES IS IMPORTANT. HOWEVER, THIS IS NOT AN ASSUMPTION OF THE PEARSON R. IT IS ASSUMED IN THE USE OF THIS CORRELATION.

COEFFICIENT THAT THE UNDERLYING RELATIONSHIP BETWEEN THE VARIABLES UNDER CONSIDERATION IS LINEAR IN NATURE. USE OF THIS COEFFICIENT WHEN THE RELATIONSHIP IS NOT LINEAR WILL GIVE A FALSE REPRESENTATION OF THE MAGNITUDE OF THE TRUE RELATIONSHIP. FURTHERMORE, IT IS ASSUMED THAT A HOMOSCEDASTIC RELATIONSHIP EXISTS BETWEEN THE VARIABLES. THAT IS, IT IS ASSUMED THAT THE SPREAD (VARIANCE) OF SCORES ABOUT THE BEST FITTING STRAIGHT LINE IS APPROXIMATELY THE SAME AT ALL LEVELS OF BOTH VARIABLES. AS YOU MAY RECALL, IF THE SPREAD IS NOT ROUGHLY THE SAME, THE RELATIONSHIP IS MORE COMPLEX THAN "ASSUMED" TO BE AND THE PRODUCT-MOMENT CORRELATION WOULD FAIL TO REVEAL THIS INFORMATION. ALSO, INTERPRETATION OF THIS PARTICULAR COEFFICIENT WOULD BE DIFFICULT UNDER THESE CIRCUMSTANCES.

CA C]

TY > PARtLY CORRECT. FOR GENERALIZATION OR FOR INFERENTIAL PURPOSES, THE FACT THAT THE SAMPLE OF SCORES ON WHICH THE CORRELATION IS COMPUTED BE REPRESENTATIVE OF SOME POPULATION OF SCORES IS IMPORTANT. HOWEVER, THIS IS NOT AN ASSUMPTION OF THE PEARSON R. IT IS ASSUMED THAT A HOMOSCEDASTIC RELATIONSHIP EXISTS BETWEEN THE VARIABLES. THAT IS, IT IS ASSUMED THAT THE SPREAD (VARIANCE) OF SCORES ABOUT THE BEST FITTING STRAIGHT LINE IS APPROXIMATELY THE SAME AT ALL LEVELS OF BOTH VARIABLES. AS YOU MAY RECALL, IF THE SPREAD OF SCORES IS NOT ROUGHLY THE SAME,
The relationship is more complex than "assumed" if we and the product-

moment correlation would fail to reveal this information. Furthermore-

interpretation of this particular coefficient would be difficult under-

these circumstances.

It is also assumed in the use of this correlation that the underlying-
relationship between the variables under consideration is linear.

In nature, use of this coefficient when the relationship is not linear-

will give a false representation of the magnitude of the true relationship.

un >-type: a 3 or (<-------->)

qu >-a final assumption of the Pearson R was not alluded to in the development of-

this correlation. Although somewhat related to the notion of homoscedasticity-

this assumption specifies that scores on one variable are distributed normally at each-

respective values of the other variable and vice versa. Furthermore, this would-

imply that the scores of each of the variables under consideration are-

distributed normally also.<------>

this assumption (as with the other two) is central to the mathematical-

foundations of this correlation.<-------->

do you think that departure from this assumption as indicated in actual data-

would seriously affect the use of this correlation? answer yes or no.<------>

ca no

br pr 1

can yes!

Ty >-you are incorrect, but only to a degree.<-------->

un >-type: yes or no<-------->

pr i

ty >-as with many statistics which assume normality of underlying data.<------>

departure from the normal assumption does not seriously affect the use of-

the Pearson R, however, the assumption of normality does play a major role in-

the development of the inferential statistical properties of this coefficient.<------>

correlation. That is, this assumption has greater impact with respect-

to a fuller interpretation of the coefficient than with respect to its common-

application. This notion will be dealt with shortly.<-------->

qu>-although a departure from normality in the data does not prohibit the use of-

the Pearson coefficient, an important consideration does concern the relative-

distributions of the variables. If the distributions of these variables-

differ from one another (in terms of their shapes - skewness<------>

and kurtosis), it can be shown that the maximum value of R is not 1.00 but<------>
This situation but some value less than this depending on the degree of...

Disparity between the distributions. This may only be an academic point...

Since it is rare that a perfect relationship between typical variables studied...

In the behavioral sciences is found, however, care should be taken...

To determine, at least to a degree, the relative shapes of the...

Variable distributions. In this way one has a notion as to the "worth" of his...

Coefficient.

To get away from the flow of the course for a moment, let us talk about...

A more practical value of the correlation coefficient and its limitations...

If you will recall, it was stated near the outset of this course that...

There were two major uses of the correlational coefficient (specifically Pearson r).

One of these was indirectly made reference to in the development of our...

Generalized correlation coefficient. This was to attempt to determine the...

Extent to which individual differences on two measures were due to the...

Same underlying factors. Do you recall what the other one was?

Call [Predict].

Ty >-good.

Ty B2]

Ty >-<----------->

Br cont2al

Un >-give it one more try.<---------->

Un >-which of these might be the other purpose.<---------->

A. To determine the slope of the regression line.<---------->

B. To "predict" scores on one of the variables from scores on the other.<---------->

C. To establish cause and effect relationships between the variables under...

Consideration.<---------->

Br or

Qu 1

Ca 5]

Ty >-that's it.<---------->

Ca A]

Ty >-the slope of the regression line can be determined much more efficiently...

Than by using the correlation coefficient. The second purpose is for "prediction"

Of scores on one variable from those of the other.<---------->

Ca C]

Ty >-cause and effect relationships cannot typically be determined through the...

Use of a correlation coefficient. The second purpose is for "prediction" of<------>
SCORING ON ONE VARIABLE FROM THOSE OF THE OTHER.

CONT2A

THE QUESTION NOW ARISES AS TO HOW THE CORRELATION COEFFICIENT MIGHT BE UTILIZED FOR THIS PURPOSE. LET'S TAKE A ROUND-ABOUT WAY OF DEMONSTRATING THIS.

Suppose you had a distribution of scores for a known sample of persons on a variable. Call it variable V. Suppose also that you were asked to guess each person's score from a number of them you had at your disposal. What would be your guess?

**Calling **

**TALLY & MEAN**

**Correct*** Since it was most likely shown in your first stat course that the mean of a distribution is that score from which the sum of deviations is at a minimum, the most logical or the best estimate of any person's score in the distribution would be the mean of same.

**Tally & Median**

**Not quite. You're on the right track, though. Try again.**

**Tally & Mode**

**Close, but not correct. Try again.**

**Tally.**

**Give it another go.**

**Tally.**

The answer is mean. Type: **Mean**

NG2

**Even though the mean would be the best guess in such a situation, it would not be all that accurate. Upon which of the following would the degree of accuracy of this guess depend?**

**A. The variance (or standard deviation) of the distribution of scores**

**B. The shape of the score distribution (its skewness and/or kurtosis)**

**C. The underlying scale of measurement of the scores in the distribution**

**Calling A**

**Correct.** The degree of accuracy would depend on the variance (or standard deviation) of that distribution since it is itself an index of the degree of spread of scores around the mean.

**Calling B**

**Although the shape of the distribution might be a pertinent factor with respect to the degree of accuracy of the guess, it would be so only as adjunct information to the variance (or standard deviation) of the distribution.**
SINCE THE VARIANCE ITSELF IS AN INDEX OF THE SPREAD OF SCORES AROUND THE MEAN.

TY > THE SCALE OF MEASUREMENT OF THE SCORES HAS LITTLE OR NOTHING TO DO WITH THE DEGREE OF ACCURACY OF THE GUESS IN THIS RESPECT. THE DEGREE OF ACCURACY WOULD DEPEND ON THE VARIANCE (OR STANDARD DEVIATION) OF THAT DISTRIBUTION OF SCORES SINCE IT IS ITSELF AN INDEX OF THE DEGREE OF SPREAD OF SCORES AROUND THE MEAN.

UN > TYPE: A, B OR C

QU > SUPPOSE NOW WE HAD MORE INFORMATION AT OUR DISPOSAL. SAY WE HAD ANOTHER VARIABLE (VARIABLE X). LET US ASSUME FOR OUR PURPOSES THAT THERE EXISTS A STATISTICAL RELATIONSHIP BETWEEN THE SCORES ON BOTH VARIABLES. LET US FURTHER ASSUME THAT THIS RELATIONSHIP IS LINEAR IN NATURE. WITH THIS ADDED INFORMATION, IS IT NOW POSSIBLE TO BETTER "GUESS" OR PREDICT A PERSON'S SCORE ON VARIABLE Y?

NOW FROM OUR DEVELOPMENT OF THE GENERALIZED CORRELATION COEFFICIENT, RECALL THAT THE SCATTER OF Y SCORES AROUND THE REGRESSION LINE AT ANY POINT OF X WAS LESS THAN THE SCATTER (VARIANCE) OF Y SCORES IN GENERAL OR ACROSS ALL VALUES OF X. PRACTICALLY SPEAKING, THE RANGE AND VARIANCE OF Y SCORES AT ANY VALUE OF X WERE NOT TAKEN INTO ACCOUNT. THUS, IF YOU WERE TO GUESS A PERSON'S Y SCORE KNOWING HIS SCORE ON X, WHICH OF THE FOLLOWING WOULD PROBABLY BE THE BEST GUESS?

