ANALYSIS OF CATEGORICAL DATA

By

GENE BURTON IVERSON

Bachelor of Science in Education
University of South Dakota
Vermillion, South Dakota
1960

Master of Natural Science
University of South Dakota
Vermillion, South Dakota
1964

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF EDUCATION
July, 1973

ANALYSIS OF CATEGORICAL DATA

Thesis Approved:

P. L. Claypool
Thesis Adviser

E. K. McEachran

D B Ackela

James B. Appleberry

D D Dunham
Dean of the Graduate College

ACKNOWLEDGMENTS

To Professor P. Larry Claypool, my thesis adviser, my thanks for his effort to help and guide me with this dissertation. Without his guidance, this paper would not have become a reality.

I also wish to thank Professor E. K. McLachlan for serving as my committee chairman. His encouragement and advice were invaluable to me. For their suggestions and cooperation my gratitude goes to the other member of my committee: Dr. Bruce Aichele and Dr. James Appleberry.

Finally, I wish to express my gratitude to my wife, Marjorie, and my children, Brian, Dane and Kip, for their encouragement and many sacrifices made during these years of graduate study.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Basic Concepts of Contingency Table Analysis

The least restrictive level of measurement of data is a nominal scale. The nominal scale of measurement results when each response names a class or category which identifies some characteristic of the unit observed. Since the nominal scale of measurement does not specify any order or metric relationship, the relevant statistic is the number of times a given class is named. Examples of categorizing schemes employed are "yes" or "no"; "defect" or "no defect"; and "superior", "good", "average", or "poor". Data measured on a nominal scale is generally referred to as categorical data since it represents the tallying of frequency or cell counts by the categories of one or more classification variables.

Most statistical analyses of categorical data involve testing one of the three hypotheses that the classification variables defined on a population are mutually independent, that the population sampled has a specified distribution or that the several populations sampled have identical distributions. The test statistics used to test these hypotheses are variations of the Pearson chi-square statistic.

In applying the Pearson chi-square statistic to categorical data, the data is partitioned by one or more criterion or classification

variables. A contingency table is used to summarize the categorical data to facilitate the computation of estimates of parameters and calculation of test statistic for testing hypotheses. The dimension of a contingency table is the number of classification variables by which the data is categorized or partitioned. A two dimensional contingency table is an array of natural numbers arranged into r rows and c columns, with rc cells or categories for the numbers. If we let A be a row classification and B be a column classification with r and c categories for each classification, respectively, then Table I represents a contingency table where n_{ij} denotes the number of responses naming class i within classification variable A and class j within classification variable B for $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, c$.

TABLE I

TWO-DIMENSIONAL CONTINGENCY TABLE

		B					
		1	2	. . .	j	. . .	c
A	1	n_{11}	n_{12}		n_{ij}		n_{1c}
	2	n_{21}	n_{22}		n_{2j}		n_{2c}
	⋮						
	i	n_{i1}	n_{i2}		n_{ij}		n_{ic}
	⋮						
	r	n_{r1}	n_{r2}		n_{rj}		n_{rc}

The natural numbers n_{ij} represent the counts, or frequencies, for the categorical data, which have been tabulated according to each of the classification variables. An m-dimensional contingency table represents data classified by m classification variables and is often referred to as an m-way contingency table.

A sample is a collection of objects or persons which may be a subset of the population defined by the objectives of the investigation. If inferences are desired to the population, it is necessary that the sample be random. To illustrate the above, assume the population of all students enrolled at a specified university and that a random sample of N students is obtained from the population. Suppose the sample is to be classified by the three classification variables of class, grade point average, and major field of study. This partitioning and tabulation of the sample is represented by a three-dimensional contingency table.

A sample classified by one criterion (one-way contingency table) usually involves the hypothesis that the population sampled has a specified distribution. A test of this hypothesis is often referred to as a goodness-of-fit test.

To obtain a two-way contingency table a sample may be obtained from a single population and partitioned by two classification variables or a sample may be obtained from each of two populations and each sample partitioned by one classification variable. In the first case one may test either the hypothesis that the population as partitioned by the two classification variables has a specified distribution or the hypothesis that the classification variables are independent of each

other, while in the latter case the hypothesis tested is that the two populations are identically distributed.

Multinomial Distribution

If probabilistic methods of analyses are to be applied to contingency tables, then the contingency tables must be assumed to have been generated by some probability model defined on each population sampled. The model that will be assumed is the multinomial distribution, which partitions the population. The objective is to obtain one or more samples and then to test hypotheses which involve the parameters of the population(s) samples. The multinomial distribution contains a parameter corresponding to each partition of the population which is the probability that a unit selected at random from the population will be classified into that partition. If one knows the value of each parameter, then the distribution of the population is determined.

Suppose a sample from a population is partitioned by one classification variable. Let n_i denote the observed frequency and p_i denote the probability of the i^{th} category of the classification variable. The multinomial distribution is the joint distribution of the observed frequencies and is given by

$$f(n_1, n_2, \ldots, n_r) = \frac{N!}{\prod_{i=1}^{r} n_i!} \prod_{i=1}^{r} p_i^{n_i}, \qquad (1.1)$$

for each $n_i = 1, 2, \ldots, N$, $\sum_{i=1}^{r} n_i = N$ and $\sum_{i=1}^{r} p_i = 1$.

The product symbol is defined by

$$\prod_{i=1}^{r} p_i^{n_i} = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r} \quad \text{and} \quad \prod_{i=1}^{r} n_i! = n_1! n_2! \cdots n_r! .$$

If the population sampled is partitioned by two classification variables having r and c categories, respectively, then let n_{ij} denote the number of units in the sample of size N which are classified into the i^{th} category of the first classification variable and the j^{th} category of the second classification variable. We will let p_{ij} denote the probability that a unit selected at random from the population will be classified into the partition (i,j). The data may then be summarized by a two-dimensional contingency table such as Table I.

The joint distribution of the observed frequencies n_{ij} is the multinomial distribution given by

$$f(n_{11}, n_{12}, \ldots, n_{rc}) = \frac{N!}{\prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!} \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{n_{ij}} , \qquad (1.2)$$

$$\text{for each } n_{ij} = 0, 1, \ldots, N, \quad \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} = N \quad \text{and} \quad \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = 1 .$$

For a sample of size N we shall define the marginal totals $n_{i.}$ (denoted as the row total) and $n_{.j}$ (denoted as the column total) by

$$n_{i.} = \sum_{j=1}^{c} n_{ij} ,$$

$$n_{.j} = \sum_{i=1}^{r} n_{ij} . \qquad (1.3)$$

The marginal probability totals are defined by the following equations

$$p_{i.} = \sum_{j=1}^{c} p_{ij} ,$$

$$p_{.j} = \sum_{i=1}^{r} p_{ij} . \qquad (1.4)$$

If we have r independent samples from r populations and the populations are partitioned by one classification variable, then the contingency table given in Table I could be a representation of the data, where the rows represent the r independent samples and the columns represent the c categories of the classification variable. In this case the marginal totals for the rows (samples) are assumed to be fixed or determined before the samples were obtained. The form of the distribution is the same as equation (1.3) except for the following constraint on the parameters

$$\sum_{j=1}^{c} p_{ij} = 1 , \quad \text{for each} \quad i = 1, 2, \ldots, r .$$

A similar situation exists if we have c independent samples from c populations and the populations are partitioned by one classification variable, where the columns represent the observations in the independent samples and the rows represent the categories of the classification variable. Again, the form of the distribution is given by the equation (1.3) and we have the following constraint on the parameters

$$\sum_{i=1}^{r} p_{ij} = 1 , \quad \text{for each} \quad j = 1, 2, \ldots, c .$$

The multinomial distribution for a sample partitioned by three classifications may be extended from equation (1.5) by using three product symbols and three subscripts, each subscript denoting a category of a classification variable. A corresponding statement may then be made for the distribution of an m-dimensional contingency table.

The hypothesis of independence of two classification variables defined on a population when stated in terms of a probability model for a two-dimensional contingency table becomes

$$H_0 : p_{ij} = p_{i.}\, p_{.j} \quad \text{for all} \quad i = 1, 2, \ldots, r, \quad \text{and}$$
$$j = 1, 2, \ldots, c. \tag{1.5}$$

The alternative hypothesis, denoted by H_1, is given by

$$H_1 : p_{ij} \neq p_{i.}\, p_{.j} \quad \text{for some pair} \quad (i, j). \tag{1.6}$$

The Pearson chi-square test statistic is given by

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}, \tag{1.7}$$

where E_{ij} is the expected frequency for the (i, j) cell, $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$. The test statistic T is distributed asymptotically as a chi-square random variable ([58], p. 118).

There are two situations that may arise when the null hypothesis is stated, namely: the null hypothesis specifies the parameters (probabilities) $p_{i.}$ for $i = 1, 2, \ldots, r$, and $p_{.j}$ for $j = 1, 2, \ldots, c$, ; or these parameters are not specified by the null hypothesis.

If the hypothesis H_0 specifies the all parameters $p_{i.}$ and $p_{.j}$, then $E_{ij} = N p_{ij} = N p_{i.} p_{.j}$, for all $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$. When $rc-1$ of the parameters are known in addition to the constraint

$$\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = 1 ,$$

then all of the parameters are determined. For the multinomial model, the number of parameters specified by H_0 determines the degrees of freedom associated with the test statistic. The test statistic (1.7) for this case has $rc-1$ degrees of freedom.

If the null hypothesis H_0 does not specify the parameters $p_{i.}$ and $p_{.j}$, then the parameters are estimated from the sample. An estimator for $p_{i.}$ is given by

$$\hat{p}_{i.} = \frac{n_{i.}}{N} \quad \text{for} \quad i = 1, 2, \ldots, r ;$$

similarly, $p_{.j}$ may be estimated by

$$\hat{p}_{.j} = \frac{n_{.j}}{N} \quad \text{for} \quad j = 1, 2, \ldots, c .$$

To denote that $p_{i.}$ and $p_{.j}$ are estimated from the sample, we use the notation $\hat{p}_{i.}$ and $\hat{p}_{.j}$. The expected frequency E_{ij} in the test statistic (1.7) is estimated by

$$\hat{E}_{ij} = N \hat{p}_{i.} \hat{p}_{.j} = N \frac{n_{i.}}{N} \frac{n_{.j}}{N} = \frac{n_{i.} n_{.j}}{N}$$

for all i and j. Since

$$\sum_{i=1}^{r} p_{i.} = 1 ,$$

estimating any $r-1$ of the parameters $p_{i.}$ determines the r^{th} parameter. Similarly,

$$\sum_{j=1}^{c} p_{.j} = 1$$

implies that if we estimate any $c-1$ of the parameters $p_{.j}$ then the c^{th} parameter is determined. The degrees of freedom for the test statistic (1,7) when the parameters $p_{i.}$ and $p_{.j}$ are estimated from the data are given by

$$rc-1-(r-1)-(c-1) = rc-r-r-c+1 = (r-1)(c-1) . \quad (1.8)$$

That is, one degree of freedom is subtracted from $rc-1$ for each parameter estimated.

In the hypothesis of identical distributions of a set of r populations sampled in which each is partitioned into c categories by a single classification variable, the hypotheses in terms of the probability model are given by

$$H_0 : p_{1j} = p_{2j} = \ldots = p_{rj} \quad \text{for} \quad j = 1,2,\ldots,c \quad (1.9)$$

and

H_1 : at least one population has a different multinomial distribution.

One further note about the parameters for the hypothesis of identical distributions is that the parameter p_{ij} is the probability of an

observation selected at random from the i^{th} population being classified into the j^{th} category of the classification variable. Again, the null hypothesis may or may not specify the parameters. If the null hypothesis specifies the parameters, then $E_{ij} = N p_{ij}$ for all i and j. The degrees of freedom for the test statistic (1.7) is given by $r(c-1)$. In this case the sample size $n_{i\cdot}$ for each i is considered to be specified and for each population

$$\sum_{j=1}^{c} p_{ij} = 1 ,$$

hence for each population there are $c-1$ probabilities to be determined. Thus, to determine the multinomial distribution (1.2), $r(c-1)$ probabilities must be known.

If the null hypothesis does not specify the parameters, then H_0 implies there are $c-1$ parameters to be estimated. These are estimated from the sample by

$$\hat{P}_{ij} = \frac{n_{\cdot j}}{N} \quad \text{for} \quad j = 1, 2, \ldots, c$$

and

$$\hat{E}_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{N} \quad \text{for all } i \text{ and } j .$$

Since $c-1$ parameters are estimated in this case, the number of degrees of freedom is given by $r(c-1) - (c-1) = (r-1)(c-1)$ for the test statistic (1.7).

The distribution of the statistic T (1.7) may be poorly approximated by the chi-square distribution if the following conditions are

found in a contingency table:

(1) if any E_{ij} (or $\hat{E}_{ij}$) is less than 1,

(2) if more than 20% of the E_{ij}'s or $\hat{E}_{ij}$'s are less than 5.

The exception to these conditions arises when all (or most) of the E_{ij} are nearly the same size. If r and c are not too small, then the E_{ij} may be as small as one without endangering the validity of the test ([10], p. 152).

Scope and Objectives

A common problem confronting researchers concerns devising useful methods for analyzing categorical data. Researchers familiar with the analysis of variance have well-developed techniques for quantitative variables, but must switch to a completely different set of varied techniques when they deal with qualitative data. Most of the information in textbooks, where the analysis of categorical data is discussed, covers the analysis of two-dimensional contingency tables in detail assuming the analysis can be extended easily to multi-dimensional tables. The analysis of higher order contingency tables is important in research and there are many more hypotheses that can be tested, which cannot be generalized from two-way tables.

This paper will endeavor to present the reader with a basic understanding of the topic of the analysis of categorical data. Chapter II presents a discussion of hypothesis testing for three-way contingency tables. Chapter III will present techniques that are analogous to the analysis of variance by defining a component of information for

categorical data based on information theory applied to statistics. Because the chi-square test and information statistics are based on large sample statistics, it is intended to present methods of analysis where difficulties are encountered in small, zero or missing frequency counts in Chapter IV. In Chapter V a method analogous to the analysis of variance technique is discussed by defining variation in categorical data for hypothesis testing and to obtain a measure of association (dependency) between classification variables.

It is intended to present the material in such a manner that a student or researcher with limited mathematical training would have little difficulty in understanding the paper.

Examples, definitions and theorems will be numbered serially with the first digit being the number of the chapter. Equations will be numbered in a similar manner when they may be needed for easy future reference. The tables are numered consecutively throughout the paper.

CHAPTER II

HYPOTHESES FOR MULTIDIMENSIONAL CONTINGENCY TABLES

Notation

In this chapter we will be primarily concerned with the three-dimensional contingency table. For the discussion of the three classification variables we will use the labels of row, column and depth classifications. The general case of the three-dimensional contingency table will be denoted with the symbols $r \times c \times d$ where r, c and d represent the number of categories in the row, column, and depth classifications, respectively. Let n_{ijk} denote the observed frequency in the category given by the i^{th} row, j^{th} column and k^{th} depth classifications and let p_{ijk} denote the probability of an observation occurring in cell (i, j, k).

If the observed frequencies n_{ijk} are summed over all values of i (from 1 to r), the result will be defined as the second-order marginal totals of the j^{th} column in the k^{th} depth classification. This marginal total is accordingly designated $n_{\cdot jk}$, so that

$$\sum_{i=1}^{r} n_{ijk} = n_{\cdot jk} . \qquad (2.1)$$

Similarly, by summing n_{ijk} over j or over k gives the following

second-order marginal totals:

$$\begin{aligned} \sum_{j=1}^{c} n_{ijk} &= n_{i\cdot k} \\ \sum_{k=1}^{d} n_{ijk} &= n_{ij\cdot} \end{aligned} \tag{2.2}$$

If the observed frequencies n_{ijk} are summed over all values of both i and j we obtain the first-order marginal totals of the k^{th} depth classification. This total is designated by

$$n_{\cdot\cdot k} = \sum_{j=1}^{c} \sum_{i=1}^{r} n_{ijk}, \tag{2.3}$$

with $n_{i\cdot\cdot}$ and $n_{\cdot j\cdot}$ defined in a similar manner. If the frequencies n_{ijk} are summed over all values of i, j, and k, then the result will be the total number of observations in the sample; i.e.,

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} n_{ijk} = n_{\cdots} = N . \tag{2.4}$$

A similar notation is used for the parameters p_{ijk}, where

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} p_{ijk} = 1 . \tag{2.5}$$

A summary of the formulae for summing the parameters follows for future reference where the parameters given are second-order and first-order probabilities, respectively:

$$\sum_{i=1}^{r} p_{ijk} = p_{\cdot jk}$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ijk} = p_{\cdot\cdot k} \cdot \qquad (2.6)$$

Further formulae may be obtained by permuting the role of the subscripts in equations (2.6). The entire notation for a system involving more than three classification variables may be extended with little difficulty.

Basic Hypotheses in Terms of Probability Statements

The probability statements for the hypotheses will be presented in terms of the sampling structures which give three-dimensional contingency tables. The extension of the analysis of a two-way table to a three-way contingency table poses entirely new conceptual problems. On the other hand, there are no new problems involved in making extensions from tables of three dimensions to those of four or more dimensions ([45], p. 88). The possible combinations of the hypotheses of interest become numerous for three-way and higher order contingency tables.

Additional comments need to be made in regard to the effect the sampling procedure may have on the statements of the hypotheses. In multi-classification of a sample it is usually the case that the sample size is assumed fixed, but none of the marginal totals are fixed. In the next section a discussion of the effect of fixing the marginal totals on the hypotheses will be presented.

The general form of the test statistic is given by

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} \frac{(n_{ijk} - E_{ijk})^2}{E_{ijk}} \qquad (2.7)$$

where E_{ijk} is the expected frequency of cell (i, j, k). We are assuming for the test statistic (2.7) that each population sampled has the multinomial distribution. In using the test statistic (2.7) for testing the hypotheses to follow only the method of determining or estimating E_{ijk} varies with the form of the different hypotheses. If the null hypothesis specifies all of the parameters then $E_{ijk} = N p_{ijk}$. If the null hypothesis does not specify the parameters p_{ijk} but specifies a relationship among them, then the expected frequencies must be estimated. An estimate will be denoted by $\hat{E}_{ijk}$. In this section we will be concerned with the form of null hypotheses which do not specify all parameters.

To extend the concept of independence of classification variables for a two-dimensional contingency table to independence of classification variables for a three-dimensional contingency table, suppose we obtain a random sample of size N from a population. If we partition this sample by three classification variables, then the null hypothesis for <u>mutual independence</u> is given by

$$H_0 : p_{ijk} = p_{i..}\, p_{.j.}\, p_{..k} \quad \text{for all } i = 1, 2, \ldots, r;$$
$$j = 1, 2, \ldots, c; \text{ and } k = 1, 2, \ldots, d. \qquad (2.8)$$

The alternative hypothesis, denoted by H_1, is

$$H_1 : p_{ijk} \neq p_{i..}\, p_{.j.}\, p_{..k} \quad \text{for some } i, j \text{ and } k. \qquad (2.9)$$

To determine the value of the test statistic (2.7) the estimate of the parameters $p_{i..}$, $p_{.j.}$, and $p_{..k}$ are given by

$$\hat{p}_{i..} = \frac{n_{i..}}{N} ,$$

$$\hat{p}_{.j.} = \frac{n_{.j.}}{N} , \qquad (2.10)$$

and $$\hat{p}_{..k} = \frac{n_{..k}}{N} , \quad \text{respectively.}$$

The estimate of the expected frequency E_{ijk} is given by

$$\hat{E}_{ijk} = N\hat{p}_{ijk} = N\hat{p}_{i..}\,\hat{p}_{.j.}\,\hat{p}_{..k} \qquad (2.11)$$

based on the relationship given by the null hypothesis (2.8). There are $rcd - 1 - (r-1) - (c-1) - (d-1) = (r-1)(c-1)(d-1)$ degrees of freedom for the test statistic (2.7). The degrees of freedom were determined by subtracting from $rcd - 1$ the total number of parameters estimated.

