CLASSIFICATION ERRORS IN CONTINGENCY TABLES ANALYZED
WITH HIERARCHICAL LOG-LINEAR MODELS

BY

EDWARD LEE KORN

TECHNICAL REPORT NO. 20

AUGUST 1978

STUDY ON STATISTICS AND ENVIRONMENTAL
FACTORS IN HEALTH

PREPARED UNDER SUPPORT TO SIMS FROM

ENERGY RESEARCH AND DEVELOPMENT ADMINISTRATION (ERDA)

ROCKEFELLER FOUNDATION

SLOAN FOUNDATION

ENVIRONMENTAL PROTECTION AGENCY (EPA)

NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

NOTICE
This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor Paul Switzer for his many ideas which have been incorporated into this dissertation. Conversations about contingency tables with Joe Verducci, Ray Faith, and Alice Whittemore have been helpful and exceptionally enjoyable.

I thank Richard Olshen and Brad Efron for introducing me into the Statistics Department.

Thanks are due to Ingram Olkin for being a reader and for his suggestions.

I wish also to thank my parents for their encouragement over the years.

The typing was done expertly by Sheilla Hill.

TABLE OF CONTENTS

	<u>Page</u>
I. Introduction	1
II. The Structure of Classification Error	4
1. Population Studies	4
2. Dose-Response Studies	11
3. Sampling Schemes on Contingency Tables with Classification Errors	15
III. Log-Linear Models and Misclassification	21
1. Definition of Hierarchical Log-Linear Models	22
2. The 2×2 Table and Classification Error	24
3. Models Preserved by Misclassification	27
IV. Estimation and Testing Hierarchical Log-Linear Models	33
1. Maximum Likelihood Estimation of Expected Cell Counts	35
2. Asymptotic Distributions of Maximum Likelihood Estimates	42
3. Asymptotic Distributions of Test Statistics	48
Appendix A: Proofs of the Effects of Misclassification on the u Terms of Hierarchical Log-Linear Models (Chapter 3)	64
Appendix B: Proofs of Finite-Sample Results (Chapter 4)	71
Appendix C: Algorithms for Finding the Maximum Likelihood Estimates of the Expected Cell Counts (Chapter 4)	75
Appendix D: Proofs of the Asymptotic Distributions of Maximum Likelihood estimates and Test Statistics Chapter 4)	82

	<u>Page</u>
Appendix E: Simpler Expressions for Noncentrality	
Parameters (Chapter 4)	97
References	103

CHAPTER 1

INTRODUCTION

Classification errors in contingency tables can present many problems to a statistical analysis. The problems can range from mild to severe depending upon the mechanism that is misclassifying the observations in the table, and the type of analysis that is being done. Bross [1954] proposed a model for misclassification in a 2×2 table in which observations are incorrectly classified in the rows of the table according to some fixed misclassification probabilities, the false positive and false negatives rates. These misclassification probabilities were assumed to be the same in the two columns of the table. Bross showed that the usual hypothesis test of independence of a sampled 2×2 table would have the right significance level but reduced power under these circumstances. Mote and Anderson [1965] extended this result to an $I \times J$ contingency table. There is a brief review of misclassification in contingency tables given in Fleiss [1973], a more complete review given in Goldberg [1972], and an extensive bibliography of classification errors in surveys given in Dalenius [1977].

My own interest in classification errors in contingency tables arose out of an attempt to analyze a 1973 Environmental Protection Agency data set. The data consisted of daily measurements of 7 air pollutants, 3 meteorological variables, and responses from a panel of asthmatics signifying whether or not they had an asthma attack on each day. One possible analysis consisted of putting the observations (person-days) into a high dimensional contingency table with the response variable and categorized versions of each of the independent variables making

up the different dimensions of the table. A contingency table analysis could then be used to see which, if any, of the pollution and meteorological variables were associated with increased asthma. The analysis of that particular data set has since turned in other directions (Whittemore and Korn [1978]), but not before I became concerned with the effect of the high unreliability of the pollution measurements on the conclusions of such an analysis. Would pollutants be appearing to be associated spuriously with asthma?

This thesis is concerned with the effect of classification error on contingency tables being analyzed with hierarchical log-linear models (independence in an $I \times J$ table is a particular hierarchical log-linear model). Hierarchical log-linear models provide a concise way of describing independence and partial independences between the different dimensions of a contingency table. The use of such models to analyze contingency tables can be expected to increase with the advent of many excellent books describing the subject (Cox [1970], Haberman [1974a], Plackett [1974], Bishop, Fienberg, and Holland [1975]), and the widespread availability of a computer program to perform the analyses (Dixon and Brown [1977]).

In Chapter 2 of this thesis, the structure of classification errors on contingency tables that will be used throughout is defined. This structure is a generalization of Bross' model, but here attention is paid to the different possible ways a contingency table can be sampled. Hierarchical log-linear models and the effect of misclassification on them are described in Chapter 3. Some models, such as independence in an $I \times J$ table, are preserved by misclassification, i.e., the presence of classification error will not change the fact that a specific table

belongs to that model. Other models are not preserved by misclassification; this implies that the usual tests to see if a sampled table belong to that model will not be of the right significance level.

A simple criterion will be given to determine which hierarchical log-linear models are preserved by misclassification. In Chapter 4, maximum likelihood theory is used to perform log-linear model analysis in the presence of known misclassification probabilities. It will be shown that the Pitman asymptotic power of tests between different hierarchical log-linear models is reduced because of the misclassification. A general expression will be given for the increase in sample size necessary to compensate for this loss of power and some specific cases will be examined.

CHAPTER 2

THE STRUCTURE OF CLASSIFICATION ERROR

In this chapter two general situations are examined which lead to quite different kinds of classification error. One occurs when a large population is being sampled and the observed attributes of an individual do not correspond to his true attributes. The other situation occurs when individuals are separated into groups to be given different levels of a dose, the doses and subsequent responses being recorded. If an individual assigned to receive one level of a dose is actually given a different level, there will be classification error. These two situations are considered for the 2×2 table in Sections 1 and 2, respectively. Section 3 generalizes the models to higher dimensional contingency tables and also formalizes the assumptions on the classification error that will be used.

1. Population Studies

Consider the problem of studying a large population to see if there is an association between smoking and lung cancer. The probability that a person chosen randomly from the population smokes and/or has cancer can be displayed in the following 2×2 table:

Table 1

	c_1	c_2
s_1	$\pi(11)$	$\pi(12)$
s_2	$\pi(21)$	$\pi(22)$

where

$$\pi(ij) = P(S_i, C_j)$$

= probability a person has smoking status i
and cancer status j

and

S_1 = smoker

S_2 = non-smoker

C_1 = has cancer

C_2 = does not have cancer.

The three common types of study (Fleiss [1973]) that might be conducted are the

- a) cross-sectional study
- b) retrospective study,
- or
- c) prospective study.

In a cross-sectional study, a sample of size $n(++)$ would be taken from the population and for each person his stated smoking status and a doctor's diagnosis of his cancer status would be recorded:

Table 2

	C_1	C_2	
T_1	$n(11)$	$n(12)$	$n(1+)$
T_2	$n(21)$	$n(22)$	$n(2+)$
	$n(+1)$	$n(+2)$	

where

$n(ij)$ = number of people with stated smoking status i
and cancer status j

and

$$T_1 = \text{stated smoker}, \quad T_2 = \text{stated non-smoker}.$$

In Table 2 and elsewhere in this thesis, a plus sign (+) as an index will stand for the sum over that index.

If a person's stated smoking status does not always agree with his true smoking status, then there will be said to be classification error between the rows of the contingency table. In the 1950's when actual studies were conducted to see if there was an association between smoking and cancer there was concern about the accuracy of stated smoking histories, e.g., Sadowsky, Gilliam, and Cornfield [1953], and Mantel and Haenszel [1959]. In this example it is assumed that the doctor's diagnosis is always correct.

The probability a person incorrectly states his smoking status may depend on his true smoking status and his cancer status. These four misclassification probabilities are given by:

$$\begin{aligned} P(T_2 | S_1, C_j) &= \text{conditional probability a person says} \\ &\quad \text{he is a non-smoker given he truly smokes} \\ &\quad \text{and has cancer status } j, \\ &\quad \text{for } j = 1, 2. \\ (1) \quad P(T_1 | S_2, C_j) &= \text{conditional probability a person says} \\ &\quad \text{he smokes given he is truly a non-smoker} \\ &\quad \text{and has cancer status } j, \\ &\quad \text{for } j = 1, 2. \end{aligned}$$

These misclassification probabilities are precisely the false positive rates and false negative rates in the smoking dimension of the table

in the cancer and non-cancer subgroups of the population.

The probability that a person states he smokes and/or has cancer can also be displayed in a 2×2 table:

Table 3

	C_1	C_2
T_1	$\tau(11)$	$\tau(12)$
T_2	$\tau(21)$	$\tau(22)$

where

$$\tau(ij) = P(T_i, C_j)$$

= probability a person has cancer status j

and stated smoking status i , for $i, j = 1, 2$.

The π 's and the τ 's can be simply related through the misclassification probabilities:

$$(2) \quad \tau(ij) = P(T_i | S_1, C_j) \pi(1j) + P(T_i | S_2, C_j) \pi(2j) .$$

In a retrospective study, $n(+1)$ people are sampled who have cancer and $n(+2)$ people who don't. Their stated smoking status is recorded as in Table 2. The probabilities of interest in the population are given as in Table 1, but now are probabilities conditional on cancer status. So, $\pi(+j) = 1$ for $j = 1, 2$. The τ 's in Table 3 are also thought of as conditional probabilities of stated smoking status given cancer status. It is easy to see that the relationship between the π 's and the τ 's here is given by (2), exactly the same relationship as in the cross-sectional study.

In a prospective study there is a new problem. The ideal would be to sample $n(1+)$ people who smoke and $n(2+)$ people who don't and see how many of each type have cancer. However, all that can be done is to sample $n(1+)$ people who state they smoke and $n(2+)$ people who state they don't and record their cancer status as in Table 2. The probabilities of interest are given in Table 1, but now are conditional on (true) smoking status. The τ 's in Table 3 are now thought of as conditional probabilities of cancer given stated smoking status. The relationship between the π 's and τ 's is given by:

$$(3) \quad \tau(ij) = P(T_i | S_1, C_j) \pi(1j) \frac{P(S_1)}{P(T_i)} + P(T_i | S_2, C_j) \pi(2j) \frac{P(S_2)}{P(T_i)}.$$

This looks similar to the relationship (2) in the previous types of studies except for the factors involving $P(S_1)$ and $P(T_i)$, the unconditional probabilities of being a smoker and a stated smoker in the population. Since the sampling scheme is fixing the smoking dimension of the table, there is no information about these unconditional probabilities in the data. This more complicated relationship between the π 's and τ 's arises because there is classification error in a dimension of the table that is being held fixed by the sampling scheme. If there is classification error in the cancer dimension of the table and not in the smoking dimension, then the problem occurs in the retrospective study and not in the prospective one.

In any of the three types of population study, one possible model for the misclassification probabilities given in (1) is as follows: The probability a person states his correct smoking status is the same in the subpopulation of people who have cancer as it is in the subpopulation

of people who don't. That is,

$$(4) \quad P(T_i | S_i, C_1) = P(T_i | S_i, C_2) \quad \text{for } i = 1, 2 .$$

This, of course, implies

$$(5) \quad P(T_i | S_i, C_j) = P(T_i | S_i) \quad \text{for } i, j = 1, 2 .$$

This model says that the false positive rates and false negative rates for smoking status are the same in both the cancer and non-cancer subpopulation. An equivalent formulation is that the probability a person has cancer given his true smoking status does not depend on his stated smoking status. That is,

$$(6) \quad P(C_1 | S_i, T_1) = P(C_1 | S_i, T_2) \quad \text{for } i = 1, 2 .$$

These assumptions are not always reasonable. For example, suppose the interviewer taking the smoking history from the subjects knows which subjects have cancer. Then one would not be too surprised to find less false smoking negatives among the cancer patients than among the non-cancer patients.

If we assume (4) or (6), the relationship (2) between the π 's and the τ 's in the cross-sectional and retrospective studies is given by:

$$(7) \quad \tau(ij) = P(T_i | S_1) \pi(1j) + P(T_i | S_2) \pi(2j) \quad \text{for } i, j = 1, 2 .$$

The relationship (3) in the prospective study becomes:

$$\begin{aligned}
 \tau(ij) &= P(T_i | S_1) \pi(1j) \frac{P(S_1)}{P(T_i)} + P(T_i | S_2, C_j) \pi(2j) \frac{P(S_2)}{P(T_i)} \\
 (8) \qquad &= P(S_1 | T_i) \pi(1j) + P(S_2 | T_i) \pi(2j) \quad \text{for } i, j = 1, 2.
 \end{aligned}$$

Although (7) and (8) look quite similar, there is a world of difference between $P(S_i | T_j)$ and $P(T_j | S_i)$. In view of (5), the quantities $\{P(T_j | S_i)\}$ can be measured in any subgroup of the population without regard to cancer status. This is not true for the $\{P(S_i | T_j)\}$.

Remark: The model for misclassification given here was first developed by Bross [1954] for the 2×2 table. Bross implicitly made the assumption (4), while Rubin, Rosenbaum, and Cobb [1956] stated it explicitly. A series of articles (Diamond and Lilienfeld [1962a, 1962b], Newell [1962], Keys and Kihlberg [1963], Buell and Dunn [1964]) debated the correct way to analyze a retrospective study trying to measure the association between women who have cancer of the cervix and the circumcision status of their husbands. A serious classification error was suspected when a study (Lilienfeld and Graham [1958]) had shown that self-report circumcision status disagreed with a doctor's examination in a large percentage (35%) of men sampled. The controversy in the articles was really about whether it was proper to assume (4) or not. A recent paper (Goldberg [1975]) claims that (4) is usually inappropriate in medical screening. However, the assumption (4) will be used in this thesis because:

a) In some applications it is very reasonable; for example, when misclassification is due to coding and keypunching errors. In the dose-response studies considered in the next section, there will also be little reason to doubt the equivalent assumption (6).

b) The reasons for the failure of (4) are likely to be similar to the reasons a retrospective study can be biased (Buell and Dunn [1964]). Using assumption (4) and the misclassification model may be a step closer to a reliable analysis.

c) Most of the work previously done on misclassification in contingency tables has been for 2×2 tables. For larger dimensional tables with more levels in each dimension the number of misclassification probabilities can become enormous. For example, in a $2 \times 5 \times 10$ table with classification error only in the first dimension, there are 100 misclassification probabilities to be considered if (4) is not assumed, while only 2 if it is.

2. Dose-Response Studies

In a dose-response study, different levels of a dose are given to subjects and their responses are recorded. For example, consider the hypothetical problem of measuring the effect of low and high doses of a drug on the mortality of rats. The probabilities of interest can be displayed in the following 2×2 table:

Table 4

	R_1	R_2
D_1	$\pi(11)$	$\pi(12)$
D_2	$\pi(21)$	$\pi(22)$

where

$\pi(ij)$ = probability a rat given dose D_i has response R_j

and

$$\begin{aligned} D_1 &= \text{low dose} & D_2 &= \text{high dose} \\ R_1 &= \text{alive} & R_2 &= \text{dead.} \end{aligned}$$

To estimate these probabilities a controlled comparative trial could be run (Fleiss [1973]): $n(1+)$ rats chosen randomly from n rats would be assigned to get a low dose of the drug, and the other $n(2+) = n - n(1+)$ rats a high dose. The results of the experiment could be recorded in the following 2×2 table:

Table 5

	R_1	R_2	
A_1	$n(11)$	$n(12)$	$n(1+)$
A_2	$n(21)$	$n(22)$	$n(2+)$

where

$$\begin{aligned} A_1 &= \text{assigned to get } D_1, \text{ low dose of the drug} \\ A_2 &= \text{assigned to get } D_2, \text{ high dose of the drug.} \end{aligned}$$

Classification error can occur in this type of study when rats assigned to a certain level of the drug are exposed to a different level. For example, suppose the experimenter blunders and unknowingly gives some rats a low dose of the drug when they were assigned to get a high dose. The misclassification probabilities are:

$$\begin{aligned} P(D_j : A_i) &= \text{probability a rat assigned to get dose } D_i \\ &\text{of the drug actually gets dose } D_j. \end{aligned}$$

These misclassification probabilities are not conditional probabilities since the number of rats assigned to get a specific dose of the drug is not random. The probabilities of mortality for rats assigned to the low and high dosage groups are:

Table 6

	R_1	R_2
A_1	$\tau(11)$	$\tau(12)$
A_2	$\tau(21)$	$\tau(22)$

where

$\tau(ij)$ = probability a rat assigned to get dose D_i
has response R_j .

Expressing the τ 's in terms of the misclassification probabilities, one has:

$$(9) \quad \tau(ij) = P(D_1 : A_i)P(R_j | D_1 : A_i) + P(D_2 : A_i)P(R_j | D_2 : A_i)$$

where

$P(R_j | D_k : A_i)$ = conditional probability a rat assigned to
get dose D_i of the drug has response R_j ,
given it truly received dose D_k of the
drug.

In many experimental situations it is reasonable to assume that the probability of response is a function only of the true dose given and not which group the subject was assigned to. That is,

$$(10) \quad P(R_j | D_i : A_1) = P(R_j | D_i : A_2) = \pi(ij) .$$

This assumption is completely analogous to (6) of the last section. However, there will be less reason to question it in the dose/response context.

Using (9) and (10), the τ 's can be expressed in terms of the π 's and the misclassification probabilities:

$$(11) \quad \tau(ij) = P(D_1 : A_i) \pi(1j) + P(D_2 : A_i) \pi(2j) .$$

This looks similar to (8) of the last section except that the misclassification probabilities have a different meaning here. In a population study, the misclassification probabilities $P(T_i | S_j)$ refer to the conditional probability of an observed state given the true state. The probabilities $P(S_j | T_i)$ are functions of the $P(T_i | S_j)$ and other population probabilities. In a dose-response study the observed state is set by the experimenter; the misclassification probabilities are of the true states given these observed states.

Remark: The same distinction between two kinds of errors can be made in the normal-theory regression framework when there is error in the independent variable. Berkson [1950] calls an observation that is made measuring a true population value with error an "uncontrolled observation." A "controlled observation" is one in which a dose is set with error by an experimenter. The theory developed for dealing with controlled observations (Berkson [1950], Scheffé [1959]) assumes that the error is unbiased around the value set by the experimenter. This is unreasonable in the contingency-table context, for a misclassification

error from level 1 to level 2 cannot be balanced with an error from level 1 to "level 0."

3. Sampling Schemes on Contingency Tables with Classification Errors

In this section some common sampling schemes for contingency tables are described. The effect of misclassification on the expected cell counts of the table will be examined for the population and dose-response studies discussed in Sections 1 and 2.

In an $I_1 \times I_2 \times \cdots \times I_K$ contingency table, let $(i_1 i_2 \dots i_K)$ denote the cell that has level i_ℓ of the ℓ^{th} dimension for $\ell = 1, 2, \dots, K$. Let the random variable $X(i_1 i_2 \dots i_K)$ represent the number of observations falling in cell $(i_1 i_2 \dots i_K)$. Let $T = I_1 \cdot I_2 \cdots I_K$ be the total number of cells in the table.

There are three sampling distributions on contingency tables that are commonly considered when there is no classification error. In the Poisson scheme the cell counts $\{X(i_1 i_2 \dots i_K)\}$ are distributed as T independent Poisson random variables. In the simple multinomial scheme, N is a fixed total number of observations over all cells, and for any given cell, all of the N observations have an equal probability of falling in that cell. In the product multinomial scheme, the T cells of the table are partitioned into subsets $J_1, J_2, \dots, J_r$. Total subset frequencies $N_1, N_2, \dots, N_r$ are fixed and within each subset we have a simple multinomial distribution; between subsets the multinomials are independent. In what follows the subsets will correspond to fixed margins of the table. Other sampling schemes are possible (Haberman [1974a]), but will not be used here.

Population studies

Allowing classification error in the K dimensions of the table, the two basic assumptions on this error are as follows:

a) For $\ell = 1, 2, \dots, K$, the probability an observation is observed with error at a particular level in dimension ℓ , given its true level in dimension ℓ , does not depend upon the true levels of that observation in the other dimensions of the table. This is the extension of assumption (4) of Section 1.

b) The misclassifications of an observation in the different dimensions of the table are independent. That is, the probability an observation is misclassified in dimension ℓ does not depend on whether the observation was misclassified in the other dimensions of the table.

Using these assumptions, let Q_ℓ be the matrix of probabilities of all possible misclassifications in dimension ℓ , viz:

$$(12) \quad Q_\ell = ((q_{ij}^\ell))_{i,j=1,\dots,I_\ell}$$

where

$$q_{ij}^\ell = P(\text{observe level } i \text{ in dim } \ell | \text{true level } j \text{ in dim } \ell) \\ \text{for } \ell = 1, \dots, K.$$

If there is no classification error in dimension ℓ , then Q_ℓ is an identity matrix.

If the sampling scheme is simple multinomial and the expected cell counts would have been $\{\lambda(i_1 i_2 \dots i_K)\}$ with no classification error, then with classification error the distribution of the observed contingency table will also be simple multinomial with expected cell

counts $\{m(i_1 i_2 \dots i_K)\}$, where

$$(13) \quad m(i_1 i_2 \dots i_K) = \sum_{j_1 j_2 \dots j_K} q_{i_1 j_1}^1 q_{i_2 j_2}^2 \dots q_{i_K j_K}^K \lambda(j_1 j_2 \dots j_K) .$$

The sum is over all cells of the table $(j_1 j_2 \dots j_K)$. It is convenient to consider functions of the cells as T-vectors with the cells in lexicographical order. So, (13) can be written as

$$(14) \quad \underline{m} = Q \underline{\lambda}$$

where

$$Q = Q_1 \otimes Q_2 \otimes \dots \otimes Q_K$$

and $A \otimes B$ extends for the Kroneker product of matrices A and B.

