TOPICS ON THE ESTIMATION OF SMALL PROBABILITIES

by

Wolfgang Pelz

Dissertation submitted to the Graduate Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

APPROVED:

I.J. Good, Chairman

S.K. Lee

R.G. Krutchkoff

R.H. Myers

R.A. Thompson

August, 1977

Blacksburg, Virginia

ACKNOWLEDGMENTS

I wish to thank the many people who have been involved in making this degree possible. I am especially grateful to the following people for their participation in this endeavor:

Dr. I.J. Good for his direction, inspiration, and appreciation of the beauty of mathematics and statistics,

Dr. A.V. Smith for introducing me to the area of statistics and suggesting I enter graduate school in statistics at V.P.I. and S.U.,

Dr. S.K. Lee for co-reading my dissertation and for his helpful suggestions,

The United States Government for financial assistance through NIH grants,

The B.F. Goodrich Company for granting a leave of absence to complete this dissertation,

Faye Roop for her typing and checking of the final copy,

and my parents without whose support and encouragement this degree would not have been completed.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont.)

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Many times when estimating small probabilities, one finds that the usual approaches to estimation are inadequate. For instance, the usual estimates of some parameter may be too crude, e.g., estimating cell probabilities for a multidimensional contingency table in which most of the cell frequencies are zero. In other cases the usual approach may not be the most efficient method for obtaining the desired estimates. Two topics dealing with these problems are here discussed.

In Part I, the problem of estimating cell frequencies in a multinomial or contingency-table framework is considered. Of special importance are those circumstances in which zero cell frequencies appear. Attention is centered primarily on multinomial and r x c contingency tables because the results are in a form readily understood. Extensions of derivations to multidimensional tables, though not difficult, are not emphasized because of the complexity of the results. A review of techniques used in estimation for contingency tables with some comparisons of these techniques is given in Chapter 1. The introduction and derivation of the Maximum Likelihood/Entropy (ML/E) estimation technique are given in Chapter 2. In Chapter 3 we describe the procedures necessary to obtain the ML/E estimates under varying conditions. Chapter 4 deals with properties of the ML/E estimator as well as asymptotic and small sample comparisons with other estimation techniques. In Chapter 5 we demonstrate how the ML/E estimation technique may be applied to a multidimensional contingency table.

We assume that the contingency tables are "pure" in the sense of Good (1956), that is, that there is no natural ordering of the rows nor of the columns, or, at least if there is, that this source of information is to be disregarded.

In Part II we consider the calculation of the Kolmogorov-Smirnov one-sample statistic. Chapter 1 deals with the methods which have been presented as procedures to calculate this statistic. In Chapter 2 we show that an approximation method due to Li-Chien can be reduced to evaluating a linear combination of derivatives of two basic functions. The relationship between the theta functions and this approximation method is given in Chapter 3. In Chapter 4 we present a transformation of this approximation method useful when the left tail area probability is small. The relationships between various methods, recommendations for a usable range for each method, and an analysis of errors obtained by use of the approximation method are given in Chapter 5.

PART I

Maximum Likelihood/Entropy Estimation

CHAPTER I

Probability Estimation for Contingency Tables

This chapter begins by introducing some notations, definition of terms, and the sampling distributions encountered in this work. A summary of the techniques promoted by different authors for probability estimation in multinomial and contingency table situations with some comparisons of these techniques follows.

1.1 Notation and Definitions

Historically a contingency table has been defined as a tabular representation of cross-classified non-negative integers which are considered as a sample drawn from a population. Good (1965) extended this concept to population contingency tables which are composed of a set of probabilities. Bishop et al (1975) use "contingency tables" to denote arrays of smoothed frequencies for different models, inter alia. In this work we will call all of these forms contingency tables. Therefore the entries in a contingency table may be non-negative numbers such as integers, real non-integers, or probabilities, depending on the particular system which the table represents. Though the table itself is obviously discrete, it may be used to characterize a continuous distribution. The table is described as having rows, columns, layers, etc. for each dimension or "facet".

Define n_{ij} as the sample frequency of the cell in the i^{th} row and j^{th} column in a two-dimensional table. This table would therefore be

composed of non-negative integers. Define $\{\pi_{ij}\}$ as the set of cell probabilities of the population from which the sample is drawn. Let $\{p_{ij}\}$ be the set of probability estimates determined by implementing some estimation procedure. The marginal totals may then be defined as

$$\pi_{i\cdot} = \sum_j \pi_{ij},\ \pi_{\cdot j} = \sum_i \pi_{ij} \qquad \text{for the population table}$$

$$p_{i\cdot} = \sum_j p_{ij},\ p_{\cdot j} = \sum_i p_{ij} \qquad \text{for the sample probability table}$$

and $$n_{i\cdot} = \sum_j n_{ij},\ n_{\cdot j} = \sum_i n_{ij} \qquad \text{for the sample frequency table}$$

with the corresponding grand totals $1 = \pi_{\cdot\cdot} = \sum_j \pi_{\cdot j}$, $1 = p_{\cdot\cdot} = \sum_j p_{\cdot j}$, and $N = n_{\cdot\cdot} = \sum_j n_{\cdot j}$, where $\sum_i$ is defined as the summation over all values of $i \in I$, where I is the set of numbers with a particular characteristic, e.g., all the frequencies in a specified row. Note that $n_{\cdot\cdot}^k$ means $(n_{\cdot\cdot})^k = (\sum_i \sum_j n_{ij})^k$, not $\sum_i \sum_j (n_{ij})^k$.

In the course of estimation, one may have to consider information about the margins. This information can result from previous experimentation, a priori knowledge, or restrictions due to the sampling model. Some examples of cases in which marginal totals may be fixed at some specified values are when:

1) the population margin values are known
2) the margins of another table taken from the same population are to be fitted
3) the margins of another table taken from a different population are to be fitted

4) the sample margins are to be kept fixed

5) the marginal values to be fitted are determined before sampling

or 6) the margins to be fitted are obtained through a very large sample so that these margins may be considered as a good approximation to the population margins.

Note that there is a distinction between fitting and fixing margins. Margins are fixed at the values in the sample table; they are fitted to values obtained from another source.

These different possibilities represent differences in sampling methods and estimation objectives and will be treated in more detail later.

