AGREEMENT ANALYSIS: A DECISION CHART AND ITS EVALUATION

By

JOHANNES LODEWICUS BOTHA, M.B., Ch.B.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

August 1979

AGREEMENT ANALYSIS: A DECISION CHART AND ITS EVALUATION

MASTER OF SCIENCE (1979) McMaster University
(Clinical Epidemiology) Hamilton, Ontario

TITLE: Agreement Analysis: A Decision Chart and Its Evaluation

AUTHOR: Johannes Lodewicus Botha, M.B., Ch.B. (Cape Town)

SUPERVISOR: Professor C.H. Goldsmith

NUMBER OF PAGES: xii, 181

ABSTRACT

Existing statistical measures of agreement are reviewed and organized into a series of flow charts to facilitate their selection for analysis. A subset of "useful" measures is selected on the basis of their measuring meaningful agreement, having interpretable referent values, being subject to hypothesis testing, being hand calculable and continuous (categorical) measures having categorical (continuous) analogues. "Useful" measures for continuous data sets are all intraclass correlation coefficients, applicable to different ANOVA models depending on the assumptions involved. "Useful" measures for categorical data include various chance corrected (kappa) types, sensitivity-specificity and predictive value measures. The "useful" measures are displayed in a single condensed flow chart. A strategy to evaluate the use of this flow chart is also developed.

ACKNOWLEDGEMENTS

I wish to thank Charlie Goldsmith for his inspired and inspiring instruction, and his stimulating guidance, Gary Anderson, Peter Tugwell and David Sackett for their constructive criticism of the manuscript, Linda Teiml for seemingly effortless yet errorless typing, Flo O'Dowd for taking the worry out of being a student and the South African Medical Research Council for financial support.

CONTENTS

ILLUSTRATIONS

List of Tables

ILLUSTRATIONS

List of Diagrams

To

Suzanne and Louw and John and Otto

from all of whom I have learnt to understand
and appreciate human individuality.

CHAPTER 1

Introductory Remarks

As a teaching aid a videotape of the correct procedure for recording blood pressure has been developed. When it was shown to an audience of four clinicians, four statisticians and four "others" and when they were asked to record the systolic blood pressures shown in the presentation, Table 1.1 resulted. How well do they agree in recording blood pressure?

At a gastro-intestinal disease clinic some patients undergo endoscopy for the diagnosis of oesophageal varices. On thirty-nine patients endoscopy was performed independently by two clinicians, with the resulting data as in Table 1.2. How well do they agree in their diagnoses?

Problems such as these are common in clinical practice, and to measure agreement is evidently of importance to clinicians and patients alike. Several authors have reviewed studies on agreement in several settings: clinical examination, diagnostic decisions, research surveys, health care studies, psychiatric and social science studies (Garland 1959, 1960; Fletcher 1964; Koran 1975). Common to all were two findings: firstly, that though aware of its existence, clinicians and non-clinician observers were uniformly incredulous at the extent of disagreement among and within themselves. Secondly, neither the definition nor the method of measuring agreement was consistent.

Table 1.1

Data Resulting From Blood Pressure Readings From a Videotape (mmHg)

	Patients				
Observer Type	1	2	3	4	5
Clinicians					
1	134	122	166	170	134
2	134	122	166	168	134
3	134	122	166	168	134
4	134	121	162	166	130
Statisticians					
1	134	122	178	178	134
2	134	120	166	168	134
3	132	120	168	170	134
4	134	122	168	170	134
Others					
1	132	120	168	168	134
2	134	121	168	169	134
3	134	122	166	169	134
4	136	120	168	168	134

Table 1.2

Data Resulting When 39 Patients Are Classified by Two Endoscopists On the Presence (+) or Absence (-) of Oesophageal Varices (adapted from Conn, Smith and Brodoff 1965)

	Oesophagoscopic Diagnosis	
Patient #	Endoscopist A	Endoscopist B
1	-	-
2	-	-
3	+	+
4	+	+
5	-	-
6	+	+
7	+	-
8	-	-
9	+	-
10	-	+
11	+	-
12	-	-
13	+	+
14	-	-
15	-	-
16	+	+
17	+	+
18	+	-
19	+	-
20	+	+
21	-	-
22	-	-
23	+	-
24	-	+
25	-	+
26	+	-
27	+	-
28	-	-
29	-	-
30	-	+
31	-	-
32	+	+
33	+	+
34	-	-
35	+	+
36	-	-
37	-	+
38	-	-
39	+	+

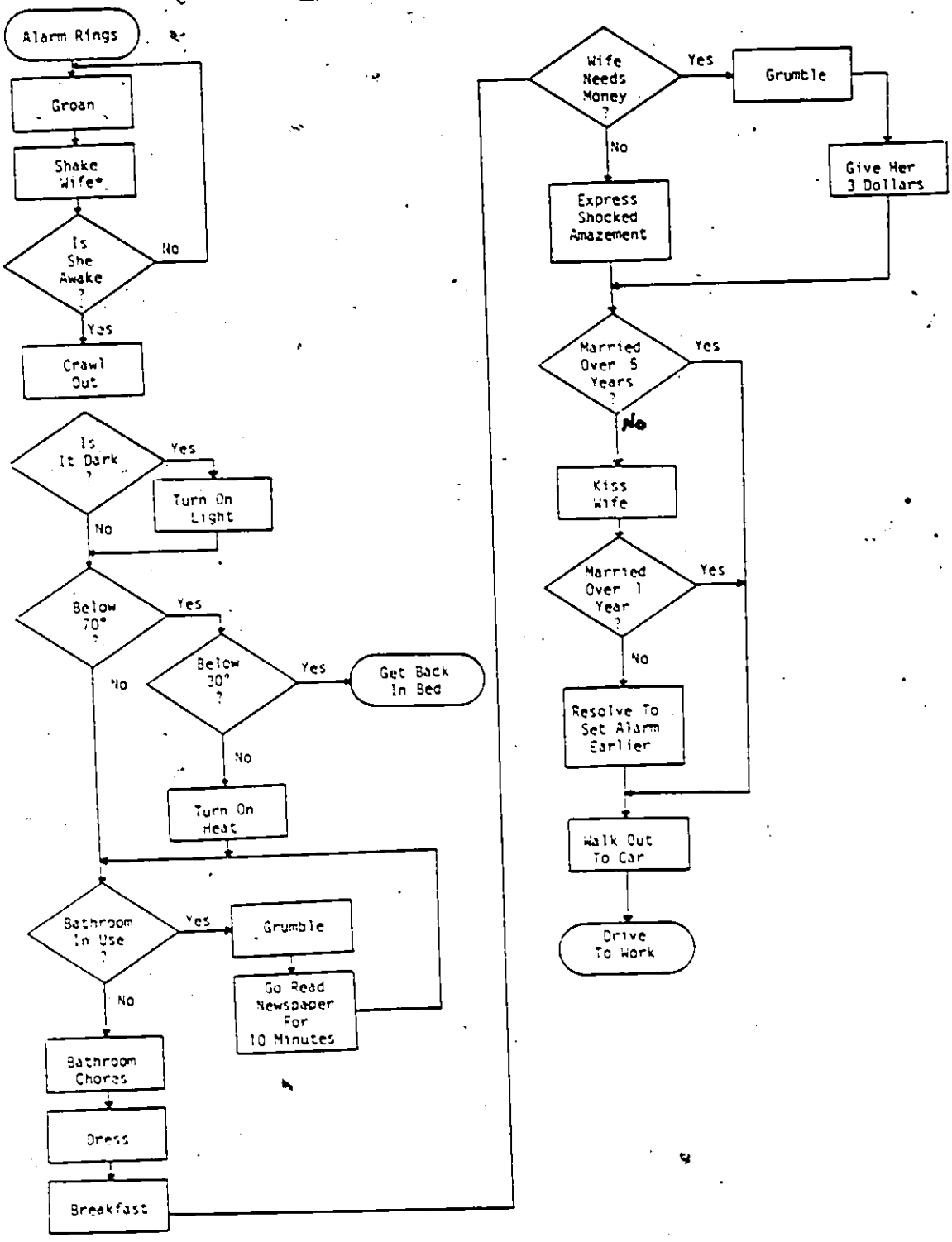
In this thesis the second of the two issues just identified will be addressed. The aim of this thesis is to review existing statistical measures of agreement, to organise them into a format facilitating the choice of an appropriate measure and to design a strategy to evaluate the use of the programmed format.

In reviewing measures of agreement, an attempt will be made to include all measures published in English until the end of 1978. Many different measures of agreement exist, but they go by names not all immediately identifying them as agreement measures to the uninitiated. Landis and Koch (1975) reviewed many of these measures within a unifying framework, but the choice of statistic is still not easy for the statistically unsophisticated person.

To facilitate the choice of an appropriate agreement measure by clinical researchers, the existing agreement measures will be organised into a flow chart to aid decision making. An example of a flow chart is shown in Diagram 1.1, and it is evident that it outlines the reasoning necessary to reach a particular conclusion in such a way that the same reasoning will repeatedly result in the same conclusion. Flow charts have been used as an approach to introductory statistics (Harshbarger 1971) and as an aid in clinical decision-making by para-medical personnel (Sackett 1978). The development of a flow chart for agreement analysis is described in Chapters 2, 3, 4 and 5 of this thesis. People awed by statistical notation may skip Chapters 3 and 4, read Chapters 2 and 5 only and still understand and be able to use the condensed flow chart of "useful" measures described in Chapter 5.

DIAGRAM 1.1

How to Get to work in the Morning



* In "Wife", readers may substitute "husband" or "cohabitant".

(from Hershberger 1971)

A strategy to evaluate the use of the flow chart as an alternative to conventional methods of instruction will be developed in Chapter 6. The evaluation is proposed to be done in this way for two reasons: it is important to know whether the presentation of agreement measures in the form of a flow chart is of more benefit to clinical researchers than existing presentations. Secondly, the idea for the flow chart resulted from my experience in a biostatistical course*. In the course, a comprehensive list of agreement measures was discussed, but the problems assigned subsequently involved the use of only one measure of agreement and its use was explicity indicated. It was felt that the course had not provided an overall approach that could be used when confronted with a raw data set and no explicit instructions.

The objective of this study is therefore to produce a flow chart of appropriate agreement measures and to propose a way of determining whether it helps clinical researchers to choose a statistic, as well as whether it is useful as an alternative to conventional teaching methods.

---

* Medical Sciences 731, McMaster University

CHAPTER 2

A Guide to the Flow Chart of Agreement Measures

There is as much difference between us and ourselves as between us and others.

Michel de Montaigne (1533-1592)

The flow chart of agreement measures will be presented as several separate sections in subsequent chapters. An overview of its construction is shown in Diagram 2.1 for orientation, with the sections indexed to the relevant chapters. In all the flow charts certain symbols (Diagram 2.2) will be consistently used to lead the user through the flow chart. The rest of this chapter serves as an explanation of the terms used on the flow chart as decision points and for descriptive purposes.

2.1 Preamble

Agreement has many guises: reliability, consistency, precision, reproducibility, correlation have all been used as alternative terms for agreement in various situations by investigators of different backgrounds. Other terms used similarly are accuracy or validity which are acceptable alternatives for agreement when comparison is with an external standard, and also association, which is acceptable only in one specific situation, the fourfold table. Association can be defined as the degree of dependence or independence which exists between two or

DIAGRAM 2.1. Overview of Flow Chart of Agreement Measures

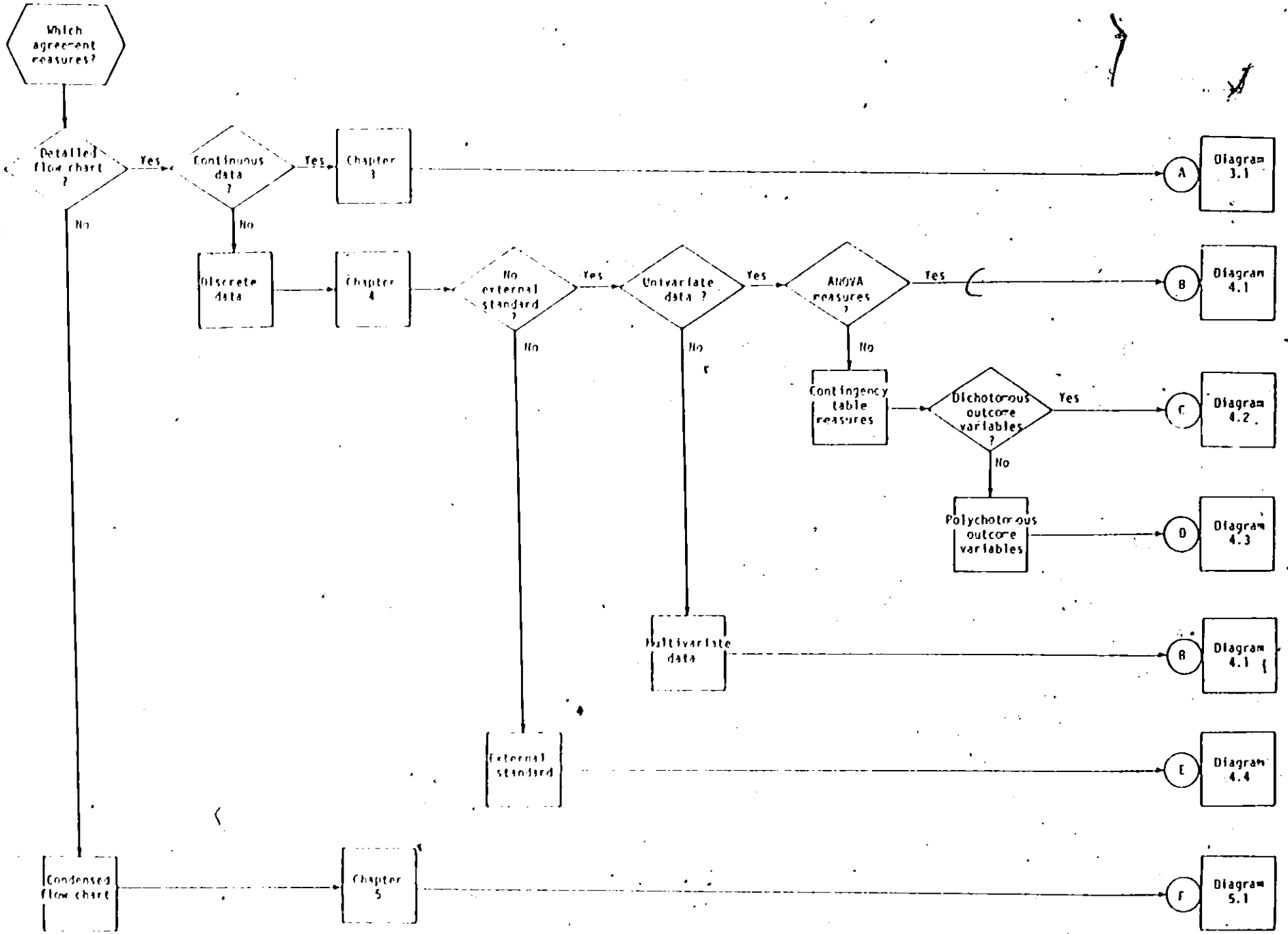
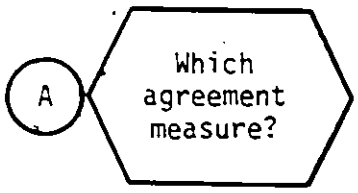
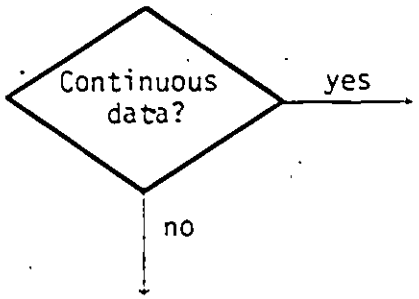
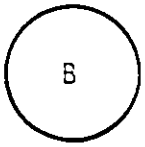
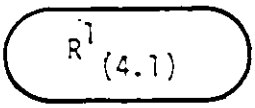


DIAGRAM 2.2

Meaning of Flow Chart Symbols

Symbol	Meaning
	Starting point for each disjointed section of the flow chart
	Decision point
	Connecting disjointed sections of flow chart
	Explanatory comment, not affecting decision flow
	Endpoints (statistical measures indexed to text)

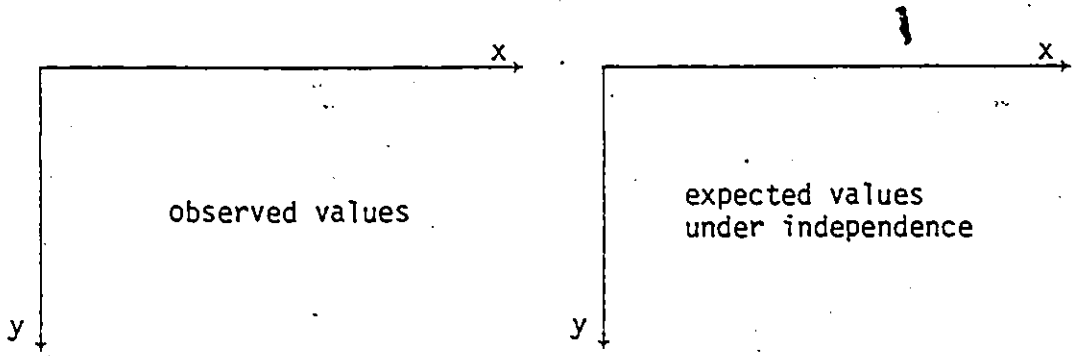
more variates whether they be measured quantitatively (on an interval or a ratio scale) or qualitatively (on a nominal or an ordinal scale). Agreement, in contrast, can be defined as "monotonic" or "diagonal" association (Diagram 2.3). In a fourfold table association, degree of dependence and agreement are all synonymous, which explains the presence of some measures of association in a flow chart and discussion of agreement analysis.

In the flow chart a distinction will consistently be made between "observer agreement" and "subject agreement" (or lack of inter-subject variation), be it "within" or "among". "Observer" and "subject" will be used and should be interpreted very broadly: "observer" as the party making the measurements and "subject" as the party on whom measurements are being made. Thus a person making judgments, a measuring instrument or method, a classification scheme or a laboratory can all be regarded as observers, whereas patients undergoing a measurement or classification or the measurements themselves could be interpreted as subjects. The idea of interchangeability of these parties is further illustrated in Diagram 2.4. The prepositions "within" or "between (among)" describe the way the agreement or disagreement is studied: with regard to observer agreement, the former refers to lack of variation in replicate measurements by the same observer (intraobserver), and the latter to the lack of variation among several observers each making at least a single observation on the same subject (interobserver). The estimable types of agreement for various study designs are shown in Table 2.1.

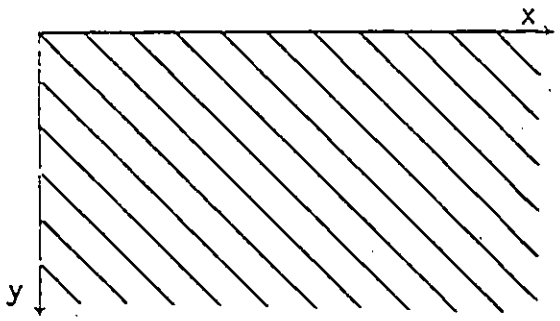
DIAGRAM 2:3

Association and Agreement

Association



Agreement



"Diagonal association" concerns values on the diagonals

DIAGRAM 2.4

"Parties" Involved in Making Observations

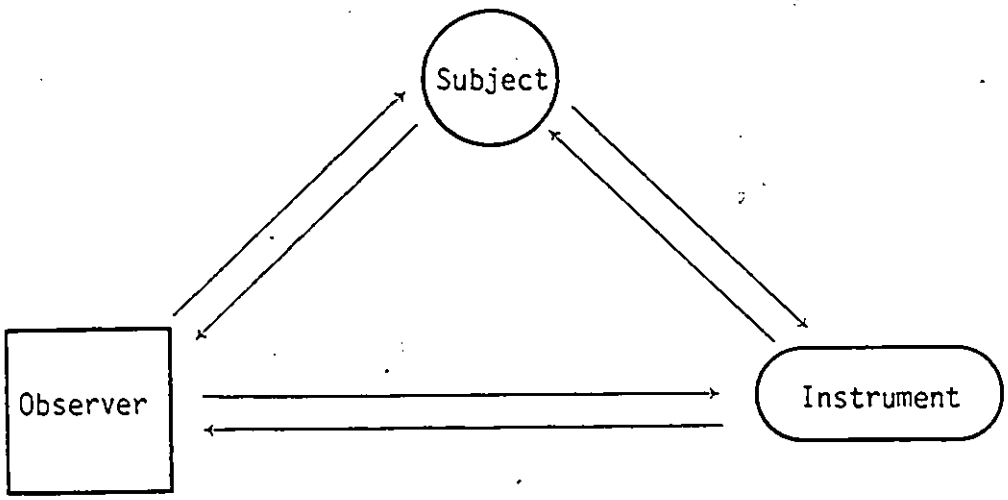


Table 2.1

Possible Study Designs and Estimable Types of Agreement

Factors		# Observations	Theoretically Estimable Agreement Components*
# Observers	# Subjects		
>1	>1	>1	Interobserver, intersubject
>1	>1	1	Intersubject, interobserver
>1	1	>1	Inter- , intraobserver, intrasubject
1	>1	>1	Inter- , intrasubject, intraobserver
>1	1	1	Interobserver
1	>1	1	Intersubject
1	1	>1	Combined intraobserver and intrasubject
1	1	1	None

* Theoretical because of limitations of estimation procedure (see Section 3.1)

2.2 Decision Points

The decision points can conveniently be divided and discussed in two broad groups: those affecting the choice and those affecting the interpretation of the agreement measure.

2.2.1 Factors Affecting the Choice of Agreement Measure

These factors are relevant because they affect the choice of agreement measure both directly and also, and importantly, via the design of the study. Their *a priori* consideration may optimize the choice of agreement measure. With the exception of those discussed in paragraph 2.2.1.5, they all appear as decision points on the flow chart.

2.2.1.1 Data Characteristics

The primary distinction according to data characteristics is between continuous and categorical (discrete) data. Continuous data refer to observations with no intervening gaps, i.e., it is possible to insert another observation between any two existing observations, irrespective of how close together they are. Categorical data refer to observations limited to discrete categories with gaps between adjacent categories. For both these types of data a further distinction will be made between single and multiple outcome variables (univariate and multivariate data). In the case of discrete data distinction will also be made between dichotomous (2 categories) and polychotomous (more than 2 categories) outcomes, and for polychotomous outcomes distinction will be made between nominal classifications and ordinal scales.

2.2.1.2 The Use of an External Standard

The use of an external standard changes the emphasis from agreement among observers (internal consistency, reliability or precision) to agreement between the observers and a standard (external consistency, validity or accuracy). The standard can be "real truth", where it is available as in chemical solutions, but more frequently a "truth indicator", a reasonable approximation of or an acceptable alternative for truth. "Truth indicators" may be independent of the observers in the study, e.g., the opinion of an "authority", or a well-established method or instrument, but need not be independent of the observers, e.g., the majority or consensus opinion of the observers.

2.2.1.3 The Study Design

The study design's importance has already been hinted at in the preamble and in Table 2.1. Table 2.2 is an extension of Table 2.1 into which formal design nomenclature used in the analysis of variance has been integrated. It illustrates a useful way* of thinking about the design of studies, also with regard to agreement analysis. Although the distinction in Table 2.2 is between single and multiple entities, in the flow chart an additional distinction between measures for two and those for more than two observers will be made. This is not as much a decision point as a point of information: most measures for more than two parties are generalizations of those for two parties.

---

* It is useful when analysis of variance is used, particularly where variance component estimation is required (see Chapter 3).

Table 2.2

Formal Design Nomenclature

Design		Factors		# Observations	Theoretically Estimable Agreement Components*
		# Observers	# Subjects		
Two-way (factor):	multiple observations per cell	>1	>1	>1	Interobserver, intersubject
	single observation per cell	>1	>1	1	Intersubject, interobserver
One-way (factor):	multiple observations per cell	>1	1	>1	Inter- , intraobserver, intrasubject
		1	>1	>1	Inter- , intrasubject, intraobserver
	single observation per cell	>1	1	1	Interobserver
		1	>1	1	Intersubject
Replicate observations		1	1	>1	Combined intraobserver and intrasubject
Single observation		1	1	1	None

* See Section 3.1

2.2.1.4 Type of Agreement

Apart from the distinction between agreement "within" and "among" parties, other points of decision are between overall (where all observations or categories are considered) and specific agreement (where only some observations or categories are considered), between unweighted (where agreement on all the observations or categories are considered equally important) and weighted agreement (where agreements or disagreement on certain observations are considered more important than others). As some agreement due to chance is possible, measures which are corrected for chance agreement will also be discussed.

2.2.1.5 Availability of Measures

The choice of statistic may be affected not only by its availability in the literature, but also by the availability of convenient computing procedures, particularly in large studies. While the flow chart is an attempt to incorporate all the published measures (in English till the end of 1978), it does not include references to the availability of computing facilities or computer programmes.

2.2.2 Factors Affecting the Interpretation of Agreement

Some of the factors mentioned in Section 2.2.1, by leading to the choice of a particular agreement measure, implicitly affects its interpretation. There are, however, some statistical considerations which affect the interpretation directly. For clarity and simplicity they are not all indicated on the flow chart, but they are all discussed here and in specifics with the relevant agreement measures.

2.2.2.1 Assumptions

Assumptions are devices which facilitate the ease with which the agreement measure is computed and usually take the form of statements about the nature of the variables, the design of the study and the nature of the observations. These assumptions need not be met to calculate the agreement measure physically, but if they are not fully justified, the distribution theory on which the calculation is based may not be satisfied and the probability statements about and the interpretation of the statistic may not be justified.

2.2.2.2 Constraints

Constraints are limits imposed by the nature of the data on the freedom with which the measure is calculated (estimated) and interpreted. Constraints are indicated and discussed in subsequent chapters when relevant.

2.2.2.3 Hypothesis Testing

Significance tests help clarify the importance that may be attached to an observed level of agreement in the contest of the study. Agreement measures for which the necessary statistical theory has been developed are indicated and the methods discussed in the text.

2.3 Notation

Standard statistical notation will be used throughout with lower case Roman letters from near the end of the alphabet indicating random variables, lower case Greek letters from the end of the alphabet

indicating the corresponding population parameters and those from the beginning of the Greek alphabet indicating constants and coefficients. Conventional statistical and mathematical symbols (e.g. n for number, p for proportion, Σ for summation) will be used, but where additional notation has to be introduced, it will be explained.

CHAPTER 3

Review: Agreement in Sets of Continuous Data

Physicians have always recognized that clinical judgments are subjective and thus liable to variation, at least in the hands of other members of the profession or men less qualified than themselves.

Fletcher and Oldham 1964, paraphrasing Todd 1955

In this chapter the estimation of agreement for continuous data will be considered. As most of the measures are based on variance component estimation, and as variance components may be conveniently estimated by analysis of variance (ANOVA) procedures, an outline of the general ANOVA procedure will first be given. It will be followed by a review of measures of agreement for univariate data first, then for multivariate data.

3.1 General ANOVA Approach

It is assumed that the reader has a basic knowledge of ANOVA. If not, a text like that by Kleinbaum and Kupper (1978) may be consulted.