If the sampling scheme is Poisson and the expected cell counts would have been $\underline{\lambda}$ with no classification error, then with classification error the distribution of the observed contingency table will also be Poisson with expected cell counts $\underline{m}$ still given by (14).

If the sampling scheme is product multinomial fixing the first L-way margin of the table, then

$$(15) \quad X(i_1 i_2 \dots i_L ++ \dots +) = N(i_1 i_2 \dots i_L)$$

for some fixed numbers $\{N(i_1 i_2 \dots i_L)\}$. If the expected cell counts would have been $\underline{\lambda}$ with no classification error, then with classification error the distribution of the observed table will still be product multinomial with (15) true and expected cell counts $\underline{m}$, where

$$m(i_1 \dots i_K) = \frac{N(i_1 \dots i_L)}{\tau(i_1 \dots i_L + \dots +)} \sum_{j_1 \dots j_K} q_{i_1 j_1}^1 \dots q_{i_K j_K}^K \frac{\pi(j_1 \dots j_L + \dots +)}{N(j_1 \dots j_L)} \cdot \lambda(j_1 \dots j_K) .$$

Here $\pi(j_1 \dots j_K)$ is the true population probability of cell $(j_1 j_2 \dots j_K)$, and $\tau = Q\pi$.

Dose-Response Studies

If the first L dimensions of the table correspond to the dose variables, then the sampling scheme is product multinomial with (15) true for some fixed numbers $\{N(i_1 i_2 \dots i_L)\}$. Allowing classification error both in the dose and response dimensions of the table, the basic assumptions on this error are as follows:

a) For the dose dimensions, the probability of a classification error in one dose dimension does not depend upon the levels of that observation in the other dose dimensions. Furthermore, the misclassifications done to an observation in the different dose dimensions of the table are independent.

b) The same as (a) but for the response dimensions of the table.

c) The probability of obtaining a certain response given the true response does not depend on the true levels of the observation in the dose variables.

d) The probability of a response given the true dose does not depend on the observed dose.

Using these assumptions, let Q_s be the matrix of probabilities of all possible misclassification in dose dimensions; viz:

$$Q_s = ((q_{ij}^s)) \quad i, j = 1, \dots, I_s$$

where

$$q_{ij}^s = P(\text{subject truly given dose level } j \text{ in dim } s : \text{subject assigned to get dose level } i \text{ in dim. } s) \\ \text{for } s = 1, 2, \dots, L.$$

For the response dimensions of the table, Q_ℓ are defined as in (12) for $\ell = L+1, \dots, K$.

If the expected cell counts would have been $\{\lambda(i_1 i_2 \dots i_K)\}$ with no classification error, then with classification error the distribution of the observed contingency table will still be product multinomial with (15) true and expected cell counts $\{m(i_1 i_2 \dots i_K)\}$, where

$$m(i_1 i_2 \dots i_K) = N(i_1 i_2 \dots i_L) \sum_{j_1 j_2 \dots j_K} q_{i_1 j_1}^1 \dots q_{i_K j_K}^K \frac{\lambda(j_1 j_2 \dots j_K)}{N(j_1 \dots j_L)}.$$

Summary

If the results for the population and dose/response studies are combined, then the before-error expected cell counts of the table, λ , and the after-error expected cell counts of the table, $\underline{m}$, are related by:

$$\underline{m} = Q_0 \lambda$$

where

$$Q_0 = D(X) Q D(\underline{z})$$

and

$$Q = Q_1 \otimes Q_2 \otimes \dots \otimes Q_K$$

and for any T-vector $\underline{w}$, $D(\underline{m})$ is the diagonal matrix with w_i on the i^{th} diagonal element. The vectors $\underline{y}$ and $\underline{z}$ are determined by the type of study and sampling scheme used. In either type of study, if the sampling scheme fixes the first L-way margin of the table as in (15), then $y(i_1 i_2 \dots i_K)$ and $z(i_1 i_2 \dots i_K)$ are functions of $(i_1 i_2 \dots i_L)$ only. Therefore, if there is no classification error across a fixed margin, i.e., no classification error in the first L dimensions of the table, then all the Q_i are column-stochastic and one may take $Q_0 = Q$.

In what follows it will frequently be assumed that the Q_ℓ are invertible. If this is not the case, then there is some redundancy in recording all the levels of the observations. For example, in a 2×2 table if the error matrix Q_1 across the rows is singular, then there is no information contained in the row classification of the observations.

CHAPTER 3

LOG-LINEAR MODELS AND MISCLASSIFICATION

In this chapter hierarchical log-linear models will be defined for the expected cell counts of a contingency table. Log-linear models refer to classes of contingency tables that have their vectors of log expected cell counts lying in particular linear spaces. Hierarchical log-linear models provide a parsimonious description of the interactions among the different dimensions of the table. A particular hierarchical log-linear model will refer to a whole class of contingency tables, and not just one table of expected cell counts. For example, the model of independence for $I \times J$ tables is a hierarchical log-linear model.

Misclassification will alter the expected cell counts λ to $Q_0 \lambda$ where $Q_0 = D(\underline{y})QD(\underline{z})$, $Q = Q_1 \otimes Q_2 \otimes \dots \otimes Q_K$, and $\underline{y}$ and $\underline{z}$ are determined by the way the table is sampled (Chapter 2, Section 3). The effect of classification error on one measure of independence of the 2×2 table will be examined and it will be shown that independence is preserved by misclassification. A log-linear model for a general table is said to be preserved by classification error if after the addition of classification error to a table in that model, the new table still belongs to that same model. When a model is preserved, tests of the hypothesis that a sampled contingency table belongs to that model will be of the correct significance level if the classification error is completely ignored. A simple criterion will be given to determine which hierarchical log-linear models are preserved by classification error.

1. Definition of Hierarchical Log-Linear Models

Log-linear models provide a concise description of the cell probabilities or expected cell counts of a contingency table. The general model (Haberman [1974a]) postulates that the T-vector of the log of the expected cell counts, $\log \lambda$, lies in an s-dimensional subspace $\mathcal{M}$ of $\mathbb{R}^T$. For any T-vector $\underline{x}$ and any univariate function f, the notation $f(\underline{x})$ will represent the T-vector with $f(x_i)$ as elements. Recall T is the number of cells in the table. A representation of the general model is:

$$(1) \quad \log \lambda = M\underline{u}, \quad \underline{u} \in \mathbb{R}^s$$

where M is a $T \times s$ design matrix.

The particular log-linear models that will be discussed in this thesis are of the "analysis of variance" type. For an $I_1 \times I_2 \times \dots \times I_K$ table, the logarithms of the expected cell counts equal the sum of the so-called "u terms":

$$(2) \quad \begin{aligned} \log \lambda(i_1 i_2 \dots i_K) = & u \\ & + u_1(i_1) + u_2(i_2) + \dots + u_K(i_K) \\ & + u_{12}(i_1 i_2) + u_{13}(i_1 i_3) + \dots + u_{K-1K}(i_{K-1} i_K) \\ & + u_{123}(i_1 i_2 i_3) + \dots \\ & \vdots \\ & + u_{123\dots K}(i_1 i_2 \dots i_K) . \end{aligned}$$

There is a close analogy between hierarchical log-linear models and the usual analysis of variance breakdown of a mean into main effects and interactions. Here u represents an overall mean, u_ℓ the main effects of dimension ℓ , u_{rt} the interactions between dimensions r

and t , etc. These effects and interactions involve the logarithms of the expected cell counts. The utility of these models lies in the fact that by postulating certain u terms to be 0, different models of partial independences among the dimensions of the table can be achieved.

In the parametrization (2), the sum over any index of any u term equals 0. For example,

$$\sum_{i_2} u_{123}(i_1 i_2 i_3) = 0.$$

The models considered will postulate certain of the u terms to be identically 0. These models can always be written in the form (1) by reparametrizing to eliminate redundant u terms. It is customary to consider only hierarchical models (Bishop, Fienberg, and Holland [1975]). The u term with subscript $\{a_1, a_2, \dots, a_r\}$ is said to be a lower-order relative of the u taken with subscript $\{b_1, b_2, \dots, b_t\}$ if $\{a_1, a_2, \dots, a_r\} \subseteq \{b_1, b_2, \dots, b_t\}$. A hierarchical model is one in which the lower-order relatives of every u term present in the model are also present in the model. The unsubscripted u term is also always assumed to be in the model. Other log-linear models using different design matrices M are not considered here because the structure for classification error described in Chapter 2 may not be appropriate. For example, in a logistic regression it makes more sense to put a continuous error distribution on the covariates rather than to assume there is classification error among the different assigned levels of the covariates.

Contingency tables which have some cells with expected counts equal to 0 (structural zeroes) are known as incomplete tables. The analysis of such tables could become more complicated than usual with the addition of classification error, since it is theoretically possible to have observations in a structural zero because of the misclassification. Incomplete tables will not be considered here.

2. The 2×2 Table and Classification Error

In the 2×2 table the log-linear models of interest are the completely saturated model where no u terms are set to 0, and the model of independence where u_{12} is set to 0. As will be seen, classification error pushes a non-independent table "towards" independence and therefore preserves an independent table.

In a 2×2 table with cell probabilities π , the completely saturated model is given by:

$$\log \pi(i_1 i_2) = u + u_1(i_1) + u_2(i_2) + u_{12}(i_1 i_2) \quad i_1, i_2 = 1, 2.$$

Since the sum of any u term over any index must equal 0, $|u_{12}(i_1 i_2)|$ must be constant for $i_1, i_2 = 1, 2$, and

$$(3) \quad u_{12}(11) = \frac{1}{4} \log \frac{\pi(11)\pi(22)}{\pi(12)\pi(21)}.$$

The ratios of the π 's given in (3) is known as the cross-product ratio for the table π . The model of independence is given by:

$$\log \pi(i_1 i_2) = u + u_1(i_1) + u_2(i_2) \quad i_1, i_2 = 1, 2.$$

This is, of course, equivalent to

$$\frac{\pi(11)}{\pi(12)} = \frac{\pi(21)}{\pi(22)}.$$

The value $u_{12}(11)$ is sometimes taken as a measure of the table's departure from independence--the further $u_{12}(11)$ is away from 0, the further the table is away from independence (Bishop, Fienberg, and Holland [1975]).

To investigate the effect of misclassification on the value of $u_{12}(11)$, the structure of classification error described in Chapter 2 is used. Let $u_{12}(\pi)$ be the $u_{12}(11)$ term associated with the without-error table π , and let $u_{12}(\tau)$ be the $u_{12}(11)$ term of the with-error table τ , $\tau = Q_0 \pi$. Since the cross-product ratio and therefore $u_{12}(11)$ is invariant under multiplication of the rows by arbitrary constants, there is no loss of generality in assuming the classification error is of the form $Q_0 = Q = Q_1 \otimes Q_2$. This is true because if the sampling scheme fixes the number of observations in each row of the table, then Q_0 and Q differ only by diagonal matrices which correspond to multiplying the row margin of the table by a constant (Section 3, Chapter 2).

Using $u_{12}(11)$ as a measure, it is seen that misclassification pushes a 2×2 table towards independence, viz:

Proposition 1: $|u_{12}(\tau)| \leq |u_{12}(\pi)|$.

The proof is straightforward and given in Appendix A.

Corollary: If $u_{12}(\pi) = 0$, then $u_{12}(\tau) = 0$.

An independent table without classification error will also be independent with such error. That is, the model of independence for 2×2 tables is preserved by classification error. For higher dimensional tables, examples of models which are preserved will be given in the next section.

The dependency ratio R ,

$$R = \frac{|u_{12}(\tau)|}{|u_{12}(\pi)|}$$

can be thought of as the reduction in dependence of the 2×2 table due to classification error. According to Proposition 1, it is less than or equal to 1. For any given table π and known classification error, the dependency ratio can easily be computed. Since the classification error is assumed to act independently in dimensions 1 and 2 of the table, it is sufficient to consider only error in dimension 1, say. We now demonstrate the interesting proposition that for a fixed error matrix Q_1 , the dependency ratio approaches a constant which is not 0 or 1 as π approaches an independent table through a sequence of tables with specified margins.

Proposition 2: Let $\pi^{(n)}$ be any sequence of 2×2 tables with constant positive margins, with $\pi^{(n)}$ approaching an independent table π . That is, each cell of $\pi^{(n)}$ approaches the corresponding cell of π as $n \rightarrow \infty$. Let $\tau^{(n)} = Q\pi^{(n)}$ where $Q = Q_1 \otimes Q_2$ and Q_2 is the identity matrix. Then

$$(1) \quad \lim_n \frac{u_{12}(\tau^{(n)})}{u_{12}(\pi^{(n)})} = \frac{\pi(1+)\pi(2+)|Q_1|}{[\pi(1+)(q_{22}^1 - q_{21}^1) - q_{22}^1][\pi(2+)(q_{11}^1 - q_{12}^1) - q_{11}^1]}$$

where $Q_1 = ((q_{ij}^1))$ and $|Q_1|$ is the determinant of Q_1 .

The proof uses L'Hospital's rule and is given in Appendix A. It is seen that the limiting dependency ratio (1) does not involve the margins of dimension 2. For a population study, q_{21}^1 is the false positive

rate, q_{12} is the false negative rate, $q_{11} = 1 - q_{21}$, and $q_{22} = 1 - q_{12}$. Figure 1 contains the values of the ratio (1) for different false positive and false negative rates. In Figure 1 each curve corresponds to a different false positive rate while the horizontal axis corresponds to the false negative rate. The limiting independent table has been chosen to have 80% negatives and 20% positives. It is seen that an increase in error rates decreases the value of the limiting dependency ratio (1). Furthermore, an increase in the false positive rate decreases the value of the limiting ratio (1) more than the same increase in the false negative rate. Heuristically, a false positive rate has a larger effect when there are fewer positives.

Remark: For more general models of classification error that allow the error rates in one dimension of the table to depend upon the levels of the observation in the other dimension of the table, the results of this section are no longer true (Keys and Kihlberg [1963], Goldberg [1975]).

3. Models Preserved by Misclassification

As stated in the last section, the model of independence is preserved by misclassification in the 2×2 table. Mote and Anderson [1965] showed that the model of independence is preserved for an $I_1 \times I_2$ table. In this section more complicated hierarchical models on higher dimensional contingency tables will be examined. It will be shown that some models are preserved under misclassification while others are not. Preservation implies that the significance levels of the usual tests of a null hypothesis that a sampled contingency table belongs

FIGURE 1

to that particular model are unaffected by classification error. A simple way to determine if classification error in certain dimensions of a table will preserve a model will be given.

A hierarchical log-linear model can be described by $\mathcal{M}$, the linear space of the log expected cell counts, or by the u terms present in the model. A model is said to be preserved by the error matrix Q_0 if whenever the log of the without-error expected cell counts, $\log \lambda$, is in $\mathcal{M}$, then the log of the with-error expected cell counts, $\log Q_0 \lambda$, is also in $\mathcal{M}$. If the sampling scheme fixes dimensions 1, 2, ..., L of the table, then allowable hierarchical models must contain the u term $u_{12 \dots L}$ and all its lower-order relatives (Bishop, Fienberg, and Holland [1975]). This implies that if a table of expected cell counts is in an allowable model, then after multiplication by arbitrary constants of any margin fixed by the sampling scheme, the table will still be in the same model. Since Q_0 and $Q = Q_1 \otimes Q_2 \otimes \dots \otimes Q_K$ differ only by diagonal matrices which correspond to multiplying margins fixed by the sampling scheme by constants (Chapter 2, Section 3), there is no loss of generality in assuming $Q_0 = Q$. It makes sense, therefore, to talk about a model being preserved by error occurring in particular dimensions of the table:

Definition: The log-linear model $\mathcal{M}$ is preserved by classification error in dimension ℓ of the table if:

$$\log \lambda \in \mathcal{M} \text{ implies that } \log(Q_\lambda) \in \mathcal{M}$$

for all $Q = Q_1 \otimes Q_2 \otimes \dots \otimes Q_K$ where Q_i is an identity matrix for $i \neq \ell$.

The model will be said to be preserved if it is preserved by classification error in all the dimensions of the table.

Using the u terms there is a simple way to determine if a model will be preserved by classification error in dimension ℓ of the table, viz:

Proposition 3: A hierarchical log-linear model is preserved by classification error in dimension ℓ of the table if and only if u_ℓ is present in the model and all u terms present in the model containing an ℓ as a subscript are lower-order relatives of a single u term present in the model.

Recall that the u term with subscript $\{a_1, a_2, \dots, a_r\}$ is a lower-order relative of the u term with subscript $\{b_1, b_2, \dots, b_t\}$ if $\{a_1, \dots, a_r\} \subseteq \{b_1, b_2, \dots, b_t\}$. The proof of Proposition 3 is given in Appendix A.

We end this section with some examples--the partial independence description of the models can be found in Birch [1963].

Example 1: $I_1 \times I_2 \times \dots \times I_K$ completely independent table--model preserved by classification error in all dimensions

$$\log \lambda(i_1 i_2 \dots i_K) = u + u_1(i_1) + u_2(i_2) + \dots + u_K(i_K) .$$

For any dimension ℓ , u_ℓ is the only u term containing an ℓ , so Proposition 3 implies the model is present in dimension ℓ .

Example 2: $I_1 \times I_2 \times \dots \times I_K$ completely saturated model--model preserved by classification error in all dimensions

$$\log \lambda(i_1 i_2 \dots i_K) = \text{sum of all } u \text{ terms} .$$

All u terms are lower-order relatives of the single u term $u_{12\dots K}$.

Example 3: $I_1 \times I_2 \times I_3$ table, dimensions 1 and 2 together independent of dimension 3--model preserved by classification error in all dimensions

$$\log \lambda(i_1 i_2 i_3) = u + u_1(i_1) + u_2(i_2) + u_3(i_3) + u_{12}(i_1 i_2) .$$

All u terms containing a 1 or 2 are lower-order relatives of u_{12} .

All u terms containing a 3 are lower-order relatives of u_3 .

Example 4: $I_1 \times I_2 \times I_3$ table, dimensions 1 and 2 conditionally independent given dimension 3--model preserved by classification error in dimensions 1 and 2, but not dimension 3

$$\log \lambda(i_1 i_2 i_3) = u + u_1(i_1) + u_2(i_2) + u_3(i_3) + u_{13}(i_1 i_3) + u_{23}(i_2 i_3) .$$

All u terms containing a 1 are lower-order relatives of u_{13} . All u terms containing a 2 are lower-order relatives of u_{23} . For dimension 3, however, u_{13} and u_{23} are not lower-order relatives of a single u term in the model. This is the first example of a model that is not preserved by classification error in all dimensions, so a specific table will be given:

		2	3
1	10	20	160
	20	40	40
			10

In the above $2 \times 2 \times 2$ table, the rows represent dimension 1, the columns dimension 2, and the two tables represent dimension 3. This table has

dimensions 1 and 2 conditionally independent given dimension 3 as is easily checked by computing the cross-product ratio to be identically 1.0 in both levels of dimension 3. If the following classification error matrix is applied to dimension 3 of the table,

$$Q_3 = \begin{pmatrix} .9 & .1 \\ .1 & .9 \end{pmatrix}$$

then the after-error expected cell counts will be:

		2	3								
1		<table><tr><td>25</td><td>22</td></tr><tr><td>22</td><td>37</td></tr></table>	25	22	22	37	<table><tr><td>145</td><td>38</td></tr><tr><td>38</td><td>13</td></tr></table>	145	38	38	13
	25	22									
22	37										
145	38										
38	13										

This table is no longer in the model as is checked by noting the cross-product ratios are 1.91 and 1.31 in levels 1 and 2, respectively, of dimension 3. When a model is not preserved by classification error, spurious non-zero values of u terms can appear. In this example, the values of u_{123} and u_{12} are non-zero in the after-error table but zero in the before-error table.

Example 5: $I_1 \times I_2 \times I_3$ table, no second order interaction--model not preserved by classification error in any dimension

$$\begin{aligned} \log \lambda(i_1 i_2 i_3) = & u + u_1(i_1) + u_2(i_2) + u_3(i_3) + u_{12}(i_1 i_2) \\ & + u_{13}(i_1 i_3) + u_{23}(i_2 i_3) . \end{aligned}$$

For dimension 1, u_{12} and u_{13} are not lower-order relatives of a single u term in the model. Similarly for dimensions 2 and 3.

CHAPTER 4

ESTIMATING AND TESTING HIERARCHICAL LOG-LINEAR MODELS

The log-linear model analysis of a contingency table $\underline{x}$ sampled with classification error is considered in this chapter. Using the structure of classification error described in Chapter 2, it is clear that if the error matrix Q is unknown, then there is an identification problem in estimating cell probabilities--many combinations of Q and cell probabilities will yield the exact same sampling distribution on $\underline{x}$. One way around this problem is to collect additional data with $\underline{x}$. Tenenbein [1969,1970] suggests using a double sampling scheme where the true classifications of a subsample of the observations falling in $\underline{x}$ can be obtained. This is the method used by Chiacchierini and Arnold [1977], and Hochberg [1977]. Koch [1969], in the context of response errors in sample surveys, suggests that observations (people) can be sampled many times to get a distribution of responses around the "true" response. The approach in this chapter is to assume Q is fixed and known. This is also the approach of Press [1968]. It is important to know the effect of a certain misclassification on a log-linear model analysis even if the exact misclassification is not known. For many analyses the effect of adding classification error will be dramatic, but the analyses will not change much as the classification error is varied.