An alternative notation for contingency tables is to express the cell values in vector format. Thus an r x c table $\{n_{ij}\}$ can be written as a vector of length rc in the following arrangement:

$$[n_{11}\ n_{12} \cdots n_{1c}\ n_{21} \cdots n_{2c}\ n_{31} \cdots n_{rc}] \quad .$$

Gower (1968) describes and lists an Algol computer program to convert such arrays to vectors. This notation is particularly useful when a multidimensional table is under consideration. For example a table with five dimensions would have typical element $n_{i_1 i_2 i_3 i_4 i_5}$ using the first notation and n_i using the second where i is a scalar. This vector notation is also useful in some theoretical work in which the first notation can again become relatively cumbersome. However, in practice, the first notation is preferred because of the visibility of the appropriate margins, and will be used here. Any deviations

from the first notation will be clearly marked.

In Chapter 3 certain terms from the area of mathematical programming will be used. For a more complete explanation of terms and/or techniques than that given below, refer to a book dealing with this area.

Many procedures are based on manipulation of a function, called an objective function, of the cell frequencies and of the probabilities. We either wish to maximize or minimize (more generally, to optimize) this function with respect to the probabilities subject to certain constraints. Sometimes these procedures are based on iterative methods which require a feasible solution, that is, a solution which satisfies all the constraints. In addition some methods require the use of a basis. Each variable is associated with a single column vector in the matrix determined by the constraints. A basis is a set of linearly independent vectors which span the column vectors in this matrix. A basic variable is a variable associated with a vector in the basis. A non-basic variable is not a basic variable.

1.2 Sampling Methods and Distributions

A contingency table composed of cell frequencies can arise from numerous sampling procedures, however, only the more common methods will be discussed here. The sampling distributions are presented in order of increasing constraints on the system.

Definition A random variable X has a Poisson distribution with parameter $m > 0$ if

$$\Pr[X=x] = \frac{e^{-m} m^{x}}{x!} \qquad x = 0,1,\ldots$$

1.2

If a contingency table is constructed of cells whose frequencies are determined by an independent Poisson process over a fixed period of time T, then the observations in the cells have independent Poisson distributions and their joint distribution is

$$\Pr[X_{11}=x_{11},\ldots, X_{rc}=x_{rc}] = \prod_i \prod_j \frac{e^{-m_{ij}} m_{ij}^{x_{ij}}}{x_{ij}!}$$

where $E(X_{ij}) = m_{ij} = n_{..} \pi_{ij}$.

Definition If an array $X = (X_{11},\ldots, X_{rc})$ has a probability function given by

$$\Pr[X_{11}=x_{11},\ldots, X_{rc}=x_{rc}] = \frac{n_{..}!}{\prod_i \prod_j x_{ij}!} \prod_i \prod_j (m_{ij}/n_{..})^{x_{ij}}$$

for nonnegative integers x_{ij} and nonnegative real $m_{ij} > 0$ for all i and j with $\sum_i \sum_j m_{ij} = \sum_i \sum_j x_{ij} = n_{..}$, then the joint distribution of X is called the multinomial distribution. This definition of the multinomial distribution is given in a non-standard notation because of its usage in connection with contingency tables in this work.

This distribution occurs when sampling until a specified total sample size is achieved. When the array X is a vector of length 2, we have the binomial distribution.

If we are given a marginal column of row totals and that the rows are independent, then the distribution of the cell frequencies given that the row margins are fixed is called a product multinomial and is expressed as

1.2

$$\Pr[X_{11}=x_{11},\ldots, X_{rc}=x_{rc} \mid x_{1\cdot},\ldots, x_{r\cdot}] = \prod_i \frac{x_{i\cdot}!}{\prod_j x_{ij}!} \prod_j (m_{ij}/x_{i\cdot})^{x_{ij}}$$

where $\sum_j m_{ij} = \sum_j x_{ij} = x_{i\cdot}$.

Definition Assume a population of N objects, M of which have a particular characteristic (say type 1), and a sample of size n with X of these objects being type 1. For $\max(0, n+M-N) \le x \le \min(n, M)$, X has the hypergeometric distribution given by

$$\Pr[X=x] = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} .$$

Definition Assume a population of N objects, N_i of which are type i ($\sum_i^t N_i = N$), and a sample of size n, taken without replacement, where $X_1,\ldots, X_t$ are the random variables representing the number of objects of types 1,..., t, respectively. The joint distribution of $X_1,\ldots, X_t$ is the multivariate hypergeometric distribution given by

$$\Pr[X_1=x_1,\ldots, X_t=x_t] = \frac{\prod_i \binom{N_i}{x_i}}{\binom{N}{n}} .$$

If an r x c contingency table is obtained through sampling with both margins fixed and entry into columns is independent of entry into rows, then the set of cell counts are jointly multivariate hypergeometrically distributed (see Yates (1934), Stevens (1938), or Fisher (1934, 21.02)).

In contingency tables of two or more dimensions, one must examine the constraints to check whether any variables are included in more than

one constraint. If so, the models become even more complicated and are closely related to the quasi-multinomial distribution of Rudolph (1967). These models have different log-likelihoods than those mentioned earlier.

When $\sum_i \sum_j m_{ij} = n_{..}$, the kernel of the likelihood function is the same for the Poisson, Multinomial, and Product Multinomial models, that is, log-likelihood $= K + \sum_i \sum_j n_{ij} \ln(m_{ij}) = K + \sum_i \sum_j n_{ij} \ln(n_{..}\pi_{ij})$ for constant K.

The estimation procedures examined in this work deal with the sampling models that have log-likelihoods containing these terms.

Though not discussed in this work, contingency tables can arise through categorizing a continuous distribution. Relevant references are Lancaster (1969) and Kendall and Stuart (1973).

1.3 Introduction to Probability Estimation for Contingency Tables

The history of probability estimation for contingency tables has been characterized by a multitude of philosophical approaches. These methods can be loosely classified into two groups:

1) methods which optimize some function of the probability estimates and cell frequencies (known as the objective function) with or without knowledge of margin totals, and
2) methods which estimate probabilities for tables with small or zero cell frequencies by using either some prior distribution or a specified adjustment procedure.

These classifications are not necessarily distinct and may have considerable overlap. This may occur when the same form for a prob-

ability estimator is developed using different approaches. Examples of overlap of different techniques are:

1) Maximum Likelihood estimation (in group 1) and Maximum Likelihood/Entropy (in groups 1 and 2) under certain conditions, and
2) the methods of Steinhaus, Bayes-Laplace, and Perks-Jeffreys (all in group 2) mentioned in Section 2.1.