A. THE MEAN Y SCORE CORRESPONDING TO THAT SPECIFIC VALUE OF X.

B. THE MEAN OF Y.

C. THE Y VALUE ON THE REGRESSION LINE CORRESPONDING TO THAT SPECIFIC VALUE OF X.

TY > IN THE MOST GENERAL SENSE THIS IS THE CORRECT ANSWER. IN EFFECT, THE BEST...

BR PR

CA CI

TY > THE MEAN Y SCORE CORRESPONDING TO THAT SPECIFIC VALUE OF X IS A LOGICAL GUESS, TAKING INTO CONSIDERATION WHAT WAS DISCUSSED ABOVE WITH RESPECT TO THE MEAN BEING THE BEST ESTIMATE OF A SCORE IN A DISTRIBUTION. HOWEVER, IN THE MOST GENERAL SENSE WITH REGARD TO REGRESSION AND CORRELATION, THE BEST...

BR PR

CA B1

TY > THE MEAN OF Y DOES NOT TAKE INTO CONSIDERATION THE FACT THAT THERE IS ADDITIONAL INFORMATION IN THE FORM OF ANOTHER VARIABLE. IN EFFECT THE BEST...
PR 1

QU 2 - AN ILLUSTRATION OF THE PREDICTIVE USE OF THE PEARSON \( r \) IS AS FOLLOWS. USING THE SCORES OF THE SCATTERPLOT IN SLIDE 2 AS THE BASE, THE \( r \) OF THESE SCORES CAN BE CALCULATED TO BE .820. ASSUMING THAT WE KNOW A PERSON HAS AN X SCORE OF 6, WHICH OF THE FOLLOWING EQUATIONS DESCRIBES THE MANNER IN WHICH AN ESTIMATE OF HIS Y SCORE CAN BE OBTAINED?

A. \( Y^\prime = .82(3.47/2.121)(6-4) + 20.0 \)
B. \( Y^\prime = .82(6.33/2.121)(6-4) + 20.0 \)
C. \( Y^\prime = .82(2.12/6.33)(6-4) + 20.0 \)
D. \( Y^\prime = .82(3.47/6.33)(6-4) + 20.0 \)

CA A1
TY 2 - RIGHT.

BR PR 1

CA A1
TY 2 - YOU HAVE USED THE PARTIAL STANDARD DEVIATION OF Y INSTEAD OF THE STANDARD DEVIATION OF Y, USING 6.33 FOR THE STANDARD DEVIATION OF Y.

BR PR 1

CA C1
TY 2 - YOU HAVE USED THE RATIO OF STANDARD DEVIATIONS INVERTED, WITH THESE INVERTED THE EQUATION NOW BECOMES THAT AS GIVEN IN ALTERNATIVE B.

BR PR 1

CA D1

BR PR 1
Applying Equation 15.3 (as was done above) and the appropriate data, we obtain an estimate of approximately 25.0 (rounding error) for his y score. Notice in slide 5 that this estimate is the mean of the y scores for x equal to 6, and that it falls on the regression line. This supports what was said before.

With reference to the more general line of best fit, that is, in using the more general equation 15.3 (in contrast to 15.1), it is still found that the predicted y value was the column mean of y at x equal 6. Since, in this case, the column means do fall on the regression line.

Notice in the illustration that not all of the y scores with associated x scores of 6 were precisely predicted. This is quite reasonable seeing that this predicted value is only a best estimate, as you might suspect. This estimate gets better as the greater the degree of relationship between the two variables. If there is a perfect correlation between x and y, then prediction is exact. More often, however, only a limited degree of correlation exists between variables, and thus, there is a measure of error associated with each estimation. Fortunately, it is possible to get some indication of the magnitude of this error of prediction, of the following, which might you consider as being a possible index:

A. The variance of y.
B. The simple difference between the person's actual score value and his predicted value.
C. The partial variance terms.

CA. Yes. The error associated with prediction can be described as the difference between the real score and the predicted score which, by definition, is on the regression line. Since this is the basis for the individual partial variance terms, the average partial variance can be used as an index of error of prediction. Turning to slide 16, you can see the partial variance term developed as a function of the degree of relationship. It is only reasonable that this be done from what was said previously with regard to the exactness of prediction. The greater the relationship between two variables, the greater the precision of prediction of one from the other.

CA. A.

Ty. No. Although used as a measure of precision, here the overall variance of y would be too rough a measure. However, the variance of y scores.
Around the regression line would be a fairly good index of error associated with this prediction since the basis for the partial variance term is nothing more than the difference between a real score and the predicted score which lies on the regression line.

CA 1

Ty > close, but not quite. This difference does serve as the basis for the index of error associated with prediction though. Since the partial variance is founded on the difference between column and/or row scores and scores that fall on the regression line (predicted scores), this term does give an index of the precision of the prediction.

UN > type: A R or

Qu > turning to slide 15

PA 51

Ty you can see the partial variance term developed as a function of the degree of relationship between variables. It is only reasonable that this be done from what was said previously — the greater the relationship between two variables, the greater the precision of prediction of one from the other. Specifically, equations 16.1 and 16.2 in slide 16 each represent what is known as the standard error of prediction or the standard error of estimate.

One for the estimation of Y from X and the other for prediction of X from Y. They describe the degree of error included in predicting one score from another.

Equation 16.3 in this same slide is also a definition of the standard error of estimate. Only given in terms of standard scores, note that in this case, both forms of this standard error are identical.

Referring to equations 16.1 and 16.2, notice that when there is no relationship between the variables (r = 0), the standard error of estimate is equal to the standard deviation of the predicted variable, what is it equal to if there is a perfect relationship (either positive or negative) between the variables?

CA (W) zero 0

Ty > correct. Since prediction should be exact in this instance.

BR QUEST11

CA (W) one 1

Ty > no. The standard error is equal to 0 in this case since prediction should be exact.

BR QUEST11

UN > I don't recognize your answer. Try again.

UN > which of the following might the standard error equal under the condition that
Let us represent the score deviation from the regression line as before: \( y_i - \bar{y}_i \) where \( \bar{y}_i \) is that \( Y \) value which falls on the line.

Then
\[
\Sigma ( y_i - \bar{y}_i ) = \sigma_{y.x} = \sigma_y - b_{y.x} \sigma_x \]
from before.

Since \( b_{y.x} = \frac{\Sigma xy}{\Sigma x^2} \) we have
\[
\sigma_{y.x} = \sigma_y - \left( \frac{\Sigma xy}{N} \right)^2 \frac{\sigma_x^2}{\Sigma x^2}
\]

Moreover, since \( r_{xy} = \frac{\Sigma xy}{\sigma_x \sigma_y} \), \( \frac{\Sigma xy}{N} = \sigma_x \sigma_y r_{xy} \)

Substituting this into the equation, we have:
\[
\sigma_{y.x} = \sigma_y - \sigma_y r_{xy} \sigma_x^2
\]
or
\[
\sigma_{y.x} = \sigma_y (1 - r_{xy}^2)
\]
or
\[
\sigma_{y.x} = \sigma_y \sqrt{1 - r_{xy}^2}
\] 16.1

Similarly
\[
\sigma_{x.y} = \sigma_x \sqrt{1 - r_{xy}^2}
\] 16.2

If we would have used standard scores instead of deviation scores,
\[
\sigma_x = \sigma_y = 1 \quad \text{and}
\]
\[
\sigma_{y.x} = \sigma_{x.y} = \sqrt{1 - r_{xy}^2}
\] 16.3
THERE IS A PERFECT RELATIONSHIP EXISTING BETWEEN VARIABLES?

**QUEST 1**

**Q 1**

A. ZERO
B. ONE
C. .5

**CA A**

**Q 2**

**CA B**

**Q 3**

**CA C**

**Q 4**

**CA D**

**UN** -> **TYPE A B OR C**

**PR 1**

TY THE STANDARD ERROR OF ESTIMATE IS ZERO HERE SINCE PREDICTION SHOULD BE EXACT.

**QUEST 2**

**Q 1**

**Q 2**

**Q 3**

**Q 4**

**CA E**

**Q 5**

**CA F**

**Q 6**

**CA G**

**Q 7**

**CA H**

**UN** -> **TYPE A B OR C**

**PR 1**

TY THE STANDARD ERROR OF ESTIMATE IS ZERO HERE SINCE PREDICTION SHOULD BE EXACT.

**QUEST 3**

**Q 1**

**Q 2**

**Q 3**

**Q 4**

**CA I**

**Q 5**

**CA J**

**Q 6**

**CA K**

**Q 7**

**CA L**

**UN** -> **TYPE A B OR C**

**PR 1**

TY THE STANDARD ERROR OF ESTIMATE IS ZERO HERE SINCE PREDICTION SHOULD BE EXACT.

**QUEST 4**

**Q 1**

**Q 2**

**Q 3**

**Q 4**

**CA M**

**Q 5**

**CA N**

**Q 6**

**CA O**

**Q 7**

**CA P**

**UN** -> **TYPE A B OR C**

**PR 1**

TY THE STANDARD ERROR OF ESTIMATE IS ZERO HERE SINCE PREDICTION SHOULD BE EXACT.

**QUEST 5**

**Q 1**

**Q 2**

**Q 3**

**Q 4**

**CA Q**

**Q 5**

**CA R**

**Q 6**

**CA S**

**Q 7**

**CA T**

**UN** -> **TYPE A B OR C**

**PR 1**

TY THE STANDARD ERROR OF ESTIMATE IS ZERO HERE SINCE PREDICTION SHOULD BE EXACT.