In the simplest formulation, classification error changes the expected cell counts $\underline{\lambda}$ of the contingency table to $\underline{m}$, where

$$\underline{m} = Q_1 \otimes Q_2 \otimes \cdots \otimes Q_K \underline{\lambda} .$$

If one knew what $\underline{m}$ was, then one could solve for $\underline{\lambda}$, viz:

$$\begin{aligned}\underline{\lambda} &= (Q_1 \otimes Q_2 \otimes \cdots \otimes Q_K)^{-1} \underline{m} \\ &= Q_1^{-1} \otimes Q_2^{-1} \otimes \cdots \otimes Q_K^{-1} \underline{m}.\end{aligned}$$

Of course, one doesn't know what $\underline{m}$ is, but hopes to estimate it from the data $\underline{x}$ which has expected cell counts $\underline{m}$.

The hierarchical log-linear model analysis of a contingency table $\underline{x}$ sampled with no classification error is concerned with estimating u terms under a specific model, and testing between alternative models. Maximum likelihood estimating and testing are the methods usually used to perform such analyses (Bishop, et al. [1975]). Weighted least squares (Grizzle, et al. [1969]) is an alternative method of estimation that has the advantage of not requiring iteration. Tenenbein [1969,1970] uses maximum likelihood and Hochberg [1977] uses weighted least squares estimation for a double sampling scheme. Maximum likelihood estimation will be used in this chapter because the simple iterative schemes used to get the maximum likelihood estimates in the no-error case can easily be extended to the with-error case when Q is specified.

In Section 1 the log-likelihood is examined for local maxima. In Section 2 the asymptotic distribution of the maximum likelihood estimate of the expected cell counts is examined as the number of observations in the table becomes large. In Section 3, the asymptotic distributions of the log-likelihood ratio and Pearson chi-square statistics for testing between different models are examined under null and alternative hypotheses. The comparison of Pitman asymptotic powers of such tests with and without classification error gives the increase in sample

size necessary to compensate for the loss of power those tests have when there is classification error. A general formula for this increase in sample size is given and some special cases are examined. Throughout this chapter attention is restricted to contingency tables sampled with no classification error across any margin being held fixed by the sampling scheme (Chapter 2, Section 3), this being the case commonly encountered in practice.

1. Maximum Likelihood Estimation of Expected Cell Counts

For a log-linear model without classification error it is known that the maximum-likelihood estimates (mle) of the expected cell counts are unique and are the same whether Poisson, simple multinomial, or product multinomial sampling is assumed (Birch [1963]). The existence of the maximum likelihood estimates is guaranteed when all the observed cell counts are positive. For Poisson sampling, the log of the likelihood is proportional to

$$(1) \quad \sum_{i=1}^T (x_i \log \lambda_i - \lambda_i)$$

where

$$\lambda = \exp\{\mu\}$$

is the T-vector of expected cell counts, and x_i is the number of observations in cell i . For notational simplicity, the single subscript i is standing for the multiple subscript $(i_1 i_2 \dots i_K)$ of the previous sections. To get the maximum likelihood estimates, expression (1) is maximized over $\mu \in \mathcal{M}$, the linear space corresponding to the log-linear

model in question. Sometimes closed-form solutions exist for the mle; other times a numerical method must be used. In any event, iterative proportional fitting, a simple numerical method, exists for finding the maximum likelihood estimates (Bishop, et al. [1975]).

With classification error matrix Q , the log of the likelihood for Poisson sampling is proportional to

$$(2) \quad \ell(\underline{x}, \underline{\mu}) \equiv \sum_{i=1}^T (x_i \log m_i - m_i),$$

where

$$\underline{m} = Q\underline{\lambda},$$

and

$$\underline{\lambda} = \exp\{\underline{\mu}\}.$$

To get the maximum likelihood estimate of $\underline{\mu}$, $\ell(\underline{x}, \underline{\mu})$ is maximized over $\underline{\mu} \in \mathcal{M}$. This is a distinct problem from the no-error case--the log expected cell counts $(\log \underline{m})$ may no longer fall in a linear manifold $\mathcal{M}$.

For this maximization problem it is useful to look at the vector of partial derivatives, $d_{\underline{\mu}} \ell(\underline{x})$, and the matrix of second partial derivatives, $d_{\underline{\mu}}^2 \ell(\underline{x})$, of $\ell(\underline{x}, \underline{\mu})$ with respect to $\underline{\mu}$:

$$(3) \quad \begin{aligned} d_{\underline{\mu}} \ell(\underline{x}) &\equiv \left(\frac{\partial \ell(\underline{\mu}, \underline{x})}{\partial \mu_i} \right)_i \\ &= D(\underline{\lambda})Q'D^{-1}(\underline{m})\underline{x} - \underline{\lambda} \end{aligned}$$

$$\begin{aligned}
d^2_{\mu} \ell(x) &\equiv \left(\left(\frac{\partial^2 \ell(\mu, x)}{\partial \mu_i \partial \mu_j} \right) \right)_{i,j} \\
(4) \qquad &= D(d\ell_{\mu}(x)) \\
&\quad - D(\hat{\lambda}) Q' D^{-1}(\hat{m}) D(\hat{x}) D^{-1}(\hat{m}) Q D(\hat{\lambda}) .
\end{aligned}$$

Recall $D(y)$, for a vector y , represents the diagonal matrix with y_i on the i^{th} diagonal element.

In the no-error case, these reduce to

$$\begin{aligned}
d\ell_{\mu}(x) &= x - \lambda \\
d^2_{\mu} \ell(x) &= -D(\lambda) .
\end{aligned}$$

So, $\ell(x, \mu)$ is strictly concave in μ and the unique maximum likelihood estimate of μ for the no-error case is given by the solution to:

$$\begin{aligned}
(5) \qquad P_{\mathcal{M}} d\ell_{\hat{\mu}}(x) &= P_{\mathcal{M}} x - P_{\mathcal{M}} \hat{\lambda} = 0 , \\
\hat{\lambda} &= \exp\{\hat{\mu}\}, \quad \hat{\mu} \in \mathcal{M},
\end{aligned}$$

where $P_{\mathcal{M}} x$ represents the orthogonal projection of x onto the linear space $\mathcal{M}$ and orthogonality refers to the usual inner product on $\mathbb{R}^T$.

With classification error the matrix $d^2_{\mu} \ell(x)$ is no longer negative definite on $\mu \in \mathcal{M}$. Nor is it true anymore that

$$\lim_{\mu_i \rightarrow -\infty} \ell(x, \mu) = -\infty .$$

This means the maxima of $\ell(x, \mu)$ may not be achieved for any finite μ . A complete investigation into the log-likelihood for finite sample sizes will not be presented here. The critical points of $\ell(x, \mu)$ are given

by:

$$(6) \quad \begin{aligned} P_{\mathcal{M}}^{d\hat{\mu}}(\underline{x}) &= P_{\mathcal{M}}^{D(\hat{\lambda})Q'D^{-1}(\hat{m})\underline{x}} - P_{\mathcal{M}}^{\hat{\lambda}} = 0 \\ \hat{m} &= Q\hat{\lambda}, \quad \hat{\lambda} = \exp(\hat{\mu}), \quad \hat{\mu} \in \mathcal{M}. \end{aligned}$$

These are the maximum likelihood equations for Poisson sampling. The maximum likelihood equations for multinomial sampling are given by

$$(7) \quad \begin{aligned} P_{\mathcal{M}}^{D(\hat{\lambda})Q'D^{-1}(\hat{m})\underline{x}} &= 0 \\ \hat{m} &= Q\hat{\lambda}, \quad \hat{\lambda} = \exp(\hat{\mu}), \quad \hat{\mu} \in \mathcal{M}, \end{aligned}$$

and $\hat{m}$ must satisfy the multinomial constraints (Section 3, Chapter 2).

Proposition 1: The maximum likelihood equations are the same whether Poisson or multinomial sampling is assumed for $\underline{x}$.

Pf: The proof, given in Appendix B, involves showing that a solution $\underline{m}$ to (6) will actually satisfy any multinomial constraints that $\underline{x}$ does.

Proposition 1 is well-known in the no-error case (Birch [1963]).

To find the maximum likelihood estimate of μ in the presence of classification error, one can use a general maximization algorithm to maximize $\ell(\underline{x}, \mu)$ over $\mu \in \mathcal{M}$. The similarity of (5) and (6), however, suggests modifying iterative proportional fitting in the no-error case to get the solutions to (6). A brief description of this method will be given here; a detailed description with examples is given in Appendix C. If $\underline{y} = D(\hat{\lambda})Q'D^{-1}(\hat{m})\underline{x}$ were known, then the solution $\hat{\lambda}$ to the equations (6) would be precisely the solution $\hat{\lambda}$ to the equations (5) when $\underline{y}$ is substituted for $\underline{x}$ in (5). Iterative

proportional fitting can always be used to solve the equations (5), sometimes closed-form solutions for $\hat{\lambda}$ exist. Since y is not known, an initial estimate $\lambda^{(0)}$ of $\hat{\lambda}$ is used to get an initial estimate $y^{(0)}$ of y . Solving the equations (5) substituting $y^{(0)}$ for x yields a new estimate $\lambda^{(1)}$ of $\hat{\lambda}$. This procedure is iterated yielding a sequence of estimates $\lambda^{(i)}$ which approach $\hat{\lambda}$ if convergent. This is Algorithm 1 given in Appendix C.

Remark: It is possible to view observing a contingency table with classification error as an incomplete data problem. For each observation in an $I_1 \times I_2 \times \dots \times I_K$ table, one imagines the with-error classification $(i_1 i_2 \dots i_K)$ and the without-error (true) classification $(j_1 j_2 \dots j_K)$. The without-error classification is unobserved. To put this in a contingency table context, one imagines an $(I_1 \times I_2 \times \dots \times I_K) \times (I_1 \times I_2 \times \dots \times I_K)$ "super" table. A "super" observation $((i_1 i_2 \dots i_K), (j_1 j_2 \dots j_K))$ has $2K$ dimensions--the first K correspond to the with-error classification and the last K to the true unobserved classification. A typical cell $((i_1 i_2 \dots i_K), (j_1 j_2 \dots j_K))$ of the super table contains the number of observations with observed levels $(i_1 i_2 \dots i_K)$ and true levels $(j_1 j_2 \dots j_K)$. When one observes x , one is observing the first K -dimensional margin of the super table summed over the last K dimensions. That is, $x(i_1 i_2 \dots i_K)$ is the sum over all $(j_1 j_2 \dots j_K)$ of the number of super observations falling in cells $((i_1 i_2 \dots i_K), (j_1 j_2 \dots j_K))$ of the super table. One is therefore observing the super table "indirectly" (Haberman [1974b]). The methods of Haberman [1974b, 1977] can be applied to the maximum likelihood estimation problem here.

In fact, Proposition 1 here can be derived as a special case of Theorem 2 of Haberman [1974b], and Algorithm 1 is a special case of one discussed in Haberman [1977]. Furthermore, observing a contingency table indirectly can be put in the framework of the incomplete data problem discussed in Dempster, et al. [1977]. Algorithm 1 is also a special case of the "EM algorithm" given in Dempster, et al. [1977].

When the log-linear model is preserved by classification error (Section 2, Chapter 3), the problem of finding maximum likelihood estimates simplifies considerably. Let $\tilde{m}$ be the mle of the expected cell counts assuming there is no classification error. This rule can be found using standard log-linear model techniques. The following proposition shows that $\tilde{m}$ will also be the unique mle of the expected cell counts in the presence of classification error provided $(Q^{-1}\tilde{m})_i > 0$ for all i , i.e., provided all the elements of the vector $Q^{-1}\tilde{m}$ are positive.

Proposition 2: Let the log-linear model $\mathcal{M}$ be preserved by classification error in dimensions $1, 2, \dots, J$ of the table. Let the error matrix Q have no classification error in dimension $J+1, \dots, K$ of the table. Let $\tilde{m}$ be the mle of the expected cell counts assuming there is no classification error, i.e., the solution to

$$\max_{\log m \in \mathcal{M}} \sum_{i=1}^T (x_i \log m_i - m_i) .$$

If $(Q^{-1}\tilde{m})_i > 0$ for all i , then $\tilde{m}$ is also the mle of the with-error expected cell counts, i.e., the solution to

$$\max_{\log Q^{-1} \tilde{m} \in \mathcal{M}} \sum_{i=1}^T (x_i \log m_i - m_i) .$$

The proof is given in Appendix B.

If $(Q^{-1} \tilde{m})_i = 0$ for some i , then the "mle" of the without-error expected cell counts λ will have some $\lambda_j = 0$. In terms of $\mu = \log \lambda$, there will be no $\mu \in \mathcal{M}$ such that $\mu = \log \lambda$. Strictly speaking, therefore, there is no mle for μ , λ , or $\tilde{m}$ in this case. For example, suppose the observed table x is:

$$x = \begin{array}{|c|c|} \hline 50 & 25 \\ \hline 11 & 25 \\ \hline \end{array} .$$

Suppose further the known classification error matrix has only error across the rows (dimension 1) given by:

$$Q_1 = \begin{array}{|c|c|} \hline .8 & .2 \\ \hline .2 & .8 \\ \hline \end{array} .$$

If the fully saturated model is fit to the data x , then $\tilde{m}$ will be precisely x . So, one has $Q^{-1} \tilde{m} = Q^{-1} x$, viz:

$$Q^{-1} \tilde{m} = \begin{array}{|c|c|} \hline 63 & 25 \\ \hline -2 & 25 \\ \hline \end{array} .$$

Allowing λ to be an arbitrary non-negative 2×2 table, it is easy to check that the likelihood given the data x is maximized at

$$\hat{\lambda} = \begin{array}{|c|c|} \hline 61 & 25 \\ \hline 0 & 25 \\ \hline \end{array} .$$

This λ does not correspond to a finite μ .

Tables of $Q^{-1}\tilde{x}$ with cells containing negative entries of large magnitude suggest Q has been misspecified. However, for a table with many cells it would not be surprising to get negative cells in $Q^{-1}\tilde{x}$ by chance even when Q is correctly specified. An ad hoc procedure to get estimates of the expected cell counts is to add $|\alpha| + .5$ to all cells in the table $\tilde{x}$, where α is the most negative value in any cell of $Q^{-1}\tilde{x}$. The mle of the expected cell counts using this new table as the observed data would then be computed. The addition of $|\alpha|$ to all the cells in $\tilde{x}$ insures $Q^{-1}\tilde{x}$ will have all non-negative cells. The further addition of .5 insures $Q^{-1}\tilde{x}$ will have all positive entries. In the no-error case, adding .5 to all cells in a table is only one of many possible procedures to smooth a table with observed zeros (Bishop, Fienberg, and Holland [1975]).

In the example described above, 2.5 would be added to all four cells of $\tilde{x}$ to form $\tilde{x}_1$, say. Since $\tilde{m}_1$ based on $\tilde{x}_1$ is precisely $\tilde{x}_1$, one has $(Q^{-1}\tilde{m}_1)_i > 0$, for all i . Therefore, the mle of $\tilde{m}$ for the fully saturated model, based on data $\tilde{x}_1$, is $\tilde{x}_1$.

2. Asymptotic Distributions of Maximum Likelihood Estimates

In this section the asymptotic distributions of the maximum likelihood estimates of the expected cell counts and the u terms for hierarchical log-linear models will be examined as the number of observations in the table becomes large. Let $\tilde{x}^{(n)}$ represent data in a contingency table with expected cell counts $\tilde{m}^{(n)}$ such that

$$(8) \quad \tilde{m}^{(n)} = Q\lambda^{(n)}$$

where

$$(9) \quad \mu^{(n)} = \log \lambda^{(n)}$$

and $\mu^{(n)}$ is in a linear space $\mathcal{M}$ corresponding to a hierarchical log-linear model (Section 1, Chapter 3). Depending on the sampling hypothesis, the contingency tables $\{\tilde{x}^{(n)} | n = 1, 2, \dots\}$ have a Poisson, simple multinomial or product multinomial distribution. The type of distribution and the error matrix Q are fixed for all n . It is assumed that

$$(10) \quad \lim_n \frac{\tilde{m}^{(n)}}{n} = \tilde{m}^*$$

and

$$\mu^* \in \mathcal{M}$$

where

$$(11) \quad \mu^* = \log \lambda^* = \log Q^{-1} \tilde{m}^*.$$

This implies that

$$(12) \quad \lim_n (\mu^{(n)} - (\log n) \underline{e}) = \mu^*,$$

where $\underline{e}$ is the T -vector of all ones. Recall T is the number of cells in the contingency table, and tables are considered as T -vectors with the cells in lexicographical order.

Based on data $\tilde{x}^{(n)}$, let

$$(13) \quad \hat{\tilde{m}}^{(n)} = Q \hat{\lambda}^{(n)} = Q \exp(\hat{\mu}^{(n)}) = Q \exp(M \hat{\mu}^{(n)})$$

represent the maximum likelihood estimates. Recall that M is the $T \times s$ design matrix for $\mathcal{M}$ (Section 1, Chapter 3). The following proposition gives the asymptotic distributions of the maximum likelihood estimates of the u terms and the expected cell counts when all the $\underline{x}^{(n)}$ have Poisson distributions.

Proposition 3: Let $\underline{x}^{(n)}$ be a sequence of contingency tables having Poisson distributions with expected cell counts satisfying (8)-(12). Then as $n \rightarrow \infty$,

$$(a) \quad n^{1/2}(\hat{\underline{\mu}}^{(n)} - \underline{\mu}^{(n)}) \xrightarrow{D} \mathcal{N}(0, \Sigma_1)$$

$$(b) \quad n^{-1/2}(\hat{\underline{m}}^{(n)} - \underline{m}^{(n)}) \xrightarrow{D} \mathcal{N}(0, \Sigma_2)$$

$$(c) \quad n^{1/2}(\hat{\underline{u}}^{(n)} - \underline{u}^{(n)}) \xrightarrow{D} \mathcal{N}(0, \Sigma_3)$$

where the D over the arrows stands for convergence in distribution, and $\mathcal{N}(0, \Sigma)$ stands for a multivariate normal distribution with mean vector 0 and covariance matrix Σ . The matrices Σ_i are given by

$$\Sigma_1 = M(M'D(\underline{\lambda}^*)Q'D^{-1}(\underline{m}^*)QD(\underline{\lambda}^*)M)^{-1}M'$$

$$\Sigma_2 = QD(\underline{\lambda}^*)\Sigma_1D(\underline{\lambda}^*)Q'$$

$$\Sigma_3 = (M'D(\underline{\lambda}^*)Q'D^{-1}(\underline{m}^*)QD(\underline{\lambda}^*)M)^{-1}$$

where $D(\underline{\gamma})$ is the diagonal matrix with $\{\gamma_i\}$ on the diagonal.

Remark: The asymptotic covariance matrix Σ_3 of the u terms is the inverse of the Fisher information for $\underline{u}$ evaluated at $\underline{u}^*$, as

is seen by taking the expected value of expression (4) and noting that $\underline{\mu} = \underline{M}\underline{\lambda}$.

For $\underline{x}^{(n)}$ having a Poisson or multinomial distribution, let $\mathcal{N}$ be the linear space of fixed margins (Appendix B). In particular, if the $\underline{x}^{(n)}$ is Poisson, then $\mathcal{N} = \langle \underline{0} \rangle$; if the $\underline{x}^{(n)}$ is simple multinomial, then $\mathcal{N} = \langle \underline{e} \rangle$.

Proposition 4: Let $\underline{x}^{(n)}$ be a sequence of contingency tables having space of fixed margins $\mathcal{N}$ with expected cell counts satisfying (8)-(12).

Then as $n \rightarrow \infty$,

$$(a) \quad n^{1/2}(\hat{\underline{\mu}}^{(n)} - \underline{\mu}^{(n)}) \xrightarrow{D} \mathcal{N}(0, \Sigma_1)$$

$$(b) \quad n^{-1/2}(\hat{\underline{m}}^{(n)} - \underline{m}^{(n)}) \xrightarrow{D} \mathcal{N}(0, \Sigma_2)$$

$$(c) \quad n^{1/2}(\hat{\underline{u}}^{(n)} - \underline{u}^{(n)}) \xrightarrow{D} \mathcal{N}(0, \Sigma_3).$$

The covariance matrices Σ_i are given by

$$\begin{aligned} \Sigma_1 &= M(M'D(\underline{\lambda}^*)Q'D^{-1}(\underline{m}^*)QD(\underline{\lambda}^*)M)^{-1}M' \\ &\quad - N(N'D(\underline{m}^*)N)^{-1}N' \end{aligned}$$

$$\Sigma_2 = QD(\underline{\lambda}^*) \Sigma_1 D(\underline{\lambda}^*)Q'$$

$$\Sigma_3 = (M'D(\underline{\lambda}^*)Q'D^{-1}(\underline{m}^*)QD(\underline{\lambda}^*)M)^{-1}$$

$$= \left(\begin{array}{c|c} (N'D(\underline{m}^*)N)^{-1} & 0 \\ \hline 0 & 0 \end{array} \right) \begin{array}{l} r \\ s-r \end{array}$$

r s-r

For Σ_1 and Σ_2 , N is defined to be any $T \times r$ matrix with range equal to η . To get the simple expression here for Σ_3 , M and N must be chosen in the following special way: If the sampling scheme fixes the $(12 \dots L)$ margin of the table, then the order in which the u terms appear in the model should be such that the r lower-order relatives of $u_{12\dots L}$ come first (this determines the order of the columns of M). The matrix N is then taken to be the first r columns of M .