In general, the ANOVA model for an individual observation is

$$y = \mu + f + e \qquad (3.1)$$

where y is an individual observation,

μ is the mean of all the observations,

f is the factor effect on that observation (here assumed to be a random effect) and

e is the random residual error (variation) associated with that observation. The factor effect, f, is dependent on the design of the study, where factor refers to observer or subject or both (Table 3.1). For example, in a balanced (equal number of observations per cell) two-way (factor) design the model containing all the components of interest is

$$y_{ijk} = \mu + s_i + d_j + (sd)_{ij} + e_{ijk} \qquad (3.2)$$

where y_{ijk} is the k-th observation in the i-th row and j-th column ($i = 1,2, \ldots, n$ subjects, $j = 1,2, \ldots, m$ observers and $k = 1,2, \ldots, \ell$ observations),

s_i is the i-th subject effect,

d_j is the j-th observer effect,

$(sd)_{ij}$ is the interaction between the i-th subject and the j-th observer,

μ and e_{ijk} are as in (3.1) and f in (3.1) is replaced by $[s_i + d_j + (sd)_{ij}]$. In a two-way design with a single observation per cell ($k = 1$) or with a single observer ($j = 1$) or subject ($i = 1$) or where either the observer or subject effect is assumed to be absent, not all the components can be estimated and the model reduces to

Table 3.1

Estimable Components of Variance

Design	Factors		# Observations	Theoretically Estimable Agreement Components*	ANOVA Model; Practically Estimable Components**
	# Observers	# Subjects			
Two-way (factor): multiple observations per cell	>1	>1	>1	Interobserver, intersubject	$y_{ijk} = u + s_i + d_j + (sd)_{ij} + e_{ijk}$
single observation per cell	>1	>1	1	Intersubject, interobserver	$y_{ij} = u + s_i + d_j + e_{ij}$
One-way (factor): multiple observations per cell	>1	1	>1	Inter- , intraobserver, intrasubjective	$y_{jk} = u + d_j + e_{jk}$
	1	>1	>1	Inter- , intrasubject, intraobserver	$y_{ik} = u + s_i + e_{ik}$
single observation per cell	>1	1	1	Interobserver	$y_j = u + e_j$
	1	>1	1	Intersubject	$y_i = u + e_i$
Replicate observations	1	1	>1	Combined intraobserver and intrasubject	$y_k = u + e_k$
Single observation	1	1	1	None	None

* See Section 3.1

$$y_{ij} = \mu + s_i + d_j + e_{ij} \tag{3.3}$$

for k = 1, i = 1,2, ..., n subjects, j = 1,2, ..., m observers and f in (3.1) is replaced by $(s_i + d_j)$, or

$$y_{ik} = \mu + s_i + e_{ik} \tag{3.4}$$

for j = 1, or no observer effect assumed, i = 1,2, ..., n subjects, k = 1,2, ..., ℓ observations and f in (3.1) is replaced by s_i, or

$$y_{jk} = \mu + d_j + e_{jk} \tag{3.5}$$

for i = 1, or no subject effect assumed, j = 1,2, ..., m observers, k = 1,2, ..., ℓ observations and f in (3.1) is replaced by d_j, respectively.

In simpler designs the model does not provide for the separation of components and is given by

$$y_i = \mu + e_i \tag{3.6}$$

for j = k = 1, i = 1,2, ..., n subjects, or

$$y_j = \mu + e_j \tag{3.7}$$

for i = k = 1, j = 1,2, ..., m observers, or

$$y_k = \mu + e_k \tag{3.8}$$

for $i = j = 1$, $k = 1,2, \ldots, \ell$ observations. By dividing the observations into groups, either arbitrarily or according to reasonable guidelines, more components may be separated out, but their interpretation is dependent on the meaningfulness of the division. In the case of a single observation by a single subject no variation is estimable.

The concept of partitioning the variation into recognizable components is important as the ANOVA measure of agreement used for continuous data is the intraclass correlation coefficient (also disguised as reliability coefficient, Cronbach's alpha, Spearman-Brown and Kuder-Richardson -20 statistics, see Bartko 1976) which has the general form

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2 + \sigma_d^2} \qquad (3.9)$$

where σ_s^2 is the variance due to subject variation,

σ_e^2 is the variance due to random variation,

σ_d^2 is the variance due to observer variation and

ρ is the intraclass correlation coefficient which expresses the proportion of the total variation due to subject variation, with a theoretical lower bound of zero when there is no subject variation at all, and a theoretical upper bound of one when both the observer and random variation terms are zero.

The assumptions underlying the models outlined before are that the response variable is a random variable with random variation, that the factor effects are random variables (i.e. each set of factor effects is a random sample from a larger population), that these sets of factor and response random effects are mutually independent and that each set

consists of independent effects which are normally distributed with mean zero and homogeneous variance for each set. These assumptions can also be written as (using the notation of model (3.1)).

$$\begin{aligned}
&\{f\} \text{ and } \{e\} \text{ are mutually independent,} \\
&\{f\} \sim N(0,\sigma_f^2) \text{ and independent (where f can be replaced by any} \\
&\qquad \text{or all of } s_i,\ d_j,\ (sd)_{ji}), \\
&\{e\} \sim N(0,\sigma_e^2) \text{ and independent}
\end{aligned} \tag{3.10}$$

and hereinafter referred to as "the usual assumptions".

If either or both factors' effects cannot be assumed to be random, their effects are regarded as fixed under the constraint that the effects always sum to zero as, e.g. $\sum_{j=1}^{m} \delta_j = 0$, where δ_j is the j-th observer effect from a fixed set of m effects.

All the usual assumptions (3.10) are necessary to fit the model and estimate the variance components by maximum likelihood (ML) procedures (Hartley and Rao 1967; Hemmerle and Hartley 1973), but not if either analysis of variance (ANOVA) procedures (Anderson and Bancroft 1952; Scheffé 1959; Searle 1971) or symmetric sum of products (SSP) procedures (Koch 1967, 1968) are used. With ANOVA and SSP procedures the assumption of normality is not necessary for fitting the model or estimating the variance components, but only for testing hypotheses about the estimators. ANOVA estimators only will be described in this chapter.

In practice, not all the properties of the general theoretical model always apply. Firstly, the theoretical upper and lower bounds are

not always achievable in practice: the realistic upper bound is often less than one as the random error term is seldom zero. A realistic upper bound of one may be created by subtracting the random error term from the denominator, a procedure for which the appropriate statistical theory has not been developed for hypotheses to be tested about agreement measures. It is an important concept to consider in the interpretation of agreement, though, particularly when the random error term is large. Due to estimation procedures (both measurement and calculation) it is conceivable that in a small proportion of studies a negative numerator may be obtained, even though it is a squared number and therefore theoretically always positive. When this occurs, the observations should be checked for measurement and calculation errors, and if none found, or if corrected and the numerator still negative, the numerator may be assigned the value of zero (or the absolute value used, or maximum likelihood estimator used), with the interpretation at the discretion of the investigator.

Secondly, a balanced design does not always result in a balanced set of results for analysis, due to withdrawals, data loss, *a posteriori* decisions on variables of interest, to name a few reasons. In unbalanced situations the usual ANOVA calculations cannot be used to estimate all the variance components exactly due to the non-orthogonality of the data: the total sum of squares cannot be partitioned exactly into its constituent components without overlap. In these situations alternative procedures have to be followed: the simplest is the method of unweighted means which may be used if no cell has a sample size more

than twice that of any other. By the method of unweighted means the factor components are estimated as before, but the estimate of σ_e^2 is replaced by

$$\sigma_{av}^2 = \frac{1}{nm}(\sigma_e^2) \sum_{i=1}^{\bar{n}} \sum_{j=1}^{m} \frac{1}{n_{ij}} \quad . \tag{3.11}$$

Should any cell have a sample size more than twice that of any other, one of several other more complex but exact methods may be used (Searle 1971, chapter 10). Multiple regression analysis may also be used for all designs (balanced, unbalanced and those more complex than mentioned thus far), but not in hand computational form. In the unbalanced situation the variance components are estimated as conditional estimates by multiple regression and the order in which the factors are entered into the regression equation is decided by the investigator (Kleinbaum and Kupper 1978). Intraclass correlation coefficients for three-way designs are discussed by Maxwell and Pilliner (1968).

There is one unbalanced design in which these special methods need not be used, but the ANOVA estimates may be used: where the cell frequencies are proportional, satisfying

$$n_{ij} = \frac{n_{i.} n_{.j}}{n_{..}} \tag{3.12}$$

where n_{ij} is the frequency in the cell in the i-th row and j-th column,

$n_{i.}$ is the i-th row total,

$n_{.j}$ is the j-th column total and

$n_{..}$ is the overall total.

3.2 Agreement Measures for Continuous Data

They are shown in flow chart form on Diagram 3.1.

3.2.1 Univariate Data With No External Standard

These measures are all variations of the intraclass correlation coefficient in (3.9) depending on the design of the study. In the model

$$y_{ij} = \mu + s_i + e_{ij} \tag{3.13}$$

as applied to a two-way design of random samples of m observers and n subjects,

$$\rho_1 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \tag{3.14}$$

which can be estimated from an ANOVA table by

$$\hat{\rho}_1 = \frac{MS_s - MS_e}{MS_s + (m-1)MS_e}, \quad (MS_s - MS_e \geq 0) \tag{3.15}$$

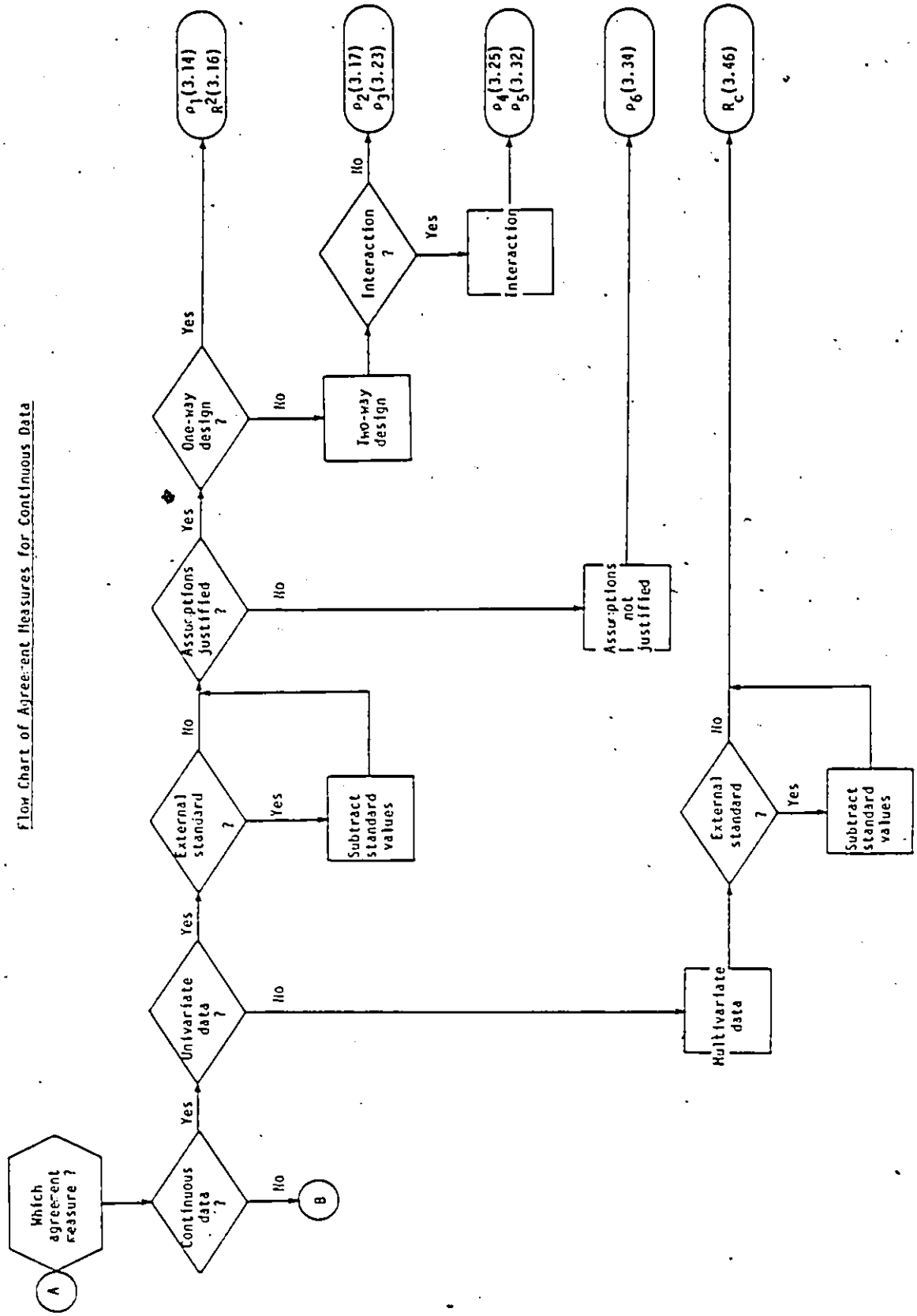
where MS_s is the mean square for subjects and

MS_e is the mean square for residual error. In this model (Ebel 1951) the absence of observer effect is assumed and the residual error includes whatever observer effect was present in the observations plus any interaction present or the inherent random variation. As such, $\hat{\rho}_1$ is a measure of intra-observer agreement, provides no information on interobserver variation and reflects the proportion of total variation attributable to the subjects.

The model in (3.13) may also be applied to the situation where one observer makes replicate observations on n subjects, or where one

DIAGRAM 3.1

Flow Chart of Agreement Measures for Continuous Data



observer makes a single observation on n subjects, with the investigator prepared to divide the subjects arbitrarily to provide "replicates" for the estimation of σ_e^2. In both cases no interobserver variation can be observed or estimated.

An additional and equivalent measure, in which the numerator consists of a single term as estimate of σ_s^2 and could therefore be used if $(MS_s - MS_e) < 0$ in (3.15), is

$$R^2 = \frac{(n-1)MS_s}{(n-1)MS_s + n(m-1)MS_e} \qquad (3.16)$$

described by Robinson (1975).

Example 3.1

Using the data from Table 1.1, assuming no observer effect, the lay-out is as shown in Table 3.2 with an ANOVA Table calculated as in Table 3.3.

$$y_{ik} = \mu + s_i + e_k \qquad (3.13)$$

where $i = 1,2, \ldots, 5$ patients, $k = 1,2, \ldots, 12$ observations on each patient and where

y_{ik} *is the k-th observation on the i-th patient*

μ *is the overall mean*

s_i *is the i-th subject or patient effect (variation)*

e_{ik} *is the random variation associated with the k-th observation on the i-th patient. Intraclass correlation coefficients are then calculated as follows for subject (patient) variation from Table 3.4:*

Table 3.2

General Data Layout for Balanced One-way Design*
(adapted from Kleinbaum and Kupper 1978)

Subjects (n)	Observations (ℓ)	Totals	Sample Means
1	$y_{11}, y_{12}, \cdots\cdots\cdots, y_{1\ell}$	$T_1 = y_{1.}$	$\bar{y}_{1.} = T_1/\ell$
2	$y_{21}, y_{22}, \cdots\cdots\cdots, y_{2\ell}$	$T_2 = y_{2.}$	$\bar{y}_{2.} = T_2/\ell$
.	. , . .	.	.
.	. , . .	.	.
.	. , . .	.	.
n	$y_{n1}, y_{n2}, \cdots\cdots\cdots, y_{n\ell}$	$T_n = y_{n.}$	$\bar{y}_{n.} = T_n/\ell$
		$G = y_{..}$	$\bar{y} = G/\ell n$

* If unbalanced, sample means are obtained by dividing the totals by the number of observations on the relevant patient (ℓ_i), and the grand total (G) by the sum of the number of observations on all patients ($\sum_{i=1}^{n} \ell_i$).

Table 3.3

General ANOVA Table For Balanced* One-way Design
(adapted from Kleinbaum and Kupper 1978)

Source	d.f.	SS	MS
Subjects	$n - 1$	$SS_s = \sum_{i=1}^{n} (T_i^2/\ell) - G^2/\ell n$	$MS_s = SS_s/(n-1)$
Error	$\ell n - n$	$SS_e = TSS - SS_s$	$MS_e = SS_e/(\ell n-n)$
Total	$\ell n - 1$	$TSS = \sum_{i=1}^{n} \sum_{k=1}^{\ell} y_{ik}^2 - G^2/\ell n$	

* If unbalanced, ℓ is replaced by ℓ_i, and ℓn is replaced by $\sum_{i=1}^{n} \ell_i$.

$$\rho_1 = \frac{MS_s - MS_e}{MS_s + (m-1)MS_e}$$

$$= \frac{5758.4333 - 5.1939}{5758.4333 + (11)5.1939}$$

$$= 0.9893$$

or

$$R^2 = \frac{(n-1)MS_s}{(n-1)MS_s + n(m-1)MS_e}$$

$$= \frac{(4)5758.4333}{(4)5758.4333 + 4(11)5.1939}$$

$$= 0.9877$$

The results suggest that nearly all the variation is due to differences (disagreements) among the patients.

Where observer effects are not assumed to be absent, but interactions are, two other intraclass correlation coefficients have been defined (Ebel 1951; Burdock, Fleiss and Hardesty 1963). When both the observer and subject effects are assumed to be random factors the model, under the usual assumptions (3.10) for a two-way design with a single observation per cell or with the additional assumption of no interaction between observers and subjects, is

$$y_{ij} = \mu + s_i + d_j + e_{ij}, \qquad \begin{matrix} i = 1,2, \ldots, n \text{ patients}, \\ j = 1,2, \ldots, m \text{ observers} \end{matrix} \qquad (3.3)$$

and

Table 3.4

ANOVA Table For Blood Pressure Data Analysed As A One-way Design
(observer effect assumed to be absent)

Source	d.f.	Sums of Squares	Mean Squares
Patients	4	23033.7333	5758.4333
Random Error	55	285.6666	5.1939
Total	59	23319.4000	

$$\rho_2 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2 + \sigma_e^2} \tag{3.17}$$

which is estimated by

$$\hat{\rho}_2 = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2} \tag{3.18}$$

with the intersubject variance estimate

$$\hat{\sigma}_s^2 = (MS_s - MS_e)/m, \ (MS_s - MS_e \geq 0), \tag{3.19}$$

the interobserver variance estimate

$$\hat{\sigma}_d^2 = (MS_d - MS_e)/n, \ (MS_d - MS_e \geq 0), \tag{3.20}$$

and the random variance estimate

$$\hat{\sigma}_e^2 = MS_e \ . \tag{3.21}$$

When the m observers are a fixed set, the ANOVA mixed model, with design and assumptions as for the random model, is

$$y_{ij} = \mu + s_i + \delta_j + e_{ij}, \ i = 1,2, \ldots, n, \ j = 1,2, \ldots, m, \tag{3.22}$$

under the constraint that $\sum_{j=1}^{m} \delta_j = 0$, then

$$\rho_3 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \tag{3.23}$$

which is estimated by

$$\rho_3 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \tag{3.24}$$

which is analogous to ρ_2 in (3.18), but the variance component estimates pertain to model (3.22).

These two measures, ρ_2 and ρ_3, therefore estimate the proportion of the total variation attributable to intersubject variation, or interobserver variation (when σ_s^2 is replaced by σ_d^2 in the numerator of ρ_2).

Example 3.2

The data in Table 1.1 may be analysed as a two-way design with a single observation per cell, as detailed in Table 3.5 and 3.6. However, to preserve continuity in building up the models in the examples in this chapter, the same data will now be analysed as a two-way design (3 types of observers, 5 patients and 4 observations by each type of observer on each patient), as detailed in Table 3.7 and the analysis done in Table 3.8. In doing this analysis interaction between type of observer and patient is assumed absent and regarded as random error and the model under consideration is

$$y_{ijk} = \mu + s_i + d_j + e_{ijk} \tag{3.3}$$

where $i = 1,2, \ldots, 5$ *patients,* $j = 1,2,3$ *types of observers,* $k = 1,2, \ldots, 4$ *observations and where*

y_{ijk} *is the* k*-th observation by the* j*-th type of observer on the*

Table 3.5

General Data Layout For A Two-way Design With A Single Observation Per Cell
(adapted from Kleinbaum and Kupper 1978)

Subjects (n)	Observers (m)	Total	Mean
1	$y_{11}, y_{12}, \cdots\cdots\cdots, y_{1m}$	R_1	$\bar{y}_{1.}$
2	$y_{21}, y_{22}, \cdots\cdots\cdots, y_{2m}$	R_2	$\bar{y}_{2.}$
.	. , . .	.	.
.	. , . .	.	.
.	. , . .	.	.
n	$y_{n1}, y_{n2}, \cdots\cdots\cdots, y_{nm}$	R_n	$\bar{y}_{n.}$
Total	$C_1, C_2, \cdots\cdots\cdots, C_m$	G	-
Mean	$\bar{y}_{.1}, \bar{y}_{.2}, \cdots\cdots\cdots, \bar{y}_{.m}$	-	$\bar{y}_{..}$

Table 3.6

General ANOVA Table For A Two-way Design With A Single Observation Per Cell
(adapted from Kleinbaum and Kupper 1978)

Source	d.f.	SS	MS
Subjects	$n - 1$	$SS_s = \frac{1}{m} \sum_{i=1}^{n} R_i^2 - \frac{G^2}{nm}$	$MS_s = SS_s/(n-1)$
Observers	$m - 1$	$SS_d = \frac{1}{n} \sum_{j=1}^{m} C_j^2 - \frac{G^2}{nm}$	$MS_d = SS_d/(m-1)$
Error	$(n-1)(m-1)$	$SS_e = TSS - SS_s - SS_d$	$MS_e = SS_e/(n-1)(m-1)$
Total	$nm - 1$	$TSS = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}^2 - \frac{G^2}{nm}$	

Table 3.7

General Data Layout For A Two-way Design With Equal Numbers of Observations Per Cell
(balanced design, adapted from Kleinbaum and Kupper 1978)

Subjects (n)	Observers (m), Replicates (ℓ)				Row Totals
	1	2	. . .	m	
1	$y_{111}, y_{112}, \cdots, y_{11\ell}$ T_{11}	$y_{121}, y_{122}, \cdots, y_{12\ell}$ T_{12}	. . .	$y_{1m1}, y_{1m2}, \cdots, y_{1m\ell}$ T_{1m}	R_1
2	$y_{211}, y_{212}, \cdots, y_{21\ell}$ T_{21}	$y_{221}, y_{222}, \cdots, y_{22\ell}$ T_{22}	. . .	$y_{2m1}, y_{2m2}, \cdots, y_{2m\ell}$ T_{1m}	R_2
. . .	. . .	. . .	. . .	. . .	. . .
n	$y_{n11}, y_{n22}, \cdots, y_{n2\ell}$ T_{n1}	$y_{n21}, y_{n22}, \cdots, y_{n2\ell}$ T_{n2}	. . .	$y_{nm1}, y_{nm2}, \cdots, y_{nm\ell}$ T_{nm}	R_n
Column Totals	C_1	C_2	. . .	C_m	G

Table 3.8

General ANOVA Table For Balanced Two-way Designs
(adapted from Kleinbaum and Kupper 1978)

Source	d.f.	SS	MS
Subject (main effect)	$n - 1$	$SS_s = \frac{1}{m\ell}\sum_{i=1}^{n} R_i^2 - \frac{G^2}{nm\ell}$	$MS_s = SS_s/(n-1)$
Observer (main effect)	$m - 1$	$SS_d = \frac{1}{n\ell}\sum_{j=1}^{m} C_j^2 - \frac{G^2}{nm\ell}$	$MS_d = SS_d/(m-1)$
Interaction (Subject × Observer)	$(n-1)(m-1)$	$SS_{(sd)} = \frac{1}{\ell}\sum_{i=1}^{n}\sum_{j=1}^{m} T_{ij}^2 - SS_s - SS_d - \frac{G^2}{nm\ell}$	$MS_{(sd)} = SS_{(sd)}/(n-1)(m-1)$
Random Error	$nm(\ell-1)$	$SS_e = TSS - SS_s - SS_d - SS_{(sd)}$	$MS_e = SS_e/[nm(\ell-1)]$
Total	$nm\ell - 1$	$TSS = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{\ell} y_{ijk}^2 - \frac{G^2}{nm\ell}$	

i-th patient,

μ *is the overall mean,*

s_i *is the i-th patient effect (subject effect),*

d_j *is the j-th type of observer effect,*

e_{ijk} *is the random error associated with* y_{ijk}, *where, both observer types and patient effects are random.*

For fixed observer effects, d_j *in (3.3) is replaced by* δ_j *such that* $\sum_{j=1}^{3} \delta_j = 0$, *and if the patient effects are regarded as fixed,* s_i *is replaced in (3.3) by* σ_i *such that* $\sum_{i=1}^{5} \sigma_i = 0$.

The resultant ANOVA Table is shown in Table 3.9, remembering that SS_e *was estimated as* $TSS - SS_s - SS_d$, *interaction being assumed absent.*

Variance Components

$$\sigma_s^2 = (MS_s - MS_e)/m = (5758.4\dot{3} - 4.8635)/3 = 1917.8566$$

$$\sigma_d^2 = (MS_d - MS_e)/n = (13.95 - 4.8635)/4 = 2.271\dot{6}$$

$$\sigma_e^2 = MS_e = SS_e/53 = \frac{56.266\dot{6} + 201.5}{53} = 4.8635$$

Random Effects Model

Variation:

$$\rho_2 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2 + \sigma_e^2} = \frac{1917.8566}{1917.8566 + 2.271\dot{6} + 4.8635} = 0.9963$$

or, 99.63% of the total variation is due to differences (disagreements) among patients.

$$\rho_2 = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_s^2 + \sigma_e^2} = \frac{2.271\dot{6}}{2.271\dot{6} + 1917.8566 + 4.8635} = 0.0012$$

Table 3.9

Two-way ANOVA Table For Blood Pressure Data, Assuming No Interaction

Source	d.f.	Sums of Squares	Mean Squares
Patients	4	23033.7333	5758.4333
Observer Types	2	27.9000	13.9500
Random Error	53	257.7666	4.8635
Total	59	23319.4000	

or, 0.12% of the total variation is due to differences (disagreements) among the 3 types of observers.

Mixed Effects Model

With observer effects fixed, the intraclass correlation coefficient for patients is

$$\rho_3 = \frac{\theta_s^2}{\theta_s^2 + \theta_e^2} = \frac{1917.8566}{1917.8566 + 4.8635} = 0.9975$$

or, 99.75% of the total variation is due to differences (disagreements) among the patients.

In a balanced two-way design situation with multiple observations per cell, the ANOVA model can be extended to include interaction effects, both for random and fixed observer effects (Landis and Koch 1975). For random effects of n subjects and m observers, the model is

$$y_{ijk} = \mu + s_i + d_j + (sd)_{ij} + e_{ijk}, \quad \begin{matrix} i = 1,2, \ldots, n, \\ j = 1,2, \ldots, m, \\ k = 1,2, \ldots, \ell \end{matrix} \qquad (3.2)$$

under the usual assumptions (3.10) and with k = 1,2, ..., ℓ observations and

$$\rho_4 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2 + \sigma_{sd}^2 + \sigma_e^2} \qquad (3.25)$$

which is estimated by

$$\rho_4 = \frac{\theta_s^2}{\theta_s^2 + \theta_d^2 + \theta_{sd}^2 + \theta_e^2} \quad . \tag{3.26}$$

with the intersubject variance estimate given by

$$\theta_s^2 = (MS_s - MS_{sd})/m\ell, \ (MS_s - MS_{sd} \geq 0), \tag{3.27}$$

the interobserver variance estimate

$$\theta_d^2 = (MS_d - MS_{sd})/n\ell, \ (MS_d - MS_{sd} \geq 0), \tag{3.28}$$

the estimate of the variance in the observers' overall rating of the same subject given by

$$\theta_{sd}^2 = (MS_{sd} - MS_e)/\ell, \ (MS_{sd} - MS_e \geq 0) \text{ and} \tag{3.29}$$

the residual variance estimate

$$\theta_e^2 = MS_e \quad . \tag{3.30}$$

The mixed model for fixed observer effects is

$$y_{ijk} = \mu + s_i + \delta_j + (s\delta)_{ij} + e_{ijk} \tag{3.31}$$

under the usual assumptions (3.10) and with the constraints $\sum_{j=1}^{m} \delta_j = 0$ and $\sum_{i=1}^{n} \sum_{j=1}^{m} (s\delta)_{ij} = 0$. The agreement measure is

$$\rho_5 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \tag{3.32}$$

which is estimated by

$$\rho_5 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \tag{3.33}$$

which is equivalent to ρ_4 in (3.26), but the variance component estimates pertain to model (3.31).