**QUEST 6**

**Q 1**

**Q 2**

**Q 3**

**Q 4**

**CA U**

**Q 5**

**CA V**

**Q 6**

**CA W**

**Q 7**

**CA X**

**UN** -> **TYPE A B OR C**

**PR 1**

TY THE STANDARD ERROR OF ESTIMATE IS ZERO HERE SINCE PREDICTION SHOULD BE EXACT.
Unfortunately, in most real situations there will be varying degrees of spread around the line of best fit.

Type: Yes or No

Thus, the standard error of estimate (the average index of variability) around the regression line is just an estimate in itself, but remember, it is the best estimate we can make as to the spread of scores around the line of best fit. If we can assume that the variability of scores around the line is fairly standard at points along the line, then our "estimate" of the standard error of estimate is fairly good.

One other assumption, has implication to the standard error of estimate in terms of more detailed use and interpretation of it. If the scores in the rows and columns are normal and assuming, again, homogeneity of variance, then we know that 68% of the scores fall between plus and minus one standard error of estimate in this case. Thus, if a person's score is predicted to be 100 and the standard error of estimate is found to be 10, what is the probability that the person's true score falls in the score range of 90-110?

CA (w) .68 .68

The probability is .68 that the person's true score will fall in the interval of 90-110 under the assumptions of normality and homoscedasticity.

Question:

Most of the measuring instruments with which we deal in the behavioral sciences are considered to be based on a continuous scale of measurement. That is, they assume that there is a multitude of scores that could be recorded between the lowest and highest scores possible on the instrument. The Pearson product-moment correlation, which we have just discussed, presumes that the variables to which it is applied possess this underlying form of measurement. Many times, however, a measure of relationship is required between a variable with an underlying continuous scale and another which allows for only two possible scores, e.g., right-wrong, yes-no, etc. The question arises as to what to do when we have one "continuous" variable and one "dichotomous" variable such as the case, for example, when we would like to correlate a total test score with an item score when the item score is given as either correct or incorrect.
THE SOLUTION TO THIS IS RELATIVELY SIMPLE AND IT FOLLOWS FROM THE PEARSON R.

NOW, IS IT TRUE THAT A DICHOTOMOUS SCALE OF MEASUREMENT IS ONLY A SPECIAL CASE OF A CONTINUOUS SCALE? ANSWER YES OR NO.

CA: YES

TY: YOU ARE CORRECT.

BR: OK.

CA: NO!

TY: I'M AFRAID YOU'RE WRONG.

BR: PR.

UN: TYPE: YES OR NO.

PR: I

TY: A DICHOTOMOUS SCALE IS DEGENERATE FORM OF A CONTINUOUS SCALE IN THAT IT ALLOWS ONLY TWO POINTS IN IT WHEREAS THE CONTINUOUS SCALE WOULD ALLOW FOR AN INFINITE NUMBER MORE.

QU: WITH THIS INFORMATION IN MIND IT SEEMS REASONABLE TO ASSUME THAT THE PEARSONIAN COEFFICIENT IS APPLICABLE UNDER THIS CIRCUMSTANCE. IN THIS CASE, ONE OF THE VARIABLES WOULD HAVE ONLY SCORES OF 0 OR 1 TO REPRESENT THE TWO POSSIBILITIES ON THIS SCALE. AS A MATTER OF INFORMATION, 0 AND 1 ARE USED ONLY FOR THE SAKE OF CONVENIENCE.

ALTHOUGH DIRECT APPLICATION OF PEARSON R IS APPROPRIATE, A MORE EXPEDIENT FORM OF THIS CORRELATION COEFFICIENT IS DERIVED AND GIVEN IN SLIDES 17 AND 18.

FOR A SPECIAL CASE OF THIS NATURE, GIVEN IN THIS FORM, THE PEARSON R IS KNOWN AS THE POINT-BISERIAL COEFFICIENT OF CORRELATION.

IF YOU WILL NOTICE, EQUATION 18.1 IS, COMPUTATIONALLY, A MUCH SIMPLER ONE THAN THAT FOR THE PEARSON R (EQUATION 17.1). IT IS FOR THIS REASON THAT THE POINT-BISERIAL WAS SOMETIMES USED IN PLACE OF R BEFORE THE DAYS OF HIGH SPEED COMPUTERS.

HOWEVER, MUCH "INFORMATION" IS LOST IN ARTIFICIALLY DICHOTOMIZING ONE OF THE CONTINUOUS VARIABLES AND THEN APPLYING THE POINT-BISERIAL. AN ILLUSTRATION OF THIS IS IN SLIDE 19. HERE THE X SCALE OF THE EXAMPLE GIVEN IN SLIDE 8 HAS BEEN DICHOTOMIZED (SCORES LESS THAN OR EQUAL THE MEAN ARE NOW 0, OR SCORES GREATER THAN THE MEAN ARE NOW 1), WHICH OF THE FOLLOWING REPRESENTS THE COMPUTATIONAL FORM OF THE POINT-BISERIAL FOR THIS DATA?

A. 10 X SQRT(3/5 x 2/5) / 6.33
B. 8.75 X (3/5 X 2/5) / 6.33
C. 8.75 X SQRT(3/5 X 2/5) / 6.33

CA: C.
Derivation of the Point-Biserial $r$ from the Pearson $r$

From the definition of the Pearson $r$, we have the computational form as follows:

$$ r_{xy} = \frac{N \Sigma X Y - (\Sigma X)(\Sigma Y)}{\sqrt{N \Sigma X^2 - (\Sigma X)^2} \cdot \sqrt{N \Sigma Y^2 - (\Sigma Y)^2}} $$

Let us assume that variable $X$ is dichotomous in nature, i.e. takes on possible values of 0 and 1. Then,

$$ \Sigma X = N_1 \text{ (number of } X \text{ scores of value 1)} $$

$$ \Sigma X^2 = N_1 $$

$$ \Sigma Y = \Sigma Y_0 + \Sigma Y_1 \text{ (sums of } Y \text{ values with corresponding } X \text{ values of 0 and 1)} $$

$$ \Sigma XY = \Sigma X_0 Y_0 + \Sigma X_1 Y_1 = \Sigma Y_1 \text{ (Since all } X_0 \text{'s are 0)} $$

Substituting,

$$ N = N_0 + N_1 $$

$$ r_{xy} = \frac{N \Sigma Y_1 - N_1 \Sigma Y}{\sqrt{N N_1 - N_1 \Sigma Y^2 - (\Sigma Y)^2}} $$

$$ = \frac{(N_0 + N_1) \Sigma Y_1 - N_1 (\Sigma Y_0 + \Sigma Y_1)}{\sqrt{(N_0 + N_1)(N_1) - N_1 \Sigma Y^2} \cdot \sqrt{N N_1 \Sigma Y^2 - (\Sigma Y)^2}} $$

$$ = \frac{N_0 \Sigma Y_1 + N_1 \Sigma Y_1 - N_1 \Sigma Y_0 - N_1 \Sigma Y_1}{\sqrt{N_0 N_1 + N_1 N_1 - N_1} \cdot \sqrt{N N_1 \Sigma Y^2}} $$

$$ = \frac{N_0 \Sigma Y_1 - N_1 \Sigma Y_0}{N \Sigma Y \sqrt{N_0 N_1}} \cdot \frac{N \Sigma Y_0 - N_1 \Sigma Y_0}{N \Sigma Y \sqrt{N_0 N_1}} $$

$$ = \frac{N_0 \Sigma Y_1 - N_1 \Sigma Y_0}{N \Sigma Y \sqrt{N_0 N_1}} \cdot \frac{N \Sigma Y_0 - N_1 \Sigma Y_0}{N \Sigma Y \sqrt{N_0 N_1}} $$

$$ = \frac{N_0 N_1 Y_1 - N_0 N_1 Y_0}{N \Sigma Y \sqrt{N_0 N_1}} \cdot \frac{N \Sigma Y_0 - N_1 \Sigma Y_0}{N \Sigma Y \sqrt{N_0 N_1}} $$

Since $N_1 = \Sigma Y_1$ and $N_0 = \Sigma Y_0$
Derivation of the Point-Biserial r from the Pearson r (cont'd)

\[ r_{\text{Bis}} = \frac{(M_1 - M_0) \sqrt{P_0 P_1}}{\sigma_y} \]

Where \( P_0 \) = proportion of scores with value of 0
and \( P_1 \) = proportion of scores with value of 1.
Slide 19
Computation of the Point-Biserial $r$ Based on the Scatterplot of Scores Contained in Slide 6

<table>
<thead>
<tr>
<th>$X$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td>11</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>11</td>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td>11</td>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td>12</td>
<td>19</td>
<td>24</td>
</tr>
<tr>
<td>13</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>13</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td>14</td>
<td>22</td>
<td>24</td>
</tr>
<tr>
<td>14</td>
<td>22</td>
<td>26</td>
</tr>
<tr>
<td>15</td>
<td>23</td>
<td>26</td>
</tr>
<tr>
<td>15</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>15</td>
<td>25</td>
<td>27</td>
</tr>
<tr>
<td>16</td>
<td>27</td>
<td>29</td>
</tr>
</tbody>
</table>

Sum $Y_0 = 495$
\[ N = 30 \]
\[ \bar{Y}_0 = 16.5 \]

Sum $Y_1 = 505$
\[ N = 20 \]
\[ \bar{Y}_1 = 25.25 \]

$r_{pbis} = .6768$
THE MEAN SCORE DIFFERENCE OF THE DICHOTOMIZED VARIABLE HERE IS 8.75.

