This research was supported under Research Grants GM70004-01, GM00038-18 and GM12868-08 from the National Institute of General Medical Sciences and Research Grant HD-03441-04 from the National Institute of Child Health and Human Development.

LINEAR MODELS ANALYSIS OF INCOMPLETE MULTIVARIATE CATEGORICAL DATA

by

Peter B. Imrey and Gary G. Koch

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 820

MAY 1972

ABSTRACT

PETER B. IMREY. Linear Models Analysis of Incomplete Multivariate Categorical Data. (Under the direction of GARY G. KOCH.)

This research deals with experiments or surveys producing multivariate categorical data which is incomplete, in the sense that not all variables of interest are measured on every subject or element of the sample. For the most part, incompleteness is taken to arise by design, rather than by random failure of the measurement process. In these circumstances, one can often assume that counts derived from appropriate disjoint subsets of the data arise from independent multinomial distributions with linearly related parameters. Best asymptotically normal estimates of these parameters may be determined by maximizing the likelihood of the observations or by minimizing Pearson's-χ^2, Neyman's χ_1^2, or a discrimination information statistic.

The four types of estimates are studied for two-way contingency tables with supplemented margins. When both margins are supplemented, all but the minimum-χ_1^2 estimates must be determined by iterative methods. A general class of "hierarchical supplementation" designs is considered. Within this class of designs, closed form estimates are obtained from each of the four methods. The estimates are distinguished by different "kernel" averaging functions through which estimates of marginal probabilities from subsets of the data are successively combined.

For general incomplete designs, maximum likelihood equations are identical to a set of weighted least-squares equations. Neyman-χ_1^2 estimates are the first step in an iterative solution of these equations.

Estimability of parameters in terms of unique solvability of these equations is related to the rank of a matrix. When parameters are not estimable, non-linear restrictions on the estimates may provide reasonable, unique solutions.

Linear models procedures for the analysis of categorical data are generalized to include quite general incomplete data. Methods are given for estimating parameters, testing fit of a model, smoothing estimates of probabilities using the model, and testing further hypotheses within the model. Applications are demonstrated to split-plot experiments, growth curve problems for categorical data, and paired comparison procedures. The "missing data" problem and its interface with this work are examined in terms of a final example.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

PART II: APPLICATIONS

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF FIGURES

PART I

ESTIMATION, TESTING AND MODELING PROCEDURES

CHAPTER I

INTRODUCTION

1.1 Incomplete Categorical Data

Multivariate statistical methods involve inference about a population using vectors of (generally correlated) measurements from the individuals of a sample. Inference is necessary because the investigator is unwilling or unable to examine the entire population with respect to the questions of interest. Commonly, it is also impractical or inadvisable to measure every important variable on every element of a random sample. Such situations, where the variables involved are continuous, have been widely dealt with from the point of view of 'missing data' (see Afifi and Elashoff [1], [2], [3], [4], Hocking and Smith [5]) on the one hand, and from that of experimental design (e.g., Kleinbaum [6], Roy, Gnanadesikan and Srivastava [7]) on the other. Investigation of problems where the variables are categorical, involving only the assignment of an individual to one of a finite number of classes, has been much less extensive, the main papers in this connection being those of Blumenthal [8], Reinfurt [9], Hocking and Oxspring [10], and Hartley and Hocking [11].

This study concerns methods of analysis for such multivariate categorical data, incomplete in the sense that not all variables of interest are observed on every individual sampled from the population. The approach will be oriented to situations in which incompleteness

arises by design, or in the context of combining independent experiments or sample surveys. However, methods and results will under certain conditions be relevant to 'missing data' problems.

To illustrate the range of discussion, we state in terms of biomedical situations three examples of the questions to be considered.

1. A sample survey is undertaken to study the relationship between certain demographic variables and prevalence of several diseases. An independent survey taken for another purpose provides information on some of the demographic variables and one of the diseases for an additional sample. How may this information be used to improve inference about the population?

2. In many surveys, it can be expected that response errors due to respondent hostility, fatigue, or other causes will increase with the size of the questionnaire. To counteract this tendency, it may be advantageous to split the questionnaire. If an incomplete block design is applied to the set of questions, each block of questions may be administered to a different sample of individuals. Can results from such a survey be used to study the relationship of responses to sets of questions that do not occur in the same block?

3. A clinical trial is designed to compare several drugs used to control serum cholesterol levels in certain high-risk cardiac patients. For each drug, it is desired to study the month to month trend in the probability (loosely speaking) that serum cholesterol will be below a threshold value for a random patient on the drug. However, the investigators judge it impractical to examine any subject monthly. Hence, for each drug, several groups of subjects are assigned to report

at various longer intervals, such that some are scheduled for examination each month of the study. How may a trend be associated with each drug and different drugs be compared? Can data from patients who report off schedule or irregularly be included in the analysis?

Categorical data problems may be naturally formulated in terms of one or several contingency tables. This chapter will review those aspects of the theory of multidimensional tables underlying this research and the outlook from which it proceeds. Only a few topics will be considered. More complete coverage of the literature may be found in Williams and Grizzle [12] or Reinfurt [9].

1.2 Sampling Distributions for Contingency Tables

A contingency table may be defined as an indexed set of integers $T_d = \{n_{\underset{\sim}{j}}\}$, where T_d describes a set S simultaneously in terms of d partitions P_k, k=1,...,d, of S. The range of the vector $\underset{\sim}{j}' = (j_1,\ldots,j_d)$ is $R = \underset{k=1}{\overset{d}{X}} R_k$, the Cartesian product of finite sets R_k, where R_k indexes P_k; we may take $R_k = \{1,2,\ldots,J_k\}$. $n_{\underset{\sim}{j}}$ is the number of elements of S in set j_k of P_k $\forall k$, ($\forall$ meaning "for every"). For instance, consider the array in Table 1.2.1, which has been studied by a number of authors (see e.g., Grizzle [13], Berkson [14]). The table gives a classification of 218 children with behavioral problems and 147 comparable children without such problems according to three different criteria:

1) behavioral problem or not

2) birth order

3) whether or not the mother lost an earlier child in pregnancy.

Here S is the total group of 365 children, P_1, P_2 and P_3 are partitions of S corresponding to 1), 2) and 3) respectively, above. Thus

R_1 = {problem, control}, R_2 = {2, 3-4, 5+}, and R_3 = {prior losses, no prior losses}, and d=3.

Call d the dimension of the table T_d, as T_d may be naturally displayed in a d-dimensional array. Neither the dimension d nor the partitions P_k are unique in a given situation, however. In particular, for d>1, the information in T_d may also be conveyed by the Table T_1 formed from the single partition $P = \bigcap_{k=1}^{d} P_k$. The multidimensional formulation is used because of its conceptual or notational convenience.

While a contingency table is a useful aid in displaying information about a fixed set S, often our interest lies in S only because $S \subset P$, some larger 'population'. The partitions P_k are restrictions to S of corresponding P_k^*, partitions of P, and our real concern is with describing P in terms of the P_k^* rather than S in terms of the P_k. If S is chosen in some sense randomly from P, the techniques of statistical inference are applicable. The entries $n_{\underset{\sim}{j}}$ of the table T_d may be modeled as realizations of random variables whose joint distribution depends on characteristics of P and the method of sampling used to choose S. Ordinarily the latter will determine the functional form of the distribution, while the make-up of P will be reflected in certain parameters; in any case it is of interest to extract information about parameters of the distribution from the observed table. Since the distribution of a statistic calculated from T_d is usually crucially dependent on the method of sampling, statistical inference from T_d can only be attempted with reference to a particular sampling model.

Pearson [15] alerted statisticians to this fact by distinguishing three different sampling methods (1a)-c)) below for a 2×2 contingency table. It is often unrecognized that the class of possible models for

this table is much, much broader. An incomplete catalog illustrates this point.

TABLE 1.2.1

DATA ON NUMBER OF MOTHERS WITH PREVIOUS INFANT LOSSES

Birth Order		No. of Mothers With Losses	No Losses	Total
2	Problems	20	82	102
	Controls	10	54	64
3-4	Problems	26	41	67
	Controls	16	30	46
5+	Problems	27	22	49
	Controls	14	23	37

Slightly altering notation, let T_2 be the general 2×2 table in Figure 1.2.1, with random cell entries $\{\eta_{ij}\}$, marginal sums $\eta_{io} = \sum_j \eta_{ij}$, $\eta_{oj} = \sum_i \eta_{ij}$, and total sample size $\eta = \sum_i \eta_{io} = \sum_j \eta_{oj}$, where i,j = 1,2. T_2 is assumed to describe a set S sampled from a population P.

T_2

η_{11}	η_{12}	η_{10}
η_{21}	η_{22}	η_{20}
η_{01}	η_{02}	η

FIGURE 1.2.1

A RANDOM 2×2 CONTINGENCY TABLE

Series 1

Suppose S is chosen by simple random sampling from an infinite P, or from a finite P with replacement.

a) If S is of predetermined, fixed size ($\eta \equiv n$), an appropriate stochastic model is given by

$$\Pr_a\{T_2=\{n_{ij}\}\} = n! \prod_i \prod_j \frac{\pi_{ij}^{n_{ij}}}{n_{ij}!} \qquad (1.2.1)$$

where π_{ij} designates the probability that a random element from P falls in cell (i,j). Thus, $\sum_i \sum_j \pi_{ij} = 1$. This is a multinomial model.

b) A conditional model derived from a), assuming fixed $\eta_{io} \equiv n_{io}$, is

$$\Pr_b\{T_2=\{n_{ij}\}\} = \prod_i \frac{n_{io}!}{\pi_{io}^{n_{io}}} \prod_j \frac{\pi_{ij}^{n_{ij}}}{n_{ij}!}, \qquad (1.2.2)$$

which is identical to the product of two binomials $B(n_{io},\pi_{i(1)})$ where

$$\pi_{i(j)} = \frac{\pi_{ij}}{\pi_{io}}, \quad \sum_j \pi_{i(j)} = 1.$$

Hence the model is also appropriate when

i) $P = P_1 \cup P_2$

ii) every element of P_k falls in (k,1) or (k,2), k=1,2, and

iii) $S = S_1 \cup S_2$, the S_k being independent simple random samples of size n_{ko} from P_k, k=1,2.

c) A conditional model derived from b), assuming fixed $\eta_{oj} \equiv n_{oj}$, is

$$Pr_c\{T_2=\{n_{ij}\}\} = \frac{Pr_b\{T_2=\{n_{ij}\}\}}{\Sigma Pr_b\{T_2=\{n_{ij}\}\}}, \qquad (1.2.3)$$

where the sum is taken over all $\{n_{ij}\}$ with the correct marginal totals. Under the further assumption that row and column classifications are statistically independent, this reduces to

$$Pr_c^*\{T_2=\{n_{ij}\}\} = \frac{\prod_i n_{io}! \prod_j n_{oj}!}{n! \prod_i \prod_j n_{ij}!},$$

a hypergeometric probability law.

An indefinite variety of further models may be formed from the basic simple random sampling situation by conditioning on different events.

d) For instance, conditional on $\eta \equiv n$, $\eta_{11} \equiv n_{11}$ we have

$$Pr_d\{T_2=\{n_{ij}\}\} = \frac{\frac{\pi_{12}^{n_{12}}}{n_{12}!} \frac{\pi_{21}^{n_{21}}}{n_{21}!} \frac{\pi_{22}^{n_{22}}}{n_{22}!}}{\frac{(1-\pi_{11})^{n-n_{11}}}{(n-n_{11})!}}$$

e) Alternatively, sampling might proceed element by element until η_{11} reaches n_{11}, and then stop. η would be distributed according to a negative binomial law, and the appropriate model for the table is

$$Pr_e\{T_2=\{n_{ij}\}\} = n_{11}(n-1)! \prod_i \prod_j \frac{\pi_{ij}^{n_{ij}}}{n_{ij}!}.$$

Series 2

Corresponding to Series 1, we have situations in which simple random sampling without replacement is used to select S from a finite P with N elements. The appropriate models here become:

a)

$$Pr'_a\{T_2=\{n_{ij}\}\} = \frac{\prod_i \prod_j \binom{N\pi_{ij}}{n_{ij}}}{\binom{N}{n}},$$

a trivariate hypergeometric distribution.

b)

$$Pr'_b\{T_2=\{n_{ij}\}\} = \prod_i \frac{\prod_j \binom{N\pi_{ij}}{n_{ij}}}{\binom{N\pi_{io}}{n_{io}}},$$

a product of hypergeometrics $(N\pi_{io}, \pi_{i(1)})$. Note that in these finite population situations π_{ij} may be interpreted directly as a relative frequency in the population P.

c)

$$Pr'_c\{T_2=\{n_{ij}\}\} = \frac{Pr'_b\{T_2=\{n_{ij}\}\}}{\sum Pr'_b\{T_2=\{n_{ij}\}\}},$$

where as before the sum is taken over all $\{n_{ij}\}$ with the correct marginal totals. Under independence the denominator reduces to

$$\prod_i \binom{N\pi_{io}}{n_{io}} \prod_j \binom{N\pi_{oj}}{n_{oj}} / \binom{N}{n}^2,$$

but no further simplification is apparent.

d) For $n \leq N$, $n_{11} \leq N\pi_{11}$,

$$Pr'_d\{T_2=\{n_{ij}\}\} = \frac{\binom{N\pi_{12}}{n_{12}}\binom{N\pi_{21}}{n_{21}}\binom{N\pi_{22}}{n_{22}}}{\binom{N(1-\pi_{11})}{n-n_{11}}} .$$

e) For $n_{11} \leq N\pi_{11}$,

$$Pr'_e\{T_2=\{n_{ij}\}\} = \frac{\binom{n-1}{n_{11}-1}\binom{N-n}{N\pi_{11}-n_{11}}\binom{N\pi_{12}}{n_{12}}\binom{N\pi_{21}}{n_{21}}\binom{N\pi_{22}}{n_{22}}}{\binom{N}{N\pi_{11}}\binom{N(1-\pi_{11})}{n-n_{11}}}$$

Other distributions may be similarly derived by conditioning on different events.

Series 3

Now suppose T_2 is constructed by classifying, as they occur, events generated by a Poisson process with parameter λ. The process is observed beginning at some time designated as 0. If each event may be regarded as randomly selected from an infinite population of events P, then the cell counts n_{ij} at any fixed time $t>0$ are independent Poisson random variables.

a) The unconditional distribution of the table T_2 at time t is then

$$Pr_a(\lambda,t)\{T_2=\{n_{ij}\}\} = \prod_i \prod_j \frac{e^{-(\lambda t\pi_{ij})}(\lambda t\pi_{ij})^{n_{ij}}}{n_{ij}!} , \qquad (1.2.4)$$

where π_{ij} is defined as in a), Series 1.

b) Fisher [16] has shown that the conditional distributions obtained from a) by fixing

i) $\eta \equiv n$

ii) $\eta_{io} \equiv n_{io}$

iii) $\eta_{io} \equiv n_{io}$, $\eta_{oj} \equiv n_{oj}$

are respectively (1.2.1)-(1.2.3). In fact, the class of distributions obtained from (1.2.4) by conditioning on fixed $\eta=n$ is thus identical with that obtained by conditioning on (1.2.1) in the manner of 1d).

c) However, distributions similar to (1.2.4) and not contained in Series 1, may also be produced by conditioning on (1.2.4) such that η remains random. For instance, conditional on fixed $\eta_{11} \equiv n_{11}$ we have

$$Pr_c(\lambda,t)\{T_2=\{n_{ij}\}\} = \prod_{(i,j)\neq(1,1)} \frac{e^{-(\lambda t \pi_{ij})} (\lambda t \pi_{ij})^{n_{ij}}}{n_{ij}!} .$$

The sampling distributions we have classified in Series 3a)-3c) are called "conditional Poisson" models by Haberman [17], who shows that these models all obey essentially the same central limit theory as the usual multinomial models (1.2.1)-(1.2.3).

Some situations may call for observing a Poisson process and classifying events until a random time τ. If a decision to stop or continue observation is made after each event on the basis of the values of η_{ij} at that time, the situation is identical to simple random sampling with a stopping rule, as in 1e). However, other models are generated by different stopping rules, and each stopping rule can generate a number

of further conditional models. For instance, one can

d) observe a process for two years, unless $\eta>10$ after one year, in which case one stops observation.

If, in d), observation continues for two years ($\tau=2$), and in that time $\eta=12$ events are observed, one can

e) analyze the data conditional on $\eta=12$, or

f) analyze conditional on $\tau=2$, $\eta=12$.

We have not written down the models here explicitly; the examples are meant to indicate the great variety and possible complexity of models obtainable from Poisson processes.

Additional series' of models can be constructed

1) as in Series 3, by observing any random stationary stream of events
2) by considering tables which are sums of other tables, such as collapsed contingency tables
3) by considering any sampling distribution given so far in a Bayesian context, i.e., with a prior distribution on certain parameters.

The reader can likely think of other models or ways to construct some; many of the more unusual models given here are surprisingly relevant to real situations. All have obvious extensions to the general multi-dimensional contingency table.

The plethora of possible models emphasizes not only the variety of situations which may generate contingency tables, but also the abundant choice of conditional analyses which may be applied to any one such situation. In all that follows we take the view that while any model chosen must basically conform to the manner in which the data are gathered and the table is constructed, many different conditional models may be appropriate to analyses of divers aspects of the same table.

Further, many different analyses may be appropriate to the same table under the same model, depending upon theory underlying the problem, the point of view of the experimenter and the questions of interest to him. For instance, the experimenter with a multidimensional table may be concerned with studying interactions under one of any number of definitions (see e.g., Bhapkar and Koch [18], Darroch [19], Lancaster [20]), fitting linear models to functions of the cell frequencies, or using the table to discriminate between certain groups. These problems may overlap, but generally indicate different types of analysis. Some remarks relating to choices of models and methods, and some references which discuss such choices in detail, are contained in the Appendix, and in Reinfurt [9].

Such flexibility of approach is desirable but impractical without a general method for producing various custom-made analyses easily and rapidly. The appearance in recent years of several methods which apply to any product-multinomial table (analogous to 1b)) is the result of an increasing trend towards studying the broad categorical data problem rather than its manifestations in tables of particular types and sizes.

1.3 Systems of Estimation and Hypothesis Testing

This trend is in some sense a reaction to the earlier tendency to view problems in continuous and categorical data as unrelated. It is not immediately apparent that different questions must be asked of continuous and categorical data arising from otherwise similar experimental situations, or that the two types of data should be treated fundamentally differently. Nevertheless, prior to Pearson's paper [15]

cited, the rapid development of normal theory for continuous data and the intriguing combinatorial problems related to contingency tables led statisticians dealing with each type of data down divergent paths of research. The result was on the one hand the general theory of multivariate linear models, and on the other a multitude of techniques developed solely to treat contingency tables of specific sizes, or to answer questions raised only by tables arising in a specific experimental context. Procedures for different size tables often had little in common with one another, or with methods for analyzing similarly obtained continuous data.

The difficulties that this situation produced stimulated a search for unifying principles for categorical data analysis that relate it clearly to multivariate analysis of continuous data. Advances in this direction were made possible by the work of S. N. Roy and his students (Bhapkar, Diamond, Kastenbaum, Mitra and Sathe) who

i) developed a terminology appropriate to product-multinomial models in an experimental context,

ii) discussed hypotheses relevant to the models they distinguished, and

iii) related contingency table problems to the analogous continuous data situations.

The papers of Bhapkar [21], [22] and Bhapkar and Koch [18], [23] reveal the flavor of this work.

We now turn to the three general methods for handling product-multinomial contingency tables which have recently stimulated the most interest. Each is based on a different member of the class of BAN (best asymptotically normal) estimators defined by Neyman [24] (see Appendix).

Suppose, for each $\underset{\sim}{i}\in I$, we observe $n_{\underset{\sim}{i}\underset{\sim}{j}}$, $\forall\, \underset{\sim}{j}\in R_{\underset{\sim}{i}}$, where

a) the $n_{\underset{\sim}{i}\underset{\sim}{j}}$ are realizations of random variables $\eta_{\underset{\sim}{i}\underset{\sim}{j}}$ and

b) for fixed $\underset{\sim}{i}\in I$, the $\eta_{\underset{\sim}{i}\underset{\sim}{j}}$ have a joint multinomial distribution, sample size $n_{\underset{\sim}{i}}$ and parameters $\pi_{\underset{\sim}{i}\underset{\sim}{j}}$, $\forall\, \underset{\sim}{j}\in R_{\underset{\sim}{i}}$. Let $p_{\underset{\sim}{i}\underset{\sim}{j}} = n_{\underset{\sim}{i}\underset{\sim}{j}}/n_{\underset{\sim}{i}}$.

Thus $\sum_{\underset{\sim}{j}\in R_{\underset{\sim}{i}}} p_{\underset{\sim}{i}\underset{\sim}{j}} = \sum_{\underset{\sim}{j}\in R_{\underset{\sim}{i}}} \pi_{\underset{\sim}{i}\underset{\sim}{j}} = 1$, $\forall\, \underset{\sim}{i}\in I$. To estimate the $\pi_{\underset{\sim}{i}\underset{\sim}{j}}$ subject to further restrictions R, we may choose estimates $\hat{\pi}_{\underset{\sim}{i}\underset{\sim}{j}}$ as values of $\pi_{\underset{\sim}{i}\underset{\sim}{j}}$ which

1. maximize $\ln L^* = \ln \prod_{\underset{\sim}{i}\in I} \prod_{\underset{\sim}{j}\in R_{\underset{\sim}{i}}} \pi_{\underset{\sim}{i}\underset{\sim}{j}}^{n_{\underset{\sim}{i}\underset{\sim}{j}}} =$

$$\sum_{\underset{\sim}{i}} \sum_{\underset{\sim}{j}} n_{\underset{\sim}{i}\underset{\sim}{j}} \ln \pi_{\underset{\sim}{i}\underset{\sim}{j}} \tag{1.3.1}$$

2. minimize $\chi_I^2 = 2 \sum_{\underset{\sim}{i}} n_{\underset{\sim}{i}} \sum_{\underset{\sim}{j}} \pi_{\underset{\sim}{i}\underset{\sim}{j}} \ln (\pi_{\underset{\sim}{i}\underset{\sim}{j}}/p_{\underset{\sim}{i}\underset{\sim}{j}})$ (1.3.2)

3. minimize $\chi_1^2 = \sum_{\underset{\sim}{i}} n_{\underset{\sim}{i}} \sum_{\underset{\sim}{j}} \frac{(p_{\underset{\sim}{i}\underset{\sim}{j}} - \pi_{\underset{\sim}{i}\underset{\sim}{j}})^2}{p_{\underset{\sim}{i}\underset{\sim}{j}}} =$

$$\sum_{\underset{\sim}{i}} \sum_{\underset{\sim}{j}} \frac{(n_{\underset{\sim}{i}\underset{\sim}{j}} - n_{\underset{\sim}{i}} \pi_{\underset{\sim}{i}\underset{\sim}{j}})^2}{n_{\underset{\sim}{i}\underset{\sim}{j}}} \tag{1.3.3}$$

4. minimize $\chi^2 = \sum_{\underset{\sim}{i}} n_{\underset{\sim}{i}} \sum_{\underset{\sim}{j}} \frac{(p_{\underset{\sim}{i}\underset{\sim}{j}} - \pi_{\underset{\sim}{i}\underset{\sim}{j}})^2}{\pi_{\underset{\sim}{i}\underset{\sim}{j}}} =$

$$\sum_{\underset{\sim}{i}} \sum_{\underset{\sim}{j}} \frac{(n_{\underset{\sim}{i}\underset{\sim}{j}} - n_{\underset{\sim}{i}} \pi_{\underset{\sim}{i}\underset{\sim}{j}})^2}{n_{\underset{\sim}{i}} \pi_{\underset{\sim}{i}\underset{\sim}{j}}} \tag{1.3.4}$$

subject to the restrictions R. Such estimates are called respectively

1. maximum likelihood

2. minimum discrimination information

3. minimum modified - (or Neyman's) chi-square

4. minimum (Pearson's) chi-square estimates.

For a wide class of R (see the Appendix) any of these estimates are BAN. Further, if the $\pi_{\underset{\sim}{ij}}$ satisfy the restrictions R, all statistics obtained by substituting any set of BAN estimates into any of the expressions (1.3.2)-(1.3.4) have the same limiting chi-square distribution as $\sum_{\underset{\sim}{i}} n_{\underset{\sim}{i}} = n \to \infty$ with $n_{\underset{\sim}{i}}/n \to Q_{\underset{\sim}{i}} > 0$. Thus, any of these statistics may be used to test many hypotheses about the $\pi_{\underset{\sim}{ij}}$. Sometimes R may be expressed as $\pi_{\underset{\sim}{ij}} = f_{\underset{\sim}{ij}}(\underset{\sim}{\theta})$ $\forall\, \underset{\sim}{i},\underset{\sim}{j}$, where the $f_{\underset{\sim}{ij}}$ are known functions and $\underset{\nu\times 1}{\underset{\sim}{\theta}}$ is a vector of unknown parameters. If R is true, and the $f_{\underset{\sim}{ij}}$ are sufficiently regular (see Chapter IV), estimates of $\underset{\sim}{\theta}$ determined using any of (1.3.1)-(1.3.4) are BAN. Further, the difference between (1.3.2) calculated under R and under R^*: R true with $\underset{\sim}{G}(\underset{\sim}{\theta}) = \underset{\sim}{0}$, for some vector of $\underset{k\times 1}{\underset{\sim}{G}}$ of linearly independent functions, has a limiting chi-square (ν-k) distribution when R^* holds. Similar differences calculated using (1.3.3)-(1.3.4) have the same asymptotic distribution. Any of them can be used to test $\underset{\sim}{G}(\underset{\sim}{\theta}) = \underset{\sim}{0}$.

Minimum chi-square (4) has been included here mainly for historical interest. The other three estimators each form the basis for a system for analyzing contingency tables, viz.,

1. the <u>maximum likelihood</u> approach (ref. Goodman [25], [26], Bishop [27], Haberman [17]) which uses maximum likelihood estimation to fit multiplicative models to cell expectations, and tests hierarchical sequences of hypotheses using likelihood ratio tests to simplify the model

(this test statistic is obtained by substituting maximum likelihood estimates (m.l.e.'s) of $\pi_{\underset{\sim}{ij}}$ into (1.3.2));

2. the minimum discrimination information approach (ref. Kullback [28], Ku, Varner and Kullback [29], Good [30]), similarly oriented, but using estimates which minimize (1.3.2), and the resulting test statistics (1.3.2).

Both these methods exploit formal relationships with the analysis of variance. A system flavored more of multiple regression uses

3. the Neyman-χ^2_1 approach (ref. Grizzle, Starmer and Koch [31], Koch and Reinfurt [32]), which fits linear models to arbitrary regular functions of the cell probabilities using estimates and tests derived from (1.3.3) and its generalization (see the Appendix) stemming from large-sample theory.

Each of these approaches has proven successful in handling a wide variety of data and models, each is adaptable to computer implementation; since in most ways the three are asymptotically equivalent, choice between them is often primarily a matter of taste. However, with the exception of Reinfurt [9], authors using these methods have considered only single multidimensional contingency tables, or 'incomplete' tables (Bishop and Fienberg [33], Williams and Grizzle [12], Mantel [34]), involving a priori zero cell entries. Many problems like those in Section 1.1 are not treatable in this way, but may be readily handled by regarding the data as several contingency tables, independently sampled, from distributions with related parameters. Any of the three general systems may be used to treat incomplete data problems formulated in this way. An early use of the basic methodology, in the context of combining experiments, appears in Mather [35].

In subsequent chapters we shall determine the estimators defined in 1-4 for some simple incomplete data situations. The comparison of estimators will shed some light on the way the different estimators combine data from several tables. The minimum-χ^2_1 system (3) will then be extended, in the manner suggested by Reinfurt [9], to handle a general class of incomplete data problems. Some particularly interesting applications will be discussed in Part II.

1.4 Initial Clarifications

We conclude preliminaries with some remarks on the limitations of this work, hoping to deal with some points of controversy before they arise.

Henceforth we confine ourselves to categorical data arising from a multinomial or product-multinomial model. However, this is not an essential restriction on the methods given, which may be adapted with appropriate modifications of the large sample theory to data arising from many of the models in Section 1.2. Johnson and Koch [36] have applied Neyman-χ^2_1 methods for complete data (Grizzle, Starmer and Koch [31]) to finite population sampling (hypergeometric models) and similar changes may be made for complete or incomplete data from other sampling models with a multivariate normal central limit theory.

Everything that follows is dependent on this central limit theory; that is, throughout we use asymptotic approximations rather than exact distributional theory. It may be useful to describe precisely where such approximations come in.

a) Estimation of parameters using (1.3.1)-(1.3.4) is justified by the fact that such estimates are BAN, that is, by properties of their

limiting distribution. Little is known about small sample properties of these estimators (see Berkson [14, 37], Odoroff [38] for small sample comparisons).

b) All significance tests are based on approximating the distribution of a statistic by its limiting chi-square distribution. The usefulness of a test is dependent on the accuracy of this approximation. A primitive rule of thumb might require, in analyzing a product multinomial model by these methods, that each component multinomial be 'well' approximated by its limiting multivariate normal density. The situation may be very complex, and a number of simulation studies would be useful in this area.

c) Frequently it will be of interest to fit linear models of form

$$\underset{\sim}{f} = X\underset{\sim}{\beta} \tag{1.4.1}$$

where $\underset{\sim}{f}$ is a vector of functions of the $\pi_{\underset{\sim}{ij}}$, $\underset{\sim}{\beta}$ a vector of unknown parameters, and X a known matrix. The Neyman-χ^2_1 methodology does this by applying weighted least-squares to the model

$$E\underset{\sim}{f}^* = X\underset{\sim}{\beta} \tag{1.4.2}$$

where

$$\underset{\sim}{f}^* = \underset{\sim}{f}\Big|_{\pi_{\underset{\sim}{ij}} = \hat{\pi}_{\underset{\sim}{ij}}},$$

and $\hat{\pi}_{\underset{\sim}{ij}}$ is some consistent estimate of $\pi_{\underset{\sim}{ij}}$. If

$$\left\| \underset{\sim}{f} - E\underset{\sim}{f}^* \right\| = O\, n^{-\frac{1}{2}} \tag{1.4.3}$$

as $n \to \infty$, $n_{\underset{\sim}{i}}/n \to Q_{\underset{\sim}{i}}$, then $\hat{\underset{\sim}{\beta}}$ obtained in this way from (1.4.2) is BAN for $\underset{\sim}{\beta}$ under the sequence (1.4.2) and also under (1.4.1), and statistics

based on $\underset{\sim}{\beta}$ have the same limiting distribution under either (see e.g., Wald [39], Davidson and Lever [40]). Thus, a linear model may be used which is only appropriate asymptotically, but $\underset{\sim}{f}$ and $\underset{\sim\sim}{\pi_{ij}}$ must be chosen to satisfy (1.4.3).

From amongst the several possible asymptotic approaches in Section 1.3, the Neyman-χ^2_1 methods have been selected for full generalization to incomplete data situations. Other methods may conceivably generalize equally well. The χ^2_1 procedures do, however, have two properties not mentioned above that some researchers may find particularly attractive. The first is that estimation is accomplished without recourse to iterative procedures, and involves only solving linear equations (matrix inversion). Maximum likelihood methods, for instance, require iteration for many complete data problems, and procedures for incomplete data can require more complex iterative procedures. (On the other hand (see Chapter IV), m.l.e.'s for incomplete data problems often can be obtained simply by iterating certain portions of the χ^2_1 computations.) The second property involves tests of hypotheses about parameters. Speaking broadly, a test of the hypothesis H_o about some parameter vector $\underset{\sim}{\theta}$ is constructed in the χ^2_1 approach by referring estimates of some functions of $\underset{\sim}{\theta}$ to an estimate $\hat{V}$ of their asymptotic covariance matrix V. $\hat{V}$ is constructed so that $\hat{V} \xrightarrow{p} V$ ($\hat{V}$ is consistent) whether H_o is true or false. Some of the pitfalls of hierarchical analyses are avoided in this way; in analogy to analysis of variance, one essentially uses an estimate of pure error in all tests, rather than pooling other sums of squares with the error term after certain non-significant test results. In this sense, the χ^2_1 tests represent conservative statistical practice.

It will be convenient to divide incomplete categorical data designs into two subclasses. Analogous to the first example of Section 1.1 are cases where samples which each measure a different proper subset of the variables of interest are obtained to supplement a sample for which information about all variables is known. The supplemental samples may be obtained because of greater interest in some variables than others, cost considerations, or a combination of the two. These situations are analogous to those which arise in the context of augmented fractional factorial designs as discussed by Box and Wilson [41], Box [42], John [43], Gaylor and Merrill [44], and sequences of fractional factorial designs as discussed by Daniel [45] and Addelman [46]. The analogy is particularly close when one considers categorical data mixed models or split-plot experiments as described by Koch and Reinfurt [47]. Examples 2 and 3 of Section 1.1 belong to the other subclass, problems for which a core sample of data on all variables is not obtained, which relate to certain incomplete block and fractional factorial type split-plot multivariate designs described by Roy, Gnanadesikan, and Srivastava [7]. Members of the first class will be called "supplemented margins designs" or "contingency tables with supplemented margins", following Reinfurt [9]; those in the second will be termed "incomplete block measurement designs".