These two measures, ρ_4 and ρ_5, therefore estimate the proportion of the total variation attributable to intersubject variation, or interobserver variation (if σ_s^2 in the numerator of ρ_4 is replaced by σ_d^2).

Example 3.3

Using the data in Table 1.1 in a two-way layout as in Table 3.7 with the observers grouped according to their training (clinicians, statisticians and others), the ANOVA Table is drawn up as in Table 3.8, for the model

$$y_{ijk} = \mu + s_i + d_j + (sd)_{ij} + e_{ijk} \tag{3.2}$$

where $i = 1,2, \ldots, 5$ patients, $j = 1,2,3$ types of observers, $k = 1,2, \ldots, 4$ observations by each type of observer on each patient and where y_{ijk}, μ, s_i, d_j and e_{ijk} are as defined in Example 3.2, but

$(sd)_{ij}$ is the interaction between the i-th patient and the j-th type of observer.

If the observer type effects are a fixed set, d_j in (3.2) is replaced by δ_j and $(sd)_{ij}$ in (3.2) by $(s\delta)_{ij}$ such that $\sum_{i=1}^{m} \delta_j = 0$ and

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (s\delta)_{ij} = 0.$$

Intraclass correlation coefficients for these models are estimated as follows, from the ANOVA Table in Table 3.10:

Variance Components

$$\theta_s^2 = (MS_s - MS_{sd})/m\ell = (5758.4333 - 7.0334)/12 = 479.2833$$
$$\theta_d^2 = (MS_d - MS_{sd})/n\ell = (13.95 - 7.0334)/20 = 0.3458$$
$$\theta_{sd}^2 = (MS_{sd} - MS_e)/\ell = (7.0334 - 4.4778)/4 = 0.6389$$
$$\theta_e^2 = (MS_e) = 4.7778$$

Random Effects Model

Patient Variation:

$$\rho_4 = \frac{\theta_s^2}{\theta_s^2 + \theta_d^2 + \theta_{sd}^2 + \theta_e^2} = \frac{479.2833}{479.2833 + 0.3458 + 0.6389 + 4.7778} = 0.9887,$$

or, 98.87% of the total variation is due to differences (disagreements) among patients.

Observer Variation:

$$\rho_4 = \frac{\theta_s^2}{\theta_d^2 + \theta_d^2 + \theta_{sd}^2 + \theta_e^2} = \frac{0.3458}{0.3458 + 479.2833 + 0.6389 + 4.7778} = 0.0007,$$

or, 0.07% of the total variation is due to differences (disagreements) among observer types.

Mixed Effects Model

Patient Variation (Observer Effect Fixed):

Table 3.10

Two-way ANOVA Table For Blood Pressure Data Testing For Interaction

Source	d.f.	Sums of Squares	Mean Squares
Patients	4	23033.7333	5758.4333
Observer Types	2	27.9000	13.9500
Interaction	8	56.2666	7.0334
Random Error	45	201.5000	4.4778
Total	59	23319.4000	

$$\rho_S = \frac{\theta_s^2}{\theta_s^2 + \theta_e^2} = \frac{479.2833}{479.2833 + 4.7778} = 0.9907,$$ *or 99.07% of the total variation is due to differences (disagreements) among patients.*

For all these measures mentioned thus far it is possible to construct confidence intervals for or test hypotheses about the individual variance component estimates and for the intraclass correlation coefficients under the assumptions of normality and independence. Their construction is discussed in Ebel (1951), Anderson and Bancroft (1952) and Scheffé (1959), the general principle being the comparison of ratios of mean squares to appropriate F distributions.

Example 3.4

To illustrate hypothesis testing about intraclass correlation coefficient estimates, the analysis in Example 3.1 will be used. The intraclass correlation coefficient ρ_1 *is estimated by*

$$\rho_1 = \frac{\theta_s^2}{\theta_s^2 + \theta_e^2} \qquad (3.16)$$

and $\rho_1 = 0$ *if the null hypothesis* $\sigma_s^2 = 0$ *is true. The null hypothesis* $\sigma_s^2 = 0$ *may be tested, by the usual variance ratio:*

$$MS_s/MS_e = F_{1-\alpha,\nu_1,\nu_2}$$

where MS_s, MS_e *are as in Table 3.3*

α *is the probability of falsely rejecting the null hypothesis,*

ν_1 *is the numerator degrees of freedom,*

ν_2 is the denominator degrees of freedom.

In Example 3.1, the F ratio is 5758.4333/5.1939 = 1108.69 (Table 3.4), with $\nu_1 = 4$, $\nu_2 = 55$, so that $p << 0.001$ and the null hypothesis is rejected.

Using Ebel's method (1951), confidence intervals for ρ_1 may be constructed as follows:

From the ANOVA Table, the observed variance ratio is $F = MS_s/MS_e$, with ν_1 and ν_2 degrees of freedom as before.

Under Ho: $\sigma_s^2 = 0$, the population variance ratio, $F_p = 1$, and its confidence limits are given by:

$$\text{upper confidence limit} = F_{pu} = F \times F_{tu}$$

where F_{tu} is the maximum ratio expected under the null hypothesis, obtained from tables of the F distribution as $F_{1-\alpha/2,\nu_1,\nu_2}$, and

$$\text{lower confidence limit} = F_{pl} \times \left(\frac{1}{F_{tl}}\right)$$

where F_{tl} is the minimum ratio expected under the null hypothesis, obtained from tables from the F distribution as $F_{1-\alpha/2,\nu_2,\nu_1}$.

Remembering that

$$\rho_1 = \frac{MS_s - MS_e}{MS_s - (m-1)MS_e} = \frac{MS_s/MS_e - MS_e/MS_e}{MS_s/MS_e - (m-1)MS_e/MS_e} \qquad (3.15)$$

$$= \frac{F - 1}{(F-1) + m} , \qquad (3.15a)$$

the upper confidence limit for ρ_1 may be obtained by substituting F_{pu} for F in (3.15a) and the lower confidence limit for ρ_1 by substituting $F_{p\ell}$ for F in (3.15a).

From Example 3.1, Ho: $\sigma_s^2 = 0$, or $\rho_1 = 0$, $F = 5758.4333/5.1939$ $= 1108.69$, $\nu_1 = 4$, $\nu_2 = 55$, assume $\alpha = 5\%$.

So, $F_{tu} = F_{0.025,4,55} = 3.03$ (from tables), therefore

$$F_{pu} = F \times F_{tu} = 1108.69 \times 3.03 = 3359.33,$$

and therefore

$$\text{upper 2.5\% confidence limit for } \rho_1 = \frac{3359.33 - 1}{(3359.33-1) + 12}$$

$$= \underline{0.9964} .$$

Similarly,

$$F_{t\ell} = F_{0.025,55,4} = 8.374, \text{ therefore}$$

$$F_{p\ell} = F(1/F_{t\ell}) = 1108.69 \times 1/8.374 = 132.40$$

and therefore

$$\text{the lower 2.5\% confidence limit for } \rho_1 = \frac{132.40 - 1}{(132.40-1) + 12}$$

$$= \underline{0.9163} .$$

The 95% confidence limits for ρ_1 based on the estimate $\hat{\rho}_1$ therefore do not span 0, so that Ho is rejected.

For more complex models the principles are the same, but as the denominator is a combination of the variance components of several factors with the random error component, the degrees of freedom have to be approximated as described by Satterthwaite (1946).

3.2.2 Univariate Data With Unjustified Assumptions

One of the assumptions that may not always hold is that of variance homogeneity for the random error set. As an alternative to the use of a linear transformation to try and stabilize the variance, an ANOVA model has been formulated to cope with this situation (Grubbs 1948), where the m observer effects are fixed:

$$y_{ij} = \mu + s_i + \delta_j + e_{ij} \qquad (3.34)$$

where i = 1,2, ..., n subjects, j = 1,2, ..., m observers under the assumptions, now, that

$\{s_i\}$ are $\sim N(0,\sigma_s^2)$ and independent,

$\{e_{ij}\}$ are $\sim N(0,\sigma_{ej}^2)$ and independent and

$\{s_i\}, \{e_{i1}\}, \ldots, \{e_{im}\}$ are all mutually independent.

For each observer a separate random variance component (σ_{ej}^2) is therefore specified and incorporated into the denominator of ρ_6, analogous to ρ_3, under the constraint that $\sum_{j=1}^{m} \delta_j = 0$. The variance components are estimated differently, though, where

$$\text{var}\{y_{ij}\} = \sigma_s^2 + \sigma_{ej}^2, \tag{3.35}$$

$$\text{cov}\{y_{ij}, y_{ij'}\} = \sigma_s^2 \text{ for } j \neq j' . \tag{3.36}$$

By setting

$$s_{jj'} = \frac{1}{n-1} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)(y_{ij'} - \bar{y}_{j'}) \tag{3.37}$$

for $j,j' = 1,2, \ldots, m$ where

$$\bar{y}_j = \sum_{i=1}^{n} y_{ij}/n \quad , \tag{3.38}$$

$$\hat{\sigma}_{ej}^2 = s_{jj} - \frac{2}{m-1} \sum_{j \neq j'}^{m} s_{jj'} + \frac{2}{(m-1)(m-2)} \sum_{\substack{k<k' \\ k,k' \neq j}}^{m} s_{kk'} \tag{3.39}$$

and

$$\hat{\sigma}_s^2 = \frac{2}{m(m-1)} \sum_{1 \leq j < j'}^{m} s_{jj'} . \tag{3.40}$$

Where $m = 2$, these expressions simplify to

$$\hat{\sigma}_{e1}^2 = s_{11} - s_{12} , \tag{3.41}$$

$$\hat{\sigma}_{e2}^2 = s_{22} - s_{12} , \tag{3.42}$$

$$\sigma_s^2 = s_{12} \qquad (3.43)$$

Example 3.5

Using the data from Table 1.1, assuming the random variance to be non-homogeneous, intraclass correlation coefficients may be estimated as shown below. To simplify the calculations, only the data from the first observer in each training group were used (see Table 3.11). The variance components are estimated from the estimates of variance for each type of observer ($s_{jj'}$, where $j = j'$) and the covariance for each pairwise combination of observer types ($s_{jj'}$, where $j \neq j'$). Instead of the formulations in (3.7) and (3.8) the variances and covariances may be estimated by the usual, more convenient computational formulae as follows:

Variances (j=j')

$$s_{jj'} = \frac{1}{n-1}\left\{\sum_{i=1}^{n} y_{ij}^2 - \frac{(\Sigma y_{ij})^2}{n}\right\}$$

so that the variance for observer type 1 is estimated by

$$s_{11} = \frac{1}{2}\left\{\sum_{i=1}^{5} y_{i1}^2 - \frac{(\Sigma y_{i1})^2}{n}\right\}$$

$$= 459.20 \ .$$

Similarly, $s_{22} = 715.20$ *and* $s_{33} = 492.80$.

Covariances ($j \neq j'$)

$$s_{jj'} = \frac{1}{n-1}\left\{\sum_{i=1}^{n} y_{ij}y_{ij'} - \frac{\left(\sum_{i=1}^{n} y_{ij}\right)\left(\sum_{i=1}^{n} y_{ij'}\right)}{n}\right\}$$

so that the covariance for observer types 1 and 2 is

$$s_{12} = s_{21} = \frac{1}{4}\left\{\sum_{i=1}^{n} y_{i1}y_{i2} - \frac{\left(\sum_{i=1}^{n} y_{i1}\right)\left(\sum_{i=1}^{n} y_{i2}\right)}{n}\right\}$$

$$= 571.20$$

Similarly, $s_{13} = s_{31} = 474.40$ *and* $s_{23} = s_{32} = 592.40$.

Variance Components

$$\hat{\sigma}_s^2 = \frac{2}{3(2)}(s_{12}+s_{13}+s_{23}) = \frac{2}{6}(571.2+474.4+592.4) = 546.00$$

Table 3.11

Table of Blood Pressure Readings by Three Selected Observers

	Observer Type		
Patient	Clinician(1)	Statistician(2)	Other(3)
1	134	134	132
2	122	122	120
3	166	178	168
4	170	178	168
5	134	134	134

$$\hat{\sigma}^2_{e1} = s_{11} - \frac{2}{2}(s_{21}+s_{31}) + \frac{2}{2(1)}(s_{23}) = 459.2 - (571.2+474.4) + 592.4$$
$$= 6.00$$

similarly,

$$\hat{\sigma}^2_{e2} = 26.00$$

$$\hat{\sigma}^2_{e3} = -2.85 \approx 0 \text{ (alternatively } \hat{\sigma}^2_{e3} \approx |\hat{\sigma}^2_{e3}| = 2.85, \text{ or maximum}$$

likelihood estimators used which are not easily calculated by hand and not shown here).

Intraclass Correlation Coefficient

Patient Variation (Observer Effects Fixed)

$$\rho_6 = \frac{\sigma^2_s}{\sigma^2_s + \sum_{j=1}^{3} \sigma^2_{ej}} = \frac{546.00}{546.00 + 6.00 + 26.00 + 0} = 0.9446 \quad (\sigma^2_{e3}=0)$$

or

$$\rho_6 = \frac{}{\sigma^2_s + \sum_{j=1}^{3} \sigma^2_{ej}} = \frac{546.00}{546.00 + 6.00 + 26.00 + 2.85} = 0.9400$$
$$(\sigma_{e3}=2.85)$$

In this situation 94.46% (or 94.00%) of the total variation is due to differences (disagreements) among patients.

This method, as well as an analogous one by Smith (1960), was developed to estimate the precision of instruments. Overall (1968) considered the situation in psychological and psychiatric research where rater reliability was the issue and by doing a separate analysis of variance for each rater, separate estimates of the random error variance

components were made for use in separate reliability measures for each rater using intraclass correlation coefficients.

Another situation in which all the assumptions need not be met, is that of survey sampling, where the objective is descriptive rather than hypothesis testing. Two measures, analogous to intraclass correlation coefficients have been developed by Hansen, et al. (1964, "the index of inconsistency") and Kish (1962).

3.2.3 Multivariate Data Without An External Standard

For the situation where m observers measure c outcome variables on n subjects, Fleiss (1966) proposed a model which is a direct extension of the ANOVA model for univariate data and fixed observer effects (3.22):

$$\underset{c\times\ell}{\underline{y}_{ij}} = \underline{\mu} + \underline{s}_i + \underline{o}_j + \underline{e}_{ij} \qquad (3.44)$$

where an underlined symbol indicates that the parameter or effect refers to a vector or a matrix and

$\underset{c\times\ell}{\underline{y}_{ij}}$ is the vector of c scores assigned to the i-th subject by the j-th observer and given by

$$\underset{\ell\times c}{\underline{y}'_{ij}} = (y_{ij}^{(1)}, y_{ij}^{(2)}, \ldots, y_{ij}^{(c)}) \qquad (3.45)$$

and $\underline{\mu}$ is an overall mean vector for the c characteristics,

$\underset{c\times\ell}{\underline{s}_i}$ is the vector of effects for the i-th subject,

$\underset{c\times\ell}{\underline{\delta}_j}$ is the vector of effects for the j-th observer and

$\underline{e}_{ij}$ is the vector of residual error for the c characteristics of the (i,j)-th observation.

The assumptions for this model are that $\underline{\mu}$ and $\underline{\delta}$ are fixed constants, that the $\{s_i\}$ and $\{e_{ij}\}$ are c-variate normal, independent and identical vectors respectively, or

$\{\underline{s}_i\} \approx N_c(\underline{0},\underline{V}_s)$ and independent, and

$\{\underline{e}_{ij}\} \approx N_c(\underline{0},\underline{V}_e)$ and independent.

The agreement is then measured by

$$R_c = \frac{1}{c} \text{ trace } \underline{V}_s(\underline{V}_s+\underline{V}_e)^{-1} \qquad (3.46)$$

When both $\underline{V}_s$ and $\underline{V}_e$ are diagonal, R_c reduces to the arithmetic mean of the univariate intraclass correlation coefficients as in (3.14) associated with each of the c characteristics. Fleiss also derived a likelihood ratio test criterion for the c-variate hypothesis of no interobserver bias. Other models were proposed by Gleser, Cronbach and Rajaratnam (1965) and Maxwell and Pilliner (1968) for the experimental situation and by Koch (1973) and Bershad (1969) for sample survey situations.

3.2.4 Agreement With An External Standard

The use of an external standard shifts the emphasis to agreement between observers and the standard, rather than among the observers. For continuous data that source of variation can be isolated by subtracting the value of the standard from each observed value,

analysing the differences in exactly the same way as described in Sections 3.2.1 and 3.2.2.

CHAPTER 4

Review: Agreement in Sets of Categorical Data

All men are forced into one of two categories; those with eleven fingers and those without.

Ned Rorem

In this chapter the estimation of agreement in sets of discrete data will be considered, first for univariate and then, briefly, for multivariate data.

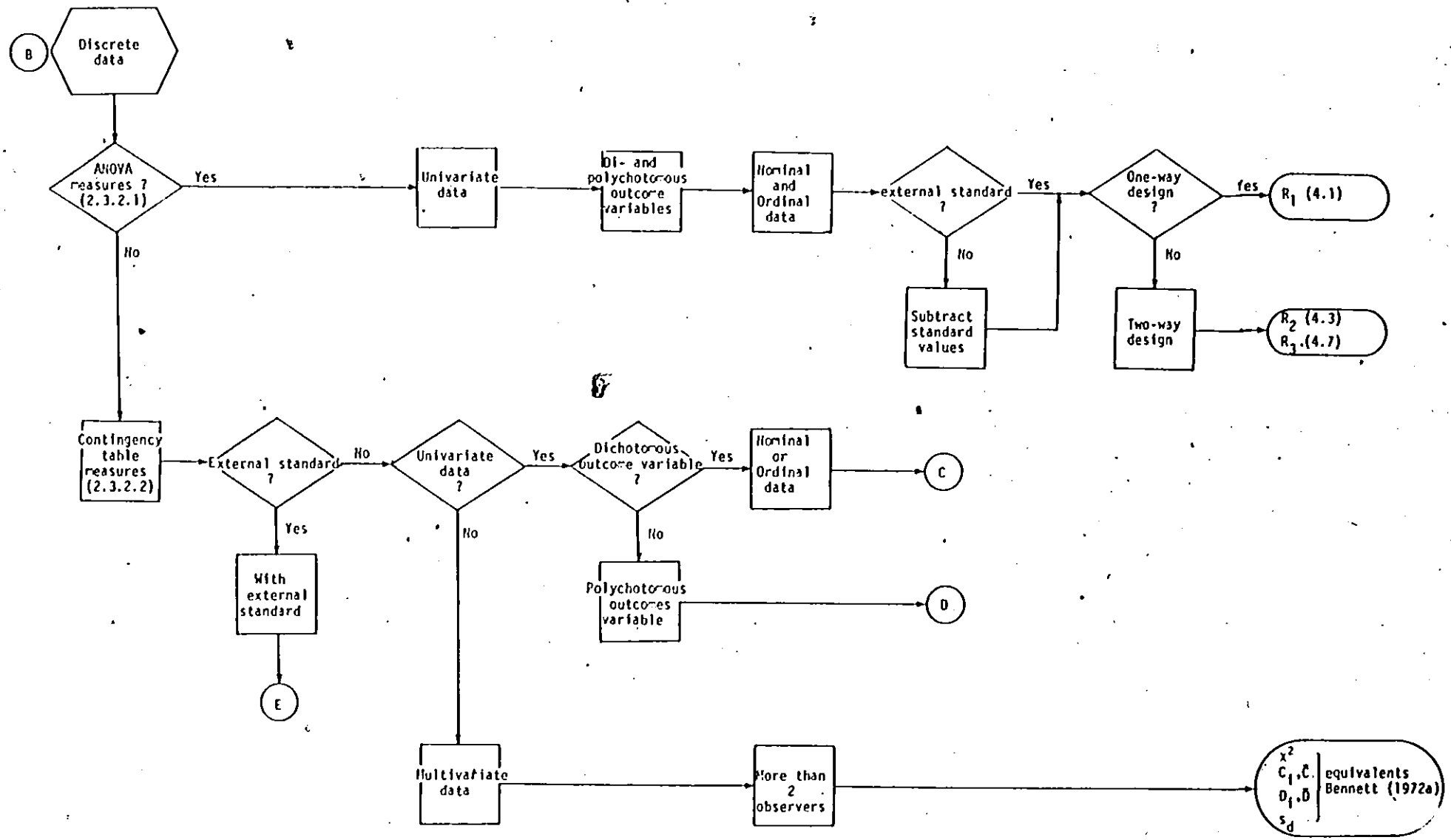
4.1 Univariate Data With No External Standard

It is possible to estimate agreement for univariate discrete data analogously to that for continuous data by using ANOVA procedures. These measures will be discussed first, followed by those measures developed from a contingency table point of view. The decision tree for choosing among these measures is shown in Diagrams 4.1, 4.2 and 4.3.

4.1.1 ANOVA Measures (Diagram 4.1)

In order to estimate the variance components in a set of discrete data, the categories have to be assigned numerical values. For nominal data meaningful numerical values can only be assigned if there are only two categories, so that the numerical values (e.g. 0,1) can serve to indicate the contrast or difference between the categories. For polychotomous nominal variables no meaningful numbers can be

DIAGRAM 4.1. Flow Chart of Agreement Measures for Discrete Data: ANOVA Measures for Univariate Data, and All Measures for Multivariate Data



assigned as, by definition, nominal categorisation conveys no information on the size of the gaps or the types of differences between adjacent categories. Ordinality implies ranking but still no information on interval size is conveyed, so that when numerical values are assigned, their meaning is questionable. For the simplest situation the proportions of subjects classified into two categories by two observers are shown in Table 4.1 and the corresponding ANOVA Table in Table 4.2 (from Landis and Koch 1975). Under the usual assumptions (3.10) as well as the assumption of no observer effect, the model is analogous to (3.13)

$$y_{ij} = \mu + s_i + e_{ij} \tag{3.13}$$

and ρ_1 estimated as in (3.15), but called R_1,

$$R_1 = \frac{MS_s - MS_e}{MS_s + (m-1)MS_e}, \quad (MS_s - MS_e \geq 0) \tag{4.1}$$

with the variance components estimated as in Table 4.2 except for the mean square for residual error which is given by

$$MS_e = \{(SS_d + SS_e)/n\} . \tag{4.2}$$

The models which do involve observer effects are analogous to the models for continuous data under the usual assumptions (3.10). Where observer effects are assumed random, the model is

Table 4.1

Observed Proportions Resulting From Two Observers Classifying n Subjects Into Two Categories

		Observer 1		
		0	1	Total
Observer 2	0	p_{11}	p_{12}	$p_{1.}$
	1	p_{21}	p_{22}	$p_{2.}$
	Total	$p_{.1}$	$p_{.2}$	1

Table 4.2

ANOVA Table for Two-way Table

Source	d.f.	Sums of Squares	Mean Squares
Observers	1	$SS_d = \frac{n}{2}[(p_{12}-p_{21})^2]$	$MS_d = SS_d/1$
Subjects	n - 1	$SS_s = \frac{n}{2}[(p_{11}+p_{22}) - (p_{11}-p_{22})^2]$	$MS_s = SS_s/(n-1)$
Residual	n - 1	$SS_e = \frac{n}{2}[(p_{12}+p_{21}) - (p_{12}-p_{21})^2]$	$MS_e = SS_e/(n-1)$
Total	2n - 1	$\frac{n}{2}[1-(p_{11}-p_{22})^2]$	

$$y_{ij} = \mu + s_i + d_j + e_{ij} \tag{3.3}$$

and the measure analogous to ρ_2 (3.18) is

$$R_2 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2 + \sigma_e^2} \tag{4.3}$$

with the intersubject variance estimate

$$\sigma_s^2 = (MS_s - MS_e)/m, \ (MS_s - MS_e \geq 0), \tag{4.4}$$

the interobserver variance estimate

$$\sigma_d^2 = (MS_d - MS_e)/n, \ (MS_d - MS_e \geq 0) \tag{4.5}$$

and the random variance estimate

$$\sigma_e^2 = MS_e \tag{4.6}$$

Where observer effects are regarded as fixed, the model

$$y_{ij} = \mu + s_i + \delta_j + e_{ij} \tag{3.22}$$

is fitted under the usual assumptions (3.10) and the constraint $\sum_{j=1}^{m} \delta_j = 0$ with an analogue for ρ_3 (3.24) estimated by

$$R_3 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \tag{4.7}$$

with the variance components estimated as in (4.4) and (4.6) but pertaining to model (3.22). In all these models m = 2 observers and as a result R_3 can also be estimated by

$$R_3 = \frac{2(p_{11}p_{22}-p_{12}p_{21})}{p_{1.}p_{2.} + p_{.1}p_{.2}} \qquad (4.8)$$

which is identical to the contingency type agreement statistic called r_{11} by Maxwell and Pilliner (1968). R_1, R_2 and R_3 have analogues among the contingency type agreement statistics as indicated in Table 4.3.

Example 4.1

The data from Table 1.2 as rearranged in Table 4.4 and the formulae from Table 4.2 yielded an ANOVA Table as in Table 4.5. Table 4.6 shows an alternative ANOVA Table for the same data, with estimators computed as in Table 3.6, a numerical value of 1 having been assigned to the presence of varices and 0 to the absence of varices. There are small differences between the tables and in the subsequent estimations of the intraclass correlation coefficients the data from Table 4.5 will be used.

No Observer Effect Assumed

The model considered here is as in (3.13),

$$y_{ij} = \mu + s_i + e_{ij} \qquad (3.13)$$

Table 4.3

Analogous Agreement Measures*

	Continuous Data	Categorical Data	
Assumptions	ANOVA Measures	ANOVA Measures	Contingency Measures
No Observer Effects	ρ_1, R^2(3.14,3.16)	R_1 (4.1)	$\hat{\pi}$ (4.29)
Fixed Observer Effects	ρ_3 (3.23)	R_3 (4.8)	r_{11} (4.8)
Random Observer Effects	ρ_2 (3.18)	R_2 (4.3)	$\hat{\kappa}$ (4.30)

* Adapted from Landis and Koch 1975

Table 4.4

Data Resulting From the Classification of 39 Patients by Two Endoscopists According to the Presence (1) or Absence (0) of Oesophageal Varices

		Endoscopist B		
		0	1	Total
Endoscopist A	0	15 (0.38)	5 (0.13)	20 (0.51)
	1	8 (0.21)	11 (0.28)	19 (0.49)
	Total	23 (0.59)	16 (0.41)	39 (1.00)

The figures in brackets indicate proportions (Conn, Smith and Brodoff, 1965).

Table 4.5

ANOVA Table for Data on the Presence of Oesophageal Varices (estimated as in Table 4.2)

Source	d.f.	Sums of Squares	Mean Squares
Patients	38	12.6750	0.3336
Observers	1	0.1248	0.1248
Random Error	38	6.5052	0.1712
Total	77	19.3050	

Table 4.6

ANOVA Table for Oesophageal Varices Data, Estimated As in Table 3.6, with 0 = Absence and 1 = Presence of Oesophageal Varices

Source	d.f.	Sums of Squares	Mean Squares
Patients	38	12.7949	0.3367
Observers		0.1154	0.1154
Random Error	38	6.3846	0.1680
Total	77	19.2949	

where y_{ij} is the classification of the i-th patient by the j-th observer,

μ is the overall mean,

s_i is the i-th subject (patient) effect,

e_{ij} is the random error associated with that observation.