In many cases the techniques followed parallel developments and influenced each other. The progress of methods in each group will be considered separately unless the overlap is of significant influence on the technique.

1.4 Estimation through Optimization

Estimation by optimization was first dominated by the methods of Minimum Chi Squared as defined by Pearson (1900) and Maximum Likelihood (see Fisher (1922)). Then Deming and Stephan (1940) considered the problem of minimizing the function $\sum_i \sum_j (n_{ij} - n_{..}p_{ij})^2/n_{ij}$ subject to fixed known margin totals. This corresponds to the method of least squares in which the sample frequencies are weighted by their own reciprocals. The authors developed the normal equations for the problem and proposed the iterative proportional fitting procedure (IPFP) as a method to solve the equations. A weakness in this method is that the cell frequencies are restricted to positive numbers, that is, zero cell counts were not permitted.

Stephan (1942) indicated two additional flaws:

1.4

1) although the marginal restrictions were satisfied, the normal equations were only approximately satisfied, and

2) a proof of the convergence of the IPFP had not yet been found.

Deming (1943) extended the procedure to situations in which only one margin is known. In addition, he considered three-dimensional tables with various combinations of margins, $\{\pi_{i\cdot\cdot}\}$, $\{\pi_{\cdot j\cdot}\}$, $\{\pi_{\cdot\cdot k}\}$, and faces, $\{\pi_{ij\cdot}\}$, $\{\pi_{\cdot jk}\}$, $\{\pi_{i\cdot k}\}$, fixed and known.

Smith (1947) obtained a system of simultaneous nonlinear equations for the probabilities in terms of the marginal restrictions. He also compared the methods of Minimum Chi Squared, Maximum Likelihood, and the method of weighted least squares as used by Deming and Stephan (1940) to show that the methods were approximately equivalent.

Neyman (1949) defined a class of estimators called Best Asymptotic Normal (BAN) which had certain desirable asymptotic properties. Included in this class are Maximum Likelihood, Minimum Chi Squared, and Minimum Modified Chi Squared (see Neyman and Pearson (1930-31, 1933), Wilks (1935, 1938), and Quade and Salama (1975)).

Lewis (1959) considered approximations to discrete probability distributions by using distributions which are products of marginal distributions of the original one in an attempt to maximize the information content in a limited amount of storage. He then chose that product distribution which was the Minimum Information (Maximum Entropy) extension of the component distributions (those distributions with the same margins). For example, in the framework of a 2^n contingency table, he wished to estimate the cell probabilities using products of the

marginal probabilities. Brown (1959) extended the argument to distributions which can employ any set of marginal distributions instead of only product approximations. He described an iterative procedure identical to IPFP and showed that the approximation improved at each step of the iteration with Minimum Information (Maximum Entropy) as the criterion.

Ireland and Kullback (1968) developed a rigorous proof of the convergence of the IPFP and showed that it converged to the Minimum Discrimination Information estimates rather than the least squares estimates. Their proof was based on an extension of Brown's work which confirmed a statement in Good (1965, p. 75) in which this extension of IPFP was called the "iterative scaling procedure".

Caussinus (1965) proved that the IPFP converges for certain contingency tables and demonstrated convergence when exactly one cell frequency is zero. Fienberg (1970a) gives a geometric proof of the IPFP for two-way contingency tables based on concepts presented in Fienberg (1968).

Bishop (1967) showed that Brown's non-rigorous proof could be extended to any multi-dimensional table by using a duality theorem of Good (1963, 1965) which gives a relationship between Maximum Likelihood estimation and Maximum Entropy estimation for contingency tables. She then showed that the IPFP could be used to derive Maximum Likelihood estimates for a variety of log-linear models developed by Birch (1963). These models were anticipated by Good (1956) in connection with Bayesian estimates for small cell frequencies in contingency tables. Birch had proved that corresponding to any particular log-linear model, there

exists a subset of all the marginal totals of a given table which is a set of sufficient conditions for that model and, as suggested by Darroch (1962), this set can be used in the IPFP to generate Maximum Likelihood estimates.

Haberman (1972) gives a Fortran program to calculate these estimates. Darroch and Ratcliff (1972) generalized this method to a larger class of models. Haberman (1974) gives a rigorous general presentation of log-linear models and indicates results for a large class of models for frequency data. Gokhale (1971) considered the IPFP as dependent on a minimal set of sufficient statistics for usual margins and generalized the method to a non-minimal set of sufficient statistics for unusual margins. This method corresponds to fixing a weighted sum of the cell frequencies, e.g., fixing the sum of the main diagonal entries at a specified value. When the underlying model is complicated, this method has the advantage of its routine applicability since a set of minimal sufficient statistics may be difficult to obtain.

The restriction that all cell counts must be positive can be relaxed so long as zero cell counts do not form a system which leads to negative probabilities. This idea is shown in Bishop, Fienberg and Holland (1975) and bears repeating. Assume we have the observed frequencies in the top 2 x 2 x 2 table given in Figure 1.4.1, where b, c, d, e, f, g > 0. If a positive estimate for the first cell is desired, say Δ, then for fixed margins we have the bottom table in the same figure, and find a negative expected frequency for the last cell. The expected or estimated values of the cells must therefore be the observed values in this example (if the marginal totals are left unchanged).

0	b
c	d

e	f
g	0

Δ	b$-\Delta$
c$-\Delta$	d$+\Delta$

e$-\Delta$	f$+\Delta$
g$+\Delta$	$-\Delta$

Figure 1.4.1 Observed and Estimated 2 x 2 x 2 Contingency Tables with Zero Cell Frequencies.

However, other less constrained situations may occur in which zero cell frequencies still result in zero estimates. This is intuitively disturbing and has led to the problem of adjustment of zero cell frequencies discussed in the next section. Comparisons of some of the methods mentioned in the present section will be made at the end of the chapter.

1.5 Adjustment of Zero Cell Frequencies

In the analysis of frequency data, it is not uncommon to encounter zeros in some cells. These zeros cause problems when various linearizing scales such as logarithms, logits, or probits are used to transform the counts. Therefore some adjustment must be made in order to use these transformations.

Another difficulty occurs when probabilities associated with particular cells in a contingency table are reported as zero. A probability of 0/5 should usually exceed a probability of 0/5000 where 5 and 5000 are total frequencies, yet many methods consider these two values as equal.