A third class of problems essentially mimics supplemented margins designs by producing similar data arrays. These are "missing data" problems, as formulated in the papers of Affifi and Elashoff [1], [2], [3], [4] and Blumenthal [8]. One attempts to observe the

same multivariate random variable on each of a collection of individuals, but the measurement or observation process is flawed and subject to some sort of random failure, so that various components of the observation vector may not be obtainable for some (random) individuals. The relationship between the analysis of supplemented margins designs and that of similar data occurring due to such random incomplete response will be briefly discussed in Chapter X.

Note the difference between all of these situations and that corresponding to "incomplete contingency tables" mentioned in Section 1.3. For the latter, every element of the sample is classified with respect to every variable, hence the data is "complete" as the term is used here. However, certain cell counts from the resulting contingency table are known to be zero a priori, that is, non-random. Hence, the remaining random data may be regarded as forming an "incomplete contingency table", missing certain cells, and a stochastic model for the data is then incomplete in the sense that it does not contain the full set of parameters appropriate to a contingency table of the given size. Such "incomplete contingency tables" will not be considered here.

Finally, estimating cell probabilities in contingency tables with supplemented margins must be distinguished from a similar problem treated by El-Badry and Stephan [48], Friedlander [49], Ireland and Kullback [50], and others. They consider estimation of cell probabilities using a contingency table formed from a sample, when the marginal probabilities or population totals are known. Essentially, one adjusts the estimates of unknown margins and population interactions from the sample to compensate for the deviation of some observed margins in the table

from their known population values. Maximum likelihood estimates may be produced using a form of "iterative proportional fitting" (IPF) as discussed by Fienberg [51]. In contrast, contingency tables with supplemented margins have additional data on the margins which is random; hence the table is not adjusted to fit the margins, but the table data and marginal data are combined to estimate all the parameters of interest. Iterative proportional fitting produces m.l.e.'s of cell probabilities if one fits the m.l.e.'s of the supplemented margins to the table (Birch [52]), but these may be as difficult to determine in their own right as the m.l.e.'s of the cell probabilities are to start with (Reinfurt [9]). In the next two chapters we consider the estimation of cell probabilities for data from some simply structured supplemented margins designs.

CHAPTER II

TWO-WAY TABLES WITH SUPPLEMENTED MARGINS

2.1 Estimation with One Supplemented Margin

The simplest supplemented margin design produces data in the form of the array T_2^* shown in Figure 2.1.1, a two-way table with one margin supplemented. The notation follows Reinfurt [9]. To the left of the dotted line is an ordinary r×c table with cell counts n_{ij}, where i=1,...,r and j=1,...,c, arising from a multinomial model with corresponding parameters π_{ij} and sample size n_{oo}. To the right is an ordinary one-way r×1 table with cell counts n_{i*}, i=1,...,r arising from a multinomial population with corresponding parameters $\pi_{io} = \sum_{j=1}^{c} \pi_{ij}$ and sample size n_{o*}. If the two tables are statistically independent, then

$$\Pr\{T_2^*\} = \left[n_{oo}! \prod_{i=1}^{r} \prod_{j=1}^{c} \frac{\pi_{ij}^{n_{ij}}}{n_{ij}!} \right] \left[n_{io}! \prod_{i=1}^{r} \frac{\pi_{io}^{n_{io}}}{n_{io}!} \right]. \qquad (2.1.1)$$

For instance, the central (core) 2×2 table might represent classification of a simple random sample of n_{oo} married women from a well-defined population along dimensions relating to family size and use of a birth control method. The supplementary data might be derived from another survey in which a simple random sample of n_{o*} married women from the same population was queried only with respect to the use of birth control procedures.

T_2^*

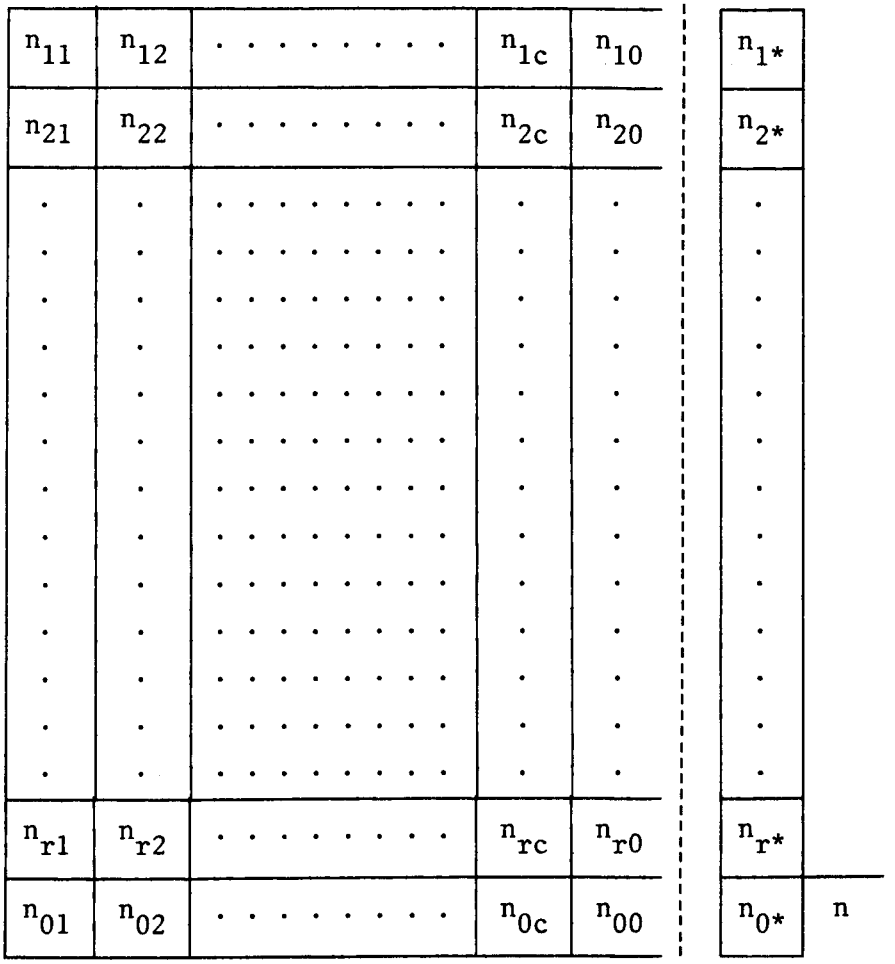


FIGURE 2.1.1

AN r×c TABLE WITH SUPPLEMENTED ROW MARGINS

Assuming that all π_{ij}, $n_{ij}>0$, the π_{ij} can be estimated using any of (1.3.1)-(1.3.4).

a) Maximum likelihood estimation

Reinfurt [9] and Hocking and Oxspring [10] have independently shown that the m.l.e. of π_{ij} is the natural estimator

$$\hat{\pi}_{ij}^{(1)} = \left(\frac{n_{ij}}{n_{io}}\right)\left(\frac{n_{io} + n_{i*}}{n_{oo} + n_{o*}}\right). \tag{2.1.2}$$

If $p_{ij} = n_{ij}/n_{oo}$, $p_{io} = \sum_{j=1}^{c} p_{ij}$, $p_{i*} = n_{i*}/n_{o*}$, and $\omega = n_{oo}/(n_{oo}+n_{o*}) = n_{oo}/n$, then $\hat{\pi}_{ij}^{(1)}$ can be rewritten as

$$\left(\frac{p_{ij}}{p_{io}}\right)(\omega p_{io} + (1-\omega)p_{i*}).$$

$\hat{\pi}_{ij}^{(1)}$ is thus the product of an estimate of π_{io}, viz., the weighted mean of p_{io} and p_{i*}, with a consistent estimate of Pr{column j|row i} = π_{ij}/π_{io} from the fully classified data. (1.3.2)-(1.3.4) also produce estimates of this form, but using different terms to estimate π_{io}, that is, different ways of combining the partially and fully classified data.

b) Minimum discrimination information estimation

From (1.3.2), these estimates $\hat{\pi}_{ij}^{(2)}$ minimize

$$\chi_I^2 = 2n_{oo}\sum_{i=1}^{r}\sum_{j=1}^{c}\pi_{ij}\ln(\pi_{ij}/p_{ij}) + 2n_{o*}\sum_{i=1}^{r}\pi_{io}\ln(\pi_{io}/p_{i*}) \tag{2.1.3}$$

subject to $\sum_{i=1}^{r} \sum_{j=1}^{c} \hat{\pi}_{ij}^{(2)} = 1$. Using a Lagrangian multiplier λ, subtracting the corresponding term $\lambda(\sum_{i=1}^{r} \sum_{j=1}^{c} \pi_{ij} - 1)$ from (2.1.3), and differentiating with respect to the π_{ij} shows that the $\hat{\pi}_{ij}^{(2)}$ and λ are roots of the set of equations

$$2(n_{oo} \ln \frac{\pi_{ij}}{p_{ij}} + n_{o*} \ln \frac{\pi_{io}}{p_{i*}} + n) = \lambda$$

or equivalently, satisfy

$$(\frac{\pi_{ij}}{p_{ij}})^{n_{oo}} (\frac{\pi_{io}}{p_{i*}})^{n_{o*}} = \text{constant } \forall i,j. \tag{2.1.4}$$

Hence, for each i, $\hat{\pi}_{ij}^{(2)}/p_{ij}$ is constant $\forall_j$. Since $\sum_{j=1}^{c} \hat{\pi}_{ij}^{(2)} = \hat{\pi}_{io}^{(2)}$, this implies that $\hat{\pi}_{ij}^{(2)} = (p_{ij}/p_{io})\ \hat{\pi}_{io}^{(2)}$, and that $\hat{\pi}_{io}^{(2)}$ must satisfy

$$\frac{\pi_{io}}{p_{io}^{\omega} p_{i*}^{1-\omega}} = k_1 \text{ constant } \forall i. \tag{2.1.5}$$

Since $\sum_{\ell=1}^{r} \hat{\pi}_{\ell o}^{(2)} = 1$, (2.1.5) gives $k_1 = (\sum_{\ell=1}^{r} p_{\ell o}^{\omega} p_{\ell *}^{1-\omega})^{-1}$, so that

$$\hat{\pi}_{io}^{(2)} = (p_{io}^{\omega} p_{i*}^{1-\omega}) (\sum_{\ell=1}^{r} p_{\ell o}^{\omega} p_{\ell *}^{1-\omega})^{-1} \tag{2.1.6}$$

and

$$\hat{\pi}_{ij}^{(2)} = (\frac{p_{ij}}{p_{io}}) (p_{io}^{\omega} p_{i*}^{1-\omega}) (\sum_{\ell=1}^{r} p_{\ell o}^{\omega} p_{\ell *}^{1-\omega})^{-1}.$$

Thus, while $\hat{\pi}_{ij}^{(1)}$ uses a weighted arithmetic mean of p_{io} and p_{i*} to

estimate π_{io}, $\hat{\pi}_{ij}^{(2)}$ uses the (normalized) weighted geometric mean with identical weights.

c) Minimum (Neyman) - χ_1^2 estimation

These estimates $\hat{\pi}_{ij}^{(3)}$ minimize

$$\chi_1^2 = n_{oo} \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(p_{ij}-\pi_{ij})^2}{p_{ij}} + n_{o*} \sum_{i=1}^{r} \frac{(p_{i*}-\pi_{io})^2}{p_{i*}}$$

with respect to the π_{ij}, subject to $\sum_{i=1}^{r} \sum_{j=1}^{c} \hat{\pi}_{ij}^{(3)} = 1$. Proceeding as in b), one finds that the $\hat{\pi}_{ij}^{(3)}$ must satisfy the equations

$$n_{oo} \frac{\pi_{ij}}{p_{ij}} + n_{o*} \frac{\pi_{io}}{p_{i*}} = \text{constant } \forall i,j. \tag{2.1.7}$$

Then, as in b), $\hat{\pi}_{ij}^{(3)} = \frac{p_{ij}}{p_{io}} \hat{\pi}_{io}^{(3)}$, and (2.1.7) reduces to

$$\left(\frac{\omega}{p_{io}} + \frac{(1-\omega)}{p_{i*}}\right)\pi_{io} = k_2 \text{ constant } \forall i. \tag{2.1.8}$$

Since

$$\sum_{\ell=1}^{r} \pi_{\ell o} = 1, \quad k_2 = \left(\sum_{\ell=1}^{r} \left(\frac{\omega}{p_{\ell o}} + \frac{(1-\omega)}{p_{\ell *}}\right)^{-1}\right)^{-1},$$

and

$$\hat{\pi}_{io}^{(3)} = \left(\frac{\omega}{p_{io}} + \frac{(1-\omega)}{p_{i*}}\right)^{-1} \left(\sum_{\ell=1}^{r} \left(\frac{\omega}{p_{\ell o}} + \frac{(1-\omega)}{p_{\ell *}}\right)^{-1}\right)^{-1}.$$

Hence,

$$\hat{\pi}_{ij}^{(3)} = \left(\frac{p_{ij}}{p_{io}}\right) \left(\frac{\omega}{p_{io}} + \frac{(1-\omega)}{p_{i*}}\right)^{-1} \left(\sum_{\ell=1}^{r} \left(\frac{\omega}{p_{\ell o}} + \frac{(1-\omega)}{p_{\ell *}}\right)^{-1}\right)^{-1}.$$

So $\hat{\pi}_{ij}^{(3)}$ uses the (normalized) weighted harmonic mean of p_{io} and p_{i*} to estimate π_{io}.

d) Minimum (Pearson) - χ^2 estimation

Finally, from (1.3.4) and following c), the estimates $\hat{\pi}_{ij}^{(4)}$ which minimize

$$\chi^2 = n_{oo} \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(p_{ij}-\pi_{ij})^2}{\pi_{ij}} + n_{o*} \sum_{i=1}^{r} \frac{(p_{io}-\pi_{io})^2}{\pi_{io}}$$

subject to $\sum_{i=1}^{r} \sum_{j=1}^{c} \hat{\pi}_{ij}^{(4)}=1$ must satisfy the sets of equations

$$\pi_{ij} = \frac{p_{ij}}{p_{io}} \pi_{io} \tag{2.1.9}$$

and

$$(\omega p_{io}^2 + (1-\omega)p_{i*}^2)^{\frac{1}{2}}/\pi_{io} = k_3 \text{ constant } \forall\, i,j, \tag{2.1.10}$$

so that

$$k_3 = \sum_{\ell=1}^{r} (\omega p_{\ell o}^2 + (1-\omega)p_{\ell *}^2)^{\frac{1}{2}},$$

$$\hat{\pi}_{io}^{(4)} = (\omega p_{io}^2 + (1-\omega)p_{i*}^2)^{\frac{1}{2}} \left(\sum_{\ell=1}^{r} (\omega p_{\ell o}^2 + (1-\omega)p_{\ell *}^2)^{\frac{1}{2}} \right)^{-1}$$

and

$$\hat{\pi}_{ij}^{(4)} = \frac{p_{ij}}{p_{io}} (\omega p_{io}^2 + (1-\omega)p_{i*}^2)^{\frac{1}{2}} \left(\sum_{\ell=1}^{r} (\omega p_{\ell o}^2 + (1-\omega)p_{\ell *}^2)^{\frac{1}{2}} \right)^{-1}.$$

This fourth type of estimate uses a fourth type of weighted average of p_{io} and p_{i*} to estimate π_{io}.

2.2 Estimation with Both Margins Supplemented

Now suppose one has data in the slightly more complicated form of the array T_2^{**} shown in Figure 2.2.1. The portion of T_2^{**} above the horizontal dotted line represents data identical to that discussed in the last section. Below the line is shown an additional $1\times c$ table with cell counts n_{*j}, $j=1,\ldots,c$, arising from a multinomial population with corresponding parameters $\pi_{oj} = \sum_{i=1}^{r}\pi_{ij}$ and sample size n_{*o}. The two tables above the line and this third table are all assumed to be statistically independent, so that

$$\Pr\{T_2^{**}\} = \left(n_{oo}! \prod_{i=1}^{r}\prod_{j=1}^{c}\frac{\pi_{ij}^{n_{ij}}}{n_{ij}!}\right)\left(n_{o*}! \prod_{i=1}^{r}\frac{\pi_{io}^{n_{i*}}}{n_{i*}!}\right)\left(n_{*o}! \prod_{j=1}^{c}\frac{\pi_{oj}^{n_{*j}}}{n_{*j}!}\right).$$

Data of this form might arise as in the example of the previous section, if a third sample of n_{*o} married women had been queried, say in a population study, only with respect to family size. Estimation of the π_{ij} by (1.3.1)-(1.3.4) is, however, considerably more difficult due to the addition of this further sample. Again it is assumed that all π_{ij}, $n_{ij}>0$. Let $n = n_{oo} + n_{o*} + n_{*o}$.

a) Maximum likelihood

No closed form expression is known for the m.l.e.'s $\tilde{\pi}_{ij}^{(1)}$ in this situation. Reinfurt [9] has shown that the $\tilde{\pi}_{ij}^{(1)}$ are roots of two separate sets of equations. The first set

T_2^{**}

n_{11}	n_{12}	. . . . .	n_{1c}	n_{1o}	n_{1*}
n_{21}	n_{22}	. . . . .	n_{2c}	n_{2o}	n_{2*}
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
.	.	. . . . .	.	.	.
n_{r1}	n_{r2}	. . . . .	n_{rc}	n_{ro}	n_{r*}
n_{o1}	n_{o2}	. . . . .	n_{oc}	n_{oo}	n_{o*}
n_{*1}	n_{*2}	. . . . .	n_{*c}	n_{*o}	n

FIGURE 2.2.1

AN r×c TABLE WITH BOTH MARGINS SUPPLEMENTED

$$\pi_{ij} = \frac{n_{ij}}{n - \frac{n_{i*}}{\pi_{io}} - \frac{n_{*j}}{\pi_{oj}}} \qquad (2.2.1)$$

is produced by using the method of Lagrange to maximize the log-likelihood function. The second set is generated by a weighted least squares algorithm and is more complex. For r=c=2, these equations become

$$\pi_{ij} = (1/N_{\tilde{\gamma}}) \left\{ n_{ij} + n_{i*} \frac{\pi_{ij}}{\pi_{io}} + n_{*j} \frac{\pi_{ij}}{\pi_{oj}} + \frac{n_{o*}}{n_{oo}} \left(\frac{n_{ij}\pi_{ij}' - n_{ij}'\pi_{ij}}{\pi_{io}} \right) \right.$$

$$+ \frac{n_{*o}}{n_{oo}} \left(\frac{n_{ij}\pi_{i'j} - n_{i'j}\pi_{ij}}{\pi_{oj}} \right) + \tilde{\gamma} \left\{ n_{ij} + \frac{\left(\frac{n_{oo}}{n_{*o}} n_{*j} - n_{i'j} \right)}{\pi_{i'j}} \right. \qquad (2.2.2)$$

$$\left. \left. + \frac{\left(\frac{n_{oo}}{n_{o*}} n_{i*} - n_{ij'} \right)}{\pi_{ij'}} + \frac{\left(n_{oo} - \frac{n_{oo}}{n_{o*}} n_{i'*} - \frac{n_{oo}}{n_{*o}} n_{*j'} + n_{i'j'} \right)}{\pi_{i'j'}} \right\} \right\}$$

where i,j=1,2; i'≠i, j'≠j;

$$\tilde{\gamma} = \frac{n_{o*}n_{*o}}{n_{oo}^2} \; \frac{\prod_{i=1}^{2} \prod_{j=1}^{2} \pi_{ij}}{\left[\prod_{i=1}^{2} \pi_{io} \right] \left[\prod_{j=1}^{2} \pi_{oj} \right]} \; ;$$

and $N_{\tilde{\gamma}} = n + n_{oo}\tilde{\gamma} \sum_{i=1}^{2} \sum_{j=1}^{2} \pi_{ij}^{-1}$. The general form of (2.2.2) will be discussed in Chapter IV. For the present, reference to (2.2.2) will be taken to include its generalization to r×c tables.

Hocking and Oxspring [10] have demonstrated the equivalence of (2.2.1) and (2.2.2); that is, (2.2.2) is simply an alternative form of the m.l. equations. They note that the log-likelihood is concave in the relevant region A: $\pi_{ij}>0$, $\sum_i \sum_j \pi_{ij} = 1$, and that provided its contours are bounded ($\pi_{ij} \geq k > 0$, say, insures this) the m.l.e.'s may be located using any strict ascent method of function maximization. Alternatively, concavity insures that any solution of (2.2.1) or (2.2.2) in A is unique and maximum likelihood. Either (2.2.1) or (2.2.2) may be iterated from the starting values $\pi_{ij} = p_{ij} = n_{ij}/n_{oo}$ as suggested by these authors. Convergence has been observed for a large number of examples, considerably more rapid using (2.2.2) than with (2.2.1). The iteration process may often be shortened considerably by using a better starting point, e.g.,

$$\pi^*_{io} = \omega_R p_{io} + (1-\omega_R)p_{i*}$$
$$\pi^*_{oj} = \omega_C p_{oj} + (1-\omega_C)p_{*j}$$
$$\pi^*_{ij} = \frac{n_{ij}}{2}\left(\frac{\pi^*_{io}}{n_{io}} + \frac{\pi^*_{oj}}{n_{oj}}\right)$$

with

$$p_{i*} = \frac{n_{i*}}{n_{o*}}, \quad p_{*j} = \frac{n_{*j}}{n_{*o}}, \quad \omega_R = \frac{n_{oo}}{n_{oo}+n_{o*}}, \quad \omega_C = \frac{n_{oo}}{n_{oo}+n_{*o}}.$$

b) Minimum χ^2_I

The m.d.i. estimates $\tilde{\pi}^{(2)}_{ij}$ also lack a closed form solution. Here

$$\chi^2_I = 2n_{oo}\sum_{i=1}^{r}\sum_{j=1}^{c}\pi_{ij}\ell n\left(\frac{\pi_{ij}}{p_{ij}}\right) + 2n_{o*}\sum_{i=1}^{r}\pi_{io}\ell n\left(\frac{\pi_{io}}{p_{i*}}\right) + 2n_{*o}\sum_{j=1}^{c}\pi_{oj}\ell n\left(\frac{\pi_{oj}}{p_{*j}}\right). \tag{2.2.3}$$

It is easily seen that χ_I^2 is convex, so that the $\tilde{\pi}_{ij}^{(2)}$ may be found using any method of strict descent on χ_I^2. However, we will propose a simple iterative scheme for solution of the m.d.i. equations.

Proceeding as in 2.1 shows that the $\tilde{\pi}_{ij}^{(2)}$ must satisfy

$$\left(\frac{\pi_{ij}}{p_{ij}}\right)^{n_{oo}}\left(\frac{\pi_{io}}{p_{i*}}\right)^{n_{o*}}\left(\frac{\pi_{oj}}{p_{*j}}\right)^{n_{*o}} = \text{constant } \forall i,j. \tag{2.2.4}$$

(2.2.4) implies, for each i,

$$\left(\frac{\pi_{ij}}{p_{ij}}\right)^{n_{oo}}\left(\frac{\pi_{oj}}{p_{*j}}\right)^{n_{*o}} = k_4(i) \text{ constant } \forall j \tag{2.2.5}$$

and, for each j,

$$\left(\frac{\pi_{ij}}{p_{ij}}\right)^{n_{oo}}\left(\frac{\pi_{io}}{p_{i*}}\right)^{n_{o*}} = k_5(j) \text{ constant } \forall_i. \tag{2.2.6}$$

From (2.2.5),

$$\frac{\pi_{ij}}{\pi_{ic}} = \left(\frac{p_{ij}}{p_{ic}}\right)\left(\frac{\pi_{oc}p_{*j}}{\pi_{oj}p_{*c}}\right)^{W_C}, \text{ where } W_C = \frac{n_{*o}}{n_{oo}}.$$

Summing over j,

$$\frac{\pi_{io}}{\pi_{ic}} = p_{ic}^{-1}\left(\frac{\pi_{oc}}{p_{*c}}\right)^{W_C} \sum_{j=1}^{c} p_{ij}\left(\frac{p_{*j}}{\pi_{oj}}\right)^{W_C},$$

so that

$$\tilde{\pi}_{ij}^{(2)} = \tilde{\pi}_{io}^{(2)}\left(p_{ij}\left(\frac{p_{*j}}{\tilde{\pi}_{oj}^{(2)}}\right)^{W_C}\right)\left\{\sum_{\ell=1}^{c} p_{i\ell}\left(\frac{p_{*\ell}}{\tilde{\pi}_{o\ell}^{(2)}}\right)^{W_C}\right\}^{-1} =$$

$$\tilde{\pi}_{io}^{(2)}\left(n_{ij}\left(\frac{n_{*j}}{\tilde{\pi}_{oj}^{(2)}}\right)^{W_C}\right)\left\{\sum_{\ell=1}^{c} n_{i\ell}\left(\frac{n_{*\ell}}{\tilde{\pi}_{o\ell}^{(2)}}\right)^{W_C}\right\}^{-1} \qquad (2.2.7)$$

Similarly, from (2.2.6),

$$\tilde{\pi}_{ij}^{(2)} = \tilde{\pi}_{oj}^{(2)}\left(n_{ij}\left(\frac{n_{i*}}{\tilde{\pi}_{io}^{(2)}}\right)^{W_R}\right)\left\{\sum_{k=1}^{r} n_{kj}\left(\frac{n_{k*}}{\tilde{\pi}_{ko}^{(2)}}\right)^{W_R}\right\}^{-1} \qquad (2.2.8)$$

where $W_R = n_{o*}/n_{oo}$.

Suppose we have initial estimates of the margins, e.g.,

$$\pi_{io}^{*} = (p_{io}^{\omega_R} p_{i*}^{1-\omega_R})\left(\sum_{\ell=1}^{r} p_{\ell o}^{\omega_R} p_{\ell *}^{1-\omega_R}\right)^{-1}$$

$$\pi_{oj}^{*} = (p_{oj}^{\omega_C} p_{*j}^{1-\omega_C})\left(\sum_{\ell=1}^{r} p_{o\ell}^{\omega_C} p_{*\ell}^{1-\omega_C}\right)^{-1}.$$

(2.2.7) and (2.2.8) suggest two iterative procedures for estimating the π_{ij}.

SCHEME 1: A step 1, use the initial estimates of the margins to compute estimates of the π_{ij} from (2.2.7). At step 2, recompute the π_{oj} using these estimates of π_{ij} and re-estimate the π_{ij} using (2.2.8). In general, at step (2n-1) use new estimates of the π_{io} from step (2n-2) in (2.2.7), and at step 2n use new estimates of the π_{oj} from step (2n-1) in (2.2.8). Iterate until successive estimates satisfy some predetermined convergence criterion.

SCHEME 2: At step 1, compute two preliminary estimates of π_{ij} by substituting the initial estimates of marginal probabilities into (2.2.7) and (2.2.8) respectively. Take as new estimates of the π_{ij} the arithmetic means of the estimates from (2.2.7) and (2.2.8) and re-estimate the margins. Repeat the process until convergence.

Scheme 1 looks like an IPF scheme with the essential difference that each margin is here updated at every other step, while IPF holds the margins constant. Unfortunately, the difference is crucial to convergence, as Scheme 1 does not converge for a number of test examples. However, Scheme 2, an obvious revision, does converge rapidly for a wide variety of examples and appears simpler to use than direct numerical methods for minimizing χ_I^2. Since χ_I^2 is convex, if Scheme 2 does converge then it must converge to the m.d.i. estimators. Finally, it is likely that Fienberg's [52] results on convergence of IPF schemes can be extended to demonstrate convergence of Scheme 2; however, the extension will not be attempted here.

c) Minimum χ_1^2

These estimates $\tilde{\pi}_{ij}^{(3)}$ satisfy the equations

$$n_{oo}\frac{\pi_{ij}}{p_{ij}} + n_{o*}\frac{\pi_{io}}{p_{i*}} + n_{*o}\frac{\pi_{oj}}{p_{*j}} = \lambda \qquad \forall\, i,j. \tag{2.2.9}$$

Hence, for fixed i,

$$n_{oo}\frac{\pi_{ij}}{p_{ij}} + n_{*o}\frac{\pi_{oj}}{p_{*j}} = k_6(i) \text{ constant } \forall j. \tag{2.2.10}$$

From (2.2.10),

$$k_6(i) = n_{oo}\frac{\pi_{io}}{p_{io}} + \sum_{\ell=1}^{c}\frac{n_{*o}p_{i\ell}}{p_{io}p_{*\ell}}\pi_{o\ell}$$

so that, for each i,j,

$$n_{oo}\frac{\pi_{ij}}{p_{ij}} + n_{*o}\frac{\pi_{oj}}{p_{*j}} = n_{oo}\frac{\pi_{io}}{p_{io}} + \sum_{\ell=1}^{c}\frac{n_{*o}p_{i\ell}}{p_{io}p_{*\ell}}\pi_{o\ell}. \tag{2.2.11}$$

Similarly, the $\tilde{\pi}_{ij}^{(3)}$ must satisfy another set of equations

$$n_{oo}\frac{\pi_{ij}}{p_{ij}} + n_{o*}\frac{\pi_{io}}{p_{i*}} = n_{oo}\frac{\pi_{oj}}{p_{oj}} + \sum_{k=1}^{r}\frac{n_{o*}p_{kj}}{p_{oj}p_{i*}}\pi_{ko}. \tag{2.2.12}$$

(2.2.11) and (2.2.12) are both sets of linear equations of rank $\leq$ (rc-1). For either set, $\Pr\{\text{rank} < (rc-1)\} \to 0$ as $n\to\infty$, $W_R \to Q_R$, $W_C \to Q_C$, where Q_R and Q_C are constants, $0<Q_R<1$, $0<Q_C<1$. When either set is of full rank (rc-1) it may be solved routinely for the $\tilde{\pi}_{ij}^{(3)}$. Writing down an explicit solution is, however, highly tedious. One way is to solve first for the estimates of the margins. From (2.2.11),

$$n_{oo}\,\pi_{oj} + n_{*o}\,\frac{p_{oj}}{p_{*j}}\,\pi_{oj} = \sum_{k=1}^{r} n_{oo}\,\frac{p_{kj}}{p_{ko}}\,\pi_{ko}$$

$$+ \sum_{k=1}^{r}\sum_{\ell=1}^{c} \frac{n_{*o}p_{k\ell}p_{kj}}{p_{*\ell}p_{ko}}\,\pi_{o\ell} \qquad (2.2.13)$$

and from (2.2.12)

$$n_{oo}\pi_{io} + n_{o*}\,\frac{p_{io}}{p_{i*}}\,\pi_{io} = \sum_{\ell=1}^{c} n_{oo}\,\frac{p_{i\ell}}{p_{o\ell}}\,\pi_{o\ell}$$

$$+ \sum_{k=1}^{r}\sum_{\ell=1}^{c} \frac{n_{o*}p_{k\ell}p_{i\ell}}{p_{k*}p_{o\ell}}\,\pi_{ko}. \qquad (2.2.14)$$

Using (2.2.13), (2.2.14) the $\tilde{\pi}^{(3)}_{io}$, $\tilde{\pi}^{(3)}_{oj}$ may be determined. The $\tilde{\pi}^{(3)}_{ij}$ are then found from either (2.2.11) or (2.2.12).

The results are very unwieldy and do not admit any obvious simplification or clear interpretation. For instance, in the simplest case, r=c=2 and

$$\tilde{\pi}^{(3)}_{10} = \frac{\alpha_2\beta_1 - \gamma_1\beta_2}{\alpha_1\alpha_2 - \gamma_1\gamma_2}, \qquad \tilde{\pi}^{(3)}_{01} = \frac{\alpha_1\beta_2 - \gamma_2\beta_1}{\alpha_1\alpha_2 - \gamma_1\gamma_2} \qquad (2.2.15)$$

where

$$\alpha_1 = n_{oo} + n_{o*}\left\{\frac{p_{10}}{p_{1*}} + \frac{p_{11}p_{21}}{p_{2*}p_{01}} + \frac{p_{12}p_{22}}{p_{2*}p_{02}} - \frac{p_{11}^2}{p_{1*}p_{01}} - \frac{p_{12}^2}{p_{1*}p_{02}}\right\}$$

$$\alpha_2 = n_{oo} + n_{*o}\left\{\frac{p_{01}}{p_{*1}} + \frac{p_{11}p_{12}}{p_{*2}p_{10}} + \frac{p_{21}p_{22}}{p_{*2}p_{20}} - \frac{p_{11}^2}{p_{*1}p_{10}} - \frac{p_{21}^2}{p_{*1}p_{20}}\right\}$$

$$\beta_1 = n_{oo}\frac{p_{12}}{p_{02}} + \frac{n_{o*}}{p_{2*}}\left\{\frac{p_{11}p_{21}}{p_{01}} + \frac{p_{12}p_{22}}{p_{02}}\right\}$$

$$\beta_2 = n_{oo}\frac{p_{21}}{p_{20}} + \frac{n_{*0}}{p_{*2}}\left\{\frac{p_{11}p_{12}}{p_{10}} + \frac{p_{21}p_{22}}{p_{20}}\right\}$$

and

$$\gamma_1 = n_{oo}\left(\frac{p_{12}}{p_{02}} - \frac{p_{11}}{p_{01}}\right), \qquad \gamma_2 = n_{oo}\left(\frac{p_{21}}{p_{20}} - \frac{p_{11}}{p_{10}}\right).$$

$\tilde{\pi}_{ij}^{(3)}$ is found using (2.2.11) for r=c=2. For larger two-way tables the results are messier. However, the routine solution of such equations involves only matrix inversion and is easily programmed (see Chapter IV).

d) Minimum χ^2

Let $\tilde{\pi}_{ij}^{(4)}$ denote the estimates which minimize

$$\chi^2 = n_{oo}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(p_{ij}-\pi_{ij})^2}{\pi_{ij}} + n_{o*}\sum_{i=1}^{r}\frac{(p_{i*}-\pi_{io})^2}{\pi_{io}} + n_{*o}\sum_{j=1}^{c}\frac{(p_{*j}-\pi_{oj})^2}{\pi_{oj}}. \tag{2.2.16}$$

It is easy to see here that χ^2 is convex with respect to the π_{ij} (as in b), because the matrix of second order partials is positive definite). Hence the $\tilde{\pi}_{ij}^{(4)}$ may be found by solving the minimum-χ^2 equations with Lagrange's method, as in a)-c), or by using a strict descent method to directly minimize χ^2.