If the test statistic for mutual independence gives a significant result (H_0 is rejected), then it should not be assumed that all three classifications interact. It might be the case that just two of the classification interact and the third is completely independent. This gives rise to three testable hypotheses, since any of the three classifications could be the independent one.

To test whether the row classification is independent of the others, the null hypothesis for a three-way contingency table is

$$H_0 : p_{ijk} = p_{i..}\,p_{.jk} \quad \text{for all } i, j \text{ and } k. \qquad (2.12)$$

The alternative hypothesis is

$$H_1 : p_{ijk} \neq p_{i..}\, p_{.jk} \quad \text{for some } i,\ j \text{ and } k.$$

Note that the null hypothesis (2.12) implies

$$p_{ij.} = \sum_{k=1}^{d} p_{ijk} = p_{i..} \sum_{k=1}^{d} p_{.jk} = p_{i..}\, p_{.j.}\ , \tag{2.13}$$

and

$$p_{i.k} = \sum_{j=1}^{c} p_{ijk} = p_{i..} \sum_{j=1}^{d} p_{.jk} = p_{i..}\, p_{..k}\ . \tag{2.14}$$

That is, the row and column classifications are independent (row x column interaction is zero) and the row and depth classifications are independent (row x depth classification interaction is zero) if H_0 given by (2.12) is true. Equations (2.13) and (2.14) do not imply the null hypothesis (2.12) (see Kullback [39], p. 163).

Since H_0 given by (2.12) does not specify the values of the parameters $p_{i..}$ and $p_{.jk}$ the test statistic given by (2.7) would have $(cd-1)(r-1)$ degrees of freedom. If the $p_{i..}$'s and $p_{.jk}$'s are estimated by $n_{i..}/N$ and $n_{.jk}/N$, respectively, then $(r-1)$ and $(cd-1)$ degrees of freedom are lost by the estimation process.

The following example will be used to illustrate the test statistic (2.7) for the various hypotheses discussed thus far.

Example 2.1. The data in Table II is the result of an experiment involving the repression of failure. A sample of $N = 106$ boys were given a series of 16 tests. The measure of repression was the difference between the number of complete and incomplete tests that

were recalled by the individuals. The subjects were classified as to social class and the type of discipline used by the parents. The categories of discipline are psychological, mixed (psychological and corporal) and corporal. The textual discussion of this study is found in Miller ([48], Chapter 10).

TABLE II

REPRESSION OF FAILURE

Social / Recall	Working Class			Middle Class		
	Psychol.	Mixed	Corp.	Psychol.	Mixed	Corp.
	Discipline			Discipline		
Positive	6	3	6	19	6	5
Zero	9	4	0	7	3	3
Negative	7	3	11	12	1	1

Some preliminary calculations are given below for use in evaluating the test statistics used in testing for mutual independence and for the independence of one classification variable. The first-order and second-order marginals are given by

First-Order Marginals

Recall	Social Class	Discipline
$n_{1..} = 45$	$n_{.1.} = 49$	$n_{..1} = 60$
$n_{2..} = 26$	$n_{.2.} = 57$	$n_{..2} = 20$
$n_{3..} = 35$		$n_{..3} = 26$

Second-Order Marginals

Recall x Social Class	Recall x Discipline	Social Class x Discipline
$n_{11.} = 15$	$n_{1.1} = 25$	$n_{.11} = 22$
$n_{12.} = 30$	$n_{1.2} = 9$	$n_{.12} = 10$
$n_{21.} = 13$	$n_{1.3} = 11$	$n_{.13} = 17$
$n_{22.} = 13$	$n_{2.1} = 16$	$n_{.21} = 38$
$n_{31.} = 21$	$n_{2.2} = 7$	$n_{.22} = 10$
$n_{32.} = 14$	$n_{2.3} = 3$	$n_{.23} = 9$
	$n_{3.1} = 9$	
	$n_{3.2} = 4$	
	$n_{3.3} = 12$	

To test the hypothesis of mutual independence

$$H_0 : p_{ijk} = p_{i..}\, p_{.j.}\, p_{..k} \quad \text{for all } i, j \text{ and } k$$

versus the alternative hypothesis

$$H_1 : \text{not } H_0 ;$$

the test statistic is given by formula (2.7) where E_{ijk} is estimated by (2.11) for all i, j and k. This follows from the assumption that marginals were not fixed by the sampling technique. Thus estimates of the parameters $p_{i..}$, $p_{.j.}$ and $p_{..k}$ are given by (2.10) for $i = 1, 2, 3$ $j = 1, 2$ and $k = 1, 2, 3$.

The value of the test statistic is $T = 28.8560$ and the calculated value of T is compared with the chi-square distribution with 12 degrees of freedom. The critical level $\hat{\alpha}$, which is defined to be the smallest significance level at which the null hypothesis would be rejected for the observed value of T ([10], p. 81), is given by $\hat{\alpha} \approx .005$.

If we reject H_0 based on $\hat{\alpha} = .005$ being less than any of the commonly used significance levels, then one might be interested in whether the recall classification is independent of both social class and discipline. The null hypothesis would then be given by

$$H_0 : p_{ijk} = p_{i..}\, p_{.jk} \quad \text{for all } i = 1, 2;\ j = 1, 2, 3;\ \text{and } k = 1, 2, 3.$$

where the alternative hypothesis is a simple negation of H_0. The test statistic is (2.7) where $\hat{E}_{ijk} = N \hat{p}_{i..}\, \hat{p}_{.jk}$ for all i, j and k since the marginals were not assumed to be given. Here the parameters are estimated by

$$\hat{p}_{i..} = \frac{n_{i..}}{N} \quad \text{for } i = 1, 2, 3;$$

and

$$\hat{p}_{.jk} = \frac{n_{.jk}}{N} \quad \text{for } j = 1, 2, \text{ and } k = 1, 2, 3.$$

The computed value of T is 20.044 which is compared to the chi-square distribution with

$$rcd - 1 - (r-1) - (cd-1) = (r-1)(cd-1) = 10$$

degrees of freedom. The critical level is $\hat{\alpha} \approx .025$. In a similar manner one could also test the hypotheses that social class is independent of the recall and discipline classifications or that the discipline classification is independent of the social class and recall classifications.

In some three-way tables it is of interest to test the hypothesis that given any depth classification, for example, the row and column classifications are independent. This hypothesis is referred to as an hypothesis of <u>conditional independence</u>. For the three-dimensional case in which a random sample of size N is taken from a single population the null and alternative hypotheses may be written as

$$H_0 : p_{ijk} = p_{i\cdot k} p_{\cdot jk} / p_{\cdot\cdot k} \quad \text{for all } i, j \text{ and } k;$$

and

$$H_1 : p_{ijk} \neq p_{i\cdot k} p_{\cdot jk} / p_{\cdot\cdot k} \quad \text{for some } i, j \text{ and } k.$$

Since the above hypotheses involve $p_{i\cdot k}$ and $p_{\cdot jk}$ it is relevant to point out that if $p_{i\cdot k} = p_{i\cdot\cdot} p_{\cdot\cdot k}$ (i.e., there is no $r \times d$ interaction) or if $p_{\cdot jk} = p_{\cdot j\cdot} p_{\cdot\cdot k}$ (i.e., there is no $c \times d$ interaction) then the hypothesis given by (2.15) becomes

$$p_{ijk} = p_{i\cdot\cdot} p_{\cdot jk} \quad \text{(complete independence of rows)} \qquad (2.16)$$

or

$$p_{ijk} = p_{\cdot j\cdot}\, p_{i\cdot k} \quad \text{(complete independence of columns)}, \qquad (2.17)$$

respectively. Both imply

$$p_{ij\cdot} = p_{i\cdot\cdot}\, p_{\cdot j\cdot} \qquad (2.18)$$

by summing (2.16) and (2.17) over the index k.

Consider the Example 2.1 and suppose we test the hypothesis (2.15). We will assume no marginals are fixed so the parameters $p_{i\cdot k}$, $p_{\cdot jk}$ and $p_{\cdot\cdot k}$ are estimated by the quantities:

$$\hat{p}_{i\cdot k} = \frac{n_{i\cdot k}}{N} \quad \text{for} \quad i = 1,2,3\ , \quad j = 1,2,3\ ;$$

$$\hat{p}_{\cdot jk} = \frac{n_{\cdot jk}}{N} \quad \text{for} \quad j = 1,2\ , \quad k = 1,2,3\ ;$$

$$\hat{p}_{\cdot\cdot k} = \frac{n_{\cdot\cdot k}}{N} \quad \text{for} \quad k = 1,2,3\ ,$$

respectively. The estimated expected frequency is

$$\hat{E}_{ijk} = N\hat{p}_{i\cdot k}\hat{p}_{\cdot jk}/\hat{p}_{\cdot\cdot k} = \frac{n_{i\cdot k}n_{\cdot jk}}{n_{\cdot\cdot k}}$$

for all i, j and k. The computed value of the test statistic is $T = 46.89$ and the degrees of freedom are given by the formula $(r-1)(c-1)d = 4$. The critical level $\hat{\alpha}$ is much less than .001.

The formulation of hypotheses up to this point has assumed one random sample of size N. Suppose we obtain d independent samples from d populations of size N_k, $k = 1,2,\ldots,d$ where each population is partitioned by two classification variables. We may

represent the d $r \times c$ contingency tables which result as an $r \times c \times d$ three-way contingency table with suitable hypothesis and restrictions. For the restrictions in this case it is reasonable to assume that each sample size is fixed; that is, $n_{..k} = N_k$ for $k = 1, 2, \ldots, d$ is determined before the samples are obtained. Thus, suppose we want to test the hypothesis that the d samples were taken from populations having identical distributions. The parameters p_{ijk} denote the probability that an observation taken at random from population k will be classified into the i^{th} and j^{th} category by the two classification variables, respectively. There are rc parameters p_{ijk} associated with population k where

$$\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ijk} = p_{..k} = 1 \quad \text{for} \quad k = 1, 2, \ldots, d .$$

The joint distribution of the observed frequencies n_{ijk} associated with a sample of size N_k from population k is the multinomial distribution given by (1.2) for each value $k = 1, 2, \ldots, d$ with $n_{ij.}$ and p_{ij} replaced by n_{ijk} and p_{ijk}, respectively, to identify the specific population k.

The hypotheses for <u>identical distributions</u> are given by

$$H_0: p_{ijk} = p_{ij.} \text{ for all i, j and k where } \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij.} = 1 \quad (2.19)$$

and

$$H_1: p_{ijk} \neq p_{ij.} \text{ for some i, j and k.}$$

We can estimate $p_{ij.}$ by

$$\hat{p}_{ij.} = \frac{n_{ij.}}{N} \quad \text{for all } i \text{ and } j. \tag{2.20}$$

Under the null hypothesis (2.19) the estimated expected value is

$$\hat{E}_{ijk} = N_k \hat{p}_{ij.} = N_k \frac{n_{ij.}}{N} \quad \text{for all } i, j \text{ and } k, \tag{2.21}$$

for determining the value of the statistic (2.7).

Example 2.2. Let us consider the experimental terminology of Example 2.1 and assume that the data in Table II gives the results of taking a sample from each of the three populations defined by the three forms of discipline parents use with their children. Assume, further that the sample sizes are $N_1 = 60$, $N_2 = 20$, and $N_3 = 26$ and that each population is partitioned by the two classification variables: recall and social class. We will test the hypothesis (2.19). Based on the estimates given by equations (2.20) and (2.21) and the marginal totals computed in Example 2.1, the value of the test statistic is $T = 15.70$ which is compared with the chi-square distribution with $(rc-1)d = 15$ degrees of freedom. The critical level $\hat{\alpha} \approx .22$.

Fixed Marginals

Sometimes when a random sample might produce disproportionately low frequencies in some section of a contingency table, the experimenter might decide to specify not only the sample size N, but also marginal totals. For example in a three-dimensional contingency table the marginal totals $n_{..k}$ for $k = 1, 2, \ldots, d$ might be fixed. In addition to the above sampling constraint, many other kinds can be

envisaged. For instance, it is possible to fix $n_{i..}$ as well as the $n_{.jk}$ marginal totals, or to fix all or some of the first-order marginals $n_{i..}$, $n_{.j.}$ and $n_{..k}$. Since such restrictions are likely to be rare in practice, they are not discussed here. In any event it will nearly always turn out that the chi-square computation is the same whether the marginals are fixed or not. In such cases it will be only the power function and the form of the hypothesis that vary ([45], p. 93).

Lewis [45] would modify the null hypothesis (2.8) when the marginals $n_{..k}$ are given to

$$H_0 : p_{ijk} = p_{i..}\, p_{.j.}\, \frac{n_{..k}}{N} \quad \text{for all } i, j \text{ and } k.$$

The estimates of the $p_{..k}$ from the sample would not have been any different in the above case, but the important point is they were determined before the sample had been taken.

For testing the hypothesis of mutual independence (2.8) in a three-dimensional contingency table Kullback [39] would modify the null hypothesis if the $n_{..k}$ totals are fixed in advance. It might be reasoned that there are in effect d distinct tables of size $r \times c$. In these circumstances p_{ijk} denotes the probability that an observation falls in the (i,j) cell of the k^{th} two-way table. Moreover, if each two-way table is considered separately, then

$$\hat{p}_{ijk} = \frac{n_{ijk}}{n_{..k}} \quad \left(\text{not } \frac{n_{ijk}}{N} \right) .$$

and $p_{..k} = 1$ for all k (not $n_{..k}/N$). Hence, the null hypothesis

(2.8) would be modified to

$$H_0 : p_{ijk} = p_{i\cdot\cdot} p_{\cdot j\cdot} \quad \text{for all } i, j \text{ and } k.$$

Complexities of Formulating Hypotheses

In the analysis of contingency tables obtained from a single population we are usually interested in the relationship between one classification and one or more of the other classifications. Suppose we consider the contingency table resulting from Example 2.1 for illustrative purposes. One could consider the row classification as representing the response of an experiment on these individuals, the column classification as a distinguishable characteristic of the sampled individuals, and the depth classification as types of treatment. Then in many respects the hypothesis of interest are analogous to those of independence and correlation in normal multivariate analysis. For example:

1. Response is independent of treatment, or

   $$H_0 : p_{i\cdot k} = p_{i\cdot\cdot} p_{\cdot\cdot k} \quad \text{for all } i \text{ and } k.$$

   This case corresponds to simple correlation. That is, H_0 corresponds to the hypothesis that response and treatment are uncorrelated.

2. Response is independent of treatment and social class, or

   $$H_0 : p_{ijk} = p_{i\cdot\cdot} p_{\cdot jk} \quad \text{for all } i, j \text{ and } k.$$

   This case corresponds to multiple correlation.

3. Response is independent of treatment given the social class, or

$$H_0 : p_{ijk} = \frac{p_{ij.}\, p_{.jk}}{p_{.j.}} \quad \text{for all } i, j \text{ and } k .$$

This hypothesis is the conditional independence given the column classification and corresponds to partial correlation ([37], p. 160).

Not all contingency tables can be interpreted in a straightforward manner. In some cases all three classifications can be considered as responses; then we may be interested in independence or association among the responses. In other cases a classification may be viewed either as a factor or a response. For convenience, we may group all the concepts of association or dependence under the general term of interaction.

The reader may have noted that up to this point no attempt has been made to define interaction among the classification variables defined on one or more populations. We have indicated only that if classification variables are independent, then there is no interaction between classification variables. With reference to Table II we may also say that the interaction between response and treatment does not interact with social class, meaning the degree of association (measure of dependency) between response and treatment is the same for both categories of the social class classification. In the following discussion some elementary concepts of interaction will be presented.

Interaction and Summary

The formulation of a meaningful hypothesis of no interaction among the classification variables in a multi-way table is not as simple as one might expect. There have been several attempts to arrive at a logical and intuitively acceptable definition of interaction that could be derived from a wider framework of hypothesis formulation. The main lines of thought for "no interaction" hypothesis can be grouped into the following classifications:

1. The original definition due to Bartlett [2] and its extension.
2. The formulation of Darrock [13] and Roy and Kastenbaum [57] based on symmetrical functions of the cell probabilities.
3. Good's definition [22] based on maximum entropy and Goodman's modification [25].

The testing of mutual independence of classification variables may be regarded as testing for significance of "no first-order interaction." For the simplest case for defining "no second-order interaction," Bartlett [2] defined for a $2 \times 2 \times 2$ table "no second-order interaction" as implying:

$$H_0 : \frac{p_{111} p_{221}}{p_{121} p_{211}} = \frac{p_{112} p_{222}}{p_{122} p_{212}} . \qquad (2.22)$$

Note that (2.22) may also be written as either:

$$\frac{p_{111}p_{122}}{p_{121}p_{112}} = \frac{p_{211}p_{222}}{p_{221}p_{212}} \quad \text{or} \quad \frac{p_{111}p_{212}}{p_{112}p_{211}} = \frac{p_{121}p_{222}}{p_{122}p_{221}}. \tag{2.23}$$

The hypothesis (2.22) and the alternative forms in (2.23) give the equality of association between row and column classifications within the two categories of the depth classification, between column and depth classifications within the two categories of the row classification, and between row and depth classifications within the two categories of the column classifications, respectively. This definition becomes difficult to interpret and involves solution of lengthy interative equations when the number of the levels of the classification variables are extended.

Roy and Kastenbaum [57] derived a set of constraints implying no interaction for a three-way contingency table of the form:

$$\frac{p_{rcd}p_{ijd}}{p_{icd}p_{rjd}} = \frac{p_{rck}p_{ijk}}{p_{ick}p_{rjk}} \quad \text{where} \quad \begin{array}{l} i = 1, 2, \ldots, r-1, \\ j = 1, 2, \ldots, c-1, \\ k = 1, 2, \ldots, d-1. \end{array} \tag{2.24}$$

The constraints (2.24) were based on the fact that the two hypotheses

$$H_1 : p_{i.k} = p_{i..}\, p_{..k} \quad \text{for all } i \text{ and } k$$

and

$$H_2 : p_{ij.} = p_{i..}\, p_{.j.} \quad \text{for all } i, j \text{ and } k$$

do not usually imply the hypothesis

$$H : p_{ijk} = p_{i..}\, p_{.jk} \quad \text{for all } i, j \text{ and } k,$$

Darroch's [13] formulation of no interaction led him to define

$$p_{ijk} = u\, a_{jk}\, b_{ki}\, c_{ij} \quad \text{for all } i,\ j \text{ and } k \tag{2.25}$$

where

$$\sum_{k=1}^{d} a_{jk} = \sum_{i=1}^{r} b_{ki} = \sum_{j=1}^{c} c_{ij} = 1$$

and

$$u \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} a_{ij}\, b_{ki}\, c_{ij} = 1 .$$

The formulation of the "no interaction" hypothesis up to this point have been extensions of Bartlett's definition. Good's [22] and Goodman's [25] formulations of the no interaction hypotheses are entirely general and physical interpretation of their meanings are extremely difficult.

Ku and Kullback [37] have developed a method of determining the cell probabilities in a multi-way contingency table. Hence, this is equivalent to a goodness-of-fit test since knowing the probabilities in a multinomial distribution determines the distribution.

Their procedure for determining the probabilities is based on a definition of the no-interaction hypothesis (marginal totals must be given) similar to formula (2.25) given by Darroch [13]. The process is an interative technique estimating cell probabilities and the coefficients in the constraints under a tenable hypothesis ([37], p. 168).