AND NOT 10. THEREFORE, C ABOVE IS THE CORRECT COMPUTATIONAL FORM.

YOU MUST HAVE READ EQUATION 13.1 INCORRECTLY. NOTICE THE SQUARE ROOT TERM.

THIS MUST BE TAKEN INTO CONSIDERATION. TRY AGAIN.

YOU MUST HAVE READ EQUATION 19.1 INCORRECTLY. NOTICED THE SQUARE ROOT TERM. THIS MUST BE TAKEN INTO CONSIDERATION. TRY AGAIN.

THE RESULTANT COEFFICIENT AFTER ACTUAL CALCULATION IS .68. IF YOU WILL

RECALL, THE PEARSON R ON THIS SAME DATA WAS .82. THUS, THERE IS IN FACT A LOSS OF "INFORMATION." ARTIFICIALLY DICHOTOMIZING THE DATA, ALTHOUGH SAVING A BIT OF COMPUTATIONAL EFFORT, RESULTS IN A SPURIOUS MUTATION AS TO THE DEGREE OF RELATIONSHIP BETWEEN THE VARIABLES OF INTEREST. THUS, WHEN APPLICABLE, THE PEARSON R SHOULD BE UTILIZED.


YES!

YOU ARE CORRECT.

YOU MUST HAVE READ EQUATION 19.1 INCORRECTLY. NOTICED THE SQUARE ROOT TERM. THIS MUST BE TAKEN INTO CONSIDERATION. TRY AGAIN.


THE PHI COEFFICIENT.

CASE WHERE BOTH OF THE VARIABLES ARE OF A DICHOTOMOUS SORT, THE MEASURE OF
RELATIONSHIP IN THIS INSTANCE IS KNOWN AS THE PHI COEFFICIENT.
WOULD IT SEEM REASONABLE TO YOU THAT THE PHI WOULD FOLLOW DIRECTLY FROM THE
PEARSON R AS DID THE POINT-BIserial?  
CA YES!
TY >> IT DOES, AND THUS IT IS FUNDAMENTALLY A PRODUCT-MOMENT CORRELATION.
CA NO!
TY >> THE ONLY DIFFERENCE IS THAT NOW THERE ARE TWO DICHOTOMOUS VARIABLES. THUS,
THE PHI DOES FOLLOW DIRECTLY FROM THE PEARSON R AND IS FUNDAMENTALLY A PRODUCT-
MOMENT CORRELATION.
UN >> TYPE: YES OR NO
QU >> AS IN THE CASE OF THE POINT-BIserial, one could directly apply the PearSon R TO
DICHOTOMOUS DATA ON TWO VARIABLES. However, a simplified version of the
PEARSON R SPECIFICALLY APPLICABLE TO SITUATIONS AS THIS IS DERIVED AND
AND GIVEN IN SLIDE 20.
UNLIKE THE POINT-BIserial, THE PHI COEFFICIENT EXHIBITS A POSSIBLE
RANGE OF VALUES FROM -1.00 TO +1.00, WHICH OF THE FOLLOWING DO YOU
THINK MIGHT BE THE REASON FOR THIS LACK OF RESTRICTION ON THE RANGE OF PHI?
IN COMPARISON TO THE RESTRICTION OF RANGE IN THE CASE OF THE POINT-
BIserial?
A. BECAUSE OF THE DIFFERENT NATURE OF THE PHI COEFFICIENT RELATIVE TO THAT
OF THE POINT-BIserial
B. BECAUSE BOTH VARIABLES ARE DICHOTOMOUS AND IT IS MORE LIKELY THAT
THEIR DISTRIBUTIONS WOULD BE SIMILAR THAN WOULD BE THOSE OF THE
POINT-BIserial.
C. BECAUSE WE ARE NOT, IN EFFECT, DEALING WITH CELL FREQUENCY SCORES ON
BOTH VARIABLES AND THUS CAPABLE OF ATTAINING THE MAXIMUM DEGREES OF
STATISTICAL RELATIONSHIP.
CA B)
BR PR 
CA A)
TY >> IF YOU CONSIDER THE DIFFERENT NATURE OF THE PHI AND POINT-BIserial TO INCLUDE THE
RELATIVE VARIABLE DISTRIBUTIONS OF EACH, THEN YOU ARE CORRECT.
BR PR 
CA C)
TY >> THE USE OF ALL CELL FREQUENCIES FOR BOTH VARIABLES HAS RELATIVELY LITTLE TO
DO WITH THE MAXIMUM VALUES ATTAINABLE BY PHI.
Let the table for X and Y values of 0 and 1 be given as follows:

\[
\begin{array}{|c|c|}
\hline
X & 0 & 1 \\
\hline
Y & c & d \\
\hline
\end{array}
\]

Let \( a, b, c, d \) represent the number of scores falling into the four possible combinations of X and Y scores.

\[
\begin{align*}
\sum X &= a \cdot 0 + c \cdot 0 + b \cdot 1 + d \cdot 1 = b + d \\
\sum Y &= a \cdot 0 + b \cdot 0 + c \cdot 1 + d \cdot 1 = c + d \\
\sum XY &= a(0 \cdot 0) + b(1 \cdot 0) + c(0 \cdot 1) + d(1 \cdot 1) = d \\
N &= a + b + c + d
\end{align*}
\]

The Pearson correlation coefficient \( r_{XY} \) can be derived as:

\[
r_{XY} = \frac{N \sum XY - \sum X \sum Y}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}}
\]

Substituting,

\[
r_{XY} = \frac{(a + b + c + d)d - (b + d)(c + d)}{\sqrt{(a + b + c + d)(b + d) - (b + d)^2(a + b + c + d)(c + d)}}
\]

\[
= \frac{ad + bd + cd + d^2 - bc - dc - bd - d^2}{\sqrt{(b + d)[a + b + c + d - b - d][c + d][a + b + c + d - c - d]}}
\]

\[
= \frac{ad - bc}{\sqrt{(b + d)(a + c)(a + b)(c + d)}}
\]

\[
\Phi = \frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}}
\]
The maximum values of the Phi coefficient, as with the other two, are still dependent on the relative shapes of the variable distributions. However, since both variables are of a dichotomous nature in the phi, it is therefore more feasible to make a comparison to the point-biserial case for their distributions to be similar. Thus, phi is capable, in this situation, of attaining maximum values.

As with the point-biserial, use of the phi coefficient on data which is suitable for application of the Pearson r results in loss of information from the user's point of view since there is a curtailment of information by dichotomizing otherwise continuous variables. For illustration, let's artificially dichotomize both variables of our original example (that given in slide 8).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>7</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>23</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The above table represents the scores of our original example after dichotomizing each variable on the basis of those scores which were less than or equal to the mean (0) and those greater than the mean (1). Now which of the following does the frequency score of 17 in the above table?

A. 17 scores had X and Y values greater than the X and Y means, respectively.
B. 17 scores had X values greater than the X mean and Y values less than or equal to the Y mean.
C. 17 scores had Y values greater than the Y mean and X values less than or equal to the X mean.

You are correct.
TY > NOT QUITE. A FREQUENCY SCORE OF 17 IN THE TABLE INDICATES THAT THERE WERE X SCORES WHICH HAD X AND Y VALUES OF 1 AFTER DICHTOMIZATION. THIS WOULD IMPLY, BY DEFINITION OF WHAT SCORES OF 0 AND 1 REPRESENT, THAT 17 SCORES HAD BOTH X AND Y VALUES GREATER THAN THEIR RESPECTIVE MEANS.

PHICOI

QU > NOW USING THE DATA IN THE ABOVE TABLE, EQUATION 20.1 OF SLIDE 20, AND THE "CALC" FUNCTION IF NECESSARY, CALCULATE THE VALUE OF PHI FOR THESE DATA. WHEN YOU HAVE CALCULATED IT, TYPE OUT THE ANSWER YOU HAVE FOUND.

CA - .73356

TY > VERY GOOD.

QU > NOTICE THE MAGNITUDE (.734) OF THE PHI COEFFICIENT RELATIVE TO THE PEARSON R CALCULATED ON THE SAME DATA (.58). BY ARTIFICIALLY DICHTOMIZING BOTH VARIABLES AND APPLYING PHI, WE LOST INFORMATION. THUS, THE PHI COEFFICIENT, IN THIS CASE WOULD NOT GIVE A VALID MEASUREMENT OF THE RELATIONSHIP OF X AND Y AS ORIGINALY DEFINED.