In this case, the latter procedure is the more rewarding. For the minimum-χ^2 equations may be written as

$$n_{oo} \frac{p_{ij}^2}{\pi_{ij}^2} + n_{o*} \frac{p_{i*}^2}{\pi_{io}^2} + n_{*o} \frac{p_{*j}^2}{\pi_{oj}^2} = \lambda \qquad \forall i,j, \tag{2.2.17}$$

which implies, for fixed i,

$$n_{oo} \frac{p_{ij}^2}{\pi_{ij}^2} + n_{*o} \frac{p_{*j}^2}{\pi_{oj}^2} = k_7(i) \text{ constant } \forall j \tag{2.2.18}$$

analogous to (2.2.5) and (2.2.10); and, for fixed j,

$$n_{oo} \frac{p_{ij}^2}{\pi_{ij}^2} + n_{o*} \frac{p_{i*}^2}{\pi_{io}^2} = k_8(j) \text{ constant } \forall i \tag{2.2.19}$$

also similar to counterparts above. The analogy ceases, however, and beyond this point it is difficult to proceed because from these equations the π_{ij} cannot be disentangled and expressed individually as functions of the margins. This makes solution of (2.2.17), a set of simultaneous sixth-degree equations, much harder than solution of the corresponding sets (2.2.1), (2.2.4), (2.2.9). Several iterative schemes using (2.2.17)-(2.2.19) were tried, but failed to converge to a solution of (2.2.17) in test examples. Thus, direct minimization of (2.2.17) by strict descent numerical methods appears to be the best bet for locating the $\hat{\pi}_{ij}^{(4)}$.

2.3 An Example

Reinfurt [9] found m.l.e.'s for the π_{ij}'s from the array T_2^{**} in Table 2.3.1. In Table 2.3.2a are the results of applying the methods of Section 2.1 to only the data above the dotted line; that is, supplementation on the column variable is ignored. Table 2.3.2b gives the estimates of the π_{ij} ignoring supplementation on the row variable; all

methods agree here to four decimal places. Table 2.3.2c gives the values of competing estimates where both margins are supplemented.

TABLE 2.3.1

A 2×2 TABLE WITH BOTH MARGINS SUPPLEMENTED

1	2	3	5
4	5	9	9
5	7	12	14
4	6	10	36

TABLE 2.3.2a

ESTIMATES OF PROBABILITIES FOR TABLE 2.3.1
USING VARIOUS METHODS - ROW MARGINS ONLY SUPPLEMENTED

	M.L.E.	MIN χ^2_I	MIN χ^2_1	MIN χ^2
π_{11}	.1026	.1017	.1008	.1034
π_{12}	.2051	.2033	.2015	.2068
π_{21}	.3077	.3089	.3101	.3066
π_{22}	.3846	.3861	.3876	.3832

TABLE 2.3.2b

ESTIMATED PROBABILITIES FOR TABLE 2.3.1 -
COLUMN MARGINS ONLY SUPPLEMENTED - FOR ALL METHODS

.0818	.1688
.3273	.4221

TABLE 2.3.2c

ESTIMATED PROBABILITIES FOR TABLE 2.3.1 USING VARIOUS
METHODS - BOTH MARGINS SUPPLEMENTED

	M.L.E.	MIN χ_I^2	MIN χ_1^2	MIN χ^2
π_{11}	.1010	.1004	.0997	.1015
π_{12}	.2069	.2048	.2028	.2089
π_{21}	.3046	.3054	.3062	.3038
π_{22}	.3875	.3894	.3914	.3857

CHAPTER III

HIERARCHICAL SUPPLEMENTATION DESIGNS

3.1 Definition

Suppose one chooses K independent samples S_k, k=1,...,K, of predetermined sizes n_k, from a single population P. Measurements are then taken on the elements of S_k to determine the values, for each element, of each variate in the set of variates V_k, where the V_k are distinct. If the S_k can be ordered and renamed so that $V_K \supset V_{K-1} \supset \cdots \supset V_1$, then the design has the properties that:

i) an element measured on a variate in $V_k - V_{k-1}$ is necessarily measured on all variates in V_{k^*}, $\forall\, k^* \geq k$; and so

ii) the data collected from S_k comes from a marginal distribution of that collected from S_{k^*}, $\forall\, k^* > k$.

For instance, in a situation involving four variables A, B, C, and D, one might have three independent samples from the same population measured on the three sets of variables $V_1 = \{A\}$, $V_2 = \{A,B\}$, $V_3 = \{A,B,C,D\}$. In this case K=3, S_3 is the sample measured on all four variables while S_2 and S_1 are the samples measured on variables A and B, and A only, respectively.

Such designs have been discussed by a number of authors in the context of multivariate linear models for normally distributed random variables. Thus, Srivastava and Roy [53] have considered the

"hierarchical multivariate" model, Srivastava [54], [55] the "hierarchical incomplete multivariate" model, and Hocking and Smith [5] the "nested multivariate" model. When the variates are all categorical, Hocking and Oxspring [10] have shown that these designs produce what they call "nested likelihoods", allowing explicit solution of likelihood equations for the cell probabilities of the full contingency table formed from the sample S_K.

In the categorical framework, S_K may be regarded as the core sample in a supplemented margins design. Hence it is appropriate to call the designs just described "hierarchical supplementation designs", and the resulting data a "hierarchically supplemented contingency table". In the next section a result of Hocking and Oxspring on maximum likelihood will be derived in simpler form, and the structure of these designs will be seen to imply cognate results for other methods of estimation.

3.2 Estimation of Cell Probabilities

3.2.1 A Canonical Form

It is convenient to represent hierarchically supplemented contingency tables in a standard form. If data consists of an observed core table with supplementary observations from (d-1) of its marginal distributions then, by defining new variates in terms of the dimensions of the core table and relabeling the observed cell counts in terms of these new variates, we may write the data, in the notation of Chapter 2, as $\{n_{j_1\cdots j_d}\}$, $\{n_{j_1\cdots j_{d-1}*}\}$, $\{n_{j_1\cdots j_{d-2}**}\}$, ..., $\{n_{j_1*\cdots *}\}$, where $1 \leq j_k \leq c_k$, k=1,...,d. In this formulation the $n_{j_1\cdots j_d}$ are simply the observed frequencies in the core table relabeled, the $n_{j_1\cdots j_{d-1}*}$ are

the relabeled observed frequencies of the highest order supplemented margin, and so on. The cell probabilities from the core table are relabeled correspondingly, and the sampling distribution of all the data may be written as

$$\Pr\{\{n_{j_1\cdots j_d}\},\ldots,\{n_{j_{1*\cdots*}}\}\} =$$

$$\left(n_{o\ldots o}!\prod_{j_1=1}^{c_1}\cdots\prod_{j_d=1}^{c_d}\frac{\pi_{j_1\cdots j_d}^{n_{j_1\cdots j_d}}}{n_{j_1\cdots j_d}!}\right)\left(n_{o\ldots o*}!\prod_{j_1=1}^{c_1}\cdots\prod_{j_{d-1}=1}^{c_{d-1}}\frac{\pi_{j_1\cdots j_{d-1}o}^{n_{j_1\cdots j_{d-1}*}}}{n_{j_1\cdots j_{d-1}*}!}\right)$$

$$\cdots\left(n_{o*\cdots*}!\prod_{j_1=1}^{c_1}\frac{\pi_{j_1o\ldots o}^{n_{j_{1*\cdots*}}}}{n_{j_{1*\cdots*}}!}\right) \qquad (3.2.1)$$

where zero in a subscript indicates summation over that subscript.

For the example involving variables A, B, C, D, suppose a, b, c, d responses are possible for the corresponding variables. The data from S_3 would then be written as $\{n_{j_1j_2j_3}\}$, $j_1=1,\ldots,a$; $j_2=1,\ldots,b$; and where j_3 indexes the combined observations on C and D (that is, j_3 indexes the set $\{1,\ldots,c\}\times\{1,\ldots,d\}$, the Cartesian product of index sets corresponding to C and D). The data from S_2 would be written as $\{n_{j_1j_2}\}$, $j_1=1,\ldots,a$; $j_2=1,\ldots,b$; and from S_1 as $\{n_{j_1}\}$, $j_1=1,\ldots,a$.

Transformation to this canonical form thus involves merely a renaming of the observed cell counts and the parameters of their sampling

distribution. The discussion that follows will be entirely in terms of hierarchically supplemented contingency tables in canonical form; analogous results for all hierarchically supplemented tables will be immediate and obvious.

3.2.2 Closed Form Estimation

In Section 2.1 it was shown that the estimation procedures (1.3.1)-(1.3.4) differed, in the simplest hierarchical design, only in the method of combining p_{io} and p_{i*} to estimate π_{io}. In each case, estimates $\hat{\pi}_{ij}$ were determined by

$$\hat{\pi}_{ij} = \frac{p_{ij}}{p_{io}} \times \frac{A(p_{i*},p_{io})}{\sum_{i=1}^{r} A(p_{i*},p_{io})} ,$$

where

$$\frac{A(p_{i*},p_{io})}{\sum_{i=1}^{r} A(p_{i*},p_{io})} = \hat{\pi}_{io}$$

and A is an averaging function peculiar to the estimation procedure. For hierarchically supplemented tables of higher order things are somewhat more complex, but in each case closed form expressions for the estimators may be obtained by stepwise application of the relevant averaging function defined in Section 2.1 to certain functions of the data. The following notational simplifications will be employed in describing a core table with (d-1) supplemented margins:

$n_\ell = n_{\underbrace{o\ldots o}_{\ell \text{ zeroes}}\overbrace{*\ldots *}^{d-\ell \text{ stars}}}$ = sample size for the ℓ-dimensional margin

$n_\ell(j_1,\ldots,j_\ell) = n_{j_1\ldots j_\ell *\ldots *}$ = observed count in cell $(j_1,\ldots,j_\ell)$ for this sample

$$n_\ell(j_1,\ldots,j_k) = \sum_{j_{k+1}=1}^{c_{k+1}} \cdots \sum_{j_\ell=1}^{c_\ell} n_\ell(j_1,\ldots,j_\ell), \quad k<\ell$$

$$p_\ell(j_1,\ldots,j_\ell) = \left(\frac{1}{n_\ell}\right) n_\ell(j_1,\ldots,j_\ell)$$

$$p_\ell(j_1,\ldots,j_k) = \sum_{j_{k+1}=1}^{c_{k+1}} \cdots \sum_{j_\ell=1}^{c} p_\ell(j_1,\ldots,j_\ell), \quad k<\ell$$

$\pi_{j_1\ldots j_d}$ = probability associated with cell $(j_1,\ldots,j_d)$ in sample S_d

$$\pi_{j_1\ldots j_k} = \sum_{j_{k+1}=1}^{c_{k+1}} \cdots \sum_{j_d=1}^{c_d} \pi_{j_1\ldots j_d}, \qquad k<d.$$

Maximum likelihood

The m.l. equations are

$$\sum_{\ell=1}^{d} n_\ell \frac{p_\ell(j_1,\ldots,j_\ell)}{\pi_{j_1\ldots j_\ell}} = \text{constant} \qquad (3.2.2.1)$$

or

$$n_d \frac{p_d(j_1,\ldots,j_d)}{\pi_{j_1\cdots j_d}} = \lambda - \sum_{\ell=1}^{d-1} n_\ell \frac{p_\ell(j_1,\ldots,j_\ell)}{\pi_{j_1\cdots j_\ell}} \quad .$$

Multiplying by $\pi_{j_1\cdots j_d}$, summing over j_d, and dividing by $\pi_{j_1\cdots j_{d-1}}$ gives

$$\sum_{\ell=d-1}^{d} n_\ell \frac{p_\ell(j_1,\ldots,j_{d-1})}{\pi_{j_1\cdots j_{d-1}}} = \lambda - \sum_{\ell=1}^{d-2} n_\ell \frac{p_\ell(j_1,\ldots,j_\ell)}{\pi_{j_1\cdots j_\ell}} \quad .$$

Proceeding stepwise, multiplying the result of step (k-1) by $\pi_{j_1\cdots j_{d-k+1}}$, summing over j_{d-k+1}, and dividing by $\pi_{j_1\cdots j_{d-k}}$ gives, after step k,

$$\frac{\sum_{\ell=d-k}^{d} n_\ell p_\ell(j_1,\ldots,j_{d-k})}{\pi_{j_1\cdots j_{d-k}}} = \lambda - \sum_{\ell=1}^{d-k-1} \frac{n_\ell p_\ell(j_1,\ldots,j_\ell)}{\pi_{j_1\cdots j_\ell}} . \qquad (3.2.2.2)$$

In particular, $\lambda = N = \sum_{\ell=1}^{d} n_\ell$, and $\hat{\pi}_{j_1}^{(1)}$ = m.l.e. of π_{j_1} is

$$\hat{\pi}_{j_1}^{(1)} = \frac{\sum_{\ell=1}^{d} n_\ell p_\ell(j_1)}{N} \quad .$$

Successively using (3.2.2.2) with k=d-2, d-3,...,0, one may solve for the m.l.e.'s $\hat{\pi}_{j_1\cdots j_d}^{(1)}$. In fact, if $N_{j_1\cdots j_\ell} = \sum_{i=\ell}^{d} n_i(j_1,\ldots,j_\ell) = \sum_{i=\ell}^{d} n_i p_i(j_1,\ldots,j_\ell)$, then $\hat{\pi}_{j_1\cdots j_d}$ may be written down directly as

$$\hat{\pi}^{(1)}_{j_1\cdots j_d} = \frac{N_{j_1}}{N} \prod_{\ell=2}^{d} \frac{N_{j_1,\ldots,j_\ell}}{N_{j_1,\ldots,j_{\ell-1}}{}^{-n_{\ell-1}(j_1,\ldots,j_{\ell-1})}} . \qquad (3.2.2.3)$$

For, from (3.2.2.2),

$$\frac{\sum_{\ell=k+1}^{d} n_\ell p_\ell(j_1,\ldots,j_{k+1})}{\pi_{j_1,\ldots,j_{k+1}}} = N - \sum_{\ell=1}^{k} \frac{n_\ell p_\ell(j_1,\ldots,j_\ell)}{\pi_{j_1,\ldots,j_\ell}} =$$

$$\frac{\sum_{\ell=k}^{d} n_\ell p_\ell(j_1,\ldots,j_k)}{\pi_{j_1,\ldots,j_k}} - \frac{n_k p_k(j_1,\ldots,j_k)}{\pi_{j_1,\ldots,j_k}} ,$$

so that

$$\pi_{j_1,\ldots,j_{k+1}} = \frac{N_{j_1,\ldots,j_{k+1}}}{N_{j_1,\ldots,j_k}{}^{-n_k(j_1,\ldots,j_k)}} \pi_{j_1,\ldots,j_k},$$

which gives (3.2.2.3).

Suppose x_i, i=1,2 are estimates of some probability p calculated from different samples, and associated with x_i is n^i, the size of the sample from which x_i is constructed. Define a function $A^{(1)}(x_1,x_2) = (n^1+n^2)^{-1}(n^1x_1+n^2x_2)$, a weighted average of x_1 and x_2, and conventionally let $A^{(1)}(x)\equiv x$. Now write $A^{(1)}_{j_1,\ldots,j_k} =$

$$A^{(1)}(p_k(j_1,\ldots,j_k), \sum_{j_{k+1}=1}^{c_{k+1}} A^{(1)}(p_{k+1}(j_1,\ldots,j_{k+1}),\ldots,$$

$$\sum_{j_d=1}^{c_d} A^{(1)}(p_d(j_1,\ldots,j_d))\ldots).$$

Then (3.2.1.2) may be written as

$$\frac{(\sum_{\ell=d-k}^{d} n_\ell)A^{(1)}_{j_1,\ldots,j_{d-k}}}{\pi_{j_1,\ldots,j_{d-k}}} = N - \sum_{\ell=1}^{d-k-1} \frac{n_\ell p_\ell(j_1,\ldots,j_\ell)}{\pi_{j_1,\ldots,j_\ell}}$$

and (3.2.1.3) as

$$\hat{\pi}^{(1)}_{j_1\cdots j_d} = \frac{A^{(1)}_{j_1}}{\sum_{j_1=1}^{c_1} A^{(1)}_{j_1}} \prod_{\ell=2}^{d} \frac{(\sum_{i=\ell}^{d} n_i)A^{(1)}_{j_1\cdots j_\ell}}{(\sum_{i=\ell-1}^{d} n_i)A^{(1)}_{j_1\cdots j_{\ell-1}} - n_{\ell-1}(j_1\cdots j_{\ell-1})}. \quad (3.2.2.4)$$

Minimum χ^2_I

Here the equations to be solved are

$$\prod_{\ell=1}^{d}\left(\frac{\pi_{j_1,\ldots,j_\ell}}{p_\ell(j_1,\ldots,j_\ell)}\right)^{n_\ell} = \lambda \text{ constant} \quad (3.2.2.5)$$

or

$$\left(\frac{\pi_{j_1,\ldots,j_d}}{p_d(j_1,\ldots,j_d)}\right)^{n_d} \prod_{\ell=1}^{d-1} \left(\frac{\pi_{j_1,\ldots,j_\ell}}{p_\ell(j_1,\ldots,j_\ell)}\right)^{n_\ell} = \lambda \text{ constant.}$$

Multiplying by $(p_d(j_1,\ldots,j_d))^{n_d}$, taking the n_d^{th} root of both sides, summing from $j_d=1,\ldots,c_d$, dividing by $p_d(j_1,\ldots,j_{d-1})$, and taking the n_d^{th} power of both sides gives

$$\frac{\pi_{j_1\cdots j_{d-1}}^{n_d+n_{d-1}}}{p_d(j_1\cdots j_{d-1})^{n_d} p_{d-1}(j_1\cdots j_{d-1})^{n_{d-1}}} \prod_{\ell=1}^{d-2} \left(\frac{\pi_{j_1\cdots j_\ell}}{p_\ell(j_1\cdots j_\ell)}\right)^{n_\ell} = \lambda.$$

Now, for two estimates x_1, x_2 define $A^{(2)}(x_1,x_2) = (x_1^{n^1} x_2^{n^2})^{\frac{1}{n^1+n^2}}$, and $A^{(2)}(x)\equiv x$. Define $A^{(2)}_{j_1\cdots j_\ell}$ similarly as $A^{(1)}_{j_1\cdots j_\ell}$, using $A^{(2)}$ rather than $A^{(1)}$. Proceeding as above, multiply the result of step (k-1) by

$$\left(A^{(2)}_{j_1,\ldots,j_{d-k+1}}\right)^{\left(\sum_{\ell=d-k+1}^{d} n_\ell\right)},$$

take the $\left(\sum_{\ell=d-k+1}^{d} n_\ell\right)^{th}$ root of both sides, sum over j_{d-k+1}, divide by

$$\left(\sum_{j_{d-k+1}=1}^{c_{d-k+1}} A^{(2)}_{j_1,\ldots,j_{d-k+1}}\right),$$

and restore the exponent. After this procedure, step k, the result may be written as

$$\left(\frac{\pi_{j_1,\ldots,j_{d-k}}}{A^{(2)}_{j_1,\ldots,j_{d-k}}}\right)^{\left(\sum_{\ell=d-k}^{d} n_\ell\right)} \prod_{\ell=1}^{d-k-1} \left(\frac{\pi_{j_1,\ldots,j_\ell}}{p_\ell(j_1,\ldots,j_\ell)}\right)^{n_\ell} = \lambda \text{ constant.} \quad (3.2.2.6)$$

In particular,

$$\lambda = \left(\sum_{j_1=1}^{c_1} A^{(2)}_{j_1}\right)^{-N}, \text{ and } \hat{\pi}^{(2)}_{j_1} = \frac{A^{(2)}_{j_1}}{\sum_{j_1=1}^{c_1} A^{(2)}_{j_1}}$$

is the minimum χ^2_I estimator of π_{j_1}. Further, we have from (3.2.2.6)

$$\left(\frac{\pi_{j_1,\ldots,j_k}}{A^{(2)}_{j_1,\ldots,j_k}}\right)^{\left(\sum_{\ell=k}^{d} n_\ell\right)} \prod_{\ell=1}^{k-1}\left(\frac{\pi_{j_1,\ldots,j_\ell}}{p_\ell(j_1,\ldots,j_\ell)}\right)^{n_\ell} = \lambda =$$

$$\left(\frac{\pi_{j_1,\ldots,j_{k-1}}}{A^{(2)}_{j_1,\ldots,j_{k-1}}}\right)^{\left(\sum_{\ell=k-1}^{d} n_\ell\right)} \prod_{\ell=1}^{k-2} \left(\frac{\pi_{j_1,\ldots,j_\ell}}{p_\ell(j_1,\ldots,j_\ell)}\right)^{n_\ell}$$

or

$$\pi_{j_1\cdots j_k} = \left(\frac{p_{k-1}(j_1\cdots j_{k-1})}{A^{(2)}_{j_1\cdots j_{k-1}}}\right)^{\left(\frac{n_{k-1}}{\sum_{\ell=k}^{d} n_\ell}\right)} \left(\frac{A^{(2)}_{j_1\cdots j_k}}{A^{(2)}_{j_1\cdots j_{k-1}}}\right) \pi_{j_1\cdots j_{k-1}},$$

so that the minimum χ^2_I estimators can be written explicitly as

$$\hat{\pi}^{(2)}_{j_1\cdots j_d} = \frac{A^{(2)}_{j_1}}{\sum_{j_1=1}^{c_1} A_{j_1}} \prod_{k=2}^{d} \left(\frac{p_{k-1}(j_1\cdots j_{k-1})}{A^{(2)}_{j_1\cdots j_{k-1}}}\right)^{\left(\frac{n_{k-1}}{\sum_{\ell=k}^{d} n_\ell}\right)} \left(\frac{A_{j_1\cdots j_k}}{A^{(2)}_{j_1\cdots j_{k-1}}}\right). \quad (3.2.2.7)$$

Minimum χ^2_1

The estimation equations are

$$\sum_{\ell=1}^{d} n_\ell \frac{\pi_{j_1,\ldots,j_\ell}}{p_\ell(j_1,\ldots,j_\ell)} = \lambda \text{ constant} \quad (3.2.2.8)$$

or

$$n_d \frac{\pi_{j_1,\ldots,j_d}}{p_d(j_1,\ldots,j_d)} = \lambda - \sum_{\ell=1}^{d-1} n_\ell \frac{\pi_{j_1,\ldots,j_\ell}}{p_\ell(j_1,\ldots,j_\ell)}.$$

Let $A^{(3)}(x_1,x_2) = (n^1 x_1^{-1} + n^2 x_2^{-1})^{-1}(n^1+n^2)$, $A_3(x)\equiv x$, and define $A^{(3)}_{j_1\cdots j_\ell}$ in the usual manner from A_3. The stepwise procedure which multiplies the result of step (k-1) by $A^{(3)}_{j_1\cdots j_{d-k+1}}$, sums over

j_{d-k+1}, and then divides by

$$\sum_{j_{d-k+1}=1}^{c_{d-k+1}} A^{(3)}_{j_1,\ldots,j_{d-k+1}},$$

gives as the result of the k^{th} step

$$\frac{\left(\sum_{\ell=d-k}^{d} n_\ell\right)\pi_{j_1\cdots j_{d-k}}}{A^{(3)}_{j_1\cdots j_{d-k}}} = \lambda - \sum_{\ell=1}^{d-k-1} n_\ell \frac{\pi_{j_1\cdots j_\ell}}{p_\ell(j_1\cdots j_\ell)} . \qquad (3.2.2.9)$$

Thus

$$\lambda = N\left(\sum_{j_1=1}^{c_1} A^{(3)}_{j_1}\right)^{-1}, \quad \text{and} \quad \hat{\pi}^{(3)}_{j_1} = \frac{A^{(3)}_{j_1}}{\sum_{j_1=1}^{c_1} A^{(3)}_{j_1}} .$$

Further, from (3.2.2.9),

$$\frac{\left(\sum_{\ell=k+1}^{d} n_\ell\right)\pi_{j_1\cdots j_{k+1}}}{A^{(3)}_{j_1}\cdots j_{k+1}} =$$

$$\lambda - \sum_{\ell=1}^{k} n_\ell \frac{\pi_{j_1\cdots j_\ell}}{p_\ell(j_1\cdots j_\ell)} = \frac{\left(\sum_{\ell=k}^{d} n_\ell\right)\pi_{j_1\cdots j_k}}{A^{(3)}_{j_1\cdots j_k}} - \frac{n_k\pi_{j_1\cdots j_k}}{p_k(j_1\cdots j_k)} ,$$

so that $\pi_{j_1\cdots j_{k+1}}$ =

$$\left\{ \frac{\left(\sum_{\ell=k}^{d} n_\ell\right)}{\left(\sum_{\ell=k+1}^{d} n_\ell\right) A^{(3)}_{j_1\cdots j_k}} - \frac{n_k}{\left(\sum_{\ell=k+1}^{d} n_\ell\right) p_k(j_1\cdots j_k)} \right\} A^{(3)}_{j_1\cdots j_{k+1}} \pi_{j_1\cdots j_k}.$$

Hence, the minimum-χ^2_1 estimates $\hat{\pi}^{(3)}_{j_1\cdots j_d} =$

$$\frac{A^{(3)}_{j_1}}{\sum_{j_1=1}^{c_1} A^{(3)}_{j_1}} \prod_{\ell=2}^{d} A^{(3)}_{j_1\cdots j_\ell} \left\{ \frac{\left(\sum_{i=k}^{d} n_i\right)}{\left(\sum_{i=\ell}^{d} n_i\right) A^{(3)}_{j_1\cdots j_{\ell-1}}} - \frac{n_{\ell-1}}{\left(\sum_{i=\ell}^{d} n_i\right) p_{\ell-1}(j_1\cdots j_{\ell-1})} \right\}. \tag{3.2.2.10}$$

Minimum-χ^2

The minimization equations for these estimates are

$$\sum_{\ell=1}^{d} n_\ell \left[\frac{p_\ell(j_1,,,.j_\ell)}{\pi_{j_1,\ldots,j_\ell}} \right]^2 = \lambda \text{ constant}$$

or

$$n_d \left[\frac{p_d(j_1,\ldots,j_d)}{\pi_{j_1,\ldots,j_d}} \right]^2 = \lambda - \sum_{\ell=1}^{d-1} n_\ell \left[\frac{p_\ell(j_1,\ldots,j_\ell)}{\pi_{j_1,\ldots,j_\ell}} \right]^2 .$$

Defining $A^{(4)}_{j_1\cdots j_\ell}$ from $A^{(4)}(x_1,x_2) = \frac{(n^1x_1^2+n^2x_2^2)^{\frac{1}{2}}}{(n^1+n^2)^{\frac{1}{2}}}$, $A^{(4)}(x)\equiv x$, we solve these equations with the following stepwise procedure: At step k, multiply by $\pi^2_{j_1\cdots j_{d-k+1}}$, take square roots of both sides, sum over j_{d-k+1}, divide by $\pi_{j_1\cdots j_{d-k}}$, and square both sides.

Step k gives

$$\frac{\left(\sum_{\ell=d-k}^{d} n_\ell\right) A^{(4)2}_{j_1\cdots j_{d-k}}}{\pi^2_{j_1\cdots j_{d-k}}} = \lambda - \sum_{\ell=1}^{d-k-1} n_\ell \left(\frac{p_\ell(j_1\cdots j_\ell)}{\pi_{j_1\cdots j_\ell}}\right)^2 . \qquad (3.2.2.11)$$

Hence,

$$\lambda = N\left(\sum_{j_1=1}^{c_1} A^{(4)}_{j_1}\right)^2, \quad \text{and} \quad \hat{\pi}^{(4)}_{j_1} = \frac{A^{(4)}_{j_1}}{\sum_{j_1=1}^{c_1} A^{(4)}_{j_1}} .$$

Finally, (3.2.2.11) gives

$$\frac{\left(\sum_{\ell=k+1}^{d} n_\ell\right) A^{(4)2}_{j_1\cdots j_{k+1}}}{\pi^2_{j_1\cdots j_{k+1}}} = \lambda - \sum_{\ell=1}^{k} n_\ell \left(\frac{p_\ell(j_1\cdots j_\ell)}{\pi_{j_1\cdots j_\ell}}\right)^2 =$$

$$= \frac{\left(\sum_{\ell=k}^{d} n\right) A^{(4)2}_{j_1\cdots j_k}}{\pi^2_{j_1\cdots j_k}} - n_k \left(\frac{p_k(j_1\cdots j_k)}{\pi_{j_1\cdots j_k}}\right)^2 =$$

$$\frac{\left(\sum_{\ell=k}^{d} n_\ell\right) A^{(4)2}_{j_1\cdots j_k} - n_k\, p^2_k(j_1\cdots j_k)}{\pi^2_{j_1\cdots j_k}} .$$

Thus, $\pi_{j_1\cdots j_{k+1}} =$

$$\left\{ \frac{\left(\sum_{\ell=k+1}^{d} n_\ell \right) A^{(4)2}_{j_1 \cdots j_{k+1}}}{\left(\sum_{\ell=k}^{d} n_\ell \right) A^{(4)2}_{j_1 \cdots j_k} - n_k \, p_k^2(j_1 \cdots j_k)} \right\}^{\frac{1}{2}} \pi_{j_1 \cdots j_k},$$

so that $\hat{\pi}^{(4)}_{j_1 \cdots j_d} =$

$$\frac{A_{j_1}}{\sum_{j_1=1}^{c_1} A^{(4)}_{j_1}} \prod_{\ell=2}^{d} \left\{ \frac{\left(\sum_{i=\ell}^{d} n_i \right) A^{(4)2}_{j_1 \cdots j_\ell}}{\left\{ \left(\sum_{i=\ell-1}^{d} n_i \right) A^{(4)2}_{j_1 \cdots j_{\ell-1}} - n_{\ell-1} p^2_{\ell-1}(j_1 \cdots j_{\ell-1}) \right\}} \right\}^{\frac{1}{2}}.$$

CHAPTER IV

ESTIMATION ALGORITHMS FOR GENERAL DESIGNS

4.1 Notation

The maximum likelihood and minimum Neyman-χ^2_1 estimates of cell probabilities derived in Chapters II and III, as well as those for more complex incomplete categorical data configurations, may be calculated using a set of weighted least-squares equations which have been discussed by Reinfurt [9] and Hocking and Oxspring [10]. These equations will now be considered in somewhat more detail.

It is necessary first to generalize our notation one final time. In so doing, we will set up the problem as one of simultaneous estimation of cell probabilities from several (I) sets of incomplete data from (I) different populations. In this chapter the populations will be assumed unrelated, and the estimates for parameters of any one population will not depend on data from any of the others. Hence, this chapter may be read assuming I=1 throughout. In Chapter 5 the multi-population notation will become of importance.

Let $M(\nu,\underset{\sim}{\pi})$ represent the multinomial distribution with sample size ν and parameter vector $\underset{\sim}{\pi} = (\pi_1,\ldots,\pi_{c-1})'$, $\pi_i > 0 \ \forall_i$, $\sum_{i=1}^{c-1} \pi_i < 1$. Let $\underset{\sim}{j}_k$ be a k-vector of 1's. Thus, the probability of observing $\underset{\sim}{n} = (n_1,\ldots,n_{c-1})'$, integer $n_i \geq 0 \ \forall_i$, $\sum_{i=1}^{c-1} n_i \leq \nu$, from $M(\nu,\underset{\sim}{\pi})$ is

$$M(\underset{\sim}{n},\nu,\underset{\sim}{\pi}) = \nu! \prod_{i=1}^{c} \frac{\pi_i^{n_i}}{n_i!},$$

where

$$\pi_c = 1-\underset{\sim}{j}'_{c-1}\underset{\sim}{\pi}, \quad n_c = \nu-\underset{\sim}{j}'_{c-1}\underset{\sim}{n}.$$

If $\underset{\sim}{n}$ is a random observation vector from $M(\nu,\underset{\sim}{\pi})$, we have $\mathrm{Cov}(\underset{\sim}{n}) = \nu(D_{\underset{\sim}{\pi}}-\underset{\sim}{\pi}\underset{\sim}{\pi}') = \nu v(\underset{\sim}{\pi})$, where $D_{\underset{\sim}{\pi}} = \{d_{ij}(\underset{\sim}{\pi})\}$ is the diagonal matrix with $d_{ii}(\underset{\sim}{\pi}) = \pi_i$. Hence, if $\underset{\sim}{p} = \underset{\sim}{n}/\nu$, $\mathrm{Cov}(\underset{\sim}{p}) = \frac{1}{\nu} v(\underset{\sim}{\pi})$.