A distinction between two types of zeros must be made. A fixed zero is a zero cell frequency which arises due to a true zero probability for that cell. For example a cell whose frequency is defined as the number of "Pregnant Males" should have zero cell probability (as well as zero cell frequency). A random zero is defined as a zero occurring in a rare but non-empty category, e.g., the number of chimney sweeps from Rhode Island. An estimation procedure might change a random zero

1.5

to something positive, but must leave fixed zeros unaltered.

The history of zero cell estimation is especially replete with different approaches for varying circumstances.

The problem of zero frequencies dates back to Laplace (1774) who reasoned that in the binomial situation with r successes out of N trials, the probability of a success at the next trial is $(r + 1)/(N + 2)$. This inference is known as Laplace's Law of Succession. Lidstone (1920) generalized this method to adding one pseudo count to each cell in a multinomial problem.

Jeffreys (1961) and Perks (1947) independently proposed invariance theories leading to the estimate $(r + \frac{1}{2})/(N + 1)$ for binomial sampling. For a t-category multinomial sampling problem, Jeffreys' theory produced the same flattening constant of $\frac{1}{2}$, whereas Perks advocated $1/t$. Perks (1967), after extensive correspondence with Good, agreed that using a flattening constant of $1/t$, which he had previously advocated, was inadequate.

Imrey and Koch (1972) and Koch et al (1972) used $1/t$ in place of observed zero cell counts to avoid singularities in the estimated covariance matrix of the probability vector when data are missing in the paired comparison of two political polls.

Gart (1962) and Haldane (1955) generalized the work of Jeffreys to adding $\frac{1}{2}$ of a pseudo count to each cell of a contingency table.

Berkson (1955), in an analysis of Minimum Logit Chi Square, used $(2N)^{-1}$ as a working value for cells with zero frequencies in order to avoid the necessity of ignoring these observations. Goodman (1963,1964)

1.5

and Plackett (1962) following Berkson prefer adding ½ count only to the empty cells in testing the null hypothesis of zero second-order interaction in three-dimensional tables for which the value of $e_{ijk} = \ln(n_{ijk})$ is required.

Lindley (1964) by way of contrast suggested subtracting ½ from each multinomial cell when using an improper prior proportional to $\prod_j \pi_j^{-1}$ with all $n_j > 0$. Bartlett (1935) suggested adding ¼ of a pseudo count to each empty cell.

Bishop and Mosteller (1969) add an arbitrary constant to all the cell frequencies, derive a model using the log-linear framework mentioned previously, and then subtract this constant from the expected values to overcome difficulties with empty cells.

Good (1956) presented the association-factor method (as well as weighted lumping methods) which follows from the definition of the amount of information in one proposition concerning another, i.e., $\log \{P(E.F)/[P(E)P(F)]\}$ where E and F are the two propositions. The equivalence to the log-linear models developed independently by Birch (1963) is immediate, though this model was used in a Bayesian way.

Goodman (1968) in an analysis of quasi-independence, that is, independence for a specified subset of cells, arrived at estimated expected frequencies by ignoring the empty cells and using an extension of the IPFP. Fienberg (1970b) and Savage (1973) provided conditions to ensure the existence of unique nonzero Maximum Likelihood estimates for the cells of incomplete tables obtained by deleting the missing or a priori zero cells, even when other cells contain zero counts due to

1.5

sampling variation. Young and Young (1975, 1976) use a weighted least squares estimator with weights of a constant and a zero assigned to nonzero and zero cells, respectively. This is equivalent to simply ignoring fixed zero cell frequencies.

Huber and Lellouch (1974) wished to estimate the distribution of a discrete vector variable when the size of the sample is small compared to the size of the contingency table representing all possible values of the variable. They proposed a variant of the log-linear models based on a generalization to the assumption of multinormality in the continuous case.

For situations with a total sample size of zero, Good (1963) suggested the use of Maximum Entropy with appropriate marginal constraints for hypothesis formulation and showed that the derived probability estimates must be positive for all cells. Zellner (1971) regards the amount of information in a distribution as minus its entropy and prefers to maximize the gain in information associated with an observation by maximizing the expected amount of information that observation provides.

Trybula (1958) suggested using the estimator that minimizes the maximum value of its risk function under quadratic loss.

Goodman (1959) estimated all the cell probabilities in two-way contingency tables using only the margins without assuming independence. Instead he made certain assumptions about a set of two-way tables, such as identical transition probabilities, and obtained his estimates by least squares. For example, let $q_{11} = n_{11}/n_{1.}$ and $q_{21} = n_{21}/n_{2.}$ be

1.5

the transitional probabilities leading from row 1 to column 1 and from row 2 to column 1 respectively. Then $n_{.1} = q_{11}n_{1.} + q_{21}n_{2.}$ + error. With such an equation for each 2 x 2 table, we can estimate q_{11} and q_{21}.

Makeham (1891) assumed that a sample of m white balls and n black balls has been chosen with replacement from an urn with probability p of choosing a white ball. He found that the probability of a white ball being chosen on the next trial was between $m/(m + n)$ and p and wrote this as $(m + rp)/(m + n + r)$ for arbitrary r, independent of m + n. Stabler (1892) found that r is not independent of m + n, but was in fact equal to m + n - 1.

Johnson (1932) considered the postulate that the expectation of p_j is dependent only on N, n_j, and t; the remaining cell frequencies n_i, $i \neq j$ being irrelevant. From this he deduced that the estimate of p_j should be $(n_j + k)/(N + tk)$ where k depends only on t and is strictly positive.

Good (1965, 1967) developed an argument for estimating k in the probability estimator $(n_i + k)/(N + tk)$ by extending Johnson's postulate to the concept of Type II Maximum Likelihood with symmetric Dirichlet priors. The estimated value of k is found to be essentially a function of the roughness of the sample. Fienberg and Holland (1972) discuss this estimator in terms of the maximum number of "small" cell probabilities for which the smoothed estimator has smaller risk than the unsmoothed estimator. Good (1965, 1967, 1976b) and Good and Crook (1974) consider placing a Type III distribution on k to obtain essentially linear combinations of symmetric Dirichlet distributions. This method

1.5

gives a form for mixed Dirichlet priors which can be used for both significance testing as well as probability estimation.