Variance Components

MS_s as in Table 4.5,

$MS_e = (SS_d+SS_e)/n = (0.1248+6.5052)/39 = 0.1700$

Patient Variation

$$R_1 = \frac{MS_s - MS_e}{MS_s + (m-1)MS_e} = \frac{0.3336 - 0.1700}{0.3336 + (1)(0.1700)} = 0.3249, \text{ i.e.}$$

32.49% of the total variation is due to differences among patients.

Both Patient and Observer Effects in Model

The model under consideration is

$$y_{ij} = \mu + s_i + d_j + e_{ij} \qquad (3.3)$$

where y_{ij}, μ, s_i, e_{ij} are as in (3.13), d_j is the j-th observer effect, where both patient and observer effects are random. Where observer effects are regarded as fixed, d_j in (3.3) is replaced by δ_j such that $\sum_{j=1}^{2} \delta_j = 0$.

The components of variation are:

$\sigma_s^2 = (MS_s-MS_e)/m = (0.3336-0.1712)/2 = 0.0812$

$\theta_d^2 = (MS_d - MS_e)/n = (0.1248 - 0.1712)/39 = -0.0012 \approx 0$ or 0.0012

$\theta_e^2 = MS_e = 0.1712.$

Random Effects Model

Patient variation:

$$R_2 = \frac{\theta_s^2}{\theta_s^2 + \theta_d^2 + \theta_e^2} = \frac{0.0812}{0.0812 + 0 + 0.1712} = 0.3217 \quad (\theta_d^2 = 0)$$

or

$$R_2 = \frac{0.0812}{0.0812 + 0.0012 + 0.1712} = 0.3202 \quad (\theta_d^2 = 0.0012)$$

i.e., about 32% of the total variation is due to differences among patients.

Observer Variation

$$R_2 = \frac{\theta_d^2}{\theta_d^2 + \theta_s^2 + \theta_e^2} = 0 \quad (\theta_d^2 = 0)$$

or

$$R_2 = \frac{0.0012}{0.0012 + 0.0812 + 0.1712} = 0.0047 \quad (\theta_d^2 = 0.0012),$$

i.e., virtually none of the variation is due to disagreement between the observers.

Mixed Effects Model

With observer effects fixed,

$$R_3 = \frac{\sigma_s^2}{\sigma_\epsilon^2 + \sigma_e^2} = \frac{0.0812}{0.0812 + 0.1712} = 0.3214$$

Another estimate is r_{11}

$$= \frac{2(p_{11}p_{22}-p_{12}p_{21})}{p_{1.}p_{2.} + p_{.1}p_{.2}} = \frac{2[(0.38)(0.28)-(0.13)(0.21)]}{(0.51)(0.49) + (0.59)(0.41)} = 0.3214 ,$$

which also suggests that 32.19% of all variation is due to patient differences.

Fleiss (1966) and Landis and Koch (1975) discuss intraclass correlation coefficients R_1, R_2 and R_3 as applied to the more complicated situations of polychotomous ordinal variables and more than two observers. The principles remain as before and the interpretation of these measures for discrete data is the same as for continuous data. However, the assumptions of homogeneous variance necessary to fit the ANOVA model is not justified for binomial or multinomial proportions, nor is the normality assumption for testing hypotheses justified for discrete data. The existence of contingency table methods for measuring agreement therefore argues against the use of intraclass correlation coefficients for discrete data.

4.1.2 Contingency Table Measures

The general principle in measuring agreement in contingency tables can be illustrated by using a two-way table as in Table 4.1. Agreement was defined as diagonal association in Chapter 2, the diagonals in a two-way table consisting of either the two concordant cells ($p_{11}+p_{22}$) representing agreement, or the two discordant cells ($p_{12}+p_{21}$) representing disagreement. Agreement and disagreement are therefore complements of each other.

This rudimentary way of assessing agreement forms the basis of virtually all the measures to be discussed here, varying with the sophistication of the estimation procedure which in turn may be dictated by either the type of agreement sought by the investigator, e.g. conditional, weighted or chance corrected, or by the complexity of the design, e.g. polychotomous outcome or more than two observers (forming a multiple contingency table).

4.1.2.1 Measures of Association (Diagrams 4.2, 4.3)

The goodness-of-fit chi-square statistic is probably the most commonly used test for the significance of association, but it does not measure the degree of association which can be regarded as an agreement equivalent in fourfold tables. The numerical value of the chi-square statistic is also not entirely independent of the sample size, making meaningful comparisons across studies difficult. The odds ratio, or cross products ratio, is commonly used to measure the degree of association and is estimated by

$$O = \frac{p_{11}p_{22}}{p_{12}p_{21}} \tag{4.9}$$

applied to Table 4.1 and can be interpreted as the ratio of agreement to disagreement. The bounds of the odds ratio are zero (either p_{11} or p_{22} zero) and infinity (either p_{12} or p_{21} zero) with one indicating independence. Although invariant under row and column multiplication and interchange, the odds ratio is not symmetrical around independence, a problem that can be solved by using its logarithm. Meaningful inter-

DIAGRAM 4.2. Flow Chart of Agreement Measures for Discrete Data: Dichotomous Outcome Variables

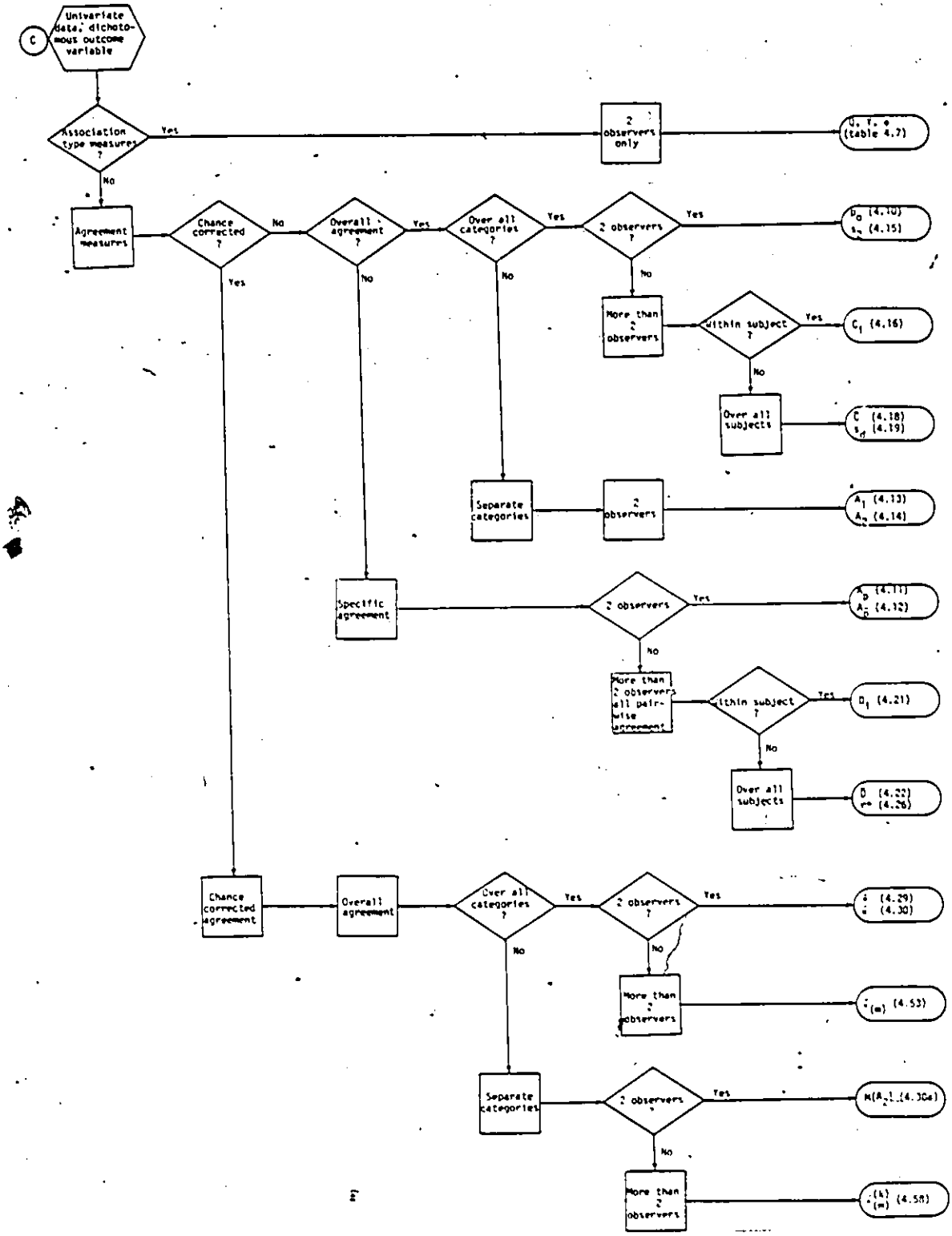
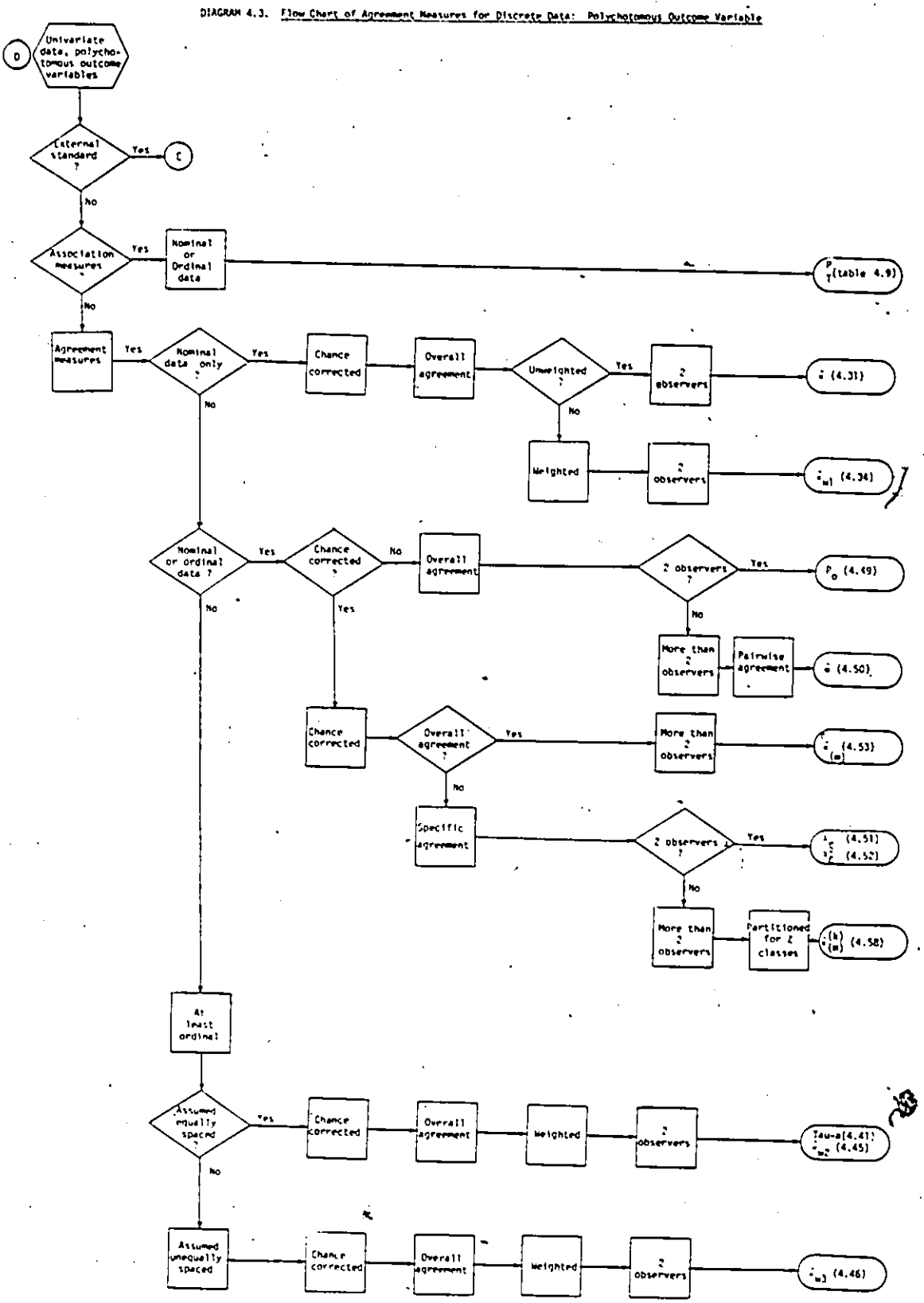


DIAGRAM 4.3. Flow Chart of Agreement Measures for Discrete Data: Polychotomous Outcome Variable



pretation is still difficult because the measurement is that of the ratio of agreement to disagreement.

Several other measures of degree of association based on either the odds ratio or the chi-square statistic are shown in Table 4.7. The first two, Yule's Q and Yule's Y (Kendall and Stuart 1961), are both symmetrical around independence, but the upper and lower bounds do not correspond with perfect agreement and disagreement respectively. For perfect agreement both p_{12} and p_{21} should be zero, but Q = Y = 1 if only one of them is zero. The same applies to disagreement.

Example 4.2

Using the data from Table 4.4 measures of association are calculated as:

$$X^2 = \frac{(|n_{12}-n_{21}|-1)^2}{n_{12}+n_{21}} = \frac{(|5-8|-1)^2}{5+8} = 0.3077$$

$$\textit{Odds ratio} = \frac{(15)(11)}{(5)(8)} = 3.9$$

The X^2 value, when referred to a table of a χ^2-distribution, is not suggestive of a significant association, which in an agreement sense may be interpreted as no significant disagreement. The odds ratio suggests that the ratio between agreement and disagreement is 3.9.

$$\textit{Yule's } Q = \frac{(.38)(0.28) - (.13)(0.21)}{(0.38)(0.28) + (0.13)(0.21)} = 0.5916$$

$$\textit{Yule's } Y = \frac{(0.38)(0.28)^{\frac{1}{2}} - (0.13)(0.21)^{\frac{1}{2}}}{(0.38)(.28)^{\frac{1}{2}} + (0.13)(0.21)^{\frac{1}{2}}} = 0.3275$$

Table 4.7

Comparison of Measures of Association
(2×2 Table)

Measure	Formulation	Bounds	Interpretation
Yule's Q	$\frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}}$	+1, if $p_{12}p_{21} = 0$, i.e. p_{12} or $p_{21} = 0$	"semi-perfect" agreement
		0, if $p_{11}p_{22} = p_{12}p_{21}$	independence
		-1, if $p_{11}p_{22} = 0$, i.e. p_{11} or $p_{22} = 0$	"semi-perfect" disagreement
Yule's Y	$\frac{(p_{11}p_{22})^{1/2} - (p_{12}p_{21})^{1/2}}{(p_{11}p_{22})^{1/2} + (p_{12}p_{21})^{1/2}}$	+1, 0, -1,	as for Yule's Q	as for Yule's Q
ϕ = Tau - b = $(\frac{\chi^2}{n})^{1/2}$	$\frac{p_{11}p_{22} - p_{12}p_{21}}{(p_{1.}p_{.1}p_{2.}p_{.2})^{1/2}}$	+1, if $p_{12} = p_{21} = 0$	perfect agreement
		0, if $p_{11}p_{22} = p_{12}p_{21}$	independence
		-1, if $p_{11} = p_{22} = 0$	perfect disagreement

$$\phi = \frac{(0.38)(0.28) - (0.13)(0.21)}{(0.51)(0.59)(0.49)(0.41)^{\frac{1}{2}}} = 0.3217 \quad .$$

The values of Yule's Q and Y are both inbetween independence and "semi-perfect" agreement and that for ϕ inbetween independence and perfect agreement.

The ϕ coefficient, which is identical to the Tau-b coefficient of concordance (Kendall 1955), is symmetric around independence and measures true perfect agreement and disagreement at the upper and lower bounds respectively, as the denominator involves marginal proportions. However, it is not invariant under row and column multiplications (Reynolds 1977) and as such is not useful for comparisons across studies. It is therefore advisable to select from among those measures designed to measure agreement specifically, as discussed in Section 4.1.2.2.

The measures of association discussed thus far apply to the situation where n subjects are classified by two observers into two categories. When more than two observers are used it becomes a multiple contingency table to be analysed either as a succession of fourfold tables or by multiple contingency table analysis in which the interpretation of agreement becomes complicated. Where there are more than two categories for classification as in Table 4.8, measures of degree of association exist which are similar to those for a fourfold table (Kendall and Stuart 1961). They are based on the chi-square statistic and are compared in Table 4.9. It is evident that for both, the lower bound indicates independence and not complete disagreement, and that

Table 4.8

Observed Proportions Resulting From Two Observers Classifying n Subjects into ℓ Categories

		Observer 2				
		1	2	.....	ℓ	Total
Observer 1	1	p_{11}	p_{12}	.....	$p_{1\ell}$	$p_{1.}$
	2	p_{21}	p_{22}	.....	$p_{2\ell}$	$p_{2.}$
	.	.	.		.	.
	.	.	.		.	.
	.	.	.		.	.
	ℓ	$p_{\ell 1}$	$p_{\ell 2}$	.....	$p_{\ell\ell}$	$p_{\ell.}$
	Total	$p_{.1}$	$p_{.2}$	.....	$p_{.\ell}$	1

Table 4.9

Comparison of Measures of Degree of Association for Polychotomous Outcome Variables

Measure	Formulation	Bounds	Interpretation
Pearson's P or C (coefficient of contingency)	$[\frac{\phi}{\phi+1}]^{1/2} = [\frac{\chi^2}{\chi^2+n}]^{1/2}$	0, if $\chi^2 = 0$	independence
		$((m-1)/m)^{1/2}$	perfect agreement
Tschuprow's T	$[\frac{\chi^2}{n(m-1)}]^{1/2} = \phi$ if m = 2	0, if $\chi^2 = 0$	independence
		+1	perfect agreement

for Pearson's P the upper bound is dependent on the number of categories. Tschuprow's T has a fixed upper bound, but as it is identical to ϕ if there are two categories, it is also not invariant to row and column marginal multiplications. As for fourfold tables, it is more appropriate to use specific agreement measures as discussed in Section 4.1.2.2.

4.1.2.2 Measures of Agreement

4.1.2.2.1 Dichotomous Outcome Variables (Diagram 4.2)

For univariate data with a dichotomous outcome variable, the table of observed proportions using two observers is shown in Table 4.1, and the data resulting from the same classification by more than two observers in Table 4.10.

The most elementary agreement index is that already mentioned in Section 4.1.2 as the concordant cells in a two-by-two table: the "crude index of agreement" described by Rogot and Goldberg (1966) which is given by

$$p_o = p_{11} + p_{22} \tag{4.10}$$

which looks at agreement in both categories equally. Its bounds are zero when both p_{11} and p_{22} are zero, and one when all the subjects fall in both concordant cells. The crude index of agreement does therefore not distinguish between the categories, but Dice (1945) defined

$$A_p = \frac{2p_{22}}{(p_{2.} + p_{.2})} \tag{4.11}$$

Table 4.10

The Data Resulting From n Subjects Being Classified Into Two on More Categories by m Observers

		Observers			
		1	2	m	Total
Subjects	1	y_{11}	y_{12}	y_{1m}	$y_{1.}$
	2	y_{21}	y_{22}	y_{2m}	$y_{2.}$
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	n	y_{n1}	y_{n2}	y_{nm}	$y_{n.}$
	Total	$y_{.1}$	$y_{.2}$	$y_{.m}$	$y_{..}$

and also its complementary analogue

$$A_{\bar{p}} = \frac{2p_{11}}{(p_{1.}+p_{.1})} \qquad (4.12)$$

for the situation where one category is of specific interest, or very much more prevalent than the other. These measures range from zero where either p_{22} or p_{11} respectively is zero, to one where all the subjects in the relevant margins are classified into that cell. Rogot and Goldberg (1966) used these conditional probability measures (4.11, 4.12) to define

$$A_1 = \frac{1}{2}(A_p+A_{\bar{p}}) \qquad (4.13)$$

as an altertative to $\bar{p}_o$. Using a different set of conditional probabilities they also defined another alternative for p_o, namely

$$A_2 = \frac{1}{4}\left[\frac{p_{11}}{p_{1.}} + \frac{p_{11}}{p_{.1}} + \frac{p_{22}}{1_{2.}} + \frac{p_{22}}{p_{.2}}\right] \qquad (4.14)$$

Both A_1 and A_2 range from zero (complete disagreement) to one (complete agreement) and are more sophisticated expressions of agreement than p_o. All these agreement measures discussed thus far have even more sophisticated chance-corrected analogues (see 4.29, 4.30, 4.31).

There is another uncorrected agreement measure for use with two observers which was defined by Fleiss (1973a) as a special case of the standard deviation agreement index (SDAI) developed by Armitage, Blendis and Smyllie (1966) for the case of multiple observers (see

(4:19) subsequently) and given by

$$S_2 = \{\frac{n}{n-1}[(p_{11}+p_{22})-(p_{11}-p_{22})^2]\}^{\frac{1}{2}} = [2(MS_s)]^{\frac{1}{2}} \qquad (4.15)$$

where MS_s is as shown in Table 4.2. S_2 is bounded by zero at its lower limit (complete disagreement) and by $\{n/(n-1)\}^{\frac{1}{2}}$ at its maximum only when $p_{11} + p_{22} = 1$ and $p_{11} = p_{22} = \frac{1}{2}$. It is therefore more difficult to interpret than p_o, A_1 or A_2.

Example 4.3

Using the data of Table 4.4 again, measures of agreement are as follows:

$$p_o = p_{11} + p_{22} = 0.38 + 0.28 = 0.66$$

$$A_p = \frac{2(0.38)}{(0.41+0.49)} = 0.6222$$

$$A_{\bar{p}} = \frac{2(0.38)}{(0.69+0.51)} = 0.6909$$

Overall agreement is therefore 66% with agreement on the presence of varices 69.09% and on the absence of varices 62.22%.

The values for other estimates of crude overall agreement are:

$$A_1 = \frac{1}{2}(0.6222+0.6909) = 0.6566$$

$$A_2 = \frac{1}{2}[\frac{0.38}{0.51} + \frac{0.38}{0.69} + \frac{0.28}{0.49} + \frac{0.28}{0.41}] = 0.6619$$

and they agree well with the value of p_o.

As a special case of the standard deviation agreement index $s_2 = 0.8168$ which is higher than the crude indices, but difficult to interpret as its upper bound is not known.

Where multiple observers are used the data are arranged as in Table 4.10 with the objective of defining an overall measure of the extent to which observers agree in classifying the subjects into the same category. By denoting the two categories by 0 and 1 respectively, the cell frequency y_{ij} represents the category into which the i-th subject was classified by the j-th observer, the marginal row total $y_{i.}$ represents the total number of observers that classified the i-th subject into the category denoted by 1, and the marginal column total $y_{.j}$ represents the total number of subjects classified into category 1 by the j-th observer.

Armitage, Blendis and Smyllie (1966) defined the majority agreement index (MAI) as

$$C_i = |2\bar{y}_{i.} - 1| \text{ for } i = 1,2, \ldots, n \text{ subjects} \qquad (4.16)$$

where $\bar{y}_{i.}$ is the mean number of classifications into category 1 for the i-th subject and

$$\bar{y}_{i.} = \frac{y_{i.}}{m} \qquad (4.17)$$

for m observers.

C_i is therefore an agreement index for all observers on a single subject at a time and ranges from zero (when the observers are evenly

divided, i.e. $\bar{y}_{i.} = \frac{1}{2}$) to one (when all the observers classify the subject into the same category). An unusual, and undesirable, property of C_i is that it can assume the value zero only if the number of observers, m, is even, otherwise its lower bound is 1/m. As a summary statistic expressing agreement over all observers and all subjects, the mean MAI was defined as

$$\bar{C} = \frac{1}{n} \sum_{i=1}^{n} C_i \tag{4.18}$$

which has the same bounds as C_i and is identical to p_o if m = 2.

Armitage, Blendis and Smyllie (1966) also defined another agreement index analogous to $\bar{C}$. Called the standard deviation agreement index (SDAI), it is based on the variation of the $\{y_{i.}\}$ from subject to subject and given by

$$S_d = \left\{ \frac{\sum_{i=1}^{n} (y_{i.} - m\bar{t})^2}{n-1} \right\}^{\frac{1}{2}} \tag{4.19}$$

where $\bar{t}$ is the overall proportion of subjects classified into the same category (1) by all the observers and is given by

$$\bar{t} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_{i.} = \frac{1}{nm} \sum_{i=1}^{n} y_{i.} \tag{4.20}$$

An increase in S_d therefore reflects greater deviations from the overall expected agreement ($m\bar{t}$) and therefore an increase in observer agreement. It has a minimum value of zero which indicates

complete disagreement except in the special case of complete agreement with all the subjects classified into the same single category. Other cases of complete agreement are not recognizable by an upper bound, so that interpretation is difficult. When $m = 2$, S_d takes the form of S_2 in (4.15).

In some situations where multiple observers perform the classification, the investigator may be interested in agreement between pairs of observers rather than among all observers. Armitage, Blendis and Smyllie (1966) defined the pair disagreement index (PDI) by

$$D_i = \frac{y_{i.}(m-y_{i.})}{\binom{m}{2}} \quad \text{for } i = 1,2, \ldots, n \text{ subjects} \qquad (4.21)$$

which is based on the fact that there are $\binom{m}{2}$ possible pairs of observers, of which $y_{i.}(m-y_{i.})$ disagree on the classification of the i-th subject. D_i has zero as its lower limit (if all observers agree on the classification) and $\frac{1}{2}m/(m-1)$ as its upper limit (if the observers are evenly divided). D_i is therefore analogous to C_i in that the agreement refers to a single subject. Due to the dependence of the upper limit on the number of observers, interpretation is not easy. A summary statistic for agreement over all subjects (like $\bar{C}$), called the mean pair disagreement index, was also defined as

$$\bar{D} = \frac{1}{n} \sum_{i=1}^{n} D_i \qquad (4.22)$$

with the same bounds as D_i and which, if $m = 2$, is identical to the

crude index of disagreement $(1-p_o)$, where p_o is defined as in (4.10).

None of the agreement indices mentioned thus far has had its distributional properties studied and the necessary theory developed for hypothesis testing. Fleiss (1965) suggested that Cochran's Q statistic be used to test the null hypothesis of no interobserver bias or marginal homogeneity among the proportions classified into the same category by each observer. In Table 4.10, if the expected values of the $\bar{y}_{.j}$ are denoted by π_j for $j = 1,2, \ldots, m$ observers, the null hypothesis is

$$H_o: \ \pi_1 = \pi_2 = \ldots = \pi_m \tag{4.23}$$

and

$$Q = \frac{2 \sum_{j=1}^{m} (y_{ij} - \bar{y}_{..})^2}{n\bar{D}} \quad \text{for } \bar{D} > 0 \tag{4.24}$$

and

$$\bar{y}_{..} = \frac{1}{m} \sum_{j=1}^{m} y_{.j} \tag{4.25}$$

where $\bar{D}$ is the mean pair disagreement index as in (4.22). Q has an approximate χ^2 distribution with (m-1) degrees of freedom under the null hypothesis of marginal homogeneity, it increases as $\bar{D}$ decreases and it is affected by both agreement and interobserver bias. Fleiss (1965) also showed that the common intraclass correlation between all pairs of classification on the same subjects can be estimated by

$$r^* = 1 - (B/Q) \tag{4.26}$$

where

$$B = \frac{\sum_{j=1}^{m} (y_{.j} - \bar{y}_{..})^2}{\bar{y}_{..} - (y_{..}^2/n)} \quad \text{and} \qquad (4.27)$$

which simplifies to $\hat{\pi}$ in (4.29) when m = 2. In addition to the Q test for marginal homogeneity, Bennett (1967, 1968) described two goodness-of-fit statistics which have both approximately χ^2 distributions in large samples with (m-1) degrees of freedom.