We will be concerned with inference about a vector $\underset{\sim}{\Pi} = (\underset{\sim}{\pi}'_1,\ldots,\underset{\sim}{\pi}'_I)'$, where

$$\underset{1\times(c_i-1)}{\underset{\sim}{\pi}_i} = (\pi_{i,1},\ldots,\pi_{i,c_{i-1}})'$$

satisfies the conditions on $\underset{\sim}{\pi}$ above, from a data vector

$$\underset{\sim}{N} = (\underset{\sim}{n}'_{11},\ldots,\underset{\sim}{n}'_{1k_1},\underset{\sim}{n}'_{21},\ldots,\underset{\sim}{n}'_{2k_2},\ldots,\underset{\sim}{n}'_{I1},\ldots,\underset{\sim}{n}'_{Ik_I})'.$$

The $\underset{(c_{ik}-1)\times 1}{\underset{\sim}{n}_{ik}}$ are assumed to be observation vectors chosen independently from the distributions

$$M(\nu_{ik}, \underset{(c_{ik}-1)\times(c_i-1)}{M_{ik}} \underset{\sim}{\pi}_i),$$

where M_{ik} is a matrix which groups disjoint sets of the elements of $\underset{\sim}{\pi}_i$. Thus, $M_{ik} = \{m_{\alpha\beta}^{(ik)}\}$ is such that $m_{\alpha\beta}^{(ik)} = 0$ or 1, and

$$m_{\alpha\beta}^{(ik)} = m_{\alpha'\beta}^{(ik)} \Rightarrow m_{\alpha\beta}^{(ik)} = 0.$$

Under the model, the observed vector $\underset{\sim}{N}$ arises from the distribution

$$\prod_i \prod_k M(\nu_{ik}, M_{ik}\underset{\sim}{\pi}_i),$$

and has probability

$$\prod_i \prod_k M(\underset{\sim}{n}_{ik}, \nu_{ik}, M_{ik}\underset{\sim}{\pi}_i). \tag{4.1.1}$$

Let

$$\underset{\sim}{\pi}_{ik} = M_{ik}\underset{\sim}{\pi}_i, \quad \underset{\sim}{\pi}_d = (\underset{\sim}{\pi}'_{11}, \ldots, \underset{\sim}{\pi}'_{Ik_I})',$$

$$\underset{\sim}{p}_{ik} = \frac{1}{\nu_{ik}} \underset{\sim}{n}_{ik}, \quad \underset{\sim}{p}_i = (\underset{\sim}{p}'_{i1}, \ldots, \underset{\sim}{p}'_{ik_i})',$$

$$\underset{\sim}{p}_d = (\underset{\sim}{p}'_1, \ldots, \underset{\sim}{p}'_I)', \quad \mathcal{V}_{ik} = \nu_{ik} I_{c_{ik}-1},$$

$\mathcal{V} = \text{diag}(\mathcal{V}_{ik})$, where I_s is the identity of order s, and $\text{diag}(\mathcal{V}_{ik})$ is the block diagonal matrix with the $\mathcal{V}_{ik}$ on the diagonal, ordered as in $\underset{\sim}{p}_d$.

Then

$$E\,\underset{\sim}{p}_d = \underset{\sim}{\pi}_d, \tag{4.1.2}$$

$$\text{Cov}(\underset{\sim}{p}_d) = V(\underset{\sim}{\pi}_d) = \mathcal{V}^{-1}\text{diag}(v(\underset{\sim}{\pi}_{ik})), \tag{4.1.3}$$

and $\underset{\sim}{p}_d$ is a consistent estimator of $\underset{\sim}{\pi}_d$, that is, $\underset{\sim}{p}_d \xrightarrow{p} \underset{\sim}{\pi}_d$ (in probability) as the $\nu_{ik} \to \infty$. By Slutsky's theorem, also

$$V(\underset{\sim}{p}_d) = \mathcal{V}^{-1}\text{diag}(v(\underset{\sim}{p}_{ik})) \xrightarrow{p} V(\underset{\sim}{\pi}_d). \tag{4.1.4}$$

If $M_i = (M'_{i1},\ldots,M'_{ik_i})'$, $M = \text{diag}(M_i)$, then $\underset{\sim}{\pi}_d = M\underset{\sim}{\Pi}$ and (4.1.2) gives

$$E\underset{\sim}{p}_d = M\underset{\sim}{\Pi}. \tag{4.1.5}$$

Lastly, write π_{ikj} for the j^{th} element of the vector $\underset{\sim}{\pi}_{ij}$.

4.2 Estimation by Weighted Least-Squares

Suppose we formally apply weighted least-squares procedures to the expectation model given by (4.1.5) and (4.1.4). The normal equations for estimating $\underset{\sim}{\Pi}$ are

$$(M'V^{-1}(\underset{\sim}{\pi}_d)M)\underset{\sim}{\Pi} = M'V^{-1}(\underset{\sim}{\pi}_d)\underset{\sim}{p}_d. \tag{4.2.1}$$

(4.2.1) can be expressed alternately as

$$M'V^{-1}(\underset{\sim}{\pi}_d)(\underset{\sim}{\pi}_d-\underset{\sim}{p}_d) = \underset{\sim}{0} \text{ or, } \forall_i,$$

$$\sum_{k=1}^{k_i} \nu_{ik}V^{-1}(\underset{\sim}{\pi}_{ik})(\underset{\sim}{\pi}_{ik}-\underset{\sim}{p}_{ik}) = \underset{\sim}{0}. \tag{4.2.2}$$

Now, since

$$V(\underset{\sim}{\pi}_{ik}) = (D_{\underset{\sim}{\pi}_{ik}} - \underset{\sim}{\pi}_{ik}\underset{\sim}{\pi}'_{ik}),$$

$$V^{-1}(\underset{\sim}{\pi}_{ik}) = D^{-1}_{\underset{\sim}{\pi}_{ik}} + \frac{1}{\pi_{ikc_{ik}}} J_{c_{ik}-1},$$

where J_d is a d×d matrix of 1's. It is readily seen from this that the elements of $\nu_{ik}M'_{ik}V^{-1}(\underset{\sim}{\pi}_{ik}) = \{\gamma^{(ik)}_{uv}\}$ can be written as

$$\gamma_{uv}^{(ik)} = \nu_{ik}\left(\frac{m_{vu}^{(ik)}}{\pi_{ikv}} + \frac{1}{\pi_{ikc_{ik}}}\right),$$

so that the u^{th} element of $\nu_{ik}M_{ik}'v^{-1}(\underset{\sim}{\pi}_{ik})\underset{\sim}{\pi}_{ik}$ is

$$\nu_{ik}\left(\frac{\pi_{ikv}}{\pi_{ikv}} + \frac{(1-\pi_{ikc_{ik}})}{\pi_{ikc_{ik}}}\right) = \frac{\nu_{ik}}{\pi_{ikc_{ik}}}.$$

Similarly, the u^{th} element of

$$\nu_{ik}M_{ik}'v^{-1}(\underset{\sim}{\pi}_{ik})\underset{\sim}{p}_{ik} = \nu_{ik}\left(\frac{p_{ikv_{iku}}}{\pi_{ikv_{iku}}} + \frac{(1-p_{ikc_{ik}})}{\pi_{ikc_{ik}}}\right),$$

where v_{iku} is such that $m_{v_{iku}u}^{(ik)} = 1$. Summing we find that (4.2.2) reduces, for each i, to (c_i-1) equations

$$\sum_{k=1}^{k_i} \nu_{ik}\frac{p_{ikv_{iku}}}{\pi_{ikv_{iku}}} = \sum_{k=1}^{k_i} \nu_{ik}\frac{p_{ikc_{ik}}}{\pi_{ikc_{ik}}}.$$

But these are precisely the equations obtained from differentiating the log of the likelihood (4.1.1). Hence, as Hocking and Oxspring [10] have shown, weighted least-squares and maximum likelihood estimation coincide for this broad class of problems relating to product-multinomial distributions. Berkson [37], among others, has noted this for less general situations. This result can also be obtained from the point of view of "scoring" procedures for maximum likelihood estimation.

Now consider (4.2.2) when the weights are unknown and estimated by $V^{-1}(\underset{\sim}{p}_d)$. We have

$$\sum_{k=1}^{k_i} \nu_{ik} M'_{ik} V^{-1}(\underset{\sim}{p}_{ik})(\underset{\sim}{\pi}_{ik} - \underset{\sim}{p}_{ik}) = 0, \quad \forall_i. \tag{4.2.3}$$

The (u,v) element of $\nu_{ik} M'_{ik} V^{-1}(\underset{\sim}{p}_{ik})$ is then

$$\delta_{uv}^{(ik)} = \nu_{ik}\left(\frac{m_{vu}^{(ik)}}{p_{ikv}} + \frac{1}{p_{ikc_{ik}}}\right)$$

so that the u^{th} element of $\nu_{ik} M'_{ik} V^{-1}(\underset{\sim}{p}_{ik})\underset{\sim}{p}_{ik} = \nu_{ik}/p_{ikc_{ik}}$, while the corresponding element of $\nu_{ik} M'_{ik} V^{-1}(\underset{\sim}{p}_{ik})\underset{\sim}{\pi}_{ik}$ is

$$\nu_{ik}\left(\frac{\pi_{ikv_{iku}}}{p_{ikv_{iku}}} + \frac{(1-\pi_{ikc_{ik}})}{p_{ikc_{ik}}}\right).$$

Summing (4.2.3) gives, for each i, the $(c_i - 1)$ equations

$$\sum_{k=1}^{k_i} \nu_{ik} \frac{\pi_{ikv_{iku}}}{p_{ikv_{iku}}} = \sum_{k=1}^{k_i} \nu_{ik} \frac{\pi_{ikc_{ik}}}{p_{ikc_{ik}}}. \tag{4.2.4}$$

These are easily seen to be the minimum-Neyman-χ_1 equations for estimating $\underset{\sim}{\Pi}$, confirming the well-known result of Bhapkar [56].

We now prove a simple

Lemma. Let $\hat{\underset{\sim}{\pi}}_d$ be any consistent estimator of $\underset{\sim}{\pi}_d$, that is, $\Pr\{\|\hat{\underset{\sim}{\pi}}_d - \underset{\sim}{\pi}_d\| > \varepsilon\} \to 0$ as the $\nu_{ik} \to \infty$, $(\sum_{k=1}^{k_i} \nu_{ik})^{-1} \nu_{ik} \to Q_{ik}$ constant, for all $\varepsilon > 0$. Then $\Pr\{M' V^{-1}(\hat{\underset{\sim}{\pi}}_d) M \text{ singular}\} \to 0$ iff rank $M_i = c_i - 1 \quad \forall$ i.

Proof. Since $\hat{\underset{\sim}{\pi}}_d$ is consistent, $\Pr\{V^{-1}(\hat{\underset{\sim}{\pi}}_d)$ non-singular and positive-definite$\} \to 1$. But under these circumstances $M'V^{-1}(\hat{\underset{\sim}{\pi}}_d)M$ is singular $\iff$ rank $M_i < c_i - 1$ for some i. Suppose $M'V^{-1}(\hat{\underset{\sim}{\pi}}_d)M$ singular. Then there is $\underset{\sim}{\lambda} \neq 0$ such that $\underset{\sim}{\lambda}'M'V^{-1}(\hat{\underset{\sim}{\pi}}_d)M = 0$, hence $(M\underset{\sim}{\lambda})'V^{-1}(\hat{\underset{\sim}{\pi}}_d)(M\underset{\sim}{\lambda}) = 0$, so that $M\underset{\sim}{\lambda} = \underset{\sim}{0}$ due to the positive definite property of $V^{-1}(\hat{\underset{\sim}{\pi}}_d)$. Since $M = \text{diag}(M_i)$, rank $M_i < c_i - 1$ for some i. Conversely, if this last is true, rank $M < \sum_{i=1}^{I}(c_i - 1)$, hence rank $(M'V^{-1}(\hat{\underset{\sim}{\pi}}_d)M) \leq$ rank $M < \sum_{i=1}^{I}(c_i - 1)$.

The condition of the lemma is automatically satisfied by any supplemented margin design, and also is fulfilled for many incomplete block measurement designs. When the condition is satisfied, the m.l. equations (4.2.1) are equivalently

$$\underset{\sim}{\Pi} = (M'V^{-1}(\underset{\sim}{\pi}_d)M)^{-1}M'V^{-1}(\underset{\sim}{\pi}_d)\underset{\sim}{p}_d. \qquad (4.2.5)$$

A closed form solution of (4.2.5) is not generally available when the design is not hierarchical. The summation procedures which eliminate cell probabilities and higher order margins from the various equations in Chapter III fail for non-hierarchical designs, so that the equations are not subject to ready simplification. The remarks of this section suggest the use of the following iterative scheme, proposed in [9] and [10], to determine the m.l.e.'s. Take

$$\underset{\sim}{\Pi}^{(1)} = (M'V^{-1}(\underset{\sim}{p}_d)M)^{-1}M'V^{-1}(\underset{\sim}{p}_d)\underset{\sim}{p}_d;\ \underset{\sim}{\pi}_d^{(1)} = M\underset{\sim}{\Pi}^{(1)}$$

$$\underset{\sim}{\Pi}^{(2)} = (M'V^{-1}(\underset{\sim}{\pi}_d^{(1)})M)^{-1}M'V^{-1}(\underset{\sim}{\pi}_d^{(1)})\underset{\sim}{p}_d;\underset{\sim}{\pi}_d^{(2)} = M\underset{\sim}{\Pi}^{(2)}$$

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

$$\underset{\sim}{\Pi}^{(n+1)} = (M'V^{-1}(\underset{\sim}{\pi}_d^{(n)})M)^{-1}M'V^{-1}(\underset{\sim}{\pi}_d^{(n)})\underset{\sim}{p}_d;\ \underset{\sim}{\pi}_d^{(n+1)} = M\underset{\sim}{\Pi}^{(n+1)}$$

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

and so on, provided the appropriate matrices are non-singular, until some suitable convergence criterion is satisfied. As noted in [10], concavity of the log-likelihood implies that whenever this procedure converges, it converges to the m.l. estimator of $\underset{\sim}{\Pi}$. Moreover, continuity of the function

$$f(\underset{\sim}{\hat{\Pi}}) = (M'V^{-1}(M\underset{\sim}{\hat{\Pi}})M)^{-1}M'V^{-1}(M\underset{\sim}{\hat{\Pi}})\underset{\sim}{p}_d$$

whenever the elements of $\underset{\sim}{\hat{\Pi}}$ are all non-zero, coupled with (4.2.5) and the lemma, implies in an obvious manner the result:

Theorem. Let $\Pi_k^{(m)}$, Π_k be the k^{th} elements of $\underset{\sim}{\Pi}^{(m)}$, $\underset{\sim}{\Pi}$ respectively. Then, for any m, $Pr\{\underset{\sim}{\Pi}^{(m)}$ exists$\} \to 1$ asymptotically iff rank $M_i = (c_i - 1)\ \forall_i$. In this case the estimators: $\tilde{\Pi}_k^{(m)} \equiv \Pi_k^{(m)}$ when $\underset{\sim}{\Pi}^{(m)}$ exists (arbitrary otherwise), are consistent and BAN for Π_k; a consistent estimate of the asymptotic covariance matrix of the $\tilde{\Pi}_k^{(m)}$ (m fixed) is given by $(M'V^{-1}(M\underset{\sim}{\tilde{\Pi}}^{(t)})M)^{-1}$, for any t.

Hocking and Oxspring present the iterative procedure given here in a computational form which produces expressions and estimates for

the gain in precision resulting from the addition of the data vector $\underset{\sim}{n}_{ik}$ to the vector $(\underset{\sim}{n}'_{i1},\ldots,\underset{\sim}{n}_{i,k-1})'$. For any two subsets $A \subset B$ of the set of vectors $\underset{\sim}{n}_{i1},\ldots,\underset{\sim}{n}_{ik_i}$ the least-squares equations may be broken down in the manner they describe to study the improvement in estimation from adding the data in (B-A) to that in A. When $A \neq A \cap B \neq B$, one may still easily compare elements and estimates of elements of the covariance matrices obtained by analyzing the subsets separately.

Sometimes it may be necessary or even advantageous to use designs in which some M_i is not of full rank. In this situation the vector $\underset{\sim}{\pi}_i$ is no longer estimable in the sense that the maximum likelihood (or, equivalently, least-squares) equations have an infinite number of solutions for $\underset{\sim}{\pi}_i$ within the simplex $\pi_{ij}>0 \ \forall j$, $\sum_{j=1}^{c_i-1} \pi_{ij}<1$. To estimate $\underset{\sim}{\pi}_i$ uniquely it is necessary to further restrict the set of admissible estimates. The usual least-squares theory does not help here, for it suggests augmenting M_i with linear restraints on $\underset{\sim}{\pi}_i$ to make M_i of full rank, while such <u>linear</u> restraints on the multinomial parameters are seldom appropriate to the experimental situation or its interpretation. However, we can still estimate $\underset{\sim}{\pi}_i$ fairly simply under some non-linear restrictions, such as the absence of certain higher order interactions, which are often reasonable.

For the sake of clarity, we assume now that I=1 and eliminate i from subscripts. Suppose rank M = m<c-1. Let $\underset{(\sum_{\ell=1}^{k}(c_\ell-1))\times m}{M^*}$ be a matrix of linearly independent columns of M.

Let $\underset{m\times(c-1)}{T}$ be a matrix of full rank satisfying M*T = M, and let $\underset{\sim}{\Pi}^* = T\underset{\sim}{\Pi}$, so $E\underset{\sim}{p}_d = M^*\underset{\sim}{\Pi}^*$. The equations (4.2.2) are then

$$T'M^{*\prime}V^{-1}(\underset{\sim}{\pi}_d)(M^*\underset{\sim}{\Pi}^*-\underset{\sim}{p}_d) = \underset{\sim}{0}. \qquad (4.2.6)$$

Now, since T is of full rank, there is a matrix $U^{m\times(c-1)}$ such that $UT' = I_m$. Thus, (4.2.6) is equivalent to $M^{*\prime}V^{-1}(\underset{\sim}{\pi}_d)(M^*\underset{\sim}{\Pi}^*-\underset{\sim}{p}_d) = \underset{\sim}{0}$, or

$$(M^{*\prime}V^{-1}(\underset{\sim}{\pi}_d)M^*)\underset{\sim}{\Pi}^* = M^{*\prime}V^{-1}(\underset{\sim}{\pi}_d)\underset{\sim}{p}_d. \qquad (4.2.7)$$

From (4.2.7) $\underset{\sim}{\Pi}^*$ may be estimated just as $\underset{\sim}{\Pi}$ when M is of full rank. Then if R is any set of restraints which determine a unique solution for $\underset{\sim}{\Pi}$ in terms of $\underset{\sim}{\Pi}^*$, the estimate for $\underset{\sim}{\Pi}^*$ determined from (4.2.7) can be used to solve for an estimate of $\underset{\sim}{\Pi}$. If the a) minimum-χ^2_1 or b) maximum likelihood estimate of $\underset{\sim}{\Pi}^*$ is determined from (4.2.7), then R uniquely determines the a) minimum-χ^2_1 or b) maximum likelihood estimate of $\underset{\sim}{\Pi}$, respectively, under R. For example, in an experiment dealing with three dichotomous response variables A, B, and C, A and B may be measured on one sample, B and C on another, A and C on a third. Under the assumption R of no second order interaction (see e.g., Bhapkar and Koch [18]), the joint probabilities corresponding to the three responses are uniquely determined by $\underset{\sim}{\Pi}^* = \underset{\sim}{\pi}_d$.

CHAPTER V

SMOOTHING PROCEDURES USING LINEAR MODELS

5.1 Fitting Models to Incomplete Data

So far we have focused on the problem of estimating the cell probabilities of an underlying contingency table using samples from a number of its marginal distributions. Often this is only the first, or even an unnecessary step, in obtaining an analysis of the data which answers the questions of real interest. Quite generally these questions can be formulated in terms of an "explanatory" linear model for the cell probabilities or certain functions thereof. We now present methods for fitting such models to general categorical data arrays, using Neyman's-χ^2_1 and related statistics. First the basic approach to complete data, as found in Grizzle, Starmer, and Koch [31], will be summarized; an extension to incomplete data, essentially due to Reinfurt [9], will then be discussed. The Appendix contains some remarks on the historical background of these methods.

In the notation of Chapter IV, the complete data situation is characterized by $k_i=1$, $M_i=I_{c_i-1}$, $\forall$ i. Thus, $\underset{\sim}{\pi}_d = \underset{\sim}{\Pi}$ and we may write $\underset{\sim}{p}_d=\underset{\sim}{p}$. Grizzle, Starmer and Koch assume additionally that $c_i=c$ $\forall$ i, an inessential restriction we omit. Consider the linear model

$$\underset{m\times 1}{\underset{\sim}{F}} \quad (\underset{\sim}{\Pi}) = \underset{m\times b}{X} \quad \underset{b\times 1}{\underset{\sim}{\beta}} \tag{5.1.1}$$

where X is an arbitrary known matrix of full rank $b \leq m$, $\underset{\sim}{\beta}$ a vector of unknown parameters, and $\underset{\sim}{F} = (f_1,\ldots,f_m)'$ a vector of functions of Π, $m \leq \sum_{i=1}^{I} (c_i-1) \equiv C$, each possessing partial derivatives up to the second order with respect to the elements of $\underset{\sim}{\Pi}$ within some neighborhood of $\underset{\sim}{\Pi}$. Let $\underset{m\times C}{H}(\underset{\sim}{z})$ be the matrix

$$\left.\frac{d\,\underset{\sim}{F}(\underset{\sim}{x})}{d\underset{\sim}{x}}\right|_{\underset{\sim}{x}=\underset{\sim}{z}}.$$

It is assumed that the functions $f_1,\ldots,f_m$ are independent in the sense that rank $H(\underset{\sim}{z}) = m$ for $\underset{\sim}{z}$ in some neighborhood of $\underset{\sim}{\Pi}$. Let

$$S(\underset{\sim}{z}) = H(\underset{\sim}{z})V(\underset{\sim}{z})H'(\underset{\sim}{z}).$$

Then:

i) $\underset{\sim}{p}$ is the maximum likelihood estimate of $\underset{\sim}{\Pi}$, with covariance matrix $V(\underset{\sim}{\Pi})$;

ii) $V(\underset{\sim}{p})$ is asymptotically non-singular and consistent for $V(\underset{\sim}{\Pi})$;

iii) $\underset{\sim}{F}(\underset{\sim}{p})$ is BAN for $\underset{\sim}{F}(\underset{\sim}{\Pi})$, with asymptotic covariance matrix $S(\underset{\sim}{\Pi})$; and

iv) $S(\underset{\sim}{p})$ is asymptotically non-singular and consistent for $S(\underset{\sim}{\Pi})$.

Further, if (5.1.1) holds, then:

v) $\underset{\sim}{b} = (X'S^{-1}(\underset{\sim}{p})X)^{-1}X'S^{-1}(\underset{\sim}{p})\underset{\sim}{F}(\underset{\sim}{p})$ is BAN for $\underset{\sim}{\beta}$, with asymptotic covariance matrix $\mathrm{Cov}(\underset{\sim}{b}) = (X'S^{-1}(\underset{\sim}{\Pi})X)^{-1}$;

v') in the special case $X=0$, $\tilde{\underset{\sim}{\Pi}} = \underset{\sim}{p} - V(\underset{\sim}{p})H'(\underset{\sim}{p})S^{-1}(\underset{\sim}{p})\underset{\sim}{F}(\underset{\sim}{p})$ is BAN for $\underset{\sim}{\Pi}$;

vi) $(X'S^{-1}(\underset{\sim}{p})X)^{-1}$ is asymptotically non-singular and consistent for $\mathrm{Cov}(\underset{\sim}{b})$;

vii) $\underset{\sim}{F}'(\underset{\sim}{p})S^{-1}(\underset{\sim}{p})\underset{\sim}{F}(\underset{\sim}{p}) - \underset{\sim}{b}'(X'S^{-1}(\underset{\sim}{p})X)^{-1}\underset{\sim}{b}$ is asymptotically χ^2_{m-b};

viii) for arbitrary $\underset{d\times b}{C}$ of rank $d<b$, $C\underset{\sim}{b}$ is BAN for $C\underset{\sim}{\beta}$, with asymptotic covariance matrix $C(X'S^{-1}(\underset{\sim}{\Pi})X)^{-1}C' = \text{Cov}\,(C\underset{\sim}{b})$;

ix) $C(X'S^{-1}(\underset{\sim}{p})X)^{-1}C'$ is asymptotically non-singular and consistent for $\text{Cov}(C\underset{\sim}{b})$;

x) hence, if

$$C\underset{\sim}{\beta} = \underset{\sim}{0} \tag{5.1.2}$$

then $(C\underset{\sim}{b})'[C(X'S^{-1}(\underset{\sim}{p})X)^{-1}C']^{-1}(C\underset{\sim}{b})$ is asymptotically χ^2_d;

xi) tests of (5.1.1) using vi) or (5.1.2) using x) are asymptotically equivalent to the corresponding likelihood ratio tests in the sense that Pr{test disagrees with likelihood ratio test} → 0 whether (5.1.1) or (5.1.2), respectively, hold or not.

In the general formulation to include incomplete data problems, $\underset{\sim}{p}$ is replaced by $\underset{\sim}{p}_d$, I_{c_i-1} by M_i as defined in Chapter IV. The observation vector $\underset{\sim}{p}_d$ is no longer the m.l.e. for $\underset{\sim}{\Pi}$, though $V(\underset{\sim}{p}_d)$ remains a consistent estimator for $V(\underset{\sim}{\pi}_d)$, the covariance matrix of $\underset{\sim}{p}_d$. Furthermore, $\underset{\sim}{F}(\underset{\sim}{\Pi})$ no longer has a natural BAN estimator, since $\underset{\sim}{F}(\underset{\sim}{p}_d)$ may not even be defined. However, let $\hat{\underset{\sim}{\Pi}}$ be any BAN estimator of $\underset{\sim}{\Pi}$ in the general situation, and write $\hat{\underset{\sim}{\pi}}_d = M\hat{\underset{\sim}{\Pi}}$. Then i) - iv) may be modified to:

i)' $\hat{\underset{\sim}{\Pi}}$ is BAN for $\underset{\sim}{\Pi}$, with asymptotic covariance matrix $(M'V^{-1}(\underset{\sim}{\pi}_d)M)^{-1} = V_d(\underset{\sim}{\pi}_d)$;

ii)' $V_d(\hat{\underset{\sim}{\pi}}_d)$ is asymptotically non-singular and consistent for $V_d(\underset{\sim}{\pi}_d)$;

iii)' $\underset{\sim}{F}(\hat{\underset{\sim}{\Pi}})$ is BAN for $\underset{\sim}{F}(\underset{\sim}{\Pi})$, with asymptotic covariance matrix $H(\underset{\sim}{\Pi})V_d(\underset{\sim}{\pi}_d)H'(\underset{\sim}{\Pi}) = S_d(\underset{\sim}{\Pi})$;

iv)' $S_d(\hat{\underset{\sim}{\Pi}})$ is asymptotically non-singular and consistent for $S_d(\underset{\sim}{\Pi})$;

while v) to xi) are valid if $\underset{\sim}{p}$ is replaced by $\hat{\underset{\sim}{\Pi}}$ and $S(\underset{\sim}{p})$ by $S_d(\hat{\underset{\sim}{\Pi}})$. Hence, any BAN estimator from the previous chapters may be used in the incomplete data situation to perform an analysis similar to any given by [31] for complete data.

Sometimes a given model $\underset{\sim}{F}(\underset{\sim}{\Pi}) = X\underset{\sim}{\beta}$ may be expressed equivalently as a second model in terms of $\underset{\sim}{\pi}_d$, viz., $F^*(\underset{\sim}{\pi}_d) = X^*\underset{\sim}{\beta}$. In such cases a second approach is possible, by which one can fit the model without explicitly estimating $\underset{\sim}{\Pi}$. Specifically, $\underset{\sim}{F}^*(\underset{\sim}{p}_d)$ is a consistent estimator of $F^*(\underset{\sim}{\pi}_d)$, with asymptotic covariance matrix $S^*(\underset{\sim}{\pi}_d) = H^*(\underset{\sim}{\pi}_d)V(\underset{\sim}{\pi}_d)H^{*\prime}(\underset{\sim}{\pi}_d)$, where

$$H^*(\underset{\sim}{z}) = \left.\frac{d\underset{\sim}{F}^*(\underset{\sim}{x})}{d\underset{\sim}{x}}\right|_{\underset{\sim}{x}=\underset{\sim}{z}}.$$

$S^*(\underset{\sim}{p}_d)$ is also consistent for $S^*(\underset{\sim}{\pi}_d)$. Hence, analysis may proceed essentially as in i) - xi) with $\underset{\sim}{p}_d$, $\underset{\sim}{\pi}_d$ replacing $\underset{\sim}{p}$, $\underset{\sim}{\Pi}$.

The simplest case is when $\underset{\sim}{F}(\underset{\sim}{\Pi}) = \underset{\sim}{\Pi}$, $\underset{\sim}{F}^*(\underset{\sim}{\pi}_d) = \underset{\sim}{\pi}_d$, $X^* = MX$. Directly fitting

$$E\underset{\sim}{p}_d = (MX)\underset{\sim}{\beta}$$

we obtain

$$\hat{\underset{\sim}{\beta}} = \underset{\sim}{b}_1 = \{(MX)'V^{-1}(\underset{\sim}{p}_d)(MX)\}^{-1}(MX)'V^{-1}(\underset{\sim}{p}_d)\underset{\sim}{p}_d,$$

and $\hat{\hat{\underset{\sim}{\Pi}}}_1 = X\underset{\sim}{b}_1$, with estimated covariance matrix $X\{(MX)'V^{-1}(\underset{\sim}{p}_d)MX\}^{-1}X'$,

to estimate $\underset{\sim}{\Pi}$ under the model. Alternatively, using a BAN estimate $\hat{\Pi}$ obtained as in Chapter IV gives

$$\hat{\underset{\sim}{\beta}} = \underset{\sim}{b}_2 = \{(MX)'V^{-1}(\hat{\underset{\sim}{\pi}}_d)(MX)\}^{-1}(MX)'V^{-1}(\hat{\underset{\sim}{\pi}}_d)\hat{\underset{\sim}{\pi}}_d,$$

$\hat{\hat{\underset{\sim}{\Pi}}}_2 = X\underset{\sim}{b}_2$, with estimated covariance matrix $X\{(MX)'V^{-1}(\hat{\underset{\sim}{\pi}}_d)(MX)\}^{-1}X'$ to estimate $\underset{\sim}{\Pi}$ under the model. Similarly, the test statistics relevant to $\underset{\sim}{b}$, use $\underset{\sim}{p}_d$ to estimate $\underset{\sim}{\pi}_d$, those related to $\underset{\sim}{b}_2$ use $\hat{\underset{\sim}{\pi}}_d$. All these estimates are BAN for the related parameters. It might be expected that $\underset{\sim}{b}_2$ and $\hat{\hat{\underset{\sim}{\pi}}}_2$ would be superior in that they use a BAN estimate of the covariance matrix rather than just a consistent estimate; however, considerations of bias and rate of approach to the limiting distribution as affected by bias may vitiate this advantage, and further study is needed to validly compare the two procedures. The "quick and dirty" method utilizing $\underset{\sim}{b}_1$ has the obvious advantages of great conceptual simplicity along with a savings in computation.

5.2 An Example

The first eight rows of Table 5.2.1 display some marginal totals from a set of data previously analyzed by Dyke and Patterson [57], Bishop [27], Goodman [25], and several others. 1729 persons are classified as to their knowledge of cancer, and whether or not they read newspapers or engage in other "solid" reading (books or magazines). The persons described were selected from a large population by simple random sampling. The remainder of Table 5.2.1 represents artificial results from two hypothetical pilot studies relating cancer knowledge to newspaper and solid reading individually. A question of interest is how exposure to newspapers and/or solid reading related to knowledge

of cancer in the population, in terms of how the probability of good cancer knowledge varies between groups with different reading habits.