It is often the case that a researcher needs to summarize the results of a higher order contingency table. It is important that one is aware of the assumption being made when contingency tables are

condensed and then analyzed, since the use of two-way tables to summarize multi-way classification data is a rather common practice.

A summary of important facts involving no interaction are as follows:

1. If there is no first-order interaction; i.e., independence of all classifications, then the information is contained in the first-order marginals in the sense that given these marginals, the complete table can be constructed to within sampling error.

2. If the first-order interaction is significant, but there is no second-order interaction, then the set of second-order marginals will be required to summarize the data adequately ([38], p. 184).

A direct consequence of this interpretation is that the analysis can be reduced to that of the set of marginal tables if there is no interaction of the same order.

Conclusion

The role of the row, column, and depth classifications in the various hypotheses can be permuted as the experimenter may desire.

There are alternative ways of stating the hypothesis in many cases and they are apt to be confusing, but most of them are carefully followed through by Ku and Kullback [37], Kullback [39] and Lewis [45]. The objective in formulating hypotheses in this paper is to obtain an exactly additive analysis. Lancaster [42] and Kullback [39] obtain additive components in their analyses; while Lancaster

partitions the total chi-square into additive components, Kullback uses the information theory approach to obtain exact additive components. The value of obtaining additive components is to allow the experimenter to test additional hypotheses.

CHAPTER III

ANALYSIS OF CATEGORICAL DATA USING INFORMATION THEORY

Introduction

Information in a technically defined sense was first introduced into statistics by R. A. Fisher in 1925 in his work on the theory of estimations. According to Kullback [39], Fisher defined the information contained in a random sample of size n taken from a population with probability density function $f(x;\theta)$ as

$$I = n\,E\left\{\left[\frac{\partial \log_e f(x;\theta)}{\partial \theta}\right]^2\right\}.$$

Shannon and Wiener, independently, published in 1948 works describing logarithmic measures of information for use in communication theory. These papers stimulated a tremendous amount of study in engineering circles on the subject of information theory [39].

Information theory is a branch of the mathematical theory of probability and mathematical statistics. As such, it can be and is applied in a wide variety of fields. The subject of this exposition is of logarithmic measures of information and their application to the testing of statistical hypotheses in contingency tables.

In this chapter a consistent and simple approach based on the principles of information theory is used in developing the various test procedures for categorical data and the results are analyzed in the form of a component of information table.

Five examples are given illustrating the computation of the test statistic and the construction of the component of information table for testing each of several possible hypotheses. The procedures proposed depend on the use of a minimum discrimination information statistic (m.d.i.s.) and its asymptotic distribution properties. The examples will also be used to illustrate the conceptual simplicity of this approach to the statistical analysis of contingency tables. The calculation of the minimum discrimination information statistic, denoted by 2I, involves the basic operations of addition and subtraction; when a tabulation of $n \log_e n$ is available. However these calculations need to be carried through using more significant digits than the Pearson chi-square statistic.

Definitions

The minimum discrimination information statistic is based in principle on a technical meaning of information. Information in a technical sense is not radically different from the everyday meaning; it is merely more precise. Information can be gained about a matter in which we are to some degree uncertain; thus information may be defined as that which removes or reduces uncertainty. In statistics information is obtained by taking an observation or a sample from a population which is used to estimate parameters or to test hypotheses.

We shall, again, assume the multinomial model and will let the subscripted letter x represent cell frequencies instead of the subscripted letter n. This form is given to be consistent with the more standard notation used in discussions of information theory.

The multinomial is given here for future reference in defining the statistic of interest in this chapter:

$$f(x_1, x_2, x_3, \ldots, x_c) = \frac{N!}{\prod_{i=1}^{c} x_i!} \prod_{i=1}^{c} p_i^{x_i} \quad \text{where} \quad p_i > 0 \qquad (3.1)$$

$$\text{for} \quad i = 1, 2, \ldots, c, \quad \sum_{i=1}^{r} p_i = 1 \quad \text{and} \quad \sum_{i=1}^{r} x_i = N .$$

Definition 3.1: **Let H be a population of m partitions with probability density $f(x) = p_i$ for $x = x_1, \ldots, x_n$, then the mean information of** an observation selected at random from H is

$$I = E\left(\log_a \frac{1}{f(x)}\right) = \sum_{i=1}^{m} \left(\log_a \frac{1}{f(x_i)}\right) f(x_i) = \sum_{i=1}^{m} p_i \log_a \frac{1}{p_i} ,$$

or

$$I = - \sum_{i=1}^{m} p_i \log_a p_i . \qquad (3.2)$$

When $a = 2$, the form of the information in (3.2) is called the Shannon-Wiener measure of information ([1], p. 8). When logarithms to the base 10 are used, the mean information of an observation in (3.2) is termed a "Hartley" statistic and when the logarithms to the base e are used (3.2) is called a "nit" statistic. The base a is determined to facilitate the determination of the distribution of I.

In the absence of the proper knowledge of the system to determine a, the base e will be used and the distribution of 2I will be approximated. Therefore, unless otherwise indicated, the base e will be assumed throughout.

An intuitive example of information may be explained by considering a game in which two persons are playing. Suppose a person is thinking about a particular square on a checkerboard and the task of the other person is to discover which of the 64 possible squares it is. It can be shown that exactly six questions are necessary and sufficient to locate the square, if the questions are asked in the same manner with six answers of yes or no. For example, the first question might be "Is the square in the upper half of the board?" With the answer of either yes or no, the questioner has now limited the location of the unknown square to the 32 remaining squares. The second question could be "Is it in the left half of the remaining squares?" and so on for the other questions. Since the answers are of the yes or no form, there are two responses for each question and altogether $2^6 = 64$ different responses. For this set of responses call it H, a relation of $m = 2^I$ is suggested; where m is the number of equally likely responses from which a choice is made and I is the amount of uncertainty or information. Now, if $m = 2^I$, then $I = \log_2 m$; thus information involves the logarithm of the number of responses. The responses are expressed in the form of probabilities for testing hypothesis. If the m outcomes are equally likely, each with probability $p_i = \frac{1}{m}$ for $i = 1, 2, \ldots, m$ then the information (answer to one question) of a response expressed in terms of probabilities is given by definition (3.1) using equation (3.2) we have

$$I = -\sum_{i=1}^{m} p_i \log_2 p_i = -\sum_{i=1}^{m} \frac{1}{m} \log_2 \frac{1}{m}$$

$$= -\frac{m}{m} \log_2 \frac{1}{m} = \log_2 m .$$

Thus applying definition (3.1) we note that if the population has m equally likely partitions each with probability $p_i = \frac{1}{m}$ for $i = 1, 2, \ldots, m$, then the mean information of an observation selected at random is given by

$$I = \log_a m = \log_a \frac{1}{p_i} .$$

Application of Information to Statistics

We shall now apply the definitions of information theory to the analysis of contingency tables. The development of information used in the rest of this chapter is patterned after the derivations of Kullback [39], but is less mathematical. The more mathematical treatment of information theory as given by Kullback is given in Appendix A.

Suppose we have a contingency table with r categories resulting from a sample taken from a population partitioned by a single classification variable. Consider the two simple hypotheses H_0 and H_1 which specify the value of each parameter as follows

$$H_0 : p_i = p_{0i} \quad \text{for} \quad i = 1, 2, \ldots, r \quad \text{where} \quad \sum_{i=1}^{r} p_{0i} = 1 \tag{3.3}$$

and

$$H_1 : p_i = p_{1i} \quad \text{for} \quad i = 1, 2, \ldots, r \quad \text{where} \quad \sum_{i=1}^{r} p_{1i} = 1$$

$$\text{and} \quad p_{1i} \neq p_{0i} \quad \text{for at least one } i .$$

We note the two hypotheses H_0 and H_1 each partitions the population and specifies the probability of an observation occurring in each cell. We would desire to be able to take observations from this population and gain some information as to which, if either, hypothesis is correct. By applying information theory we make the following definition.

Definition 3.2: The mean information per observation from the population hypothesized by H_0, for discriminating in favor of H_1 against H_0 is

$$I(H_1:H_0) = \sum_{i=1}^{r} p_{1i} \log \frac{p_{1i}}{p_{0i}} . \tag{3.4}$$

If we let O_N be the set of N observations obtained from a population with a multinomial distribution (3.1) then the amount of information obtained from the sample is given by the following definition.

Definition 3.3: The mean discrimination information for a random sample of N independent observations for discrimination in favor of H_1 against H_0 is

$$I(H_1:H_0;O_N) = N \sum_{i=1}^{r} p_{1i} \log \frac{p_{1i}}{p_{0i}} , \tag{3.5}$$

If the equation (3.5) is multiplied by 2 and the natural logarithm is used, then the distribution of $2I(H_1:H_0;O_N)$ is approximated by the chi-square distribution with $r-1$ degrees of freedom ([39], p. 113). We restate Definition 3.3 so that the mean discrimination information (3.5) is in the proper form for a statistic.

Definition 3.4: The mean discrimination information statistic for a random sample of size N for discrimination in favor of H_1 against H_0 is

$$2I(H_1:H_0;O_N) = 2N \sum_{i=1}^{r} p_{1i} \log \frac{p_{1i}}{p_{0i}} . \qquad (3.6)$$

For a random sample which is partitioned by three classification variables into rcd categories by a simple null hypothesis H_0 given by

$$H_0 : p_{ijk} = p_{0ijk} \quad \text{for} \quad i = 1, 2, \ldots, r ; \; j = 1, 2, \ldots, c ; \; \text{and}$$
$$k = 1, 2, \ldots, d ;$$

against a simple alternative hypothesis H_1 given by

$$H_1 : p_{ijk} = p_{1ijk} \quad \text{for} \quad i = 1, 2, \ldots, r ; \; j = 1, 2, \ldots, c ;$$
$$k = 1, 2, \ldots, d ; \; \text{and} \; p_{1ijk} \neq p_{0ijk} \; \text{for some cell}$$
$$(i, j, k),$$

the equation (3.6) becomes

$$2I(H_1:H_0;O_N) = 2N \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} p_{1ijk} \log \frac{p_{1ijk}}{p_{0ijk}} . \qquad (3.7)$$

The null hypothesis H_0, usually specifies a relationship among the parameters and in some cases specifies the value of each parameter. Since the null hypothesis usually specifies a general relationship among the parameters in the multinomial distribution and the alternative hypothesis is a negation of H_0, the sample values may be used to minimize the discrimination information statistic. Thus in the

case where H_0 specifies the parameters p_{ijk} for all i, j and k we are "speculating" that the population is of the form

$$f(x) = \frac{N!}{x_{111}!x_{112}!\cdots x_{rcd}!} p_{111}^{x_{111}} p_{112}^{x_{112}} \cdots p_{rcd}^{x_{rcd}}$$

where x represents the sampled data.

By applying Theorem 3 in Appendix A, we know the minimum of the discrimination information statistic is obtained by using the best unbiased sample estimates for the parameters p_{1ijk} in the statistic (3.7). We are really estimating the probabilities from the sample for the distribution of the alternative hypothesis H_1. The objective is to obtain the smallest possible value for the statistic (3.7) so that if it is "sufficiently large" this would give us evidence that the sample does not resemble the distribution under the null hypothesis.

It is supposed that the sample, properly obtained, "resembles" the population. Thus, the population parameters under the alternative hypothesis are replaced by the best unbiased estimates based on the sample. The minimum of the test statistic (3.7) for a random sample, O_N, of size N would become in the above case

$$2I(H_1:H_0;O_N) = 2N \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} \frac{x_{ijk}}{N} \log \frac{x_{ijk}}{N p_{0ijk}} \qquad (3.8)$$

where x_{ijk}/N for all i, j and k are the best unbiased estimates of the parameters p_{ijk} for the composite alternative hypothesis. The statistic is distributed asymptotically as a chi-square random variable with rcd - 1 degrees of freedom.

If the null hypothesis does not specify the values of the parameters, but describes a relationship among them such as the hypothesis that the three classification variables are independent, then a degree of freedom is lost for each parameter estimated. To illustrate that the parameters not specified by the composite null hypothesis are estimated from the sample, a "hat" is placed over the I; that is,

$$2\hat{I}(H_1 : H_0; O_N) = 2 \sum \sum \sum x_{ijk} \log \frac{N^2 x_{ijk}}{x_{i..} x_{.j.} x_{..k}} \qquad (3.9)$$

where the parameters $p_{i..}$, $p_{.j.}$, and $p_{..k}$ are estimated by

$$\hat{p}_{i..} = \frac{x_{i..}}{N}, \quad \hat{p}_{.j.} = \frac{x_{.j.}}{N} \quad \text{and} \quad \hat{p}_{..k} = \frac{x_{..k}}{N},$$

respectively, for all i, j and k under the null hypothesis and the parameters p_{ijk} for all i, j, and k are estimated by $\hat{p}_{ijk} = \frac{x_{ijk}}{N}$ under the alternative hypothesis. Degrees of freedom are lost only for estimating those parameters not specified by the null hypothesis; thus for the statistic (3.9) we have the following degrees of freedom

$$rcd-1 - (r-1) - (c-1) - (d-1) = (r-1)(c-1)(d-1).$$

The following examples are given to illustrate primarily the computational procedure for the minimum discrimination information statistic. The first example represents the case where a sample is taken from a population partitioned into two categories or cells by one classification variable. In this example, we will assume the null hypothesis specifies the probabilities of the two categories.

Example 3.1: Suppose that in a sequence of 55 independent tosses of a coin we observe 37 heads and 18 tails. Let p_1 = probability of a head and p_2 = probability of a tail on a single toss of the coin. To test the null hypothesis that the coin is unbiased, formulated symbolically as $H_0: p_1 = p_2 = 1/2$, we compute the minimum discrimination information statistic as follows

$$2I = 2 \Sigma x_i \log \frac{x_i}{N p_i}$$

$$= 2x_1 \log x_1 + 2x_2 \log x_2 - 2N \log N + 2N \log 2 \qquad (3.10)$$

since $p_1 = p_2 = 1/2$ under the null hypothesis. From the data one may note that $x_1 = 37$, $x_2 = 18$ and $N = 55$. Substitution of these quantities into (3.10) gives

$$2I = 3(37 \log 37) + 2(18 \log 18) - 2(55 \log 55) + 2(55 \log 2)$$

$$= 6.700 .$$

Since $2I$ is distributed, approximately, as a chi-square random variable with one degree of freedom; the critical level $\hat{\alpha} \approx .01$. Comparing the statistic $2I$ with the statistic

$$T = \sum_{i=1}^{2} \frac{(x_i - E_i)^2}{E_i} ,$$

where $E_i = N/2 = 27.5$, the calculated value of T is 6.55. Comparing T with the chi-square distribution with one degree of freedom $\hat{\alpha} \approx .0108$. Thus, the two statistics give similar results.

The next two examples illustrate tests for independence and identical distributions, respectively, in two-way contingency tables.

Example 3.2: In an investigation of the nature and consequences of social stratification in a small mid-western community, it was found that the members of the community divided themselves into four social classes ([58], p. 177). The research centered on the correlates of this stratification among the youth of the community and one of the predictions was that adolescents in the different social classes would enroll in different curricula at the high school. A sample of 390 high school students were classified by the social class to which their family belongs and by the curriculum in which they are enrolled.

We are assuming in this example that we have a random sample from a single population partitioned by two classification variables. The data is given in Table III.

TABLE III

FREQUENCY OF ENROLLMENT FROM FOUR SOCIAL CLASSES IN THREE ALTERNATIVE HIGH SCHOOL CURRICULUMS

Curriculum \ Class	I	II	III	IV	Totals $x_{i\cdot}$
College Preparatory	23	40	16	2	81
General	11	75	107	14	207
Commercial	1	31	60	10	102
Totals $x_{\cdot j}$	35	146	183	26	390

The null hypothesis is

$$H_0 : p_{ij} = p_{i.}\, p_{.j} \quad \text{for all } i \text{ and } j ;$$

that is, the curriculum a student pursues is independent of the social class. The class of alternatives is given by

$$H_1 : p_{ij} \neq p_{i.}\, p_{.j} \quad \text{for some } i \text{ and } j .$$

The null hypothesis of independence does not specify $p_{i.}$ for $i = 1, 2, 3$ nor $p_{.j}$ for $j = 1, 2, 3, 4$. The minimum discrimination information statistic is given by

$$2I = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} \log \frac{x_{ij}}{N p_{i.}\, p_{.j}}$$

where the best unbiased estimates of $p_{i.}$ and $p_{.j}$ are given by $x_{i.}/N$ and $x_{.j}/N$, respectively, for $i = 1, 2, 3$ and $j = 1, 2, 3, 4$. Thus, the minimum discrimination information statistic becomes

$$2\hat{I} = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} \log \frac{x_{ij}}{N \frac{x_{i.}}{N} \frac{x_{.j}}{N}}$$

$$= 2N \log N + 2 \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} \log x_{ij} - 2 \sum_{i=1}^{r} x_{i.} \log x_{i.}$$

$$- 2 \sum_{j=1}^{c} x_{.j} \log x_{.j} . \qquad (3.11)$$

Based on the information in Table III, we find

$$2N \log N = 2(390 \log 390) = 4653.59446 ;$$

$$2 \sum_{i=1}^{3} \sum_{j=1}^{4} x_{ij} \log x_{ij} = 3055.77464 ;$$

$$2 \sum_{i=1}^{3} x_{i.} \log x_{i.} = 383.14080 ;$$

and

$$2 \sum_{j=1}^{4} x_{.j} \log x_{.j} = 3780.38044 .$$

The statistic $2\hat{I}$ given by (3.11) is approximated by the chi-square distribution with $(r-1)(c-1) = 6$ degrees of freedom. The critical level of the statistic $2\hat{I} = 65.6$ is much less than .001. The statistic (1.7) has a calculated value of 69.2 and is distributed, asymptotically, as chi-square with 6 degrees of freedom. The critical level $\hat{\alpha}$ is also much less than .001.

Example 3.3: Suppose, now the data given in Table III gives the results of taking a random sample from each of the four social classes discussed in Example 3.2 where each social class is partitioned into three categories by the curriculum classification. Assuming the samples are mutually independent, the objective is to determine whether or not the four populations of social classes are identically distributed. Table III is presented below transposed so that the notation developed on page 9 in Chapter I corresponds to the statements of this example.

TABLE IV

SOCIAL CLASS VERSUS CURRICULUMS

Curriculum / Class	General Preparatory	General	Commercial	Totals N_i
I	23	11	1	35
II	40	75	31	146
III	16	107	60	183
IV	2	14	10	26
Totals $x_{\cdot j}$	81	207	102	390

In this example we would like to consider the hypotheses given by

$$H_0 : p_{1j} = p_{2j} = p_{3j} = p_{4j} \quad \text{for all} \quad j = 1, 2, 3$$

(the samples are from the same population)

and

$$H_1 : p_{ij} \neq p_{i'j} \quad \text{for some} \quad i \neq i' \quad \text{and} \quad j = 1, 2, \ldots, c$$

(the samples are from different populations).