Fienberg and Holland (1970, 1973) consider a different form of mixed Dirichlet priors mentioned in passing by Good (1967) to estimate the cell probabilities in a contingency table. The Pseudo-Bayes estimator presented by Fienberg and Holland is given by

$$p_i^* = (N\hat{p}_i + \hat{K}\lambda_i)/(N + \hat{K})$$

where $\hat{p}_i$ is the Maximum Likelihood estimator. It should be noted that in the Multinomial case with $\lambda_i = t^{-1}$, this estimator becomes $p_i^* = (n_i + \hat{K}/t)/(N + \hat{K})$ which is exactly the form of the estimator given in Good (1965). The difference between the two estimators is that Fienberg and Holland calculate $\hat{K}$ by minimizing the expected risk based on a squared error risk function. Good prefers to estimate $\hat{K}$ using the Type II Maximum Likelihood method. Since both methods are examples of procedures using Dirichlet priors, only one will be chosen for later comparison with other estimators. In any subsequent discussion, the estimator described as the Pseudo-Bayes estimator will in fact be the estimator p^* proposed by Fienberg and Holland. Sutherland, Holland, and Fienberg (1974) extend the Pseudo-Bayes concept using different estimators for $\hat{K}$ and λ_i .

The Pseudo-Bayes method, since it involves the estimation of a hyperparameter, is an example of what Good (1965) calls the Bayes/non-Bayes compromise. The notion of a compromise between Bayesian and non-Bayesian methods dates back at least to Good (1952) where it

was expressed in terms of probabilities of higher and higher "types".

Leonard (1972, 1973) proposed estimation methods for row, column, and interaction effects in a two-way contingency table when the use of exchangeable priors is considered reasonable. These methods lead to the estimation of "shrinking parameters" that cause Maximum Likelihood estimates to be shrunk towards a central value. The procedure is useful even when zero cell frequencies are present and does not force certain effects to be zero in such a case.

For estimating the Binomial parameter, Schafer (1976) considered modifying the Minimum Variance Unbiased estimator of Arnold (1972) for situations in which certain values which the estimator can attain are excluded from the set of possible estimates. For instance, it is possible to restrict the parameter space to the open interval (0,1) a priori, that is, making estimates of 0 or 1 impossible. Farebrother (1977) compares this estimator to Maximum Likelihood and finds conditions for preferring ML or Schafer's estimator on the basis of Mean Squared Error.

1.6 Detailed Comparisons of Certain Methods

Some of the methods presented in Section 1.4 are easy to compare on the basis of their objective functions, yet it is not immediately seen how these different criteria affect the probability estimates. Procedures developed by Smith (1947) and El-Badry and Stephan (1955) offer this opportunity. The development for Maximum Likelihood Estimation is given in great detail; the results for other methods are found in a similar manner. Note that the method of Maximum Entropy

mentioned in Section 1.5 because of its relation to estimation of zero cell counts is also described here due to the similarity to the optimization procedures.

Definition The method of Maximum Likelihood for an r x c contingency table is defined as obtaining the set of probability estimates $\{p_{ij}\}$ which maximizes the objective function $L = \sum_i \sum_j (n_{ij}/n_{..}) \ln \pi_{ij}$ and satisfies the particular marginal constraints imposed by the sampling method as well as satisfying $\sum_i \sum_j p_{ij} = 1$.

Theorem 1.6.1 The Maximum Likelihood estimator of π_{ij} can be expressed as

$$p_{ij} = (n_{ij}/n_{..})(\alpha_i + \beta_j)^{-1} \text{ for all } (i,j), \qquad (1.6.1)$$

where α_i and β_j are functions of the marginal constraints

$$\sum_i p_{ij} = b_j \qquad j = 1,..,c$$

and

$$\sum_j p_{ij} = a_i \qquad i = 1,..,r.$$

Note: it appears at first that there are r + c independent constraints on 1.6.1. However there are only r + c - 1 constraints since $\sum_j b_j = \sum_i a_i = 1$.

Proof:

The objective function and constraints may be written in one equation using the Lagrangian multipliers α_i and β_j as

1.6

$$L^* = \sum_i \sum_j (n_{ij}/n_{..}) \ln \pi_{ij} - \sum_i \alpha_i (\sum_j \pi_{ij} - a_i) - \sum_j \beta_j (\sum_i \pi_{ij} - b_j). \quad (1.6.2)$$

Taking partial derivatives with respect to π_{ij} yields

$$\partial L^*/\partial \pi_{ij} = n_{ij}/(n_{..}\pi_{ij}) - \alpha_i - \beta_j \quad \text{for all } i,j. \quad (1.6.3)$$

Setting the expression on the right equal to zero and substituting the estimators for the probabilities gives

$$0 = n_{ij}/(n_{..}p_{ij}) - \alpha_i - \beta_j \quad \text{for all } i,j, \quad (1.6.4)$$

and finally

$$p_{ij} = (n_{ij}/n_{..})(\alpha_i + \beta_j)^{-1} . \quad (1.6.5)$$

Corollary If the only constraint on the probabilities is that their sum be equal to 1, then

$$p_{ij} = n_{ij}/n_{..} \quad \text{for all } i,j. \quad (1.6.6)$$

Proof:

The single constraint is achieved by putting $\alpha_i = \alpha$(constant) and $\beta_j = 0$ for all j in equation 1.6.2. Then 1.6.4 becomes

$$0 = n_{ij}/(n_{..}p_{ij}) - \alpha$$

or

$$\alpha n_{..}p_{ij} = n_{ij} .$$

1.6

Summation over i and j gives $\alpha = 1$ and equation 1.6.6 immediately follows. Note: Equation 1.6.6 gives of course the familiar Maximum Likelihood estimator for the multinomial distribution.