Example 4.1

Continuing with the data in the paper by Conn, Smith and Brodoff (1965) a radiologist (third observer) reported on the presence or absence of varices in the same 50 patients as seen on radiographic examination. With the data arranged as in Table 4.11 several agreement measures may be calculated:

For each patient a mean majority agreement index may be calculated, the mean of which is given by $\bar{C} = 0.68$. Similarly, for each patient a pairwise disagreement index may be estimated, the mean of which is $\bar{D} = 0.32$. The interpretation of $\bar{C}$ is similar to the crude indices of agreement, and $\bar{D}$ indicates the overall pairwise disagreement (they appear complementary to each other).

The standard deviation agreement index in this situation is $S_d = 1.1175$, which is difficult to interpret as it has no upper bound that is fixed.

To test for marginal homogeneity, Q = 2.03, which, when referred

Table 4.11

Data Resulting From the Classification of 39 Patients by 3 Observers According to the Presence (1) or Absence (0) of Oesophageal Varices

Patient	Endoscopist A	Endoscopist B	Radiologist	Total
1	0	0	0	0
2	0	0	0	0
3	1	1	1	2
4	1	1	1	3
5	0	0	1	1
6	1	1	1	3
7	1	0	0	1
8	0	0	0	0
9	1	0	0	1
10	0	1	1	2
11	1	0	0	1
12	0	0	0	0
13	1	1	1	3
14	0	0	0	0
15	0	0	0	0
16	1	1	1	3
17	1	1	1	3
18	1	0	0	1
19	1	0	0	1
20	1	1	1	3
21	0	0	0	0
22	0	0	0	0
23	1	0	0	1
24	0	1	1	2
25	0	1	0	1
26	1	0	0	1
27	1	0	1	2
28	0	0	0	0
29	0	0	0	0
30	0	1	0	1
31	0	0	0	0
32	1	1	0	2
33	1	1	1	3
34	0	0	1	1
35	1	1	1	3
36	0	0	1	1
37	0	1	0	1
38	0	0	0	0
39	1	1	0	2
Total	19	16	14	49

to χ^2 tables with 2 degrees of freedom does not lead to rejection of the null hypothesis. In terms of agreement that suggests no evidence of significant disagreement. The common intraclass correlation coefficient, $r^* = 0.34$, but the interpretation of that number is not clear in this situation.

With the exception of S_d and Cochran's Q the agreement measures discussed thus far in Section 4.1.2, do not take into account any agreement that can be expected to occur under certain baseline constraints involving the marginal proportions. However there are chance corrected measures of agreement which are analogues of p_o, A_p, A_1, and A_2. They can be derived from the general form as described by Fleiss (1973b):

$$M(I) = \frac{I_o - I_e}{1 - I_e} \qquad (4.28)$$

where I_o denotes the observed value of the agreement index and I_e the value of the same index expected under the assumption of independence of the marginal proportions. Therefore $(I_o - I_e)$ represents the excess agreement beyond chance and $(1 - I_e)$ the maximum possible excess agreement beyond chance. In this context M(I) has as lower bound the quantity $-I_e/(1-I_e)$, as upper bound one and as indicator of independence zero. These properties can be interpreted as follows:

+1, if the observers are in complete agreement,

0, if the level of agreement is equal to the agreement that can be expected to occur due to chance and

-1, if $I_e = \frac{1}{2}$.

The lower bound is therefore difficult to interpret, but it is seldom of interest to the investigator, who is usually interested in the performance of the observers beyond the level of chance agreement, as observers are not really of use if all they can achieve is the level of chance agreement. The distributional characteristics for a number of these chance-corrected indices have been described and the theory formulated for hypothesis testing about the difference from chance and complete agreement.

The way I_o and I_e are estimated for the analogues of p_o, A_p, A_1 is shown in Table 4.12. Scott (1955) derived a chance corrected analogue for p_o under the constraints of marginal homogeneity and independence, and it can be estimated by

$$\hat{\pi} = \frac{4(p_{11}p_{22}-p_{12}p_{21}) - (p_{12}-p_{21})^2}{(p_{1.}+p_{.1})(p_{2.}+p_{.2})} \quad . \tag{4.29}$$

Cohen (1960) developed a chance corrected analogue to p_o, A_p and A_1, under the single constraint of independence and it can be estimated by

$$\hat{\kappa} = \frac{2(p_{11}p_{22}-p_{12}p_{21})}{p_{1.}p_{.2} + p_{.1}p_{2.}} \quad . \tag{4.30}$$

The chance corrected analogue of A_2 is not equivalent to any of the others and is estimated by

$$\hat{M}(A_2) = (p_{11}p_{22}-p_{12}p_{21})\left(\frac{1}{2p_{1.}p_{2.}} + \frac{1}{2p_{.1}p_{.2}}\right) \quad . \tag{4.30a}$$

Table 4.12

Chance Corrected Agreement Statistics for Dichotomous Variables

I_o	I_e	$\hat{M}(I)$
p_o	$\frac{1}{2}\{(p_{1.}+p_{.1})^2 + (p_{2.}+p_{.2})^2\}$	$\hat{\pi}$ (4.29)
p_o	$p_{1.}p_{.1} + p_{2.}p_{.2}$	$\hat{\kappa}$ (4.30)
A_p	$\frac{2p_{2.}p_{.2}}{p_{2.} + p_{.2}}$	$\hat{\kappa}$
A_1	$\frac{p_{1.}p_{.1}}{p_{1.} + p_{.1}} + \frac{p_{2.}p_{.2}}{p_{2.} + p_{.2}}$	$\hat{\kappa}$

It is interesting to note that they all contain the contrast $(p_{11}p_{22}-p_{12}p_{21})$ in their numerators, but that their denominators differ. Scott's $\hat{\pi}$ tends to be used when the investigator is specifically interested in one category at the expense of the other and when the marginal distributions tend towards homogeneity. Cohen's $\hat{\kappa}$ ranges from $\{-2p_{1.}p_{2.}/(p_{1.}^2+p_{2.}^2)\}$ as lower bound through zero for independence to one for complete agreement and $M(A_2)$ from -1 for complete disagreement to +1 for complete agreement. Where multiple observers are used, an extension of $\hat{\kappa}$ (Fleiss 1971) can be used and it, $\hat{\kappa}_{(m)}$, will be fully discussed under polychotomous outcomes, as are tests of hypotheses about kappa.

Example 4.5

Returning to the data as displayed in Table 4.4, the following chance corrected agreement measures may be calculated:

$$\hat{\pi} = \frac{4[(0.38)(0.28)-(0.13)(0.21)] - (0.13-0.21)^2}{(0.51+0.59)(0.49+0.41)} = 0.31$$

$$\hat{\kappa} = \frac{2[(0.38)(0.28)-(0.13)(0.21)]}{(0.51)(0.41) + (0.59)(0.49)} = 0.32$$

$$(Or,\ p_o = 0.66,\ p_e = (0.3024+0.1999) = 0.50,\ \hat{\kappa} = \frac{0.66-0.50}{1-0.50} = 0.32)$$

$$M(A_2) = \{(0.38)(0.28)-(0.13)(0.21)\}\{\frac{1}{2(0.51)(0.49)} + \frac{1}{2(0.59)(0.41)}\} = 0.32$$

The proportion of agreement achieved over and above that of chance (which was 0.5), is estimated as 0.32 by all these measures.

4.1.2.2.2 Polychotomous Outcome Variables (Diagram 4.3)

The data resulting from two observers classifying n subjects into ℓ categories are displayed in Table 4.8. Measures exist for use with either nominal or ordinal data, nominal data only and ordinal data only. These three groups of measures will be discussed separately and within each group measures that take chance agreement into account will be discussed in addition to those that do not should they exist.

4.1.2.2.2.1 Nominal Data

Cohen (1960, 1968) developed two measures of chance corrected agreement specifically for polychotomous nominal data. The first is a measure of overall agreement which is an extension of $\hat{\kappa}$ (4.30) for dichotomous variables and it is estimated by

$$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e} \tag{4.31}$$

where p_o is the observed proportion of agreement given by

$$p_o = \sum_{k=1}^{\ell} p_{kk} \tag{4.32}$$

and p_e is the expected proportion of agreement under the contraint of independence and given by

$$p_e = \sum_{k=1}^{\ell} p_{k.} p_{.k} \tag{4.33}$$

As defined here, $\hat{\kappa}$ has the same properties as M(I) in (4.28) and simplifies to $\hat{\kappa}$ in (4.30) when $\ell = 2$.

The second measure allows for disagreement to be scaled, by

allocating weights to the patterns of disagreement according to their importance as assessed by the investigator. Cohen (1968) defined a weighted kappa statistic as

$$\hat{\kappa}_{wl} = \frac{p_o^* - p_e^*}{1 - p_e^*} \tag{4.34}$$

where p_o^* is the weighted observed proportion of agreement given by

$$p_o^* = \sum_{k=1}^{\ell} \sum_{k'=1}^{\ell} w_{kk'} p_{kk'} \tag{4.35}$$

and p_e^* is the weighted expected proportion of agreement under the constraint of complete independence given by

$$p_e^* = \sum_{k=1}^{\ell} \sum_{k'=1}^{\ell} w_{kk'} p_{k.} p_{k'.} \tag{4.36}$$

where $\{w_{kk'}\}$ are a set of weights for $k,k' = 1,2, \ldots, \ell$ as stipulated by the investigator reflecting the contribution of each cell, as in Table 2.8, to the measure of agreement, and $0 \leq w_{kk'} \leq 1$ for all k,k'. Everitt (1968) and Cohen (1960, 1968) derived the distributional properties of $\hat{\kappa}$ and $\hat{\kappa}_{wl}$ under the assumptions of independence and fixed margins. Subsequently Fleiss, Cohen and Everitt (1969) described the distributional properties of $\hat{\kappa}$ and $\hat{\kappa}_{wl}$ under the single constraint of a fixed number of subjects so that hypotheses about independence and complete agreement can be tested using the unconditional large sample variance of weighted kappa estimated by

$$\text{var}(\hat{\kappa}_w) = \frac{1}{n(1-p_e^*)^4}\{\sum_{k=1}^{\ell}\sum_{k'=1}^{\ell} p_{kk'}[w_{kk'}(1-p_e^*)-(\bar{w}_{k.}+\bar{w}_{.k'})(1-p_o^*)]^2 - (p_o^*p_e^*-2p_e^*+p_o^*)^2\} \tag{4.37}$$

with

$$\bar{w}_{k.} = \sum_{k'=1}^{\ell} w_{kk'}p_{.k'} \tag{4.38}$$

and

$$\bar{w}_{.k'} = \sum_{k=1}^{\ell} w_{kk'}p_{k.} \tag{4.39}$$

It is important to note that choosing

$$w_{kk'} = \begin{cases} 1, & k = k' \\ 0, & k \neq k' \end{cases} \tag{4.40}$$

$\hat{\kappa}_{w1}$ in (4.34) simplifies to $\hat{\kappa}$ in (4.31) which for m = 2 simplifies in turn to $\hat{\kappa}$ in (4.30) so that hypotheses about all these measures can be tested. Furthermore three other kappa variations still to be discussed, $\hat{\kappa}_{(m)}$ (4.53), $\hat{\kappa}_{w2}$ (4.45) and $\hat{\kappa}_{w3}$ (4.66) can be tested in the same way. The test statistic is $Z_o(\hat{\kappa}_w)^* \approx N(0,1)$ under independence hypothesis or $Z_1(\hat{\kappa}_w) \approx N(1,1)$ under the null hypothesis of complete agreement. Hypothesis testing can be done as long as the sample size, n, is equal to or greater than twice the square of the number of categories in the outcome

---

* $Z_o(\hat{\kappa}_w) = \dfrac{\hat{\kappa}_w - 0}{\sqrt{\text{var}\ \hat{\kappa}_w}}$, $Z_1(\hat{\kappa}_w) = \dfrac{\hat{\kappa}_w - 1}{\sqrt{\text{var}\ \hat{\kappa}_w}}$.

variable (Fleiss and Cicchetti 1975, see Table 4.13).

Example 4.6

Continuing with the data from Conn, Smith and Brodoff (1965), the endoscopists also noted the presence or absence of prominent mucosal folds at the gastro-oesophageal junction. By reclassifying their data as in Table 4.14, the following chance corrected measures of agreement for polychotomous nominal data may be calculated

$$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e} = \frac{0.39 - 0.30}{1 - 0.30} = 0.13$$

Although the crude index of agreement suggests 39% agreement between the two endoscopists, the agreement over and above that which can be expected to occur by chance is 13%. In this estimation, all disagreements and agreements were considered equally important. However, allocating weights to the various cells as in Table 4.15 (where agreement is weighted as one, and most extreme disagreement as zero):

$\hat{\kappa}_{w_1} = \frac{p_o^* - p_e^*}{1 - p_e^*} = \frac{0.5970 - 0.4991}{1 - 0.4991} = 0.1954$. *The weights were allocated according to the importance of the misclassification to this author, and the higher agreement index suggests that the endoscopists' disagreements occurred in cells of lesser importance.*

To illustrate hypothesis testing, the null hypothesis about unweighted $\hat{\kappa}$ *is*

$$Ho: \kappa = 0 \text{ (independence)}$$

Table 4.13

Minimum Sample Sizes Necessary for Reliable Tests of Significance About Kappa in an $\ell \times \ell$ Table for Selected Values of ℓ

# Categories (ℓ)	Sample Size (n)*
2	8
3	18
4	32
5	50

* $n = 2\ell^2$ (Fleiss and Cicchetti 1975)

Table 4.14

Table Resulting From the Classification of Patients According to the Presence or Absence of Oesophagel Varices and/or Prominent Mucosal Folds

		Endoscopist B				
		Both	Varices	Mucosal folds	Neither	Total
Endoscopist A	Both	1 (0.03)	1 (0.03)	1 (0.03)	0 (0.00)	3 (0.08)
	Varices	1 (0.03)	7 (0.18)	7 (0.18)	1 (0.03)	16 (0.41)
	Mucosal folds	1 (0.03)	4 (0.03)	2 (0.05)	5 (0.13)	12 (0.31)
	Neither	0 (0.00)	0 (0.00)	3 (0.08)	5 (0.13)	8 (0.21)
	Total	3 (0.08)	12 (0.31)	13 (0.33)	11 (0.28)	39 (1.00)

Table 4.15

Table of Weights Used in Calculation of $\hat{\kappa}_{wl}$ in Example 4.6

		Endoscopist B			
		Both	Varices	Mucosal folds	Neither
Endoscopist A	Both	1.00	1.00	0.00	0.00
	Varices	1.00	1.00	-0.50	0.00
	Muscosal Folds	0.00	0.50	1.00	0.20
	Neither	0.00	0.00	0.20	1.00

with the alternative hypothesis

$$H_1: \kappa > 0 .$$

The test statistic is

$$Z_0(\hat{\kappa}) = \frac{\hat{\kappa} - 0}{\sqrt{var\ (\hat{\kappa})}}$$

and var $(\hat{\kappa}) = 0.0180$ *(from formulation in (83), where weights are taken as*

$$w_{11} = w_{22} = w_{33} = w_{44} = 1$$

$$w_{12} = w_{21} = w_{13} = w_{31} = w_{14} = w_{41} = w_{23} = w_{32} = w_{24} = w_{41} = w_{34} = w_{43} = 0,)$$

$Z_0(\hat{\kappa}) = \frac{0.13 - 0}{\sqrt{0.0180}} = 0.9690$. *Referring* $Z = 0.9690$ *to the one-tailed standardized normal distribution (Colton 1974),* $0.166 < p < 0.169$, *so that* H_0 *can only be rejected at a 16.9% level, and there is insufficient evidence to say that agreement was not due to chance. On the other hand, were* H_0: $\kappa = 1$ *(perfect agreement) and* H_1: $\kappa < 1$, $Z_1(\hat{\kappa}) = \frac{0.13-1}{\sqrt{0.0180}} = -6.48$, $p << 0.001$, *suggesting that the agreement is significantly different from perfect agreement.*

The same procedure may be followed for $\hat{\kappa}_{w1}$ *using weights as in Table 2.28 in the estimation of variance of* $\hat{\kappa}_{w1}$.

4.1.2.2.2.2 Ordinal Data

Kendall (1955) proposed a measure estimated by

$$\text{Tau-a} = M(S) = \frac{S - 0}{\max(S) - 0} \tag{4.41}$$

with S a measure of disarray such that

$$S = \sum_{\substack{i=1 \\ i>j}}^{\ell} \sum_{j=1}^{\ell} \lambda_{ij} \tag{4.42}$$

and $\lambda_{ij} = 1$, if $r_i > r_j$

0, if $r_i = r_j$ (4.43)

-1, if $r_i < r_j$

where r_i is the i-th ranking of an observer and r_j is the j-th ranking of the same subject by another observer and

$$\max(S) = \binom{\ell}{2} = \tfrac{1}{2}\ell(\ell-1) \tag{4.44}$$

Tau-a does not use the distance between rankings, but two other kappa measures do, either by assuming knowledge about the distances between ordinal ranks, or for interval data. The first, $\hat{\kappa}_{w2}$, is used where the ℓ categories can be assumed to be equally spaced so that discrete numerical integers such as $1,2, \ldots, \ell$ can be assigned to the categories. If the weights are chosen such that

$$w_{kk'} = 1 - (k-k')^2 \tag{4.45}$$

and $\hat{\kappa}_{w2}$, calculated by substituting these weights into $\hat{\kappa}_{w1}$ (4.34), $\hat{\kappa}_{w2}$, under observed marginal symmetry, is equal to the product-moment correlation coefficient calculated on the integer-valued categories

(Cohen 1968). Furthermore, if the random effects model is assumed, the estimate of the intraclass correlation coefficient ρ_2 in (3.18) is asymptotically equal to $\hat{\kappa}_{w2}$.

The second kappa-type measure using distances between rankings is $\hat{\kappa}_{w3}$, which is $\hat{\kappa}_{w1}$ (4.34) with weights selected as recommended by Cicchetti (1972) and Cicchetti and Allison (1973)

$$w_{kk'} = 1 - \frac{|k-k'|}{(\ell-1)} \tag{4.46}$$

which can be used if the ordinal classes can be assumed not to be equally spaced, because the weights are based on the distances between categories. In addition to the test statistic $Z_0(\hat{\kappa}_w)$ referred to before, Cicchetti also developed another

$$Z_c = \frac{p_o^* - p_e^*}{(\text{var}(p_o^*))^{\frac{1}{2}}} \tag{4.47}$$

with

$$\text{var}(p_o^*)) = \frac{1}{n-1}\left\{\sum_{k=1}^{\ell}\sum_{k'=1}^{\ell} w_{kk'}^2 p_{kk'} - p_o^{*2}\right\} \tag{4.48}$$

with $Z_c \sim N(0,1)$ under independence. Fleiss and Cicchetti (1975) compared Z_c with $Z_0(\hat{\kappa}_w)$ and recommended that the latter be used.

Example 4.7

To illustrate the use of agreement measures in ordinal data sets, data from a study in which observers rated the physical function

of patients on a three level scale, where 1 = major impairment, 2 = partial impairment and 3 = minimal impairment. The data are displayed in Table 4.16 and the measures of agreement are calculated as follows:

$$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e} = \frac{0.60 - 0.40}{1 - 0.40} = 0.33 \quad .$$

Using an unweighted kappa measure, the observers showed 33% agreement on the scores beyond chance. Assuming that these ranks are equally spaced, weights may be allocated according to the formula $w_{kk'} = 1 - (k-k')^2$, so that

$$w_{11} = w_{22} = w_{33} = 1, \qquad 1$$
$$w_{12} = w_{21} = w_{23} = w_{32} = 0, \quad \text{standardized:} \quad 0.75$$
$$w_{13} = w_{31} = -3. \qquad 0$$

Then, $\hat{\kappa}_{ws} = \frac{p_o^* - p_e^*}{1 - p_e^*} = \frac{0.64 - 0.10}{1 - 0.10} = 0.48$ which is higher than $\hat{\kappa}$. If for the purpose of illustration, the classes are assumed to be unequally spaced such that the interval between classes one and two is one unit wide, and that between classes 2 and 3 three units wide, the following weights may be used ($w_{kk'} = 1 - \frac{|k-k'|}{L-1}$):

$$w_{11} = w_{22} = w_{33} = 1,$$
$$w_{12} = w_{21} = 0.5,$$
$$w_{13} = w_{31} = -1$$
$$w_{23} = w_{32} = -0.5 \qquad \text{and}$$

Table 4.16

Table Resulting From Patients Being Classified As Having Major Impairment (1), Partial Impairment (2) or Minimal Impairment (3) of Physical Function by Two Observers

		Observer B			
		1	2	3	Total
Observer A	1	10 (0.21)	3 (0.06)	0 (0.00)	13 (0.28)
	2	4 (0.09)	16 (0.34)	6 (0.13)	26 (0.55)
	3	1 (0.02)	5 (0.11)	2 (0.04)	8 (0.17)
	Total	15 (0.32)	24 (0.51)	8 (0.17)	47 (1.00)

Figures in brackets are proportions. (Data are a personal communication from F. Fortin, Faculty of Nursing, University of Montreal and C.H. Goldsmith, Department of Clinical Epidemiology and Biostatistics, McMaster University.)

$$\hat{\kappa}_{\omega 3} = \frac{P_O^* - P_e^*}{1 - P_e^*} = \frac{0.54 - 0.37}{1 - 0.37} = 0.27 .$$

Hypotheses may be specified and tested by the same procedure demonstrated in Example 4.6.

4.1.2.2.2.3 Nominal Or Ordinal Data

The simplest measure of overall agreement is not chance corrected and is an extension of p_o (4.10) which describes the proportion of subjects classified identically by two observers and is estimated by

$$\hat{p}_o = \sum_{k=1}^{\ell} p_{kk} \quad . \tag{4.49}$$

For more than two observers Cartwright (1956) described a measure of pairwise agreement over all observers estimated by

$$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} d_i / \binom{m}{2} \tag{4.50}$$

where d_i is the number of pairs of observers in agreement on the classification of the i-th subject. The coefficient, $\hat{\alpha}$, is therefore an uncorrected index which does not distinguish among agreement in the various categories and which is a complementary analogue of $\bar{D}$ (4.22) for dichotomous variables and identical to p_o (4.10) for two observers. It has a lower bound of zero (no agreement) and an upper bound of one (perfect agreement) as the $\{d_i\}$ can range between zero and $\binom{m}{2}$ for each subject. Its estimation is based on "equiprobable categories" which

restricts its use.

There are several chance corrected measures for use with either nominal or ordinal data which consider the cells on the main diagonal of a square table for two observers. Goodman and Kruskal (1954) proposed a measure based on optimal prediction estimated by

$$\lambda_r = \frac{\sum_{k=1}^{\ell} p_{kk} - \frac{1}{2}(p_{M.}+p_{.M})}{1 - \frac{1}{2}(p_{M.}+p_{.M})} \qquad (4.51)$$

where $p_{M.}$ and $p_{.M}$ are the marginal proportions corresponding to a hypothesized modal class. In this situation, λ_r has a lower bound of -1 (when all the diagonal cells contain no subjects) and an upper bound of +1 (when both observers are in complete agreement and the diagonal cells contain all the subjects). Other measures are based on χ^2 statistics, but consider only the cells on the diagonal. Light (1971) described one such index which reflects deviations from the expected pattern under independence on the diagonal and is estimated by

$$x^2_{\ell} = n\{\sum_{k=1}^{\ell} \frac{(p_{kk}-p_{k.}p_{.k})^2}{p_{k.}p_{.k}} + [\sum_{k=k'}^{\ell}\sum^{\ell} p_{k.}p_{.k'}]^{-1} [\sum_{k=k'}^{\ell}\sum^{\ell} (p_{kk'}-p_{k.}p_{.k'})]^2\} \qquad (4.52)$$

which is asymptotically chi-square with 1 degree of freedom under the hypothesis of independence. Mantel and Crittenden (in Chen, Crittenden, Mantel and Cameron 1961) proposed a chi-square statistic with one degree of freedom as a test of agreement on the main diagonal cells and Spiers and Quade (1970) use the method of minimum χ^2 to estimate expected values

for the (k,k')-th cells as the weighted averages of the expected values under independence and the expected values with the diagonals inflated to the greatest possible extent for a test of independence to be performed. These χ^2 based measures afford one the opportunity of testing hypotheses about them, but as for λ_r these measures concern agreement on the main diagonal, treating all disagreement cells equally and total disagreement as the complement of agreement. There are some kappa measures that can differentiate between degrees of disagreement by using weights: they were discussed earlier (4.34, 4.45, 4.46).

For a chance corrected measure of overall agreement among several observers, Fleiss (1971) proposed an extension of kappa which can be estimated by

$$\hat{\kappa}_{(m)} = \frac{\bar{p}_o - \bar{p}_e}{1 - \bar{p}_e} \tag{4.53}$$

where $\bar{p}_o$ is the overall observed proportion of agreement given by

$$\bar{p}_o = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{\ell} \binom{u_{ik}}{2} / \binom{m}{2} , \tag{4.54}$$

where u_{ik} is the number of assignments of the i-th subject to the k-th category by m observers, which means that $\bar{p}_o$ is identical to p_o (4.32). The estimate of chance agreement in (4.35) is

$$\bar{p}_e = \sum_{k=1}^{\ell} q_k^2 \tag{4.55}$$

where q_k is the overall proportion of assignments to the k-th category and given by

$$q_k = \frac{1}{nm} \sum_{i=1}^{n} u_{ik} \quad . \tag{4.56}$$

$\hat{\kappa}_{(m)}$ therefore accounts for expected agreement under the baseline constraints of pairwise independence and marginal homogeneity and its bounds are as described for M(I) in (4.28).

The agreement measure, $\hat{\kappa}_{(m)}$, can also be used to measure agreement specific to a single category. As such, the overall agreement is partitioned into components for each category by expressing $\hat{\kappa}_{(m)}$ as shown by Fleiss (1971) and given by

$$\hat{\kappa}_{(m)} = \frac{\sum_{k=1}^{\ell} q_k(1-q_k)\hat{\kappa}_{(m)}^{(k)}}{\sum_{k=1}^{\ell} q_k(1-q_k)} \tag{4.57}$$

where q_k is the overall proportion of assignments to the k-th category as in (4.56) and $\{\hat{\kappa}_{(m)}^{(k)}\}$ for $k = 1,2, \ldots, \ell$ are separate estimates of interobserver agreement for each of the ℓ categories and given by

$$\hat{\kappa}_{(m)}^{(k)} = \frac{Q_k - q_k}{1 - q_k} \tag{4.58}$$

where Q_k is an estimate of the conditional probability of agreement between two randomly selected observers on the assignment of a particular subject to the k-th category, given that the first observer classified the subject into the k-th category and is given by

$$\hat{Q}_k = \frac{\sum_{i=1}^{\ell} \binom{u_{ik}}{2} / \binom{m}{2}}{nq_k} \qquad (4.59)$$

Therefore, $\hat{\kappa}_{(m)}$ can be regarded as a weighted average of the $\hat{\kappa}_{(m)}^{(k)}$. Tests of hypotheses can be performed as discussed for $\hat{\kappa}_{w1}$ (4.34).