An alternate form of these hypotheses may be stated for notational convenience as

$$H_0 : p_{ij} = p_j \quad \text{for all } i \text{ and } j \tag{3.13}$$

and

$$H_1 : p_{ij} \neq p_j \quad \text{for some } i \text{ and } j .$$

The discrimination information statistic for the samples is given by the expression

$$2 \sum_{i=1}^{4} N_i \sum_{j=1}^{3} p_{ij} \log \frac{p_{ij}}{p_j} , \tag{3.14}$$

where N_i is the sample size and corresponds to a fixed value of $x_{i\cdot}$. The best unbiased estimates of p_{ij} for discriminating against H_0 to minimize the discrimination information statistic in (3.14) is given by $\hat{p}_{ij} = x_{ij} / N_i$. The minimum discrimination information statistic is given by

$$2I = 2 \sum_{i=1}^{4} \sum_{j=1}^{3} x_{ij} \log \frac{x_{ij}}{N_i p_j} . \tag{3.15}$$

The null hypothesis (3.13) does not state the values of the parameters p_j, $j = 1,2,3$, therefore we estimate the parameters from the sample and they are given by

$$\hat{p}_j = \frac{x_{\cdot j}}{N} \quad \text{for} \quad j = 1,2,3 .$$

The minimum discrimination information statistic for testing hypothesis (3.13) becomes

$$2\hat{I} = \sum_{i=1}^{4} \sum_{j=1}^{3} x_{ij} \log \frac{N x_{ij}}{N_i x_{\cdot j}} , \tag{3.16}$$

which is distributed asymptotically as a chi-square random variable with $(r-1)(c-1)$ degrees of freedom. Equation (3.16) may be simplified for computation purposes to give

$$2\hat{I} = 2 \sum_{i=1}^{4} \sum_{j=1}^{3} x_{ij} \log x_{ij} - 2 \sum_{j=1}^{3} x_{\cdot j} \log x_{\cdot j} + 2N \log N$$

$$- \sum_{i=1}^{4} N_i \log N_i \, . \qquad (3.17)$$

The computations are given by

$$\sum_{i=1}^{4} \sum_{j=1}^{3} x_{ij} \log x_{ij} = 1527.687331 \, ,$$

$$\sum_{j=1}^{3} x_{\cdot j} \log x_{\cdot j} = 1931.570398 \, ,$$

$$\sum_{i=1}^{4} N_i \log N_i = 1890.090224 \, ,$$

and

$$N \log N = 390 \log 390 = 2326.79722825 \, .$$

The calculated value is $2\hat{I} = 65.648$ with 6 degrees of freedom. The critical level $\hat{\alpha}$ is much less than .001.

The application of the minimum discrimination statistic will be extended to a three-way contingency table where we will discuss and present examples corresponding to the hypotheses discussed in Chapter II. The main purpose in using the minimum discrimination information statistic is to obtain an additive analysis similar to the analysis of variance for quantitative data. We will be using what is termed a component of information table to obtain a "complete" analysis of a contingency table. Since each entry in the component of information table represents the formulation of a tenable hypothesis, we will in the following section define some symbols and discuss

component of information tables for the hypotheses of independence, conditional independence and identical distributions of populations. For the component of information table of independence and conditional independence we will assume we have one sample from a population partitioned by three classification variables (three-dimensional contingency table). For the component of information table for identical distribution we will assume we have two or more independent samples from populations which are partitioned by two criterion variables to obtain a three-dimensional contingency table.

Independence of Classification Variables

The hypothesis of independence of classification variables based on a sample presented as a three-dimensional contingency table may be partitioned into additive components by noting that the

$$p_{ijk} = p_{i..}\, p_{.j.}\, p_{..k} \quad \text{for all } i, j \text{ and } k, \tag{3.18}$$

implies the conditions

$$p_{ijk} = p_{i..}\, p_{.jk} \quad \text{for all } i, j \text{ and } k \tag{3.19}$$

and

$$p_{.jk} = p_{.j.}\, p_{..k} \quad \text{for all } j \text{ and } k.$$

The converse of the above statement also follows; that is, if we have conditions (3.19) and (3.20), then (3.18) holds. The three classification variables are independent if and only if the row classification is independent of both the column and the depth classifications and the column and depth classifications are independent. We will let the

symbols $R \times C \times D$ denote the hypothesis represented by equation (3.18), $R \times (CD)$ denote the hypothesis represented by equation (3.19) and $C \times D$ denote the hypothesis represented by equation (3.20). When one analyzes a contingency table, the designation of row, column and depth classifications may be replaced by more descriptive terms in the application. In Table V we will give the component of information table for independence where the column denoted Component will symbolize the hypotheses being tested, Information will give the formulae for calculating the test statistics, and d.f. will give the degrees of freedom associated with each test. Note, that it is always the last component listed that has been partitioned into the additive components listed above it; thus the last row of a component of information table is analogous to the "Total" row of an analysis of variance table. The minimum discrimination information statistic for each component is additive and is distributed asymptotically as a chi-square random variable with the degrees of freedom indicated.

Using the table below one is able to permute the role of the classification variables to test other hypotheses, such as row and column classifications independence and the depth classification is independent of both the row and column classifications denoted by $R \times C$ and $D \times (RC)$, respectively.

TABLE V

COMPONENT OF INFORMATION FOR INDEPENDENCE

Component	Information	d.f.
C × D	$2 \sum_{j=1}^{c} \sum_{k=1}^{d} x_{\cdot jk} \log \frac{N x_{\cdot jk}}{x_{\cdot j\cdot} x_{\cdot\cdot k}}$	(c-1)(d-1)
R × (CD)	$2 \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log \frac{N x_{ijk}}{x_{i\cdot\cdot} x_{\cdot jk}}$	(r-1)(cd-1)
R × C × D	$2 \sum \sum \sum x_{ijk} \log \frac{N^2 x_{ijk}}{x_{i\cdot\cdot} x_{\cdot j\cdot} x_{\cdot\cdot k}}$	rcd - r - c - d + 2

Conditional Independence of Classification Variables

For conditional independence we note that

$$p_{ijk} = \frac{p_{i\cdot k} p_{\cdot jk}}{p_{\cdot\cdot k}} \quad \text{for all } i, j \text{ and } k \tag{3.21}$$

and

$$p_{i\cdot k} = p_{i\cdot\cdot} p_{\cdot\cdot k} \quad \text{for all } i \text{ and } k,$$

if and only if

$$p_{ijk} = p_{i\cdot\cdot} p_{\cdot jk} \quad \text{for all } i, j \text{ and } k.$$

For the component of information of conditional independence we will let the symbol $(R|D) \times (C|D)$ denote the hypothesis (3.21) that the row and column classification are independent given the depth classification. The equations given by (3.22) and (3.23) will be denoted as previously defined by $R \times D$ and $R \times CD$, respectively. Again, since the minimum discrimination information statistic is additive, a component of information table for conditional independence may be formed as given by Table VI.

TABLE VI

COMPONENT OF INFORMATION FOR CONDITIONAL INDEPENDENCE

Component	Information	d.f.
$R \times D$	$2 \sum_{i=1}^{r} \sum_{k=1}^{d} x_{i\cdot k} \log \frac{N x_{i\cdot k}}{x_{i\cdot\cdot} x_{\cdot\cdot k}}$	$(r-1)(d-1)$
$(R\|D) \times (C\|D)$	$2 \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk} x_{\cdot\cdot k}}{x_{i\cdot k} x_{\cdot jk}}$	$d(r-1)(c-1)$
$R \times (CD)$	$2 \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log \frac{N x_{ijk}}{x_{i\cdot\cdot} x_{\cdot jk}}$	$(r-1)(c\,d-1)$

Again the role of the classification variables may be permuted with a corresponding interchange of marginal totals appearing in the

computation formulae. Suppose we want to test an hypothesis of conditional independence given the column classification instead of the depth classification, then the component of information table would be as given in Table VII.

TABLE VII

COMPONENT OF INFORMATION FOR CONDITIONAL INDEPENDENCE GIVEN THE COLUMN CLASSIFICATION

Component	Information	d.f.
$R \times C$	$2 \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij\cdot} \log \frac{N x_{ij\cdot}}{x_{i\cdot\cdot} x_{\cdot j\cdot}}$	$(r-1)(c-1)$
$(R\|C) \times (D\|C)$	$2 \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk} x_{\cdot j\cdot}}{x_{ij\cdot} x_{\cdot jk}}$	$c(r-1)(d-1)$
$R \times (CD)$	$2 \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log \frac{N x_{ijk}}{x_{i\cdot\cdot} x_{\cdot jk}}$	$(r-1)(cd-1)$

This table is given because we will now consider an example illustrating each of the components of information tables presented thus far. The statistics in the information column need to be expanded to perform the calculations using the properties of logarithms similar to the proceedures in equation (3.11) in order to perform the calculations.

Example 3.4: A committee composed of a representative from each of the four major manufacturers of tape recorders has employed two consumer reporting agencies to test the product being marketed by the four manufacturers for defects in both electronic and mechanical components. Each testing agency, T_1 and T_2, is assigned a fixed proportion of the total production of units from each of the manufacturers M_1, M_2, M_3, and M_4, of which half of the units are sent to the electronics division to be tested for electronic defects D_1 and the remaining half are sent to the mechanical division to be tested for mechanical defects D_2. Each agency writes a report on each unit tested. At the end of the testing period the committee selects a random sample of size 124 reports from among those which report the existence of a defect. The data partitioned according to manufacturer, testing agency and type of defect are given in Table VIII.

The major objectives are to test the null hypothesis that the three classification variables are mutually independent $(M \times T \times D)$ and that the manufacturer is independent of both the testing agency and the type of defect $(M \times (TD))$. The minimum discrimination information statistic will be used to test these hypotheses. The following marginal totals and calculation are necessary. The critical level of each test will be in the fourth column of the following component of information tables.

TABLE VIII

TESTING FOR MANUFACTURING DEFECTS

Manufacturer	Defect	Test 1	Test 2
1	1	24	11
	2	8	13
2	1	7	2
	2	13	8
3	1	7	7
	2	2	7
4	1	5	3
	2	2	5

First-Order Marginals

Manufacturer	Test	Defect
$x_{1..} = 56$	$x_{.1.} = 68$	$x_{..1} = 66$
$x_{2..} = 30$	$x_{.2.} = 56$	$x_{..2} = 58$
$x_{3..} = 23$		
$x_{4..} = 15$		

Second-Order Marginals

Manufacturer x Test	Manufacturer x Defect	Test x Defect
$x_{11\cdot} = 32$	$x_{1\cdot1} = 35$	$x_{\cdot11} = 43$
$x_{12\cdot} = 24$	$x_{1\cdot2} = 21$	$x_{\cdot12} = 23$
$x_{21\cdot} = 20$	$x_{2\cdot1} = 9$	$x_{\cdot21} = 25$
$x_{22\cdot} = 10$	$x_{2\cdot2} = 21$	$x_{\cdot22} = 33$
$x_{31\cdot} = 9$	$x_{3\cdot1} = 14$	
$x_{32\cdot} = 14$	$x_{3\cdot2} = 9$	
$x_{41\cdot} = 7$	$x_{4\cdot1} = 8$	
$x_{42\cdot} = 8$	$x_{4\cdot2} = 7$	

Calculations needed from the data are:

$$\sum_{i=1}^{4}\sum_{j=1}^{2}\sum_{k=1}^{2} x_{ijk}\log x_{ijk} = 280.642, \qquad \sum_{i=1}^{4}\sum_{j=1}^{2} x_{ij\cdot}\log x_{ij\cdot} = 357.097,$$

$$\sum_{i=1}^{4}\sum_{k=1}^{2} x_{i\cdot k}\log x_{i\cdot k} = 359.061, \qquad \sum_{j=1}^{2}\sum_{k=1}^{2} x_{\cdot jk}\log x_{\cdot jk} = 429.705,$$

$$\sum_{i=1}^{4} x_{i\cdot\cdot}\log x_{i\cdot\cdot} = 440.193, \qquad \sum_{j=1}^{2} x_{\cdot j\cdot}\log x_{\cdot j\cdot} = 512.347,$$

$$\sum_{k=1}^{2} x_{\cdot\cdot k}\log x_{\cdot\cdot k} = 512.023,$$

and

$$N\log N = 124\log 124 = 597.715.$$

The component of information table for testing $(M \times T \times D)$ is given by Table IX which is based on the formulation given in Table V. Since the hypothesis of mutual independence would likely be rejected ($\hat{\alpha} \approx .01$), the tests given by partitioning the test statistic into additive components are performed. The hypothesis denoted by $(M \times (T\,D))$ sates that the manufacturer classification variable is independent of the composite test-defect classification variable having four categories which are the combinations of the two testing agencies with the two types of defects. Since $\hat{\alpha} \approx .05$, no clear-cut decision to reject or not reject H_0 would be reached at the .05 significance level. The null hypothesis that the testing agency is independent of the type of defect $(T \times D)$, which ignores the manufacturer classification variable, would most likely be rejected since $\hat{\alpha} \approx .015$.

TABLE IX

COMPONENT OF INFORMATION FOR INDEPENDENCE OF MANUFACTURER, TEST AND DEFECT CLASSIFICATIONS

Component	Information	d.f.	$\hat{\alpha}$
$T \times D$ [test × defect]	6.100	1	$\approx .015$
$M \times (T\,D)$ [manufacturer × test, defect]	16.918	9	$\approx .05$
$M \times T \times D$ [manufacturer × test × defect]	23.018	10	$\approx .01$

If we assume that the hypothesis $(M \times (T\,D))$ would be rejected then the test statistic of this test could be partitioned into components either as in Table X or as in Table IX. The hypothesis denoted by $(M|D) \times (T|D)$ in Table X states that the manufacturer classification is independent of the testing agency classification given an arbitrary category with the defect classification. This hypothesis would not be rejected at any of the more commonly used levels of significance since $\hat{\alpha} \approx .25$. However the hypothesis that the manufacturer $(M \times D)$, which ignores the testing agency classification, would most likely be rejected.

The analysis for conditional independence based on the calculations on page 57 and the analysis in Table VI yields Table X.

TABLE X

COMPONENT OF INFORMATION FOR CONDITIONAL INDEPENDENCE GIVEN THE DEFECT

Component	Information	d.f.	$\hat{\alpha}$
$M \times D$	9.120	3	$\approx .029$
$(M\|D) \times (T\|D)$	7.798	6	$\approx .25$
$M \times (T\,D)$	16.918	9	$\approx .05$

In Table XI it should be noted that the test denoted by $(M \times T)$ would not lead to a rejection of the null hypothesis which is in

agreement with the conclusion reached for the hypothesis $(M|D) \times (T|D)$ from Table X. That is, if we have concluded that the manufacturer and testing agency classification variables are independent given an arbitrary category of the defect classification, then it should follow that they are independent ignoring the defect classification.

Similarly, the calculated values on page 57 and the analysis in Table VII yields Table XI.

TABLE XI

COMPONENT OF INFORMATION FOR CONDITIONAL INDEPENDENCE GIVEN THE TEST

Component	Information	d. f.	$\hat{\alpha}$
$M \times T$	4.544	3	$> .25$
$(M\|T) \times (D\|T)$	12.374	6	$\approx .055$
$M \times (TD)$	16.918	9	$\approx .05$

Identical Distribution

Suppose we consider r independent random samples from r populations which are partitioned by two classification variables. We will consider in this three-dimensional contingency table the rows as being the independent random samples and the column and depth

classifications as the two criteria variables. The hypothesis of identical distributions is tested to determine if the two dimensional contingency tables generated by the samples are representative of identically distributed populations. Considering the hypothesis for identical distributions in Chapter III we note that

$$p_{ijk} = p_{\cdot jk} \quad \text{for all } i, j \text{ and } k \tag{3.24}$$

if and only if

$$p_{ij\cdot} = p_{\cdot j\cdot} \quad \text{for all } i \text{ and } j, \tag{3.25}$$

and

$$\frac{p_{ijk}}{p_{ij\cdot}} = \frac{p_{\cdot jk}}{p_{\cdot j\cdot}} \quad \text{for all } i, j \text{ and } k. \tag{3.26}$$

We will then use the above probability statements to obtain a component of information table for identical distributions. We will use the symbol $(C,D)I$ to denote the hypothesis that the r populations sampled are identically distributed which implies equation (3.24). The equation given by (3.25) is implied by the hypothesis that the r populations partitioned only by the column classification are identically distributed. This hypothesis will be denoted by $C(I)$ and has the effect of completely ignoring the presence of a depth classification. The equation (3.26) may be restated as

$$p_{ijk} = \frac{p_{\cdot jk}\, p_{ij\cdot}}{p_{\cdot j\cdot}} \quad \text{for all } i, j \text{ and } k, \tag{3.27}$$

which is the conditional hypothesis that the depth classifications are identically distributed, given the column classification among the r

independent samples. We will use the symbol $(D|C)I$ to denote the hypothesis expressed by equation (3.26). Each of the equations (3.24), (3.25) and (3.26) represent tenable hypotheses and the hypothesis $(C, D)I$ may be partitioned into additive components as shown in Table XII. The minimum discrimination information statistic given for each component is distributed asymptotically as a chi-square random variable with the indicated degrees of freedom.

TABLE XII

COMPONENT OF INFORMATION FOR IDENTICAL DISTRIBUTIONS

Component	Information	d.f.
$(C)I$	$2 \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij\cdot} \log \frac{N x_{ij\cdot}}{x_{i\cdot\cdot} x_{\cdot j\cdot}}$	$(r-1)(c-1)$
$(D\|C)I$	$2 \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk} x_{\cdot j\cdot}}{x_{ij\cdot} x_{\cdot jk}}$	$c(r-1)(d-1)$
$(C, D)I$	$2 \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log \frac{N x_{ijk}}{x_{i\cdot\cdot} x_{\cdot jk}}$	$(r-1)(cd-1)$

Let us now consider an example illustrating the analysis of identical distributions of several populations.

Example 3.5: Suppose independent random samples of items were obtained from each of four manufacturers where each population of items is partitioned by the two classification variables given in Example 3.4. Assume the four samples of size $N_i = x_{i..}$ for $i = 1, 2, 3, 4$, respectively, given the data shown in Table VIII. The procedure of analysis given in Table XII applied to the calculations given on page 57 yields Table XIII.

TABLE XIII

COMPONENT OF INFORMATION FOR IDENTICAL DISTRIBUTION OF MANUFACTURERS

Component	Information	d.f.	$\hat{\alpha}$
(D)I [defect]	4.544	3	$\approx .22$
(T|D)I [test|defect]	12.374	6	$\approx .05$
(T,D)I [test, defect]	16.918	9	$\approx .05$

The analysis again may be changed to some extent by permuting the role of the test and defect classifications. However, the random samples are from the manufacturers and we assumed the sample size was determined before the sampling was performed. For the interpretation in this example we cannot permute the role of the rows of the contingency table with either of the classification variables.

Summary

The analyses presented above may be extended to higher ordered contingency tables; however, these procedures for analysis are not complete in terms of all the various hypotheses that may be tested. Kullback, Kupperman, and Ku ([41], p. 575), claim the above procedures of analysis are simpler in statistical practice than the techniques based on the chi-square statistic. However, there is no theoretical reason to prefer the chi-square statistic over the minimum discrimination information statistic except one of taste and convenience ([41], p. 576). The utility of the minimum discrimination information statistic lies in its additivity and computational properties. The partitioning of the total component of information into several additive components is similar to the analysis of variance. Each component of the information in the table provides a minimum discrimination information statistic whose distribution is approximated by the chi-square distribution with the appropriate degrees of freedom ([40], p. 218).

CHAPTER IV

SMALL, ZERO AND MISSING FREQUENCIES IN CONTINGENCY TABLES

Introduction

The various tests for the analysis of contingency tables are usually based on large sample theory. In particular the chi-square, the likelihood ratio, and the minimum discrimination information statistics are approximated by the chi-square distribution when the sample size is large.

We will make several references to the likelihood ratio statistic, denoted as $-2 \log_e \lambda$ in this chapter. The test procedure based on the likelihood ratio statistic is defined and illustrated for two-way contingency tables in Appendix B. Under the assumption of sampling from a multinomial distribution, the likelihood ratio statistic is identical to the minimum discrimination information statistic ([37], p. 114). Thus in the developments of the test statistic presented in this chapter the minimum discrimination information statistic will be used to be consistent with previous discussions, even though one would find that frequently the original development is in the terminology of the likelihood ratio statistic.