Proposition 3 is a special case of Proposition 4. The proof of Proposition 4 is given in Appendix D and uses the implicit function theorem to find consistent roots of the maximum likelihood equations. Taylor series arguments are used to get the asymptotic distributions of the maximum likelihood estimates.

If the sampling scheme is simple multinomial, then $\eta = \langle e \rangle$ and Σ_3 is the same as in Poisson sampling except the asymptotic variance of the unsubscripted u term is reduced. In general, if the sampling scheme fixes the $(12 \dots L)$ margin, then the asymptotic covariance of u_{θ_1} and u_{θ_2} will be the same as in the Poisson case, except when both u_{θ_1} and u_{θ_2} are lower-order relatives of $u_{12\dots L}$.

Example: Let the $\tilde{x}^{(n)}$ be $2 \times 2 \times 2$ tables having without-error expected cell counts with no second order interaction (Example 5 of Section 3, Chapter 3). If the order of the nonredundant u terms is

$$\tilde{u} = (u, u_1, u_2, u_3, u_{12}, u_{13}, u_{23}),$$

then the design matrix M is given by

$$M = (\underline{e}, \underline{e}_1, \underline{e}_2, \underline{e}_3, \underline{e}_{12}, \underline{e}_{13}, \underline{e}_{23})$$

where

$$\underline{e}' = (1 \ 1 \ 1 \ 1 \quad 1 \ 1 \ 1 \ 1)$$

$$\underline{e}_1' = (1 \ 1 \ 1 \ 1 \quad -1 \ -1 \ -1 \ -1)$$

$$\underline{e}_2' = (1 \ 1 \ -1 \ -1 \quad 1 \ 1 \ -1 \ -1)$$

$$\underline{e}_3' = (1 \ -1 \ 1 \ -1 \quad 1 \ -1 \ 1 \ -1)$$

$$\underline{e}_{12}' = (1 \ 1 \ -1 \ -1 \quad -1 \ -1 \ 1 \ 1)$$

$$\underline{e}_{13}' = (1 \ -1 \ 1 \ -1 \quad -1 \ 1 \ -1 \ 1)$$

$$\underline{e}_{23}' = (1 \ -1 \ -1 \ 1 \quad 1 \ -1 \ -1 \ 1) .$$

If, for example, the sampling scheme of the $\underline{x}^{(n)}$ was product multinomial fixing the dimension 1 - dimension 3 margin of the table, then in order to get the simple expression here for $\sum_3$, the order of the u terms in the model should be

$$\underline{u} = (u \ u_1 \ u_3 \ u_{13} \ u_2 \ u_{12} \ u_{23}) ,$$

giving a design matrix M such that

$$M = (\underline{e}, \underline{e}_1, \underline{e}_3, \underline{e}_{13}, \underline{e}_2, \underline{e}_{12}, \underline{e}_{23})$$

and

$$N = (\underline{e}, \underline{e}_1, \underline{e}_3, \underline{e}_{13}) .$$

Remark: In the no-error case (Haberman [1974a]), $\sum_1$ reduces to

$$\sum_1 = M(M'D(\underline{m}^*)M)^{-1}M' - N(N'D(\underline{m}^*)N)^{-1}N' .$$

3. Asymptotic Distributions of Test Statistics

This section considers two testing situations. The first is a simple null hypothesis versus a composite alternative, i.e., the expected cell counts are hypothesized to equal a specific table versus lying in a particular log-linear model (Propositions 5,6,7). The second testing situation is a composite null hypothesis versus a larger composite alternative, i.e., the expected cell counts are hypothesized to lie in a specific log-linear model versus lying in a particular larger log-linear model (Propositions 9,10,11). Propositions 8 and 12 compare the Pitman asymptotic power of these hypothesis tests with and without classification error in both testing situations, respectively.

For testing the null hypothesis

$$H_0 : \mu = \mu^0$$

against the alternative hypothesis

$$H_A : \mu \in \mathcal{M}$$

where $\mu^0 \in \mathcal{M}$ is a fixed table, one rejects for large values of the likelihood-ratio statistic

$$(14) \quad -2\Delta(\hat{x}, \hat{\mu}, \mu^{(0)}) = -2 \sum_{i=1}^T x_i \log \frac{m_i^{(0)}}{\hat{m}_i}$$

or the Pearson chi-square statistic

$$c(\hat{\mu}, \mu^{(0)}) = \sum_{i=1}^T \frac{(\hat{m}_i - m_i^{(0)})^2}{m_i^{(0)}}.$$

To compute the asymptotic distribution of these statistics, let $\underline{x}^{(n)}$ again represent a sequence of contingency tables with expected cell counts $\underline{m}^{(n)}$ satisfying (8)-(12). The sampling scheme of the $\underline{x}^{(n)}$ is characterized by the space of fixed margins $\mathcal{N}$ (Appendix B). Recall s is the dimension of $\mathcal{M}$, and r is the dimension of $\mathcal{N}$ which equals the number of sampling constraints on the tables $\underline{x}^{(n)}$.

Proposition 5: Consider a sequence of null hypotheses

$H_0 : \underline{\mu}^{(n)} = \underline{\mu}^{(n,o)} \in \mathcal{M}$ satisfying (12), and a sequence of contingency tables $\underline{x}^{(n)}$ with expected cell counts satisfying (8)-(12), and with space of fixed margins $\mathcal{N}$. If these null hypotheses are true, then $-2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n,o)})$ and $C(\hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n,o)})$ are asymptotically equivalent, that is, their difference converges in probability to 0 as $n \rightarrow \infty$. Here $\hat{\underline{\mu}}^{(n)}$ is the maximum likelihood estimate of $\underline{\mu} \in \mathcal{M}$ based on data $\underline{x}^{(n)}$. Furthermore,

$$\begin{aligned} & \lim_{n \rightarrow \infty} P\{-2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n,o)}) > \chi_{s-r}^2(\alpha)\} \\ &= \lim_{n \rightarrow \infty} P\{C(\hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n,o)}) > \chi_{s-r}^2(\alpha)\} \\ &= \alpha \end{aligned}$$

where $\chi_{\nu}^2(\alpha)$ is the upper α -point of a χ^2 distribution with ν degrees of freedom.

The proof of Proposition 5 uses Proposition 4 and a Taylor series argument and is given in Appendix D.

If the true $\underline{\mu}$ is not in the null hypothesis, then one would like both the likelihood-ratio statistic and Pearson chi-square statistic

to have large power, i.e., a large probability of rejecting the null hypothesis. For a sequence of incorrect null hypotheses, there are two cases of interest. One is when the true $\mu^{(n)}$ and null hypotheses $\mu^{(n,o)}$ are converging to different limiting values μ^* and $\mu^{(*,o)}$, respectively. Proposition 6 shows that in this case both tests are consistent. In the second case, the true $\mu^{(n)}$ and null hypotheses $\mu^{(n,o)}$ are converging to the same limiting value μ^* . Proposition 7 shows that if the rate of convergence is chosen properly, then both test statistics will converge in distribution to a noncentral chi-square distribution.

Proposition 6: Let $\tilde{x}^{(n)}$ be a sequence of contingency tables with expected cell counts satisfying (8)-(12), and with space of fixed margins $\mathcal{M}$. Suppose that a sequence of null hypotheses $H_0 : \mu^{(n)} = \mu^{(n,o)} \in \mathcal{M}$ is given such that

$$\lim(\mu^{(n,o)} - (\log n)\varepsilon) = \mu^{(*,o)}$$

and

$$\mu^{(*,o)} \neq \mu^*$$

where μ^* is defined by (12). Then

$$\lim_n P\{-2\Delta(\tilde{x}^{(n)}, \hat{\mu}^{(n)}, \mu^{(n,o)}) > \chi_{s-r}^2(\alpha)\} = 1$$

and

$$\lim_n P\{C(\hat{\mu}^{(n)}, \mu^{(n,o)}) > \chi_{s-r}^2(\alpha)\} = 1$$

where $\hat{\mu}^{(n)}$ is the maximum likelihood estimate of $\mu \in \mathcal{M}$ based on

data $\underline{x}^{(n)}$. That is, both the likelihood-ratio test and the Pearson chi-square test are consistent.

The proof is given in Appendix D.

Proposition 7: Let $\underline{x}^{(n)}$ be a sequence of contingency tables with expected cell counts satisfying (8)-(12), and with space of fixed margins $\mathcal{M}$. Suppose that a sequence of null hypotheses $H_0: \underline{\mu}^{(n)} = \underline{\mu}^{(n,0)} \in \mathcal{M}$ is given such that

$$\lim_n n^{1/2}(\underline{\mu}^{(n)} - \underline{\mu}^{(n,0)}) = \underline{\zeta}^*$$

where $\underline{\mu}^{(n)}$ is defined by (9). Then

$$\begin{aligned} & \lim_n P\{-2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n,0)}) > \chi_{s-r}^2(\alpha)\} \\ &= \lim_n P\{C(\hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n,0)}) > \chi_{s-r}^2(\alpha)\} \\ &= P\{X_{s-r, \delta^2}^2 > \chi_{s-r}^2(\alpha)\} \end{aligned}$$

where X_{s-r, δ^2}^2 has a noncentral chi-square distribution with $s-r$ degrees of freedom and noncentrality parameter δ^2 given by

$$(15) \quad \delta^2 = \underline{\zeta}^{*'} D(\underline{\lambda}^*) Q' D^{-1}(\underline{m}^*) Q D(\underline{\lambda}^*) \underline{\zeta}^*$$

and where $\hat{\underline{\mu}}^{(n)}$ is the maximum likelihood estimate of $\underline{\mu} \in \mathcal{M}$ based on data $\underline{x}^{(n)}$.

The proof is given in Appendix D.

Remark: The limiting power is known as the Pitman asymptotic power.

With no classification error the noncentrality parameter is given by

$$\delta^2 = \xi^{*'} D(\lambda^*) \xi^*$$

which is derived by substituting an identity matrix for Q in expression (15).

In Proposition 7 it is seen that the Pitman asymptotic power depends on the direction ξ^* in which the null hypotheses approach the true $\mu^{(n)}$, the dimension s of the alternative space $\mathcal{M}$, the dimension r of the space of fixed margins $\mathcal{N}$, the limiting table of expected cell counts λ^* , and the error matrix Q . Since the larger the non-centrality parameter, the greater the asymptotic power of the test, the following proposition shows that the power will always be reduced in the presence of misclassification.

Proposition 8: For all $\xi \in \mathbb{R}^T$,

$$\xi' D(\lambda) Q' D^{-1}(\underline{m}) Q D(\lambda) \xi \leq \xi' D(\lambda) \xi$$

for any λ and $\underline{m} = Q\lambda$.

The proof uses Cauchy's inequality and is given in Appendix D.

For testing the composite null hypothesis

$$H_0 : \mu \in \mathcal{M}_1$$

against the alternative hypothesis

$$H_A : \mu \in \mathcal{M}_2$$

where $\mathcal{M}_1 \subseteq \mathcal{M}_2$, one rejects for large values of the generalized likelihood-ratio statistic

$$-2\Delta(\tilde{x}, \hat{\mu}^{(2)}, \hat{\mu}^{(1)}) = -2 \sum_{i=1}^T x_i \log \frac{\hat{m}_i^{(1)}}{\hat{m}_i^{(2)}}$$

or the Pearson chi-square statistic

$$c(\hat{\mu}^{(2)}, \hat{\mu}^{(2)}) = \sum_{i=1}^T \frac{(\hat{m}_i^{(2)} - \hat{m}_1^{(i)})^2}{\hat{m}_1^{(i)}},$$

where $\hat{\mu}^{(i)}$ is the maximum likelihood estimate of μ under the model $\mu \in \mathcal{M}_i$, $i = 1, 2$. To compute the asymptotic distribution of these statistics, let $\tilde{x}^{(n)}$ again represent a sequence of contingency tables satisfying (8)-(12). The following proposition computes this asymptotic distribution under a sequence of true null hypotheses. Let s_i be the dimension of $\mathcal{M}_i$, $i = 1, 2$, and recall that $\mathcal{N}$ is the space of fixed margins.

Proposition 9: Let $\tilde{x}^{(n)}$ be a sequence of contingency tables with expected cell counts satisfying (8)-(12), and with space of fixed margins $\mathcal{N}$. Suppose that a sequence of null hypotheses $H_0 : \mu^{(n)} \in \mathcal{M}_1$ and alternative $H_A : \mu^{(n)} \in \mathcal{M}_2$ is given such that $\mathcal{N} \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_2$. If for all n

$$\mu^{(n)} \in \mathcal{M}_1,$$

where $\mu^{(n)}$ is defined by (9), then the generalized likelihood-ratio statistic and Pearson chi-square statistic are asymptotically equivalent, and

$$\begin{aligned}
& \lim_n P\{-2\Delta(\hat{x}^{(n)}, \hat{\mu}^{(n,2)}, \hat{\mu}^{(n,1)}) > \chi^2_{s_2-s_1}(\alpha)\} \\
&= \lim_n P\{C(\hat{\mu}^{(n,2)}, \hat{\mu}^{(n,1)}) > \chi^2_{s_2-s_1}(\alpha)\} \\
&= \alpha
\end{aligned}$$

where $\hat{\mu}^{(n,i)}$ is the maximum likelihood estimate of $\mu \in \mathcal{M}_i$ based on data $\hat{x}^{(n)}$, $i = 1, 2$.

The proof is given in Appendix D.

If the true $\mu^{(n)}$ is not in the null hypothesis, there are again two cases of interest. Proposition 10 shows that when the true $\mu^{(n)}$ are converging to a point in $\mathcal{M}_2$ but outside $\mathcal{M}_1$, then both the generalized likelihood-ratio test and the Pearson chi-square tests are consistent. In Proposition 11, the true $\mu^{(n)}$ are taken outside $\mathcal{M}_1$ but converge to a point in $\mathcal{M}_1$ as $n \rightarrow \infty$. If the rate of convergence is chosen properly, then both statistics will converge in distribution to a noncentral chi-square distribution.

Proposition 10: Let $\hat{x}^{(n)}$ be a sequence of contingency tables with expected cell counts satisfying (8)-(12), and with space of fixed margins $\mathcal{N}$. Suppose that a sequence of null hypotheses $H_0 : \mu^{(n)} \in \mathcal{M}_1$ and alternatives $H_A : \mu^{(n)} \in \mathcal{M}_2$ is given such that $\mathcal{N} \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_2$. If

$$\mu^* \in \mathcal{M}_2, \text{ but } \mu^* \notin \mathcal{M}_1$$

where μ^* is defined by (11), then

$$\begin{aligned}
& \lim_n P\{-2\Delta(\tilde{x}^{(n)}, \hat{\mu}^{(n,2)}, \hat{\mu}^{(n,1)}) > \chi^2_{s_2-s_1}(\alpha)\} \\
&= \lim_n P\{C(\hat{\mu}^{(n,2)}, \hat{\mu}^{(n,1)}) > \chi^2_{s_2-s_1}(\alpha)\} \\
&= 1
\end{aligned}$$

where $\hat{\mu}^{(n,i)}$ is the maximum likelihood estimate of $\mu \in \mathcal{M}_i$ based on data $\tilde{x}^{(n)}$, $i = 1, 2$.

The proof is given in Appendix D.

Proposition 11: Let $\tilde{x}^{(n)}$ be a sequence of contingency tables with expected cell counts satisfying (8)-(12), and with space of fixed margins $\mathcal{N}$. Suppose that a sequence of null hypotheses $H_0 : \mu^{(n)} \in \mathcal{M}_1$ and alternatives $H_A : \mu^{(n)} \in \mathcal{M}_2$ is given such that $\mathcal{N} \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_2$. If for all n

$$\begin{aligned}
& \mu^{(n)} \in \mathcal{M}_2, \text{ but } \mu^{(n)} \notin \mathcal{M}_1, \\
& \mu^{(n)} - (\log n)\xi - n^{-1/2}\xi^{(n)} \in \mathcal{M}_1,
\end{aligned}$$

and

$$\lim_n \xi^{(n)} = \xi^*$$

where $\mu^{(n)}$ is defined by (9), then

$$\begin{aligned}
& \lim_n P\{-2\Delta(\tilde{x}^{(n)}, \hat{\mu}^{(n,2)}, \hat{\mu}^{(n,1)}) > \chi^2_{s_2-s_1}(\alpha)\} \\
&= \lim_n P\{C(\hat{\mu}^{(n,2)}, \hat{\mu}^{(n,1)}) > \chi^2_{s_2-s_1}(\alpha)\} \\
&= P\{\chi^2_{s_2-s_1, \delta^2} > \chi^2_{s_2-s_1}(\alpha)\}
\end{aligned}$$

where $\hat{\mu}^{(n,i)}$ is the maximum likelihood estimate of $\mu \in \mathcal{M}_i$ based on data $\underline{x}^{(n)}$, $i = 1, 2$. The noncentrality parameter δ^2 is given by

$$(17) \quad \delta^2 = \|\underline{c}^* - \mathcal{P}_{\mathcal{M}_1}(-d^2 \ell_{\mu}^*(\underline{m}^*)) \underline{c}^*\|_{(2)}^2$$

where $\mathcal{P}_{\mathcal{M}_1}(-d^2 \ell_{\mu}^*(\underline{m}^*)) \underline{c}^*$ represents the projection of $\underline{c}^*$ onto $\mathcal{M}_1$, and $\|\underline{z}\|_{(2)}$ represents the norm of $\underline{z}$, both taken with respect to the inner product given by $-d^2 \ell_{\mu}^*(\underline{m}^*)$, viz:

$$\begin{aligned} ((\underline{x}, \underline{z})) &= \underline{x}' (-d^2 \ell_{\mu}^*(\underline{m}^*)) \underline{z} \\ &= \underline{y}' D(\underline{\lambda}^*) Q' D^{-1}(\underline{m}^*) Q D(\underline{\lambda}^*) \underline{z}. \end{aligned}$$

The proof is given in Appendix D.

Remark: With no classification error the noncentrality parameter is given by

$$\delta^2 = \|\underline{c}^* - \mathcal{P}_{\mathcal{M}_1}(D(\underline{\lambda}^*)) \underline{c}^*\|_{(1)}^2$$

where $\mathcal{P}_{\mathcal{M}_1}(D(\underline{\lambda}^*)) \underline{c}^*$ represents the projection of $\underline{c}^*$ onto $\mathcal{M}_1$, and $\|\underline{z}\|_{(1)}$ represents the norm of $\underline{z}$, both taken with respect to the inner product given by $D(\underline{\lambda}^*)$. Propositions 5, 6, 7, 9, 10, and 11 are well-known in the no-error case (Haberman [1974a]).

The Pitman asymptotic power is seen to depend on the limiting value of the expected cell counts $\underline{\lambda}^*$, the null hypothesis model $\mathcal{M}_1$, the direction $\underline{c}^*$ in which the true $\mu^{(n)}$ approach the null hypothesis,

and the alternative hypothesis $\mathcal{M}_2$ since $\xi^* \in \mathcal{M}_2$ and s_2 equals the dimension of $\mathcal{M}_2$. In any event, the following proposition shows that classification error reduces Pitman asymptotic power.

Proposition 12: The asymptotic power of the generalized likelihood ratio test and Pearson chi-square test between alternative models is reduced in the presence of misclassification. That is, for all ξ

$$\|\xi - P_{\mathcal{M}_1}(-d^2_{\xi}(\mu^*))\xi\|_{(2)}^2 \leq \|\xi - P_{\mathcal{M}_1}(D(\lambda^*))\xi\|_{(1)}^2$$

where the projections and norms are defined in Proposition 11 and its following remark.

The proof uses Proposition 8 and the Pythagorean theorem, and is given in Appendix D.

Remark: When ξ is restricted to be perpendicular to $\mathcal{M}_1$ with respect to both the inner products given by $D(\lambda^*)$ and $-d^2_{\xi}(\mu^*)$, then Proposition 12 reduces to Proposition 11. In this case, Mote and Anderson [1965] showed this inequality of noncentrality parameters for testing independence in an $I \times J$ table with classification error.

Remark: There is a situation when there can be an increase in asymptotic power due to the presence of classification error. In Proposition 10, the $\mu^{(n)}$ are constrained to lie in $\mathcal{M}_2$, i.e., the alternative model contains the true $\mu^{(n)}$. If the true $\mu^{(n)}$ are outside the alternative model, then Propositions 10 and 11 as given here do not apply, and there can be an increase in power with classification error. An example is given at the end of Appendix E.

To compute the projections used in Proposition 11 and Proposition 12, it is useful to note that (Haberman [1974a])

$$(18) \quad P_{\mathcal{M}_1}(A)\xi = M_1(M_1^*AM_1)^{-1}M_1^*A\xi$$

where $\mathcal{M}_1$ is spanned by the columns of the matrix M_1 , and the projection is taken with respect to the inner product given by A . For models $\mathcal{M}_1$ which have closed-form maximum likelihood estimates of the expected cell counts, simpler expressions for the noncentrality parameter can frequently be derived (see Appendix E for some examples).

The ratio of the noncentrality parameters with and without classification error given in Proposition 12 is easily seen to be the asymptotic ratio of sample sizes necessary to achieve the same power with classification error as without. Assakul and Proctor [1967] give some examples of this ratio for testing independence in an $I \times J$ table.