A. A MANIFESTATION OF THE REDUCED MAXIMUM VALUE ATTAINABLE BY THE POINT-BISERIAL.

B. A MANIFESTATION OF THE FACT THAT MORE INFORMATION IS LOST IN USING THE POINT-BISERIAL ON CONTINUOUS DATA.

C. A MANIFESTATION OF THE FACT THAT THE DATA WAS MORE AMENABLE TO DICHTOMIZATION OF BOTH VARIABLES.

CA A

TY > EXACTLY

CA B

TY > NOT TRUE. MORE INFORMATION IS LOST IN DICHTOMIZING TWO CONTINUOUS VARIABLES THAN ONE.
CA (1)
TY >-THIS WOULD NOT BE TOO LIKELY.<
UN >-TYPE: A B OR C
PR ___
TY >-THE MOST PROBABLE REASON FOR THE SMALLER POINT-BISERIAL IS THE FACT THAT VARIABLE DISTRIBUTION SIMILARITIES ARE VERY IMPORTANT WITH RESPECT TO THE SIZE OF CORRELATION COEFFICIENTS. SINCE WE KNOW THAT THE DISTRIBUTIONS OF X AND Y WILL NECESSARILY BE DIFFERENT TO SOME DEGREE IN THE CASE OF THE POINT-BISERIAL, WE WOULD EXPECT IT TO BE SMALLER THAN A PHI (WHERE THE DISTRIBUTIONS HAVE A GREATER LIKELIHOOD OF BEING SIMILAR).<
PA 71

SPERHO ___
QU >-SPEARMAN'S RHOD MENTIONED IS THE LAST CORRELATION COEFFICIENT TO BE DISCUSSED IN THIS COURSE. IT IS KNOWN AS SPEARMAN'S RHOD OR THE RANK-DIFFERENCE CORRELATIONAL METHOD. IT IS DESIGNED TO HANDLE DATA WHICH ARE IN THE FORM OF RANKS ON BOTH VARIABLES. IT DESCRIBES THE DEGREE OF RELATIONSHIP BETWEEN TWO SETS OF RANKED DATA ON N INDIVIDUALS OR OBJECTS. A DERIVATION OF RHOD FROM THE PEARSON PRODUCT-MOMENT CORRELATION IS PRESENTED IN SLIDES 21 AND 22. BECAUSE RHOD DESCRIBES A RELATIONSHIP BETWEEN TWO SETS OF RANKED DATA, AN UNDERLYING SCALE OF MEASUREMENT FOR VARIABLES WHICH WOULD PERMIT RANKING, WHAT IS THE NAME OF A SCALE OF MEASUREMENT THAT WOULD ALLOW THIS?<
CA(I) ORD&1
TY >-RIGHT, YOU MIGHT HAVE ANSWERED INTERVAL OR RATIO AND STILL HAVE BEEN CORRECT. SINCE THESE LIKESewise PERMIT RANKING OF DATA.<
CA(I) INT&1
TY >-RIGHT, ACTUALLY AN ORDINAL SCALE OF MEASUREMENT IS ALL THAT IS NECESSARY TO RANK DATA. HOWEVER, RANKING IS APPROPRIATE FOR THE INTERVAL AND ALSO THE RATIO SCALES OF MEASUREMENT.<
CA(I) RAT&1
TY >-RIGHT, ACTUALLY AN ORDINAL SCALE OF MEASUREMENT IS ALL THAT IS NECESSARY TO RANK DATA. HOWEVER, RANKING IS APPROPRIATE FOR THE RATIO AND ALSO THE
Derivation of Spearman's Rho from the Pearson r

Organization of data when in terms of ranks for two variables might look as follows:

<table>
<thead>
<tr>
<th>PERSONS</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>K</td>
<td>K</td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>6</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>7</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>8</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>9</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>10</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>11</td>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>N</td>
<td>S</td>
<td>S</td>
</tr>
</tbody>
</table>

\[ X = \frac{N(N+1)}{2} \quad (\text{Sum of ranks } 1-N) \]
\[ Y = \frac{N(N+1)}{2} \]
\[ X = \frac{N(N+1)(2N+1)}{6} \]
\[ Y = \frac{N(N+1)(2N+1)}{6} \]

Since \( (X-Y)^2 = X^2 + Y^2 - 2XY \)
and \( \sum (X-Y)^2 = \sum X^2 + \sum Y^2 - 2 \sum XY \)

\[ \sum XY = \frac{\sum X^2 + \sum Y^2 - \sum (X-Y)^2}{2} \]

Let \( (X-Y) \), the difference of a pair of ranks, be signified as \( D \).

\[
 r_{xy} = \frac{N \sum XY - \sum X \sum Y}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}}
\]

\[
 r_{xy} = \frac{N \left[ \frac{\sum X^2 + \sum Y^2 - \sum D^2}{2} - \sum X \sum Y \right]}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}}
\]

\[
 = \frac{N \left[ \frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)(2N+1)}{6} - \sum D^2 \right]}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}} - \left[ \frac{N(N+1)}{2} \right] \left[ \frac{N(N+1)}{2} \right]
\]

\[
 \frac{N(N+1)(2N+1) - N^2(N+1)^2}{6} \sqrt{N^2(N+1)(2N+1) - N^2(N+1)^2} \]

\[
 \frac{N(N+1)(2N+1)}{6} - \frac{N^2(N+1)^2}{4} \sqrt{\frac{N(N+1)(2N+1)}{6} - \frac{N^2(N+1)^2}{4}}
\]
Derivation of Spearman's Rho from the Pearson r (cont'd)

\[
\begin{align*}
\rho &= \frac{N\left[\frac{2N(N+1)(2N+1) - \sum D^2}{12} \right] - \frac{N^3(N+1)^2}{6}}{N^2(N+1)(2N+1) - \frac{N^3(N+1)^2}{4}} \\
&= \frac{2N^2(N+1)(2N+1) - 6N \sum D^2 - 3N^3(N+1)^2}{12} \\
&\quad \div \frac{2N^2(N+1)(2N+1) - 3N^3(N+1)^2}{12} \\
&= \frac{\sum N^4 + \sum N^3 + 2N^3 - 6N \sum D^2 - 3N^3 - 6N^2 - 3N^2}{\sum N^4 + \sum N^3 + 2N^3 - 3N^2 - 6N^2 - 3N^2} \\
&= \frac{N\left[\frac{N^3 - N - 6 \sum D^2}{N(N^2 - N)}\right]}{N(N^2 - N)} \\
&= \frac{(N^3 - N) - 6 \sum D^2}{N(N^2 - 1)} \\
\rho &= 1 - \frac{6 \sum D^2}{N(N^2 - 1)}
\end{align*}
\]
INTERVAL SCALES OF MEASUREMENT.

UN > NOT QUITE, GIVE IT ANOTHER TRY.

UN > OF THE FOLLOWING TYPES OF SCALES OF MEASUREMENT, WHICH WOULD YOU CONSIDER AS ALLOWING THE RANKING OF DATA?

- NOMINAL
- ORDINAL
- INTERVAL
- RATIO

CHOOSE ONE OF THEM AS YOUR ANSWER.

QU > AS AS WITH THE POINT-BISERIAL AND PHI, RHO IS SOMETIMES USED IN PLACE OF THE PEARSON R AS A MEASURE OF RELATIONSHIP. LIKE THE OTHER TWO, RHO SHOULD NOT BE EMPLOYED WHERE THE APPLICATION OF R CAN BE JUSTIFIED.

WHEREAS THE PEARSON R DEALS WITH BOTH THE MAGNITUDE OF SCORES AND THEIR RELATIVE POSITIONS IN THE SERIES, RHO IS CONCERNED ONLY WITH THE RELATIVE POSITIONS OF THE DATA. IN SPITE OF THIS, RHO DOES PRODUCE A VALUE WHICH IS QUITE SIMILAR TO R IN MAGNITUDE IF BOTH WERE TO BE EMPLOYED ON THE SAME CONTINUOUS DATA (RANKED FOR RHO, OF COURSE) AND IF THESE CONTINUOUS DATA WERE APPROXIMATELY NORMALLY DISTRIBUTED. HOWEVER, THIS IS NOT A LICENSE TO USE RHO WHERE R IS APPLICABLE.


CAIW) .827 & .827 & .93 & .83

TY > GOOD WORK.

CAIW) .173 & .173

TY > YOU FORGOT TO SUBTRACT THIS FROM THE (1). THUS, RHO IS EQUAL TO .827.

UN > GIVE IT ANOTHER TRY.

UN > IF YOU USED THE "CALC" FUNCTION CORRECTLY AND ALSO THE PROPER DATA, RHO WOULD BE EQUAL TO .927.

BR OR |

QU > AS WAS MENTIONED ABOVE, THE VALUES OF PHI AND THE PEARSON R CALCULATED ON THE SAME CONTINUOUS DATA SHOULD BE SIMILAR (ASSUMING THE DATA ARE APPROXIMATELY NORMALLY DISTRIBUTED). THIS IS ILLUSTRATED SUFFICIENTLY HERE IN THAT...
Computations of Spearman's Rho Based on the Scatterplot of Scores Contained in Slide 8