In this chapter some general procedures for analyzing contingency tables with small or zero frequency counts will be presented.

A procedure for estimating missing frequencies and some of the effects of misclassification will be noted.

Zero or Small Frequencies

The Pearson chi-square statistic T illustrated in Chapter II and the minimum discrimination information statistic $2\hat{I}$ for testing hypotheses involving multi-way contingency tables are large sample tests. Even though most experiments are designed so that the probability of an observed cell frequency being zero is quite small, such an occurrence will be observed occasionally. Empirical evidence suggests that the presence of zero frequencies tends to inflate the value of either test statistic ([36], p. 398). When the test statistic is increased the critical level is decreased. This smaller critical level would cause the experimenter to reject the null hypothesis more frequently than he should.

The chi-square approximation to the distribution of the statistic $2\hat{I}$ is based on the assumption that all parameters involved in the constraints of the null hypothesis are greater than zero. If we assume that the alternative hypothesis also includes only alternatives for which these parameters are greater than zero and that an observed frequency of zero is the result of an insufficiently large sample size, then one could infer that the evidence provided by the sample is that the probability of observing a zero frequency in a cell is greater than the probability of observing a nonzero frequency in that cell. Assuming the multinomial model, Ku [36] proposes that one unit be subtracted from the computed value of the minimum discrimination information statistic for each zero cell count observed.

When an observed frequency of zero occurs it is frequently the case that other cells will have small frequencies (less than five) which may also endanger the validity of the test as well. It is generally agreed that the use of the Pearson chi-square test usually requires that each cell have an expected cell frequency of at least five ([68], p. 217). However, it has been shown by Zehna ([69], p. 553), that the chi-square distribution still provides an "adequate" approximation to the distribution of the likelihood ratio statistic (or equivalently, the minimum discrimination information statistic) even for relatively small sample sizes in the presence of small cell frequencies. Thus when small frequencies occur, the minimum discrimination information statistic should be used instead of the classical chi-square statistic.

To illustrate Ku's correction for zero frequencies, let us recall Example 2.1 where a random sample was taken from a population partitioned by three classification variables. The data is presented in Table II and we note the cell with a zero count and the seven cells with counts of less than five. It is desired to test the hypothesis that the classification variables are mutually independent.

The preliminary calculations that are needed for the analysis presented in Table V are

$$\sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{3} x_{ijk} \log x_{ijk} = 218.1852165 ,$$

$$\sum_{i=1}^{3}\sum_{j=1}^{2} x_{ij\cdot} \log x_{ij\cdot} = 310.2271315 ,$$

$$\sum_{i=1}^{3}\sum_{k=1}^{3} x_{i\cdot k}\log x_{i\cdot k} = 239.3664246\,,$$

$$\sum_{j=1}^{2}\sum_{k=1}^{3} x_{\cdot jk}\log x_{\cdot jk} = 320.2225579\,,$$

$$\sum_{i=1}^{3} x_{i\cdot\cdot}\log x_{i\cdot\cdot} = 380.4475041\,, \qquad \sum_{j=1}^{2} x_{\cdot j\cdot}\log x_{\cdot j\cdot} = 421.11531168\,,$$

$$\sum_{k=1}^{3} x_{\cdot\cdot k}\log x_{\cdot\cdot k} = 390.286318\,, \qquad N\log N = 494.3245439\,.$$

It should be noted in the calculations above that $0\log 0$ is defined to be 0.

In the component of information table for independence, Table V, we note the minimum discrimination information statistic for $C \times D$ (social class $\times$ discipline) involves only marginal totals; however for the $R \times (CD)$ (recall $\times$ social class, discipline) and $R \times C \times D$ (recall $\times$ social class $\times$ discipline) hypotheses, we have the frequency for each cell involved in the computation of the test statistic. Thus, the value of one would be subtracted from the statistics for the components $R \times (CD)$ and $R \times C \times D$ to correct for the single occurrence of a zero frequency. Table XIV shows the corrected value of the test statistic for each of these components of information. The critical level for mutual independence in Example 2.1 is $\hat{\alpha} \approx .005$ using the test statistic (2.7).

For $r \times c$ contingency tables there are methods for correcting statistical tests with small frequency counts. Suguira and Ôtake [62] have made numerical comparisons of improved methods for testing the hypothesis of independence in a contingency table with small frequencies

TABLE XIV

COMPONENT OF INFORMATION FOR INDEPENDENCE CORRECTED FOR A ZERO FREQUENCY

Component	Information	d.f.	$\hat{\alpha}$
Social class x discipline	6.321	2	$\approx .046$
Recall x (Social class, discipline)	22.574	10	$\approx .015$
Recall x social class x discipline	28.895	12	$\approx .005$

by the exact method; that is, comparing of the test statistics with the exact probability distribution in the tests of independence. The exact probability distribution assumes the marginal totals are fixed [68] and is expressed by

$$P(x_{ij} \mid x_{i.}, x_{.j}) = \frac{\prod_{i=1}^{r} x_{i.}! \prod_{j=1}^{c} x_{.j}!}{x_{..}!} \prod_{i=1}^{r} \prod_{j=1}^{c} \frac{1}{x_{ij}!}, \qquad (4.1)$$

which is the probability of cell frequency x_{ij} given the row and column marginals, ($x_{i.}$ and $x_{.j}$). One of the techniques that is applicable to the general $r \times c$ contingency table is the corrected minimum discrimination information statistic. The correction proposed by Suguira and Ôtake involves a constant K, so that the test statistic is $2K\hat{I}$, where

$$K = 1 - [6N(r-1)(c-1)]^{-1} \left(N \sum_{i=1}^{r} x_{i.}^{-1} - 1 \right) \left(N \sum_{j=1}^{c} x_{.j}^{-1} - 1 \right). \qquad (4.2)$$

The correction factor K for the minimum discrimination information statistic is obtained by calculating the first and the second conditional moments of the statistic $2K\hat{I}$ for given marginals in the exact distribution (4.1) and equating them to those of the chi-square with $(r-1)(c-1)$ degrees of freedom up to terms of order $1/N$; that is, assuming the statistic $-2K\hat{I}$ is approximated by the chi-square distribution with $(r-1)(c-1)$ degrees of freedom. Gart [21] made comparisons with the corrected minimum discrimination information test and an exact test with given marginals for 2×2 and 2×3 contingency tables with zero cell frequency, and concluded that one may use the corrected minimum discrimination information statistic with zero frequency counts as well as small frequency counts.

The use of the correction proposed by Suguira and Ôtake will be illustrated by altering Example 2.1. Assume we select a random sample of 49 boys from the working social class and classify the sample by the discipline and recall classifications discussed in Example 2.1. The data is then given in Table XV.

Using the data below we will test the hypothesis of independence given by (1.8). The value of K as given by the formula (4.2) for $N = 49$, $r = 3$, and $c = 3$ is

$$K = 1 - [6 \cdot 49 \cdot 2 \cdot 2]^{-1}\left[49\left(\frac{1}{15} + \frac{1}{13} + \frac{1}{21} - 1\right) \quad 49\left(\frac{1}{22} + \frac{1}{10} + \frac{1}{17} - 1\right)\right]$$
$$= .93588 .$$

The test statistic for independence is given by

$$-2K\hat{I} = 2K \sum_{i=1}^{3} \sum_{j=1}^{3} x_{ij} \log \frac{N x_{ij}}{x_{i\cdot}\, x_{\cdot j}} = 13.134 .$$

The critical level $\hat{\alpha} \approx .10$ were the test statistic, $-2K\hat{I}$, has 4 degrees of freedom.

TABLE XV

ZERO FREQUENCY DATA

Recall	Discipline			Marginals $x_{i.}$
	Psycho.	Mixed	Corp.	
Positive	6	3	6	15
Zero	9	4	0	13
Negative	7	3	11	21
Marginals $x_{.j}$	22	10	17	

Missing Frequencies

Missing frequencies in the analysis of contingency tables can result from a number of situations in a study or experiment. In a paper by Watson [66] procedures are presented for estimating missing cell frequencies associated with a sample of unknown size taken from a population partitioned by two classification variables. The procedure is based on the maximum likelihood estimates generated from the frequencies which are available under the null hypothesis of independence. The maximum likelihood estimates in such a two-way contingency table are found from the likelihood function subject to the

constraints $\sum_{i=1}^{r} p_{i.} = 1$ and $\sum_{j=1}^{c} p_{.j} = 1$. It is assumed the observed cell frequencies represent a sample from a multinomial population with parameters p_{ij}. Under the hypothesis of independence the parameters may be written as

$$p_{ij} = \frac{p_{i.}\, p_{.j}}{1 - p_{u.}\, p_{v.}} \quad \text{for } i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, c;\ (i, j) \neq (u, v)$$

where (u, v) is the missing cell and the total of the available frequencies is denoted by N'. This procedure is similar to the development of the maximum likelihood estimates given in Appendix B, except for the constraints on the parameters. The null hypothesis that the two classification variables are independent implies H_0 is given by

$$H_0 : p_{ij} = \frac{p_{i.}\, p_{.j}}{1 - p_{u.}\, p_{.v}} \quad \text{for all } (i, j) \neq (u, v)$$

and the alternative hypothesis is given by

$$H_1 : \text{not } H_0 .$$

The formula for estimating the missing frequency in cell (u, v) of a two dimensional contingency table is given by

$$x_{uv} = \frac{x_{u.}\, x_{.v}}{N' - x_{u.} - x_{.v}}, \tag{4.3}$$

where $x_{u.}$ and $x_{.v}$ are the row and column marginal totals and N' is the total of the recorded frequencies. With the frequency count

missing in cell (u, v), the marginal totals $x_{u.}$ and $x_{.v}$ are obtained by omitting the unknown cell. The formulae needed to estimate the unspecified parameters are given as follows

$$\hat{p}_{u.} = \frac{x_{u.} + x_{uv}}{N' + x_{uv}}$$

$$\hat{p}_{.v} = \frac{x_{.v} + x_{uv}}{N' + x_{uv}} \tag{4.4}$$

$$\left.\begin{aligned} \hat{p}_{i.} &= \frac{x_{i.}}{N' + x_{uv}} \quad , \quad i = 1, 2, \ldots, r \\ \hat{p}_{.j} &= \frac{x_{.j}}{N' + x_{uv}} \quad , \quad j = 1, 2, \ldots, c \\ \hat{E}_{ij} &= \frac{N' \hat{p}_{i.} \hat{p}_{.j}}{1 - \hat{p}_{u.} \hat{p}_{.v}} \end{aligned}\right\} (i, j) \neq (u, v)$$

The test statistic is

$$T = \Sigma' \frac{x_{ij}^2}{\hat{E}_{ij}} - N' \tag{4.5}$$

where Σ' is taken over all cells except the missing cell. The degrees of freedom associated with this test is given by $(r-1)(c-1) - 1$ ([66], p. 49); that is, one degree of freedom is lost in estimating the missing frequency.

The application of formulae (4.3), (4.4) and (4.5) will be illustrated for the data in Table XV where we will assume the sample size is unknown and the zero cell frequency which appears is actually

a missing cell frequency. The missing frequency occurs in cell (2,3), so $x_{2.} = 9 + 4 = 13$, $x_{.3} = 17$ and $N' = 49$. The calculations needed for determining the test statistic for the hypothesis of independence are

$$x_{23} = \frac{(13)(17)}{49 - 13 - 17} \approx 12$$

$$\left.\begin{aligned} \hat{p}_{2.} &= \frac{x_{2.} + x_{23}}{N' + x_{23}} = \frac{25}{61} \\ \hat{p}_{.3} &= \frac{x_{.3} + x_{23}}{N' + x_{23}} = \frac{29}{61} \end{aligned}\right\} \text{ probability estimates involving the missing cell}$$

$$\hat{p}_{1.} = \frac{x_{1.}}{N' + x_{23}} = \frac{15}{61} \qquad \hat{p}_{.1} = \frac{x_{.1}}{N' + x_{23}} = \frac{22}{61}$$

$$\hat{p}_{3.} = \frac{x_{3.}}{N' + x_{23}} = \frac{21}{61} \qquad \hat{p}_{.2} = \frac{x_{.2}}{N' + x_{23}} = \frac{10}{61} .$$

Using the above estimates of the parameters we calculate

$$\hat{E}_{ij} = \frac{N \hat{p}_{.i} \hat{p}_{.j}}{1 - \hat{p}_{2.} \hat{p}_{.3}}$$

for $i = 1,2,3$; $j = 1,2,3$; and $(i,j) \neq (2.3)$. The test statistic

$$T = \Sigma' \frac{x_{ij}^2}{\hat{E}_{ij}} - N' = .5713$$

with $(r-1)(c-1) - 1 = 3$ degrees of freedom. The critical level for this test is $\hat{\alpha} \approx .90$.

When there are several missing frequencies the correct analysis varies with their disposition. The above analysis can be extended using the formula (4.3) and the ordinary method of computing the chi-square statistic. The formula for estimating a missing frequency given by (4.3) is applied to each of the missing cells in succession. Once a missing frequency has been estimated, this estimate may be used where applicable in estimating the frequency of other cells. The iteration process is continued until the estimates obtained for each missing cell in the last two iterations differ by less than a pre-determined amount. The statistic would be calculated using the formula (4.5) and the number of degrees of freedom would be determined by the expression $(r-1)(c-1)$ less the number of cells with missing frequencies.

To illustrate the above discussion suppose (s,t) and (u,v) are the missing cells in a $r \times c$ contingency table. For a null hypothesis of independence, we will assume the observed cell frequencies total N' and represent a sample from a multinomial population with probabilities

$$p_{ij} = \frac{p_{i.}\, p_{.j}}{1 - p_{s.}\, p_{.t} - p_{u.}\, p_{.v}} \quad \text{for all } (i,j) \neq (s,t) \text{ and } (i,j) \neq (u,v).$$

To estimate the missing cell frequencies we will denote the k^{th} iterates of the cell frequencies (s,t) and (u,v) by $x_{st}^{(k)}$ and $x_{uv}^{(k)}$, respectively. There are two cases to consider, namely

(1) the two missing cells are not in the same row or column; i.e., $s \neq u$ and $t \neq v$.

(2) the two missing cells have a row or column in common; i.e., $s = u$ or $t = v$.

The estimates in case (1) for the cell frequencies are

$$x_{st} = \frac{x_{s.}\, x_{.t}}{N' - x_{s.} - x_{.t}} \quad ;$$

and

$$x_{uv} = \frac{x_{u.}\, x_{.v}}{N' - x_{u.} - x_{.v}} \quad ;$$

and for the parameters are

$$\hat{p}_{s.} = \frac{x_{s.}\, x_{st}}{N' + x_{st} + x_{uv}} \quad ; \qquad \hat{p}_{u.} = \frac{x_{u.} + x_{uv}}{N' + x_{st} + x_{uv}} \quad ;$$

$$\hat{p}_{.t} = \frac{x_{.t} + x_{st}}{N' + x_{st} + x_{uv}} \quad ; \qquad \hat{p}_{.v} = \frac{x_{.v} + x_{uv}}{N' + x_{st} + x_{uv}} \quad ;$$

and

$$\left.\begin{aligned} \hat{p}_{i.} &= \frac{x_{i.}}{N' + x_{st} + x_{uv}} \qquad i = 1, 2, \ldots, r \\ \hat{p}_{.j} &= \frac{x_{.j}}{N' + x_{st} + x_{uv}} \qquad j = 1, 2, \ldots, c \end{aligned}\right\} \begin{array}{l} (i,j) \neq (s,t) \\ \text{or} \\ (i,j) \neq (u,v) \end{array} .$$

The estimates in case (2) are more involved and we will illustrate the process by assuming the missing frequencies occur in the same row, with a similar technique if the missing frequencies are in the same column. Let the missing cells be denoted by (s,t) and (s,v) and we will estimate the frequency of cell (s,t) first. The first iterates

of the cell frequencies are given by

$$x_{st}^{(1)} = \frac{x_{s\cdot}\, x_{\cdot t}}{N' - x_{s\cdot} - x_{\cdot t}} \quad \text{where} \quad x_{s\cdot}\,,\ x_{\cdot t} \ \text{and} \ N'$$

are the initial marginals and total number of observations recorded and

$$x_{sv}^{(1)} = \frac{x_{s\cdot}^{(1)}\, x_{\cdot v}}{N' - x_{s\cdot}^{(1)} - x_{\cdot v}} \quad \text{where} \quad x_{s\cdot}^{(1)} = x_{s\cdot} + x_{st}^{(1)}\,.$$

The second iterates are found by the formulae

$$x_{st}^{(2)} = \frac{x_{s\cdot}^{(2)}\, x_{\cdot t}^{(1)}}{N' - x_{s\cdot}^{(2)} - x_{\cdot t}^{(1)}}$$

where

$$x_{s\cdot}^{(2)} = x_{s\cdot}^{(1)} + x_{sv}\,; \quad x_{\cdot t}^{(1)} = x_{\cdot t} + x_{st}^{(1)} \quad \text{and} \quad x_{sv}^{(2)} = \frac{x_{s\cdot}^{(3)}\, x_{\cdot v}^{(1)}}{N' - x_{s\cdot}^{(3)} - x_{\cdot v}^{(1)}}$$

where

$$x_{s\cdot}^{(3)} = x_{s\cdot}^{(2)} + x_{st}^{(2)}\,; \quad x_{\cdot v}^{(1)} = x_{\cdot v} + x_{sv}^{(1)}\,.$$

The k^{th} iterates of cell frequencies are

$$x_{st}^{(k)} = \frac{x_{s\cdot}^{(2k-2)}\, x_{\cdot t}^{(k-1)}}{N' - x_{s\cdot}^{(2k-2)} - x_{\cdot t}^{(k-1)}}$$

where

$$x_{s\cdot}^{(2k-2)} = x_{s\cdot}^{(2k-1)} + x_{sv}^{(k-1)}\,; \quad x_{\cdot t}^{(k-1)} = x_{\cdot t}^{(k-2)} + x_{st}^{(k-1)}\,;$$

and

$$x_{sv}^{(k)} = \frac{x_{s\cdot}^{(2k-1)}\, x_{\cdot v}^{(k-1)}}{N' - x_{s\cdot}^{(2k-1)} - x_{\cdot v}^{(k-1)}}$$

where

$$x_{s\cdot}^{(2k-1)} = x_{s\cdot}^{(2k)} + x_{st}^{(k)} \; ; \quad x_{\cdot v}^{(k-1)} = x_{\cdot v}^{(k-2)} + x_{sv}^{(k-1)} \; .$$

This process can be terminated when the difference between successive iterates is less than a desired quantity. The estimated probabilities at the end of the k^{th} iterate are

$$\hat{p}_{s\cdot} = \frac{x_{s\cdot} + x_{st}^{(k)} + x_{sv}^{(k)}}{N' + x_{st}^{(k)} + x_{sv}^{(k)}}$$

$$\hat{p}_{\cdot t} = \frac{x_{\cdot t} + x_{st}^{(k)}}{N' + x_{st}^{(k)} + x_{sv}^{(k)}}$$

$$\hat{p}_{\cdot v} = \frac{x_{\cdot v} + x_{sv}^{(k)}}{N' + x_{st}^{(k)} + x_{sv}^{(k)}}$$

$$\left.\begin{aligned} \hat{p}_{i\cdot} &= \frac{x_{i\cdot}}{N' + x_{st}^{(k)} + x_{sv}^{(k)}} \qquad i = 1, 2, \ldots, r \\ \hat{p}_{\cdot j} &= \frac{x_{\cdot j}}{N' + x_{st}^{(k)} + x_{st}^{(k)}} \qquad j = 1, 2, \ldots, c \end{aligned}\right\} \begin{array}{l} (i,j) \neq (s,t) \\ \text{or} \\ (i,j) \neq (u,v) \\ \text{where} \quad u = s \; . \end{array}$$

For either of the two cases the estimate of the expected values are given by

$$\hat{E}_{ij} = \frac{N' \hat{p}_{i.} \hat{p}_{.j}}{1 - \hat{p}_{s.} \hat{p}_{.t} - \hat{p}_{u.} \hat{p}_{.v}}$$

for all (i,j) such that $(i,j) \neq (s,t)$ and $(i,j) \neq (u,v)$ and the test statistic is given again by formula (4.5) with $(r-1)(c-1) - 2$ degrees of freedom.