For a $2 \times 2 \times 2$ table, Figure 1 shows the asymptotic ratio of sample sizes necessary to achieve the same power for testing the null hypothesis of no second order interaction ($u_{123} = 0$), against the alternative of the fully saturated model. The left half of Figure 1 refers to the classification error assumed on the table; the right half gives the ratio of the sample sizes. Since there are only two levels in each dimension of the table, the false positive and negative rates for each dimension completely describe the misclassification. To compute the ratio of noncentrality parameters, one usually must specify the direction ξ^* in which the true $\mu^{(n)}$ are approaching the null hypothesis, and the limiting table of expected cell counts λ^* . Since, in this case, the models of the null and alternative hypothesis differ by only one

FIGURE 1

ASYMPTOTIC RATIO OF SAMPLE SIZES: $2 \times 2 \times 2$ TABLETesting H_0 : No Second Order Interaction ($u_{123} = 0$)vs. H_A : Fully Saturated Model

False	Classification Errors						λ^* Completely Independent With $\%/\%+$ in Each Dimension =		
	Dim 1		Dim 2		Dim 3		50%/50%	20%/80%	80%/20%
	+	-	+	-	+	-			
	0	0	0	0	0	0	1.00	1.00	1.00
	.05	0	0	0	0	0	1.11	1.07	1.26
	.1	0	0	0	0	0	1.22	1.14	1.56
	.2	0	0	0	0	0	1.50	1.31	2.25
	.05	.05	0	0	0	0	1.23	1.37	1.37
	.1	.1	0	0	0	0	1.56	1.88	1.88
	.2	.2	0	0	0	0	2.78	3.78	3.78
	.05	0	.05	0	.05	0	1.35	1.21	2.02
	.1	0	.1	0	.1	0	1.83	1.48	3.76
	.2	0	.2	0	.2	0	3.38	2.26	11.39
	.05	.05	.05	.05	.05	.05	1.88	2.55	2.55
	.1	.1	.1	.1	.1	.1	3.81	6.63	6.63
	.2	.2	.2	.2	.2	.2	21.43	53.91	53.91

dimension, it will be shown in Appendix E that the ratio does not depend on ξ^* .

In general, λ^* can be any table in the null hypothesis. For simplicity, it is assumed in Figure 1 that λ^* is a completely independent table with the percent positive in each dimension of the table given by 50%, 80%, or 20%. For example, in the third line of Figure 1, the classification error consists entirely of a false positive rate of .1 across dimension 1 of the table. If the limiting table is completely independent with 50% positives in each dimension of the table, then the asymptotic ratio of sample sizes is seen to be 1.11. That is, 11% more observations are required with classification error to get the same asymptotic power as without. If the limiting table is completely independent with 80% positives in each dimension, then only 7% more observations would be required to achieve the same power. On the other hand, if the limiting table has only 20% positives, then 26% more observations would be required. This difference in asymptotic ratios, because of the difference in the limiting tables, corresponds with the notion that a false positive rate is more serious when there are less overall positives (cf. Section 2, Chapter 3). One can see that the ratios in Figure 1 become quite large as one inverses the classification error. Since λ^* is taken to be completely independent, it will be shown in Appendix E that the ratios in Figure 1 are given by the simple formula:

$$(19) \quad \underline{d}' D^{-1} (\underline{\lambda}^*) \underline{d} / (Q \underline{d})' D^{-1} (Q \underline{\lambda}^*) (Q \underline{d})$$

where

$$\underline{d}' = (1 \ -1 \ -1 \ 1 \ -1 \ 11 \ -1)$$

and

$$Q = Q_1 \otimes Q_2 \otimes Q_3$$

is the matrix containing the misclassification probabilities.

For a $2 \times 2 \times 2$ table, Figure 2 shows the asymptotic ratio of sample sizes necessary to achieve the same power for testing the null hypothesis of complete independence against the alternative model where dimensions 1 and 2 together are independent of dimension 3 ($u_{123} = u_{13} = u_{23} = 0$). Again, since there is a one dimensional difference between the two models, the ratio does not depend on ξ^* . The ratios in Figure 2 appear similar to the ratios in Figure 1. However, now it is seen in the last line of Figure 2 that there is no reduction of asymptotic power when there is classification error in dimension 3 alone. In fact, no matter what the classification error in dimensions 1 and 2, there will be no additional loss of power when misclassification is added across dimension 3. Heuristically, this is because both the null and alternative models specify that dimension 3 of the table is independent of dimensions 1 and 2 together. More formally, it is derived from the following simple formula given in Appendix E for the asymptotic ratio of sample sizes shown in Figure 2, viz:

$$(20) \quad \underline{d}' D(\underline{\lambda}^*) \underline{d} / (Q \underline{d})' D^{-1}(Q \underline{\lambda}^*) (Q \underline{d})$$

where

$$\underline{d}' = (\gamma, 1-\gamma, -\gamma, -(1-\gamma), -\gamma, -(1-\gamma), \gamma, (1-\gamma))$$

and γ is the proportion of negatives in dimension 3 of the table $\underline{\lambda}^*$, and Q is the matrix containing the misclassification probabilities.

FIGURE 2

ASYMPTOTIC RATIO OF SAMPLE SIZES: $2 \times 2 \times 2$ TABLETesting H_0 : Complete Independencevs. H_A : Dim(1,2) Indep. of Dim 3 ($u_{123} = u_{13} = u_{23} = 0$)

False	Classification Errors						λ^* Completely Independent with %-%+ in Each Dimension =		
	Dim 1		Dim 2		Dim 3		50%/50%	20%/80%	80%/20%
	+	-	+	-	+	-			
	0	0	0	0	0	0	1.00	1.00	1.00
	.05	0	0	0	0	0	1.11	1.07	1.26
	.1	0	0	0	0	0	1.22	1.14	1.56
	.2	0	0	0	0	0	1.50	1.31	2.25
	.05	.05	0	0	0	0	1.23	1.37	1.37
	.1	.1	0	0	0	0	1.56	1.88	1.88
	.2	.2	0	0	0	0	2.78	3.78	3.78
	.05	0	.05	0	.05	0	1.22	1.14	1.60
	.1	0	.1	0	.1	0	1.49	1.30	2.42
	.2	0	.2	0	.2	0	2.25	1.72	5.06
	.05	.05	.05	.05	.05	.05	1.52	1.87	1.87
	.1	.1	.1	.1	.1	.1	2.44	3.53	3.53
	.2	.2	.2	.2	.2	.2	7.71	14.27	14.27
	0	0	0	0	Any	Any	1.00	1.00	1.00

In this chapter it has been shown how to take into account a specified error matrix Q to compute the maximum likelihood estimates of the expected cell counts of a table under different hierarchical log-linear models. The generalized likelihood-ratio statistic or Pearson chi-square statistic for testing between alternative models can be calculated using the maximum likelihood estimates computed under these alternative models. When the models being tested are preserved by classification error, these tests are precisely the usual no-error generalized likelihood-ratio test and Pearson chi-square test completely ignoring the classification error. This is because in this case, the with-error maximum likelihood estimates of the expected cell counts are precisely the no-error maximum likelihood estimates of the expected cell counts (Section 1). In any event, it is seen that the increase in sample size in these tests necessary to compensate for the loss of (asymptotic) power due to misclassification can be substantial.

APPENDIX A:

Proofs of the Effects of Misclassification on the u Terms of Hierarchical Log-Linear Models (Chapter 3)

Proof of Proposition 1: Since the error matrix $Q = Q_1 \otimes Q_2$, it is sufficient to show the result separately for classification error only in dimension ℓ , $\ell = 1, 2$. The argument is completely symmetric in the two dimensions, so we show the result assuming there is classification error in dimension 1 alone. Then

$$\begin{aligned} \exp(4u_{12}(\tau)) &= \frac{\tau(11)\tau(22)}{\tau(12)\tau(21)} \\ &= \frac{\alpha\pi(11) + (1-\alpha)\pi(21)}{\beta\pi(11) + (1-\beta)\pi(21)} \cdot \frac{\beta\pi(12) + (1-\beta)\pi(22)}{\alpha\pi(12) + (1-\alpha)\pi(22)} \\ &= g(\alpha)/g(\beta) \end{aligned}$$

for some $\alpha, \beta \in [0,1]$, where g is defined by

$$g(\gamma) = [\gamma\pi(11) + (1-\gamma)\pi(21)] / [\gamma\pi(12) + (1-\gamma)\pi(22)] .$$

Taking the derivative of $g(\gamma)$ with respect to γ , one finds

$$g'(\gamma) \geq 0 \text{ if and only if } u_{12}(\tau) \geq 0 .$$

Therefore, if $u_{12}(\tau) \geq 0$, then

$$\begin{aligned} \exp(-4u_{12}(\tau)) &= \frac{g(0)}{g(1)} \leq \exp(4u_{12}(\tau)) = \frac{g(\alpha)}{g(\beta)} \leq \frac{g(1)}{g(0)} \\ &= \exp(4u_{12}(\tau)) . \end{aligned}$$

That is, if $u_{12}(\tau) \geq 0$, then

$$|u_{12}(\tau)| \leq u_{12}(\tau) .$$

Similarly, if $u_{12}(\pi) \leq 0$, then

$$|u_{12}(\tau)| \leq -u_{12}(\pi) .$$

Putting these together yields

$$|u_{12}(\tau)| \leq |u_{12}(\pi)| .$$

Q.E.D.

Proof of Proposition 2: Let

$$\pi = \begin{array}{cc|c} \pi(11) & \pi(12) & \pi(1+) \\ \hline \pi(21) & \pi(22) & \pi(2+) \\ \hline \pi(+1) & \pi(+2) & \end{array}$$

and let

$$\delta^{(n)} = \pi^{(n)}(11) - \pi(11) .$$

Since the $\pi^{(n)}$ have the same margins as π , $u_{12}(\pi^{(n)})$ and $u_{12}(\tau^{(n)})$ are functions of n through only $\delta^{(n)}$, say $g(\delta^{(n)})$ and $h(\delta^{(n)})$, respectively. Then

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{u_{12}(\tau^{(n)})}{u_{12}(\pi^{(n)})} &= \lim_{n \rightarrow \infty} \frac{h(\delta^{(n)})}{g(\delta^{(n)})} \\ &= \lim_{\delta \rightarrow 0} \frac{h(\delta)}{g(\delta)} \end{aligned}$$

and when the last limit is evaluated using L'Hospital's rule, the desired result is obtained.

Q.E.D.

Proof of Proposition 3:

The idea of the proof is quite simple. If the table λ belongs

to a given model, then to prove $Q\lambda$ belongs to that same model it will be shown that the u terms associated with that model can be assigned values which correspond to $Q\lambda$. This will be done by exhibiting a set of linear equations that can be solved to get values for these u terms. For the converse, it will be shown that this set of linear equations cannot, in general, be solved.

First some notation: Let θ represent a generalized index. For example, if $\theta = (13)$, then $u_{\theta}(i_1 i_2 \dots i_K) = u_{13}(i_1 i_3)$. The notation $\theta_1 \subseteq \theta_2$ will mean the numbers appearing in θ_1 are a subset of those appearing in θ_2 . For a log-linear model $\mathcal{M}$, let

$$\mathcal{S} = \{\theta | u_{\theta} \text{ is present in the model } \mathcal{M}\},$$

that is, the set of main effects and interactions present in the model.

The constraints on the u terms can be written:

(1) For any $\theta \in \mathcal{S}$, for any $h \in \theta$,

$$\sum_{i_h} u_{\theta}(i_1 i_2 \dots i_K) = 0.$$

If $\log \lambda \in \mathcal{M}$, then there exists $\{u_{\theta}^{(\lambda)} | \theta \in \mathcal{S}\}$ satisfying (1) such that

$$\log \lambda(i_1 i_2 \dots i_K) = \sum_{\theta \in \mathcal{S}} u_{\theta}^{(\lambda)}(i_1 i_2 \dots i_K)$$

for all cells $(i_1 i_2 \dots i_K)$.

Without loss of generality, let the classification error in dimension ℓ of the table described in Proposition 3 be in dimension 1 of the table. If $\underline{m} = Q\lambda$, then the model will be preserved if and only if

there exists $\{u_{\theta}^{(m)} | \theta \in \mathcal{J}\}$ satisfying (1) such that

$$\log m(i_1 i_2 \dots i_K) = \sum_{\theta \in \mathcal{J}} u_{\theta}^{(m)}(i_1 i_2 \dots i_K)$$

for all cells $(i_1 i_2 \dots i_K)$.

To see when this is true, let θ_0 be the union of all indices in $\mathcal{J}$ containing a 1, i.e.,

$$\theta_0 = \bigcup \{\theta \in \mathcal{J} | 1 \in \theta\}.$$

Let

$\mathcal{J}_1$ = all lower-order relatives of θ_0 which are in $\mathcal{J}$.

$\mathcal{J}_2 = \{\theta \in \mathcal{J} | \theta \notin \mathcal{J}_1\}$.

Now,

$$\begin{aligned} m(i_1 i_2 \dots i_K) &= \sum_{\ell_1} q_{i_1 \ell_1}^1 \lambda(\ell_1 i_2 \dots i_K) \\ &= \sum_{\ell_1} q_{i_1 \ell_1}^1 \exp\left[\sum_{\theta \in \mathcal{J}} u_{\theta}^{(\lambda)}(\ell_1 i_2 \dots i_K)\right] \\ (2) \quad &= \exp\left[\sum_{\theta \in \mathcal{J}_2} u_{\theta}^{(\lambda)}(i_1 i_2 \dots i_K)\right] \\ &\quad \cdot \sum_{\ell_1} q_{i_1 \ell_1}^1 \exp\left[\sum_{\theta \in \mathcal{J}_1} u_{\theta}^{(\lambda)}(\ell_1 i_2 \dots i_K)\right] \end{aligned}$$

since $u_{\theta}^{(\lambda)}(\ell_1 i_2 \dots i_K)$ is not a function of ℓ_1 for all $\theta \in \mathcal{J}_2$.

< IF > Suppose

- (3) u_1 is present in the model, and all u terms present in the model containing a 1 as a subscript are lower-order relatives of a single u term present in the model.

Let $u_{\theta}^{(m)} = u_{\theta}^{(\lambda)}$ for all $\theta \in \mathcal{J}_2$, so that

$$(4) \quad m(i_1 i_2 \dots i_K) = \exp \left[\sum_{\theta \in \mathcal{J}_2} u_{\theta}^{(m)}(i_1 i_2 \dots i_K) \right] \\ \cdot \sum_{\ell_1} q_{i_1 \ell_1}^1 \exp \left[\sum_{\theta \in \mathcal{J}_1} u_{\theta}^{(\lambda)}(\ell_1 i_2 \dots i_K) \right].$$

Consider the linear equations in the unknown $\{u_{\theta}^{(m)} | \theta \in \mathcal{J}_1\}$:

$$(5) \quad \sum_{\theta \in \mathcal{J}_1} u_{\theta}^{(m)}(i_1 i_2 \dots i_K) = \log \left(\sum_{\ell_1} q_{i_1 \ell_1}^1 \exp \left[\sum_{\theta \in \mathcal{J}_1} u_{\theta}^{(\lambda)}(\ell_1 i_2 \dots i_K) \right] \right) \\ \text{for all } (i_1 i_2 \dots i_K).$$

The model will be preserved if these equations can be solved for $\{u_{\theta}^{(m)} | \theta \in \mathcal{J}_1\}$ satisfying (1).

Since the model is of the special form (3), $\mathcal{J}_1$ consists of an index $\theta_0 = (1h_1 \dots h_s)$ and all its lower-order relatives. The equations (5) can be solved because these are precisely the equations to fit a fully saturated model to an $I_1 \times I_{h_1} \times \dots \times I_{h_s}$ contingency table: The equations (5) consist of $I_1 I_{h_1} \dots I_{h_s}$ distinct equations since only $(i_1, i_{h_1} \dots i_{h_s})$ appear as arguments for $\{u_{\theta} | \theta \in \mathcal{J}_1\}$. The $\{u_{\theta}^{(m)} | \theta \in \mathcal{J}_1\}$ appearing in (5) can be rewritten in terms of $I_1 I_{h_1} \dots I_{h_s}$ unconstrained variables using the linear constraints given in (1). Since there are the same number of linear equations as unknowns, the solutions for these unconstrained variables can be found

and used to solve back for the $\{u_\theta^{(m)} | \theta \in \mathcal{S}_1\}$. Therefore the model is preserved.

It will be useful later to have a slightly stronger result than just proved. An examination of the proof reveals that the only property of the matrix $Q_1 = ((q_{i_1 i_2}^1))$ used was that the arguments of the logarithms in (5) are positive. So, in fact, it has just been proved that if the model is of the special form (3), then:

$$\log \lambda \in \mathcal{M} \text{ and } A_\lambda \text{ has all positive entries implies} \\ \text{that } \log(A_\lambda) \in \mathcal{M}$$

for all $A = A_1 \otimes A_2 \otimes \dots \otimes A_K$ where A_i is an identity matrix for $i \neq 1$, and A_1 is arbitrary.

< ONLY IF > Assume the model is not of the special form (3) and is preserved by classification error. It will be shown that a contradiction is reached. First, if u_1 is not in the model, then it is clear that classification error in dimension 1 can make u_1 non-zero. Therefore, the model is not preserved. In what follows, it is assumed that u_1 is in the model.

It will now be shown that if Q_1 is column stochastic, and λ and $\underline{m}$ are in the model, then $u_\theta^{(m)} = u_\theta^{(\lambda)}$ for all $\theta \in \mathcal{S}_2$: Every index in $\mathcal{S}_2$ must contain a number that appears in no index of $\mathcal{S}_1$. If $\mathcal{A}$ is the set of these numbers, i.e.,

$$\mathcal{A} = \{\ell | \exists \theta \in \mathcal{S}_2, \ell \in \theta \text{ and } \ell \notin \phi \text{ for all } \phi \in \mathcal{S}_1\}$$

then

$$u_{1\ell} = 0 \text{ for all } \ell \in \mathcal{A}.$$

By a collapsability theorem (Bishop, Fienberg, and Holland [1975]), the u term involving $\mathcal{A}$, which are $\{u_\theta | \theta \in \mathcal{S}_2\}$, are the same whether based on the original table or the table collapsed over dimension 1. The collapsed table for λ is the same as the collapsed table for $\underline{m} = Q\lambda$ since Q_1 is column stochastic. Since λ and $\underline{m}$ are both in the model, the $u_\theta^{(\lambda)}$ and $u_\theta^{(\underline{m})}$, for $\theta \in \mathcal{S}_2$, can both be based on the same collapsed table. Therefore $u_\theta^{(\underline{m})} = u_\theta^{(\lambda)}$ for all $\theta \in \mathcal{S}_2$.

For Q_1 column stochastic, equations (4) are therefore true. So, $\underline{m}$ will be in the model if and only if equations (5) can be solved for $\{u_\theta^{(\underline{m})} | \theta \in \mathcal{S}_1\}$. Since the definition of model-preserving would require $\underline{m}$ to be in the model in particular for all Q_1 column stochastic, there will be a contradiction when it is shown that the equations (5) cannot be solved: Since the model is not of the special form (3), θ_o will not be in $\mathcal{S}$. Let $\theta_o = (1h_1 \dots h_s)$. The equations (5) still consist of $I_1 \cdot I_{h_1} \dots I_{h_s}$ distinct equations, but now the $\{u_\theta^{(\underline{m})} | \theta \in \mathcal{S}_1\}$ appearing in (5) can be rewritten in terms of b unconstrained variables using the linear constraints 1). Here b is a number strictly less than $I_1 \cdot I_{h_1} \cdot \dots \cdot I_{h_s}$ --it would equal $I_1 \cdot I_{h_1} \cdot \dots \cdot I_{h_s}$ if $\theta_o \in \mathcal{S}_1$. The right-hand side of the equations (5) can be thought of as a b dimensional manifold in Euclidean space of dimension $I_1 \cdot I_{h_1} \cdot \dots \cdot I_{h_s}$ as λ ranges over possible values in the model. This b dimensional manifold is not a linear manifold provided Q_1 is not an identity matrix. Therefore, (5) cannot in general be solved for $\{u_\theta^{(\underline{m})} | \theta \in \mathcal{S}_1\}$ satisfying (1), and a contradiction is reached.

Q.E.D.

APPENDIX B:

Proofs of Finite-Sample Results in Chapter 4

Product multinomial sampling schemes: If the sampling scheme fixes the first L -way margin of the $I_1 \times I_2 \times \dots \times I_K$ contingency table, then there exists fixed numbers $\{N(i_1 i_2 \dots i_L)\}$ such that

$$\{x(i_1 i_2 \dots i_L \dots i_K) | \text{all } (i_{L+1} \dots i_K)\}$$

has a simple multinomial distribution with total sample size $N(i_1 i_2 \dots i_L)$, for all $(i_1 i_2 \dots i_L)$ (cf. Chapter 2, Section 3). Let the T -vectors

$\{v^{(i_1 i_2 \dots i_L)} | \text{all } (i_1 i_2 \dots i_L)\}$ be defined so that

$$v^{(i_1 i_2 \dots i_L)}(j_1 \dots j_L \dots j_K) = \begin{cases} 1 & \text{if } (j_1 j_2 \dots j_L) = (i_1 i_2 \dots i_L) \\ 0 & \text{otherwise} \end{cases}$$

where as usual the T cells of the table are considered in lexicographical order. If one further considers the superscripts of the v 's to be in lexicographical order, then

$$\{v^1, v^2, \dots, v^r\} = \{v^{(i_1 i_2 \dots i_L)} | \text{all } (i_1 i_2 \dots i_L)\}$$

where $r = I_1 \cdot I_2 \cdot \dots \cdot I_L$. The space of fixed margins $\mathcal{N}$ will be defined to be

$$\mathcal{N} = \langle v^1, v^2, \dots, v^r \rangle,$$

the linear space spanned by the vectors $\{v^1, v^2 \dots v^r\}$. For simple multinomial sampling, one takes $r = 1$ and $v^1 = \underline{e}$, the vector of all ones, and $\mathcal{N} = \langle \underline{e} \rangle$. For Poisson sampling, one defines $\mathcal{N} = \langle \underline{0} \rangle$.