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Rx</th>
<th>Ry</th>
<th>D^2</th>
<th>X</th>
<th>Y</th>
<th>Rx</th>
<th>Ry</th>
<th>D^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>32</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>19</td>
<td>25.5</td>
<td>28</td>
<td>6.25</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>18</td>
<td>25.5</td>
<td>30.5</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>28</td>
<td>2</td>
<td>7</td>
<td>25</td>
<td>4</td>
<td>17</td>
<td>25.5</td>
<td>32</td>
<td>42.5</td>
</tr>
<tr>
<td>7</td>
<td>31</td>
<td>5.5</td>
<td>2</td>
<td>12.25</td>
<td>4</td>
<td>15</td>
<td>25.5</td>
<td>38</td>
<td>156.25</td>
</tr>
<tr>
<td>7</td>
<td>29</td>
<td>5.5</td>
<td>5</td>
<td>25</td>
<td>4</td>
<td>13</td>
<td>25.5</td>
<td>42.5</td>
<td>289</td>
</tr>
<tr>
<td>7</td>
<td>26</td>
<td>5.5</td>
<td>10.5</td>
<td>25</td>
<td>3</td>
<td>24</td>
<td>34.5</td>
<td>16.5</td>
<td>324</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>5.5</td>
<td>16.5</td>
<td>121</td>
<td>3</td>
<td>22</td>
<td>34.5</td>
<td>20.5</td>
<td>196</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
<td>10</td>
<td>5</td>
<td>25</td>
<td>3</td>
<td>20</td>
<td>34.5</td>
<td>25.5</td>
<td>61</td>
</tr>
<tr>
<td>6</td>
<td>26</td>
<td>10</td>
<td>10.5</td>
<td>25</td>
<td>3</td>
<td>19</td>
<td>34.5</td>
<td>28</td>
<td>42.25</td>
</tr>
<tr>
<td>6</td>
<td>25</td>
<td>10</td>
<td>13</td>
<td>9</td>
<td>3</td>
<td>16</td>
<td>34.5</td>
<td>34.5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td>10</td>
<td>16.5</td>
<td>42.25</td>
<td>3</td>
<td>15</td>
<td>34.5</td>
<td>38</td>
<td>12.25</td>
</tr>
<tr>
<td>6</td>
<td>21</td>
<td>10</td>
<td>23</td>
<td>169</td>
<td>3</td>
<td>13</td>
<td>34.5</td>
<td>42.5</td>
<td>64</td>
</tr>
<tr>
<td>5</td>
<td>29</td>
<td>16.5</td>
<td>5</td>
<td>132.25</td>
<td>3</td>
<td>11</td>
<td>34.5</td>
<td>46</td>
<td>132.25</td>
</tr>
<tr>
<td>5</td>
<td>27</td>
<td>16.5</td>
<td>8.5</td>
<td>64</td>
<td>2</td>
<td>19</td>
<td>41</td>
<td>28</td>
<td>169</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>16.5</td>
<td>13</td>
<td>12.25</td>
<td>2</td>
<td>16</td>
<td>41</td>
<td>34.5</td>
<td>42.25</td>
</tr>
<tr>
<td>5</td>
<td>24</td>
<td>16.5</td>
<td>16.5</td>
<td>0</td>
<td>2</td>
<td>15</td>
<td>41</td>
<td>38</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>16.5</td>
<td>23</td>
<td>42.25</td>
<td>2</td>
<td>14</td>
<td>41</td>
<td>40.5</td>
<td>.25</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>16.5</td>
<td>25.5</td>
<td>81</td>
<td>2</td>
<td>11</td>
<td>41</td>
<td>46</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>16.5</td>
<td>30.5</td>
<td>196</td>
<td>1</td>
<td>16</td>
<td>45.5</td>
<td>34.5</td>
<td>121</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>16.5</td>
<td>34.5</td>
<td>324</td>
<td>1</td>
<td>14</td>
<td>45.5</td>
<td>40.5</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>27</td>
<td>25.5</td>
<td>8.5</td>
<td>289</td>
<td>1</td>
<td>11</td>
<td>45.5</td>
<td>46</td>
<td>.25</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>25.5</td>
<td>13</td>
<td>156.25</td>
<td>1</td>
<td>9</td>
<td>45.5</td>
<td>49</td>
<td>12.25</td>
</tr>
<tr>
<td>4</td>
<td>23</td>
<td>25.5</td>
<td>19</td>
<td>42.25</td>
<td>0</td>
<td>12</td>
<td>49</td>
<td>44</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>25.5</td>
<td>20.5</td>
<td>25</td>
<td>0</td>
<td>10</td>
<td>49</td>
<td>48</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>25.5</td>
<td>23</td>
<td>6.25</td>
<td>0</td>
<td>8</td>
<td>49</td>
<td>50</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \sum D^2 = 3603 \]

\[ N = 50 \]
Who is equal to .827 as found to be .827?

As with the Pearson R and Phi, the Spearman Rank-Difference Coefficient does not suffer from a reduction in maximum values, most probably because of the tendency toward similarity of the two rank distributions. Therefore, functionally, the range of values that rho may assume runs from -1.00 to +1.00.

The use of any one of the presented correlation coefficients for analysis is of course dependent upon a particular data situation. Assuming the ideal conditions of data collection and the applicability of one of the figures, coefficients presented in this program, determine which of these correlations is best applicable in the hypothetical situations which follow.

Note: For each case, type the number corresponding to the correlation that you would apply:

1. Pearson Product-Moment R
2. Point-Biserial R
3. Phi Coefficient
4. Spearman Rho
5. None of the above

Applications of the coefficients

A sociologist asks you to analyze some data he has collected on persons' political affiliation (assumed Democrat or Republican). He also knows the racial background (assumed Negro or Caucasian) of these persons, he would like to determine whether there is a statistical relationship between political affiliation and race for the persons in his sample. What coefficient would you recommend?

Call (l) 361. You correctly, since he has assumed dichotomous categories of political affiliation and race, the phi coefficient would be properly applied here.
Although the Pearson $R$ could be applied here, it most likely would not be for expedience purposes. Phi was derived for $R$ for data situations as this and is the coefficient most properly applied here since the sociologist has assumed dichotomous categories of political affiliation and race.

No. The application of the point-biserial is not justified under the data situation given, since the sociologist has assumed dichotomous categories of political affiliation and race. Phi would be most properly applied.

No. The application of the Spearman rank coefficient is not justified under the data situation given, since the sociologist has assumed dichotomous categories of political affiliation and race. Phi would be most properly applied.

Since the sociologist has assumed dichotomous categories of political affiliation and race, Phi would be most properly applied.

You, as a teacher-researcher, are interested in determining how well a number of subject-matter units were presented to a class (in terms of clarity, meaningfulness, etc.). Thus, you select two evaluators to rate the units, relative to each other, in terms of effectiveness of presentation. Now you would like to see of there is a relationship between the ratings of both evaluators, which correlation would you use?

Very good. The relative ratings of the units by the evaluators should be taken as rankings of the units in terms of their effectiveness under two conditions (the two evaluators). Thus, the Spearman coefficient is the correlation of choice.
THE PEARSON $r$ IS NOT APPROPRIATELY APPLIED HERE EVEN THOUGH ITS APPLICATION IS POSSIBLE. SPEARMAN'S RHO IS THE PROPER AND MOST EFFICIENT COEFFICIENT TO USE SINCE THE RELATIVE RATINGS OF THE UNITS BY THE EVALUATORS SHOULD BE TAKEN AS RANKINGS OF THE UNITS IN TERMS OF THEIR EFFECTIVENESS UNDER TWO CONDITIONS (THE TWO EVALUATORS).


PHI CANNOT FIND PROPER APPLICATION HERE. THE DATA SITUATION STRONGLY IMPLIES THE USE OF SPEARMAN'S RHO SINCE THE RELATIVE RATINGS OF THE UNITS BY THE EVALUATORS SHOULD BE TAKEN AS RANKINGS OF THE UNITS IN TERMS OF THEIR EFFECTIVENESS UNDER TWO CONDITIONS (THE TWO EVALUATORS). THUS, THERE IS NO INDICATION OF TWO DICHOTOMOUS SCALES OF MEASUREMENT WITH RESPECT TO THE DATA GIVEN TO WARRANT THE USE OF PHI.

SPEARMAN'S RHO IS NOT ONLY APPLICABLE BUT THE PROPER COEFFICIENT TO USE IN THE DATA SITUATION PRESENTED. THE RELATIVE RATINGS OF THE UNITS BY THE EVALUATORS SHOULD BE TAKEN AS RANKINGS OF THE UNITS IN TERMS OF THEIR EFFECTIVENESS UNDER TWO CONDITIONS (THE TWO EVALUATORS).

UN ANSWER EITHER 1, 2, 3, 4, OR 5.

AN ADMISSIONS OFFICE AT A SMALL LIBERAL ARTS COLLEGE WOULD LIKE TO SEE IF
THERE IS A STATISTICAL RELATIONSHIP BETWEEN HIGH SCHOOL GRADE POINT AVERAGES (GPA) OF INCOMING FRESHMEN WITH THEIR GPA'S AT THE END OF THEIR FIRST YEAR IN COLLEGE. IN THE FUTURE HE WOULD ALSO LIKE TO DETERMINE THE STATISTICAL RELATIONSHIP BETWEEN THIS GROUP'S H. S. GPA'S AND THEIR FOUR YEAR COLLEGE GPA'S.

WHAT CORRELATION COEFFICIENT SHOULD HE USE IN EITHER INSTANCE?

RIGHT. THE ASSUMPTION IN THIS SITUATION IS THAT THE GPA'S ARE MEASURED ON A CONTINUOUS SCALE OF MEASUREMENT. SINCE THE CORRELATION COEFFICIENTS ARE CORRELATION COEFFICIENTS, THE DEGREE OF CORRELATION BETWEEN THE TWO VARIABLES OF INTEREST.

IT IS ASSUMED THAT BOTH GPA'S HERE ARE MEASURED ON CONTINUOUS SCALES OF MEASUREMENT. IN THIS CASE THE PEARSON R IS THE APPROPRIATE CORRELATION TO USE. THE POINT-BISEKIAL WOULD ONLY BE OF VALUE IF ONE OF THE VARIABLES WAS MEASURED ON A DICHTOMOUS SCALE AND THE OTHER VARIABLE ON A CONTINUOUS SCALE.

YOU MIGHT HAVE MISSED THE ASSUMPTION ABOUT THE DATA UNDER CONSIDERATION. IT IS ASSUMED THAT BOTH GPA'S HERE ARE MEASURED ON CONTINUOUS SCALES OF MEASUREMENT. THE PHII COEFFICIENT WOULD ONLY BE OF VALUE IF BOTH OF THE VARIABLES WERE MEASURED DICHTOMOUSLY.