Corollary If the row margins alone are constrained, equation 1.6.5 reduces to

$$p_{ij} = a_i n_{ij}/n_{i\cdot} \quad \text{for all } i,j. \tag{1.6.7}$$

Proof:

By placing $\beta_j = 0$ for all j in 1.6.5, we obtain

$$n_{ij} = p_{ij}\alpha_i n_{\cdot\cdot} \quad \text{for all } i,j. \tag{1.6.8}$$

Summation over j gives

$$n_{i\cdot} = p_{i\cdot}\alpha_i n_{\cdot\cdot}$$

$$= a_i \alpha_i n_{\cdot\cdot}$$

$$\text{and} \qquad \alpha_i = n_{i\cdot}/(a_i n_{\cdot\cdot}) \quad \text{for all } i,j. \tag{1.6.9}$$

Combining 1.6.8 and 1.6.9 gives

$$p_{ij} = a_i n_{ij}/n_{i\cdot} \quad \text{for all } i,j.$$

Corollary The Maximum Likelihood estimator p_{ij} given in equation 1.6.5 must satisfy

$$n_{ij}/p_{ij} - n_{ic}/p_{ic} - n_{rj}/p_{rj} + n_{rc}/p_{rc} = 0 \text{ for all } i,j. \tag{1.6.10}$$

1.6

Proof:

Equation 1.6.10 follows at once from equation 1.6.5 since the left side of equation 1.6.10 equals

$$n_{..}(\alpha_i + \beta_j) - n_{..}(\alpha_i + \beta_c) - n_{..}(\alpha_r + \beta_j) + n_{..}(\alpha_r + \beta_c).$$

Tables 1.6.1 and 1.6.2 give these results in tabular form and show the corresponding results for the other indicated methods. Note that while a_i, b_j, n_{ij}, $n_{i.}$, $n_{..}$, r, and c are the same for each method, α_i and β_j may be quite different.

The sets of equations derived for each of the estimation methods are necessary, but not sufficient, conditions for the estimators. For example assume that we have a 2 x 2 contingency table with cell frequencies $\{n_{ij}\}$ and given margin totals $\{m_{i.}, m_{.j}\}$ to be fitted such that

$$m_{i.} \neq n_{i1} + n_{i2} \quad i = 1,2$$

$$m_{.j} \neq n_{1j} + n_{2j} \quad j = 1,2$$

If the method of Maximum Likelihood estimation is used, we have that the estimators must satisfy the equation

$$n_{11}/p_{11} - n_{12}/p_{12} - n_{21}/p_{21} + n_{22}/p_{22} = 0.$$

Define $p_{ij} = n_{ij}/n_{..}$ for i, j = 1, 2. These estimators obviously satisfy the above equation, yet cannot be the Maximum Likelihood estimators since the margin constraints are not satisfied. The same estimators can be used to disprove sufficiency for Minimum Chi Squared, Minimum Modified

Chi Squared, and Minimum Discrimination Information. For Maximum Entropy, we need an estimator with the form $p_{ij} = (rc)^{-1}$.

If we include the condition that the marginal constraints must be satisfied, then this condition and the necessary equations (such as (1.6.10)) are jointly necessary and sufficient.

Theorem 1.6.2 Given an r x c contingency table with cell frequencies $\{n_{ij}\}$ such that $n_{ij} \neq 0$ for all i,j and a set of marginal constraints on the rows and columns. Then the Minimum Discrimination Information method yields probability estimates which satisfy to 0(1) the equation in Table 1.6.2 for the method of Maximum Likelihood.

Proof:

Let $\{p_{ij}\}$ be the set of probability estimators which is obtained through the method of Minimum Discrimination Information. It is therefore a set of estimators which satisfies

$$\ln(n_{ij}/p_{ij}) - \ln(n_{ic}/p_{ic}) - \ln(n_{rj}/p_{rj}) + \ln(n_{rc}/p_{rc}) = 0 \quad \text{for all } i,j. \qquad (1.6.11)$$

For $0 \leq n_{ij}/p_{ij} \leq 2n_{..}$ which will be true with high probability for $n_{..}$ large enough and $p_{ij} \neq 0$ for all i and j, we can expand each of the logarithmic terms in a Taylor series about $n_{..}$. Equation 1.6.11 can then be written as

1.6

$$0 = \ln(n_{..}) + \sum_{k=1}^{\infty} (-1)^{k+1}(n_{ij}/p_{ij} - n_{..})^k/(kn_{..}^k)$$

$$- \ln(n_{..}) - \sum_{k=1}^{\infty} (-1)^{k+1}(n_{ic}/p_{ic} - n_{..})^k/(kn_{..}^k)$$

$$- \ln(n_{..}) - \sum_{k=1}^{\infty} (-1)^{k+1}(n_{rj}/p_{rj} - n_{..})^k/(kn_{..}^k)$$

$$+ \ln(n_{..}) + \sum_{k=1}^{\infty} (-1)^{k+1}(n_{rc}/p_{rc} - n_{..})^k/(kn_{..}^k) \ .$$

This can be reduced to

$$0 = (n_{ij}/p_{ij} - n_{..})/n_{..} - (n_{ic}/p_{ic} - n_{..})/n_{..} - (n_{rj}/p_{rj} - n_{..})/n_{..}$$

$$+ (n_{rc}/p_{rc} - n_{..})/n_{..} + \sum_{k=2}^{\infty} [(-1)^{k+1}/(kn_{..}^k)][(n_{ij}/p_{ij} - n_{..})^k$$

$$- (n_{ic}/p_{ic} - n_{..})^k - (n_{rj}/p_{rj} - n_{..})^k + (n_{rc}/p_{rc} - n_{..})^k]$$

$$0 = n_{ij}/p_{ij} - n_{ic}/p_{ic} - n_{rj}/p_{rj} + n_{rc}/p_{rc} + \sum_{k=2}^{\infty} [(-1)^{k+1}/(kn_{..}^{k-1})]$$

$$[(n_{ij}/p_{ij} - n_{..})^k - (n_{ic}/p_{ic} - n_{..})^k - (n_{rj}/p_{rj} - n_{..})^k$$

$+ (n_{rc}/p_{rc} - n_{..})^k]$. Since we have $0 \le n_{ij}/p_{ij} \le 2n_{..}$ and

$$(n_{ij} - n_{ij}^{\frac{1}{2}})/n_{..} \le p_{ij} \le (n_{ij} + n_{ij}^{\frac{1}{2}})/n_{..} \quad \text{for all } i \text{ and } j,$$

$(n_{ij}/p_{ij} - n_{..})^k$ is of the order of $n_{..}^k/n_{ij}^{\frac{1}{2}k}$ or $n_{..}^{\frac{1}{2}k}$.