TABLE 5.2.1

RESPONSES TO CANCER SURVEYS
(* DENOTES UNMEASURED RESPONSE)

PATTERNS OF RESPONSE			NUMBER OF RESPONDENTS	
NEWSPAPERS	SOLID READING	CANCER KNOWLEDGE		TOTAL = 2639
YES	YES	GOOD	353	
YES	YES	POOR	270	
YES	NO	GOOD	125	
YES	NO	POOR	225	
NO	YES	GOOD	87	
NO	YES	POOR	110	
NO	NO	GOOD	103	
NO	NO	POOR	456	$n_I = 1729$
YES	*	GOOD	90	
YES	*	POOR	100	
NO	*	GOOD	40	
NO	*	POOR	110	$n_{II} = 340$
*	YES	GOOD	150	
*	YES	POOR	120	
*	NO	GOOD	80	
*	NO	POOR	220	$n_{III} = 570$

Let $\underset{\sim}{\Pi}' = (\pi_{11G}, \pi_{11P}, \pi_{10G}, \ldots, \pi_{00G})$ be the vector of probabilities corresponding to the first seven rows of Table 5.2.1, where 0 represents no, 1 represents yes, and G,P represent good and poor respectively. One way to attack the problem is to first obtain an estimate of $\underset{\sim}{\Pi}$ from all the data using the model

$$\underset{\sim}{p}_d = M\underset{\sim}{\Pi}$$

where

$$\underset{\sim}{p}_d' = \left(\frac{353}{1729}, \ldots, \frac{103}{1729}, \frac{90}{340}, \ldots, \frac{40}{340}, \frac{150}{570}, \ldots, \frac{80}{570}\right),$$

$$M = \begin{bmatrix} I_7 \\ X_1 \end{bmatrix}$$

and

$$X_1 = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

For example, the minimum-χ_1^2 estimator of $\underset{\sim}{\Pi}$ is

$$\hat{\underset{\sim}{\Pi}} = (.20433,.15529,.07278,.13024,.05127,.06284,.06105)'.$$

$V(M\hat{\underset{\sim}{\Pi}})$ may be used to estimate $\mathrm{Cov}(\hat{\underset{\sim}{\Pi}})$ by $(M'V^{-1}(M\hat{\underset{\sim}{\Pi}})M)^{-1}$. One next considers a model

$$\underset{\sim}{\pi}^c = X^c\underset{\sim}{\beta}, \tag{5.2.1}$$

where $(\underset{\sim}{\pi}^c)' = (\pi_{11}^c,\pi_{10}^c,\pi_{01}^c,\pi_{00}^c)$, π_{ij}^c denoting the conditional probability of good cancer knowledge gives categories i,j of newspaper and solid reading, respectively, $\underset{\sim}{\beta}' = (\mu,\beta_1,\beta_2)$, and

$$X^c = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix}.$$

Thus, μ is a mean and β_1, β_2 are newspaper and solid reading effects. If $\underset{\sim}{\Pi}^{*\prime} = (\underset{\sim}{\Pi}', 1-\underset{\sim}{j}_7'\underset{\sim}{\Pi})$, $\underset{\sim}{j}_k$ indicating a k-vector of 1's, then $\underset{\sim}{\pi}^c$ may be written as

$$\underset{\sim}{\pi}^c = \exp\{K \log A\underset{\sim}{\Pi}^*\}$$

with

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \otimes I_4, \qquad K = [1 \quad -1] \otimes I_4,$$

$\otimes$ indicating Kronecker matrix product, and the vector functions of a vector log and exp taking their usual elementwise definitions. So $\underset{\sim}{\pi}^c$ may be estimated by

$$\hat{\underset{\sim}{\pi}}^c = \exp\{K \log A\hat{\underset{\sim}{\pi}}^*\}, \qquad \hat{\underset{\sim}{\Pi}}^* = (\hat{\underset{\sim}{\pi}}', 1-\underset{\sim}{j}'\hat{\underset{\sim}{\pi}})'.$$

Further, the asymptotic covariance matrix of $\hat{\underset{\sim}{\Pi}}^*$ is then

$$V^*(\underset{\sim}{\pi}_d) = \begin{bmatrix} V(\underset{\sim}{\pi}_d) & x(\underset{\sim}{\pi}_d) \\ y'(\underset{\sim}{\pi}_d) & z(\underset{\sim}{\pi}_d) \end{bmatrix},$$

for $x(\underset{\sim}{\pi}_d) = -V(\underset{\sim}{\pi}_d)\underset{\sim}{j}_7$, $\underset{\sim}{y}'(\underset{\sim}{\pi}_d) = -\underset{\sim}{j}_7'V(\underset{\sim}{\pi}_d)$, $z(\underset{\sim}{\pi}_d) = \underset{\sim}{j}_7'(\underset{\sim}{\pi}_d)\underset{\sim}{j}_7$. Hence the limiting covariance matrix of $\hat{\underset{\sim}{\pi}}^c$ is

$$\mathrm{Cov}(\hat{\underset{\sim}{\pi}}^c) = D_2(\underset{\sim}{\Pi})K_1D_1^{-1}(\underset{\sim}{\Pi})AV^*(\underset{\sim}{\pi}_d)A'D_1^{-1}(\underset{\sim}{\Pi})K_1'D_2'(\underset{\sim}{\Pi}),$$

where D_1 and D_2 are diagonal matrices formed respectively from the vectors $A\underset{\sim}{\Pi}^*$, $\underset{\sim}{\pi}^c$. This may be estimated by

$$D_2(\hat{\underset{\sim}{\Pi}})K_1D_1^{-1}(\hat{\underset{\sim}{\Pi}})AV^*(\hat{\underset{\sim}{\pi}}_d)A'D_1^{-1}(\hat{\underset{\sim}{\Pi}})K_1'D_2^{-1}(\hat{\underset{\sim}{\Pi}}),$$

and the estimate used to fit the model (5.2.1). Test statistics for various hypotheses may be obtained as described in Section 1. The results of fitting (5.2.1) to the first eight rows of the table, and then to all the data, are shown in Table 5.2.2. Estimates in Table 5.2.2 are

BAN for the relevant parameters.

An alternative approach which is quicker involves fitting the model

$$\exp K \log A^* \, p_d = X\underset{\sim}{\beta} \tag{5.2.2}$$

where

$$A^* = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \otimes I_8$$

$$K = (1 \quad -1) \otimes I_8$$

$$\underset{\sim}{p}_d^* = \left(\frac{353}{1729}, \cdots, \frac{456}{1729}, \frac{90}{340}, \cdots, \frac{110}{340}, \frac{150}{570}, \cdots, \frac{220}{570}\right)$$

and

$$X^* = \begin{bmatrix} X \\ X_1 \end{bmatrix}$$

$$X_1 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix} .$$

The results of fitting this model are in Table 5.2.3. Either analysis shows that the data is consistent with the view that either type of reading has a linear, positive effect on the conditional probability of good knowledge of cancer, and that the effect of one type of reading is the same for both levels of the other type.

TABLE 5.2.2

ESTIMATED MODEL PRAMETERS AND COVARIANCE MATRIX FOR CANCER SURVEY, WITH TEST STATISTICS: FIRST METHOD

		ESTIMATES	ESTIMATED COVARIANCES × 10^4		
			μ	β_1	β_2
COMPLETE DATA ONLY	μ	.382	1.230	-.076	.308
	β_1	.078		1.480	-.695
	β_2	.115			1.553

χ_1^2 (residual) = .089

χ_1^2 (β_1=0) = 40.92

χ_1^2 (β_2=0) = 84.45

			μ	β_1	β_2
ALL DATA	μ	.385	.830	-.101	.228
	β_1	.076		1.275	-.623
	β_2	.115			1.220

χ_1^2 (residual) = 1.01

χ_1^2 (β_1=0) = 44.74

χ_1^2 (β_2=0) = 108.50

TABLE 5.2.3

ESTIMATED MODEL PARAMETERS AND COVARIANCE MATRIX FOR CANCER SURVEY, WITH TEST STATISTICS: SECOND METHOD

		ESTIMATES	ESTIMATED COVARIANCES × 10^4		
			μ	β_1	β_2
ALL DATA	μ	.387	.818	-.034	.183
	β_1	.079		1.147	-.410
	β_2	.121			1.074

χ^2_5 (residual) = 5.26

χ^2_1 (β_1=0) = 55.09

χ^2_1 (β_2=0) = 135.85

PART II

APPLICATIONS

CHAPTER VI

BALANCED INCOMPLETE BLOCK AND RELATED DESIGNS

6.1 Introduction

The remainder of this work consists principally of applications of the methods of Chapters 4 and 5 to some categorical data configurations less familiar than the standard multi-dimensional contingency table. The large-sample linear model approach has the potential to be as useful for analyzing categorical data as the general linear models approach is for continuous data. The broad scope of the method opens the door to consideration of a wide variety of designs for large-scale surveys and experiments and raises questions as to the possible uses and efficiencies of known designs, and the construction of new designs, which have not been touched upon in the categorical data literature. Each of the examples given will suggest problems that beg further research not attempted here. The goal is to present a currently practical method of analysis for some complicated data sets, while indicating lines of research to clarify principles of design.

We begin with a brief discussion of incomplete block measurement designs. These designs make no provision for sampling from the core distribution involving jointly all variables of interest, but rather dictate sampling from some set of its lower dimensional margins. Such designs are useful when it is impossible to measure all variables on an

individual, or when doing so is excessively costly or tends to introduce response errors. The theorem of Chapter IV divides all designs into two classes, those which allow estimation of the core distribution, and those which do not. It can easily be seen that balanced incomplete block measurement designs, for instance, fall into the second category unless certain higher order interactions are known to be zero. On the other hand, augmenting such a design with a small sample from the core distribution makes all joint probabilities estimable without any extra assumptions on the interactions. It is therefore necessary to view any design in the light of the parameters of interest and the additional assumptions, if any, necessary to make them uniquely estimable.

These remarks imply that balanced incomplete block measurement designs without augmentation, and without restrictive assumptions on joint probabilities, are of most interest in the context of split-plot experiments such as those described by Koch and Reinfurt [32], where the parameters of interest are functions of marginal probabilities estimable under the design. In the next section we illustrate by analyzing some artificial data.

6.2 A Split-Plot Drug Comparison

Grizzle, Starmer and Koch [31] consider the analysis of data from a drug experiment, and apply a test due to Bhapkar [22]. Favorable or unfavorable responses to separate administrations of three different drugs without carry-over effects were noted for each of 46 women. This data is given in Table 6.2.1 and is complete in the sense that response to each drug was measured for each of the women in the experiment.

When the primary objective is to compare the probabilities of favorable response associated with each of the three drugs, analysis involves using the three "matched" sample proportions of favorable response to make inferences about their expectations.

TABLE 6.2.1

RESPONSES TO DRUGS A, B, C
(1 DENOTES FAVORABLE RESPONSE, 0 DENOTES UNFAVORABLE RESPONSE)

PATTERNS OF RESPONSE			NUMBER OF RESPONDENTS
DRUG			
A	B	C	
1	1	1	6
1	1	0	16
1	0	1	2
1	0	0	4
0	1	1	2
0	1	0	4
0	0	1	6
0	0	0	6
TOTAL			$n_{ABC} = 46$

In the general drug comparison situation, medical reasons and/or limitations of time may preclude administering all drugs of interest sequentially to each of the members of a sample. It may be more appropriate to test on each individual only a proper subset of the set of drugs under study. The artificial data in Table 6.2.2 might represent the results of such a trial, collected with the same goal as the complete data above, and where an asterisk indicates that such persons did not receive the indicated drug. The design is obviously a very simple example of a balanced incomplete block design for three treatments (drugs), with the same number of persons receiving the set of drugs in any given block.

TABLE 6.2.2

RESPONSES TO DRUGS A, B, C
(1 DENOTES FAVORABLE, 0 DENOTES UNFAVORABLE,
* DENOTES THAT THIS DRUG WAS NOT ADMINISTERED)

PATTERNS OF RESPONSE			NUMBER OF RESPONDENTS
DRUG			
A	B	C	
1	1	*	12
1	0	*	3
0	1	*	3
0	0	*	7
SUBTOTAL			$n_{AB} = 25$
1	*	1	5
1	*	0	10
0	*	1	4
0	*	0	6
SUBTOTAL			$n_{AC} = 25$
*	1	1	4
*	1	0	12
*	0	1	4
*	0	0	5
SUBTOTAL			$n_{BC} = 25$
TOTAL			75

The problem in analyzing the data remains one of comparing "matched" proportions, but the matching design here is considerably more complex than for the data in Table 6.2.1. However, the linear models approach may be used to handle such a matching design fairly easily.

Specifically, if we write the marginal probabilities of interest as a vector $\underset{\sim}{\pi} = (\pi_A, \pi_B, \pi_C)'$, the design implies that

$$E A \underset{\sim}{p}_d = X\underset{\sim}{\pi} \qquad (6.2.1)$$

where $\underset{\sim}{p}_d$ is the vector of random variables underlying the observed data vector $\underset{\sim}{p}^*_d = \frac{1}{25}(10, 4, 4, 5, 10, 4, 4, 11, 5)'$, $A = \tilde{A} \otimes I_3$, $\tilde{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$, and

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Using $V(\underset{\sim}{p}^*_d)$ to estimate the covariance matrix of $\underset{\sim}{p}_d$, weighted least-squares applied to (6.2.1) produces the minimum-χ^2_1 estimate $\hat{\underset{\sim}{\pi}}$ of $\underset{\sim}{\pi}$, given in Table 6.2.3 with its estimated asymptotic covariance matrix $(X'V^{-1}(\underset{\sim}{p}^*_d)X)^{-1}$. A test to detect differences between any of the drugs may be generated from the matrix $C = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$, as under H_o: $\pi_A = \pi_B = \pi_C$ the statistic $(C\hat{\underset{\sim}{\pi}})'(C(X'V^{-1}(\underset{\sim}{p}^*_d)X)^{-1}C')^{-1}(C\hat{\underset{\sim}{\pi}})$ is asymptotically chi-square with two degrees of freedom. For this data the resulting test statistic is $\chi^2_2 = 9.12$, indicating significant differences between the drugs. Pairwise comparisons may be made using the vectors $C_1 = (1 \quad -1 \quad 0)$, $C_2 = (1 \quad 0 \quad -1)$, and $C_3 = (0 \quad 1 \quad -1)$ to generate analogous test statistics each with limiting chi-square distribution (one d.f.). Alternately, a multiple comparison procedure given by Goodman [58] may be used. This procedure will detect a difference between some pair of drugs exactly when the overall (two d.f.) significance test above rejects H_o.

TABLE 6.2.3

ESTIMATED PROBABILITIES OF FAVORABLE RESPONSE
TO DRUGS A, B, C, AND THEIR ESTIMATED COVARIANCE MATRIX

	ESTIMATES	ESTIMATED COVARIANCES $\times 10^3$		
		π_A	π_B	π_C
π_A	.606	4.47	1.26	-.27
π_B	.618		4.36	-.48
π_C	.341			4.43

More general designs based on balanced incomplete block designs may also be of interest, and in many instances can be handled with essentially the same analysis. For instance, suppose the data of Table 6.2.2 is augmented by data on other individuals treated with only a single drug, as in Table 6.2.4. Essentially this is data from a BIB design with block size two and another BIB design with block size one, applied to the same set of treatments. The problem is to make a reasonable combined analysis of all the data. There is no difficulty in doing this with the least-squares approach. Analogous to 6.2.1, the model is

$$E\ A_1\ \underset{\sim}{p}_d^{(1)} = X_1 \underset{\sim}{\pi} \tag{6.2.2}$$

where $\underset{\sim}{p}_d^{(1)}$ is the vector of random variables underlying the observed data vector

$$\underset{\sim}{p}_d^{(1)*} = \left(\frac{10}{25}, \frac{4}{25}, \frac{4}{25}, \frac{5}{25}, \cdots, \frac{5}{25}, \frac{9}{15}, \frac{11}{15}, \frac{5}{15}\right)',$$

TABLE 6.2.4

RESPONSES TO DRUGS A, B, C
(1 DENOTES FAVORABLE, 0 DENOTES UNFAVORABLE,
* DENOTES THAT THIS DRUG WAS NOT ADMINISTERED)

PATTERNS OF RESPONSE			NUMBER OF RESPONDENTS
DRUG			
A	B	C	
1	*	*	9
0	*	*	6
SUBTOTAL			$n_A = 15$
*	1	*	11
*	0	*	4
SUBTOTAL			$n_B = 15$
*	*	1	5
*	*	0	10
SUBTOTAL			$n_C = 15$
TOTAL			45

$$A_1 = \begin{bmatrix} \tilde{A} \otimes I_3 & \underset{6\times 3}{0} \\ \underset{3\times 9}{0} & I_3 \end{bmatrix}, \quad \text{and} \quad X_1 = \begin{bmatrix} X \\ I_3 \end{bmatrix}.$$

$V(\underset{\sim}{p}_d^{(1)*})$ is consistent for the covariance matrix of $\underset{\sim}{p}_d^{(1)}$, and weighted least-squares using these estimates of the weights produces $\hat{\underset{\sim}{\pi}}^{(1)}$, the minimum-$\chi_1^2$ estimate of $\underset{\sim}{\pi}$ using all the data. The estimate $\hat{\underset{\sim}{\pi}}^{(1)}$ and related estimated asymptotic covariance matrix for this data are shown in Table 6.2.5. The test statistic for H_o above is

$$(C\hat{\underset{\sim}{\pi}}^{(1)})'(C(X_1'V^{-1}(\underset{\sim}{p}_d^{(1)*})X_1)C')^{-1}(C\hat{\underset{\sim}{\pi}}^{(1)}) = 14.76,$$

which may be referred to the chi-square distribution (2 d.f.), and indicates significant differences between the drugs.

TABLE 6.2.5

ESTIMATED PROBABILITIES AND COVARIANCES, DRUGS A, B, C, OBTAINED FROM AUGMENTED DATA

ESTIMATES		ESTIMATED COVARIANCES × 10^3		
		$\underset{\sim}{\pi}_A$	$\underset{\sim}{\pi}_B$	$\underset{\sim}{\pi}_C$
$\underset{\sim}{\pi}_A$	.611	3.44	.74	-.14
$\underset{\sim}{\pi}_B$	.646		3.22	-.27
$\underset{\sim}{\pi}_C$	.337			3.40

6.3 Remarks on Asymptotic Covariances

We now calculate inverses of the asymptotic covariance matrices of $\underset{\sim}{\pi}$ and $\hat{\underset{\sim}{\pi}}^{(1)}$, or any other BAN estimators corresponding to the two designs of Section 6.2. For any known set of joint probabilities of responses to pairs of drugs, the asymptotic covariance matrix of the appropriate design may be calculated easily by matrix inversion; however, the general expressions for these matrices are too cumbersome to write down. Nevertheless, the form of the inverses is of some interest in itself, and in particular clarifies a point about $\hat{\underset{\sim}{\pi}}$ and BAN estimators in general. For simplicity we work things out only in terms of the specific two designs in the example, but some generality will be obvious.

The notation of Chapter II is used. Thus, π_{111} refers to the probability that a random individual would respond favorably to all three drugs A, B, and C if they were administered singly and under the conditions of the experiment, π_{222} to the probability of three unfavorable responses, and $\pi_A \equiv \pi_{100}$, $\pi_B \equiv \pi_{010}$, $\pi_C \equiv \pi_{001}$. Let n_k be the number of subjects in each block of the BIB design with block size k.

Now write $\underset{\sim}{p}_d$ in (6.2.1) as

$$\underset{1\times 9}{\underset{\sim}{p}_d} = ({}_1\underset{\sim}{p}_d', \underset{1\times 3}{{}_2\underset{\sim}{p}_d'}, \underset{1\times 3}{{}_3\underset{\sim}{p}_d'})'.$$

Then the ${}_i\underset{\sim}{p}_d$, i=1,2,3 are independent. The covariance matrix of ${}_1\underset{\sim}{p}_d$ is

$$v_1 = \frac{1}{n_2}\begin{pmatrix} \pi_{110}(1-\pi_{110}) & -\pi_{110}\pi_{120} & -\pi_{110}\pi_{210} \\ -\pi_{110}\pi_{120} & \pi_{120}(1-\pi_{120}) & -\pi_{120}\pi_{210} \\ -\pi_{110}\pi_{210} & -\pi_{120}\pi_{210} & \pi_{210}(1-\pi_{210}) \end{pmatrix},$$

and the (exact) covariance of $(\tilde{A}\ {}_1\underset{\sim}{p}_d)$ is thus

$$v_1^* = \tilde{A}\, v_1\, \tilde{A}' = \frac{1}{n_2}\begin{pmatrix} \pi_A(1-\pi_A) & \pi_{110}\pi_{220}-\pi_{120}\pi_{210} \\ \pi_{110}\pi_{220}-\pi_{120}\pi_{210} & \pi_B(1-\pi_B) \end{pmatrix}. \quad (6.3.1)$$

From similar results for ${}_2\underset{\sim}{p}_d$, ${}_3\underset{\sim}{p}_d$ one obtains

$$\mathrm{Cov}(A\underset{\sim}{p}_d) = \mathrm{diag}(v_i^*) \qquad i=1,2,3 \quad (6.3.2)$$

where

$$v_2^* = \frac{1}{n_2}\begin{pmatrix} \pi_A(1-\pi_A) & \pi_{101}\pi_{202}-\pi_{102}\pi_{201} \\ \pi_{101}\pi_{202}-\pi_{102}\pi_{201} & \pi_C(1-\pi_C) \end{pmatrix} \tag{6.3.3}$$

and

$$v_3^* = \frac{1}{n_2}\begin{pmatrix} \pi_B(1-\pi_B) & \pi_{011}\pi_{022}-\pi_{012}\pi_{021} \\ \pi_{011}\pi_{022}-\pi_{012}\pi_{021} & \pi_C(1-\pi_C) \end{pmatrix} . \tag{6.3.4}$$

Now the (asymptotic) covariance of the estimate $\hat{\underset{\sim}{\pi}} = (\hat{\pi}_A, \hat{\pi}_B, \hat{\pi}_C)'$ in 6.2 is the inverse of

$$X'(\mathrm{Cov}(A\underset{\sim}{p}_d))^{-1}X.$$

From (6.3.2), $(\mathrm{Cov}(A\underset{\sim}{p}_d))^{-1} = (\mathrm{diag}(v_i^*))^{-1}$, i=1,2,3 and obviously

$$(v_1^*)^{-1} = \frac{1}{\Delta_1}\begin{pmatrix} \pi_B(1-\pi_B) & \pi_{120}\pi_{210}-\pi_{110}\pi_{220} \\ \pi_{120}\pi_{210}-\pi_{110}\pi_{220} & \pi_A(1-\pi_A) \end{pmatrix} \tag{6.3.5}$$

with similar expressions holding for $(v_2^*)^{-1}$, $(v_3^*)^{-1}$, where $\Delta_i = \det(v_i^*)$, the determinant of the matrix v_i^*. Using these expressions and performing the indicated matrix operations determines that $(\mathrm{Cov}(\hat{\underset{\sim}{\pi}}))^{-1}$ =

$$n_2\begin{bmatrix} \frac{1}{\Delta_1}\pi_B(1-\pi_B)+\frac{1}{\Delta_2}\pi_C(1-\pi_C) & \frac{1}{\Delta_1}(\pi_{120}\pi_{210}-\pi_{110}\pi_{220}) & \frac{1}{\Delta_2}(\pi_{102}\pi_{201}-\pi_{101}\pi_{202}) \\ \frac{1}{\Delta_1}(\pi_{120}\pi_{210}-\pi_{110}\pi_{220}) & \frac{1}{\Delta_1}\pi_A(1-\pi_A)+\frac{1}{\Delta_3}\pi_C(1-\pi_C) & \frac{1}{\Delta_3}(\pi_{012}\pi_{021}-\pi_{011}\pi_{022}) \\ \frac{1}{\Delta_2}(\pi_{102}\pi_{201}-\pi_{101}\pi_{202}) & \frac{1}{\Delta_3}(\pi_{012}\pi_{021}-\pi_{011}\pi_{022}) & \frac{1}{\Delta_2}\pi_A(1-\pi_A)+\frac{1}{\Delta_3}\pi_B(1-\pi_B) \end{bmatrix}.$$

In the same way it is easily seen that

$$(\mathrm{Cov}(\hat{\underset{\sim}{\pi}}^{(1)}))^{-1} = (\mathrm{Cov}(\hat{\underset{\sim}{\pi}}))^{-1} + n_1 V^{(1)},$$

where

$$V^{(1)} = \begin{pmatrix} (\pi_A(1-\pi_A))^{-1} & 0 & 0 \\ 0 & (\pi_B(1-\pi_B))^{-1} & 0 \\ 0 & 0 & (\pi_C(1-\pi_C))^{-1} \end{pmatrix}.$$

As has been noted, general expressions for the matrices $\mathrm{Cov}(\hat{\underset{\sim}{\pi}})$ and $\mathrm{Cov}(\hat{\underset{\sim}{\pi}}^{(1)})$ are easy to obtain at this point but exceedingly complex to write down, and we will not do so. However, in the unlikely situation that the response of an individual to one of the drugs, say drug A, is pairwise independent of response to the other two drugs, B and C, the results are very simple. It is immediate from (6.3.6) and (6.3.7) that $\hat{\pi}_A$ is then (asymptotically) independent of $\hat{\pi}_B$ and $\hat{\pi}_C$, with (limiting) variance

$$\left\{ \frac{n_2}{\Delta_1} \pi_B(1-\pi_B) + \frac{n_2}{\Delta_2} \pi_C(1-\pi_C) \right\}^{-1} =$$

$$\left\{ \frac{n_2\pi_B(1-\pi_B)}{\pi_A\pi_B(1-\pi_A)(1-\pi_B)} + \frac{n_2\pi_C(1-\pi_C)}{\pi_A\pi_C(1-\pi_A)(1-\pi_C)} \right\}^{-1} = \frac{\pi_A(1-\pi_A)}{2n_2} \quad ,$$

while $\pi_A^{(1)}$ is (asymptotically) independent of $\pi_B^{(1)}$ and $\pi_C^{(1)}$, with (limiting) variance

$$((\mathrm{Var}(\hat{\pi}_A))^{-1} + \frac{n_1}{\pi_A(1-\pi_A)})^{-1} = \frac{\pi_A(1-\pi_A)}{n_1+2n_2} .$$

These are precisely the results obtainable, in each design, from estimating π_A by the proportion of all individuals who receive drug A that respond favorably to it, ignoring the matching design and all data referring to drugs B and C. Thus, one cannot lose in large samples by using the split-plot design and the estimate $\hat{\underset{\sim}{\pi}}^{(1)}$ even when the responses are independent and simple estimators of π_A, π_B, and π_C are available. This illustrates the well-known fact (see e.g., Kleinbaum [6]) that BAN estimators in a linear model with unknown covariance matrix achieve asymptotically the minimum variance obtainable using an unbiased linear estimator (linear function of the observations) in the situation when the covariance matrix is known.

CHAPTER VII

A CATEGORICAL ANALOGUE OF GROWTH-CURVE ANALYSIS

7.1 The General Growth-Curve Problem

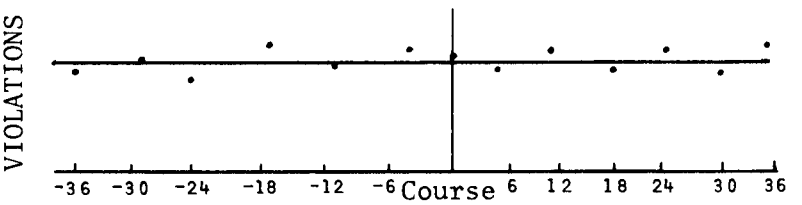
In many areas of research it is of interest to study the change in certain measurements over a period of time. The measurements on any experimental unit, or subject, may be regarded as functions of an underlying stochastic process associated with the unit. One usually wishes to make inferences about some of the margins, or finite-dimensional distributions, of the processes associated with certain single units or sets of units. The problem is complicated by the fact that, as in a split-plot situation, measurements on the same subject at different times will usually be correlated, and the correlation structure will seldom be known.

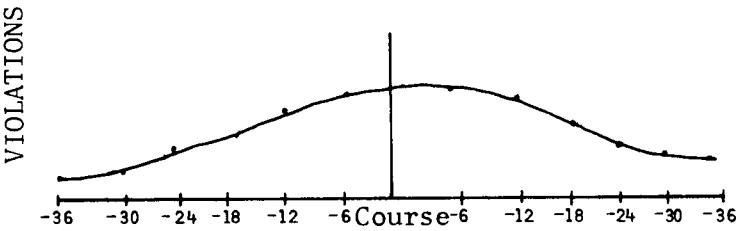
Longitudinal studies of this sort involving continuous data have been discussed by many authors, Box [59], Elston and Grizzle [60], Allen and Grizzle [61], Khatri [62], Kleinbaum [6], Potthoff and Roy [63], Roy [64], Roy, Gnanadesikan and Srivastava [7], and Rao [65, 66, 67] being several among them. A common, attractive procedure for handling such data involves initially a reduction and reparametrization by fitting polynomials or other regular functions of time to the series of measurements associated with each individual, and subsequently an analysis of the coefficients of these functions in terms of factors differentiating

sets of individuals. Grizzle [68] gives a particularly clear example of such an analysis.

Categorical data of this type may arise just as easily as continuous data. For instance, North Carolina drivers with a record of repeated accidents and/or violations are encouraged to take a driver training course administered by the state. In an attempt to evaluate the effect of this course, the University of North Carolina Highway Safety Research Center is compiling records of the numbers of accidents and violations incurred by drivers in each six-month period for three years before and after taking this course. Drivers may be separated by demographic variables such as sex and race into populations with possibly different driving characteristics. The Center is interested in overall response to the course as measured by accidents and violations, and in differences in response between the demographic groups. Since the numbers of violations and accidents incurred in any six-month period, even by problem drivers, is likely to be small and often zero, this is most easily studied in a categorical framework. Some possible response curves that might result from this study, plotting frequency of accidents and/or violations over time, are shown in Figure 7.1.1.

When the relevant measurements are categorical in nature, and all measurements are determined for each subject at each of a number of times, data for any set of subjects may be expressed as a multidimensional contingency table in which time is one of the dimensions. The data may then be analyzed as a categorical data mixed model or split-plot situation as defined by Reinfurt [9], or by maximum-likelihood methods such as those of Bishop [27] for dichotomous variables. The full table may also be





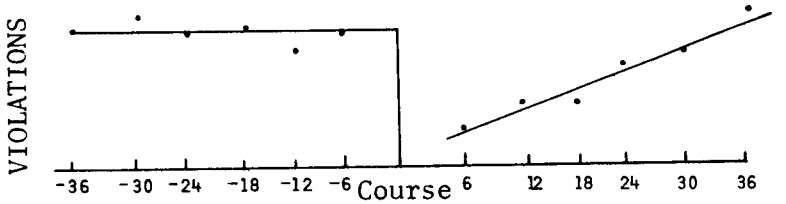


FIGURE 7.1.1

POSSIBLE TRENDS OF VIOLATIONS
FOR DRIVERS TAKING COURSE

split into several tables corresponding to different time periods and analyzed separately at each time. The vector of cell counts for a subgroup of individuals at any given time period corresponds to the multivariate observation on an individual at a given time period in the analysis for continuous data. The question of interest involving the time dimension is how the probabilities underlying these cell counts, or how particular functions of these probabilities, vary over time.

Within the framework of such a problem the concept of incompleteness has two aspects:

i) at some fixed time or set of times, not all individuals may be measured with respect to all variables of interest; and

ii) not all individuals may be measured, on each variable of interest, of every time point at which measurements are taken.

Incompleteness in either of these two senses or in combinations of them makes the data difficult to handle as a full or split-contingency table, but analysis using a Neyman-χ^2_1 linear modeling approach can be fairly straightforward.

7.2 Application to Categorical Data

To amplify on the preceding discussion, and clarify how the growth curve concept may be applied in a categorical framework, we will consider some artificial data which is complete and show how linear model techniques may be used to provide a compact analysis. The data will then be augmented in such a way as to make it incomplete in

the sense of ii) above, and a modified analysis will be given to account for incompleteness of the full set of data.

Suppose the elements of a random sample of individuals from some population whose members have abnormally high serum cholesterol readings are randomly assigned to one of four diets (I, II, III, IV) designed to reduce their cholesterol counts. After one, two, and three months on the respective diets, blood samples are taken from each person and classified as to whether cholesterol is present in normal or abnormal quantity. The results may be displayed as in Table 7.2.1. The problem is an example of a categorical data mixed model. Interest centers on the marginal probabilities of having a normal cholesterol reading at the end of each month on a diet, how these change from month to month, and on the differences of behavior of serum cholesterol for persons adhering to the different diets.

TABLE 7.2.1

RESULTS OF SERUM CHOLESTEROL TESTS
(N=NORMAL, A=ABNORMAL)

PATTERNS OF RESPONSE TIME PERIOD			NUMBER OF RESPONDENTS DIET			
0	1	2	I	II	III	IV
N	N	N	2	7	16	31
N	N	A	2	2	13	0
N	A	N	8	5	9	6
N	A	A	9	2	3	0
A	N	N	9	31	14	22
A	N	A	15	5	4	2
A	A	N	27	32	15	9
A	A	A	28	6	6	0
TOTAL (ν_i)			100	90	80	70

To analyze this data one may model the logits of the marginal probabilities of normal cholesterol at the end of a given month, for each month and each of the four diets. The vector of these logits can be expressed as

$$\underset{12\times 1}{\underset{\sim}{\ell}} = K \log A \underset{\sim}{p}$$

where

$$\underset{24\times 32}{A} = A^* \otimes I_4,$$

$$\underset{6\times 8}{A^*} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix},$$

$\underset{12\times 24}{K} = \{k_{ij}\}$ where

$$k_{ij} = \begin{cases} 1 & \text{if } j=2i-1 \\ -1 & \text{if } j=2i \\ 0 & \text{otherwise} \end{cases},$$

and

$$\underset{\sim}{p} = (\underset{8\times 1}{\underset{\sim}{p}_1'}, \underset{8\times 1}{\underset{\sim}{p}_2'}, \underset{8\times 1}{\underset{\sim}{p}_3'}, \underset{8\times 1}{\underset{\sim}{p}_4'})' = (\frac{2}{100}, \frac{2}{100}, \ldots, \frac{28}{100}, \frac{7}{90}, \ldots, \frac{9}{70}, \frac{0}{70})'.$$

The asymptotic covariance matrix of the vector $\underset{\sim}{\ell}$ may thus be estimated by

$$S = KD^{-1}_{(A\underset{\sim}{p})} AVA'D^{-1}_{(A\underset{\sim}{p})} K'$$

where $V = \text{diag}(v_i)$, $i=1,2,3,4$ and $v_i = \frac{1}{\nu_i}(D_{\underset{\sim}{p}_i} - \underset{\sim}{p}_i\underset{\sim}{p}_i')$. If one makes the preliminary assumption that the time trend of the logits for each diet

is linear, it is reasonable to fit a model $\underset{\sim}{\ell} = X\underset{\sim}{\beta}$, where

Model I:

$$\underset{12\times 8}{X} = \underset{3\times 2}{X^*} \otimes I_4, \quad X^* = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix},$$

and

$$\underset{8\times 1}{\underset{\sim}{\beta}} = (\beta_1, \beta_2, \ldots, \beta_8)'.$$

Under this model the minimum-χ^2_1 estimate of $\underset{\sim}{\beta}$ is

$$\underset{\sim}{b} = (-1.411, .597, -1.548, 1.570, .045, .335, .043, 1.441)',$$

and the Neyman-χ^2_1 statistic testing fit of the model is χ^2 (lack of fit) = 1.60 with one degree of freedom. Since the data thus do not contradict the preliminary hypothesis of linearity in the logits, it is appropriate to make tests directed at simplifying the model. Pairwise tests for differences of slopes in the regression lines are shown in Table 7.2.2. The first statistic is obtained using the C vector (1,0,-1,0,0,0,0,0) and the rest are obtained similarly.