Recall (Chapter 3, Section 3) that allowable hierarchical log-linear models with the sampling scheme that fixes the first L -way margin of the table must have $u_{12\dots L}$ present. In the notation here, the log-linear model $\mathcal{M}$ is allowable with the sampling scheme $\mathcal{N}$ if $\mathcal{N} \subseteq \mathcal{M}$ (Haberman [1974a]).

Lemma: Let $\{v^1, v^2, \dots, v^r\} = \{v^{(i_1 i_2 \dots i_L)}\}$ as defined above represent a product multinomial sampling scheme. Let $Q = Q_1 \times Q_2 \times \dots \times Q_L \times \dots \times Q_K$ be an error matrix with no error across a margin being held fixed by the sampling, i.e., Q_i is an identity matrix for $i = 1, 2, \dots, L$. Then

$$QD(\lambda)v^i = D(\underline{m})v^i \quad \text{for } i = 1, 2, \dots, r$$

where $\underline{m} = Q\lambda$.

$$\begin{aligned} \text{Pf:} \quad & [QD(\lambda)v^{(i_1^0 i_2^0 \dots i_L^0)}](i_1 i_2 \dots i_K) \\ &= \sum_{j_{L+1} \dots j_K} q_{i_{L+1} j_{L+1}}^{L+1} \dots q_{i_K j_K}^K \lambda(i_1 \dots i_L j_{L+1} \dots j_K) \\ & \quad \cdot v^{(i_1^0 i_2^0 \dots i_L^0)}(i_1 i_2 \dots i_L j_{L+1} \dots j_K) \\ &= \begin{cases} \sum_{j_{L+1} \dots j_K} q_{i_{L+1} j_{L+1}}^{L+1} \dots q_{i_K j_K}^K \lambda(i_1 \dots i_L j_{L+1} \dots j_K) & \text{if } (i_1 \dots i_L) = (i_1^0 \dots i_L^0) \\ 0 & \text{otherwise} \end{cases} \\ &= (Q\lambda)(i_1 \dots i_K)v^{(i_1^0 \dots i_L^0)}(i_1 \dots i_K) \\ &= [D(\underline{m})v^{(i_1^0 \dots i_L^0)}](i_1 \dots i_K). \end{aligned}$$

Q.E.D.

Proof of Proposition 1: It is sufficient (Birch [1963]) to show that the solution $\hat{m}$ to the maximum likelihood equations assuming Poisson sampling on $\underline{x}$, viz:

$$\mathcal{P}_{\mathcal{M}} D(\hat{\lambda}) Q' D^{-1}(\hat{m}) \underline{x} = \mathcal{P}_{\mathcal{M}} \hat{\lambda} ,$$

satisfies the multinomial constraints

$$\begin{aligned} \hat{m}(i_1 \dots i_L + \dots +) &= N(i_1 \dots i_L) \\ &= x(i_1 i_2 \dots i_L + \dots +) . \end{aligned}$$

Since $\mathcal{N} \subseteq \mathcal{M}$, the maximum likelihood equations imply

$$\mathcal{P}_{\mathcal{N}} D(\hat{\lambda}) Q' D^{-1}(\hat{m}) \underline{x} = \mathcal{P}_{\mathcal{N}} \hat{\lambda} .$$

Since the $\underline{v}$'s form an orthogonal basis for $\mathcal{N}$, this implies

$$(\underline{v}^i)' D(\hat{\lambda}) Q' D^{-1}(\hat{m}) \underline{x} = (\underline{v}^i)' \hat{\lambda} \quad \text{for } i = 1, \dots, r .$$

By the lemma this yields

$$(\underline{v}^i)' \underline{x} = (\underline{v}^i)' \hat{\lambda} = (\underline{v}^i)' \hat{m}$$

where the last equality holds because Q has no error across a fixed margin.

Q.E.D.

Proof of Proposition 2: Let

$$\mathcal{G}_T = \{ \underline{y} \in \mathbb{R}^T \mid y_i > 0 \text{ all } i \} ,$$

$$f(\underline{\pi}) = \sum_{i=1}^T (x_i \log \pi_i - \pi_i) ,$$

and

$$\mathcal{B} = \exp(\mathcal{M}) \equiv \{\exp(\underline{z}) \mid \underline{z} \in \mathcal{M}\}.$$

Would like to show if $\tilde{\underline{m}} \in \mathcal{B}$ achieves the maximum

$$\max_{\underline{m} \in \mathcal{B}} f(\underline{m})$$

and $Q^{-1} \tilde{\underline{m}} \in \mathcal{O}_T$, then $\tilde{\underline{m}}$ also achieves the maximum of

$$\max_{\underline{m} \in Q(\mathcal{B})} f(\underline{m}).$$

By the strict concavity of $f(\underline{m})$ it is sufficient to show

$$Q(\mathcal{B}) = \mathcal{B} \cap Q(\mathcal{O}_T).$$

By the proof of Proposition 3, Chapter 3, given in Appendix A, one has $\log \underline{\lambda} \in \mathcal{M}$ and $A \underline{\lambda} \in \mathcal{O}_T$ implies that

$$\log(A \underline{\lambda}) \in \mathcal{M}$$

for all $A = A_1 \otimes \cdots \otimes A_K$ where A_i is an identity matrix for $i = J+1, \dots, K$. Letting $A = Q^{-1}$ yields

$$\mathcal{B} \cap Q(\mathcal{O}_T) \subseteq Q(\mathcal{B}).$$

But

$$Q(\mathcal{B}) \subseteq \mathcal{B} \cap Q(\mathcal{O}_T)$$

so

$$Q(\mathcal{B}) = \mathcal{B} \cap Q(\mathcal{O}_T).$$

Q.E.D.

APPENDIX C:

Algorithms for Finding the Maximum Likelihood Estimates of the Expected Cell Counts (Chapter 4)

Let $\underline{x}$ be the observed table, Q the known classification error matrix, and $\mathcal{M}$ the log-linear model to be fit.

Algorithm 1

- (a) Begin with an initial estimate $\lambda^{(0)}$ of λ , the without-error expected cell counts.
- (b) Calculate a new table of "corrected" cell counts $D(\lambda^{(0)})Q'D^{-1}(Q\lambda^{(0)})\underline{x}$. Using the model $\mathcal{M}$, compute the standard log-linear maximum likelihood estimate of λ , assuming no classification error, based on the "corrected" cell counts.
- (c) Iterate step (b) with the current estimate of λ replacing $\lambda^{(0)}$.

Thus one obtains a sequence of estimators $\lambda^{(i)}$ satisfying

$$(1) \quad P_{\mathcal{M}D(\lambda^{(i)})Q'D^{-1}(Q\lambda^{(i)})\underline{x}} = P_{\mathcal{M}\lambda^{(i+1)}}.$$

If these estimates converge, they will converge to a solution of the maximum likelihood equations since both sides of (1) are continuous functions of λ .

Implementing Algorithm 1 requires doing a standard log-linear model estimation at step (b), that is solving

$$(2) \quad P_{\mathcal{M}\underline{y}} = P_{\mathcal{M}\lambda}$$

for λ , where $\underline{y}$ is the corrected cell counts at that step. Depending

on the model $\mathcal{M}$, closed-form estimates may exist for λ , otherwise a numerical method must be used. For the initial estimate $\lambda^{(0)}$, I recommend doing step (b) with the "corrected" cell counts given by $Q^{-1}x$. That is, let $\tilde{x} = Q^{-1}x$ and let $\lambda^{(0)}$ be the solution to (2). This again requires finding a standard (no-error) log-linear model estimate.

Remark: This algorithm is a special case of one given in Haberman [1977] and Dempster, Laird, and Rubin [1977]; some convergence properties are discussed there. Other more "efficient" algorithms for finding the maximum of the likelihood function may exist for this problem that, for example, make use of the second derivatives of the log likelihood (Haberman [1977]).

Example 1: Consider the following observed $2 \times 2 \times 2$ table:

Table 1

		2	3		
	1	10	20	30	30
		20	40	40	50

The log-linear model to be fit is that of dimensions 1 and 2 being conditionally independent given dimension 3 (Example 4 of Section 3, Chapter 3). Assume there is classification error in dimension 3 of the table only, known to be

$$Q_3 = \begin{pmatrix} .9 & .1 \\ .1 & .9 \end{pmatrix}.$$

In the no-error case, this model has the following closed-form expression

for the maximum likelihood estimate of the expected cell counts:

$$\lambda(i_1 i_2 i_3) = \frac{\lambda(i_1 + i_3) \lambda(+ i_2 i_3)}{\lambda(++ i_3)}.$$

Algorithm 1 is implemented for this model and data in Figure 1. The eight cells of the table $\{(i_1 i_2 i_3)\}$ are laid out across the page.

At each iteration the corrected cell counts (called x_i) are printed out, along with $\lambda^{(i)}$ (called L_i) and $\mu^{(i)} = Q\lambda^{(i)}$. The log likelihood (actually $\ell(x, \mu^{(i)})$) is also printed out at each iteration.

The initial estimate is computed at iteration 0, from the "corrected" data $Q^{-1}x$. We see that the convergence of the $\lambda^{(i)}$ are quite rapid.

In applying Algorithm 1 to a hierarchical log-linear model that does not have a closed-form expression for the maximum likelihood estimates, one can use iterative proportional fitting to solve for λ in step (b). Iterative proportional fitting starts with an initial estimate of the expected cell counts that is in the model. It then forces this estimate to match certain margins of the observed table in sequence to get a new estimate of the expected cell counts. This new estimate is forced to match the same certain margins, and the procedure is iterated. The particular log linear model being fit determines which margins are matched. Usually the initial estimate of all ones is taken for convenience to start the procedure, although any estimate in the model will work. See Bishop, Fienberg, and Holland [1975] for a complete description of iterative proportional fitting.

Two possible improvements are available for Algorithm 1 when iterative proportional fitting is necessary to get the no-error maximum likelihood estimate of the expected cell counts in step (b). The first

FIGURE 1

CELL		111	121	211	221	112	122	212	222
Q ⁻¹	X 0	10.000	20.000	20.000	40.000	30.000	30.000	40.000	50.000
	X 0	7.500	18.750	17.500	38.750	32.500	31.250	42.500	51.250
	L 0	7.954	18.295	17.045	39.204	30.357	33.392	44.642	49.107
	M 0	10.194	19.805	19.805	40.194	28.116	31.883	41.883	48.116
	LOG LIKELIHOOD = 837.43643737								
	X 1	7.871	18.349	17.119	39.186	32.128	31.650	42.880	50.813
	L 1	7.940	18.280	17.050	39.255	30.380	33.399	44.629	49.064
	M 1	10.184	19.792	19.808	40.236	28.136	31.887	41.871	48.083
	LOG LIKELIHOOD = 837.43655385								
	X 2	7.863	18.344	17.122	39.204	32.136	31.655	42.877	50.795
	L 2	7.934	18.274	17.052	39.275	30.389	33.402	44.624	49.048
	M 2	10.179	19.786	19.809	40.252	28.143	31.889	41.867	48.071
	LOG LIKELIHOOD = 837.43657092								
	X 3	7.860	18.342	17.123	39.211	32.139	31.657	42.876	50.788
	L 3	7.931	18.271	17.052	39.282	30.393	33.403	44.622	49.042
	M 3	10.177	19.784	19.809	40.258	28.146	31.890	41.865	48.066
	LOG LIKELIHOOD = 837.43657345								
	X 4	7.859	18.342	17.124	39.213	32.140	31.657	42.875	50.786
	L 4	7.930	18.270	17.052	39.285	30.394	33.404	44.622	49.040
	M 4	10.177	19.783	19.809	40.260	28.148	31.890	41.865	48.064
	LOG LIKELIHOOD = 837.43657383								
	X 5	7.858	18.341	17.124	39.214	32.141	31.658	42.875	50.785
	L 5	7.930	18.270	17.052	39.286	30.395	33.404	44.621	49.039
	M 5	10.176	19.783	19.809	40.261	28.148	31.890	41.865	48.064
	LOG LIKELIHOOD = 837.43657389								
	X 6	7.858	18.341	17.124	39.214	32.141	31.658	42.875	50.785
	L 6	7.930	18.270	17.052	39.286	30.395	33.404	44.621	49.039
	M 6	10.176	19.783	19.809	40.261	28.148	31.890	41.864	48.063
	LOG LIKELIHOOD = 837.43657390								

is to use the previous estimate of λ in step (b) as the initial estimate to start the iterative proportional fitting rather than the table of all ones. The second is to do only one round of iterative proportional fitting in step (b) rather than actually finding the no-error maximum likelihood estimate to some specified precision. It seems wasteful to spend a lot of time estimating λ precisely in step (b), when the corrected data is going to be changed quite a bit in the next iteration of step (b). These two changes lead to Algorithm 2.

Algorithm 2

- (a) Begin with an initial estimate $\lambda^{(0)}$ of λ , the without-error expected cell counts.
- (b) Calculate a new table of "corrected" cell counts $D(\lambda^{(0)})Q'D^{-1}(Q\lambda^{(0)})x$. Using the model $\mathcal{M}$, do one round of iterative proportional fitting using the "corrected" cell counts as the observed data, $\lambda^{(0)}$ as the initial estimate of the expected cell counts, and solving for λ , the new estimate of the expected cell counts.
- (c) Iterate step (b) with the current estimate of λ replacing $\lambda^{(0)}$.

For the initial estimate $\lambda^{(0)}$, I recommend computing the standard (no-error) maximum likelihood estimate of λ based on the "corrected" cell counts $Q^{-1}x$. This will require iterative proportional fitting.

Example 2: Consider the $2 \times 2 \times 2$ table of observed cell counts and error structure both as given in Example 1 of this section. The model to be fit now is that of no second order interaction (Example 5 of Section 3, Chapter 3). In the no-error case, this model does not have

a closed-form expression for the maximum likelihood estimate of the expected cell counts. One round of iterative proportional fitting consists of going from $\lambda^{(i-1)}$ to $\lambda^{(i)}$ via the following steps (Bishop, Fienberg, and Holland [1975]):

$$(i) \quad \text{let } \lambda^{(i-1,0)} = \lambda^{(i-1)}$$

$$(ii) \quad \text{let } \lambda^{(i-1,1)}(j_1 j_2 j_3) = \lambda^{(i-1,0)}(j_1 j_2 j_3) \frac{x^{(i)}(j_1 j_2 +)}{\lambda^{(i-1,0)}(j_1 j_2 +)}$$

$$\text{for } j_1, j_2, j_3 = 1, 2$$

$$(iii) \quad \text{let } \lambda^{(i-1,2)}(j_1 j_2 j_3) = \lambda^{(i-1,1)}(j_1 j_2 j_3) \frac{x^{(i)}(j_1 + j_3)}{\lambda^{(i-1,1)}(j_1 + j_3)}$$

$$\text{for } j_1, j_2, j_3 = 1, 2$$

$$(iv) \quad \text{let } \lambda^{(i-1,3)}(j_1 j_2 j_3) = \lambda^{(i-1,2)}(j_1 j_2 j_3) \frac{x^{(i)}(+j_2 j_3)}{\lambda^{(i-1,2)}(+j_2 j_3)}$$

$$\text{for } j_1, j_2, j_3 = 1, 2$$

$$(v) \quad \text{let } \lambda^{(i)} = \lambda^{(i-1,3)}.$$

This round is done for each iteration of step (b) of Algorithm 2. In Figure 2, Algorithm 2 is implemented for this model and data. The layout is similar to Figure 1. At each iteration, the "corrected" cell counts (called x_i), $\lambda^{(i)}$ (called L_i), and $\mu^{(i)} = Q\lambda^{(i)}$ are printed out along with the log likelihood (actually $\ell(x, \mu^{(i)})$). The initial estimate $\lambda^{(0)}$ (called L_0) required 4 rounds of iterative proportional fitting to get 3 decimal places accuracy. We see again that the convergence of the $\lambda^{(i)}$ is quite rapid.

FIGURE 2

CELL		111	121	211	221	112	122	212	222
Q ⁻¹	X 0	10.000	20.000	20.000	40.000	30.000	30.000	40.000	50.000
	X 0	7.500	18.750	17.500	38.750	32.500	31.250	42.500	51.250
	L 0	8.596	17.653	16.237	40.012	32.162	31.587	43.032	50.717
	L 0	8.441	17.808	16.555	39.694	31.569	32.180	43.434	50.315
	L 0	8.439	17.810	16.560	39.689	31.560	32.189	43.439	50.310
	L 0	8.439	17.810	16.560	39.689	31.560	32.189	43.439	50.310
	M 0	10.751	19.248	19.248	40.751	29.248	30.751	40.751	49.248
	LOG LIKELIHOOD =		837.54417208						
X 1	7.930	18.392	17.112	39.091	32.069	31.607	42.887	50.908	
L 1	8.474	17.849	16.569	39.633	31.512	32.164	43.443	50.353	
M 1	10.777	19.280	19.257	40.705	29.208	30.732	40.755	49.281	
LOG LIKELIHOOD =		837.54447274							
X 2	7.946	18.405	17.114	39.073	32.053	31.594	42.885	50.926	
L 2	8.491	17.860	16.570	39.617	31.503	32.144	43.435	50.377	
M 2	10.792	19.289	19.256	40.693	29.201	30.716	40.748	49.301	
LOG LIKELIHOOD =		837.54453140							
X 3	7.953	18.411	17.115	39.065	32.046	31.588	42.884	50.934	
L 3	8.498	17.866	16.570	39.611	31.498	32.136	43.431	50.386	
M 3	10.798	19.293	19.256	40.688	29.198	30.709	40.745	49.309	
LOG LIKELIHOOD =		837.54454157							
X 4	7.956	18.414	17.115	39.063	32.043	31.585	42.884	50.936	
L 4	8.501	17.868	16.570	39.608	31.497	32.132	43.430	50.390	
M 4	10.801	19.294	19.256	40.686	29.197	30.706	40.744	49.312	
LOG LIKELIHOOD =		837.54454334							
X 5	7.957	18.415	17.116	39.062	32.042	31.584	42.883	50.937	
L 5	8.503	17.869	16.570	39.607	31.496	32.130	43.429	50.391	
M 5	10.802	19.295	19.256	40.686	29.197	30.704	40.743	49.313	
LOG LIKELIHOOD =		837.54454365							
X 6	7.958	18.415	17.116	39.061	32.041	31.584	42.883	50.938	
L 6	8.503	17.869	16.570	39.607	31.496	32.130	43.429	50.392	
M 6	10.802	19.295	19.256	40.685	29.196	30.704	40.743	49.314	
LOG LIKELIHOOD =		837.54454370							
X 7	7.958	18.415	17.116	39.061	32.041	31.584	42.883	50.938	
L 7	8.503	17.870	16.570	39.607	31.496	32.129	43.429	50.392	
M 7	10.803	19.296	19.256	40.685	29.196	30.703	40.743	49.314	
LOG LIKELIHOOD =		837.54454371							
X 8	7.958	18.415	17.116	39.061	32.041	31.584	42.883	50.938	
L 8	8.503	17.870	16.570	39.607	31.496	32.129	43.429	50.392	
M 8	10.803	19.296	19.256	40.685	29.196	30.703	40.743	49.314	
LOG LIKELIHOOD =		837.54454371							

APPENDIX D:

Proofs of Asymptotic Distributions of Maximum Likelihood Estimates and Test Statistics (Chapter 4)

Heuristic proof: The reason the usual no-error theorems do not apply when there is classification error is that as the u terms run over their possible values, the log of the expected cell counts, $\log \underline{m}$, is not falling in a necessarily linear manifold. Let $\mathcal{B}$ be the (possibly) non-linear manifold containing $\log \underline{m}$. Recall

$$\lim_n \log \frac{\underline{m}^{(n)}}{n} = \log \underline{m}^* \in \mathcal{B}.$$

Where the expected cell counts go, the maximum likelihood estimates cannot be far behind, so as n gets large, both $\log(\underline{m}^{(n)}/n)$ and $\log(\hat{\underline{m}}^{(n)}/n)$ are falling with high probability in a decreasingly small neighborhood of $\log \underline{m}^*$ on $\mathcal{B}$. Any smooth non-linear manifold looks linear as one confines attention to a smaller and smaller neighborhood around a fixed point. In particular, the linear space that passes through $\log \underline{m}^*$ and is tangent to $\mathcal{B}$ at $\log \underline{m}^*$ is given by

$$\underline{m}^* = D^{-1}(\underline{m}^*)_{QD(\underline{\lambda}^*)} \underline{m}$$

where $\underline{m}$ is the log-linear model being considered (Section 1, Chapter 3). Substituting this linearized problem for the actual problem and applying the no-error log-linear model theorems will yield the propositions involving asymptotic results in Chapter 4.

Unfortunately, to make the above heuristic arguments precise requires as much work as proving the results from scratch, which is done here.

The proofs are similar to the no-error case as given by Theorems 4.1, 4.3, 4.4, 4.5, 4.6, 4.7, and 4.8 of Haberman [1974a] corresponding here to Propositions 4, 5, 6, 7, 9, 10, and 11, respectively. Arguments which are identical to those given there will be omitted.