An investigation of the effects of misclassification on the properties of the chi-square test reveals that misclassification reduces the power of the test ([50], p. 99). The power of a test is defined as the probability of rejecting the null hypothesis when a specified one of the alternatives included in the alternative hypothesis is true.

Conclusion

It appears from a review of the literature and studies on the analysis of contingency tables that the minimum discrimination information statistic is more reliable than the Pearson chi-square statistic for small cell frequencies. The likelihood ratio statistic is identical to the minimum discrimination information statistic for the multinomial distribution.

The corrected minimum discrimination information statistic may be used when small and/or zero frequency counts are present in two-dimensional contingency tables. The minimum discrimination information statistic with corrections for zero cell frequencies is also

an appropriate statistic for small frequency counts occurring in contingency tables.

For a method of analyzing contingency tables with missing cells in a section or a diagonal of a contingency table, Goodman [24] has a specialized and detailed discussion involving applications in biology. Estimation of missing cells or testing hypothesis of quasi-independence and interaction are discussed.

Some final comments about the Pearson chi-square test applied to contingency tables with small frequencies are: if any E_{ij} is less that one, or if more than 20% of the E_{ij} are less than five, then the approximation by the chi-square distribution may be poor. If all (or most) of the E_{ij} are nearly the same size, and if r and c are not too small, then Conover [10] indicates that the E_{ij} may be as small as one without endangering the validity of the test.

If some of the E_{ij} are too small, several cells may be combined to eliminate the E_{ij} which are too small. Just which cells should be combined is a matter of judgement. Generally, categories are combined only if they are similar in some respects, so that the hypothesis retain their meaning.

CHAPTER V

ANALYSIS OF VARIANCE FOR CATEGORICAL DATA

Introduction

A one-way classification of data originates from an experiment involving one independent variable and a response (dependent) variable. In this chapter we will be concerned with the analysis when the response variable is measured on a categorical or nominal scale.

Recall the parametric one-way classification design model where a random sample of size n_j is taken from treatment population j for $j = 1, 2, \ldots, t$; the populations are independent; each is normally distributed; and they have a common variance. Let y_{ij} = value of the i^{th} observation from population j. The objective is to test the null hypothesis

H_0 : Treatment populations have equal means,

H_1 : Not H_0 .

The one-way analysis of variance table is given by Table XVI.

TABLE XVI

ONE-WAY AOV TABLE

Source	d.f.	SS
Total	$n - 1$	$\sum_{i=1}^{n_t} \sum_{j=1}^{t} (y_{ij} - \bar{y}_{..})^2 = TSS$
Between Treatments	$t - 1$	$\sum_{j=1}^{t} n_j (\bar{y}_{.j} - \bar{y}_{..})^2 = BSS$
Within Treatments	$n - t$	$\sum_{i=1}^{n_j} \sum_{j=1}^{t} (y_{ij} - \bar{y}_{.j})^2 = WSS$

The test statistic is

$$F = \frac{BMS}{WMS} = \frac{BSS}{t-1} \div \frac{WSS}{n-t} .$$

Reject H_0 at the level α if $F_{calc} > F_{1-\alpha, t-1, n-t}$.

The above illustrates a well developed technique for handling quantitative data, namely partitioning the sums of squares to explain the variation in the data. This type of analysis supplies a measure of association between the response variable and the treatment (independent) variable which is used to estimate the proportion of the total variation in the response variable which is attributed to the predictor variable. This measure of association is given by

$$R^2 = \frac{BSS}{TSS} . \qquad (5.1)$$

Objectives

The main objective is to define a measure of variation for categorical data and to partition the total variation into an explainable component and an unexplainable component to test the hypotheses that the data came from identical populations. In terms of our multinomial model we need to investigate whether the c populations have the same multinomial probability structures.

The second purpose is to determine the degree of association between the independent variable and response variable. There are several procedures to calculate a number to represent a measure of association; however, none can be given a "proportion of the explained variation" interpretation for categorical data since the concept of partitioning variation has not been applied.

With these two objectives in mind, attention will be focused on the application of a general method of the one-way analysis of variance to categorical data or the equivalent two-way contingency table. The concept of variation for categorical data will be defined and the partitioning of the variation into additive components to give corresponding procedures for categorical data as the analysis of variance for quantitative data as described in the introduction for the parametric technique.

Assessing Variation in Categorical Data

Variation is very often thought of as a measure of deviation of a set of individual observations about their mean. For categorical data the mean is an undefined concept. The following procedure provides

a method for defining total variation for categorical data, then partitioning this variation into between group and within group sources of variation. First let us define variation within a sample of size n from a multinomial population in which the measurement scale is nominal. Let the n responses be $X_1, X_2, \ldots, X_n$ in which each X_i names one of r possible categories or classes. Define d_{ij} for all i and j such that

$$\begin{aligned} d_{ij} &= 1 \text{ if } X_i \text{ and } X_j \text{ name different categories} \\ &= 0 \text{ if } X_i \text{ and } X_j \text{ name the same category .} \end{aligned} \tag{5.2}$$

Then the variation for the categorical responses $X_1, X_2, \ldots, X_n$ is defined as

$$\begin{aligned} \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 &= \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} \\ &= \frac{1}{2n} \sum_{\substack{i=1 \\ i \neq j}}^{r} \sum_{j=1}^{r} n_i n_j \end{aligned} \tag{5.3}$$

where n_i is the number of observations identifying category i for $i = 1, 2, \ldots, r$ and $n = \sum_{i=1}^{r} n_i$. Since

$$n^2 = (n_1 + n_2 + \ldots + n_r)^2 = \sum_{i=1}^{r} n_i^2 + \sum_{\substack{i=1 \\ i \neq j}}^{r} \sum_{j=1}^{r} n_i n_j \tag{5.4}$$

implies that

$$\sum_{\substack{i=1 \\ i \neq j}}^{r} \sum_{j=1}^{r} n_i n_j = n^2 - \sum_{i=1}^{r} n_i^2 , \tag{5.5}$$

the formula for total variation may be written as

$$TSS = \frac{n}{2} - \frac{1}{2n} \sum_{i=1}^{r} n_i^2 \qquad (5.6)$$

upon substitution of (5.5) into (5.3).

Two lemmas are stated without proof ([46], p. 535) which exhibit properties one would reasonably expect of a measure of variation for categorical data.

Lemma 5.1: The variation of n categorical responses is minimized if and only if they all belong to the same category.

Lemma 5.2: The variation of n responses, where $n = rS + L$, $0 \leq L < r$, is maximized for any vector $(n_1, n_2, \ldots, n_r)$ of category counts such that L counts equal $S+1$, and $r-L$ counts equal S, i.e., the variation of n responses is maximized when the responses are distributed among the available categories as "evenly as possible."

Lemma 5.1 corresponds to the usual concept of the absence of variation when all of the responses are identical and Lemma 5.2 has no explicit counterpart for quantitative data.

To motivate the definition of variation further, note that if we have n quantitative measurements the sum of squares of deviation from the mean can be expressed solely as a function of the squares of the pairwise difference for all $\binom{n}{2}$ pairs. If $X_1, X_2, \ldots, X_n$ denotes the measurements and if

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

then

$$\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i^2 - 2X_iX_j + X_j)^2$$

$$= \frac{1}{2n}\left[n\sum_{i=1}^{n}X_i^2 - 2\sum_{i=1}^{n}X_i\sum_{j=1}^{n}X_j + n\sum_{j=1}^{n}X_j\right]$$

$$= \frac{1}{2n}\left[n\sum_{i=1}^{n}X_i^2 - 2n^2\bar{X}^2 + n\sum_{i=1}^{n}X_i^2\right]$$

$$= \frac{1}{2n}\left[2n\sum_{i=1}^{n}X_i^2 - 2n^2\bar{X}^2\right]$$

$$= \sum_{i=1}^{n}X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 . \qquad (5.7)$$

For quantitative data d_{ij} is interpreted as the deviation between X_i and X_j while for categorical data the concept of a deviation is meaningful only in terms of the presence or absence of a "difference." Thus, for categorical data, if $d_{ij} = X_i - X_j$ for all i and j, then

$$\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}d_{ij}^2$$

$$= \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}d_{ij}$$

$$= \text{TSS} .$$

Partitioning Total Variation

In Independent Samples

The general approach to categorical data which is proposed is to compute the total variation in the data and then partition this variation into specific components. The distributions of the various components are derived under the multinomial model and the analysis then proceeds in a direction dependent on this model ([46], p. 536).

Suppose a random sample of size $n_{\cdot j}$ has been taken from population j for $j = 1, 2, \ldots, c$; each population (sample) has been partitioned into r categories; and the c populations are independent. Let

n_{ij} = number of observations from sample j belonging to category i;

$n_{i\cdot} = \sum_{j=1}^{c} n_{ij}$ be the i^{th} row total;

and

$n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} = \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{c} n_{\cdot j}$ is the total number of observations taken.

The objective is then to test the null hypothesis:

$H_0 : p_{ij} = p_i$, for all i and j

(the c populations are identically distributed)

$H_1 : p_{ij} \neq p_i$ for some i and j.

Note that a total of $n_{i\cdot}$ responses have been identified as belonging to category i for $i = 1, 2, \ldots, r$. Thus, using equation (5.6), the total variation in the response variable or "total sum of squares" is given by

$$TSS = \frac{n}{2} - \frac{1}{2n} \sum_{i=1}^{r} n_{i\cdot}^2 . \tag{5.9}$$

The variation in the response variable within the j^{th} group (or sample) is then

$$WSS_j = \frac{n_{\cdot j}}{2} - \frac{1}{2n_{\cdot j}} \sum_{i=1}^{r} n_{ij}^2 \quad \text{for} \quad j = 1, 2, \ldots, c .$$

Adding over all samples, the "within sum of squares" is

$$WSS = \sum_{j=1}^{c} WSS_j = \frac{n}{2} - \frac{1}{2} \sum_{j=1}^{c} \frac{1}{n_{\cdot j}} \sum_{i=1}^{r} n_{ij}^2 . \tag{5.10}$$

Finally the between sample variation or "between sum of squares" is defined as the difference TSS - WSS. That is

$$BSS = \frac{1}{2} \sum_{j=1}^{c} \left[\frac{1}{n_{\cdot j}} \sum_{i=1}^{r} n_{ij}^2 \right] - \frac{1}{2n} \sum_{i=1}^{r} n_{i\cdot}^2$$

$$= TSS - WSS .$$

In the standard analysis of variance, BSS and WSS are independent and hence BSS and TSS are not. The following theorem states that just the opposite is true for the components of variation defined for categorical data. This indicates that we are at a point of

departure from the standard ANOVA theory. For the proof of the following theorem consult ([46], p. 537).

Theorem 5.1: Asymptotically with large $n_{\cdot j}$, TSS and BSS are independent under $H_0 : p_{ij} = p_i$ for all i and j.

The previously proposed method of partitioning categorical variation is referred to as a categorical analysis of variance, or CATANOVA. The test statistic is

$$C = \frac{(n-1)(r-1)\,BSS}{TSS} \tag{5.11}$$

for testing $H_0 : p_{ij} = p_i$. C is asymptotically approximated as a chi-square random variable with $(r-1)(c-1)$ degrees of freedom ([46], p. 540).

Measure of Association

We now turn to the problem posed in the introduction on measures of association for categorical data. The three components of variation defined enable us to define a measure of association between the grouping and the response variables which may be given a "proportion of the variation explained" interpretation. The measure of association is

$$R^2 = \frac{\left(\sum_{j=1}^{c} \frac{1}{n_{\cdot j}} \sum_{i=1}^{r} n_{ij}^2 \right) - \frac{1}{n} \sum_{i=1}^{r} n_{i\cdot}^2}{n - \frac{1}{n} \sum_{i=1}^{r} n_{i\cdot}^2}$$

$$= \frac{BSS}{TSS} \; . \tag{5.12}$$

This measure of association has the property that

$$R^2 = 0 \quad \text{if} \quad \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n} \quad \text{for} \quad i = 1, 2, \ldots, r, \quad j = 1, 2, \ldots, c,$$

i.e., if there is no association. $R^2 = 1$ if for each j, $j = 1, 2, \ldots, c$, there exist an i such that $n_{ij} = n_{.j}$, i.e., if there is perfect predictability. Otherwise $0 < R^2 < 1$. R^2 then is the proportion of total variation in the response variable which is accounted for by the knowledge of the grouping variable. Multiplying all entries in the contingency table by a positive constant leaves R^2 invariant.

Examples

Example 5.1: Suppose now the data given in Table XVII gives the results of taking a random sample from each of the four social classes discussed in Example 3.2 where each social class is partitioned into three categories by the curriculum classification. Assuming the samples are mutually independent, the objective is to determine whether or not the four populations of social classes are identically distributed.

We will use the CATANOVA to test the hypothesis (5.8). Using the data in Example 5.1 and the equation (5.9) yields the total variation given by

$$TSS = \frac{n}{2} - \frac{1}{2n} \sum_{i=1}^{r} n_{i.}^2 = 195 - \frac{1}{2(390)} (81^2 + 207^2 + 102^2)$$

$$= 195 - \frac{60294}{2(390)} = 195 - 76.685 = 118.315 .$$

TABLE XVII

EDUCATION ASPIRATIONS BY SOCIO-ECONOMIC LEVEL

Curriculum	Class I	Class II	Class III	Class IV	Totals $n_{i.}$
College Preparatory	23	40	16	2	81
General	11	75	107	14	207
Commercial	1	31	60	10	102
Totals $n_{.j}$	35	146	183	26	390

The "within sum of squares" is determined by using equation (5.10) and the calculation is given by

$$WSS = \frac{n}{2} - \frac{1}{2} \sum_{j=1}^{c} \frac{1}{n_{.j}} \Sigma n_{ij}^2$$

$$= \frac{390}{2} - \frac{1}{2}\left[\frac{1}{35}(23^2 + 11^2 + 1^2) + \frac{1}{146}(40^2 + 75^2 + 31^2)\right.$$

$$\left. + \frac{1}{183}(16^2 + 107^2 + 60^2) + \frac{1}{26}(2^2 + 14^2 + 10^2)\right]$$

$$= 195 - \frac{1}{2}[18.600 + 56.068 + 83.634 + 11.538]$$

$$= 195 - 84.920 = 110.080 .$$

The between sample variation is found by subtraction and it is given by

$$BSS = TSS - WSS = 118.315 - 110.080 = 8.235 .$$

A summary of the above computations is given below in Table XVIII.

TABLE XVIII

CATANOVA FOR EDUCATIONAL ASPIRATIONS

Source		SS	R^2	C
Between Classes	3	8.235	.070	54.46
Within Classes	8	110.08		
Total	11	118.315		

$$R^2 = \frac{BSS}{TSS} \approx .070$$

$$C = \frac{(n-1)(r-1)BSS}{TSS} = (n-1)(r-1)R^2 = (389)(2)(.070)$$

$$= 54.46 .$$

Comparing C with chi-square distribution with $(r-1)(c-1) = 6$ degrees of freedom, the observed critical level $\hat{\alpha} < .001$. The Pearson chi-square statistic for testing the hypothesis (5.8) is 69.2. The critical level for this value of the Pearson chi-square statistic with 6 degrees of freedom is much less than .001.

This example serves to illustrate a fact well known to researchers who work with large sets of data. Weakly related variables can

often exhibit very statistically significant dependencies.

In this study of independent samples from four social classes, $R^2 \approx .070$ implies that approximately 7% of the variation of educational aspiration is explained by the knowledge of the respondents social class.

Example 5.2: Suppose we have independent samples from three secondary schools in a lower socioeconomic metropolitan school district in which the number of yearly truancy reports for each sample is cited in Table XIX.

TABLE XIX

NUMBER OF TRUANCY REPORTS BY SCHOOL

Number of Truancy Reports By Individuals	Schools A	B	C	Totals $n_{i\cdot}$
None	400	300	100	800
1 - 3	100	50	25	175
4 - 6	50	25	25	100
More than 6	50	25	50	125
Totals $n_{\cdot j}$	600	400	200	1200

Using CATANOVA to test

$H_0: p_{ij} = p_i$ for all i and j versus $H_1: p_{ij} \neq p_i$ for some i and j,

the total variation is given by the following calculation

$$TSS = \frac{n}{2} - \frac{1}{2n} \sum_{i=1}^{3} n_{i\cdot}^{2}$$

$$= 600 - \frac{1}{2400} [800^2 + 175^2 + 100^2 + 125^2]$$

$$= 600 - \frac{216250}{2400} = 600 - 90.104$$

$$= 509.896 .$$

The "within sum of squares" is given by

$$WSS = \frac{n}{2} - \frac{1}{2} \sum_{j=1}^{4} \frac{1}{n_{\cdot j}} \sum_{i=1}^{3} n_{ij}^{2}$$

$$= 600 - \frac{1}{2} \left[\frac{1}{600} (400^2 + 100^2 + 50^2 + 50^2) \right.$$

$$+ \frac{1}{400} (300^2 + 50^2 + 25^2 + 25^2)$$

$$\left. + \frac{1}{200} (100^2 + 25^2 + 25^2 + 50^2) \right]$$

$$= 600 - 298.521 = 301.479 .$$

Again, the between sample variation is found by subtraction and is given by

$$BSS = TSS - WSS = 509.896 - 301.479 = 208.417 .$$

$$R^2 = \frac{BSS}{TSS} = \frac{208.417}{509.896} \approx .609$$

$$C = (n-1)(r-1)R^2 = (1199)(3)(.509) = 1830.873 .$$

Comparing C with the chi-square distribution with 6 degrees of freedom the critical level is $\hat{\alpha} < .001$. $R^2 \approx .509$ implies that approximately 50% of the variation of the number of truancy reports is explained by the knowledge of the respondent's school. A summary of the above computations is given below in Table XX.

TABLE XX

CATANOVA FOR TRUANCY REPORTS

Source	df	SS	R^2	C
Between Schools	2	208.417	.509	1830.873
Within Schools	9	301.479		
Total	11	509.896		

The Pearson chi-square statistic for testing the hypothesis $H_0 : p_{ij} = p_i$ for all i and j (that the three samples are drawn from the same population) is 72.45. Hence, the Pearson chi-square statistic with 6 degrees of freedom has a critical level $\hat{\alpha} < .001$.

Final Observation and Prospects

Empirical sampling experiments were run to see how well the approximate asymptotic null hypothesis theory holds for some specific cases. The purpose of these studies was to analyze how accurately

the mean and variance of the empirically generated test statistic matched its asymptotically approximated values for small and moderately large $n_{\cdot j}$ and for different $r \times c$ contingency tables and cell probability structures. The experiments under the multinomial model indicate for various r, c, probabilities, and small sample sizes the CATANOVA statistic is referenced quite well under H_0 by the chi-square distribution with $(r-1)(c-1)$ degrees of freedom ([46], p. 540).