Since the series converges, all terms beyond k = 2 must be smaller than the terms at k = 2. Therefore the series is $0(n_{..}/n_{..}) = 0(1)$. The

terms of the form n_{ij}/p_{ij} are all of order $0(n_{..})$, so, as $n_{..} \to \infty$, the terms of order 0(1) become relatively insignificant. Thus we have

$$0 = n_{ij}/p_{ij} - n_{ic}/p_{ic} - n_{rj}/p_{rj} + n_{rc}/p_{rc} + 0(1). \qquad (1.6.12)$$

We therefore have that the Minimum Discrimination Information estimates satisfy the equation for Maximum Likelihood to 0(1). Note that "large" in this situation means that the minimum n_{ij} must be greater than or equal to 4 for the theorem to hold. This lower limit is obtained through combining $n_{ij}/p_{ij} \leq 2n_{..}$ and $(n_{ij} - n_{ij}^{\frac{1}{2}})/n_{..} \leq p_{ij} \leq (n_{ij} + n_{ij}^{\frac{1}{2}})/n_{..}$, both of which will be true with high probability for $n_{..}$ large enough and $p_{ij} \neq 0$ for all i and j.

Theorem 1.6.3 Given an r x c contingency table with cell frequencies $\{n_{ij}\}$ such that $n_{ij} \neq 0$ for all i and j, and a set of constraints on the margins. Then the Minimum Discrimination Information method yields probability estimates which satisfy to 0(1) the equation in Table 1.6.2 for the method of Minimum Modified Chi Squared.

Proof:

We can write the necessary equations for Minimum Discrimination Information as given in Table 1.6.2. Expanding each of the logarithmic terms in a Taylor series about $n_{..}^{-1}$ similar to the proof of theorem 1.6.2 yields the desired result.

This theorem is also a large-sample result, although not as restrictive as in the previous theorem. The necessary conditions are now $0 \leq p_{ij}/n_{ij} \leq 2/n_{..}$ which reduces to $n_{ij} \geq 1$.

TABLE 1.6.1

Comparison of Estimator Forms for Optimization Methods under Varying Conditions

(α_i and β_j different for each method)

Method	Objective Function	Estimator Forms (p_{ij} = ...) no margins fixed	row margin fixed	two margins fixed
Maximum Likelihood	$\sum_i \sum_j (n_{ij}/n_{..}) \ln(\pi_{ij})$	$n_{ij}/n_{..}$	$a_i n_{ij}/n_{i\cdot}$	$(n_{ij}/n_{..})(\alpha_i + \beta_j)^{-1}$
Minimum χ^2 (Pearson's χ^2)	$\sum_i \sum_j (n_{ij} - n_{..}\pi_{ij})^2/(n_{..}\pi_{ij})$	$n_{ij}/n_{..}$	$a_i n_{ij}/n_{i\cdot}$	$(n_{ij}/n_{..})(\alpha_i + \beta_j)^{-\frac{1}{2}}$
Minimum Modified χ^2 (Neyman's χ^2)	$\sum_i \sum_j (n_{ij} - n_{..}\pi_{ij})^2/n_{ij}$	$n_{ij}/n_{..}$	$a_i n_{ij}/n_{i\cdot}$	$(n_{ij}/n_{..})(\alpha_i + \beta_j)$
Minimum Discrimination Information	$\sum_i \sum_j \pi_{ij} \ln(n_{..}\pi_{ij}/n_{ij})$	$n_{ij}/n_{..}$	$a_i n_{ij}/n_{i\cdot}$	$(n_{ij}/n_{..})\alpha_i\beta_j$
Maximum Entropy	$-\sum_i \sum_j \pi_{ij} \ln(\pi_{ij})$	$(rc)^{-1}$	a_i/c	$\alpha_i\beta_j$

TABLE 1.6.2

Necessary Condition Equations for Varying Methods of Estimation for Probabilities in a r x c Contingency Table with Fixed Margins

Method	Equation
Maximum Likelihood	$n_{ij}/p_{ij} - n_{ic}/p_{ic} - n_{rj}/p_{rj} + n_{rc}/p_{rc} = 0$
Minimum χ^2 (Pearson's χ^2)	$(n_{ij}/p_{ij})^2 - (n_{ic}/p_{ic})^2 - (n_{rj}/p_{rj})^2 + (n_{rc}/p_{rc})^2 = 0$
Minimum Modified χ^2 (Neyman's χ^2)	$p_{ij}/n_{ij} - p_{ic}/n_{ic} - p_{rj}/n_{rj} + p_{rc}/n_{rc} = 0$
Minimum Discrimination Information	$\ln(n_{ij}/p_{ij}) - \ln(n_{ic}/p_{ic}) - \ln(n_{rj}/p_{rj}) + \ln(n_{rc}/p_{rc}) = 0$
Maximum Entropy	$\ln(p_{ij}) - \ln(p_{ic}) - \ln(p_{rj}) + \ln(p_{rc}) = 0$

CHAPTER 2

The Maximum Likelihood/Entropy (ML/E) Estimator

This chapter deals with Maximum Likelihood/Entropy estimation - its history, forms, and derivation. Two different estimators for the hyper-parameter λ are derived and discussed. Finally the ML/E method is defined in terms of optimizing a specified function of the probabilities and sample frequencies subject to the particular constraints being imposed.

2.1 History of Maximum Likelihood/Entropy Estimation

The purpose of the Maximum Likelihood/Entropy estimation procedure was to generalize the Maximum Likelihood (ML) technique to situations when small probabilities are to be estimated and the standard ML estimator is inadequate. In addition when no sample exists the technique should give meaningful results by reducing to a method of some interest, e.g., the method of Maximum Entropy.

The idea of maximizing some linear combination of the entropy and the log-likelihood first appeared in Good (1963) in which it is presented as

$$\text{Maximize} \quad \sum_i \sum_j (n_{ij} - \pi_{ij}) \ln(\pi_{ij})$$

subject to the restraints. He mentioned that this method would presumably estimate π_{ij} somewhere between $\pi_{i\cdot}\,\pi_{\cdot j}$ and n_{ij}/n , thereby resembling the methods of Good (1956).

Good (1965) describes this procedure as equivalent to the selection of the distribution of maximum final credibility, assuming the logarithm

2.1

of the initial density to be proportional to the entropy, that is, the initial density is proportional to $\prod_i \pi_i^{-k\pi_i}$. This estimation procedure was compared to the Bayes-Laplace estimates and Perks-Jeffreys estimates for varying cases in the binomial situation. Good (1975b) mentions that this prior does not depend on the size of the sample. Good (1969a) suggests the use of this technique when sampling word digraphs. Good and Gaskins (1972) generalize the concept to continuous distributions by using a "roughness penalty" based on derivatives of the density function rather than on entropy.