Example 4.8

To illustrate the use of agreement measures for polychotomous nominal or ordinal data sets in which more than two observers classify patients, the data set used in Table 4.16, has been extended to include a third observer (Table 4.17).

$$\hat{\kappa}_{(m)} = \frac{\bar{p}_o - \bar{p}_e}{1 - \bar{p}_e} = \frac{0.58 - 0.37}{1 - 0.37} = 0.34$$

For each category, may be calculated:

$$\hat{\kappa}_{(m)}^{(1)} = \frac{Q_1 - q_1}{1 - q_1} = \frac{0.65 - 0.32}{1 - 0.32} = 0.49$$

$$\hat{\kappa}_{(m)}^{(2)} = \frac{0.58 - 0.46}{1 - 0.20} = 0.19$$

$$\hat{\kappa}_{(m)}^{(3)} = \frac{0.49 - 0.20}{1 - 0.20} = 0.36 \text{ and the agreement beyond}$$

chance is greatest on category 1, major physical impairment, and smallest on partial impairment. Hypotheses about these indices may be specified and tested as shown in example 4.6.

Table 4.17

Table Resulting From Patients Being Classified As Having Major Impairment (1), Partial Impairment (2) or Minimal Impairment (3) Of Physical Function by Three Observers

Patients	Observers A	B	C
1	3	3	3
2	2	2	3
3	3	2	2
4	2	2	2
5	2	2	1
6	3	3	2
7	1	2	1
8	2	2	2
9	2	2	1
10	2	2	1
11	1	1	1
12	3	3	3
13	2	2	2
14	2	2	1
15	3	3	3
16	1	1	1
17	1	1	1
18	2	2	1
19	2	2	1
20	1	2	1
21	2	1	1
22	2	2	2
23	1	1	1
24	1	1	2
25	1	1	2
26	3	2	2
27	1	1	1
28	1	1	1
29	3	2	3
30	3	3	2
31	2	2	2
32	2	2	3
33	2	1	3
34	3	3	3
35	1	2	3
36	2	2	1
37	2	2	2
38	2	2	3
39	1	1	1
40	2	2	2
41	1	1	1
42	2	2	2
43	3	2	2
44	2	2	2
45	3	3	3
46	3	3	3

4.2 Univariate Data With An External Standard (Diagram 4.4)

4.2.1 Dichotomous Outcome Variables

The comparison, when an external standard exists, becomes that of each observer with the standard. It has been well characterised for the comparison of a test for the presence or absence of a sign with a standard. The data can be displayed as in Table 4.18. Yerushalmy (1947) introduced the term "sensitivity" for the proportion of true positives given by

$$\hat{\xi} = \frac{n_{11}}{n_{1.}} \tag{4.60}$$

and the term "specificity" for the proportion of true negatives given by

$$\hat{\eta} = \frac{n_{22}}{n_{2.}} \tag{4.61}$$

Both these measures are bounded by zero for no agreement and one for perfect agreement. It is pertinent to note that they provide separate estimates for agreement on the two categories. It was pointed out by Fleiss (1973b) and Feinstein (1975) that to use the procedure for predicting purposes some information on misclassification is needed and they proposed alternative reliability measures, namely the positive predictive value (PPV) or the "true positive rate" of the test

$$\hat{\rho}+ = \frac{n_{11}}{n_{.1}} \tag{4.62}$$

and the negative predictive value (NPV) or the "true negative rate" of

DIAGRAM 4.4. Flow Chart of Agreement Measures for Discrete Data: External Standard

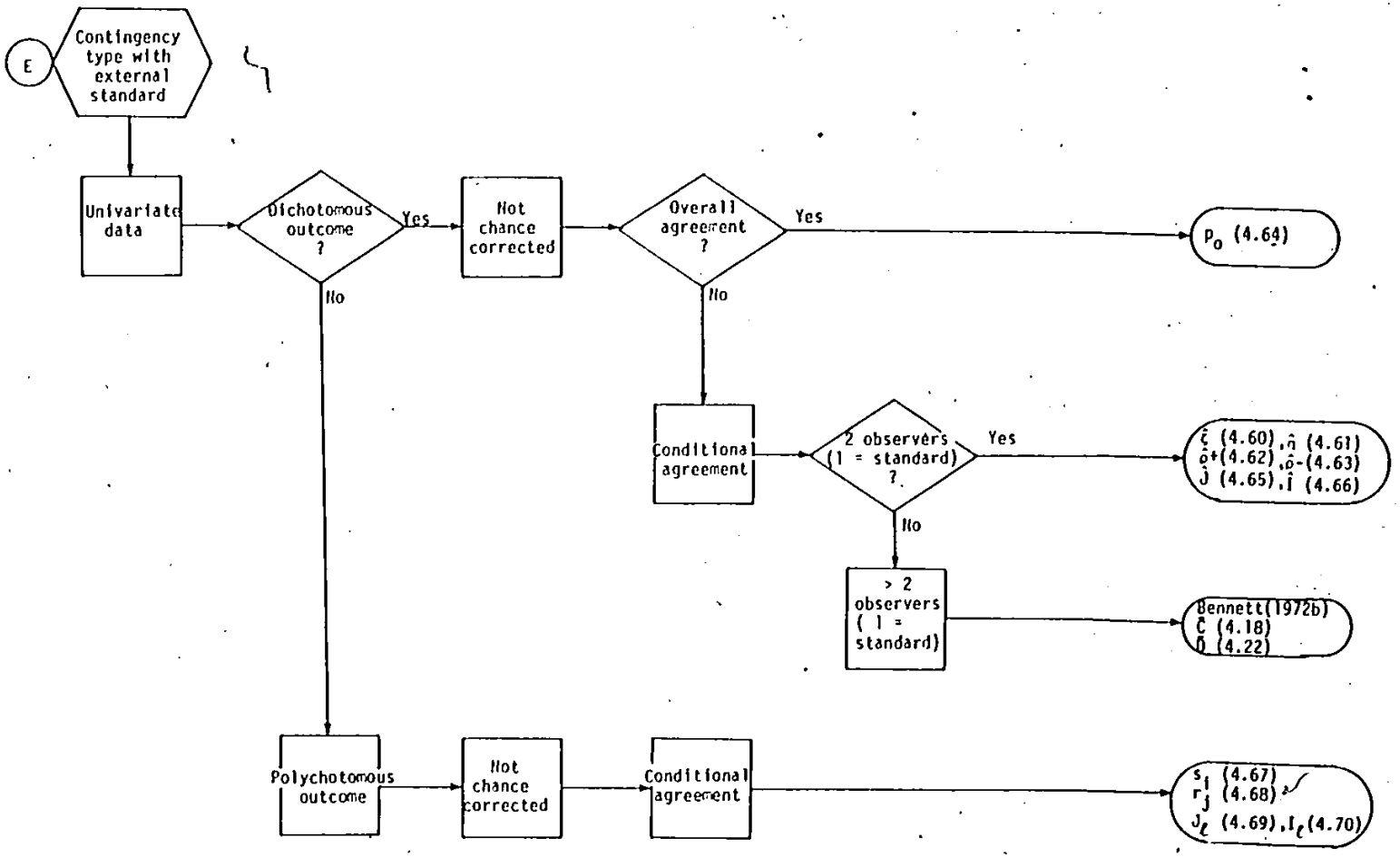


Table 4.18

Classification by Standard or Test Procedure On A Dichotomous Scale

		Test		
		+	-	Total
Standard	+	n_{11}	n_{12}	$n_{1.}$
	-	n_{21}	n_{22}	$n_{2.}$
	Total	$n_{.1}$	$n_{.2}$	$n_{..}$

test

$$\hat{\rho}- = \frac{n_{22}}{n_{.2}} \qquad (4.63)$$

These latter two measures also vary from zero to one, but they are markedly affected by the prevalence of the category of interest (the commoner the category (+ or -), the higher the PPV or NPV respectively). All four measures mentioned thus far can be regarded as conditional probabilities, binomially distributed conditional on the relevant marginal frequency. Hypotheses can therefore be tested and confidence intervals constructed for the true values.

Overall summary estimates of the validity of the test results with respect to the standard giving equal importance to the two categories also exist. The simplest is

$$p_o = \frac{1}{n}(n_{11}+n_{22}) \qquad (4.64)$$

which is directly analogous to the crude index of agreement in (4.10). Youden (1950) proposed a combined sensitivity/specificity index, called Youden's J and estimated by

$$\hat{J} = \hat{\xi} + \hat{\eta} - 1 \quad . \qquad (4.65)$$

Layhew and Goldsmith (1975) proposed a similar index for the combined predictive values, estimated by

$$\hat{I} = (\hat{\rho}+) + (\hat{\rho}-) - 1 \quad . \qquad (4.66)$$

Both these combined indices range from -1 (when both conditional probabilities are equal to zero) to +1 (when both conditional probabilities are equal to one), both have been shown to be related to kappa as in (4.30) and for both confidence intervals can be constructed (Layhew and Goldsmith 1975).

Example 4.8

To illustrate the use of sensitivity-specificity and predictive value measures, data have been taken from a publication by Kilpatrick (1976), in which exercise cardiography is compared with coronary arteriography in the diagnosis of coronary artery disease. The results from coronary arteriography will be regarded as a truth indicator or standard. The data are displayed in Table 4.19.

Sensitivity, $\hat{\xi} = \frac{56}{69} = 0.84$

Specificity, $\hat{\eta} = \frac{15}{23} = 0.65$

Positive predictive value, $\hat{\rho}+ = \frac{56}{66} = 0.88$

Negative predictive value, $\hat{\rho}- = \frac{15}{26} = 0.58$

in a population with a disease prevalence of $69/92 = 0.75$.

Overall indices are:

$$\hat{P}_o = \frac{1}{92}(56+15) = 0.79$$

$$\hat{J} = (0.84+0.65-1) = 0.49$$

$$\hat{I} = (0.88+0.58-1) = 0.46 \quad .$$

4.2.2 Polychotomous Outcome Variables

Where the outcome of the test and standard involves more than

Table 4.19

Data Display To Illustrate Agreement With A Standard

		Test		
		+	-	Total
Truth	+	58	11	69
	-	8	15	23
	Total	66	26	92

Truth is indicated by coronary arteriography, test by exercise vectorcardiography, where + indicates presence of coronary artery disease and - its absence.

two categories, the $\ell \times \ell$ table resulting (Table 4.20) can be collapsed into one of the $(\ell-1)^2$ possible two-by-two tables, but the decision may be arbitrary and some information lost in the process. In this regard Layhew and Goldsmith (1975) have generalized the measures of sensitivity, specificity and predictive values to $\ell \times \ell$ tables by defining the estimates of the i-th sensitivity of the test as

$$s_i = \frac{n_{ii'}}{n_{i.}}, \quad (i,i' = 1,2, \ldots, \ell) \tag{4.67}$$

and the j-th predictive value of the test as

$$r_j = \frac{n_{jj'}}{n_{.j}}, \quad (j,j' = 1,2, \ldots, \ell) \tag{4.68}$$

These measures are also conditionally binomially distributed, and it is relevant to note that with $\ell = 2$, these measures reduce to the original sensitivity, specificity and predictive measures in a two-by-two table. No specificity measures for the $\ell \times \ell$ table were defined, as no single set of cells can be equated with the cell of true negatives in the fourfold table. However, should the investigator be able to make a value judgment in this regard, it may be possible to define them. Layhew and Goldsmith (1975) showed that $\hat{\kappa}$ and $\hat{\kappa}_w$ can be expressed in terms of $\{s_i\}$, $\{r_j\}$ and the marginal proportions of the $\ell \times \ell$ table, with the kappa measures for each category bounded by zero (if no agreement) and one (if the test agrees perfectly with the standard). Layhew and Goldsmith (1975) also proposed generalized indices to summarize the information for ease of interpretation. The generalized

Table 4.20

Classification by Standard or Test Procedure on Polychotomous Scale

		Test			
		1	2	ℓ	Total
Standard	1	n_{11}	n_{12}	$n_{1\ell}$	$n_{1.}$
	2	n_{21}	n_{22}	$n_{2\ell}$	$n_{2.}$
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	ℓ	$n_{\ell 1}$	$n_{\ell 2}$	$n_{\ell\ell}$	$n_{\ell.}$
	Total	$n_{.1}$	$n_{.2}$	$n_{.\ell}$	$n_{..}$

sensitivity index for an $\ell \times \ell$ table is given by

$$\hat{J}_\ell = \frac{\sum_{i=1}^{\ell} s_i - 1}{\ell - 1} \tag{4.69}$$

such that $\hat{J}_\ell$ = Youden's $\hat{J}$ if $\ell = 2$, and the generalized predictive value index is given by

$$\hat{I}_\ell = \frac{\sum_{j=1}^{\ell} r_j - 1}{\ell - 1} \tag{4.70}$$

such that $I_\ell = I$ (4.66) if $\ell = 2$. Both these latter combined indices, J_ℓ and I_ℓ, take chance agreement into account and as such

$$J_\ell = I_\ell = \begin{cases} -1/(\ell-1), \text{when } n_{ii'} = n_{jj'} = 0,\ i,i',\ j,j' = 1,2,\ \ldots,\ \ell \\ 0, \text{ if the test and standard are independent} \\ 1, \text{ if full agreement, when all } s_i = 1, \text{ all } r_j = 1, \\ \qquad\qquad i,j = 1,2,\ \ldots,\ \ell. \end{cases} \tag{4.71}$$

For both J_ℓ and I_ℓ their large sample variances can be estimated as shown by Layhew and Goldsmith (1975) under the hypothesis of no agreement and for each measure a test statistic was developed for testing hypotheses. By specifying weights according to his judgment, the investigator can differentiate between categories, both for dichotomous and polychotomous tests and standards.

For multiple tests or observers Bennett (1972a) showed how the

sensitivity, specificity and predictive values for the m tests or observers on the same set of n subjects can be compared. Also, the measures $\bar{C}$ (4.18) and $\bar{D}$ (4.22) can be used to evaluate each of the m observers compared with the majority opinion (Armitage, Blendis and Smyllie 1966).

4.3 Multivariate Data (Diagram 4.1)

Multivariate problems arise when, e.g. m observers classify n subjects according to c outcome variables. Bennett (1972b) extended the contingency table agreement measures developed by Armitage, Blendis and Smyllie (1966) into multivariate indices in terms of the average proportion put into each category for each of c signs. No more details will be supplied in this thesis about multivariate analysis.

CHAPTER 5

Reprise: "Useful" Measures of Agreement

There is no need for the writer to eat a whole sheep to be able to tell what mutton tastes like. It is enough if he eats a cutlet.

Somerset Maugham

The purpose of this reprise is twofold: firstly, to select a subset of "useful" measures of agreement from among all those discussed in Chapters 3 and 4, and secondly, to summarise their attributes.

5.1 Selection of "Useful" Measures

In the context of this thesis, "useful" refers to the fulfilment of at least four of the following five criteria (one of which must be meaningful agreement):

1. The agreement measured should be clinically meaningful and justified in terms of statistical assumptions.
2. The agreement statistic should have readily interpretable bounds or referent values.
3. Hypothesis testing and confidence interval construction should be possible.
4. The agreement measure should be relatively easy to calculate as availability of computer programmes is not covered in this thesis.
5. A categorical (continuous) analogue for a continuous (cate-

gorical) measure of agreement should exist.

The criteria are based on the subjective judgment of this author and are intended to limit the number of measures from which to choose, while retaining sufficient variety to offer a choice appropriate to commonly occurring situations. The selected measures (Table 5.1), as well as some of the measures not selected (Table 5.2), will be summarily discussed in Section 5.2.

5.2 Summary of "Useful" Measures

Should the reader be in awe of statistical notation, it is suggested that this chapter be read in conjunction with Chapters 1 and 2 with the hope that the condensed flow chart in Diagram 5.1 will be readily understandable. For computational formulae and numerical examples, the reader is referred to the relevant parts of Chapters 3 and 4, as indexed in Diagram 5.1.

5.2.1 Continuous Data

Four similar measures were selected for use with continuous data sets: ρ_1, ρ_2, ρ_3, and ρ_6. They are all intraclass correlation coefficients, where class refers to the factor of interest, usually a set of observers. Due to the interchangeability of the parties involved in making observations (Diagram 2.2), the factor of interest need not be a set of observers, but may also be a set of subjects or instruments.

In general, intraclass correlation coefficients are formulated by expressing the variation due to the factor of interest (and therefore

Table 5.1

Agreement Measures Selected As "Useful"

Agreement Measure	Criteria				
	Meaningful Agreement	Interpretable Bounds	Hypothesis Testing	Hand Calculable	Analogues
Continuous Data					
ρ_1(3.14)*, R^2(3.16)	+	+	+	+	$\hat{\pi}$
ρ_2(3.17)	+	+	+	+	$\hat{\kappa}$
ρ_3(3.23)	+	+	+	+	r_{11}
ρ_6(3.34)	+	+	+	+	-
Categorical Data					
r_{11}(4.8)	+	+	-	+	ρ_3
$\hat{\kappa}$(4.30)	+	+	+	+	ρ_2
$\hat{\pi}$(4.29)	+	+	-	+	ρ_1, R^2
$\hat{\xi}$(4.60), $\hat{\eta}$(4.61)	+	+	+	+	ρ_2 (as J(4.65))
$\hat{\rho}$+(4.62), $\hat{\rho}$-(4.63)	+	+	+	+	ρ_2 (as I(4.66))
s_j(4.62), r_j(4.68)	+	-	+	+	ρ_2 (as J_k(4.69) and I_k(4.70) respectively)

+ = fulfilment of criteria

- = non-fulfilment of criteria

* reference numbers (Chapters 3 and 4)

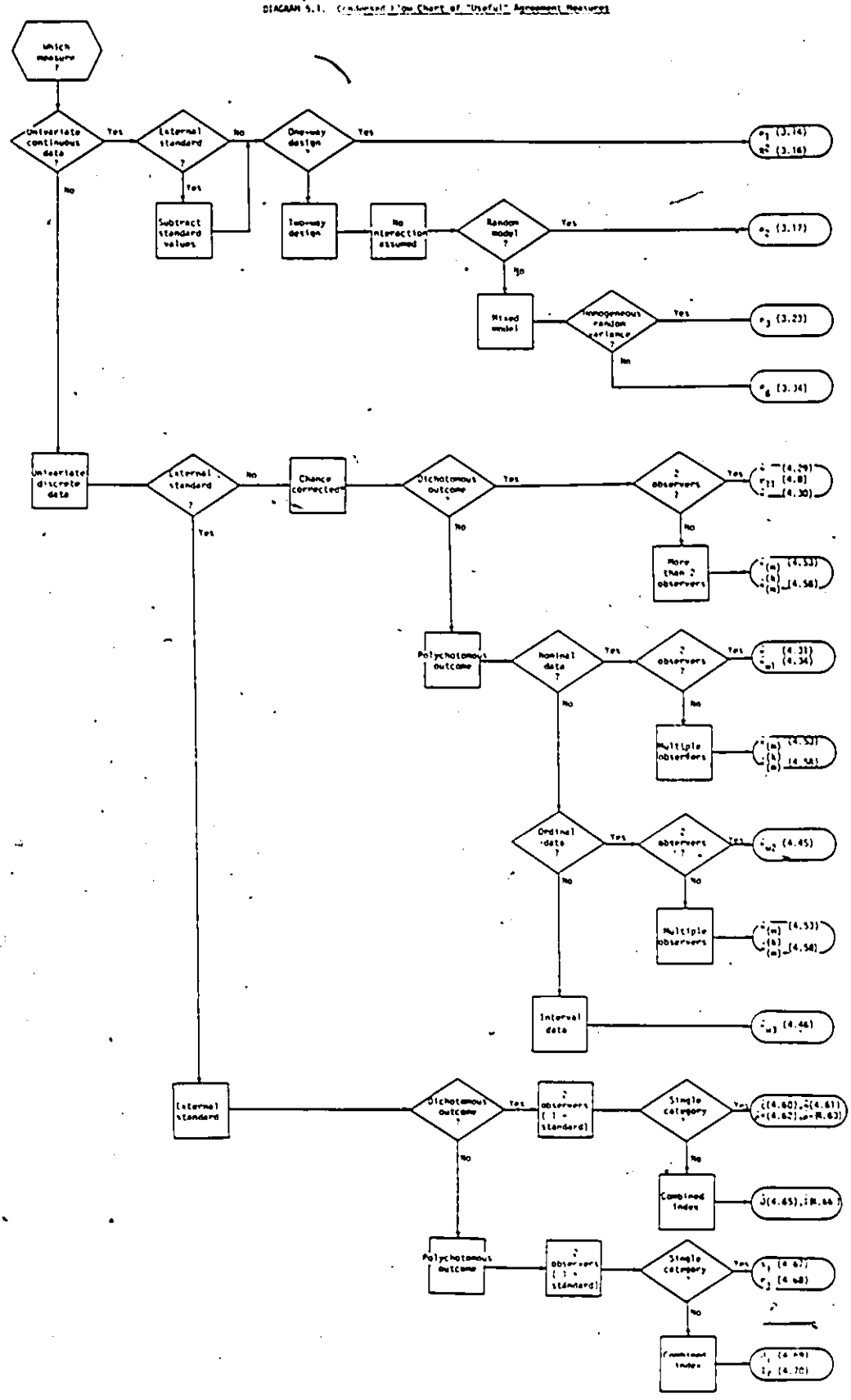
Table 5.2

Agreement Measures Not Selected As "Useful"

Agreement Measure	Criteria				
	Meaningful* Agreement	Interpretable* Bounds	Hypothesis Testing	Hand Calculable	Analogues
Continuous Data					
ρ_4(3.25)	-	+	+	+	-
ρ_5(3.32)	-	+	+	+	-
R_c(3.46)	-	+	+	-	-
Discrete Data					
R_1(4.1)	-	+	+	+	+
R_2(4.3)	-	+	+	+	+
R_3(4.7)	-	+	+	+	+
Q(4.9)	-	-	+	+	-
Yule's Q,Y (Table 4.7)	-	-	+	+	-
ϕ (Table 4.7)	-	+	+	+	-
Pearson's P (Table 4.8)	-	-	-	+	-
Tschuprow's T (Table 4.8)	-	+	-	+	-
p_o(4.10)	-	+	-	+	+
A_p, $A_{\bar{p}}$ (4.11,4.12)	-	+	+	+	+
A_1, A_2 (4.13,4.16)	-	+	-	+	+
C_1, D_1 (4.16,4.21)	-	-	-	+	-
$\bar{C}$, $\bar{D}$ (4.18,4.22)	-	-	-	+	+
S_d (4.19)	-	-	-	+	+
Q (4.24)	-	-	+	+	-
$H(A_2)$(4.31)	+	+	-	+	-
Tau-a (4.41)	-	+	-	+	-
$\bar{\alpha}$ (4.50)	+	+	-	+	-
λ_r (4.51)	+	+	-	+	-
Multivariate	-	-	-	-	+

* The arguments against meaningful agreement or interpretable bounds may be found in the relevant section as indexed.

DIAGRAM 5.1. Condensed Flow Chart of "Useful" Agreement Measures



* cannot r_{11}

the disagreement) as a proportion of the total variation. As the selected four measures are all intraclass correlation coefficients, they share the same properties, fulfilling the selection criteria as follows:

1. The proportion of variation attributed to the factor of interest is directly applicable to clinical and investigative situations. Three of the measures (ρ_1,ρ_2,ρ_3) are used when all the statistical assumptions (3.10 in Chapter 3) are justified, and the other, ρ_6, is used when the assumption of homogeneous random variance is not met.

2. All four selected measures are similarly bounded: by zero when there is no variation due to the factor of interest (perfect agreement among members of that set), and by one when all the variation is due to the factor of interest (complete disagreement). Values in between the bounds can therefore also be readily interpreted: the greater the proportion of variation due to the factor of interest, the greater the disagreement.

3. It is possible to construct confidence intervals and test hypotheses for all four of these measures, the general principle being the comparison of ratios of mean squares to appropriate F distributions.

4. All four intraclass correlation coefficients can be calculated by hand.

5. Categorical analogues have been developed for all these measures, with the exception of ρ_6.

When the assumption of variance homogeneity cannot be made, ρ_6 is used to measure agreement. The choice among the other three depends on the study design, which determines in turn the variance components that can be estimated. Estimation of the variance components can conveniently be accomplished by the analysis of variance procedure for balanced designs, and by other methods if unbalanced (Chapter 3). The appropriate choice of measures by design is:

- ρ_1 for one-way designs, where the variation due to a single factor is the only estimable component. R^2 is an alternative formulation, in which the numerator cannot assume a negative value, something that may happen with ρ_1,

- ρ_2 and ρ_3 both in two-way designs, where the variance components due to two factors can be separately estimated. Interaction between the two factors is either non-existent, or may be assumed not to exist. When the factor of interest is a random subset, of subjects for example, ρ_2 applies and when the factor of interest is a fixed set, of observers for example, ρ_3 applies.

- ρ_6 for situations where variance homogeneity cannot be assumed, and separate random variation terms for each observer-subject combination are estimated.

The four selected measures have all been described for use in assessing agreement among a set of observers, but they may also be applied to measure agreement with an external standard. By subtracting the standard values from the observed values and modelling the intraclass correlation coefficients on the differences between observed and

standard values, agreement with an external standard may be measured. If there were complete agreement with the standard, the expected differences would be zero, as would the agreement measured.

Not included in the condensed flow chart are agreement measures for situations where interaction between factors is present, i.e. where observer effects are not the same for all subjects. When interaction is present, meaningful interpretation of the "main factor effects" and agreement becomes questionable. For this reason, and because categorical analogues have not been developed, they were not selected for the condensed flow chart.

5.2.2 Discrete Data

In general, nine different measures were selected to measure agreement in categorical data sets, three of which (r_{11}, $\hat{\pi}$, $\hat{\kappa}$) are applicable when no external standard is used, and the other six (ξ, $\hat{\eta}$, $\hat{\rho}+$, $\hat{\rho}-$, s_i and r_j) are used when an external standard is employed.

The three measures for use among a set of observers (r_{11}, $\hat{\pi}$, $\hat{\kappa}$) were selected as they fulfilled the criteria as follows:

1. Two of them ($\hat{\pi}$ and $\hat{\kappa}$) are chance corrected measures, measuring the agreement that occurs over and above that expected to occur by chance. In clinical terms this portion of the observed agreement is indeed the most relevant. The third, r_{11}, measures the average agreement between two raters.
2. The chance corrected measures ($\hat{\pi}$ and $\hat{\kappa}$) have a meaningful value of zero, when the observed agreement is the same as

that expected by chance and an upper bound of one, when agreement is perfect. They have a lower bound which is not readily interpretable, but as the portion of agreement of meaningful interest lies between zero and one, the lower bound is irrelevant for most clinical purposes. The third, r_{11}, varies from zero for perfect agreement to one for no agreement.

3. Hypothesis testing and confidence interval construction are possible for $\hat{\kappa}$.
4. All three measures can be calculated by hand.
5. Continuous analogues exist for all three measures.

The three measures are used in the following situations:

- r_{11} is used when there are two observers and they form a fixed set.

- $\hat{\pi}$ is used when the marginal distributions are similar for both observers (no observer effects) and $\hat{\pi}$ is limited for use with two observers.

- $\hat{\kappa}$ can be used in many situations. It may be used for nominal ($\hat{\kappa}$,$\hat{\kappa}_{w1}$), ordinal ($\hat{\kappa}_{w2}$) and interval data ($\hat{\kappa}_{w3}$) by choosing appropriate weights. For more than two observers $\hat{\kappa}_{(m)}$ is applicable and for each category an individual $\hat{\kappa}^{(k)}_{(m)}$ for agreement on that category may be estimated. $\hat{\kappa}_{(m)}$ and $\hat{\kappa}^{(k)}_{(m)}$ may also be applied when the outcome variable has more than two categories.