Proof of Proposition 4: As in the no-error case,

$$(1) \quad \frac{1}{n} \tilde{x}^{(n)} \xrightarrow{P} \tilde{m}^*$$

where the P over the arrow stands for convergence in probability.

Also,

$$(2) \quad n^{-1/2}(\tilde{x}^{(n)} - \tilde{m}^{(n)}) \xrightarrow{D} \mathcal{N}(0, D(\tilde{m}^*)[I - P_{\mathcal{L}}(D(\tilde{m}^*))])$$

where for any linear space $\mathcal{L}$ and positive definite matrix A , $P_{\mathcal{L}}(A)$ is the projection onto $\mathcal{L}$ orthogonal with respect to the inner product given by A , viz:

$$((x, y)) = x' A y.$$

If $\mathcal{L}$ is spanned by the columns of the matrix L , then (Haberman [1974a])

$$(3) \quad P_{\mathcal{L}}(A) = L(L'AL)^{-1}L'A.$$

The idea of the proof, as in the no-error case, is to use the implicit function theorem to define a function F which gives the mle of μ when the data is sufficiently close to m^* . The chain rule evaluates the derivatives of F in terms of derivatives of the log likelihood. The mean value theorem can be used to express $\hat{\mu}^{(n)} - \mu^{(n)}$ in terms of the derivatives of F and $\tilde{x}^{(n)} - \tilde{m}^{(n)}$. This with (2)

will yield the asymptotic distribution of $\hat{\mu}^{(n)} - \mu^{(n)}$.

Recall the vector of derivatives of the log likelihood with respect to μ :

$$[d\ell_{\mu}(\tilde{x})] = D(\tilde{\lambda})Q'D^{-1}(\tilde{m})\tilde{x} - \tilde{\lambda}.$$

Let $[d^1\ell_{\mu}(\tilde{x})]$ be the matrix of partial derivatives of $[d\ell_{\mu}(\tilde{x})]$ with respect to the first variable (x) :

$$\begin{aligned} [d^1\ell_{\mu}(\tilde{x})] &\equiv \left(\left(\frac{\partial [d\ell_{\mu}(\tilde{x})]_i}{\partial x_j} \right) \right)_{i,j=1,\dots,T} \\ &= D(\tilde{\lambda})Q'D^{-1}(\tilde{m}). \end{aligned}$$

Let $[d^2\ell_{\mu}(\tilde{x})]$ be the matrix of partial derivatives of $[d\ell_{\mu}(\tilde{x})]$ with respect to the second variable (μ) :

$$\begin{aligned} [d^2\ell_{\mu}(\tilde{x})] &\equiv \left(\left(\frac{\partial [d\ell_{\mu}(\tilde{x})]_i}{\partial \mu_j} \right) \right)_{i,j=1,\dots,T} \\ &= D([d\ell_{\mu}(\tilde{x})]) - \\ &\quad D(\tilde{\lambda})Q'D^{-1}(\tilde{m})D(\tilde{x})D^{-1}(\tilde{m})QD(\tilde{\lambda}). \end{aligned}$$

Since

$$[d\ell_{\mu^*}(\tilde{m}^*)] = 0$$

and

$$-[d^2\ell_{\mu^*}(\tilde{m}^*)]$$

is positive definite, the implicit function theorem can be applied to

$[d\ell_\mu(x)]$ as a function on $\mathbb{R}^T \times \mathcal{M}$ at $(\underline{m}^*, \underline{\mu}^*)$. There exists open balls $A \subseteq \mathbb{R}^T$ and $B \subseteq \mathcal{M}$ such that $\underline{m}^* \in A$ and $\underline{\mu}^* \in B$ with the following property: For each $\underline{x} \in A$ there is a unique $F(\underline{x}) \in B$ such that

$$(4) \quad \underline{v}'[d\ell_{F(\underline{x})}(\underline{x})] = 0 \quad \text{for all } \underline{v} \in \mathcal{M}.$$

That is, $F(\underline{x})$ is a maximum likelihood estimate of $\underline{\mu}$ given data $\underline{x}$. Taking the partial derivatives of (4) with respect to $\underline{x}$ and using the chain rule yields:

$$(5) \quad \underline{v}'[-d^2\ell_{F(\underline{x})}(\underline{x})][dF_{\underline{x}}] = \underline{v}'[d^1\ell_{F(\underline{x})}(\underline{x})]$$

for all $\underline{x} \in A, \underline{v} \in \mathcal{M}$,

where $[dF_{\underline{x}}]$ is the matrix of partial derivatives of $F(\underline{x})$ with respect to $\underline{x}$,

$$[dF_{\underline{x}}] = \left(\left(\frac{\partial F_i(\underline{x})}{\partial x_j} \right) \right)_{i,j=1,\dots,T}.$$

A linear algebra argument yields from (5):

$$[dF_{\underline{x}}] = \mathcal{P}_{\mathcal{M}}(-d^2\ell_{F(\underline{x})}(\underline{x}))[-d^2\ell_{F(\underline{x})}(\underline{x})]^{-1}[d^1\ell_{F(\underline{x})}(\underline{x})].$$

In particular,

$$(6) \quad [dF_{\underline{m}^*}] = \mathcal{P}_{\mathcal{M}}(-d^2\ell_{\underline{\mu}^*}(\underline{m}^*))[-d^2\ell_{\underline{\mu}^*}(\underline{m}^*)]^{-1}[d^1\ell_{\underline{\mu}^*}(\underline{m}^*)].$$

Let $\hat{\underline{\mu}}^{(n)} = F(\underline{x}^{(n)})$. The mean value theorem shows that if $\underline{x}^{(n)}/n$ and $\underline{m}^{(n)}/n$ are in A , then

$$(\hat{\mu}^{(n)} - \mu^{(n)})_i = ([dF_{\tilde{z}}^{(n,i)}][\frac{1}{n} \tilde{x}^{(n)} - \frac{1}{n} \tilde{m}^{(n)}])_i$$

for $i = 1, 2, \dots, T$

where $\tilde{z}^{(n,i)}$ is on the line segment joining $\tilde{x}^{(n)}/n$ and $\tilde{m}^{(n)}/n$.

As $n \rightarrow \infty$,

$$\tilde{z}^{(n,i)} \xrightarrow{P} \tilde{m}^* \text{ for } i = 1, \dots, T$$

by (1). Using (2), as $n \rightarrow \infty$

$$(7) \quad n^{1/2}(\hat{\mu}^{(n)} - \mu^{(n)}) - [dF_{\tilde{m}}^*](n^{-1/2}(\tilde{x}^{(n)} - \tilde{m}^{(n)})) \xrightarrow{P} 0.$$

(This argument is slightly incorrect in Haberman [1974a].) Applying (2) once more yields as $n \rightarrow \infty$

$$(8) \quad n^{1/2}(\hat{\mu}^{(n)} - \mu^{(n)}) \xrightarrow{D} \eta(0, \Sigma)$$

where

$$(9) \quad \Sigma = [dF_{\tilde{m}}^*] D(\tilde{m}^*) [I - P_{\eta}(D(\tilde{m}^*))] [dF_{\tilde{m}}^*]'$$

Using the symmetry of $P_{\eta}(-d^2 \ell_{\tilde{m}}^*(\tilde{m}^*))[-d^2 \ell_{\tilde{m}}^*(\tilde{m}^*)]^{-1}$, the lemma in Appendix B, and (3) shows that (9) reduces to the expression given for Σ_1 in Proposition 4(a).

A Taylor series argument shows

$$(10) \quad n^{-1/2}(\hat{m}^{(n)} - \tilde{m}^{(n)}) - Q[\frac{1}{n} D(\tilde{\lambda}^{(n)})][n^{1/2}(\hat{\mu}^{(n)} - \mu^{(n)})] \xrightarrow{P} 0.$$

Proposition (4)(b) follows immediately.

To prove Proposition 4(c), imbed M in a $T \times T$ fully saturated design matrix M_0 , that is, let

$$M_o = (M \begin{smallmatrix} \vdots \\ Z \end{smallmatrix})$$

and

$$\mathbb{R}^T = \{M_o \tilde{x} \mid \tilde{x} \in \mathbb{R}^T\}.$$

Letting

$$\tilde{v}^{(n)} = M_o^{-1} \hat{u}^{(n)}$$

it follows that

$$\tilde{v}^{(n)} = \begin{pmatrix} \hat{u}^{(n)} \\ 0 \end{pmatrix}.$$

As $n \rightarrow \infty$, Proposition 4(a) implies

$$(11) \quad n^{1/2}(\tilde{v}^{(n)} - M_o \mu^{(n)}) \xrightarrow{D} \eta(Q, M_o^{-1} \sum_1 (M_o^{-1})').$$

But

$$\begin{aligned} M_o^{-1} \sum_1 (M_o^{-1})' &= \begin{pmatrix} (M'D(\lambda^*)Q'D^{-1}(\underline{m}^*)QD(\lambda^*)M)^{-1} & \vdots & 0 \\ \cdots & \vdots & \cdots \\ 0 & \vdots & 0 \end{pmatrix} \begin{matrix} s \\ T-s \end{matrix} \\ &\quad \begin{matrix} s & T-s \end{matrix} \\ &- \begin{pmatrix} (N'D(\underline{m}^*)N)^{-1} & \vdots & 0 \\ \cdots & \vdots & \cdots \\ 0 & \vdots & 0 \end{pmatrix} \begin{matrix} r \\ T-r \end{matrix} \\ &\quad \begin{matrix} r & T-r \end{matrix} \end{aligned}$$

Restricting attention to the first s dimensions of (11) yields Proposition 4(c). Q.E.D.

Testing Results: The likelihood ratio Δ as given by expression (14) of Chapter 4 implicitly assumes that the total number of observations in the null hypothesis table, $m^{(o)}$, is the same as the total number of observations in the observed table, x . That is,

$$(12) \quad \sum_{i=1}^T m_i^{(o)} = \sum_{i=1}^T x_i.$$

In general, the log of the ratio of the likelihoods is proportional to

$$\Delta(x, \hat{\mu}, \mu_o) = \sum_{i=1}^T [x_i \log \frac{m_i^{(o)}}{\hat{m}_i} - (m_i^{(o)} - \hat{m}_i)].$$

In the proofs of the propositions that follow, it will be convenient to use this new function Δ rather than the old function Δ ; we continue to assume (12) so that they coincide when evaluated for a hypothesis test. We further assume that for product multinomial sampling that the null hypothesis table satisfies the same marginal constraints as the observed table.

Proof of Proposition 5: Considering $\Delta(x^{(n)}, \hat{\mu}^{(n)}, \mu^{(n)})$ as a function of $\mu^{(n)}$, a Taylor series expansion around $\hat{\mu}^{(n)}$ shows

$$(13) \quad -2\Delta(x^{(n)}, \hat{\mu}^{(n)}, \mu^{(n)}) = [n^{1/2}(\hat{\mu}^{(n)} - \mu^{(n)})]' \frac{1}{n} [-d^2 \ell_{\gamma^{(n)}}(x^{(n)})] \cdot [n^{1/2}(\hat{\mu}^{(n)} - \mu^{(n)})]$$

for some $\gamma^{(n)}$ on the line segment joining $\mu^{(n)}$ and $\hat{\mu}^{(n)}$. Using (10) shows the asymptotic equivalence of $-2\Delta(x^{(n)}, \hat{\mu}^{(n)}, \mu^{(n)})$ and $c(\hat{\mu}^{(n)}, \mu^{(n)})$.

Using (7) and (13) yields

$$(14) \quad -2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n)}) = \{[\underline{dF}_{*}^{\underline{\mu}}][n^{-1/2}(\underline{x}^{(n)} - \underline{m}^{(n)})]\}' \cdot$$

$$[-d^2\ell_{*}^{\underline{\mu}}(\underline{m}^{*})][[\underline{dF}_{*}^{\underline{\mu}}][n^{-1/2}(\underline{x}^{(n)} - \underline{m}^{(n)})]] \succeq 0.$$

Since $[-d^2\ell_{*}^{\underline{\mu}}(\underline{m}^{*})]$ is positive definite, there exists (Rao [1973])

an invertible matrix $[-d^2\ell_{*}^{\underline{\mu}}(\underline{m}^{*})]^{1/2}$ such that

$$(15) \quad [-d^2\ell_{*}^{\underline{\mu}}(\underline{m}^{*})] = ([-d^2\ell_{*}^{\underline{\mu}}(\underline{m}^{*})]^{1/2})' [-d^2\ell_{*}^{\underline{\mu}}(\underline{m}^{*})]^{1/2}.$$

Using (14), (15), and the definition of $[\underline{dF}_{*}^{\underline{\mu}}]$ in (6), one has

$$(16) \quad -2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n)}) = \underline{z}^{(n)'} A_{\underline{m}}([-d^2\ell_{*}^{\underline{\mu}}(\underline{m}^{*})]) \underline{z}^{(n)} \succeq 0$$

where

$$(17) \quad \underline{z}^{(n)} = ([-d^2\ell_{*}^{\underline{\mu}}(\underline{m}^{*})]^{-1/2})' [d^1\ell_{*}^{\underline{\mu}}(\underline{m}^{*})] n^{-1/2} (\underline{x}^{(n)} - \underline{m}^{(n)})$$

and for any linear space $\mathcal{L}$,

$$(18) \quad A_{\mathcal{L}}([-d^2\ell_{*}^{\underline{\mu}}(\underline{m}^{*})]) = [-d^2\ell_{*}^{\underline{\mu}}(\underline{m}^{*})]^{1/2} P_{\mathcal{L}}([-d^2\ell_{*}^{\underline{\mu}}(\underline{m}^{*})]) \cdot$$

$$[-d^2\ell_{*}^{\underline{\mu}}(\underline{m}^{*})]^{-1/2}$$

is the projection onto the linear space $[d^2\ell_{*}^{\underline{\mu}}(\underline{m}^{*})]^{1/2} \mathcal{L}$ orthogonal

with respect to the usual inner product on $\mathbb{R}^T$.

Using (2) and the lemma in Appendix B shows

$$(19) \quad \underline{z}^{(n)} \stackrel{D}{\rightarrow} \{I - A_{\mu} \eta([-d^2 \ell_{*}(\underline{m}^*)])\} \underline{z}$$

where $\underline{z}$ has a standard multivariate normal distribution. Combining (16) and (19) shows $-2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n)})$ converges in distribution to a chi-square random variable with $s-r$ degrees of freedom. Q.E.D.

Proof of Proposition 6: As in the no-error case (Haberman [1974a])

$$\frac{1}{n} \Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n,o)}) = \frac{1}{n} \Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n)}) + \frac{1}{n} \Delta(\underline{x}^{(n)}, \underline{\mu}^{(n)}, \underline{\mu}^{(n,o)}) .$$

The first term converges in probability to 0 by Proposition 5. The second term converges to

$$\Delta(\underline{m}^*, \underline{\mu}^*, \underline{\mu}^{(*,o)}) < 0 .$$

Also

$$\begin{aligned} \frac{1}{n} c(\hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n,o)}) &= \frac{1}{n} (\hat{\underline{m}}^{(n)} - \underline{m}^{(n)})' D^{-1}(\underline{m}^{(n,o)}) (\hat{\underline{m}}^{(n)} - \underline{m}^{(n)}) \\ &\quad + \frac{2}{n} (\hat{\underline{m}}^{(n)} - \underline{m}^{(n)})' D^{-1}(\underline{m}^{(n,o)}) (\underline{m}^{(n)} - \underline{m}^{(n,o)}) \\ &\quad + \frac{1}{n} c(\underline{\mu}^{(n)}, \underline{\mu}^{(n,o)}) . \end{aligned}$$

The first two terms converge in probability by Proposition 4; the last term converges to

$$c(\underline{\mu}^*, \underline{\mu}^{(*,o)}) > 0 . \quad \text{Q.E.D.}$$

Proof of Proposition 7: Considering $\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n,o)})$ as a function of $\underline{\mu}^{(n,o)}$, a Taylor series argument around $\hat{\underline{\mu}}^{(n)}$ shows

$$-2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n,o)}) = [n^{1/2}(\hat{\underline{\mu}}^{(n)} - \underline{\mu}^{(n,o)})]' \frac{1}{n} [-d^2 \ell_{\gamma^{(n)}(\underline{x}^{(n)})}] \cdot \quad (20)$$

$$[n^{1/2}(\hat{\underline{\mu}}^{(n)} - \underline{\mu}^{(n,o)})]$$

for some $\gamma^{(n)}$ on the line segment joining $\hat{\underline{\mu}}^{(n)}$ and $\underline{\mu}^{(n,o)}$. Using (20) and a Taylor series expansion similar to (10) shows the asymptotic equivalence of the two test statistics.

To find the asymptotic distribution of the test statistics, note that (7) and (20) imply

$$-2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n,o)}) - G' [-d^2 \ell_{\underline{\mu}^*}(\underline{m}^*)] G \xrightarrow{P} 0$$

where

$$G = \{[dF_{\underline{m}^*}][n^{-1/2}(\underline{x}^{(n)} - \underline{m}^{(n)})] + \underline{\xi}^{(n)}\}$$

and where

$$\underline{\xi}^{(n)} = n^{1/2}(\underline{\mu}^{(n)} - \underline{\mu}^{(n,o)}) .$$

Since $\underline{\xi}^* \in \mathcal{M}$, it follows that

$$[dF_{\underline{m}^*}] QD(\underline{\lambda}^*) \underline{\xi}^* = \underline{\xi}^*$$

so therefore

$$(21) \quad -2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n)}, \underline{\mu}^{(n,o)}) - (\underline{z}^{(n)} + \underline{\chi})' A_{\mathcal{M}}(-d^2 \ell_{\underline{\mu}^*}(\underline{m}^*)) (\underline{z}^{(n)} + \underline{\chi}) \xrightarrow{P} 0$$

where $\underline{z}^{(n)}$ and $A_{\mathcal{M}}(-d^2 \ell_{\underline{\mu}^*}(\underline{m}^*))$ are defined by (17), (18), respectively, and

$$\chi = [-d^2 \ell_{\mu}^{*}(\underline{m}^*)]^{1/2} \underline{c}^*$$

where $[-d^2 \ell_{\mu}^{*}(\underline{m}^*)]^{1/2}$ is defined by (15).

The sampling constraints on $\underline{m}^{(n)}$ and $\underline{m}^{(n,o)}$ force

$$A_{\mu}(-[d^2 \ell_{\mu}^{*}(\underline{m}^*)])\chi = \underline{0}.$$

This combined with (19) and (21) yields

$$-2\Delta(\underline{x}^{(n)}, \underline{\mu}^{(n)}, \underline{\mu}^{(n,o)}) \stackrel{D}{\rightarrow} (\underline{z} + \underline{\chi})' [A_{\mu}(-d^2 \ell_{\mu}^{*}(\underline{m}^*)) - A_{\mu}(-d^2 \ell_{\mu}^{*}(\underline{m}^*))](\underline{z} + \underline{\chi})$$

where $\underline{z}$ has a standard multivariate normal distribution. This implies that the two test statistics are distributed as noncentral chi-squares with $s-r$ degrees of freedom, and noncentrality parameter

$$\begin{aligned} \delta^2 &= \underline{\chi}' [A_{\mu}(-d^2 \ell_{\mu}^{*}(\underline{m}^*)) - A_{\mu}(-d^2 \ell_{\mu}^{*}(\underline{m}^*))] \underline{\chi} \\ &= \underline{\chi}' \underline{\chi} \\ &= \underline{c}^{*'} [-d^2 \ell_{\mu}^{*}(\underline{m}^*)] \underline{c}^*. \end{aligned}$$

Q.E.D.

Proof of Proposition 8: Let

$$N = \sum_{i=1}^T \lambda_i,$$

$$\pi = \frac{1}{N} \lambda, \text{ and}$$

$$\underline{d} = D(\pi) \underline{c}.$$

Then it is sufficient to show that for all $\underline{d} \in \mathbb{R}^T$,

$$\underline{d}' Q' D^{-1} (Q \underline{\pi}) Q \underline{d} \leq \underline{d}' D^{-1} (\underline{\pi}) \underline{d} ,$$

that is,

$$\sum_{i=1}^T \frac{(Q \underline{d})_i^2}{(Q \underline{\pi})_i} \leq \sum_{i=1}^T \frac{d_i^2}{\pi_i} .$$

Fix i and let

$$z(s) = \sqrt{q_{is}} \quad \text{and} \quad y(s) = \sqrt{q_{is}} \frac{d_s}{\pi_s} .$$

Taking expected values with respect to the probability distribution $\underline{\pi}$,

$$Ez^2 = \sum_s q_{is} \pi_s, \quad Ey^2 = \sum_s \frac{q_{is} d_s^2}{\pi_s}$$

and

$$Ezy = \sum_s q_{is} d_s .$$

Cauchy's inequality, $(Ezy)^2 / Ez^2 \leq Ey^2$, yields

$$\frac{(\sum_s q_{is} d_s)^2}{\sum_s q_{is} \pi_s} \leq \sum_s \frac{q_{is} d_s^2}{\pi_s} .$$

Summing both sides over i gives the conclusion since Q' is stochastic.

Q.E.D.