The results suggest fitting a condensed

Model II:

$$\underset{12\times 6}{X_1} = [{}_1X_1, {}_2X_1], \quad \underset{12\times 4}{{}_1X_1} = \underset{\sim}{j}_3 \otimes I_4,$$

where $\underset{\sim}{j}_i$ is an i-vector of 1's,

TABLE 7.2.2

PAIRWISE COMPARISONS OF REGRESSION COEFFICIENTS
FOR UNSUPPLEMENTED DATA:
NEYMAN-χ^2_1 STATISTICS WITH 1 D.F.

DIETS COMPARED	COMPARISON OF INTERCEPTS	SLOPES
MODEL I		
I-II		14.06
I-III		1.29
I-IV		8.25
II-III		21.21
II-IV		.16
III-IV		13.45
MODEL II		
(I,III)-(II,IV)		28.02
I-II	.62	
I-III	43.26	
II-IV	22.10	
II-III	24.90	
II-IV	34.91	
III-IV	.16	
MODEL III		
(I,III)-(II,IV)		59.12
(I,II)-(III,IV)	77.02	

$$_2X_1 = \begin{pmatrix}0\\1\\2\end{pmatrix} \otimes I_2 \otimes \underset{\sim}{j}_2,$$

and $\underset{\sim}{\beta} = (\beta_1,\ldots,\beta_6)'$. For this model, the estimated parameter vector is

$$\underset{\sim}{b}_1 = (-1.278,-1.496,-.103,-.005,.477,1.517)',$$

and χ^2 (lack of fit) = 3.04 with six degrees of freedom, which is non-significant.

The two remaining slope parameters in Model II are significantly different, but pairwise comparisons of the intercepts, shown in

Table 7.2.2, suggest a further simplification to

Model III:

$$X_{2} = ({}_1X_2, {}_2X_2), \quad {}_1X_2 = j_6 \otimes I_2, \quad {}_2X_2 = {}_2X_1,$$
$$12\times 4$$

and $\underset{\sim}{\beta} = (\beta_1,\beta_2,\beta_3,\beta_4)'$. Here χ^2 (lack of fit) = 4.20 with eight degrees of freedom, so that this parsimonious model still provides a good fit to the data. The estimate of $\underset{\sim}{\beta}$ is $\underset{\sim}{b}_2 = (-1.352,-.072,.494,1.455)'$, and the two slope parameters are significantly different, as are the intercepts. This final model compactly describes the data and indicates that diet IV is preferable for reducing serum cholesterol to normal levels by virtue of its strong initial effect and rapid continuing gain over the three month period, relative to the other three diets.

Now, suppose we have incomplete information available on a further 100 persons adhering to diet I, in the sense that, by design, 25 people give blood samples only after the first month, while the remaining 75 are tested only two times, 25 for each possible pair of months. Such data is shown in Table 7.2.3, where * indicates that such people gave no sample at the time specified. To incorporate this data into the above models for the complete data, one may proceed in either of two ways, as described in Chapter V. One approach would be to first obtain estimates of the joint probabilities of different sets of three responses for the people on diet I from the model

$$E\ \underset{\sim}{n} = Z\underset{\sim}{\pi}_I$$

where $\underset{\sim}{n}$ is the vector of random variables underlying the observed data

vector $\underset{\sim}{n}^* = (2,2,8,9,9,15,27,1,4,6,1,3,7,3,4,8,6)'$, $\underset{\sim}{\pi}_I$ is the corresponding vector of response probabilities for people on diet I, and Z is a matrix of zeros and the population totals 100 and 25 in the appropriate places. The least-squares solution for $\underset{\sim}{\pi}_I$ from this model is the minimum-χ^2_I estimate of $\underset{\sim}{\pi}_I$; iterating the least-squares solution as in Chapter IV produces the maximum-likelihood estimate of $\underset{\sim}{\pi}_I$. The covariance matrix of either estimate may be itself estimated as in Chapter IV. If $\underset{\sim}{p}_1$ above is replaced by a vector whose first seven elements are those of $\hat{\underset{\sim}{\pi}}_I$ obtained as above, and whose eighth element is $(1-\underset{\sim}{j}_3\hat{\underset{\sim}{\pi}}_I)$, and v_1 is replaced by the estimated covariance of this vector, logits may be generated and modeled precisely as above for the complete, unaugmented data.

TABLE 7.2.3

RESULTS OF SERUM CHOLESTEROL TESTS:
SUBJECTS ON DIET I
(N=NORMAL, A=ABNORMAL, *=NO SAMPLE TAKEN)

PATTERNS OF RESPONSE TIME PERIOD			NUMBER OF RESPONDENTS
0	1	2	
N	N	*	1
N	A	*	4
A	N	*	6
A	A	*	14
N	*	N	1
N	*	A	3
A	*	N	7
A	*	A	14
*	N	N	3
*	N	A	4
*	A	N	8
*	A	A	10
N	*	*	6
A	*	*	19

A second approach, characterized in Chapter V as "quick and dirty", will be explicitly carried out here. Let $\underset{\sim}{p}_G$ be the vector $\underset{\sim}{p}_G = (\underset{\sim}{p}', \underset{\sim}{p}_5')'$, where $\underset{\sim}{p}_5 = (\frac{1}{25}, \ldots, \frac{14}{25}, \frac{1}{25}, \ldots, \frac{14}{25}, \frac{3}{25}, \ldots, \frac{10}{25}, \frac{6}{25}, \frac{19}{25})$, and let $\underset{\sim}{\ell}_G = K_G \log A_G \underset{\sim}{p}_G$, $\underset{19\times 38}{K_G}$ defined as was K, and

$$A_G = \begin{bmatrix} A & 0 & 0 \\ 0 & A_1 & 0 \\ 0 & 0 & I_2 \end{bmatrix}, \quad A_1 = A_1^* \otimes I_3,$$

$$A_1^* = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Then the considerations which suggest fitting Model I to $\underset{\sim}{p}$ imply that an appropriate model for $\underset{\sim}{p}_G$ is

Model IA

Adjoin to the bottom of X in Model I

$$X_A = \begin{bmatrix} 1 & 0 & \\ 1 & 1 & \\ 1 & 0 & 0 \\ 1 & 2 & 7\times 6 \\ 1 & 1 & \\ 1 & 2 & \\ 1 & 0 & \end{bmatrix}.$$

Obtain χ^2 (lack of fit) = 3.61 with 11 degrees of freedom, $\underset{\sim}{b}_A = (-1.400, .549, -1.548, 1.570, .045, .335, .043, 1.441)'$. The model evidently fits rather well. Comparisons of the trends, shown in Table 7.2.4 leads to

TABLE 7.2.4

PAIRWISE COMPARISONS OF REGRESSION COEFFICIENTS FOR SUPPLEMENTED DATA: NEYMAN-χ^2_1 STATISTICS WITH 1 D.F.

DIETS COMPARED	COMPARISON OF INTERCEPTS	SLOPES
MODEL IA		
I-II		17.81
I-III		1.03
I-IV		10.26
II-III		21.21
II-IV		.16
III-IV		13.45
MODEL IIA		
(I,III)-(II,IV)		30.64
I-II	.46	
I-III	55.18	
I-IV	27.16	
II-III	26.15	
II-IV	34.92	
III-IV	.15	
MODEL IIIA		
(I,III)-(II,IV)		66.09
(I,II)-(III,IV)	89.74	

Model IIA:

Adjoin to the bottom of X_1

$$X_{1A} = \begin{bmatrix} 1 & & 0 & \\ 1 & & 1 & \\ 1 & & 0 & \\ 1 & 0_{7\times3} & 2 & 0_{7\times1} \\ 1 & & 1 & \\ 1 & & 2 & \\ 1 & & 0 & \end{bmatrix}.$$

Obtain χ^2 (lack of fit) = 4.80 with 13 degrees of freedom,

$$\underset{\sim}{b}_{1A} = (-1.320,-1.496,-.100,.005,.474,1.517)'.$$

Results of comparisons (Table 7.2.4) leads to

Model IIIA

Adjoin to the bottom of X_2

$$X_{2A} = \begin{bmatrix} 1 & & 0 & \\ 1 & & 1 & \\ 1 & & 0 & \\ 1 & 0 \ 7\times 1 & 2 & 0 \ 7\times 1 \\ 1 & & 1 & \\ 1 & & 2 & \\ 1 & & 0 & \end{bmatrix}.$$

Obtain χ^2 (lack of fit) = 5.69 with 15 degrees of freedom,

$\underset{\sim}{b}_{2A} = (-1.365,-.065,.485,1.461)$, with estimated covariance matrix

$$V_{\underset{\sim}{b}_{2A}} = \begin{bmatrix} 1.426 & .628 & -.700 & -.818 \\ .628 & 1.713 & -.614 & -.564 \\ -.700 & -.614 & .763 & .445 \\ -.818 & -.564 & .445 & 1.569 \end{bmatrix} \times 10^{-2}.$$

This is the final model. Conclusions are the same here as for the unaugmented data.

CHAPTER VIII

A GRAECO-LATIN SQUARE SPLIT-PLOT DESIGN

8.1 A Hypothetical Experiment

To demonstrate how some generally sophisticated principles of experimental design might be applied in the context of categorical rather than continuous data, we briefly consider in this chapter the analysis of a hypothetical experiment to test effectiveness of several educational procedures. The design is typical of a number in use at the University of North Carolina Highway Safety Research Center, so we will phrase our discussion in terms related to this application.

Consider an experiment whose object is to assess the effects of three different training methods and three examining procedures on ability to pass a driving test. Subjects in the experiment normally are examined at three separate times, once using each examining procedure, and each time driving a different one of the three cars used in the experiment. The combinations of car and examining procedure are chosen according to the Graeco-Latin square of Table 8.1.1, where rows represent the driver training method; columns, the three trials; Latin letters, the three cars; and Greek letters, the three testing procedures. Suppose that due to cost considerations some subjects are not asked to participate in all three trials. The data might look like that of Table 8.1.2.

TABLE 8.1.1

DESIGN OF DRIVING TEST EXPERIMENT

	I	II	III
A	$X\alpha$	$Y\beta$	$Z\gamma$
B	$Z\beta$	$X\gamma$	$Y\alpha$
C	$Y\gamma$	$Z\alpha$	$X\beta$

TABLE 8.1.2

RESULTS OF DRIVING TESTS
(P=PASS, F=FAIL, *=NOT TAKEN)

PATTERN OF RESPONSE TRIAL			NUMBER OF RESPONDENTS METHOD		
I	II	III	A	B	C
P	P	P	15	12	6
P	P	F	24	3	14
P	F	P	9	22	6
P	F	F	15	9	6
F	P	P	5	16	10
F	P	F	13	13	25
F	F	P	10	13	13
F	F	F	9	12	20
SUBTOTAL			100	100	100
P	*	P	6	14	3
P	*	F	21	8	6
F	*	P	8	10	11
F	*	F	5	8	20
SUBTOTAL			40	40	40
P	P	*	18	8	5
P	F	*	10	11	5
F	P	*	7	8	18
F	F	*	5	13	12
SUBTOTAL			40	40	40

Some comments on this design are in order. We have three split-plot factors (cars, trials, and examining procedure) and one whole-plot factor (training method), each factor with three levels, combined into only nine "treatments". If a linear model is applied to functions of the data generated by different cells of the Graeco-Latin square (each cell representing a treatment), such a model must thus be limited to only nine parameters. Interactions among any set of factors are thus confounded by the design with main effects of the remaining factors. For instance, the first order interaction of training method and examining procedure, the variables of primary interest here, is confounded with the main effects of the blocking variables car and trial. The design is thus only appropriate when there is solid reason to believe, prior to the experiment, that no interactions exist involving the blocking variables car and trial, so that the main effects of the primary variables may be measured. Further, it is necessary to know either that the main effects of car and trial are negligible, so that the interaction between training method and examining procedure may be studied, or that this interaction is zero, so that a main effects model is appropriate. The analysis that follows assumes that use of the design is justified in the sense that there is strong reason to expect no interactions at all among the four factors with respect to a linear model for the observed proportions of passes under the various treatments.

8.2 Analysis

Let

$$\underset{1\times 40}{\underset{\sim}{p}} = (\underset{1\times 8}{\underset{\sim}{p}_1'}, \underset{1\times 8}{\underset{\sim}{p}_2'}, \underset{1\times 8}{\underset{\sim}{p}_3'}, \underset{1\times 4}{\underset{\sim}{p}_4'}, \underset{1\times 4}{\underset{\sim}{p}_5'}, \underset{1\times 4}{\underset{\sim}{p}_6'}, \underset{1\times 4}{\underset{\sim}{p}_7'})'$$

be a data vector constructed from Table 8.1.2, viz.,

$$\underset{\sim}{p} = (\tfrac{15}{100}, \tfrac{24}{100}, \ldots, \tfrac{9}{100}, \tfrac{12}{100}, \ldots, \tfrac{12}{100}, \tfrac{6}{100}, \ldots, \tfrac{20}{100}, \tfrac{6}{40}, \ldots, \tfrac{12}{40})'.$$

The vector whose elements are the proportions of subjects passing, among subjects given the same sequence of treatments, may be written as $\underset{\sim}{F}=A\underset{\sim}{p}$, where

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix},$$

$$A_1 = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} \otimes I_3,$$

$A_2 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \otimes I_6$. Since all of these probabilities are between .3 and .7 it is reasonable to analyze them on the natural linear scale, rather than apply a logit or other transformation. By the remarks of the last section it is thus appropriate to initially fit a main effects model

$$E\underset{\sim}{F} = X\underset{\sim}{\beta} \qquad (8.2.1)$$

where

$\underset{\sim}{\beta} = (\mu, \tau_1, \tau_2, \varphi_1, \varphi_2, \gamma_1, \gamma_2, \xi_1, \xi_2)'$,

$$
\underset{21\times 9}{X} = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & -1 & 0 & -1 & 0 & -1 \\
1 & 1 & 1 & -1 & 0 & -1 & 0 & -1 & 0 \\
1 & 0 & -1 & 1 & 1 & -1 & 0 & 0 & -1 \\
1 & 0 & -1 & 0 & -1 & 1 & 1 & -1 & 0 \\
1 & 0 & -1 & -1 & 0 & 0 & -1 & 1 & 1 \\
1 & -1 & 0 & 1 & 1 & 0 & -1 & -1 & 0 \\
1 & -1 & 0 & 0 & -1 & -1 & 0 & 1 & 1 \\
1 & -1 & 0 & -1 & 0 & 1 & 1 & 0 & -1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & -1 & 0 & -1 & 0 & -1 & 0 \\
1 & 0 & -1 & 1 & 1 & -1 & 0 & 0 & -1 \\
1 & 0 & -1 & -1 & 0 & 0 & -1 & 1 & 1 \\
1 & -1 & 0 & 1 & 1 & 0 & -1 & -1 & 0 \\
1 & -1 & 0 & -1 & 0 & 1 & 1 & 0 & -1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & -1 & 0 & -1 & 0 & -1 \\
1 & 0 & -1 & 1 & 1 & -1 & 0 & 0 & -1 \\
1 & 0 & -1 & 0 & -1 & 1 & 1 & -1 & 0 \\
1 & -1 & 0 & 1 & 1 & 0 & -1 & -1 & 0 \\
1 & -1 & 0 & 0 & -1 & -1 & 0 & 1 & 1
\end{bmatrix},
$$

and the elements of $\underset{\sim}{\beta}$ may be described as

μ = general mean effect

τ_1 = differential effect of training method A versus method C

τ_2 = effect of method A versus B

ρ_1 = effect of trial I versus trial III

ρ_2 = effect of trial I versus trial II

γ_1 = effect of car X versus car Z

γ_2 = effect of car X versus car Y

ξ_1 = effect of examination procedure α versus procedure γ

ξ_2 = effect of procedure α versus β.

A consistent estimate for the covariance matrix $V_{\underset{\sim}{F}}$ of $\underset{\sim}{F}$ is given by $\tilde{V}_{\underset{\sim}{F}} = AV(\underset{\sim}{p})A'$, where $V(\underset{\sim}{p})$ is the block diagonal matrix with multiples of $v(\underset{\sim}{p}_i) = D_{\underset{\sim}{p}_i} - \underset{\sim}{p}_i\underset{\sim}{p}_i'$, i=1,...,7 on the diagonal. Weighted least-squares

using $\tilde{V}_{\underset{\sim}{F}}$ provides a means of fitting the model. Estimates of the parameters and their standard errors are shown in Table 8.2.1. For purposes of comparison, this table also shows results from applying the model to only the 300 subjects who take three driving tests, and from analysis of only the data on the 240 remaining subjects who take only two tests each. These results are obtained from models involving only the first nine and last twelve rows of $\underset{\sim}{F}$ and X, respectively.

TABLE 8.2.1

DRIVING TEST EXPERIMENT: ESTIMATES AND STANDARD ERRORS FOR FULL MAIN EFFECTS MODEL

	Subjects in All Three Trials (n=300)		Subjects in Two Trials Only (n=240)		All Subjects (n=540)	
Parameter	estimate	s.e.	estimate	s.e.	estimate	s.e.
μ	.482	.016	.481	.023	.482	.013
τ_1	.076	.023	.094	.032	.088	.018
τ_2	-.028	.023	-.022	.033	-.029	.019
ρ_1	.026	.023	.051	.025	.033	.019
ρ_2	-.038	.024	-.053	.034	-.040	.019
γ_1	.016	.024	.003	.033	.011	.019
γ_2	-.024	.021	-.005	.032	-.015	.017
ξ_1	.099	.024	.153	.033	.121	.019
ξ_2	.022	.023	-.014	.032	.009	.019

The residual χ^2 (lack of fit) value for the full data is small, indicating adequate fit, so that it is proper to test significance of parameters of the model. Chi-square statistics generated by C matrices appropriate to hypotheses of interest are shown in Table 8.2.2.

TABLE 8.2.2

DRIVING TEST EXPERIMENT: TEST STATISTICS
FOR FULL MAIN EFFECTS MODEL

Hypothesis	χ^2-statistic Subjects in All Three Trials	χ^2-statistic Subjects in Two Trials Only	χ^2-statistic All Subjects
$\tau_1=\tau_2=0$	11.22	9.95	25.50
$\rho_1=\rho_2=0$	2.56	2.77	4.61
$\gamma_1=\gamma_2=0$	1.34	0.02	0.79
$\xi_1=\xi_2=0$	33.84	25.24	62.52
Residual	--	0.58 (3 d.f.)	4.52 (12 d.f.)

For example, the test for H_o: $\tau_1=\tau_2=0$ is generated for each set of data using the matrix

$$C = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The statistics in Table 8.2.2 indicate that car effects and trial effects are negligible. Note that had this been known in advance, it would not have been necessary to assume no interaction between training and examination methods, and the joint test of H_o: $\rho_1=\rho_2=\gamma_1=\gamma_2=0$ generated by

$$C = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

would be directed at the hypothesis of no interaction between these two variables. In any case, the analysis of the model indicates that

it may be feasible to fit a reduced model

$$E\underset{\sim}{F} = X_1 \underset{\sim}{\beta}_1$$

where $\underset{\sim}{\beta}_1 = (\mu, \tau_1, \tau_2, \xi_1, \xi_2)'$ and X_1 is X with the fourth-seventh columns deleted. Applying this model to data on all subjects gives estimates and test statistics shown in Table 8.2.3.

TABLE 8.2.3

DRIVING TEST EXPERIMENT: FIVE PARAMETER MODEL

Parameter	Estimate	s.e.	Hypothesis	χ^2	D.F.
μ	.480	.013			
τ_1	.087	.018			
τ_2	-.031	.019	$\tau_1=\tau_2=0$	24.81	2
ξ_1	.119	.019			
ξ_2	.007	.018	$\xi_1=\xi_2=0$	61.72	2
			Residual	9.80	16

The table suggests further reduction of the model by eliminating the parameters τ_2 and ξ_2. Fitting this third model gives results shown in Table 8.2.4. Conclusions that may be drawn from the analysis are:

i) car and trial order have no significant effect on ability to pass the driving test;

ii) training method C is less successful in teaching ability to pass the test than either A or B;

iii) examination procedure γ makes the test harder to pass than either α or β.

TABLE 8.2.4

DRIVING TEST EXPERIMENT: THREE PARAMETER MODEL

Parameter	Estimate	s.e.	Hypothesis	χ^2	D.F.
μ	.479	.013			
τ_1	.071	.015	$\tau_1=0$	31.32	1
ξ_1	.121	.016	$\xi_1=0$	59.89	1
			Residual	12.31	18

Finally, note that we have again used a "quick and dirty" analysis, and that the two-step procedure described in Chapter V can also be applied in this situation with some additional time and effort.

CHAPTER IX

PAIRED COMPARISON EXPERIMENTS

9.1 Analysis of Bradley-Terry Models

Paired comparison experiments and related models have a long history of use, particularly in psychometrics, in the study of variables and underlying concepts which are hard to quantify or scale in an objective manner. Direct measurement of such variables in terms of any absolute reference is thus impossible; nevertheless, if the variable or concept is to have any operational significance, a minimal requirement is that subjective comparisons are possible between certain units in terms of the variable, and that a large number of such subjective comparisons reveal some identifiable pattern. The pattern may be taken as generated by some underlying objective structure, or, alternatively, the concept may be defined empirically through the patterns elicited from some "standard" sets of subjective comparisons. Paired comparison experiments are also valuable for statistical reasons in some situations where the variable involved is well-defined and even directly measurable (see David [69]).

In the general paired comparison experiment, a number of "items" are submitted in pairs to a set of "judges". For each pair assigned to a judge, the judge indicates a "preference" for one member of the pair. Judges generally receive more than one pair, and the same pair is presented to a number of different judges. Judges may also be asked

to compare the same two items at more than one time. The declared preferences of the judges are then used to produce a partial or full ordering among the items, and sometimes in addition a partial or full distance function among them. The literature on paired comparisons is large, and a good account of the problem through 1962 may be found in David [69]. A number of articles have appeared since then; much recent work (e.g., Rao and Kupper [70], Davidson [71]) has been focused on generalized experiments in which judges are allowed to abstain from stating a preference in certain pairs, declaring that such pairs are "tied".

Bradley and Terry [72] proposed a model for paired comparison experiments that has proven particularly useful. Suppose that judges are selected randomly from some population, and that the probability that a random judge from the population will prefer i to j, in a given direct comparison, where $i,j=1,\ldots,V$ index a set of V items, and $i \neq j$, may be written as

$$\pi_{ij} = H(m_i - m_j) = \tfrac{1}{4} \int_{-(m_i-m_j)}^{\infty} \operatorname{sech}^2 \tfrac{1}{2}y \, dy$$

for a set of numbers m_k, $k=1,2,\ldots,V$ satisfying $\sum_{k=1}^{V} e^{m_k} = 1$. In this situation $\pi_{ij} = \pi_i/(\pi_i+\pi_j)$, where $\pi_k = e^{m_k}$. The m_k, $k=1,\ldots,V$ may be regarded as intrinsic "merits" which are not directly measurable, but often suitably characterize or scale the items in terms of the quality or variable of interest. The π_k, $k=1,\ldots,V$, represent a monotonic transformation of the m_k, and the vector $\underset{\sim}{\pi} = (\pi_1,\ldots,\pi_{V-1})'$ may be regarded as the underlying parameter vector of a family of multinomial

distributions $M(\nu,\underset{\sim}{\pi})$, $\nu=1,\ldots,\infty$. An interesting empirical problem whenever this model is applied is the relationship of this family to the distribution of answers to the hypothetical question "of the V items 1,2,...,V, which do you prefer above all others?" The distribution of answers to this question for given sample size ν may be quite different from $M(\nu,\underset{\sim}{\pi})$.

Linear model procedures for the analysis of categorical data can be used to analyze many sets of paired comparison data when an underlying Bradley-Terry model is assumed. Koch, Abernathy, Caltagirone and Johnson [73] have carried out such an analysis for a simple set of data in which each judge compares only one pair of items. Their analysis can be extended to more complex designs in which judges may consider a number of pairs. The linear model approach avoids the restrictive assumption, necessary for the usual maximum likelihood methods of analysis, that preferences of a judge within several pairs are statistically independent random variables.

Consider an experiment in which items or objects $O_1,\ldots,O_V$ are to be studied in terms of paired comparisons made by J judges, where J is large. Let S be the set of all possible pairs of the objects $O_1,\ldots,O_V$ and let S_k, $k=1,\ldots,K$, be lists of pairs in S, not necessarily disjoint or exhaustive of S, of respective sizes $s_1,\ldots,s_k$. Within each list pairs may occur more than once. The set of judges is partitioned into sets $U_1,\ldots,U_K$ of sizes $n_1,\ldots,n_K$. During the course of the experiment each judge in U_k is asked to compare exactly those pairs in S_k, as often as they appear in S_k, stating a preference within each pair at each time he examines it. It is assumed that $n_1,\ldots,n_k$ are all relatively large.

For example, $O_1,\ldots,O_V$ might be a set of soft drinks O_1 = Coca-Cola, O_2 = Pepsi-Cola,...,O_V = Royal Crown Cola to be compared in a taste test by a sample of shoppers. The set S of $\binom{V}{2}$ possible pairs of Colas might be partitioned into blocks S_k of size three, using an incomplete block design with $\binom{V}{2}$ treatments and block size three. Corresponding to each block of the design a sample of 100 shoppers might be asked to choose successively between samples of the two Colas in each pair of that block. Here K is the number of blocks b in the incomplete block design, U_k is the set of shoppers who judge pairs in the block S_k, and $s_k=3$, $n_k=100$, for k=1,...,b.

In any experiment of this type, the preferences of the set of judges U_k may be summarized in a 2^{s_k} contingency table with each cell representing a different "pattern of preference" determined by choices of judges within the pairs in S_k. The cell probabilities corresponding to the table from U_k may be estimated by the observed proportions of judges in U_k falling into the corresponding cells, and the covariance matrix of the strung-out vector $\underset{\sim}{p}_k$ of these probabilities by $v_k = \frac{1}{n_k}(D_{\underset{\sim}{p}_k} - \underset{\sim}{p}_k\underset{\sim}{p}_k')$. The one-way marginal probabilities of the table are the "preference probabilities" π_{xy}, π_{yx} for the pairs (O_x,O_y) in S_k. Consistent, unbiased estimates for the various π_{xy}, π_{yx} for pairs in S_k may be written as $A_k\underset{\sim}{p}_k$ for an appropriate premultiplier matrix A_k, and the covariance matrix of $A_k\underset{\sim}{p}_k$ estimated by $A_kv_kA_k'$. We write p_{xy} for the estimated preference probability obtained in this manner. Note that up to this point separate comparisons of the pair (O_x,O_y) by the same judges are treated as comparisons of different pairs, successive comparisons of O_x and O_y by the judges in U_k relating to different dimensions of the resulting table and producing separate estimates of

π_{xy}. Write $p_{xy}^{(1)}, p_{xy}^{(2)}, \ldots$ for such different estimates of π_{xy}.

Now under the Bradley-Terry model, $\pi_{xy} = \pi_x/(\pi_x+\pi_y)$, so that logit $\pi_{xy} = \log\left(\frac{\pi_{xy}}{1-\pi_{xy}}\right) = \log\left(\frac{\pi_x}{\pi_y}\right) = m_x - m_y$. Since p_{xy} is consistent for π_{xy}, logit p_{xy} is consistent for (m_x-m_y). This observation implies that the merits m_i may be estimated by least-squares using an appropriate linear model. The vector of observed logits may be produced as

$$\underset{\sim}{\ell} = K \log A\underset{\sim}{p}$$

where $\underset{\sim}{p}$ is the strung-out vector of the $\underset{\sim}{p}_k$'s, $A = \mathrm{diag}(A_k)$, and K is a vector of 1's, -1's, and zeros. The covariance matrix of the observed logits $\underset{\sim}{\ell}$ may be estimated consistently by

$$V_{\underset{\sim}{\ell}} = KD_{(A\underset{\sim}{p})}^{-1}\left(\sum_{k=1}^{b} A_k v_k A_k'\right)D_{(A\underset{\sim}{p})}^{-1}K'. \qquad (9.1.1)$$

A suitable linear model for $\underset{\sim}{\ell}$ is then

$$\underset{\sim}{\ell} = z\underset{\sim}{\beta} \qquad (9.1.2)$$

for $\underset{\sim}{\beta} = (\beta_1,\ldots,\beta_6)'$, $\beta_i = (m_i - m_o)$, $z = \{z_{ij}\}$, and where, if the i^{th} element of $\underset{\sim}{\ell}$ estimates logit π_{xy}, $z_{ix}=1$, $z_{iy}=-1$, $z_{ij}=0$ for $j\neq x,y$. $\underset{\sim}{\beta}$ may be estimated using weighted least-squares applied to (9.1.1) and (9.1.2), and the covariance matrix of the resultant estimate $\hat{\underset{\sim}{\beta}}$ estimated by $V_{\hat{\underset{\sim}{\beta}}} = (Z'V_{\underset{\sim}{\ell}}^{-1}Z)^{-1}$. Estimates of $\pi_0,\ldots,\pi_6$ can then be obtained as

$$\hat{\pi}_o = \left\{1 + \sum_{j=1}^{V} e^{-\hat{\beta}_j}\right\}^{-1}, \quad \hat{\pi}_j = \hat{\pi}_o\, e^{-\hat{\beta}_j}, \quad j=1,\ldots,V \qquad (9.1.3)$$

with estimated covariance matrix $HV_{\hat{\beta}}H' = V_{\hat{\underset{\sim}{\pi}}}$ for

$$H = \begin{bmatrix} \hat{\pi}_o \hat{\underset{\sim}{\pi}}^{*\prime} \\ \\ \hat{\underset{\sim}{\pi}}^* \hat{\underset{\sim}{\pi}}^{*\prime} - D_{\hat{\underset{\sim}{\pi}}^*} \end{bmatrix}, \quad \hat{\underset{\sim}{\pi}}^* = (\hat{\pi}_1, \ldots, \hat{\pi}_V)'.$$

The merits $m_o, \ldots, m_V$ are estimated by $\hat{m}_j = \log \hat{\pi}_j$, $j=0,\ldots,V$, and the covariance matrix of $\hat{\underset{\sim}{m}} = (m_o, \ldots, m_V)'$ by $D^{-1}_{\hat{\underset{\sim}{\pi}}} V_{\hat{\underset{\sim}{\pi}}} D^{-1}_{\hat{\underset{\sim}{\pi}}}$, $\hat{\underset{\sim}{\pi}} = (\hat{\pi}_o, \hat{\underset{\sim}{\pi}}^{*\prime})'$. Many tests of hypotheses concerning the merit vector $\underset{\sim}{m}$ and the "preference distribution" give by $\underset{\sim}{\pi}$ may be obtained using appropriate C matrices to form the relevant test statistics. Alternatively, maximum likelihood estimates $\hat{\hat{\underset{\sim}{m}}}$ and $\hat{\hat{\underset{\sim}{\pi}}}$ may be produced by the iterative procedure of Chapter IV; in the process m.l.e.'s of the related covariance matrices are also generated.

Finally, we note that models for multivariate paired comparison experiments with underlying structures similar to that of the univariate Bradley-Terry model (see e.g., Davidson and Bradley [74, 75, 76]) can also be analyzed using generalizations of the linear models just discussed.

9.2 Paired Choice in Population Studies

In this section we depart somewhat from our main problem in a brief discussion of one use of paired comparison methodology. Remarks will be centered on the paper Koch, Abernathy, Caltagirone and Johnson [73], which uses a paired comparison approach to study "desired family size" of North Carolina women.