In comparisons of the CATANOVA and the Pearson chi-square test statistic for independence in a two-dimensional contingency table, the tests are highly correlated with rank correlation coefficient applied to the mean of the test statistics. When there are two response categories $(r=2)$, regardless of the number of experimental groups (number of populations sampled), the CATANOVA and Pearson chi-square are identical ([46], p. 542).

Although the CATANOVA and chi-square test statistics have an identical reference distribution under the null hypothesis, the question arises as to their comparative behavior under various specific alternative hypothesis. General analytic results for the $r \times c$ table are not yet available, however in power studies with 3×2 contingency tables with various probabilities the power of CATANOVA exceeds the power of the chi-square in some cases and conversely in others ([46], p. 543). The 3×2 tables (three response categories for two experimental groups) were chosen because this is the simplest case for which the CATANOVA differs from the chi-square statistic.

Further research and the extension of the CATANOVA statistic is being extended and studied for higher dimensional contingency tables by Light and Margolin [46].

CHAPTER VI

SUMMARY AND PROSPECTS

The two main themes of this paper have been to present methods of analysis of categorical data that are analogous to the analysis of variance of quantitative data and to present topics to help the experimenter in the analysis of contingency tables with small, zero and missing frequencies.

Chapter I provides an introduction into the basic concepts and definitions of the probabilistic model for analyzing categorical data for one and two dimensional contingency tables. Chapter II develops some of the concepts of formulating hypotheses in terms of the probability model in three-dimensional contingency tables. Examples are presented to illustrate the application of hypothesis testing using the Pearson chi-square statistic to determine the critical level.

The third chapter illustrates the use of information theory applied to categorical data. The minimum discrimination information statistic is presented and used to test hypotheses in a component of information table. Component of information tables for hypotheses of mutual independence, conditional independence and identical distribution of samples are presented. The component of information table is analogous to the analysis of variance table for quantitative data. The main advantage in using the procedure associated with the component of information table is that the table presents an additive analysis of

the complete contingency table, rather than just a special segment of the analysis. There is no theoretical reason why the widely applied chi-square statistic should be preferred over the minimum discrimination information statistic. The minimum discrimination information statistic can be computed with fewer algebraic operations, when a tabulation of $n \log n$ is available. An $n \log n$ table is found in references [39] and [40]. The disadvantages of the minimum discrimination information statistic are that more significant digits must be carried through in the calculations and the chi-square statistic is a simpler mathematical function of the observations.

Chapter IV is a potpourri of results involving problem areas in the analysis of contingency tables. The purpose of this chapter is to present some of the elementary methods of handling the analysis when zero frequencies occur and for estimating missing frequencies under the hypothesis of independence. For small frequency counts it is recommended that the minimum discrimination information statistic be used for the test statistic.

In Chapter V an analysis of variance for categorical data is presented and a measure of association between the response and predictor variable is presented by estimating the per cent of variation of the response variable attributed to the predictor variable. The source of this chapter is a paper presented by Light and Margolin [46]. They are in the process of extending the procedure to multi-dimensional contingency tables.

We have presented a mathematical expository outline and development of information theory in Appendix A. This introductory summary of results are used to explain the procedures to obtain a

minimum for the discrimination information statistic in the analysis of contingency tables with the probability model. Appendix B contains a development of the likelihood ratio procedures for testing hypothesis under the multinomial distribution model. Methods of determining the maximum likelihood estimates are presented for estimating the parameters used in the likelihood ratio statistic. The maximum likelihood estimates are the best unbiased estimates for the parameters under the assumed multinomial model.

In conclusion, a few ideas of further studies and research are suggested. One could pursue the study of information theory to populations of other assumed models for example, data originating from Poisson processes or from normal populations. There is a need for research involving power studies dealing with zero of small frequencies in contingency table of higher dimension. Most of the power studies in the literature deal with 2×2 or $2 \times c$ contingency tables.

BIBLIOGRAPHY

[1] Attneave, Tred. Application of Information Theory to Psychology: A Summary of Basic Concepts, Methods, and Results. New York: Holt, Rinehart, and Winston, 1959.

[2] Bartlett, M. S. "Contingency table interactions." Journal Royal Statistics Society, 2 (1935), pp. 248-252.

[3] Bhapkar, V. P. "Some tests for categorical data." The Annals Mathematical Statistics, 32 (1961), pp. 72-83.

[4] Billingsley, P. "Statistical methods in Markov claims." The Annals Mathematical Statistics, 32 (1961) , pp. 12-40.

[5] Birch, M. W. "The detection of partial association, I: the 2 x 2 case." Journal Royal Statistics Society, B 26 (1964), pp. 313-324.

[6] Chernoff, H. "On the distribution of the likelihood ratio." The Annals Mathematical Statistics, 25 (1954), pp. 573-578.

[7] Chernoff, H. "Query: Degrees of freedom for chi-square." Technometrics, 9 (1967), pp. 489-490.

[8] Chernoff, H. and Lehmann, E. L. "The Use of maximum likelihood estimates in χ^2 tests for goodness of fit." The Annals of Mathematical Statistics, 25 (1954), pp. 579-686.

[9] Cochran, W. G. "The χ^2 test of goodness of fit." The Annals of Mathematical Statistics, 23 (1952), pp. 315-345.

[10] Conover, W. J. Practical Nonparametric Statistics. New York: John Wiley and Sons, Inc., 1971.

[11] Cramer, H. Mathematical Methods of Statistics. Princeton: Princeton University Press, 1946.

[12] Daly, C. "A simple test for trends in a contingency table." Biometrics, 18 (1962), pp. 114-119.

[13] Darroch, J. N. "Interactions in multi-factor contingency tables." Journal Royal Statistics Society, B 24 (1962), pp. 251-263.

[14] Darroch, J.N. and Silvey, S.D. "On testing more than one hypothesis." The Annals of Mathematical Statistics, 34 (1963), pp. 555-567.

[15] Deming, W.E. and Stephan, F.F. "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known." The Annals of Mathematical Statistics, 11 (1940), pp. 427-444.

[16] Diamond, E.L. "The limiting power of categorical data chi-square tests analogous to normal analysis of variance." The Annals of Mathematical Statistics, 34 (1963), pp. 1432-1441.

[17] Diamond, E.L. and Lilienfield, A.M. "Misclassification errors in 2 x 2 tables with one margin fixed: some further comments." American Journal of Public Health, 52 (1962), pp. 2106-2110.

[18] Edwards, A.W.F. "The measure of association in a 2 x 2 table." Journal Royal Statistics Society, A 126 (1963), pp. 109-114.

[19] Elashoff, R.M. and Afifi, A. "Missing values in multivariate statistics - I review of the literature." Journal American Statistics Association, 61 (1966), pp. 595-604.

[20] Gabriel, K.R. "Simultaneous test procedures for multiple comparisons on categorical data." Journal American Statistics Association, 61 (1966), pp. 1081-1096.

[21] Gart, J.J. "Alternative analyses of contingency tables." Journal Royal Statistics Society, B 28 (1966), pp. 164-179.

[22] Good, I.J. "On the estimation of small frequency in contingency tables." Journal Royal Statistics Society, B 18 (1956), pp. 113-124.

[23] Goodman, L.A. "On methods for comparing contingency tables." Journal Royal Statistics Society, A 126 (1963), pp. 94-108.

[24] Goodman, L.A. "The analysis of cross-classified data: Independence, quasi-independence, and intersections in contingency tables with or without missing entries." Journal American Statistics Association, 63 (1968), pp. 1091-1131.

[25] Goodman, L.A. "The multivariate analysis of qualitative data: Interactions among multiple classifications." Journal American Statistics Association, 65 (1970), pp. 226-256.

[26] Guenther, W.C. and Terrango, P.J. "A review of the literature on a class of coverage problems." The Annals Mathematical Statistics, 35 (1964), pp. 232-260.

[27] Guenther, W.C. and Thomas, P.O. "Some graphs useful for statistical inference." Journal American Statistics Association, 60 (1965), pp. 334-343.

[28] Haldane, J.B.S. "The first six moments of χ^2 for the n-fold table with n degrees of freedom when some expectation are small." Biometrika, 29 (1937), pp. 389-391.

[29] Haldane, J.B.S. "The cumulants and moments of the binomial distribution and the cumulants χ^2 for a $(n \times 2)$-fold table." Biometrika, 31 (1939), pp. 392-396.

[30] Haldane, J.B.S. "A problem in the significance of small numbers." Biometrika, 42 (1955), pp. 266-267.

[31] Hamdan, M.A. "The number and width of classes in the chi-square test." Journal American Statistics Association, 58 (1963), pp. 678-689.

[32] Hoyt, C.J., Krishnaiah, P.R. and Torrance, E.P. "Analysis of complex contingency data." Journal Experimental Education, 27 (1959), pp. 187-194.

[33] Ireland, C.T. and Kullback, S. "Contingency tables with given marginals." Biometrika, 55 (1968), pp. 179-188.

[34] Irwin, J.O. "Test of Significance for differences between percentages based on small numbers." Metron, 12 (1935), pp. 83-94.

[35] Kincaid, W.M. "The combination of $2 \times m$ contingency tables." Biometrics, 18 (1962), pp. 224-228.

[36] Ku, H.H. "A note on contingency tables involving zero frequencies and the 2I test." Technometrics, 5 (1963), pp. 398-400.

[37] Ku, H.H. and Kullback, S. "Interaction in Multidimensional Contingency Tables: An Information Theoretic Approach." Journal Research National Bureau of Standards, 72 (1968), pp. 159-199.

[38] Ku, H.H., Varner, Ruth N., and Kullback, S. "On the Analysis of Multidimensional Contingency tables." Journal American Statistics Association, 66 (1971), pp. 55-64.

[39] Kullback, S. Information Theory and Statistics. New York: John Wiley and Sons, Inc., 1959.

[40] Kullback, S., Kupperman, M. and Ku, H.H. "An Introduction of information theory to the analyses of contingency tables, with a table of $2n\ell nn$, $n = 1(1)\,10,000$." Journal Research National Bureau Standards, 66 (1962), pp. 217-243.

[41] Kullback, S., Kupperman, M. and Ku, H.H. "Test for Contingency Tables and Markov Chains." Technometrics, 4 (1962), pp. 573-608.

[42] Lancaster, H.O. "Complex contingency tables treated by the partition of χ^2." Journal Royal Statistics Society, 13 (1951), pp. 242-249.

[43] Lancaster, H.O. The Chi-Squared Distribution. New York: John Wiley, 1969.

[44] Leslie, P.H. "A simple method of calculating the exact probability in 2 x 2 contingency tables with small marginal totals." Biometrika, 42 (1955), pp. 522-523.

[45] Lewis, B.N. "On the analysis of interaction in multi-dimensional contingency tables." Journal Royal Statistics Society, 125 (1962), pp. 88-117.

[46] Light, Richard J. and Margolin, Barry H. "An Analysis of Variance for Categorical Data." Journal American Statistical Association, 66 (1971), pp. 534-544.

[47] Lindley, D.V. "The Bayesian Analysis of Contingency tables." The Annals Mathematics Statistics, 35 (1964), pp. 1622-1643.

[48] Miller, Daniel R. and Swanson, Guy E. Inner Conflict and Defense. New York: Henry Holt and Company, 1960.

[49] Mosteller, Frederick. "Association and Estimation in Contingency Tables." Journal American Statistical Association, 63 (1968), pp. 1-29.

[50] Mote, V.L. and Anderson, R.L. "An Investigation of the effects of misclassification on the properties of χ^2-tests analysis of categorical data." Biometrika, 52 (1965), pp. 95-110.

[51] Pearson, E.S. "The choice of Statistical tests illustrated on the interpretation of data classed in a 2 x 2 table." Biometrika, 34 (1947), pp. 139-167.

[52] Pearson, E.S. and Merrington, M. "2 x 2 tables, the power function of the test on a randomized experiment." Biometrika, 35 (1948), pp. 331-345.

[53] Peters, C.C. "The misuse of chi-square - a replay to Lewis and Burke." Psychological Bulletin, (1950), pp. 331-337.

[54] Plackett, R.L. "A note on interaction in contingency tables." Journal Royal Statistics Society, 24 (1962), pp. 162-166.

[55] Plackett, R.L. "The continuity correction in 2 x 2 tables." Biometrika, 51 (1969), pp. 327-337.

[56] Popham, W. James. Educational Statistics. New York: Harper and Row Publishers, 1967.

[57] Roy, S.N., and Kastenbaum, M.A. "A generalization of analysis of variance and multivariate analysis to data based on frequencies in qualitative categories or class intervals," Inst. Statist. Univ. North Carolina Mimeo Series, 131 (1955), p. 27.

[58] Siegel, Sidney. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill, 1956.

[59] Silvey, S.D. Statistical Inference. Baltimore: Penguin Books, 1968.

[60] Slater, P. "The factor analysis of a matrix of 2 x 2 tables." Journal Royal Statistics Society, 9 (1947), pp. 114-127.

[61] Smith, H.F. "On Comparing contingency tables." Philippine Statistician, 6 (1957), pp. 71-81.

[62] Sugiura, Nariaki and Ôtake, Masanovi. "Numerical comparison of improved methods of testing in contingency tables with small frequencies." Annals of the Institute of Statistical Mathematics, 20 (1968), pp. 505-517.

[63] Sutcliffe, J.P. "A general method of analysis of frequency data for multiple classification designs." Psychological Bulletin, 54 (1957), pp. 134-137.

[64] Vajda, S. "The algebraic analysis of contingency tables." Journal Royal Statistics Society, 106 (1943), pp. 333-342.

[65] Walsh, J.E. "Loss in test efficiency due to misclassification for 2 x 2 tables." Biometrics, 19 (1963), pp. 158-162.

[66] Watson, G.S. "Missing and mixed up frequency tables." Biometrics, 12 (1956), pp. 47-50.

[67] Wilks, S.S. "The likelihood test of independence in contingency tables." The Annals Mathematical Statistics, 6 (1935), pp. 190-196.

[68] Yates, F. "Contingency tables involving small numbers of the test χ^2 test." Journal Royal Statistics Society, 1 (1934), pp. 217-235.

[69] Zehna, Peter W. Probability Distribution and Statistics. Boston: Allyn and Bacon, 1970.

[70] Zehna, Peter W. and Barr Donald R. Probability. Belmont: Brooks-Cole Publishing Company, 1971.

APPENDIX A

INFORMATION THEORY

Definitions

Consider the probability space $(\Omega, \mathcal{C}, \mu_i)$ $i = 1, 2$ where Ω is the sample space, $\mathcal{C}$ is a σ-algebra of subsets of Ω, and μ_i, $i = 1, 2$ are probability measures defined on $\mathcal{C}$.

We assume the probability measures μ_1 and μ_2 are absolutely continuous with respect to one another, denoted $\mu_1 \equiv \mu_2$. Recall that μ_1 is absolutely continuous with respect to μ_2, $\mu_1 << \mu_2$, if $\mu_1(E) = 0$ for all $E \in \mathcal{C}$ whenever $\mu_2(E) = 0$. If λ is a probability measure such that $\lambda \equiv \mu_1$, $\lambda \equiv \mu_2$, then by the Radon-Nikodym theorem there exist functions $f_1(x)$ and $f_2(x)$, called generalized probability densities, unique up to sets of probability zero in λ, $0 < f_i(x) < \infty\ [\lambda]$, $i = 1, 2$, such that

$$\mu_i(E) = \int_E f_i(x)\, d\lambda(x)$$

$i = 1, 2$, for all $E \in \mathcal{C}$. The function $f_i(x)$ is called the Radon-Nikodym derivative, and we note the following equations

$$d\mu_i = f_i(x)\, d\lambda(x)$$

or

$$f_i(x) = \frac{d\mu_i}{d\lambda}, \quad i = 1, 2.$$

If H_1 and H_2 are statistical hypotheses and the set X is from the statistical population with probability measures μ_1 and μ_2, then it follows from Bayes' Theorem that

$$P(H_i \mid x) = \frac{P(H_i) f_i(x)}{P(H_1) f_1(x) + P(H_2) f_2(x)} \tag{1}$$

where $P(H_i)$, $i = 1, 2$, is the prior probability of H_i and $P(H_i \mid x)$ is the conditional probability of H_i given $X = x$. Note, since $f_i(x)$, $i = 1, 2$ is the Radon-Nikodym derivative, $f_i(x)$ is the conditional probability density at $X = x$ under the hypothesis H_i. From equation (1) for $i = 1, 2$ we can obtain the following equation

$$\frac{P(H_1 \mid x)}{P(H_2 \mid x)} = \frac{P(H_1) f_1(x)}{P(H_2) f_2(x)} .$$

Solving the latter equation for

$$\frac{f_1(x)}{f_2(x)}$$

we obtain the formula

$$\frac{f_1(x)}{f_2(x)} = \frac{P(H_1 \mid x)\, P(H_2)}{P(H_2 \mid x)\, P(H_2)} . \tag{2}$$

Now, we take the natural logarithm of (2) and get

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P(H_1 \mid x)}{P(H_2 \mid x)} - \log \frac{P(H_1)}{P(H_2)} . \tag{3}$$

The right hand side of the equation (3) is a measure of the difference between the logarithm of the odds in favor of H_1 after the observation of $X=x$ and logarithm of odds before the observation. This difference can be positive or negative and is considered to be the information resulting from the observation $X=x$. We define the logarithm of the likelihood ratio, $\log \frac{f_1(x)}{f_2(x)}$ as the information in $X=x$ for discrimination in favor of H_1 against H_2.

Definition 1: The information in an observation $X=x$ for discrimination in favor of H_1 against H_2 is

$$\log \frac{f_1(x)}{f_2(x)} .$$

Definition 2: The mean information for discrimination in favor of H_1 against H_2 given $x \in E \in \mathcal{C}$, for μ_1, is

$$I(1:2;E) = \frac{1}{\mu_1(E)} \int_E \log \frac{f_1(x)}{f_2(x)} \, d\mu_1(x)$$

$$= \begin{cases} \frac{1}{\mu_1(E)} \int_E f_1(x) \log \frac{f_1(x)}{f_2(x)} \, d\lambda(x) & \text{for } \mu_1(E) > 0 \\ 0 \quad \text{for } \mu_1(E) = 0 , \end{cases} \tag{4}$$

with $d\mu_1(x) = f_1(x)\, d\lambda(x)$. If E is the sample space Ω, then equation (4) becomes

$$I(1:2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} \, d\lambda(x) \tag{5}$$

since $\mu_1(\Omega) = 1$.

Theorem 1: $I(1:2)$ is additive for independent random events; that is for X and Y independent random variables under both H_1 and H_2,

$$I(1:2; E_1, E_2) = I(1:2; E_1) + I(1:2; E_2)$$

where $E_1, E_2 \in \mathcal{C}$ are events associated with the observations X and Y, respectively.

The following theorem is important and is needed to establish the form of the minimum of $I(1:2)$ used in the application of contingency table analysis. For a proof and discussion of this theorem refer to Kullback ([39], pp. 36-39). First we need a definition.

Definition 3: A set M of probability measures on $\mathcal{C}$ is called dominated if there exists a measure λ on $\mathcal{C}$, λ not necessarily a member of M, such that every member of the set M is absolutely continuous with respect to λ.