Proof of Proposition 9: Since

$$\Delta(\underline{x}^{(n)}, \hat{\mu}^{(n,2)}, \hat{\mu}^{(n,1)}) = \Delta(\underline{x}^{(n)}, \hat{\mu}^{(n,2)}, \mu^{(n)}) - \Delta(\underline{x}^{(n)}, \hat{\mu}^{(n,1)}, \mu^{(n)}) ,$$

by (16)

$$-2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n,2)}, \hat{\underline{\mu}}^{(n,1)}) - \underline{z}^{(n)*} [A_{\mathcal{M}_2}(-d^2_{\underline{\mu}^*}(\underline{m}^*)) - A_{\mathcal{M}_1}(-d^2_{\underline{\mu}^*}(\underline{m}^*))] \\ \cdot \underline{z}^{(n)} \xrightarrow{P} 0$$

where $\underline{z}^{(n)}$ is defined by (17). Using (19) shows the asymptotic distribution of $-2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n,2)}, \hat{\underline{\mu}}^{(n,1)})$ is chi-square with $s_2 - s_1$ degrees of freedom.

The two test statistics are shown to be asymptotically equivalent using (7) and (10). Q.E.D.

Proof of Proposition 10: As in the no-error case (Haberman [1974a]), the arguments used in Proposition 4 can be used to show

$$[\hat{\underline{\mu}}^{(n,1)} - (\log n)\underline{e}] \xrightarrow{P} \underline{y}^*$$

where $\underline{y}^*$ is the location of the maximum for $\underline{\mu} \in \mathcal{M}_1$ of $\ell(\underline{m}^*, \underline{\mu})$.

Furthermore,

$$\frac{1}{N} \Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n,2)}, \hat{\underline{\mu}}^{(n,1)}) = \frac{1}{n} \Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n,2)}, \underline{\mu}^{(n)}) \\ + \frac{1}{n} \Delta(\underline{x}^{(n)}, \underline{\mu}^{(n)}, \hat{\underline{\mu}}^{(n,1)}) .$$

The first term converges in probability to 0 since by Proposition 5, $-2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n,2)}, \underline{\mu}^{(n)})$ converges in distribution to a chi-square random variable with $s_2 - r$ degrees of freedom. The second term converges in probability to

$$\Delta(\underline{m}^*, \underline{\mu}^*, \underline{y}^*) < 0 .$$

An argument similar to the one used in the proof of Proposition 6 shows

$$\frac{1}{n} c(\hat{\mu}^{(n,2)}, \hat{\mu}^{(n,1)}) \xrightarrow{P} c(\mu^*, \nu^*) > 0. \quad \text{Q.E.D.}$$

Proof of Proposition 11: A slightly more general result than Proposition 11 will be proved here. Namely, if one does not require $\mu^{(n)} \in \mathcal{M}_2$, then the limiting distribution of the two test statistics will still be noncentral chi-square with $s_2 - s_1$ degrees of freedom, but with noncentrality parameter δ^2 given by

$$(22) \quad \delta^2 = \| [P_{\mathcal{M}_2}(-d^2 \ell_{\mu^*}(\tilde{m}^*)) - P_{\mathcal{M}_1}(-d^2 \ell_{\mu^*}(\tilde{m}^*))] \tilde{c}^* \|^2_{(1)}.$$

This corresponds to the case when both the null hypothesis and the alternative hypothesis are incorrect (Haberman [1974a]). When the $\mu^{(n)} \in \mathcal{M}_2$, then $\tilde{c}^* \in \mathcal{M}_2$ and the above expression reduces to the one given in Proposition 11, Chapter 4.

As similar to the no-error case (Haberman [1974a]), a Taylor series expansion shows

$$(23) \quad -2\Delta(\tilde{x}^{(n)}, \hat{\mu}^{(n,2)}, \hat{\mu}^{(n,1)}) = [n^{1/2}(\hat{\mu}^{(n,2)} - \hat{\mu}^{(n,1)})]^T \cdot \frac{1}{n} [d^2 \ell_{\tilde{\nu}^{(n)}}(\tilde{x}^{(n)})] [n^{1/2}(\hat{\mu}^{(n,2)} - \hat{\mu}^{(n,1)})]$$

for some $\tilde{\nu}^{(n)}$ on the line segment joining $\hat{\mu}^{(n,2)}$ and $\hat{\mu}^{(n,1)}$. It follows that as in the proof of Proposition 5,

$$(24) \quad n^{1/2}(\hat{\mu}^{(n,2)} - \hat{\mu}^{(n,1)}) - G \xrightarrow{P} 0$$

where

$$G = [P_{m_2}(-d^2 \ell_{\mu}^*(\underline{m}^*)) - P_{m_1}(-d^2 \ell_{\mu}^*(\underline{m}^*))][d^2 \ell_{\mu}^*(\underline{m}^*)]^{-1} \cdot \\ \cdot [d^2 \ell_{\mu}^*(\underline{m}^*)][n^{-1/2}(\underline{x}^{(n)} - \underline{m}^{(n)}) + QD(\underline{\lambda}^*)\underline{\xi}^{(n)}] .$$

Therefore

$$-2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n,2)}, \hat{\underline{\mu}}^{(n,1)}) - G^*[d^2 \ell_{\mu}^*(\underline{m}^*)]G \xrightarrow{P} 0 .$$

An argument similar to that used in the proof of Proposition 7 then shows that $-2\Delta(\underline{x}^{(n)}, \hat{\underline{\mu}}^{(n,2)}, \hat{\underline{\mu}}^{(n,1)})$ converges in distribution to a chi-square random variable with $s_2 - s_1$ degrees of freedom and noncentrality parameter given by (22). Using (23) and a Taylor series expansion similar to (10) shows the two test statistics are asymptotically equivalent. Q.E.D.

Proof of Proposition 12:

$$\|\underline{\xi} - P_{m_1}(d^2 \ell_{\mu}^*(\underline{m}^*))\underline{\xi}\|_{(2)}^2 \\ = \|[\underline{\xi} - P_{m_1}(D(\underline{\lambda}^*))\underline{\xi}] - P_{m_1}(d^2 \ell_{\mu}^*(\underline{m}^*))[\underline{\xi} - P_{m_1}(D(\underline{\lambda}^*))\underline{\xi}]\|_{(2)}^2 \\ = \|\underline{\xi} - P_{m_1}(D(\underline{\lambda}^*))\underline{\xi}\|_{(2)}^2 \\ - \|P_{m_1}(d^2 \ell_{\mu}^*(\underline{m}^*))[\underline{\xi} - P_{m_1}(D(\underline{\lambda}^*))\underline{\xi}]\|_{(2)}^2$$

by the Pythagorean theorem

$$\leq \|\underline{\xi} - P_{m_1}(D(\underline{\lambda}^*))\underline{\xi}\|_{(2)}^2 \\ \leq \|\underline{\xi} - P_{m_1}(D(\underline{\lambda}^*))\underline{\xi}\|_{(1)}^2 \quad \text{by Proposition 8.}$$

Q.E.D.

APPENDIX E:

Simpler Expressions for Noncentrality Parameters (Chapter 4)

For a particular log-linear model $\mathcal{M}_1$, the computation of the noncentrality parameter (17) of Chapter 4 for general λ^* and ξ^* by evaluating the projection using (18), could be rather tedious. Haberman [1974a] shows in the no-error case how one computes the noncentrality parameter as a limit of maximum likelihood estimates. This is simple to do when the models have closed-form expressions for the maximum likelihood estimate. The same approach can be used with classification error but will not be pursued here.

If the model in question is preserved by classification error, then testing the without-error expected cell counts to be in that model is equivalent to testing the with-error expected cell counts to be in that model (Section 3, Chapter 3). Therefore, if one has an expression for the noncentrality parameter when there is no classification error, say

$$\delta^2(\text{no error}) = g(\lambda^*, \xi^*, \mathcal{M}_1),$$

then the noncentrality parameter when there is classification error Q is given by

$$\delta^2(\text{error}) = g(\underline{m}^*, D^{-1}(\underline{m}^*) Q D(\lambda^*) \xi^*, \mathcal{M}_1)$$

where

$$\underline{m}^* = Q \lambda^*.$$

This is because if $\mu^{(n)}$ satisfies (8)-(12) of Chapter 4, and if

$$\mu^{(n)} - n^{-1/2} \xi^{(n)} \in \mathcal{M}_1, \text{ and}$$

$$\lim \xi^{(n)} = \xi^*,$$

then

$$\log \tilde{m}^{(n)} - n^{-1/2} \tilde{z}^{(n)} \in \mathcal{M}_1$$

where

$$\lim \tilde{z}^{(n)} = D^{-1}(\tilde{m}^*) QD(\lambda^*) \xi^*$$

(cf. heuristic proof in Appendix D).

Example: Testing complete independence in an $I_1 \times I_2 \times I_3$ table versus the fully saturated model. Without loss of generality, let $\lambda^*(+++)=1$. From Diamond [1958],

$$\delta^2(\text{no error}) = \sum_{i_1 i_2 i_3} \frac{1}{\lambda^*(i_1 i_2 i_3)} \cdot$$

$$[d(i_1 i_2 i_3) - \lambda(+i_2+) \lambda(++i_3) d(i_1++)$$

$$- \lambda(i_1++) \lambda(++i_3) d(+i_2+) - \lambda(i_1++) \lambda(+i_2+) d(++i_3)]^2$$

where

$$\underline{d} = D(\lambda^*) \xi^*.$$

Therefore,

$$\delta^2(\text{error}) = \sum_{i_1 i_2 i_3} \frac{1}{m^*(i_1 i_2 i_3)} \cdot [(Qd)(i_1 i_2 i_3) - m(+i_2+)m(++i_3)(Qd)(i_1++) - m(i_1++)m(++i_3)(Qd)(+i_2+) - m(i_1++)m(+i_2+)(Qd)(++i_3)]^2$$

where $\underline{d}$ is as given above. When $d(i_1++) = d(+i_2+) = d(++i_3) = 0$, the above expressions reduce to

$$\begin{aligned} \delta^2(\text{no error}) &= \underline{d}^* D^{-1} (\underline{\lambda}^*) \underline{d} = \underline{c}^* D (\underline{\lambda}^*) \underline{c}^* \\ \delta^2(\text{error}) &= (Q\underline{d})^* D^{-1} (\underline{m}^*) (Q\underline{d}) \\ &= \underline{c}^* [d^2 \ell_* (\underline{m}^*)] \underline{c}^* . \end{aligned}$$

Choosing $\underline{d}$ this way is in fact equivalent to choosing $\underline{c}^*$ perpendicular to $\mathcal{M}_1$ with respect to the inner products given by both $D(\underline{\lambda}^*)$ and $[d^2 \ell_* (\underline{m}^*)]$. So these expressions are also immediate from Proposition 11 and its following remark.

When the dimensions of the models $\mathcal{M}_1$ and $\mathcal{M}_2$ differ by only 1, then the ratio of noncentrality parameters with and without classification error does not depend on the direction $\underline{c}^* \in \mathcal{M}_2$:

Proposition: Let $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \mathbb{R}^n$ be linear spaces such that the dimension of $\mathcal{M}_2$ is one larger than the dimension of $\mathcal{M}_1$. Let A and B be two inner products on $\mathbb{R}^n$. Then there exists a constant K such that

$$\|\underline{c} - \mathcal{P}_{\mathcal{M}_1}^{(A)}(\underline{c})\|_{(A)}^2 = K \|\underline{c} - \mathcal{P}_{\mathcal{M}_1}^{(B)}(\underline{c})\|_{(B)}^2$$

for all $\underline{c} \in \mathcal{M}_2$.

Proof: Without loss of generality, let $\mathcal{M}_2 = \mathbb{R}^n$ and

$$\mathcal{M}_1 = \{(x_1, x_2, \dots, x_{n-1}, 0) | x_i \in \mathbb{R}\}.$$

Let

$$\tilde{x}^{(A)} = (x_1^{(A)}, x_2^{(A)}, \dots, x_n^{(A)}) \text{ and}$$

$$\tilde{x}^{(B)} = (x_1^{(B)}, x_2^{(B)}, \dots, x_n^{(B)})$$

be the unit normal vectors to $\mathcal{M}_1$ with respect to the inner products given by A and B, respectively. Let

$$K = (x_n^{(B)} / x_n^{(A)})^2. \quad \text{Q.E.D.}$$

Example: Testing no second order interaction versus a fully saturated model in a $2 \times 2 \times 2$ table.

Since the difference of the dimensions of the models is 1, one is free to choose ξ^* conveniently to get the ratio of the noncentrality parameters. Let

$$\xi^* = D^{-1}(\lambda^*)d$$

where

$$d' = (1 \ -1 \ -1 \ 1 \ -1 \ 1 \ 1 \ -1).$$

Then it is easy to see that ξ^* is perpendicular to $\mathcal{M}_1$ with respect to the inner product given by $D(\lambda^*)$, so that

$$\rho_{\mathcal{M}_1}(D(\lambda^*))\xi^* = 0.$$

It is also true that if λ^* is a completely independent table, then ξ^*

is perpendicular to $\mathcal{M}_1$ with respect to the inner product given by $[-d^2 \ell_{\mu}^*(\underline{m}^*)]$, so that

$$\mathcal{P}_{\mathcal{M}_1}([-d^2 \ell_{\mu}^*(\underline{m}^*)])c^* = 0.$$

The ratio of the noncentrality parameters is therefore given by

$$\|\underline{c}^*\|_{(1)}^2 / \|\underline{c}^*\|_{(2)}^2$$

where the norms $\|\cdot\|_{(1)}^2$ and $\|\cdot\|_{(2)}^2$ are given with respect to the inner products given by $D(\lambda^*)$ and $[-d^2 \ell_{\mu}^*(\underline{m}^*)]$, respectively. This is exactly expression (19) of Chapter 4.

Example: Testing complete independence versus the model of dimensions 1 and 2 together being independent of dimension 3.

Again, since the difference in the dimensions of the models is 1, one is free to choose c^* . Let

$$\underline{c}^* = D^{-1}(\lambda^*)\underline{d}$$

where

$$\underline{d} = (\gamma, (1-\gamma), -\gamma, -(1-\gamma), -\gamma, -(1-\gamma), \gamma, (1-\gamma))$$

and γ is the proportion of negatives in dimension 3 of the completely independent table λ^* . Then $\underline{c}^*$ is in $\mathcal{M}_2$ but perpendicular to $\mathcal{M}_1$ with respect to both the inner products given by $D(\lambda^*)$ and $[-d^2 \ell_{\mu}^*(\underline{m}^*)]$. The ratio of the noncentrality parameters is therefore given by

$$\|\xi^*\|_{(1)}^2 / \|\xi^*\|_{(2)}^2$$

which is expression (20) of Chapter 4.

Example of increased power with classification error: In order for this to happen, the alternative hypothesis must be misspecified. The proof of Proposition 11 in Appendix D allows for this situation. Suppose one is testing in a $2 \times 2 \times 2$ table the model dimensions 1 and 2 together being independent of dimension 3 against the model dimensions 2 and 3 being conditionally independent given dimension 1. The dimensions of these models are 5 and 6, respectively. Suppose the direction ξ^* is

$$\xi^* = D^{-1}(\lambda^*) \underline{z}$$

where

$$\underline{z}' = (1 \ -1 \ -1 \ 1 \ -1 \ 1 \ 1 \ -1) .$$

This ξ^* is not in $\mathcal{M}_2$ so expression (18) of Appendix D must be used to compute the noncentrality parameter δ^2 . It is easily checked that ξ^* is perpendicular to $\mathcal{M}_2$ (and therefore $\mathcal{M}_1$) with respect to the inner product given by $D(\lambda^*)$. This implies $\delta^2 = 0$ when there is no classification error. It is also easy to check that for general

$\lambda^* \in \mathcal{M}_1$, ξ^* is perpendicular to $\mathcal{M}_1$ but not to $\mathcal{M}_2$ with respect to the inner product given by $[-d^2 \ell_{*}(\underline{\mu}^*)]$. This means $\delta^2 > 0$ when

there is classification error. That is, the asymptotic power is positive when there is classification error and zero when there is none.

REFERENCES

- Assakul, K., and Proctor, C. H. [1967]. Testing Independence in Two-Way Contingency Tables with Data Subject to Misclassification, Psychometrika 32: 67-76.
- Berkson, J. [1950]. Are There Two Regressions? J. Amer. Stat. Assoc. 45: 164-180.
- Birch, M. W. [1963]. Maximum Likelihood in Three-Way Contingency Tables, J. Roy. Stat. Soc. Ser. B 25: 220-223.
- Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. [1975]. Discrete Multivariate Analysis, M.I.T. Press, Cambridge, Mass.
- Bross, I. [1954]. Misclassification in 2×2 Tables, Biometrics 10: 478-486.
- Buell, P., and Dunn, Jr., J. E. [1964]. The Dilution Effect of Misclassification, Amer. J. Public Health 54: 598-602.
- Chiacchierini, R. P., and Arnold, J. C. [1977]. A Two-Sample Test for Independence in 2×2 Contingency Tables with Both Margins Subject to Misclassification, J. Amer. Stat. Assoc. 72: 170-174.
- Cox, D. R. [1970]. The Analysis of Binary Data, Methuen, London.
- Dalenius, T. [1977]. Bibliography on Non-Sampling Errors in Surveys, International Statistical Review 45: 71-89, 181-197, 303-317.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. [1977]. Maximum Likelihood from Incomplete Data via the EM Algorithm, J. Roy. Stat. Soc. Ser. B 39: 1-22.
- Diamond, E. L. [1958]. Asymptotic Power and Independence of Certain Classes of Tests on Categorical Data, University of North Carolina Institute of Statistics, Mimeograph Series No. 196.

- Diamond, E. L., and Lilienfeld, A. M. [1962a]. Effects of Errors in Classification and Diagnosis in Various Types of Epidemiological Studies, Amer. J. Public Health 52: 1137-1144.
- Diamond, E. L., and Lilienfeld, A. M. [1962b]. Misclassification Errors in 2×2 Tables with One Margin Fixed, Some Further Comments, Amer. J. Public Health 52: 2106-2110.
- Dixon, W. J., and Brown, M. B. [1977]. BMDP-77, Biomedical Computer Programs, University of California Press, Berkeley, Ca.
- Fleiss, J. L. [1973]. Statistical Methods for Rates and Proportions, Wiley, N.Y.
- Goldberg, J. D. [1972]. The Effects of Misclassification on the Analysis of Data in 2×2 Tables, Harvard School of Public Health Doctor of Science thesis, Boston.
- Goldberg, J. D. [1975]. The Effects of Misclassification on the Bias in the Difference Between Two Proportions and the Relative Odds in the Fourfold Table, J. Amer. Stat. Assoc. 70: 561-567.
- Grizzle, J. E., Starmer, C. F., and Koch, G. G. [1969]. Analysis of Categorical Data by Linear Models, Biometrics 25: 489-504.
- Haberman, S. J. [1974a]. The Analysis of Frequency Data, University of Chicago Press, Chicago.
- Haberman, S. J. [1974b]. Log-Linear Models for Frequency Tables Derived by Indirect Observations: Maximum Likelihood Equations, Annals of Stat. 2: 911-924.
- Haberman, S. J. [1977]. Product Models for Frequency Tables Involving Indirect Observation, Annals of Stat. 5: 1124-1147.

- Hochberg, Y. [1977]. On the Use of Double Sampling Schemes in Analyzing Categorical Data with Misclassification Errors, J. Amer. Stat. Assoc. 72: 914-921.
- Keys, A., and Kihlberg, J. K. [1963]. Effect of Misclassification on Estimated Relative Prevalence of a Characteristic, Amer. J. Public Health 53: 1656-1665.
- Koch, G. G. [1969]. The Effect of Non-Sampling Errors on Measures of Association in 2×2 Contingency Tables, J. Amer. Stat. Assoc. 64: 852-863.
- Lilienfeld, A. M., and Graham, S. [1958]. Validity of Determining Circumcision Status by Questionnaire as Related to Epidemiological Studies of Cancer of the Cervix, J. Nat. Cancer Inst. 21: 713-720.
- Mantel, N., and Haenszel, W. [1959]. Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease, J. Nat. Cancer Inst. 22: 719-748.
- Mote, V. L., and Anderson, R. L. [1965]. An Investigation of the Effect of Misclassification on the Properties of χ^2 -Tests in the Analysis of Categorical Data, Biometrics 52: 95-109.
- Newell, D. J. [1962]. Errors in the Interpretation of Errors in Epidemiology, Amer. J. Public Health 52: 1925-1928.
- Plackett, R. L. [1974]. The Analysis of Categorical Data, Griffin, London.
- Press, S. J. [1968]. Estimating from Misclassified Data, J. Amer. Stat. Assoc. 63: 123-133.
- Rao, C. R. [1973]. Linear Statistical Inference and Its Applications, second edition, Wiley, New York.

- Rogot, E. [1961]. A Note on Measuring Errors and Detecting Real Differences, J. Amer. Stat. Assoc. 56: 314-319.
- Rubin, T., Rosenbaum, J., and Cobb, S. [1956]. The Use of Interview Data for the Detection of Associations in Field Studies, J. of Chronic Diseases 4: 253-266.
- Sadowsky, D. A., Gilliam, A. G., and Cornfield, J. [1953]. The Statistical Association between Smoking and Carcinoma of the Lung, J. Nat. Cancer Inst. 13: 1237-1258.
- Scheffe, H. [1959]. The Analysis of Variance, Wiley, New York.
- Tenenbein, A. [1969]. Estimation from Data Subject to Measurement Error, Harvard Statistics Department Ph.D. thesis, Cambridge, Mass.
- Tenenbein, A. [1970]. A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications, J. Amer. Stat. Assoc. 65: 1350-1361.
- Whittemore, A.S. and Korn, E.L. [1978]. Methods for Analyzing Panel Studies of Acute Health Effects of Air Pollution, Stanford Technical Report (to appear).