Various concepts such as "desired", "expected", "ideal" and "intended" family size have been in use for many years in population

studies, where they are intended to measure attitudes and serve as indicators of future population trends. Numerous articles involving measurement of these variables, and the appropriateness of such uses for them, are available in the population literature. For a sampling see Meyers and Roberts [77], Blake [78, 79, 80, 81], Bumpass and Westoff [82], Freedman, Coombs and Bumpass [83], and Ryder and Westoff [84]. Measurements relating to any of these variables may vary strikingly in the aggregate between populations with disparate demographic characteristics (e.g., race, religion, age distribution), and frequently change over time for an individual woman. Alteration over time may even be marked and sudden, as with changes in health, economic, social or marital status. What follows will be phrased in the context of desired family size, but will usually be applicable to the related concepts of expected, intended and ideal family size as well.

Desired family size has usually been defined operationally in terms of answers to a direct question, such as "How many children do you want to have in your reproductive years?" It is assumed that most women have available a mental answer to such a question but underlying models in terms of intellectual, emotional, social or other motivational factors which determine both that answer and its relationship to completed family size are usually not considered. The statistical problem involved is thus limited to measuring accurately a woman's formulated mental response to a direct question, and studying the empirical relationship of these measurements to other indicators of attitude and to her future reproductive action, or that of her cohort. Measuring a woman's mental concept of desired family size by asking her a direct question about it is, however, a hazardous operation and problems of response error and

bias in such a procedure mitigate against successful use of the resulting data. Koch et al. [73] suggest that paired choice procedures may alleviate the problem; nevertheless, such procedures may introduce biases of their own.

Specifically, Koch et al. [73] study the responses of approximately 4625 women in the 1968 North Carolina Abortion Survey to a question as follows:

> "Let's suppose, for a moment, that you have just been married and that you are given the choice of having, during your entire lifetime, either $\underline{x}$ or $\underline{y}$ children. Which would you choose, $\underline{x}$ or $\underline{y}$?"

For a given woman, x and y were randomly selected from the set of all integer pairs (x,y), $0 \leq x,y \leq 6$, $x \neq y$. There are 42 such pairs, so that the question relating to each pair was asked of about 100 women. The rationale behind this question was that presenting respondents with only two specific alternatives to choose from, by limiting the scope and difficulty of the question, and perhaps focussing attitudes clearly, might make answering easier and cut down on response error due to pressures of social conformity. Ordering of pairs was randomized to reduce the influence of response set on the aggregate data. Results of questions containing pairs (x,y) and (y,x) were then pooled. Since each woman was queried about only one pair, responses to (x,y) and (y,x) questions are independent of responses to (x',y') and (y',x') questions if $(x',y') \neq (x,y)$ or (y,x). These responses are analyzed by the methods of the last section, which simplify due to this independence, to estimate intrinsic merits $m_0,m_1,\ldots,m_6$ and related parameters $\pi_0,\pi_1,\ldots,\pi_6$ pertaining to the numbers of children zero through six, as judged by North Carolina women. A preliminary hypothesis entertained in the paper is that the

"preference distribution" given by $\pi_0,\pi_1,\ldots,\pi_6$ is identical to the distribution of actual preferences when response is not constrained by paired choice. The paper is inconclusive with regard to this possibility.

Several observations seem relevant to this study and its analysis. Firstly, it is not really important that a theoretical Bradley-Terry model fit the underlying probability structure of the study. Nor is there an overriding necessity that the "preference probabilities" $\hat{\pi}_0,\hat{\pi}_1,\ldots,\hat{\pi}_6$, determined by the analysis, be good estimates of the probabilities of corresponding mental answers to a direct question about desired family size; indeed, there seems little substantive reason to suspect they would be. What is of interest is the usefulness of $\pi_0,\pi_1,\ldots,\pi_6$ and $m_0,m_1,\ldots,m_6$ in describing attitudes and predicting population trends. All that can be said about these numbers is that they provide a group measure on the relative desirability of various family sizes as judged by the sample, or the population from which it is chosen. An empirical question that remains is whether this measure is operationally useful.

In order for the measure to be useful, it would be desirable to derive it from data relating to the actual choices women will make in their reproductive lives. From this point of view, the study in [73] is open to some criticism. In particular, it might be suspected that presenting women with the unrealistic choices represented by pairs (x,y) with very different elements would produce response errors and misleading answers quite as often as they occur in direct questioning. As an extreme example, consider women forced to choose between zero and six as their preferred number of children. If responses could be assumed

regular in the sense that women desiring fewer than four children would respond zero, and others give six as the answer, then the paired choice question could be regarded as providing reliable information on desired family size. However, a woman desiring four children might quite rationally choose zero in this situation from the knowledge that six children would be too many to handle, due to economic or to her own physical and emotional limitations. Likewise, many women wanting one or two children badly might choose six in this comparison in preference to being childless. Whether one is really measuring desired family size here, or anything else related to the usual real choices a woman is faced with, one or two children, four or five, is problematic. On the other hand, a woman faced with a choice between x and $y>x$ similar numbers of children, even if her desired family size is $<x$ or $>y$, can usually choose x or y closest to her desired number of children without distorting her real feelings.

However, choices such as these last present a further problem. Even if x and y are close together, the question remains of how relevant to a woman's future behavior her choice between five and six children is when she only has one or two children at the time of decision. This is an empirical question which can only be answered by actual data, but it is reasonable to assume that a woman who has not experienced the difficulties in managing three children might revise her desires downward after the third or fourth child more often than upwards. Similarly, a choice between one or two children might be irrelevant to prediction of future population when elicited from a mother of seven.

These considerations all suggest that one might want to limit the set of pairs considered in a paired choice study of desired family size.

If the data is to be used to estimate the parameters $m_0, m_1, \ldots, m_6$, $\pi_0, \pi_1, \ldots, \pi_6$ as in 9.1, the analysis places only minor limitations on the way in which this is done. To estimate six independent parameters one must simply retain at least six pairs, and do it in such a way that the design matrix X in 9.1 is of full rank (six). This is a practical restriction of no significance in this problem.

Many peripheral studies may be necessary to pin down the concept of desired family size and to determine whether the methods of [73] and 9.1 are useful for its study. The paper [73] appears to have raised more questions than it answers, as is to be expected in an exploratory study. If further attempts to use paired choice to study desired family size are in order, then the remarks above, combined with the results of 9.1, suggest some modifications of the procedures of Koch et al.

i) Economies or increased precision can be obtained by presenting each woman with two or three, rather than a single, paired choice questions. Asking one woman more than three such questions invites confusion and carelessness in answering which could vitiate any tendency of paired choice to reduce response errors.

ii) Only pairs (x,y) with $|x-y| = 1$ or 2 should be used in the study, and not necessarily all such pairs.

iii) The pairs a woman is questioned on should have elements close to the number of children she has at the time of questioning. Thus a particular preference probability π_x or merit m_x would be primarily determined by preferences of women who have made such choices or are likely to confront them relatively soon.

Modification i) increases the efficiency of paired choice studies in estimating parameters; ii) and iii) essentially alter the definition of parameters to make them more relevant to actual reproductive behavior.

CHAPTER X

MISSING DATA

Up to now we have postulated the absence of 'interaction' between the measurement process by which incomplete data is obtained and the actual values of sampled individuals with respect to variables of interest. Thus, the choice of random variables measured on an individual was assumed independent of any of the variables under study. Missing data problems, where such is often not the case, have been omitted from the discussion. Here we tentatively explore ways in which techniques applicable to incomplete data designs might be used when incompleteness refers to observations which were planned into the design, but could not be collected.

Table 10.1 shows results of an opinion poll conducted prior to the 1968 Florida gubernatorial election by Independent Research Associates, Inc. of Washington, D. C. On September 1 a primary sample (for simplicity, we treat this as a simple random sample of Florida residents eligible to vote) was queried as to preference between the candidates Collins, a Democrat, and Gurney, a Republican. On October 1 an attempt was made to contact and repoll members of the primary sample, but some individuals could not be reached at that time. These individuals were discarded from the working sample. An attempt was made to contact the October respondents on November 1 and poll them for a third time, but the sample suffered further attrition since some could not be

reached at that time.

TABLE 10.1

PREFERENCES FOR GOVERNOR OF FLORIDA (1968)
SEPTEMBER, OCTOBER, NOVEMBER

PATTERN OF RESPONSE			NO. RESPONDENTS
SEPT.	OCT.	NOV.	
1	1	1	116
1	1	2	15
1	1	3	6
1	2	1	9
1	2	2	30
1	2	3	4
1	3	1	10
1	3	2	2
1	3	3	2
2	1	1	10
2	1	2	6
2	1	3	3
2	2	1	6
2	2	2	142
2	2	3	5
2	3	1	1
2	3	2	4
2	3	3	1
3	1	1	9
3	1	2	2
3	1	3	4
3	2	1	3
3	2	2	19
3	2	3	3
3	3	1	5
3	3	2	4
3	3	3	6
SUBTOTAL			427

PATTERN OF RESPONSE			NO. RESPONDENTS
SEPT.	OCT.	NOV.	
1	1	*	32
1	2	*	10
1	3	*	5
2	1	*	3
2	2	*	25
2	3	*	7
3	1	*	0
3	2	*	11
3	3	*	2
SUBTOTAL			95
1	*	*	145
2	*	*	84
3	*	*	47
SUBTOTAL			276
TOTAL			798

Key:

1 = COLLINS (D)
2 = GURNEY (R)
3 = NEITHER, DON'T KNOW
* = NOT CONTACTED

Suppose one wishes to study the probabilities of various patterns of candidate preference over the three time periods. If H_o: persons responding only one or two times exhibit the same distribution of candidate preference patterns as persons responding fully (necessarily so if non-response occurs independently of preference pattern), then a

supplemented margins model applies conditional on the numbers of respondents on the first day of each month. Since the structure of non-response is hierarchical, we can estimate probabilities of the twenty-seven response patterns as in Chapter III. Any of the corresponding test statistics in (1.3.2) - (1.3.4) then provides a test of the hypothesis H_o.

To avoid some arbitrariness in a minimum-χ^2_1 analysis due to a zero cell count in the data, we use maximum likelihood to estimate the cell probabilities of the 3×3×3 core table of response patterns. These estimates may be used to calculate an ordinary Pearson-χ^2 statistic as a test of H_o. The results are shown in Table 10.2. Since $\chi^2_{10} = 31.16$, one must conclude that retention in the sample was associated with candidate preference, so that the estimated probabilities in Table 10.2 are invalid due to selection bias in the sampling method. This suggests that the data from this poll should not be used to estimate probabilities of preference patterns for the total population of eligible Florida voters. It remains to ask what portions of the data may be used, in what ways, to salvage the maximum amount of information from this survey.

A careful examination of the data, using any appropriate supplementary statistics, is enlightening in this regard. Evidently, lack of fit of the supplemented margins model is due to:

i) the greater preference of September-only respondents for Collins, which might have occurred because working and rural people, largely Democratic, were harder to reach in callbacks than the remainder of the sample;

TABLE 10.2

MAXIMUM LIKELIHOOD ESTIMATES OF CELL PROBABILITIES
UNDER FULL SUPPLEMENTED MARGINS MODEL

PATTERN OF RESPONSE			ESTIMATED PROBABILITY
SEPT.	OCT.	NOV.	
1	1	1	.287
1	1	2	.037
1	1	3	.015
1	2	1	.022
1	2	2	.074
1	2	3	.010
1	3	1	.027
1	3	2	.005
1	3	3	.005
2	1	1	.020
2	1	2	.012
2	1	3	.006
2	2	1	.012
2	2	2	.289
2	2	3	.010
2	3	1	.004
2	3	2	.015
2	3	3	.003
3	1	1	.019
3	1	2	.004
3	1	3	.008
3	2	1	.009
3	2	2	.058
3	2	3	.009
3	3	1	.012
3	3	2	.009
3	3	3	.014

χ^2_{10} (lack of fit) = 31.16*

ii) differing interactions of November respondents and November drop-outs (responding in October) in the 2×3 subtables

		October	
		Collins	Gurney	neither, don't know
September	Gurney			
	neither, don't know			

On the other side of the coin

iii) the one-way marginal distributions of choices in September and October are similar for the groups of respondents polled twice and three times;

iv) persons choosing Collins in September, and responding again in October, gave October preferences similar to those choosing Collins in September and responding to both repolls.

Hence, for this data it is possible to isolate the four degrees of freedom causing lack of fit of the supplemented margins model. If there is some theoretical justification for singling out these degrees of freedom as miscreants and assuming that the model fits otherwise, one might wish to attempt further analysis. Of necessity this analysis can only be directed at answering limited questions, viz.

a) What were the probabilities of various preference patterns within the population of eligible voters who would have been reached all three times by the survey technique?

b) What were the probabilities of September-October preference patterns within the population which would have been reached twice?

c) What were the probabilities of each response in September for the entire population of eligible voters?

Questions a) and b) may be answered by fitting a linear model incorporating iii) and iv), and eliminating September only respondents from consideration. One such model is

$$E\, \underset{\sim}{p}_d^* = M\underset{\sim}{\Pi}^* \qquad (10.1)$$

where $\underset{\sim}{p}_d$ is the random vector underlying the observation vector

$$p_d^* = \left(\frac{116}{427}, \frac{15}{427}, \ldots, \frac{4}{427}, \frac{32}{95}, \ldots, \frac{\frac{1}{9}}{15}, \frac{11}{15}\right)', \qquad (10.2)$$

$$M = \left[\begin{array}{cccccccccccccccccccccccccccc}
\multicolumn{26}{c}{I_{26}} & \multicolumn{2}{c}{\underset{26\times 2}{0}} \\
1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\
0&0&0&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\
0&0&0&0&0&0&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0 \\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1 \\
0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&-1&-1 \\
0&0&0&0&0&0&0&0&0&1&1&1&0&0&0&0&0&0&1&1&1&0&0&0&0&0&-1&0 \\
0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&0&0&0&0&0&0&1&1&1&0&0&0&-1
\end{array}\right],$$

$\underset{\sim}{\Pi}^* = (\pi_{111}, \ldots, \pi_{332}, \pi_{21*}, \pi_{22*})'$ and π_{ijk} refers to probabilities mentioned in a), while π_{ij*} refers to those in b). Note that in (10.2) $\frac{1}{9}$ $(= \frac{1}{r}$, r = no. of responses) has been substituted for the observed zero cell count, following Grizzle, Starmer and Koch [31], to avoid singularity of the estimated covariance matrix of $\underset{\sim}{p}_d$. If $\hat{\underset{\sim}{\Pi}}^*$ is the estimate of $\underset{\sim}{\Pi}^*$ obtained in this manner, $\hat{\underset{\sim}{\pi}}_d^* = M\hat{\underset{\sim}{\Pi}}^*$ provides estimates for all probabilities in a) and b). These are shown with their estimated standard errors (formed from $\underset{\sim}{p}_d^*$) in Table 10.3. Finally, c) may obviously be answered by taking $\hat{\pi}_i$, the estimated probability of response i in September for the whole sampled population, equal to p_i, the observed proportion of i's among all 798 September respondents. Thus, $p_1 = .484$, $p_2 = .372$, $p_3 = .144$.

We emphasize in conclusion that the analysis given here is somewhat risky and is meant as an illustration rather than justification of the method. The statistician must resist the temptation to play such numerical games without some substantive backing for any modified supplemented margins model he employs. Further research is needed to

formalize heuristic aspects of the procedure, and to study properties of the use of partial supplementation as described.

TABLE 10.3

ESTIMATED PROBABILITIES AND STANDARD ERRORS
FROM MODIFIED SUPPLEMENTED MARGINS MODEL

PATTERN			PROBABILITY	STANDARD ERROR*
SEPT.	OCT.	NOV.	THREE RESPONSES	
1	1	1	.278	.020
1	1	2	.036	.009
1	1	3	.014	.006
1	2	1	.022	.007
1	2	2	.072	.012
1	2	3	.010	.005
1	3	1	.025	.007
1	3	2	.005	.003
1	3	3	.005	.003
2	1	1	.019	.007
2	1	2	.011	.006
2	1	3	.006	.004
2	2	1	.014	.006
2	2	2	.328	.020
2	2	3	.012	.005
2	3	1	.003	.002
2	3	2	.010	.005
2	3	3	.003	.002
3	1	1	.017	.007
3	1	2	.004	.003
3	1	3	.008	.005
3	2	1	.007	.004
3	2	2	.045	.010
3	2	3	.007	.004
3	3	1	.013	.005
3	3	2	.011	.005
3	3	3	.016	.006
			TWO RESPONSES	
1	1	*	.328	.020
1	2	*	.103	.013
1	3	*	.036	.008
2	1	*	.062	.011
2	2	*	.297	.025
2	3	*	.045	.014
3	1	*	.002	.003
3	2	*	.116	.017
3	3	*	.010	.013

CHAPTER XI

SUMMARY COMMENTS

This research has proceeded from the primary assumption that an experiment or survey is conducted and analyzed in order to answer certain questions which basically are independent of the discrete or continuous nature of the data. The statistician attempts to elicit information from the data particularly pertinent to these questions. Thus, it is desirable to develop flexible methods which guide the construction of analyses relevant to different questions and types of data. The theory of multivariate linear models provides such methods for continuous data, and no less so for sets of categorical data large enough to be "continuized" by application of suitable transformations and appropriate central limit theory. This is the thrust of Grizzle, Starmer and Koch [31] and papers that have built on it. The bulk of our work demonstrates that the flexibility of linear models in handling "continuized" categorical data extends quite generally to situations in which the data is incomplete, due to the linear relationship between parameters of the distributions of various subsets of the data.

In Part I, the problem was formulated in terms of a very general view of categorical data. Specializing to the product-multinomial situation, four methods were used to estimate cell probabilities in a two-way contingency table with supplemented margins. When both margins were supplemented, it was found that three of the methods required

iterative procedures to determine the estimates. In Chapter III we studied the general class of "hierarchically supplemented" designs. For such designs, all four types of estimates can be explicitly found without iteration. The estimation methods are distinguished by the use of different "kernel" averaging functions, through which estimates of marginal probabilities from subsets of the data are successively combined. Further research is needed to compare small sample distributions of the estimates, and their large sample properties in terms of a finer criterion than asymptotic variance. We likely need Monte-Carlo studies for the small sample problem; the work of Rao [85] and of Bahadur [86] on efficiency is relevant for large samples.

In Chapter IV, minimum Neyman-χ_1^2 and maximum likelihood estimates of cell probabilities are related through a set of least-squares equations. A condition for least-squares estimability of cell probabilities is given in terms of the rank of a matrix, and a way of imposing restrictions on these probabilities to make them estimable is suggested for cases when the condition is not satisfied. In Chapter V, least-squares procedures are provided for fitting linear models to functions of the data. For some situations two methods are available. Within a set of individuals measured on the same variables, one method estimates relevant probabilities and their covariances using just the data from these individuals; the other uses estimates obtained from all the data. Estimates of parameters of fitted linear models are BAN in either case; a comparison of estimates and related test statistics in small samples would be of interest. An example involving estimation of cell probabilities and subsequent linear models analysis concludes Part I.

Part II attempts to illustrate the breadth and power of the methods of Chapters IV and V by application to a number of different problems. Nevertheless, VI, VII and VIII are related, as they all consider split-plot situations. Chapter VI analyzes drug comparison experiments constructed with incomplete block designs. In VII, the split-plot factor is time, and a growth curve type analysis is demonstrated. Chapter VIII shows how effects of different factors may be isolated from a Graeco-Latin square split-plot design. In each case, incomplete aspects of the data provide no bar to a convenient analysis.

Chapter IX, in contrast, gives a new approach to a different problem, paired comparison experiments and surveys. The suggested analysis eliminates an independence assumption necessary for other methods. An application in demography is critically discussed in terms of the new approach. In Chapter X we have boldly tackled a missing data problem. Such data should be treated rather more gingerly than our analysis indicates.

Part II as a whole raises questions which invite much further investigation.

i) What are the variances of estimates and the powers of related test statistics used above and for other designs?

ii) How does the allocation of a sample to different measurement blocks affect the answers to i)?

iii) Taking into account cost structure, what principles should guide choice of a design in situations where incompleteness is necessary? What designs should be used? If incompleteness is not mandatory, is it ever preferable? If so, to what extent,

and what designs should be used?

At this point, these questions appear quite formidable.

APPENDIX

REMARKS ON ANALYZING CATEGORICAL DATA WITH WALD STATISTICS

Introduction

The analysis of categorical data, of special importance in the interpretation of research in the health and medical sciences, has long been limited in scope by the basic dependence of available methods on the dimensionality and sampling scheme of the data under consideration. Many procedures derive optimal qualities, or computational simplicity, from combinatorial properties peculiar to a multi-dimensional contingency table of certain fixed size. Additionally, the tendency to regard categorical data problems from a combinatorial viewpoint has often led to the development of techniques completely unrelated to those used to treat continuous data arising from an identical sampling scheme.

The past several years have seen widespread effort directed at the development of an analytical procedure for categorical data comparable in generality to multivariate linear models in the continuous case.* Here we give historical remarks on the development of one such approach, linear modeling of functions of the data using Neyman-χ^2_1 estimation.

We consider in what follows doubly indexed data $N = \{n_{\underset{\sim}{i}\underset{\sim}{j}}\}$, where $\underset{\sim}{i}$ and $\underset{\sim}{j}$ are themselves multiple subscripts. We shall assume here that N is generated by an underlying probability model in which $n_{\underset{\sim}{i}\underset{\sim}{j}}$ is the frequency of occurrence of a response called $\underset{\sim}{j}$ in a sample of size $n_{\underset{\sim}{i}}$ from a multinomial population called $\underset{\sim}{i}$. Additionally, we regard samples from each population as taken independently, so that the probability of the data N is specified by

---

*See, for example, Kullback [28], Lewis [87], Ku and Kullback [88], Bishop [27], Grizzle, Starmer and Koch [31], Goodman [26], Koch and Reinfurt [32].

$$\Pr(N) = \prod_{\underset{\sim}{i}} \frac{n_{\underset{\sim}{i}}!}{\prod_{\underset{\sim}{j}} n_{\underset{\sim}{ij}}!} \prod_{\underset{\sim}{j}} \prod_{\underset{\sim}{i}} \pi_{\underset{\sim}{ij}}^{n_{\underset{\sim}{ij}}}$$

where $\pi_{\underset{\sim}{ij}}$ is the probability of a response corresponding to $\underset{\sim}{j}$ in the population corresponding to $\underset{\sim}{i}$. The usual restrictions $\sum_{\underset{\sim}{j}} \pi_{\underset{\sim}{ij}} = 1$, $\sum_{\underset{\sim}{j}} n_{\underset{\sim}{ij}} = n_{\underset{\sim}{i}} \quad \forall \underset{\sim}{i}$, govern the above. Henceforth we will equate a subscript with the indicated population or response.

The method we will study developed out of two separate lines of work dealing with the above product-multinomial model. One line has been concerned with general methodological considerations, i.e., how to apply this model to some given set of data in terms of the population-response set-up, and how to formulate hypotheses appropriate to each model.

Pearson [15] noticed that several different sampling schemes could produce data expressible in identical contingency tables, but that knowledge of the sampling scheme remained crucial to a correct analysis. Pearson distinguished his sampling schemes according to how many and which margins of his contingency table were fixed or random. This is equivalent to asking which variables under study are population variables, as distinguished from response variables, in the above formulation.

S. N. Roy and his students (Mitra, Diamond, Sathe, Bhapkar and Kastenbaum) used the terms "variate" and "way of classification" to refer to variables with random or fixed margins in a multi-dimensional categorical set-up. Following Pearson, they treat any way of classification as a population, any variate as a response, in the

product-multinomial model. Several of these authors also used the terms "factor" and "response" to distinguish variables with fixed and random margins.

Bhapkar and Koch [18, 23] use the factor-response terminology, but define it from a different point of view. They liken any sampling scheme to an experiment. Certain information is known beforehand about the subjects, perhaps even controlled by the experimenter, while other information becomes available only as a result of the experiment. Usually interest centers on whether what was known prior to experimentation provides information on what occurs as a result. The situation is then analogous to determining how an experimental "treatment" affects a "response" in the sense of experimental design terminology. "Factor" denotes a generalization of "treatment" to include blocking variables. Factor levels will usually be designated as populations in a product-multinomial model.

Recent work by Koch and his collaborators recognizes some types of categorical data which are not appropriately described by the simple factor-response dichotomy. One example is a situation in which the same subjects respond successively to different levels of some treatment. In this case, the different treatment levels do not represent independent multinomial populations. Koch calls such a treatment a "split-plot factor." A distinct situation may occur in a multi-stage experiment where the response at one stage determines treatment at the next, so that the treatment margins are random. With a split-plot factor, it may be possible to answer certain questions by applying the product-multinomial model to a rearranged form of the data. In the

second situation one may usually perform an analysis conditional on the treatment margins.

The trend we have been describing involves the development of terminology for categorical data that is in basic conformity with that of continuous multivariate analysis. Concurrently, the major work of these same authors has been concerned with the translation of various concepts of continuous multivariate analyses into the language of multi-dimensional contingency tables. This approach is summed up in the two papers of Bhapkar and Koch [18, 23]*. They classify multidimensional problems into four types determined by the numbers of factors and responses involved, and discuss for each type the various problems conceivably of interest. In the continuous case, assuming multinormality, these problems are usually expressed and tested in terms of hypotheses about various correlation coefficients, or in terms of parameters of certain linear models. Analogous tests are suggested and expressed in terms of the $\pi_{\underset{\sim}{ij}}$. In particular, an attempt is made to relate the term "interaction", as used in the categorical data literature, to analogous terms used in the continous case. In some sampling schemes, interaction in a contingency table may be equivalent to the absence of an interaction parameter in a certain linear model.

On the other hand, multinormality in the continuous case provides us with convenient parameters of use as indices for many concepts of interest, e.g., correlation coefficients to test "interaction". In the

---

*For the development, see Roy and Mitra [89], Roy and Kastenbaum [90], Roy [64], Roy and Bhapkar [91], and Bhapkar [92, 21, 93, 22].

categorical case such ready-made indices may not be available. Often a hypothesis unambiguous in the multinormal situation will have several possible versions in the categorical framework. Bhapkar and Koch generate tests for these, using the Neyman-χ^2 procedures on which this research has focussed.

Wald's Paper

The development of this method, which constitutes the second line of work referred to earlier, begins with the remarkable 1943 paper by Wald [39]. Dealing with independent, identically distributed (i.i.d.) random variables (r.v.'s) from a distribution with unknown vector parameter $\underset{\sim}{\theta}$, he uses the maximum likelihood (m.l.) estimator $\hat{\underset{\sim}{\theta}}$ to construct a test procedure competitive with likelihood ratio. He begins with a set $\mathcal{A}$ of strong technical assumptions about the underlying distribution. These assumptions will not be considered individually here; they are quite difficult to verify in general situations.

Under $\mathcal{A}$, Wald shows that $\hat{\underset{\sim}{\theta}}$ converges uniformly in distribution to a specific multinormal law and, crucially, that this convergence is uniform over the possible values of $\underset{\sim}{\theta}$. This result leads to the very powerful

Theorem (Wald): Let $(\Omega_n, \mathcal{B}_n, P_n)$ be the probability space of a random sample of size n from the r.v. $\underset{\sim}{X}$, with distribution satisfying the assumptions referred to above. Let $(\Omega, \mathcal{B}, \Phi_n^*)$ be the probability space of $\hat{\underset{\sim}{\theta}}$, the m.l. estimate of the parameter vector $\underset{\sim}{\theta}$, under its limiting multinormal distribution. Then there is a sequence of mappings

W_n^*: $\mathcal{B}_n \to \mathcal{B}$ such that $|P_n(W_n|\underset{\sim}{\theta}) - \Phi_n^*(W_n^*(W_n)|\underset{\sim}{\theta})| \underset{n}{\to} 0$ uniformly in W_n and $\underset{\sim}{\theta}$.

The theorem implies that any asymptotic problem of interest involving the sequence $(\Omega_n, \mathcal{B}_n, P_n)$ may be solved by considering appropriate events in the measurable space $(\Omega, \mathcal{B})$, and their limiting probabilities under the sequence of measures Φ_n^*. This very special role of the m.l. estimator and its limiting distribution suggests that tests of hypotheses based on this statistic may also have worthwhile properties. Wald considers several related test statistics for various hypotheses in the light of three asymptotic optimality properties.

We introduce some notation, assuming the existence of the derivatives mentioned.

$\underset{\sim}{F}(\underset{\sim}{\theta})$ = vector of functions of $\underset{\sim}{\theta}$

$h_{ij}(\underset{\sim}{\theta}) = \dfrac{\partial F_i(\underset{\sim}{\theta})}{\partial \theta_j}$

$H(\underset{\sim}{\theta}) = \|h_{ij}(\underset{\sim}{\theta})\|$

$p(\underset{\sim}{x},\underset{\sim}{\theta})$ = probability mass function if $\underset{\sim}{X}$ discrete

= probability density function if $\underset{\sim}{X}$ continuous

$c_{ij}(\underset{\sim}{\theta}) = -E_{\underset{\sim}{\theta}} \dfrac{\partial^2 \log p(\underset{\sim}{x},\underset{\sim}{\theta})}{\partial\theta_i \partial\theta_j}$

$C(\underset{\sim}{\theta}) = \|c_{ij}(\underset{\sim}{\theta})\|$ $\qquad$ $S(\underset{\sim}{\theta}) = C^{-1}(\underset{\sim}{\theta}) = \|\sigma_{ij}(\underset{\sim}{\theta})\|$

$\underset{\sim}{\theta}^*$ = proper subvector of $\underset{\sim}{\theta}$ $\qquad$ $S_1(\underset{\sim}{\theta}) = H(\underset{\sim}{\theta})\, S(\underset{\sim}{\theta})\, H'(\underset{\sim}{\theta})$

$C^*(\underset{\sim}{\theta}^*)$ = submatrix of $C(\underset{\sim}{\theta})$ corresponding to $\underset{\sim}{\theta}^*$

$A(S)$ = area of S, for a set S in Euclidean space

Consider the hypothesis H_1: $\underset{\sim}{F}(\underset{\sim}{\theta}) = \underset{\sim}{0}$. Wald suggested the test defined by the critical region $T = (\underset{\sim}{F}(\hat{\underset{\sim}{\theta}}))'(H(\hat{\underset{\sim}{\theta}})C^{-1}(\hat{\underset{\sim}{\theta}})H'(\hat{\underset{\sim}{\theta}}))^{-1}\underset{\sim}{F}(\hat{\underset{\sim}{\theta}}) \geq d$, where d is chosen so that $\sup_{\underset{\sim}{\theta}\in H_1} \Pr\{T \geq d\} = \alpha$, the desired size of the test. Restricting $\underset{\sim}{F}$ to an appropriate set of functions we obtain, as a special case of H_1, H_2: $\underset{\sim}{\theta}^* = \underset{\sim}{\theta}_o^*$ and the corresponding test $(\hat{\underset{\sim}{\theta}}^*-\underset{\sim}{\theta}_o^*)'C^*(\hat{\underset{\sim}{\theta}}^*)(\hat{\underset{\sim}{\theta}}^*-\underset{\sim}{\theta}_o^*) \geq d$. A special case of H_2 is the simple hypothesis H_3: $\underset{\sim}{\theta} = \underset{\sim}{\theta}_o$ and the corresponding test $(\hat{\underset{\sim}{\theta}}-\underset{\sim}{\theta}_o)'C(\hat{\underset{\sim}{\theta}})(\hat{\underset{\sim}{\theta}}-\underset{\sim}{\theta}_o) \geq d$.

Now, let $W(\underset{\sim}{\theta})$ be a non-negative function of $\underset{\sim}{\theta}$, and $\{W_n\}$ be a sequence of critical regions to test H: $\underset{\sim}{\theta}\in\omega$ for some subset ω of the parameter space. Let $K = \{K(c,\underset{\sim}{\theta})\}$ be a family of surfaces in the parameter space, defined for each $\underset{\sim}{\theta}\in\omega$, and let A be the area measure on these surfaces. We define three optimal properties for the sequence $\{W_n\}$. Let $A[J] = \int_J W(\underset{\sim}{\theta})\, dA$ for a given surface J.

Property I: For any critical region Z for H, call

$$\int_{K(c,\underset{\sim}{\theta})} \frac{\Pr(Z|\phi)W(\phi)dA}{A[K(c,\underset{\sim}{\theta})]} = AP(Z;W,c,\underset{\sim}{\theta})$$

the average power of Z on the surface $K(c,\underset{\sim}{\theta})$ with respect to (w.r.t.) the weight function W. Suppose $\{W_n\}$ is such that

i) for each n, $\operatorname{lub}_{\underset{\sim}{\theta}\in\omega} \Pr(W_n|\underset{\sim}{\theta}) = \alpha$;

ii) for any other sequence $\{Z_n\}$ of critical regions to test H satisfying i),

$$\limsup_{n\to\infty} \left(\operatorname{lub}_{\substack{c \\ \underset{\sim}{\theta}\in\omega}} (AP(Z_n; W,c,\underset{\sim}{\theta}) - AP(W_n; W,c,\underset{\sim}{\theta}))\right) \leq 0;$$

then $\{W_n\}$ is said to have <u>asymptotically best average power</u> w.r.t. the family K and weight function W.