Theorem 2: If $f_1(x)$ and a given $f_2(x)$ are probability densities of a dominated set of probability measures, $Y = T(x)$ is a measurable statistic such that

$$\theta = \int T(x)\, f_1(x)\, d\lambda(x)$$

exists, and

$$m_2(t) = \int f_2(x)\, e^{t\,T(x)}\, d\lambda(x)$$

exists in some interval; then

$$I(1:2) \geq \theta t - \log m_2(t) = I(1^*:2), \tag{6}$$

$$\theta = \frac{d}{dt} \log m_2(t)$$

with equality in (6) if and only if

$$f_1(x) = f_2^*(x) = e^{t\,T(x)}\, f_2(x)/m_2(t) .$$

The underlying principle in using the minimum discrimination information in statistics is that $f_2(x)$ will be associated with the set of populations of the null hypothesis and $f_1(x)$ will range over the set of the alternative hypothesis. The sample values will be used to determine the resemblance between the sample, as a possible member of the set of populations of the alternative hypothesis, and the closest population of the set of populations of the null hypothesis by an estimate of the minimum discrimination information. The null hypothesis will be rejected if the estimated minimum discrimination is significantly large.

When the maximum likelihood estimates of the parameters for $f_2^*(x)$ are used, we denote $I(1:2)$ as $\hat{I}(1:2)$ (also $2\hat{I}$) and $2\hat{I}(1:2)$ is distributed as the likelihood ratio statistic $-2 \log \lambda$ ([39], pp. 94-97). Thus $2\hat{I}$ is the minimum discrimination information statistic used to test the null hypothesis H_2 against the alternative hypothesis H_1.

Applications to Multinomial Populations

We shall now undertake the application of the principles and results developed in the preceding sections to the analysis of samples from a multinomial distribution for testing statistical hypotheses.

Since the multinomial distribution is discrete, the hypotheses we will be concerned with are those involving contingency tables.

If we assume the multinomial density of the form

$$f_i(\mathbf{x}) = \frac{N!}{\prod_{j=1}^{c} X_j} \prod_{j=1}^{c} p_{ij}^{X_j},$$

where $\sum_{j=1}^{c} p_{ij} = 1$, and $\mathbf{x}$ represents the observation classified by the c categories of a single classification variable, then the measure λ is the counting measure and the integral is replaced by the summation symbol. We note

$$\mu_i(E) = \sum_E f_i(x) \quad i = 1, 2$$

where E is the set $\{\mathbf{x} = (x_1, x_2, \ldots, x_c) : \sum x_i = N\}$. Equation (5) becomes

$$I(1:2) = \sum f_1(x) \log \frac{f_1(x)}{f_2(x)} \tag{7}$$

which is the mean discrimination information per observation.

The following theorem is an important consequence of Theorem 2 and will be stated without proof. For the proof of the theorem see the discussion in ([39], pp. 111-112).

Theorem 3: The least informative distribution on the population partitioned by one classification into c cells with given expected values, for discrimination against the multinomial distribution $f_2(x)$, is the distribution $f_1^*(x)$ such that

$$E_{f_1^*(x)}(x_j) = \theta_j \qquad j = 1, 2, \ldots, c$$

and

$$\Sigma_E f_1^*(x) \log \frac{f_1^*(x)}{f_2(x)} \tag{8}$$

is a minimum is given by the distribution

$$f_1^*(x) = e^{\sum_{j=1}^{c} t_j x_j} f_2(x) \Big/ \left(\sum_{j=1}^{c} p_{2j} e^{t_j} \right)^N$$

$$= \frac{N!}{\prod_{j=1}^{c} x_j!} \prod_{j=1}^{c} p_{1j}^{x_j} \tag{9}$$

where

$$p_{1j}^* = p_{2j} e^{t_j} \Big/ \left(p_{21} e^{t_1} + \ldots + p_{2c} e^{t_c} \right) \qquad j = 1, 2, \ldots, c$$

the t_j's are real parameters, and

$$\theta_j = \left[\frac{\partial}{\partial t_j} \log \left(p_{21} e^{t_1} + p_{22} e^{t_2} + \ldots + p_{2c} e^{t_c} \right) \right] \tag{10}$$

An important point to note in Theorem 3 is that the least informative distribution $f_1^*(x)$ is a multinomial distribution.

The minimum discrimination information of a sample E with N observations is

$$I(1:2;E) \geq t_1\theta_1 + t_2\theta_2 + \ldots + t_c\theta_c - N \log \left(p_{21} e^{t_1} + \ldots + p_{2c} e^{t_c} \right)$$

$$= I(1^*, 2; E) \tag{11}$$

by applying the result of Theorem 2. Simplifying $I(1^*:2;E)$, we note equation (10) can be differentiated and solved for t_j for $j = 1, 2, \ldots, c$ where:

$$\theta_j = \frac{\partial}{\partial t_j}\left[\log\left(p_{21}\, e^{t_1} + p_{22}\, e^{t_2} + \ldots + p_{2c}\, e^{t_c}\right)\right]$$

$$= \frac{N\, p_{1j}\, e^{t_j}}{p_{21}\, e^{t_1} + \ldots + p_{2c}\, e^{t_c}} \quad .$$

Solving for t_j we have

$$t_j = \log \frac{\theta_j}{N\, p_{1j}} + \log\left(p_{21}\, e^{t_1} + \ldots + p_{2c}\, e^{t_c}\right)$$

and $I(1^*:2;E)$ becomes

$$I(1^*:2;E) = \theta_1 \log \frac{\theta_1}{N\, p_{21}} + \theta_2 \log \frac{\theta_2}{N\, p_{22}} + \ldots + \theta_c \log \frac{\theta_c}{N\, p_{2c}}$$

$$+ \sum_{j=1}^{c} \theta_j \log\left(p_{21}\, e^{t_1} + \ldots + p_{2c}\, e^{t_c}\right)$$

$$- N \log\left(p_{21}\, e^{t_1} + \ldots + p_{2c}\, e^{t_c}\right)$$

$$= \theta_1 \log \frac{\theta_1}{N p_{21}} + \theta_2 \log \frac{\theta_2}{N p_{22}} + \ldots + \theta_c \log \frac{\theta_c}{N p_{2c}} \quad . \quad (12)$$

Suppose we want to test the simple null hypothesis H_2 that the sample is from the population specified by

$$H_2 : p_j = p_{2j} \quad \text{for} \quad j = 1, 2, \ldots, c \quad \text{and} \quad \sum_{j=1}^{c} p_{2j} = 1$$

against the alternative hypothesis H_1 that the sample is from any other possible multinomial population. From the distribution in (9) we take the parameters to be the same as the best unbaised sample estimates, that is,

$$\hat{\theta}_j = N\left(\hat{p}_{1j}^{*}\right) = N \frac{x_j}{N} = x_j , \quad j = 1, 2, \ldots, c . \tag{13}$$

Using equation (12) we substitute equations (13) for the parameters θ_j, $j = 1, 2, \ldots, c$ and we get

$$2I = \sum_{j=1}^{c} x_j \log \frac{x_j}{N p_{2j}} . \tag{14}$$

Equation (14) is of the form used as the test statistic in Chapter III for a one dimensional contingency table.

APPENDIX B

LIKELIHOOD RATIO STATISTIC

We have made frequent reference of the likelihood ratio statistic throughout this paper. In the study of the analysis of contingency tables in probabilistic terms, we have assumed the multinomial model. Our objective is to develop the likelihood ratio statistic and compare it with the minimum discrimination information statistic for two-dimensional contingency tables under the null hypothesis of independence. We shall suppose that we have a sample of N observations from a multinomial population partitioned by two classification variables. Let x_{ij} be the number of observations occurring in cell (i,j), where $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

The likelihood function for a sample of size N is defined to be

$$L(p_{ij}) = \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{x_{ij}} \tag{1}$$

where x_{ij} is the frequency in cell (i,j).

Under the null hypothesis that the classifications are independent,

$$H_0 : p_{ij} = p_{i.}\, p_{.j} \quad \text{for all } i \text{ and } j,$$

we will develop the likelihood ratio statistic to test this hypothesis against the alternative hypothesis which simply negates the null hypothesis.

The likelihood function under the null hypothesis becomes

$$L(p_{i.}\ p_{.j}) = \prod_{i=1}^{r}\prod_{j=1}^{c} p_{ij}^{x_{ij}} = \prod_{i=1}^{r}\prod_{j=1}^{c} (p_{i.}\ p_{.j})^{x_{ij}}$$

$$= \prod_{i=1}^{r} p_{i.}^{x_{i.}} \prod_{j=1}^{c} p_{.j}^{x_{.j}} \quad . \tag{2}$$

Now, since $\sum_{i=1}^{r} p_{i.} = 1$ and $\sum_{j=1}^{c} p_{.j} = 1$, we can write $p_{r.}$ and $p_{.c}$ as follows

$$p_{r.} = 1 - \sum_{i=1}^{r-1} p_{i.} \ , \tag{3}$$

$$p_{.c} = 1 - \sum_{j=1}^{c} p_{.j} \quad . \tag{3'}$$

Substitutions of (3) and (3') into (2) give

$$L(p_{i.}\ p_{.j}) = \left(1 - \prod_{i=1}^{r-1} p_{i.}\right)^{x_{r.}} \prod_{i=1}^{r-1} p_{i.}^{x_{i.}} \prod_{j=1}^{c} p_{.j}^{x_{.j}} \tag{4}$$

and

$$L(p_{i.}\ p_{.j}) = \prod_{i=1}^{r} p_{i.}^{x_{i.}} \left(1 - \sum_{j=1}^{c-1} p_{.j}\right)^{x_{.c}} \prod_{j=1}^{c-1} p_{.j}^{x_{.j}} \ , \tag{4'}$$

respectively. If we take the natural logarithm of equations (4) and (4'), we then get

$$\log L(p_{i.}\ p_{.j}) = x_{r.} \log\left(1 - \sum_{i=1}^{r-1} p_{i.}\right) + \sum_{i=1}^{r-1} x_{i.} \log p_{i.} + \sum_{j=1}^{c} x_{.j} \log p_{.j} \tag{5}$$

$$\log L(p_{i.}\ p_{.j}) = \sum_{i=1}^{r} x_{i.} \log p_{i.} + x_{.c} \log\left(1 - \sum_{j=1}^{c-1} p_{.j}\right)$$

$$+ \sum_{j=1}^{c-1} x_{.j} \log p_{.j} \ , \qquad (5')$$

respectively. The maximum likelihood estimates are the values of the parameters $p_{i.}$, $i = 1, 2, \ldots, r$ and $p_{.j}$, $j = 1, 2, \ldots, c$, which maximizes $L(p_{i.}, p_{.j})$. Thus, to find values of $p_{i.}$ and $p_{.j}$ which maximizes (5) and (5'), as well as (4) and (4'), we differentiate with respect to each of the parameters $p_{i.}$ for $i = 1, 2, \ldots, r-1$, and $p_{.j}$ for $j = 1, 2, \ldots, c-1$ and equate each expression to zero giving

$$\frac{\partial \log L(p_{i.}\ p_{.j})}{\partial p_{i.}} = \frac{x_{r.}(-1)}{1 - \sum_{i=1}^{r-1} p_{i.}} + \frac{x_{i.}}{p_{i.}} = 0 \quad \text{for}$$

$$i = 1, 2, \ldots, r-1 \ , \qquad (6)$$

$$\frac{\partial \log L(p_{i.}\ p_{.j})}{\partial p_{.j}} = \frac{x_{.c}(-1)}{1 - \sum_{j=1}^{c-1} p_{.j}} + \frac{x_{.j}}{p_{.j}} = 0 \quad \text{for}$$

$$j = 1, 2, \ldots, c-1 \ . \qquad (6')$$

Solving the equations (6) and (6') for all $p_{i.}$ and $p_{.j}$, we obtain

$$\hat{p}_{i.} = \frac{x_{i.}\left(1 - \sum_{i=1}^{r-1} \hat{p}_{i.}\right)}{x_{r.}} \quad \text{for} \quad i = 1, 2, \ldots, r-1 \qquad (7)$$

and

$$\hat{p}_{\cdot j} = \frac{x_{\cdot j}\left(1 - \sum_{j=1}^{c-1} \hat{p}_{\cdot j}\right)}{x_{\cdot c}} \quad \text{for} \quad j = 1, 2, \ldots, c-1 . \tag{7'}$$

If we substitute the equations (3) and (3') into (7) and (7'), respectively, then we obtain

$$\hat{p}_{i\cdot} = \frac{x_{i\cdot}\, \hat{p}_{r\cdot}}{x_{r\cdot}} \quad \text{for} \quad i = 1, 2, \ldots, r-1 \tag{8}$$

and

$$\hat{p}_{\cdot j} = \frac{x_{\cdot j}\, \hat{p}_{\cdot c}}{x_{\cdot c}} \quad \text{for} \quad j = 1, 2, \ldots, c-1 . \tag{8'}$$

Summing (8) with respect to i and (8') with respect to j we get

$$1 = \sum_{i=1}^{r} \hat{p}_{i\cdot} = \frac{\sum_{i=1}^{r} x_{i\cdot}\, \hat{p}_{r\cdot}}{x_{r\cdot}} = \frac{N\, \hat{p}_{r\cdot}}{x_{r\cdot}} \tag{9}$$

and

$$1 = \sum_{j=1}^{c} \hat{p}_{\cdot j} = \frac{\sum_{j=1}^{c} x_{\cdot j}\, \hat{p}_{\cdot c}}{x_{\cdot c}} = \frac{N\, \hat{p}_{\cdot c}}{x_{\cdot c}} . \tag{9'}$$

Now, solving (9) and (9') for $p_{r\cdot}$ and $p_{\cdot c}$ we obtain

$$\hat{p}_{r\cdot} = \frac{x_{r\cdot}}{N} \tag{10}$$

and

$$\hat{p}_{\cdot c} = \frac{x_{\cdot c}}{N} \tag{10'}$$

If we substitute (10) and (10') into (8) and (8'), respectively, then

$$\hat{p}_{i.} = \frac{x_{i.}}{N} \quad \text{for} \quad i = 1, 2, \ldots, r \tag{11}$$

and

$$\hat{p}_{.j} = \frac{x_{.j}}{N} \quad \text{for} \quad j = 1, 2, \ldots, c\,. \tag{11'}$$

The maximum likelihood estimators are given by (11) and (11'). Thus, the likelihood function (2) evaluated at $\hat{p}_{i.}$ and $\hat{p}_{.j}$ becomes

$$L(\hat{p}_{i.}\,\hat{p}_{.j}) = \prod_{i=1}^{r} \frac{x_{i.}^{x_{i.}}}{N^{x_{i.}}} \prod_{j=1}^{c} \frac{x_{.j}^{x_{.j}}}{N^{x_{.j}}}$$

$$= \frac{\prod_{i=1}^{r} x_{i.}^{x_{i.}} \prod_{j=1}^{c} x_{.j}^{x_{.j}}}{N^{2N}}\,. \tag{12}$$

By a similar procedure for determining the maximum likelihood estimates of $p_{i.}$ and $p_{.j}$, we can find the maximum likelihood estimates of p_{ij} using the likelihood function

$$L(p_{ij}) = \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{x_{ij}}$$

and the equation

$$p_{rc} = 1 - \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} p_{ij}\,. \tag{13}$$

The likelihood function becomes upon substitution of equation (13)

$$L(p_{ij}) = \left(1 - \sum_{i=1}^{r-1}\sum_{j=1}^{c-1} p_{ij}\right)^{x_{rc}} \prod_{i=1}^{r-1}\prod_{j=1}^{c-1} p_{ij}^{x_{ij}} . \qquad (14)$$

The natural logarithm of (14) is

$$\log L(p_{ij}) = x_{rc} \log\left(1 - \sum_{i=1}^{r-1}\sum_{j=1}^{c-1} p_{ij}\right) + \sum_{i=1}^{r-1}\sum_{j=1}^{c-1} x_{ij} \log p_{ij}$$

and the partial derivatives of this equation with respect to each parameter p_{ij} gives

$$\frac{\partial \log L(p_{ij})}{\partial p_{ij}} = \frac{x_{rc}(-1)}{1 - \sum_{i=1}^{r-1}\sum_{j=1}^{c-1} p_{ij}} + \frac{x_{ij}}{p_{ij}} = 0 \quad \text{for}$$

$i = 1, 2, \ldots, r-1$ and $j = 1, 2, \ldots, c-1$.

Solving the above equations for p_{ij} we get

$$\hat{p}_{ij} = \frac{x_{ij}\left(1 - \sum_{i=1}^{r-1}\sum_{j=1}^{c-1} \hat{p}_{ij}\right)}{x_{rc}} = \frac{x_{ij}\,\hat{p}_{rc}}{x_{rc}} \qquad (15)$$

and summing with respect to i and j we have

$$1 = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = \frac{N\hat{p}_{rc}}{x_{rc}} . \qquad (16)$$

Substituting $\hat{p}_{rc} = \frac{x_{rc}}{N}$ from (16) into (15) we get

$$\hat{p}_{ij} = \frac{x_{ij}}{N} \quad \text{for all } i \text{ and } j . \tag{17}$$

Equation (17) gives the maximum likelihood estimates and equation (2) becomes

$$L(\hat{p}_{ij}) = \frac{\prod_{i=1}^{r} \prod_{j=1}^{c} x_{ij}^{x_{ij}}}{N^N} .$$

The likelihood ratio statistic denoted by λ is given by

$$\lambda = \frac{L(\hat{p}_{i.}\,\hat{p}_{.j})}{L(\hat{p}_{ij})} = \frac{1}{N^N} \frac{\prod_{i=1}^{r} x_{i.}^{x_{i.}} \prod_{j=1}^{c} x_{.j}^{x_{.j}}}{\prod_{i=1}^{r} \prod_{j=1}^{c} x_{ij}^{x_{ij}}} , \tag{18}$$

The natural logarithm of λ gives

$$\log\lambda = \sum_{i=1}^{r} x_{i.} \log x_{i.} + \sum_{j=1}^{c} x_{.j} \log x_{.j} - N \log N$$

$$- \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} \log x_{ij} . \tag{19}$$

If we multiply (19) by -2 we get

$$-2 \log\lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} \log x_{ij} - 2 \sum_{i=1}^{r} x_{i.} \log x_{i.}$$

$$- 2 \sum_{j=1}^{c} x_{.j} \log x_{.j} + 2N \log N . \tag{20}$$

We note equation (20) is identical to the minimum discrimination information statistic for testing the hypothesis of independence. The statistic $-2 \log \lambda$ is asymptotically distributed chi-square with $(r-1)(c-1)$ degrees of freedom ([62], p. 113).

VITA

Gene Burton Iverson

Candidate for the Degree of

Doctor of Education

Thesis: ANALYSIS OF CATEGORICAL DATA

Major Field: Higher Education

Biographical:

Personal Data: Born in Vermillion, South Dakota, June 16, 1938, the son of George L. and Dora Iverson.

Education: Attended Meckling Consolidated Grade School and graduated from Meckling Consolidated High School, Meckling, South Dakota, in 1956; received the Bachelor of Science in Education degree from the University of South Dakota, Vermillion, South Dakota in 1960; received the Master of Natural Science degree in Mathematics and Physics in the summer of 1964 from the University of South Dakota, Vermillion, South Dakota; attended the University of Oklahoma, Norman, Oklahoma, summer 1965; attended New Mexico State University, Las Cruces, New Mexico, summer 1966; attended San Jose State College, San Jose, California, summer 1967; completed requirements for the Doctor of Education degree from Oklahoma State University in July, 1973.

Professional Experience: Taught high school mathematics and science at Viborg High School, Viborg, South Dakota from 1960 to 1963; taught high school mathematics at Fairmont High School, Fairmont, Minnesota in 1964 to 1965; instructor and assistant professor of mathematics at Wisconsin State University, Superior, Wisconsin from 1965 to 1968; assistant professor of mathematics and computer science at the University of South Dakota, Vermillion, South Dakota; graduate assistant in the Department of Mathematics and Statistics, Oklahoma State University, Stillwater, Oklahoma from 1971 to 1973.