The six measures listed for measuring agreement with an external standard ($\hat{\xi}$, $\hat{\eta}$, $\hat{\rho}+$, $\hat{\rho}-$, s_i and r_j) are really of two types only. The first kind are sensitivity-specificity ($\hat{\xi}$, $\hat{\eta}$, s_i) measures which

describe how well an observer (or test) classifies subject with reference to the standard. The generalized sensitivity-specificity measure, s_i, estimates the conditional probability of classification into each of i categories. When there are only two categories labelled + for presence of a specified state and - for absence of that state, s_+ is identical to $\hat{\xi}$ (sensitivity) and s_- is identical to $\hat{\eta}$ (specificity). When there are more than two categories, the category representing the presence of the specified state has to be identified by the investigator. Only then can the words "sensitivity" and "specificity" be separately and meaningfully applied to this situation.

The second kind of measures for measuring agreement with an external standard are the predictive value measures ($\hat{\rho}+$, $\hat{\rho}-$ and r_j) which describe how well the state indicated by the standard can be predicted by the observer (or test). The generalized measure, r_j, estimates a predictive value for each of j categories. When there are only two categories, with + indicating the predicted presence of the disease, and - indicating its predicted absence, r_+ is identical with the positive predictive value, $\hat{\rho}+$, and r_- with the negative predictive value, $\hat{\rho}-$.

Both types of measures for measuring agreement with an external standard comply with the selection criteria as follows:

1. They are all conditional probabilities, measuring the proportion of subjects classified into a category by an observer (or test) given that they had been similarly classified by a reference observer (or test). They find application to clinical decision making in evalu-

ating diagnostic tests or methods or measurement.

2. They all have interpretable bounds of zero for no agreement and one for complete agreement, the greater the value in between, the greater the agreement.

3. As conditional probabilities hypotheses about these measures may be tested.

4. They can all be easily calculated by hand.

5. Through their combined indices (see Chapter 4) they are analogous to $\hat{\kappa}$.

5.3 Notable Omissions From List of "Useful" Measures

Notably absent from the condensed flow chart are measures designed for multivariate data sets, specifically because of the complexity of computation and difficulty of interpretation. Although some multivariate agreement measures were mentioned in Chapter 4, a multivariate problem can be reduced to a series of univariate problems, which is the way most multivariate problems are probably handled in practice.

Another omission from the flow chart is measurement of pairwise agreement by measures purposely designed. Such measures D_i (4.21), $\bar{D}$ (4.22) and $\hat{\alpha}$ (4.50) are not chance corrected, nor are their bounds all interpretable nor can hypotheses about them be tested, so that they were not selected. Furthermore, by arranging multiple observers in pairs other agreement measures for two observers can be used.

CHAPTER 6

A Strategy For Evaluating The Flow Chart

"Whom are you?" said he, for he had been to night school.

George Ade

In Chapters 3 and 4 a flow chart of measures of agreement was developed with the aim of aiding clinical investigators in learning about and using measures of agreement. In Chapter 5 a condensed flow chart of "useful" measures was described. In this chapter a strategy to evaluate the use of the flow chart will be described.

6.1 Literature Review

The relevant literature was searched* to see if flow charts have ever been used as a method of instruction in statistics and whether instruction by flow chart has ever been compared with other modes of instruction.

The search revealed that flow charts exist for selecting basic statistical procedures (Harshbarger 1971; Andrews, Klein, Davidson, et al. 1974), but none dealing in detail with agreement analysis were found. Although flow charts have been used for decision making in a variety of fields, the search of educational literature revealed no evaluation studies of flow charts other than the abstracts of two papers

* Computerised search of ERIC (Educational Resources Information Center) file (1966 - Feb. 1979), Lockheed Information Systems, Palo Alto, California, U.S.A.

presented at meetings. In these abstracts (Coscarelli 1977; Gerlach and Schmid 1978), different types of algorithmized instruction (of which flow chart is one) were assessed in solving computational problems and in critical thinking ability. In both these studies the time taken to solve problems was apparently decreased by algorithmized instruction. Time for solving problems was therefore adopted as outcome in the flow chart assessment strategy detailed in this chapter, but no other methodologic information was found in the literature.

6.2 Objective

The objective in designing the strategy in this chapter is to obtain answers to the following questions:

Is the use of the flow chart of agreement measures of more benefit to clinical investigators than conventional teaching methods in terms of

1. Solution of problems?
2. Time spent solving problems correctly?

6.3 Choice of Design

In choosing a design with which to answer the questions posed as the objective, some features are considered desirable in the design. They refer to designs in general and do not concern problems specific to educational evaluation or agreement analysis. The desirable features are used here to choose the basic design which will subsequently be

discussed in detail with regard to the problems inherent in this study. The generally desirable features are:

a. Control Group

The design should allow for the comparison of two groups, one of which is exposed to the flow chart and the other not. The groups should ideally be identical in all other respects, so that the results will be unbiased estimates of the real values.

b. No Sampling Bias

The design should allow for the avoidance of sampling bias by the selection of representative and/or comparable groups.

c. Prospective Direction

The design should allow the investigation to be done prospectively, i.e. progressing naturally from the initial state via exposure to the investigative manoeuvre to the subsequent state (Diagram 6.1). Results from a prospective study are less liable to biases in determining a cause-effect relationship than those from a retrospective study (Feinstein 1977).

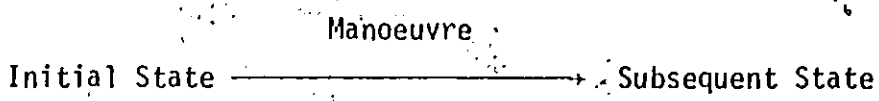
d. Experimental Level

The design should allow for the administration of the investigative manoeuvre to be under the control of the investigator.

In Table 6.1 seven basic study designs are assessed in terms of these desirable features. The more desirable features a design has, the more suitable it may be regarded. The seven designs are briefly

Diagram 6.1

The Experimental Process



Feinstein (1977)

Table 6.1

Desirable Properties of Study Designs: An Assessment of Basic Designs

Desirable features*: allowing for	Designs*						
	Descriptive Survey	Before-After Study	Case-Control Study	Analytic Survey	Cohort Analytic Study	Non-Randomised Controlled Trial	Randomised Controlled Trial
Control group	-	-	+	+	+	+	+
No sampling bias	+	+	-	-	-	-	+
Prospective direction	-	+	-	+	+	+	+
Experimental manoeuvre	-	+	-	-	-	+	+

* The desirable features and the designs are defined in the text.

+: Presence of desirable feature

-: Absence of desirable feature

defined here and the reasons for selection or non-selection mentioned.

Descriptive Survey

A descriptive survey may be defined as an investigation carried out to provide absolute information about a population rather than relative to a comparison or control group. It was not chosen, because it has only one (no sampling bias) of four desirable features:

a. Either no formal control group is used or historical control data are used for comparison.

b. It is possible to select a representative sample; comparability is not an issue when no control group is used. Historical controls are seldom strictly comparable as exposure occurred at a different period in time.

c. The survey is usually done in a crossectional way and not prospectively.

d. No manoeuvre is usually definable.

Before - After Study

A before-after study may be defined as one in which all the subjects in the study sample are exposed to the investigative manoeuvre and their status before and after exposure compared. This design was not considered suitable, because although it has three of four desirable features, it lacks an important fourth:

a. No formal control group is used such that an alternative procedure may be contrasted with the procedure under investigation.

b. A representative sample may be obtained, comparability not being an issue in the absence of a formal control group.

c. The study may be conducted in a prospective fashion.

d. The manoeuvre may be under the investigator's control.

Case - Control Study

A case-control study may be defined as one in which two representative samples are drawn, cases being subjects in whom the outcome of interest is present, controls being subjects in whom the outcome of interest is absent and their prior exposure to the putative cause determined. This design was not chosen, as it may have only one (a control group) of the four desirable features:

a. A control group is used.

b. Sampling bias resulting from incompleteness of the source list, difficulty in defining the parent population and selecting comparable control groups may affect the interpretation of the results.

c. By definition it is a retrospective study.

d. Because of its retrospective nature, the investigator has no control over the manoeuvre.

Analytic Survey

An analytic survey may be defined as a study in which a single sample of subjects is drawn and the exposure and outcome determined either simultaneously or the outcome is not present at the time the exposure is ascertained and is measured by follow-up. This design was not chosen, because it may have only two of the desirable features:

a. A control group is used.

b. Sampling bias may exist by the two groups not being comparable.

c. The study may proceed prospectively.

d. The manoeuvre is not under the control of the investigator, but exposure occurs or does not occur as part of an "experiment of nature".

Cohort Analytic Study

A cohort analytic study may be defined as one in which two samples of subjects without the outcome of interest are drawn, one group exposed to the putative cause and the other not and both groups followed up for the outcome measurement. This design was not considered suitable, because it may have only two desirable features:

a. A control group is used.

b. The cohorts may not be comparable as they are defined by the likelihood of their exposure to the putative cause in the course of an "experiment of nature".

c. The design is prospective in fashion.

d. The manoeuvre is not under the control of the investigator.

Non-Randomised Controlled Trial

A non-randomised controlled trial can be defined as a study in which two cohorts are defined by non-random allocation, one cohort then exposed to an experimental manoeuvre (under the investigator's control), the other group not exposed and both cohorts followed up for the out-

come measurement. The design was not chosen, because although it has three desirable features it lacks the fourth:

a. A control group is used.

b. Non-random allocation to experimental and control groups does not ensure comparability of the groups and the results may be confounded.

c. The design is propective in fashion.

d. The manoeuvre is under the investigator's control.

Randomised Controlled Trial

A randomised controlled trial may be defined as a study in which two identical cohorts are generated through random allocation, one cohort then exposed to an experimental manoeuvre, the other not and both followed up for the outcome of interest. This design incorporates all four desirable features, the sampling bias of the non-randomised controlled trial being avoided by random allocation. As it is the only design with all four desirable features, it has accordingly been selected for use in this strategy.

However, although the design to be described in the rest of this chapter has the basic structure of a randomised controlled trial, the classical design has been modified to make it better suited to educational evaluation. The final choice of design is therefore between the following two designs:

Pretest-Posttest Randomised Controlled Trial (PPRCT)

The heading is almost self-explanatory: this proposed design is a randomised controlled trial with the outcome measurement not a

single measurement after exposure or non-exposure, but the change that occurred from before to after the experimental procedure.

This method of measuring the outcome is thought to be desirable, because clinical researchers vary in their statistical training and experience, and it is conceivable that this factor may confound the results if a single measurement is made. By randomly allocating subjects to two groups it is hoped to achieve identical groups in terms of their backgrounds, so that this confounder is equally likely to affect results in both groups. The additional variation due to differing baseline knowledge and ability is therefore not removed by randomization, but hopefully equalised in the two groups and as such may make the statistical analysis of the results less sensitive to differences between the groups.

Full equalisation may be achieved by prognostic stratification (Feinstein 1977). Another way of handling source of variation would be to measure the statistical "background" of the subjects, isolate its component of variance in the statistical analysis and exclude it from the variance used to assess statistical significance. However, meaningful measurement of statistical "background" may be difficult, if not impossible to achieve, so that the pretest-posttest design seems more suitable: in measuring the before-after change baseline variation is removed and greater statistical sensitivity may result. The pretest-posttest design has been commonly used in educational research (Campbell and Stanley 1963; Isaac and Michael 1971; Ebel 1965).

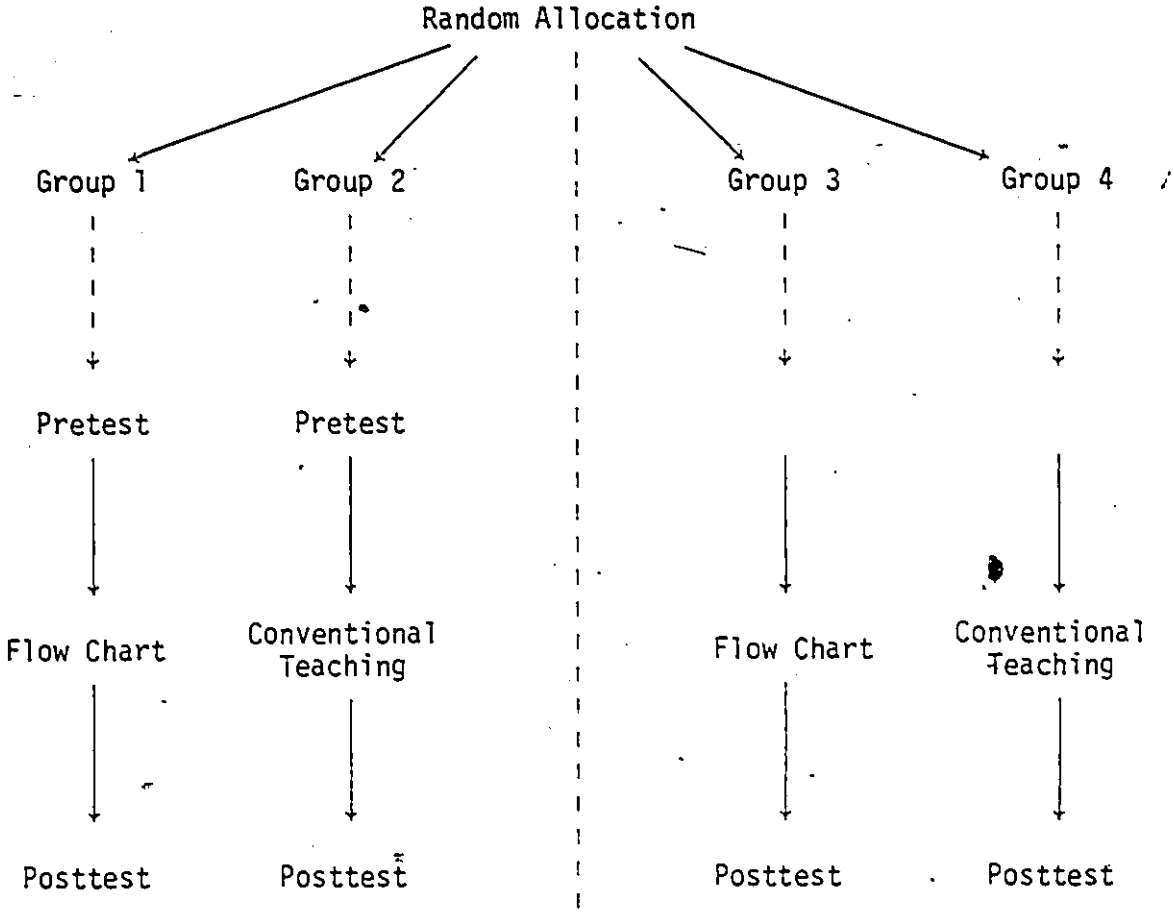
Four Group Randomised Controlled Trial (FGRCT)

This design can be described as a randomised controlled trial in which four identical groups are selected, two of which are exposed and two not, one exposed and one non-exposed group having both pre- and posttests and the other two groups only a posttest (see Diagram 6.2). This design is considered, because it is conceivable that a pretest-manoeuvre interaction may occur in which the pretest alerts the subjects to the subject matter covered by the subsequent manoeuvre. The alerted subjects may therefore perform better in the posttest than they may have without the pretest. This bias would evidently affect both groups in PPRCT equally, not affecting the comparability of the results in the two groups. In that design, however, the generalisability of the results may be affected if the manoeuvre is, or will be, used in real life without the pretest. The questionable generalisability of results from PPRCT may be overcome by the four group design in which the significance of the interaction and of the main effect of the pretest may be tested in the statistical analysis.

The final choice of design is therefore between the last two variations on a randomised controlled trial and is determined by two factors. The first is the sample sizes required, and their feasibility. Sample size estimation will only be possible after the pilot study, so that the deciding factor at this stage is the second of the two factors: the potential use of the flow chart should be considered. As the manoeuvres are envisaged as slide-tape shows which will be generally available after the study, the pretest-posttest format will be retained

Diagram 6.2

Schematic Representation of a Four Group Randomised Controlled Trial



This side on its own represents the two group pretest-posttest randomised controlled trial design.

in its subsequent use. The manoeuvres may also, however, be used in a classroom setting, but because agreement analysis is a fairly specialised field of interest, it is highly likely that the individuals interested in such a course of instruction, are interested because they have been alerted to it in some way. Even outside the teaching environment it is conceivable that when a researcher starts looking for information on agreement analysis, his/her interest stems from being alerted to it by some problem situation. For these reasons, therefore, it seems likely that the additional information to be gained from the more complex four group design, is not sufficiently relevant and meaningful for it to be the preferred design. The pretest-posttest two group randomised controlled trial will therefore be the design in the subsequent sections of this chapter.

6.4 Detailed Strategy

6.4.1 Selection of Study Groups

The selection of the study groups is a problem that may be considered on two levels: firstly, by defining the user population, and therefore the study population, and secondly, by sampling from the identified population.

6.4.1.1 User Population

There are several possible ways of defining the potential user population. It may be thought of as, in the broadest sense, comprising all clinical researchers, as such presenting insurmountable enumeration problems making it an impractical definition. Another way of thinking

about the user population is as the staff and students at institutions where epidemiology and/or biostatistics are taught, but this definition is also broad so that even multistage random sampling may still result in selected venues such that the study is geographically impractical. For these reasons it seems impossible to achieve a representative, yet practicable sample, a situation best accepted and the study focused on the validation of the teaching methodology in the best possible chunk sample. Considering the way the idea for the study arose and also of where the flow chart is most likely to be used, it seems logical to define the user population for the purposes of this study as the staff and students of the Health Sciences and Mathematical Sciences Departments at McMaster University.

Not only do they constitute current and potential clinical researchers, but they may be the only people ever to know of and use the flow chart. Should the flow chart be used elsewhere, it is intended to develop and describe the strategy in sufficient detail that the study could be repeated wherever it is thought necessary.

The departments at McMaster are listed in Table 6.2 and it is intended to contact all departmental chairmen to obtain their consent for departmental participation. A list of all eligible staff and students as at September 1, 1979 will also be obtained from the chairmen for sampling purposes. Eligibility of staff and students refer to their potential to be clinical researchers and for the purposes of this study eligibility criteria are as follows:

Table 6.2

Departments in Health Sciences Faculty, McMaster University

School of Medicine

Anatomy

Anaesthesia

Biochemistry

Clinical Epidemiology and Biostatistics

Family Medicine

Neurosciences

Obstetrics and Gynaecology

Pathology

Paediatrics

Psychiatry

Radiology

Surgery

School of Nursing

Nursing

Inclusion Criteria

a. All staff members with ranks of professor, associate professor, assistant professor, lecturer, research associate or research assistant on September 1, 1979 will be eligible.

b. All postgraduate students will be eligible, meaning masters, doctoral, postdoctoral, medical undergraduate students, residents, fellows and interns as on September 1, 1979.

Exclusion Criteria

a. Those members of the staff and students in the Department of Clinical Epidemiology and Biostatistics (C.E. & B.) exposed to the flow chart during its preparation and those participating in the pilot study.

b. All members of staff with ranks other than those mentioned under inclusion criteria.

c. All undergraduate students not mentioned under inclusion criteria.

d. All members of staff having a service appointment only (as opposed to academic appointments).

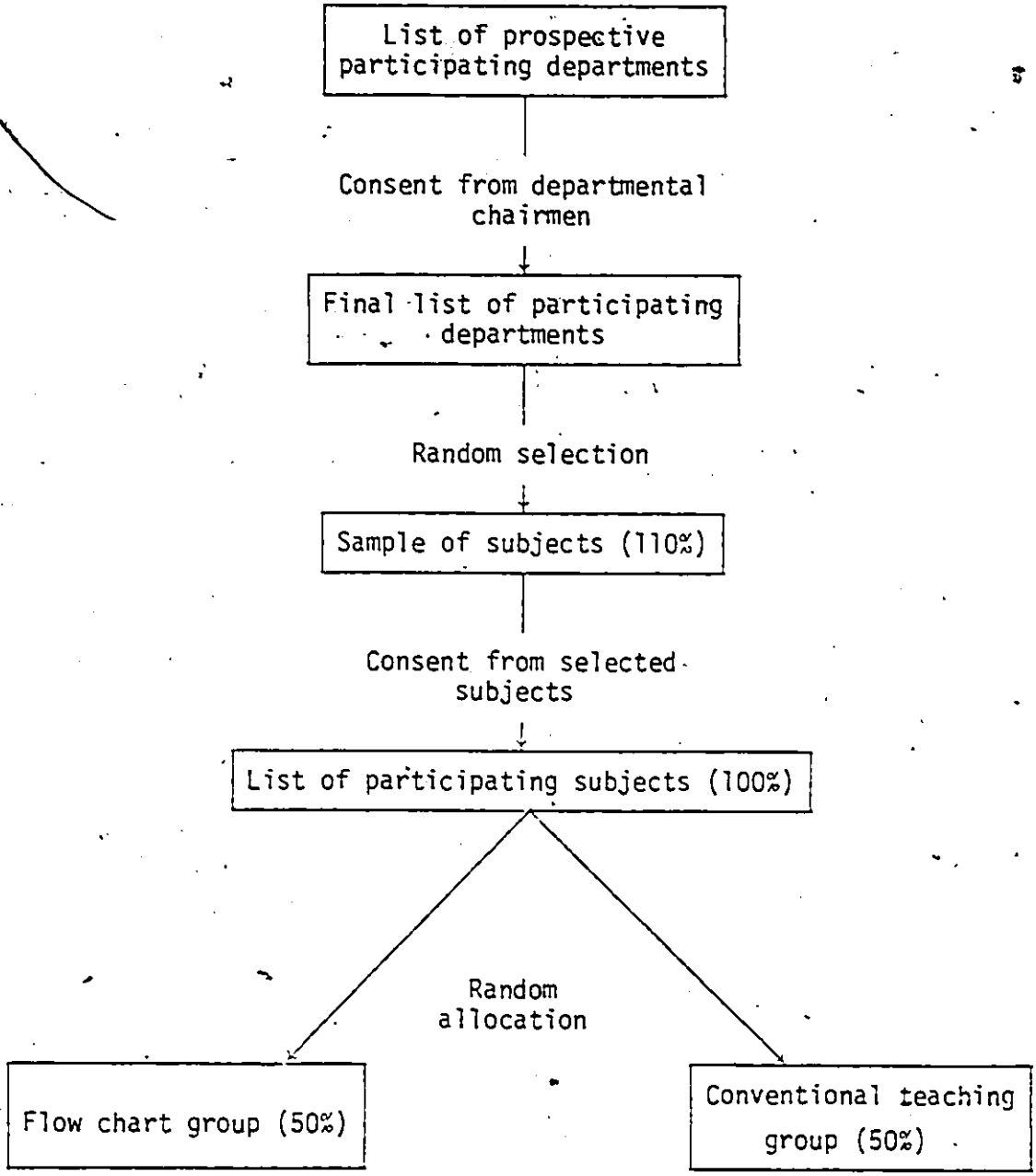
Eligibility decisions will be made by this investigator and verified by the C.E. & B. staff members assisting in this study as: reviewers of several procedures.

6.4.1.2 Sampling Procedure

The sampling and randomisation schedule is shown in Diagram 6.3. It is intended to obtain the consent of departmental chairmen for parti-

Diagram 6.3

Schematic Representation of Sample Selection



cipation, after which a list of eligible staff and students will be drawn up and a random sample drawn, allowing for a ten per cent refusal rate. All these randomly selected individuals will be contacted and their consent for participation sought. A new list of consenting, participating subjects will be compiled, from which random allocation to two groups will be made. The groups will then be randomly assigned as either the "flow chart" or "conventional teaching" groups.

At this stage in the development of the strategy, it is not possible to estimate the required sample sizes as no reliable information on the expected educational change or the expected variation in the outcome measurements is available. It is intended to estimate the sample sizes required by the methods described here, after the pilot study has been conducted.

The outcome measurements are described in Section 6.4.3. The difference between the two groups will be analysed by t-test for independent samples or one-way ANOVA (or equivalent multiple regression analysis) if covariates are to be considered. To estimate the required sample size, the formula given in Colton (1974) will be used:

$$n = 2\left[\frac{(z_\alpha - z_\beta)\sigma}{\delta}\right]^2$$

where n is the sample size in each group,

z_α, z_β are the points in the standardized normal distribution cutting off 100α or $100\beta\%$ in the two tails,

σ is the estimated standard deviation and

δ is the difference in the observed outcome between the two groups.

The unknown information to be obtained from the pilot study is the ratio (σ/δ). Estimated sample sizes for an assumed range of (σ/δ) are shown in Table 6.3.

A sample size estimation will be done for each outcome variable and the larger estimate regarded as the required sample size, provided the parent population is large enough.

Potential Problems

a. Although selected individuals will be free to refuse participation, in order to decrease the likelihood of refusal, they will be assured that the departmental chairmen have agreed to the department's participation, and that the time spent on the study will not affect the security of their employment. They will be offered a stipend of $5 per session to participate (sufficient to cover incidental expenses like parking, but insufficient to induce people to participate for the money only). The evaluation procedure will be kept as short as possible, confidentiality and anonymity of the results will be guaranteed and copies of the flow chart will be made available to all participants at the conclusion of the study. Individual results (their own) will be available to the subject concerned on request.

b. The statistical know-how of the subjects is likely to vary markedly, but it is hoped that by allocating subjects randomly to the two groups, this confounding factor will affect the groups approximately equally.

Prognostic stratification was also considered, but decided

Table 6.3

Estimated Sample Sizes For A t-Test On Two Independent Sample Means

		$(\frac{\sigma}{\delta})$*						
α	β	0.2	0.4	0.6	0.8	1.0	2.0	4.0
0.05	0.1	2	5	10	17	26	104	416
	0.2	1	4	8	14	21	84	337
0.01	0.1	2	6	13	23	23	143	570
	0.2	2	5	11	20	20	119	476

* Unknown information in $n = 2[\frac{(z_\alpha - z_\beta)\sigma}{\delta}]^2$ where n is the sample size for each group and z_α and z_β are points cutting off 100α and 100β per cent in the two tails of the standardised normal distribution respectively. (Colton 1974)

against for the following reasons:

i) No reliable prior information on statistical background will be available.

ii) Information on statistical background may be obtained at the time of the pretest from all participants, with randomisation into strata following the pretest. The time required for marking the pretest, deciding on strata and randomisation, will split the procedure into two sessions: the pretest followed later by the slide-tape show and the posttest, thus necessitating two periods of absence from work for each participant. The clinical credibility to be gained by this procedure is not felt to be justified, when considered against a possible increased refusal rate and doing covariance analysis.

iii) As a refusal to participate may result in a volunteer type bias, attempts will be made to minimise the likelihood of refusal, as outlined in a. In addition, the refusing individuals will be surveyed for information on age, sex, and employment status for comparison with the participants to see if a bias can be detected. Their reasons for refusal will be noted.

6.4.2 Manoeuvre

The experimental procedure will consist of a slide-tape show for each group, one show covering agreement analysis in the conventional way*, the other show using a flow chart approach. Both shows will be accompanied by a printed copy of the verbal presentation and, in the

---

* As taught in the Medical Sciences course 731.

case of the flow chart group, a copy of the flow chart, too. The choice of slide-tape format for presentation of the subject matter over other methods was made for the following reasons:

i) It was felt that presentation of the subject matter could be better standardized than classroom teaching in terms of

a) equal instruction time for both groups

b) the same instructor for both groups

c) the same presentation to all subjects in the same group.

ii) A slide-tape show may be of more widespread use than an interactive computer programme, and practically easier to design.

Both slide-tape shows will be drawn up according to the principles outlined by Anderson, Gent, Goldsmith and Sackett (1970). The content matter of the two shows will be reviewed for comparability by two members of C.E. & B. faculty. It is intended to develop the presentation of both shows in accordance with accepted principles of adult education (Knowles 1973), summarised by a World Health Organization study group (WHO Technical Report Series 1973):

The basic principle is that of learning by doing, i.e. application of educational principles to real problems. Principles of learning of practical use are that

i) learning is an individual process

ii) learning is facilitated if the learner understands clearly what he is expected to learn

iii) learning is facilitated if the student perceives that what

he is expected to learn has relevance to his general goals

iv) learning is facilitated by rapid and complete individual feedback on the extent to which required learning is being accomplished

v) learning rarely occurs without motivation, differentiating between internal and external motivating factors.