<u>Property II</u>: If

i) $\underset{\underset{\sim}{\theta}\in\omega}{\text{lub}} \Pr(W_n|\underset{\sim}{\theta}) = \alpha$ for all n;

ii) $\lim_{n\to\infty} \{\underset{c}{\text{lub}} \{ \underset{\underset{\sim}{\theta}\in\omega}{\text{lub}} \underset{\underset{\sim}{\phi}\in K(c,\underset{\sim}{\theta})}{\text{lub}} \Pr(W_n|\underset{\sim}{\phi}) - \underset{\underset{\sim}{\phi}\in K(c,\underset{\sim}{\theta})}{\text{glb}} \Pr(W_n|\phi)\}\} = 0$; and

iii) if $\{Z_n\}$ is a competing sequence of critical regions satisfying i), ii), then

$$\lim_{n\to\infty} \underset{\underset{\sim}{\theta}\notin\omega}{\text{lub}} \{\Pr(Z_n|\underset{\sim}{\theta}) - \Pr(W_n|\underset{\sim}{\theta})\} \leq 0;$$

$\{W_n\}$ is said to have <u>asymptotically best constant power</u> w.r.t. the surfaces K.

<u>Property III</u>: Let $P_n(\underset{\sim}{\theta},\omega,\alpha) = \underset{\underset{\sim}{\theta}\in\omega}{\text{lub}} \Pr(Z_n|\underset{\sim}{\theta})$, where the lub is taken over all measurable Z_n satisfying $\underset{\underset{\sim}{\theta}\in\omega}{\text{lub}} \Pr(Z_n|\underset{\sim}{\theta}) = \alpha$. Suppose $\{W_n\}$ is such that

i) $\exists \alpha \ni \underset{\underset{\sim}{\theta}\in\omega}{\text{lub}} \Pr(W_n|\underset{\sim}{\theta}) = \alpha \ \forall n$;

ii) For any competing sequence of critical regions $\{Z_n\}$ satisfying i) for the same α,

$$0 \geq \lim_{n\to\infty} \sup\{\underset{\underset{\sim}{\theta}}{\text{lub}}[P_n(\underset{\sim}{\theta},\omega,\alpha)-\Pr(W_n|\underset{\sim}{\theta})] - \underset{\underset{\sim}{\theta}}{\text{lub}}[P_n(\underset{\sim}{\theta},\omega,\alpha)-\Pr(Z_n|\underset{\sim}{\theta})]\}.$$

Then $\{W_n\}$ is said to be <u>asymptotically most stringent</u> for testing H.

Wald shows that under certain conditions the test suggested above for H_1 shares these properties with the likelihood ratio test for a specific weight function and family of curves. This of course holds for H_2 and H_3, but each less general case requires a weaker set of conditions. It is also shown under suitable conditions that the likelihood ratio and the Wald test criterion have the same limiting distribution irrespective of whether or not the hypothesis is true. The impact of the paper is that we may achieve some of the optimum properties of the likelihood ratio test, which requires maximizing the likelihood over the entire parameter space and also in the restricted space of the hypothesis, by using the Wald statistic which involves only the unrestricted m.l. estimate.

Neyman's Paper

Wald's procedure is directly applicable to certain one population problems involving the multinomial distribution, but the application was not made for over twenty years. Meanwhile, the Wald statistic in the categorical case was arrived at from a different route by Neyman. Wald had considered the problem of producing tests of hypotheses which share good properties of the likelihood ratio test. The analogue of this problem in estimation theory is the search for estimators which share good properties of the m.l. estimate. Neyman [24] solves this problem in a categorical situation.

We return to the product multinomial distribution discussed earlier. Suppose $\pi_{\underset{\sim}{ij}} = f_{\underset{\sim}{ij}}(\underset{\sim}{\theta})$ for some unknown vector $\underset{\sim}{\theta}$ and all $\underset{\sim}{i},\underset{\sim}{j}$ where the $f_{\underset{\sim}{ij}}$ are known functions. Let θ_i be the i^{th} component of $\underset{\sim}{\theta}$,

$\hat{\theta}_i$ the i^{th} component of $\hat{\underset{\sim}{\theta}}$. Then under fairly nonrestrictive conditions $\hat{\theta}_i$ is:

i) consistent for θ_i;

ii) approaches in distribution $N(\theta_i, \sigma_i/\sqrt{n})$ for some constant σ_i;

iii) if $p_{\underset{\sim}{ij}} = n_{\underset{\sim}{ij}}/n_{\underset{\sim}{i}}$, then $\partial\hat{\theta}_k/\partial p_{\underset{\sim}{ij}}$ exists and is continuous for each $\underset{\sim}{i}$, $\underset{\sim}{j}$, k;

iv) for any other sequence of estimators satisfying i)-iii) with constants σ_i^*, $\sigma_i^* \geq \sigma_i$.

These properties define a class of estimators which Neyman calls "best asymptotically normal" (BAN for short).

Let $\Pi = \{\pi_{\underset{\sim}{ij}}\}$. Now, the problem just stated, 1) estimate $\underset{\sim}{\theta}$ under $\pi_{\underset{\sim}{ij}} = f_{\underset{\sim}{ij}}(\underset{\sim}{\theta})$, can be reformulated equivalently as 2) estimate Π under $\underset{\sim}{F}(\Pi) = \underset{\sim}{0}$ where $\underset{\sim}{F}$ is determined by eliminating $\underset{\sim}{\theta}$ from the equations in 1). For this problem one BAN estimator may be obtained by minimizing, subject to the restrictions, the expression

$$\chi_1^2 = \sum_{\underset{\sim}{i}} n_{\underset{\sim}{i}} \sum_{\underset{\sim}{j}} \frac{(p_{\underset{\sim}{ij}} - f_{\underset{\sim}{ij}}(\underset{\sim}{\theta}))^2}{p_{\underset{\sim}{ij}}},$$

assuming all $p_{\underset{\sim}{ij}} > 0$. Specifically, Neyman shows that the equations obtained by setting the $\partial \log \chi_1^2/\partial\theta_k = 0$ have a solution which is a BAN estimator of $\underset{\sim}{\theta}$. If the $f_{\underset{\sim}{ij}}$ are linear in the θ_k, then these minimizing equations are too; hence their solution is easy. This situation is equivalent to the F_k, components of $\underset{\sim}{F}$, being linear in the $\pi_{\underset{\sim}{ij}}$ in formulation 2) above. Is there a similarly easy way to obtain BAN estimates where the F_k are not linear in the $\pi_{\underset{\sim}{ij}}$? We can solve a

linear problem that is "close" to the real one by replacing F_k with F_k^*, the linear part (first two terms) of the Taylor series expansion of F_k about the $p_{\underset{\sim}{ij}}$. We can then minimize χ_1^2 subject to $F_k^* = 0$ by solving a set of linear equations. Neyman shows that estimates obtained in this way are also BAN for $\underset{\sim}{\theta}$. Note that if $\theta_k = \pi_{\underset{\sim}{ij}}$ for some $\underset{\sim}{i}$, $\underset{\sim}{j}$, k, then we obtain BAN estimates of these $\pi_{\underset{\sim}{ij}}$. It is easily seen from this that estimates generated for all the $\pi_{\underset{\sim}{ij}}$ are BAN, and can be used to obtain BAN estimates for any θ_k expressible in terms of the $\pi_{\underset{\sim}{ij}}$.

We now return to the problem of Wald, i.e., constructing tests which share asymptotic properties of the likelihood ratio test. Let

$$Q_1 = \sum_{\underset{\sim}{i}} n_{\underset{\sim}{i}} \sum_{\underset{\sim}{j}} \frac{(p_{\underset{\sim}{ij}} - \hat{\pi}_{\underset{\sim}{ij}}(H))^2}{p_{\underset{\sim}{ij}}} \qquad Q_2 = \sum_{\underset{\sim}{i}} n_{\underset{\sim}{i}} \sum_{\underset{\sim}{j}} \frac{(p_{\underset{\sim}{ij}} - \hat{\pi}_{\underset{\sim}{ij}})^2}{p_{\underset{\sim}{ij}}}$$

where $\hat{\pi}_{\underset{\sim}{ij}}$ is an unrestricted BAN estimator of $\pi_{\underset{\sim}{ij}}$, and $\hat{\pi}_{\underset{\sim}{ij}}(H)$ is a BAN estimator of $\pi_{\underset{\sim}{ij}}$ under restrictions implied by the hypothesis H. Let $Q_3 = Q_1 - Q_2$. Neyman shows:

i) Q_2 is asymptotically χ^2,

ii) Q_3 is asymptotically χ^2 provided H is true,

iii) the likelihood ratio test and the test which rejects H for Q_3 on the upper tail of χ^2 are asymptotically equivalent, in the sense that the probability that they contradict each other tends to zero. Both tests are consistent.

iv) i), ii), iii) also hold for Q_1, Q_2, Q_3 based on the Pearson χ^2.

In particular, we may use for $\hat{\pi}_{\underset{\sim}{ij}}$, $\hat{\pi}_{\underset{\sim}{ij}}(H)$ the BAN estimators found by minimizing χ_1^2, so that Q_1 and Q_2 become simply the minimum values of

χ_1^2 under the hypothesis, and over the whole parameter space. Thus, $Q_1 = 0$ if there is no prior knowledge about $\pi_{\underset{\sim}{ij}}$, and to test H_4: $\pi_{\underset{\sim}{ij}} = f_{\underset{\sim}{ij}}(\underset{\sim}{\theta})$, we may use simply Q_1. To test the hypothesis H_5: $G(\underset{\sim}{\theta}) = 0$ assuming H_4 true, we may use $Q_1(H_4 \cap H_5) - Q_1(H_4)$.

We summarize the tools for testing hypotheses that have so far been developed. Under the product multinomial model we may test $\underset{\sim}{F}(\Pi) = \underset{\sim}{0}$ by either

i) (in the case of only one population) using m.l. estimators of $\pi_{\underset{\sim}{ij}}$ $(=\pi_{\underset{\sim}{j}})$ to estimate $\underset{\sim}{F}(\Pi)$ and its asymptotic covariance matrix, and calculating a test statistic based on these (Wald's method), or

ii) (in the general case) substituting any BAN estimate (under H) of $\pi_{\underset{\sim}{ij}}$ into χ_1^2, which is the test statistic itself (Neyman's method). This may always be done by using that BAN estimate which minimizes χ_1^2 under the linearized hypothesis, which may always be calculated just by solving a set of linear equations.

To test $\pi_{\underset{\sim}{ij}} = f_{\underset{\sim}{ij}}(\underset{\sim}{\theta})$ we may reformulate it as above and use i) or ii). Under the assumption that $\pi_{\underset{\sim}{ij}} = f_{\underset{\sim}{ij}}(\underset{\sim}{\theta})$ we may test $\underset{\sim}{G}(\underset{\sim}{\theta}) = 0$ using the difference of Neyman χ_1^2's as discussed above.

Note the following:

1) In their generality, these methods provide a means of fitting models $\underset{\sim}{F}(\Pi) = \underset{\sim}{G}(\underset{\sim}{\theta})$ simply by eliminating $\underset{\sim}{\theta}$ from these equations and reexpressing them as $\underset{\sim}{F}_1(\Pi) = \underset{\sim}{0}$.

2) The Neyman-χ^2_1 procedure for the model in 1) includes obtaining BAN estimates for the parameters in the model, and gives a straightforward way of doing this.

3) While the Wald statistic is defined originally only for the one-population case, it has an obvious generalization to multi-population problems. The properties of the statistic in this more general situation are unknown.

Bhapkar

The results of Neyman and Wald that we have described at length were put in their current perspective by Bhapkar [42, 56], who consolidated the two approaches by proving

Lemma: If all $n_{ij} > 0$, then Neyman's linearized χ^2_1 to test H: $\underset{\sim}{F}(\Pi) = \underset{\sim}{0}$ is identical to the generalized (multi-population) Wald statistic.

In fact, this statistic can be arrived at from a third point of view. To test the model $\underset{\sim}{F}(\Pi) = \underset{\sim}{G}(\underset{\sim}{\theta})$, the analogy to weighted least squares suggests using the statistic $T^2 = (\underset{\sim}{F}(\underset{\sim}{p}) - \underset{\sim}{G}(\underset{\sim}{\theta}))'S_1^{-1}(\underset{\sim}{p})(\underset{\sim}{F}(\underset{\sim}{p})-\underset{\sim}{G}(\underset{\sim}{\theta}))$. We have

Theorem (Bhapkar): 1) If $\underset{\sim}{G}(\underset{\sim}{\theta})$ is linear, that is, $\underset{\sim}{G}(\underset{\sim}{\theta}) = X\underset{\sim}{\theta}$ for a matrix X, then the Wald statistic to test $\underset{\sim}{F}(\Pi) = \underset{\sim}{G}(\underset{\sim}{\theta})$ is equal to the minimum of T^2 over all admissible $\underset{\sim}{\theta}$. This minimum is $\underset{\sim}{F}'S_1^{-1}\underset{\sim}{F} - \underset{\sim}{b}'(X'S_1^{-1}X)^{-1}\underset{\sim}{b}$ evaluated at $\pi_{ij} = p_{ij}$, where $b = (X'S_1^{-1}X)^{-1}X'S_1^{-1}\underset{\sim}{F}$, evaluated at $\pi_{ij} = p_{ij}$. 2) If $\underset{\sim}{F}(\Pi) = \underset{\sim}{G}(\underset{\sim}{\theta})$, then the Wald statistic to test $C\underset{\sim}{\theta} = \underset{\sim}{0}$ is $\underset{\sim}{b}'C'[C(X'S^{-1}X)^{-1}C'] C\underset{\sim}{b}$. This is the test arrived at

by the formal procedure of weighted multiple regression.*

Thus, machinery developed for multivariate linear models is seen to be applicable in analyzing categorical data. In particular, as long as care is taken in formulating a model, the approach of multivariate linear models becomes directly translatable from continuous to categorical data, and computer programs become immediately available simply by adapting the analogous programs for linear models. Programs utilizing the "generalized least squares" algorithm given by Bhapkar's theorem will produce Neyman statistics and minimum χ^2_1 estimates whenever they exist. In situations where some $\underset{\sim}{n}_{ij} = 0$ but S_1 still is non-singular, the procedure will produce "generalized Wald statistics" and "generalized weighted least-squares" estimates, about which little is known. This set of procedures has been described and applied to various types of categorical data in papers by Bhapkar and Koch [23], Grizzle, Starmer and Koch [31], Forthofer, Starmer and Grizzle [96], Johnson and Koch [36] and Koch and Reinfurt [32].

Other Optimality Considerations

We remark upon two other theoretical papers which bear on the merits of the procedures under discussion. It has been established that under appropriate circumstances the test criterion proposed by Wald and Neyman has the same limiting distribution as the corresponding likelihood ratio and, further that the probability that the two tests disagree approaches 0 as the sample size gets large, whether the hypothesis is

---

*This theorem is proved by Bhapkar [92] for linear F. A proof for general F is found in the appendix to Bhapkar and Koch [94].

true or not. Lastly, the sequence of Wald tests determined by testing at a fixed level, chosen independently of sample size, shares with the corresponding sequence of likelihood ratio tests the three asymptotic power properties of Wald.

On the other hand, choosing a sequence of tests of fixed size provides mathematical simplification at a great cost in terms of practical relevance. While for any real situation there may be a range of sample sizes within which we would choose, for given initial size and power, an improvement in power over a reduction in size, no reasonable loss function would allow us to do this indefinitely as sample size increases. We are thus led to the consideration of sequences of tests $\{T_{n,\alpha_n}\}$ of decreasing sizes $\{\alpha_n\}$.

The problem is that of finding, for a given test criterion, a sequence of sizes that will determine a test sequence with good power properties. One way of evaluating a criterion would be to determine whether such properties hold for a large class of sequences; for instance, whether some simple asymptotic property of $\{\alpha_n\}$ will guarantee good asymptotic power properties for $\{T_{n,\alpha_n}\}$. Hoeffding [96] investigated this situation for a single multinomial; we will give the crudest possible statement of his results.

Denote by $B_Z(\underset{\sim}{\pi})$ the power of the test Z at the alternative $\underset{\sim}{\pi}$. Let $\{T_{n,\alpha_n}\}$ be a sequence of tests with decreasing levels $\{\alpha_n\}$ and L_{n,α_n} be the likelihood ratio test of size α based on n observations. We have

Theorem (Hoeffding) For:

i) $\{\alpha_n\}$ asymptotically decreasing at a proper rate;

ii) $\{T_{n,\alpha_n}\}$ satisfying certain conditions;

there will exist an identifiable set of alternatives S (often "small") and a sequence $\{\alpha_n^*\}$, $\alpha_n^* \leq \alpha_n$ all n, such that outside S,

$$(1-B_{L_{n,\alpha_{n^*}}})/(1-B_{T_{n,\alpha_n}}) \to 0$$

faster than any power of n.

In particular, the result is applicable to Pearson chi-square tests in some situations. Since the class of sequences $\{\alpha_n\}$ for which the theorem works is sizable, this provides grounds for preferring a likelihood ratio test in these situations. We conjecture the same disadvantages for Neyman chi-square tests in these situations. Verification of this should be fairly easy.

From the point of view of estimation, Rao [85] concludes that Neyman chi-square procedures also compare unfavorably with maximum likelihood. For the single multinomial problem, he compares the mean-square errors of estimates of one underlying parameter obtained by a number of different methods. The mean square error is expanded in negative powers of n up to terms $O(n^{-2})$. Since all of the methods considered provide BAN estimates, the coefficient of n^{-1} is always 1/i, where i is the Fisher information (per observation) about the parameter. However, coefficients of n^{-2} differ. Maximum likelihood produces the smallest of these; Pearson chi-square is somewhat worse. Neyman chi-square is inferior to both of these, but better than the remaining three procedures. Rao [85] shows that maximum likelihood is under general conditions superior to all other methods in this respect.

Generalized Model-Building Techniques

Much recent effort has been devoted to developing methods of analysis for unorthodox arrangements of categorical data, e.g., contingency tables with certain cells a priori excluded, or with certain subjects unclassified in some of the dimensions. Maximum likelihood techniques for the first of these situations have been studied by Goodman [97], Mantel [34], Bishop and Feinberg [33] and others.

In the second situation, a suitable adaptation of the linear models approach may be used. Kleinbaum [6] formulates a model for multivariate observations in the continuous case which generalizes the usual multivariate linear model by allowing for i) incomplete data and ii) different designs for the variates. He shows that hypotheses under this "more general multivariate linear model" can be tested by using a slightly modified form of the Wald statistic. The modification consists in using any BAN estimates of the functions to be tested, and any consistent estimate of the variance-covariance matrix of the p_{ij}'s, to form the statistic that Wald obtains using only m.l. estimates. The basic approach of Reinfurt [9] and this research is that of Kleinbaum adapted to the categorical data situation.

A further extension of the linear models approach is discussed in Johnson and Koch [36], and an example given in Koch and Reinfurt [32]. These authors consider the situation of sampling from a finite population, so that the basic distributional model is no longer multinomial or product-multinomial, but hypergeometric or product-hypergeometric. Using Kleinbaum's method of constructing a Wald statistic, Johnson and Koch show that after modifying the estimated covariance matrix in accord

with the hypergeometric model we may proceed to use the usual weighted least squares procedures to estimate parameters and test hypotheses. The modifications necessary to programmed procedures are simple and straightforward. That the core of the Neyman-χ_1 approach may be used relatively easily here, and presumably with other distributional models, is a strong point in favor of this set of procedures.

REFERENCES

[1] Afifi, A. A. and Elashoff, R. M. 1966. Missing observations in multivariate statistics I. Review of the literature. Journal of the American Statistical Association 61: 595-604.

[2] Afifi, A. A. and Elashoff, R. M. 1967. Missing observations in multivariate statistics II. Point estimation in simple linear regression. Journal of the American Statistical Association 62: 10-29.

[3] Afifi, A. A. and Elashoff, R. M. 1969. Missing observations in multivariate statistics III. Large sample analysis of a simple linear regression. Journal of the American Statistical Association 64: 337-58.

[4] Afifi, A. A. and Elashoff, R. M. 1969. Missing observations in multivariate statistics IV. A note on simple linear regression. Journal of the American Statistical Association 64: 359-65.

[5] Hocking, R. R. and Smith, W. B. 1968. Estimation of parameters in the multivariate normal distribution with missing observations. Journal of the American Statistical Association 63: 159-73.

[6] Kleinbaum, D. G. 1970. Estimation and hypothesis testing for generalized multivariate linear models. University of North Carolina Institute of Statistics Mimeo Series No. 669.

[7] Roy, S. N., Gnanadesikan, R., and Srivastava, J. N. 1971. Analysis of Certain Quantitative Multiresponse Experiments. Pergamon Press, New York.

[8] Blumenthal, S. 1968. Multinomial sampling with partially categorized data. Journal of the American Statistical Association 63: 542-51.

[9] Reinfurt, D. W. 1970. The analysis of categorical data with supplemented margins including applications to mixed models. University of North Carolina Institute of Statistics Mimeo Series No. 697.

[10] Hocking, R. R. and Oxspring, H. H. 1971. Maximum likelihood estimation with incomplete multinomial data. Journal of the American Statistical Association 66: 65-70.

[11] Hartley, H. O. and Hocking, R. R. 1971. The analysis of incomplete data. Biometrics 27: 783-823.

[12] Williams, O. D. and Grizzle, J. E. 1970. Analysis of categorical data with more than one response variable by linear models. University of North Carolina Institute of Statistics Mimeo Series No. 715.

[13] Grizzle, J. E. 1961. A new method of testing hypotheses and estimating parameters for the logistic model. Biometrics 17: 372-85.

[14] Berkson, J. 1968. Application of minimum logit χ^2 estimate to a problem of Grizzle with a notation on the problem of no interaction. Biometrics 24: 75-96.

[15] Pearson, E. S. 1947. The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table. Biometrika 34: 139-67.

[16] Fisher, R. A. 1922. On the interpretation of χ^2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85: 87-94.

[17] Haberman, S. J. 1970. The general log-linear model. Unpublished Ph.D. dissertation, The University of Chicago, Chicago, Illinois.

[18] Bhapkar, V. P. and Koch, G. G. 1968. On the hypothesis of 'no interaction' in contingency tables. Biometrics 24: 567-94.

[19] Darroch, J. N. 1962. Interactions in multi-factor contingency tables. Journal of the Royal Statistical Society B 24: 251-63.

[20] Lancaster, H. O. 1969. Contingency tables of higher dimensions. Bulletin of the International Statistical Institute 43: 143-51.

[21] Bhapkar, V. P. 1966. Notes on analysis of categorical data. University of North Carolina Institute of Statistics Mimeo Series No. 477.

[22] Bhapkar, V. P. 1970. Categorical data analogs of some multivariate tests. Essays in Probability and Statistics (eds. R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao, K. J. C. Smith). Chapel Hill: The University of North Carolina Press.

[23] Bhapkar, V. P. and Koch, G. G. 1968. Hypotheses of 'no interaction' in multi-dimensional contingency tables. Technometrics 10: 107-23.

[24] Neyman, J. 1949. Contribution to the theory of the χ^2 test. Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability. Berkeley and Los Angeles: University of California Press (pp. 239-72).

[25] Goodman, L. A. 1970. The multivariate analysis of qualitative data: Interactions among multiple classifications. Journal of the American Statistical Association 65: 226-56.

[26] Goodman, L. A. 1971. The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models with multiple classifications. Technometrics 13: 33-61.

[27] Bishop, Y. M. M. 1969. Full contingency tables, logits, and split contingency tables. Biometrics 25: 383-99.

[28] Kullback, S. 1959. Information Theory and Statistics. New York: John Wiley and Sons.

[29] Ku, H. H., Varner, R., and Kullback, S. 1971. Analysis of multi-dimensional contingency tables. Journal of the American Statistical Association 66: 55-64.

[30] Good, I. J. 1963. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. Annals of Mathematical Statistics 34: 911-34.

[31] Grizzle, J. E., Starmer, C. F., and Koch, G. G. 1969. Analysis of categorical data by linear models. Biometrics 25: 489-504.

[32] Koch, G. G. and Reinfurt, D. W. 1970. The analysis of complex contingency table data from general experimental designs and sample surveys. Proceedings of the Sixteenth Conference on the Design of Experiments in Army Research, Development and Testing, Fort Lee, Virginia.

[33] Bishop, Y. M. M. and Fienberg, S. E. 1969. Incomplete two-dimensional contingency tables. Biometrics 25: 119-28.

[34] Mantel, N. 1970. Incomplete contingency tables. Biometrics 26: 291-304.

[35] Mather, K. 1938. Measurement of Linkage in Heredity. London: Methuen and Company.

[36] Johnson, W. D. and Koch, G. G. 1970. Analysis of qualitative data. Health Services Research, Winter 1970: 358-69.

[37] Berkson, J. 1955. Maximum likelihood and minimum χ^2 estimates of the logistic function. Journal of the American Statistical Association 50: 130-62.

[38] Odoroff, C. L. 1970. A comparison of minimum logit chi-square estimation and maximum likelihood estimation in 2×2×2 and 3×2×2 contingency tables: Tests for interaction. Journal of the American Statistical Association 65: 1617-31.

[39] Wald, A. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society 54: 426-82.

[40] Davidson, R. R. and Lever, W. E. 1970. The limiting distribution of the likelihood ratio statistic under a class of local alternatives. Sankhya 32: 209-24.

[41] Box, G. E. P. and Wilson, K. B. 1951. On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society B 13: 1-45.

[42] Box, G. E. P. 1966. A note on augmented designs. Technometrics 8: 184-88.

[43] John, P. W. M. 1966. Augmenting 2^{n-1} designs. Technometrics 8: 469-80.

[44] Gaylor, D. W. and Merrill, J. A. 1968. Augmenting existing data in multiple regression. Technometrics 10: 73-82.

[45] Daniel, C. 1962. Sequences of fractional replicates in the 2^{p-q} series. Journal of the American Statistical Association 57: 403-29.

[46] Addelman, S. 1969. Sequences of two level fractional factorial plans. Technometrics 11: 477-510.

[47] Koch, G. G. and Reinfurt, D. W. 1971. The analysis of categorical data from mixed models. Biometrics 27: 157-74.

[48] El-Badry, M. A. and Stephan, F. F. 1955. On adjusting sample tabulations to census counts. Journal of the American Statistical Association 50: 738-62.

[49] Friedlander, D. 1961. A technique for estimating a contingency table, given the marginal totals and some supplementary data. Journal of the Royal Statistical Society A 124: 412-420.

[50] Ireland, C. T. and Kullback, S. 1968. Contingency tables with given marginals. Biometrika 55: 179-88.

[51] Fienberg, S. E. 1970. An iterative procedure for estimation in contingency tables. Annals of Mathematical Statistics 41: 901-17.

[52] Birch, M. W. 1963. Maximum likelihood in three-way contingency tables. Journal of the Royal Statistical Society B 25: 220-33.

[53] Srivastava, J. N. and Roy, S. N. 1965. Hierarchical and p-block multiresponse designs and their analysis. Contributions to statistics, (ed. C. R. Rao). New York: Pergamon Press.

[54] Srivastava, J. N. 1966. Some generalizations of multivariate analysis of variance. Multivariate Analysis, (ed. P. R. Krishnaiah). New York: Academic Press.

[55] Srivastava, J. N. 1967. On the extension of Gauss-Markov theorem to complex multivariate linear models. Annals of the Institute of Statistical Mathematics 19: 417-37.

[56] Bhapkar, V. P. 1966. A note on the equivalence of two test criteria for hypotheses in categorical data. Journal of the American Statistical Association 61: 228-35.

[57] Dyke, G. V. and Patterson, H. D. 1952. Analysis of factorial experiments when the data are prepartions. Biometrics 8: 1-12.

[58] Goodman, L. A. 1964. Simultaneous confidence intervals for contrasts among multinomial populations. Annals of Mathematical Statistics 35: 716-25.

[59] Box, G. E. P. 1950. Problems in the analysis of growth and wear curves. Biometrics 6: 382-89.

[60] Elston, R. C. and Grizzle, J. E. 1962. Estimation of time response curves and their confidence bands. Biometrics 18: 148-59.

[61] Allen, D. M. and Grizzle, J. E. 1968. Analysis of growth and dose response curves. University of North Carolina Institute of Statistics Mimeo Series No. 576.

[62] Khatri, C. C. 1966. A note on a MANOVA model applied to problems in growth curves. Annals of the Institute of Statistical Mathematics 18: 75-86.

[63] Potthoff, R. F. and Roy, S. N. 1964. A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika 51: 122-27.

[64] Roy, S. N. 1957. Some Aspects of Multivariate Analysis. New York: John Wiley and Sons.

[65] Rao, C. R. 1959. Some problems involving linear hypotheses in multivariate analysis. Biometrika 46: 49-58.

[66] Rao, C. R. 1965. The theory of least-squares when the parameters are stochastic and its application to growth curves. Biometrika 52: 447-58.

[67] Rao, C. R. 1966. Covariance adjustment and related problems in multivariate analysis. Multivariate Analysis (ed. P. R. Krishnaiah). New York: Academic Press.

[68] Grizzle, J. E. 1970. An example of the analysis of a series of response curves and an application of multivariate multiple comparisons. Essays in Probability and Statistics (eds. R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao, K. J. C. Smith). Chapel Hill: The University of North Carolina Press.

[69] David, H. A. 1963. The Method of Paired Comparisons. New York: Hafner Publishing Company, Inc.

[70] Rao, P. V. and Kupper, L. L. 1967. Ties in paired comparison experiments: A generalization of the Bradley-Terry model. Journal of the American Statistical Association 62: 194-204.

[71] Davidson, R. R. 1970. On extending the Bradley-Terry model to accomodate ties in paired comparison experiments. Journal of the American Statistical Association 65: 317-28.

[72] Bradley, R. A., and Terry, M. B. 1952. Rank analysis of incomplete block designs I. The method of paired comparisons. Biometrika 39: 324-45.

[73] Koch, G. G., Abernathy, J. R., Caltagirone, S. A. and Johnson, W. D. 1971. The paired choice technique for measuring desired family size. University of North Carolina Institute of Statistics Mimeo Series No. 793.

[74] Davidson, R. R. and Bradley, R. A. 1969. Multivariate paired comparisons: The extension of a univariate model and associated estimation and test procedures. Biometrika 56: 81-95.

[75] Davidson, R. R. and Bradley, R. A. 1970. Multivariate paired comparisons: Some large-sample results on estimation and tests of equality of preference. Nonparametric Techniques in Statistical Inference (ed. M. L. Puri). Cambridge: Cambridge University Press.

[76] Davidson, R. R. and Bradley, R. A. 1971. A regression relationship for multivariate paired comparisons. Biometrika 58: 555-560.

[77] Meyers, G. C. and Roberts, J. M. 1968. A technique for measuring preferential family size and composition. Eugenics Quarterly 15: 164-72.

[78] Blake, J. 1966. The Americanization of Catholic reproductive ideals. Population Studies 20: 27-43.

[79] Blake, J. 1966. Ideal family size among white Americans: A quarter of a century's evidence. Demography 3: 154-73.

[80] Blake, J. 1967. Income and reproductive motivation. Population Studies 21: 185-206.

[81] Blake, J. 1967. Family size in the 1960's -- a baffling fad? Eugenics Quarterly 14: 60-74.

[82] Bumpass, L. L. and Westoff, C. F. 1970. The 'perfect contraceptive' population. Science 169: 1177-82.

[83] Freedman, R., Coombs, L. C., and Bumpass, L. L. 1965. Stability and change in expectations about family size: A longitudinal study. Demography 2: 250-75.

[84] Ryder, N. B. and Westoff, C. F. 1971. Reproduction in the United States 1965. Princeton: Princeton University Press.

[85] Rao, C. R. 1963. Criteria of estimation in large samples. Contributions to statistics (ed. C. R. Rao). New York: Pergamon Press.

[86] Bahadur, R. R. 1967. Rates of convergence of estimates and test statistics. Annals of Mathematical Statistics 38: 303-25.

[87] Lewis, J. A. 1968. A program to fit constants to multiway tables of quantitative and quantal data. Applied Statistics 17: 33-41.

[88] Ku, H. H. and Kullback, S. 1968. Interaction in multidimensional contingency tables: An information theoretic approach. Journal of Research of the National Bureau of Standards -- Mathematical Sciences 72B: 159-99.

[89] Roy, S. N. and Mitra, S. K. 1956. An introduction to some non-parametric generalizations of analysis of variance and multivariate analysis. Biometrika 43: 361-76.

[90] Roy, S. N. and Kastenbaum, M. A. 1956. On the hypothesis of no interaction in a multi-way contingency table. Annals of Mathematical Statistics 27: 749-57.

[91] Roy, S. N. and Bhapkar, V. P. 1960. Some nonparametric analogues of normal ANOVA, MANOVA, and of studies in normal association. Contributions to Probability and Statistics. Stanford: Stanford University Press.

[92] Bhapkar, V. P. 1961. Some tests for categorical data. Annals of Mathematical Statistics 32: 72-83.

[93] Bhapkar, V. P. 1968. On the analysis of contingency tables with a quantitative response. Biometrics 24: 329-38.

[94] Bhapkar, V. P. and Koch, G. G. 1965. On the hypothesis of "no interaction" in three dimensional contingency tables. University of North Carolina Institute of Statistics Mimeo Series No. 440.

[95] Forthofer, R. N., Starmer, C. F. and Grizzle, J. 1971. A program for the analysis of categorical data by linear models. Journal of Biomedical Systems 2: 3-48.

[96] Hoeffding, Wassily. 1965. Asymptotically optimal tests for multinomial distributions. Annals of Mathematical Statistics 36: 369-401.

[97] Goodman, L. A. 1968. The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. Journal of the American Statistical Association 63: 1091-131.