 THERE IS NO INDICATION THAT THE DATA ARE RANKED. IT IS ASSUMED, IN FACT, THAT THE GPA'S ARE MEASURED ON CONTINUOUS SCALES OF MEASUREMENT. THE PEARSON R IS THE APPROPRIATE CORRELATION FOR APPLICATION HERE AND NOT SPEARMAN'S PHII.

NO. AS DESCRIBED, THE SITUATION CALLS FOR THE APPLICATION OF THE PEARSON R.
THE ASSUMPTION THAT SHOULD HAVE BEEN MADE IN REACHING THIS DECISION IS THAT GPA'S ARE MEASURED ON CONTINUOUS SCALES OF MEASUREMENT - A REQUIREMENT OF THE VARIABLES FOR USE OF THE PEARSON R.

PA 101
BR ASIDE 1
UN ANSWER EITHER 1, 2, 3, 4, OR 5.

ASIDE 1
QU A LOGICAL OUTGROWTH OF THE LITTLE STUDY CONDUCTED BY THE ADMISSIONS OFFICER IN THE PREVIOUS SITUATION MIGHT BE AN ATTEMPT TO ESTABLISH AN INDICATION OF PROBABLE ACADEMIC SUCCESS IN COLLEGE BASED ON HIGH SCHOOL GPA. IN THIS WAY, THE COLLEGE COULD COUNSEL FUTURE POTENTIAL STUDENTS AS TO THEIR CHANCES OF ACADEMIC SUCCESS AT THAT INSTITUTION. ASSUMING THE USE OF THE PEARSON R TO DETERMINE THE STATISTICAL RELATIONSHIP BETWEEN H.S. GPA AND COLLEGE GPA, WHAT OTHER TECHNIQUE MIGHT THE ADMISSIONS OFFICER USE IN OBTAINING A GENERAL INDICATION OF PROBABLE ACADEMIC SUCCESS IN COLLEGE FOR ANY POTENTIAL STUDENT? TYPE A., B., OR C.

A. THE STATISTICAL RELATIONSHIP BETWEEN H.S. GPA AND COLLEGE GPA (USING "R") IS SUFFICIENT TO GIVE HIM A RELATIVELY GOOD INDICATION.

B. THE USE OF A REGRESSION EQUATION (IN CONJUNCTION WITH THE PEARSON R) WILL SERVE TO INDICATE STATISTICALLY THE STUDENTS' PROBABLE SUCCESS IN COLLEGE.

C. IN ADDITION TO THE STATISTICAL RELATIONSHIP BETWEEN THE GPA'S, THE ADMISSIONS OFFICER MIGHT CONSTRUCT STUDENT PROFILES BASED ON JUDGED IMPORTANT VARIABLES OF PAST STUDENTS AT THE COLLEGE AND THE DEGREE OF ACADEMIC SUCCESS THEY ACHIEVED. HE COULD THEN COMPARE INCOMING STUDENTS' PROFILES ON THESE VARIABLES WITH THE "STANDARD" PROFILES.

CALL 81
TY RIGHT. IN THIS INSTANCE A REGRESSION EQUATION BASED ON THE PEARSON R BETWEEN THE TWO SETS OF GPA'S WOULD SERVE THE PURPOSE OF A "PREDICTION" EQUATION BY WHICH AN INCOMING OR POTENTIAL STUDENT WOULD HAVE SOME INDICATION AS TO HIS PROBABLE ACADEMIC SUCCESS AT THIS COLLEGE. THE EQUATION THE OFFICER WOULD MOST LIKELY USE WOULD BE 15.1 OF SLIDE 15 WHERE VARIABLE "X" WOULD BE H.S. GPA AND VARIABLE "Y" WOULD BE COLLEGE GPA.

PA 51
BR SITUATION
WALL 1 A1
TY ALTHOUGH THE STATISTICAL RELATIONSHIP BETWEEN THE TWO GPA'S WOULD GIVE AN...
INDICATION OF ACADEMIC SUCCESS. IT WOULD BE ONLY WHERE SUBJECTIVELY, BY USING ...

A REGRESSION EQUATION BASED ON THE PEARSON R BETWEEN THE TWO SETS OF GPA'S...

ONE COULD OBTAIN A "PREDICTION" EQUATION BY WHICH AN INCOMING OR POTENTIAL...

STUDENT WOULD HAVE SOME INDICATION AS TO HIS PROBABLE ACADEMIC SUCCESS AT THIS...

COLLEGE. THE EQUATION THE OFFICER WOULD MOST LIKELY USE WOULD BE 15.1 ...

OF SLIDE 15 WHERE VARIABLE X WOULD BE H.S. GPA AND VARIABLE Y WOULD BE COLLEGE GPA.

PA 101

BR SITUATION

WAIT 16

TY THE PROFILE MAY BE WORTHWHILE. HOWEVER, IN TERMS OF EFFICIENCY OF LABOR...

AND OF PURPOSE AND IN KEEPING WITH THE NATURE OF THIS COURSE, A REGRESSION EQUATION...

BASED ON THE PEARSON R BETWEEN THE TWO SETS OF GPA'S WOULD SERVE THE PURPOSE OF...

A "PREDICTION" EQUATION BY WHICH AN INCOMING OR POTENTIAL STUDENT WOULD HAVE SOME...

INDICATION AS TO HIS PROBABLE ACADEMIC SUCCESS AT THIS COLLEGE. FURTHERMORE...

IT MAY BE FOUND THAT THE PROFILE APPROACH WOULD FINALLY RESOLVE ITSELF IN THE...

USE OF A NUMBER OF REGRESSION EQUATION S OR IN A TECHNIQUE THAT WASN'T DISCUSSED...

IN THE PROGRAM. THAT BEING MULTIPLE REGRESSION...

THE EQUATION THE OFFICER WOULD MOST LIKELY USE WOULD BE 15.1 OF SLIDE 15 WHERE...

VARIABLE X WOULD BE H.S. GPA AND VARIABLE Y WOULD BE COLLEGE GPA.

PA 101

BR SITUATION

WAIT 16

UN ANSWER EITHER A, B, OR C.

SITUATION

OUT A TEST CONSTRUCTOR WISHES TO DETERMINE THE APPROPRIATENESS OF ITEMS IN HIS TEST...

AT MEASURING MECHANICAL SKILLS. THUS, HE DETERMINES FIVE CRITERIA WHICH HE...

FEELS DEFINE THE CONCEPT OF MECHANICAL SKILL AND MEASURES A NUMBER OF SUBJECTS...

ON THESE CRITERIA, THESE SAME SUBJECTS ARE GIVEN HIS TEST. WHAT MEASURE...

OF RELATIONSHIP MIGHT HE USE TO CORRELATE EACH ITEM WITH HIS CRITERIA?

CALL 561

TY RIGHT. THE MEASURE OF ASSOCIATION WHICH WOULD BE REQUIRED IN THIS CASE TO DETERMINE...

THE APPROPRIATENESS OF HIS ITEMS HAS NOT BEEN PRESENTED IN THE PROGRAM. IF YOU...

ARE INTERESTED IN KNOWING WHAT THIS TECHNIQUE IS, TYPE "YES". IF YOU ALREADY...

KNOW OR DON'T WISH TO KNOW AT THIS TIME, TYPE "NO".

BR OPTION1

WAIT 161

TY NO. THE PEARSON R IS NOT APPLICABLE HERE BECAUSE OF THE NATURE OF THE DATA TO BE...
CORRELATED.  FURTHERMORE, NONE OF THE MEASURES OF CORRELATION THAT WERE PRESENTED IN THIS COURSE IS APPROPRIATE IN THIS SITUATION. IF YOU ARE INTERESTED IN KNOWING AN APPROPRIATE TECHNIQUE, TYPE "YES". IF YOU ALREADY KNOW OR DON'T WISH TO KNOW AT THIS TIME, TYPE "NO":

BR OPTION

WRAP (I) 26]

TY THE POINT-BISERIAL IS NOT FULLY APPLICABLE HERE BECAUSE OF THE NATURE OF THE DATA TO BE CORRELATED. FURTHERMORE, NONE OF THE MEASURES OF CORRELATION THAT WERE PRESENTED IN THIS COURSE IS APPROPRIATE IN THIS SITUATION. IF YOU ARE INTERESTED IN KNOWING AN APPROPRIATE TECHNIQUE, TYPE "YES". IF YOU ALREADY KNOW OR DON'T WISH TO KNOW AT THIS TIME, TYPE "NO":

BR OPTION

WRAP (I) 36]

TY THE PHI COEFFICIENT IS NOT APPLICABLE HERE BECAUSE OF THE NATURE OF THE DATA TO BE CORRELATED. FURTHERMORE, NONE OF THE MEASURES OF CORRELATION THAT WERE PRESENTED IN THIS COURSE IS APPROPRIATE IN THIS SITUATION. IF YOU ARE INTERESTED IN KNOWING AN APPROPRIATE TECHNIQUE, TYPE "YES". IF YOU ALREADY KNOW OR DON'T WISH TO KNOW AT THIS TIME, TYPE "NO":

BR OPTION

WRAP (I) 46]

TY SPEARMAN'S RHO IS NOT APPLICABLE HERE BECAUSE OF THE NATURE OF THE DATA TO BE CORRELATED. FURTHERMORE, NONE OF THE MEASURES OF CORRELATION THAT WERE PRESENTED IN THIS COURSE IS APPROPRIATE IN THIS SITUATION. IF YOU ARE INTERESTED IN KNOWING AN APPROPRIATE TECHNIQUE, TYPE "YES". IF YOU ALREADY KNOW OR DON'T WISH TO KNOW AT THIS TIME, TYPE "NO":

BR OPTION

UN TYPE 1, 2, 3, 4, OR 5.