The slide-tape shows will be developed, reviewed and tested as part of the pilot study.

The manoeuvre will be further standardised by stipulating and enforcing a single run-through of the presentation for all subjects. However, to ensure that the outcome measured will be a change in knowledge and performance and not memorising ability, the subjects will be able to refer to the handout while doing the posttest.

Potential Problems

i) Aptitude-treatment interaction bias may confound the results if the groups are not comparable in terms of the subjects' learning ability in the slide-tape show situation. In the absence of psychological testing, no prior information on this attribute will be available and no preventive or adjustive measures taken. As a source of bias it will have to be accepted in this study, although its effects may be approximately equalised in the two groups as a result of the random allocation of subjects to groups.

ii) Contamination bias is a real possibility as all the subjects will be working or studying at McMaster University. Their presence at McMaster, however, also means that they will be able to undergo the manoeuvre at a single convenient venue and collectively over a short

period, thereby reducing the likelihood of contamination. The most serious form of contamination would occur if copies of the handouts become available to subjects in the inappropriate group before they undergo the manoeuvre, something that may happen as all subjects will not undergo the manoeuvre at the same time, the number of copies of the slide-tape show being limited. To avoid this source of bias, the handouts will be handed back to the supervisor with the posttest, and not retained by the subjects. Verbal communication is another possible source of bias whose effect will be reduced by conducting the tests over as short a period as possible and asking participants not to discuss the manoeuvre with anybody until they have received copies of the handouts. These copies will be sent to all participants on completion of the whole study. The subjects will also undergo the tests under supervision in separate rooms or screened from one another.

iii) An alternative manoeuvre would be to use the same slide-tape show presentation for both groups, the difference between the two groups being the enclosure of the flow chart with the handout for one group, but not the other. The null hypothesis in this case is that no benefit is derived from the flow chart as an adjunct to conventional teaching. It is a valid hypothesis to test, but the results may be confounded by a bias against the alternative hypothesis resulting from unfamiliarity with the flow chart format.

6.4.3 Outcome Measurement

Three variables will be measured in the study:

i) The statistical background of subjects will be measured in

pretest and used as a covariable in the analysis.

ii) The performance of subjects in solving problems of agreement analysis will be measured in both pretest and posttest. The difference between the scores in the two tests will be compared between the two groups, with statistical background as covariable.

iii) The problem solving tests will be timed in both the pretest and the posttest. The difference between the times taken for the two tests will be compared between the two groups, adjusting for statistical background by covariance analysis.

The measurement of the first two variables will be by means of multiple choice questions (MCQ). The pretest will consist of two parts, one part measuring statistical background and the other part (comparable to the posttest) will measure the problem solving performance of the students. The tests will be made up as follows:

Pretest - Part One

Part one of the pretest will be designed to measure statistical background by a series of ten MCQs, each with five possible answers, one of which will be the best (assumed correct) answer. Each correct answer will carry one mark so that any integer value between and including zero and ten will be possible. The questions will cover basic statistical concepts, including (not exclusively) agreement analysis.

Pretest - Part Two

Part two of the pretest will be designed to measure the problem solving performance of subjects. It will consist of twenty questions

each with five possible answers, one of which is the best (correct) answer. Each correct answer will carry one mark, so that any integer value between and including zero and twenty will be attainable. The questions will deal with agreement analysis exclusively, and will consist of some data sets with questions designed to measure the application of knowledge about agreement analysis. Knowledge acquisition will not be measured directly.

Posttest

The posttest will be identical to the Pretest part two in the number of questions, the maximum mark attainable, the scope covered by the questions, the overall degree of difficulty and in measuring problem solving performance. The questions in the posttest will, however, not be identical to the questions in the pretest part two. It is proposed to draw up pairs of comparable questions, test them in the pilot study for the criteria mentioned, use those complying with the criteria by randomly allocating one from each pair to the pretest part two and the other to the posttest. The order in which the questions occur in each test will be randomly determined for both tests, as will the order of the possible answers.

When a subject is being tested, the pretest part two will be attempted first, completed and returned to the supervisor, its completion time recorded by the supervisor. Pretest part one will then be completed, followed by the slide-tape show and the posttest, which will again be timed by the supervisor. After that, all the material is returned to the supervisor. Whether the subjects will be tested together

or separately will be determined after the pilot study, in which the administration as well as the validity of the multiple choice test items will be pretested (see Section 6.4.6.2).

Potential Problems

i) The multiple choice questions will be drawn up by the same person who designs the flow chart and it is possible that the phrasing of the questions may be flow chart oriented, creating a bias favouring the flow chart. Two faculty reviewers will be asked to look at this specific problem as part of the development of the tests.

ii) Multiple choice questioning is the measurement of choice, because it is deemed possible to construct a sensitive and unbiased scoring system for them. However, as it is possible to guess a correct answer on the one hand and as questions may be omitted on the other, these contingencies will be taken care of in the scoring system. If all questions have the same number of possible answers, the two situations can be accounted for as described by Ebel (1965): For guessing,

$$S = C - \left(\frac{I}{k-1}\right)$$

where S is the score corrected for guessing,

C is the number of correct answers,

I is the number of incorrect answers,

k is the number of possible answers for each question,

I + C = N and N is the total number of questions.

For omission of questions

$$S' = C + \frac{O}{k}$$

where O is the number of questions omitted and $C + O \neq N$. In using S all omitted questions are counted as incorrect and all incorrect answers are assumed to be due to guessing, whereas in S' no correction for guessing is made. It is therefore proposed to combine the two indices as

$$S'' = C - \frac{I}{k-1} + \frac{O}{k}$$

where $C + I + O = N$. Participating subjects will be forewarned of the penalty for incorrect answers and the correction for omissions.

iii) The estimates of the time taken to solve the problems may be biased, as a shorter time may measure quicker solution of the problems, but not necessarily quicker solution yielding correct answers. For this reason the time variable will be weighted according to the subject's score and the variable used in the analysis will be the difference between the weighted time estimates for the posttest and pretest part two:

if t_1 is time taken for pretest part two

t_2 is time taken for posttest

S_1'' is score for pretest part two

S_2'' is score for posttest

then

$t_1' = t_1(\frac{20}{S_1''})$ is the weighted time for pretest part two

$t_2' = t_2(\frac{20}{S_2''})$ is the weighted time for posttest, both standardized

to the maximum attainable score, and the analysis will be done on

$$d_{(t)} = t'_1 - t'_2$$

iv) It is important to realise that the study is not intended to evaluate the appropriateness of the measures selected by using the flow chart - by definition, a flow chart is designed to provide the same answer to the same problem repeatedly if used as designed. The study is therefore intended for evaluating the use of the flow chart. For the purposes of the study the measures selected by using the flow chart are regarded as the "truth".

6.4.4 Statistical Analysis

Preliminary Analysis

i) If the refusal rate is greater than 10%, the respondents will first be compared with the non-respondents for the following variables: age, sex and employment status (t-test for age, chi-square for sex and employment status). Should statistically significant (5% level) and meaningful differences exist, the possibility of volunteer bias confounding the results will have to be considered in the interpretation.

ii) The comparability of the two experimental groups in terms of age, sex and employment status will also be compared. Should statistically significant (5% level) and meaningful differences exist, these variables will be treated as covariables in the main analysis.

Main Analysis

Two separate, but similar analyses will be done, one for each of problem solving score and weighted time. Remembering that both outcome variables in the analysis will be the difference (d) between pretest and posttest, the table of data resulting from the study is expected to look like Table 6.4, for both variables. The procedure to be followed in each case will be multiple regression analysis, with statistical background as covariable. Multiple regression, rather than one-way ANOVA will be used, as the covariable will take on values between 0 and 10, to be regarded as a continuous variable , unless the data distribution suggests categorisation.

The multiple regression model in each case will be

$$y = \beta_0 + \beta_1 STAT + \beta_2(d) + \beta_3(dSTAT) + e$$

where β_0 is the overall mean,

β_1 is the effect due to the covariable STATistical background,

β_2 is the effect due to the difference between the two groups in the outcome variable,

β_3 is the interaction effect.

Two hypotheses will be tested:

First:

Ho: $\beta_3 = 0$,

no interaction by partial F(dSTAT/STAT,d).

If the interaction term is not significant at 5% level, it will be assumed absent and the next hypothesis tested:

Table 6.4

Data Arrangement For the Analysis of Outcome Variables
(d = difference between pretest and posttest)

Group (2)	Observations (n)	Total
Flow Chart Group	$d_{11}, d_{12}, \ldots, d_{1n}$	$d_{1.}$
Conventional Teaching	$d_{21}, d_{22}, \ldots, d_{2n}$	$d_{2.}$

$$Ho: \beta_2 = 0,$$

no difference between the groups by partial F(d/STAT). The adjusted means for the two groups will be given by $(\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2)$ for one group and $(\hat{\beta}_0+\hat{\beta}_1-\hat{\beta}_2)$ for the other.

The sources of variation and degrees of freedom are shown in Table 6.5.

6.4.5 Ethics

Informed consent will be obtained from the departmental chairmen as well as the all selected individuals. All subjects will be free to refuse participation in the study, although a stipend will be paid to those participating. The results of the test will be confidential and not supplied to anybody other than the relevant participant and then only on request. Participants' anonymity in presentation or publication of the results is guaranteed. Handouts of the verbal presentation of both slide-tape shows will be sent to all participants at the end of the study.

6.4.6 Pilot Study

A pilot study will be conducted on some willing students and staff of the department of Clinical Epidemiology and Biostatistics. They were chosen for the pilot study as they will be readily available, have a wide range of statistical expertise and will probably be exposed to the flow chart at the thesis defence, thereby being ineligible for the definitive study.

The pilot study will have two major stages: development and

Table 6.5

Multiple Regression Analysis

Source	d.f.	Sums of Squares	Means Squares	Variance Ratio
Regression				
STAT	1	SS_{STAT}	MS_{STAT}	
d/STAT	1	SS_d	MS_d	$(MS_d/MS_e)^*$
dSTAT/d,STAT	1	SS_{dSTAT}	MS_{dSTAT}	MS_{dSTAT}/MS_e
Residual	2n-4	SS_e	MS_e	
Total	2n-1			

* $MS_d/(SS_{dSTAT}+SS_e)/(2n-3))$ if interaction assumed absent

validation.

6.4.6.1 Development of Instruments

i) Two slide-tape shows and handouts will be developed according to the principles mentioned in Section 6.4.2.

ii) The multiple choice questionnaires will be developed by drawing up approximately 15 basic statistics questions, and approximately 30 pairs of questions on agreement analysis. Each question will have five possible answers, one of which will be designated the "best" (correct) answer by this investigator. The correctness will be reviewed by two C.E. & B. faculty members with the final decision made by consensus.

All these questions will be given to one third of the selected pilot study subjects to be completed. The subsequent item analysis will be in three parts:

a. Item difficulty, which may be defined as the per cent of the group answering the item correctly (really an index of ease). As suggested in Edwards and Scannell (1968), only questions with a degree of difficulty between 40 and 70% will be selected for the final tests.

b. Their discrimination, which reflects the effectiveness with which an item distinguishes high scoring from low scoring students. The top and bottom 25% of the group scores will be used for estimating the discrimination index,

$$D = \frac{U_c - L_c}{U}$$

where U_c is the number in the upper group answering item correctly,

L_c is the number in the lower group answering item correctly,

U = L is the number in each group and

D ranges from -1 to +1 and only items with D greater than 0.30 will be selected.

c. Distracter functioning, where distracter refers to the non-correct answers, is judged to be proper if it is selected by some students, but more attractive to lower than upper group students.

Ten basic statistical and 20 pairs of agreement analysis questions will be selected if they fulfill all 3 criteria:

a) Item difficulty between 40 and 70%,

b) Item discrimination greater than 0.30,

c) Distracters selected by more lower than upper group students.

The ten basic statistical questions will make up Pretest Part One, and the members of each pair of 20 questions on agreement analysis will be randomly allocated to Pretest Part Two, or Posttest.

6.4.6.2 Validation of Instruments

The pre- and posttests thus assembled will then be given to the other two-thirds of the selected pilot study sample to test several aspects of overall validity. The procedure used will be the same as that envisaged for the definitive study, i.e. two groups will be randomly selected and randomly allocated to the two slide-tape shows. Several bits of information will be obtained.

Procedure: how long the whole procedure takes, whether it should be administered to the subjects in groups or singly?

Sample size estimation: information on the expected difference between the groups and the variation will be obtained for both outcome variables.

Instrument validation:

i) The timing procedure of the Pretest (Part two) and the posttest will be precisely defined.

ii) The validity of the test scores will be assessed by considering the following aspects of instrument validity*:

a. Internal validity (precision): no test-retest of the same set of questions on the same subject is possible, because of learning and memory effects. It is proposed to split the tests into two (odd and even numbered questions), mark them separately and estimate the intraclass correlation coefficient for the two halves: it should be no higher than 0.30 and preferably less than 0.20.

b. External validity (accuracy or concurrent criterion validity): a problem in this regard is the unavailability of a criterion or external standard. In its absence it is proposed to compare the mean score obtained to the score midway between the chance score and the maximum possible (Edwards and Scannell 1968; Ebel 1965). As there are five possible answers for each question, one of which is correct, the

---

* Other non-measurable aspects will be put to the 2 C.E. & B. faculty reviewers to decide by consensus: face validity (do the questionnaires appear valid?), content validity (are the questionnaires comprehensive and exclusive with regard to agreement analysis?) and construct validity (do the questionnaires measure the use of the flow charts?).

likelihood of guessing correctly for each item is 0.2, assuming statistical independence for all questions. For all 10 the basic statistics questions in Pretest Part One, the chance score would therefore be $(0.20)^{10} = 1.02 \times 10^{-7}$, and for the 20 agreement questions $(0.20)^{20} = 1.05 \times 10^{-14}$, both of which are very small numbers, approximating zero in the context of this study, and therefore assumed zero. The criterion for the basic statistics questions is therefore 5/10, and for the agreement questions 10/20, or 50% for both. If the mean score obtained is between 40 and 60% it would be judged acceptable.

c. Marker Validation: a score sheet with the correct answers (one per question) will be prepared for use by the marker(s). Intra-marker reliability will be assessed by getting the same marker to mark a few subjects' answer sheets more than once (twice) and inter-marker reliability by having the markers all mark the answer sheet of the same subject. Because of the score sheet supplied for use by the marker, only perfect reliability (agreement) within and among observers will be acceptable.

6.4.7 Budget

Quotations for the following have been received:

Slide-tape show: $16/hour for art work
$18/hour for taping
$1.80/slide for photography
$10 for miscellaneous expenses .

For a 40 minute show it is apparently advisable to allow two hours for taping and one hour for editing.

Printing: 2 × 100 copies, each backprinted, 10 pages long 2 × $50.

Other expenses

Time of investigators: preparing slide-tape show

reviewing flow chart and MCQ tests

supervising the execution of the study

Computer analyses

Stipend to participants: $5 per participant.

CHAPTER 7

Concluding Remarks

In Chapter 1, the objectives for this thesis were stated:

- to review existing measures of agreement
- to organise them into a programmed format to facilitate selection of agreement measures by clinical researchers
- to design a strategy to evaluate the use of the programmed format.

These objectives have now been accomplished and the next logical steps would be to evaluate the use of the flow chart and for the flow chart to be used by clinical researchers. This author is particularly keen for the chart to be used, as the practical experience gained from its use, may be expected to change the list of "useful" measures, a list selected in this thesis on largely theoretical grounds.

Other aspects of agreement analysis which the author is considering as future research projects are:

- methodologic standards for the design of studies of agreement
- sample size requirements in studies of agreement
- agreement measures for multivariate data
- computer programmes for agreement measure estimation
- agreement analysis of clinical decisions: establishing new diagnostic tests, using multiple tests or repeated observations.

REFERENCES

Ade, G., quoted in Peter, L.J. (1977), Peter's Quotations. Ideas for our time. Bantam Books, Toronto, London, New York.

Anderson, R.L., Bancroft, T.A. (1952), Statistical theory in research. McGraw-Hill, New York.

Anderson, G.D., Gent, M., Goldsmith, C.H., Sackett, D.L. (1970), The use of visual aids in teaching Biostatistics; How on earth do you teach Biostatistics and Epidemiology to medical students? Personal communication, slide-tape show CT376, Health Sciences Library, McMaster University, Hamilton, Ontario.

Andrews, F.M., Klem, L., Davidson, T.N., O'Malley, P.M., Rodgers, W.L. (1974), A guide for selecting statistical techniques for analyzing social science data. Institute for Social Research, The University of Michigan, Ann Arbor, Michigan.

Armitage, P., Blendis, L.M., Smyllie, H.C. (1966), The measurement of observer disagreement in the recording of signs. *J.R. Statist. Soc. A 129*: 98-109.

Bartko, J.J. (1976), On various intraclass correlation reliability coefficients. *Psych. Bull. 83*: 762-765.

Bennett, B.M. (1967), Tests of hypotheses concerning matched samples. *J.R. Statist. Soc. B 29*: 468-474.

Bennett, B.M. (1968), Note on χ^2 tests for matched samples. *J.R. Statist. Soc. B 30*: 368-370.

Bennett, B.M. (1972a), On comparisons of sensitivity, specificity and predictive value of a number of diagnostic procedures. *Biometrics 28*: 793-800.

Bennett, B.M. (1972b), Measures for clinicians' disagreements over signs. *Biometrics 28*: 607-612.

Bershad, M.A. (1969), The index of inconsistency for an L-fold classification system, L ≥ 2. U.S. Bureau of the Census, Technical Notes 2: 1-3.

Burdock, E.I., Fleiss, J.L., Hardesty, A.S. (1963), A new view of inter-observer agreement. *Personnel. Psychol. 16*: 373-384.

Campbell, D.T., Stanley, J.C. (1963), Experimental and quasi-experimental designs for research. Rand McNally College Publishing Company, Chicago.

Cartwright, D.S. (1956), A rapid non-parametric estimate of multi-judge reliability. *Psychometrika 21*: 17-29.

Chen, W.Y., Crittenden, L.B., Mantel, N., Cameron,W.R. (1961), Site distribution of cancer deaths in husband-wife and sibling pairs. *J. Nat. Cancer Inst. 27*: 875-892.

Cicchetti, D.V. (1972), A new measure of agreement between rank ordered variables., Proceedings, 80th Annual Convention, APA, 17-18.

Cicchetti, D.V., Allison, T. (1973), Assessing the reliability of scoring EEG sleep records: an improved method. Proceedings of the Electro-physiological Technologists' Association 20: 92-102.

Cohen, J. (1960), A coefficient of agreement for nominal scales. *Educ. and Psychol. Meas. 20*: 37-46.

Cohen, J. (1968), Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull. 70*: 213-220.

Colton, T. (1974), Statistics in Medicine. Little, Brown and Company, Boston.

Conn, H.O., Smith, H.W., Brodoff, M. (1965), Observer variation in the endoscopic diagnosis of esophageal varices. A prospective investigation of the diagnostic validity of esophagoscopy. *N. Engl. J. Med. 272*: 830-834.

Coscarelli, W.C. (1977), Inquiry in development: Efficiency and effectiveness of algorithmic representations in a laboratory situation. Paper presented at the Annual Meeting of the Association for Educational Communications and Technology, Miami, Florida, April 25-29.

De Montaigne, M., quoted in Peter, L.J. (1977), Peter's Quotations. Ideas for our time. Bantam Books, Toronto, London, New York.

Dice, L.R. (1945), Measures of the amount of ecologic association between species. *Ecology 26*: 297-302.

Ebel, R.L. (1951), Estimation of the reliability of ratings. *Psychometrika 16*: 407-424.

Ebel, R.L. (1965), Measuring educational achievement. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Edwards, A.J., Scannell, D.P. (1968), Educational Psychology. The teaching-learning process. International Textbook Company, Scranton, Pennsylvania.

Everitt, B.S. (1968), Moments of the statistics kappa and weighted

kappa. *Brit. J. Math. and Statist. Psychol. 21*: 97-103.

Feinstein, A.R. (1975), Clinical Biostatistics. XXXI. On the sensitivity, specificity and discrimination of diagnostic tests. *Clin. Pharmacol. Ther. 17*: 104-116.

Feinstein, A.R. (1977), Clinical Biostatistics. C.V. Mosby Company, St. Louis,

Fleiss, J.L. (1965), Estimating the accuracy of dichotomous judgments. *Psychometrika 30*: 469-479.

Fleiss, J.L. (1966), Assessing the accuracy of multivariate observations. *J. Amer. Statist. Assoc. 61*: 403-412.

Fleiss, J.L. (1971), Measuring nominal scale agreement among many raters. *Psychol. Bull. 79*: 378-382.

Fleiss, J.L. (1973a), Measuring agreement between two judges on the presence or absence of a trait. Paper presented at Joint Meetings of the American Statistical Association, New York City.

Fleiss, J.L. (1973b), Statistical Methods for rates and proportions. John Wiley and Sons, New York.

Fleiss, J.L., Cicchetti, D.V. (1975), The non-null distribution of weighted kappa. Invited paper presented at the Joint Central Regional Meetings of the American Statistical Association, March 1975, St. Paul, Minnesota.

Fleiss, J.L., Cohen, J., Everitt, B.S. (1969), Large sample standard errors of kappa and weighted kappa. *Psychol. Bull. 72*: 323-327.

Fletcher, C.M. (1964), The problem of observer variation in medical diagnosis with special reference to chest diseases. *Methods Inf. Med. 3*: 98-103.

Fletcher, C.M., Oldham, P.D. (1964), Diagnosis in group research. Chapter of Medical Surveys and Clinical Trials, ed. L.J. Witts, 2nd ed., Oxford University Press, London.

Garland, L.H. (1959), Studies on the accuracy of diagnostic procedures. *Am. J. Roentgenol. Radium Ther. Nucl. Med. 82*: 25-38.

Garland, L.H. (1960), The problem of observer error. *Bull. N.Y. Acad. Med. 36*: 570-584.

Gerlach, V.S., Schmid, R.F. (1978), The efficiency of algorithmized instruction. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, March 27-31.

Gleser, G.C., Cronbach, L.J., Rajaratnum, N. (1965), Generalizing of scores influenced by multiple sources of variance. *Psychometrika 30*: 395-418.

Goodman, L.A., Kruskal, W.H. (1954), Measures of association for cross classification. *J. Amer. Statist. Assoc. 49*: 732-764.

Grubbs, F.E. (1948), On estimating precision of measuring instruments and product variability. *J. Amer. Statist. Assoc. 43*: 243-264.

Hansen, M.H., Hurwitz, W.N., Pritzker, L. (1964), The estimation and interpretation of gross differences and the sample response variance. Contributions to statistics presented to Professor P.S. Mahalanobis on the occasion of his 70th birthday. Pergamon Press, Calcutta.

Harshbarger, T.R. (1971), Introductory Statistics: A Decision Map. New York, MacMillan and London, Collier-MacMillan, 1st edition.

Hartley, H.O., Rao, J.N.K. (1967), Maximum-likelihood estimation for the

mixed analysis of variance model. *Biometrika 54*: 93-108.

Hemmerle, W.J., Hartley, H.O. (1973), Computing maximum likelihood estimates for the mixed A.O.V. model using the W transformation. *Technometrics 15*: 819-831.

Isaac, S., Michael, W.B. (1971), Handbook in Research and Evaluation for education and the behavioral sciences. Edits Publishers, San Diego, California.

Kendall, M.G. (1955), Rank correlation methods. Hafner Pub. Co., New York, 2nd edition.

Kendall, M.G., Stuart, A. (1961), The advanced theory of statistics, Vol. 2. Hafner Pub. Co., New York.

Kilpatrick, D. (1976), Exercise vectorcardiography in diagnosis of ischaemic heart-disease. *Lancet ii*: 332-334.

Kish, L. (1962), Studies of interviewer variance for attitudinal variables. *J. Amer. Statist. Assoc. 57*: 92-115.

Kleinbaum, D.G., Kupper, L.L. (1978), Applied Regression Analysis and Other Multivariate Methods. Duxbury Press, North Scituate, Massachusetts.

Knowles, M. (1973), The Adult Learner: A Neglected Species. Gulf Publishing Company, Houston, Texas

Koch, G.G. (1967), A general approach to the estimation of variance components .: *Technometrics 9*: 93-118.

Koch, G.G. (1968), Some further remarks concerning 'A general approach to the estimation of variance components'. *Technometrics 10*: 551-558.

Koch, G.G. (1973), An alternative approach to multivariate response

error models for sample survey data with applications to estimators involving subclass means. *J. Amer. Statist. Assoc. 68*: 906-913.

Koran, L.M. (1975), The reliability of clinical methods, data and judgments. Part I: *New Engl. J. Med. 293*: 642-646. Part II: *New Engl. J. Med. 293*: 695-701.

Landis, J.R., Koch, G.G. (1975), A review of statistical methods in the analysis of data arising from observer reliability studies. Part I: *Statist. Neerl. 29*: 101-123. Part II: *Statist. Neerl. 29*: 151-161.

Layhew, G.S., Goldsmith, C.H. (1975), Generalized sensitivity-specificity and predictive value measures of agreement. Personal communication, Statistics Technical Report 75/1, Department of Applied Mathematics, McMaster University, Hamilton, Ontario.

Light, R.J. (1971), Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychol. Bull. 76*: 367-377.

Maugham, W. Somerset, quoted in Peter, L.J. (1977), Peter's Quotations. Ideas for our time. Bantam Books, Toronto, London, New York.

Maxwell, A.E., Pilliner, A.E.G. (1968), Deriving coefficients of reliability and agreement for ratings. *Brit. J. Math. Statist. Psychol. 21*: 105-116.

Overall, J.E. (1968), Estimating individual rater reliabilities from analysis of treatment effects. *Educ. and Psychol. Meas. 28*: 255-264.

Reynolds, H.T. (1977), Analysis of nominal data. Sage University Paper

Series on Quantitative Applications in the Social Sciences 07-007, Beverly Hills and London: Sage Pubns.

Robinson, W.S. (1975), The measurement of agreement. *Amer. Sociol. Review 22*: 17-25.

Rogot, E., Goldberg, I.D. (1966), A proposal index for measuring agreement in test-retest studies. *J. Chronic Dis. 19*: 991-1006.

Rorem, N., quoted in Peter, L.J. (1977), Peter's Quotations. Ideas for our time. Bantam Books, Toronto, London, New York.

Sackett, D.L. (1978), Clinical diagnosis and the clinical laboratory. *Clin. and Invest. Med. 1*: 37-43.

Satterthwaite, F.E. (1946), An approximate distribution of estimates of variance components. *Biometrics Bull. 2*: 110-114.

Scheffé, H. (1959), The analysis of variance. John Wiley & Sons, New York.

Scott, W.A. (1955), Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quart. 19*: 321-325.

Searle, S.R. (1971), Linear models. John Wiley & Sons, New York.

Smith, H.F. (1960), Estimating precision of measuring instruments. *J. Amer. Statist. Assoc. 45*: 447-451.

Spiers, P.S., Quade, D. (1970), On the question of an infectious process in the origin of childhood leukemia. *Biometrics 26*: 723-737.

Todd, J.W. (1953), The superior clinical acumen of the old physician. A myth. *Lancet 1*: 482.

WHO Technical Report Series number 521 (1973), Training and preparation of teachers for schools of Medicine and of Allied Health Sciences.

Yerushalmy, J. (1947), Statistical problems in assessing methods of medical diagnosis with special reference to X-ray techniques. *Publ. Hlth. Rep. (Wash.)* 62: 1432-1449.

Youden, W.J. (1950), Index of rating diagnostic tests. *Cancer* 3: 32-35.