OPTION

1<br>

CALL (I) YES 51

TY THE APPROPRIATE MEASURE OF ASSOCIATION FOR USE HERE WOULD BE THE CHI-SQUARE. MOST STATISTICS TEXTS DISCUSS THIS MEASURE.

BR SITUAS1

CALL (I) NO 61

BR SITUAS1

UN TYPE YES OR NO.
In the previous case, suppose the test constructor used total scores of another measure of mechanical skill as the criterion measure of mechanical skill.

What measure of correlation might he use now to determine the appropriateness of each of the items of his test at measuring mechanical skill?

(Try good.) Since each item of his test can be measured as either being correct or incorrect and since it is assumed that the criterion test scores are measured on a continuous scale, the point-biserial $r$ is the correct coefficient to use. In the first case, the criterion scores must be considered to be five distinct and discrete categories each measured as being attained or not attained. Thus, the point-biserial (or any of the other coefficients discussed in this course) was not applicable there even though the test items can be measured dichotomously.

(Try no.) The Pearson $r$ is not the appropriate coefficient for application here. Since each item of his test can be measured as either being correct or incorrect and since it is assumed that the criterion test scores are measured on a continuous scale, the point-biserial $r$ is the correct coefficient to use. Moreover, in the first case, the criterion scores must be considered to be five distinct and discrete categories, each measured as being attained or not attained. Thus, the point-biserial (or any of the other coefficients discussed in this course) was not applicable there even though the test items can be measured dichotomously.

(Try no.) The phi coefficient is not the appropriate coefficient for application here. Since each item of his test can be measured as either being correct or incorrect and since it is assumed that the criterion test scores are measured on a continuous scale, the point-biserial $r$ is the correct coefficient to use. Remember, phi is the coefficient of choice when there are dichotomous scales of measurement on each of the two variables.

As a matter of interest, the point-biserial (or any of the other coefficients discussed in this course) was not applicable in the previous case because the criterion scores there must be considered to be five distinct and discrete categories, each measured as being attained or not attained.
TY NO. THERE IS NO REASON TO BELIEVE THAT THE VARIABLES OF CORRELATION HERE HAVE DATA IN RANKED FORM. SINCE EACH ITEM OF THE TEST CAN BE MEASURED AS EITHER BEING CORRECT OR INCORRECT AND SINCE IT IS ASSUMED THAT THE CRITERION TEST SCORES ARE MEASURED ON A CONTINUOUS SCALE, THE POINT-BIVARIATE $r$ IS THE CORRECT COEFFICIENT TO USE.

AS A MATTER OF INTEREST, THE POINT-BIVARIATE ($r$ OR ANY OF THE OTHER COEFFICIENTS DISCUSSED IN THIS COURSE) WAS NOT APPLICABLE IN THE PREVIOUS CASE BECAUSE THE CRITERION SCORES THERE MUST BE CONSIDERED TO BE FIVE DISTINCT AND DISCRETE CATEGORIES, EACH MEASURED AS BEING ATTAINED OR NOT ATTAINED.

TY >> NO. SINCE EACH ITEM OF THE TEST CAN BE MEASURED AS EITHER BEING CORRECT OR INCORRECT AND SINCE IT IS ASSUMED THAT THE CRITERION TEST SCORES ARE MEASURED ON A CONTINUOUS SCALE, THE POINT-BIVARIATE $r$ CAN BE USED HERE.

AS A MATTER OF INTEREST, THE POINT-BIVARIATE ($r$ OR ANY OF THE OTHER COEFFICIENTS DISCUSSED IN THE COURSE) WAS NOT APPLICABLE IN THE PREVIOUS CASE BECAUSE THE CRITERION SCORES THERE MUST BE CONSIDERED TO BE FIVE DISTINCT AND DISCRETE CATEGORIES, EACH MEASURED AS BEING ATTAINED OR NOT ATTAINED.

CONCLUSION

THIS CONCLUDES THIS VERY BRIEF INTRODUCTION TO THE STUDY OF CORRELATION. BY NO MEANS HAS THE MATERIAL PRESENTED EXHAUSTED THE DOMAIN OF CORRELATIONS. WHAT WAS PRESENTED, HOWEVER, GIVES YOU A FUNDAMENTAL KNOWLEDGE OF THE THEORY OF CORRELATIONS AND A FEW METHODS FOR APPLICATION IN VARIOUS DATA SITUATIONS. ATTENTION WAS FOCUSED ON THE PRODUCT-MOMENT FAMILY OF CORRELATIONS OF WHICH THE PEARSON $r$ IS THE MOST POPULAR. ALTHOUGH LIMITED IN SOME RESPECTS BECAUSE OF ITS ASSUMPTIONS, IT AND ITS THREE DERIVATIONS PROVE TO BE QUITE USEFUL WHERE THE DATA ANALYSIS DEMANDS SOME MEASURE OF VARIABLE RELATIONSHIP, THOUGH NOT APPLICABLE TO ALL SUCH CASES, THEY DO FIND EMPLOYMENT IN A GOOD NUMBER OF THE TYPICAL EMPIRICAL ANALYSES FOUND IN THE BEHAVIORAL SCIENCES.

THANK YOU FOR TAKING THIS COURSE. PLEASE TYPE "SIGN OFF" BEFORE YOU LEAVE.

END
APPENDIX B

Opinionnaire
APPENDIX B

OPINIONNAIRE

1) How would you rate your exposure (e.g., number of courses, work experience with statistics, etc.) to statistics?

Very Limited  Limited  Average  Broad  Very Broad

2) How would you rate the choice of concepts presented in the program? That is, considering the objective of presenting the four product-moment family correlations, how would you rate their development in terms of the specific concepts used in the program?

Very Poor  Poor  Average  Good  Very Good

3) How would you rate the logical sequencing of these concepts in this program?

Very Poor  Poor  Average  Good  Very Good

4) In general, how would you rate the concepts presented in terms of difficulty?

Very Easy  Easy  Average  Difficult  Very Difficult

5) As they were presented in the program, how would you rate your difficulty in understanding these concepts?

Very Much  Much  Average  Little  Very Little
6) How would you rate the clarity of presentation of the various concepts in the development of the correlations?

Very Poor Poor Average Good Very Good

7) Which of these concepts (or sections of the program) specifically lack a clear presentation? Furthermore, how unclear are they?

8) As a result of taking this CAI course, how would you rate your understanding of those concepts with which the program dealt?

Very Poor Poor Average Good Very Good

9) In general, how would you rate the instructional value of this CAI program?

Very Poor Poor Average Good Very Good

10) Do you think this subject matter might be best presented in traditional lecture fashion, via a computer system, or some combination of both?

Traditional lecture fashion
Via a computer system
Some combination of the above two methods
Give the rationale for your choice.

11) Do you feel this program should be revised?

Yes No
If your answer is "yes", indicate where the program needs revision and any suggestions as to how these sections may be revised.

If your answer is "no", state why you think the program should not be revised.

12) As defined, the general instructional objectives of the program were as follows:

1) The student will be capable of identifying some possible but inadequate definitions of a correlation coefficient

2) The student will be capable of manifesting a comprehension of the theoretical development of an appropriate measure of correlation and that this coefficient represents a general statement of correlation

3) The student will be capable of demonstrating knowledge of some fundamental concepts of linear regression as they relate to the theory of correlation

4) The student will be capable of recognizing that, under certain conditions, the developed generalized correlation coefficient reduces to the Pearson Product-Moment coefficient

5) The student will be capable of recognizing that, under other specific conditions, the Pearson Product-Moment correlation reduces to the Point-Biserial r, the Phi Coefficient, or Spearman's Rho, depending on the particular set of conditions stated
6) The student will exhibit a sufficient knowledge of the appropriate empirical applications of the presented correlation coefficients.

To what degree do you believe each of these objectives were met by the program?

<table>
<thead>
<tr>
<th>Objective 1:</th>
<th>Very Limited</th>
<th>Limited</th>
<th>Average</th>
<th>Great</th>
<th>Very Great</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective 2:</td>
<td>Very Limited</td>
<td>Limited</td>
<td>Average</td>
<td>Great</td>
<td>Very Great</td>
</tr>
<tr>
<td>Objective 3:</td>
<td>Very Limited</td>
<td>Limited</td>
<td>Average</td>
<td>Great</td>
<td>Very Great</td>
</tr>
<tr>
<td>Objective 4:</td>
<td>Very Limited</td>
<td>Limited</td>
<td>Average</td>
<td>Great</td>
<td>Very Great</td>
</tr>
<tr>
<td>Objective 5:</td>
<td>Very Limited</td>
<td>Limited</td>
<td>Average</td>
<td>Great</td>
<td>Very Great</td>
</tr>
<tr>
<td>Objective 6:</td>
<td>Very Limited</td>
<td>Limited</td>
<td>Average</td>
<td>Great</td>
<td>Very Great</td>
</tr>
</tbody>
</table>
SELECTED BIBLIOGRAPHY

Books


**Periodicals**


Reports and Papers


Gilman, D. A. A comparison of several feedback modes in a computer-assisted adjunct auto-instruction program. Terre Haute, Indiana: School of Education, Indiana State University, 1968.