Chew (1971) wished to estimate the probability of success in a reliability problem after 14 previous trials were successes. He compared a number of techniques on the basis of reasonableness of the calculated estimates, p, of success. He found that Maximum Likelihood resulted in p = 1, intuitively too high. The Maximum Information estimate described in Good (1965) was the solution to

$$\ln(p/(1 - p)) = x/p - (N - x)/(1 - p)$$

where x is the number of successes in N trials. Chew calculated this p as 0.93. He also considered the Bayes estimate which for a uniform prior leads to Laplace's Law of Succession (p = 0.9375), as well as Beta priors with estimator form

$$p = (\alpha + x + 1)/(\alpha + \beta + N + 2)$$

where α and β are the parameters of the Beta distribution. Finally he examined an estimator due to Steinhaus (1957) which minimized the expected (long-run) loss, rather than minimizing the maximum loss at

each stage. This estimator has the form

$$p = (x + N^{\frac{1}{2}}/2)/(N + N^{\frac{1}{2}})$$

and value p = 0.8946 for this case. Note that for N = 4 this estimator has the same form as the Bayes-Laplace estimator and for N = 1 the form is the same as the Perks-Jeffreys estimator. Good (1972) discusses the historical background, logic, improvements, and extensions of Chew's results.

2.2 Derivation of Type II Likelihood Function

The derivation of the Maximum Likelihood/Entropy (ML/E) estimation procedure is based on the Inverse Probability Theorem quoted in Perks (1947), and in Jeffreys (1961) where the result is attributed to Bayes. It states that the posterior probability is proportional to the prior probability times the likelihood, that is,

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}.$$

The method of Type II Maximum Likelihood (Good (1965)) is then used to calculate the probability estimates. This particular procedure is termed "Bayesian in Mufti" by Good and Gaskins (1972) or Pseudo-Bayes by Fienberg and Holland (1973). A full dress Bayesian approach would involve the maximization of an expected utility. Here we make no assumptions concerning the utility and instead maximize the posterior probability as a function of a hyperparameter of the population probabilities.

2.2

First consider the case for a t-category multinomial. Let $P((\pi_i)|\lambda)$ and $P((n_i)|(\pi_i))$ represent the prior probability and the likelihood respectively, where λ is a hyperparameter. Then the Type II likelihood $P((n_i)|\lambda)$ can be expressed as

$$P((n_i)|\lambda) = \int\int\ldots\int_{\substack{t-1 \\ \sum_{i=1} \pi_i \leq 1}} P((n_i)|(\pi_i))P((\pi_i)|\lambda)\, d\pi_1\ldots d\pi_{t-1}. \qquad (2.2.1)$$

Under the proper sampling conditions, the likelihood is given by

$$P((n_i)|(\pi_i)) = (n_.!/\textstyle\prod_i n_i!)\prod_i \pi_i^{n_i}.$$

Now assume the prior distribution is proportional to $\prod_i \pi_i^{-\lambda\pi_i}$. Then $P((\pi_i)|\lambda) = G(\lambda)\prod_i \pi_i^{-\lambda\pi_i}$ where $[G(\lambda)]^{-1} = \int\ldots\int_{\substack{t-1 \\ \sum_{i=1} \pi_i \leq 1}} \prod_i \pi_i^{-\lambda\pi_i}\, d\pi_1\ldots d\pi_{t-1}$.

Therefore

$$P((n_i)|\lambda) = \left(\frac{n_.!}{\prod_i n_i!}\right) \frac{\int\ldots\int_{\substack{t-1 \\ \sum_{i=1} \pi_i \leq 1}} \prod_{i=1}^{t} \pi_i^{n_i-\lambda\pi_i}\, d\pi_1\ldots d\pi_{t-1}}{\int\ldots\int_{\substack{t-1 \\ \sum_{i=1} \pi_i \leq 1}} \prod_{i=1}^{t} \pi_i^{-\lambda\pi_i}\, d\pi_1\ldots d\pi_{t-1}} \qquad (2.2.2)$$

and λ^* is that value of λ which maximizes $P((n_i)|\lambda)$. λ^* is called the Type II Maximum Likelihood estimator for λ.

2.3 Derivation of Posterior Distribution for $\{\pi_i\}$ (1)

If an ordinary prior for the $\{\pi_i\}$ is assumed, then we are concerned with maximizing the posterior density of the $\{\pi_i\}$, that is,

$$\underset{\{\pi_i\}}{\text{Max}} \quad P((\pi_i)|(n_i),\lambda). \tag{2.3.1}$$

However

$$P((\pi_i)|(n_i),\lambda) = P((\pi_i),(n_i),\lambda)/P((n_i),\lambda)$$

$$= P((\pi_i),(n_i)|\lambda)P(\lambda)/P((n_i)|\lambda)P(\lambda)$$

$$= P((n_i)|(\pi_i))P((\pi_i)|\lambda)/P((n_i)|\lambda). \tag{2.3.2}$$

Now

$$\ln(P((\pi_i)|(n_i),\lambda)) = \ln(P((n_i)|(\pi_i))) + \ln(P((\pi_i)|\lambda)) - \ln(P((n_i)|\lambda))$$

$$= K_1 + \sum_{i=1}^{t} n_i \ln(\pi_i) + K_2 - \sum_{i=1}^{t} \lambda\pi_i \ln(\pi_i) - K_3 \tag{2.3.3}$$

where K_1, K_2, and K_3 do not depend on $\{\pi_i\}$. Finally maximizing the posterior with respect to $\{\pi_i\}$ is equivalent to maximizing the logarithm of the posterior with respect to $\{\pi_i\}$, so $\underset{\{\pi_i\}}{\text{Max}}\ \ln(P((\pi_i)|(n_i),\lambda)$ is equivalent to

$$\underset{\{\pi_i\}}{\text{Max}} \quad \sum_{i=1}^{t} (n_i - \lambda\pi_i)\ln(\pi_i) \ . \tag{2.3.4}$$

---

(1) Instead of using the ordinary notation for contingency tables, in this section we will use the vector notation given in Section 1.1.

2.4 Definition of Maximum Likelihood/Entropy Estimation

The method of Maximum Likelihood/Entropy for an r x c contingency table is defined as

$$\underset{\{\pi_{ij}\}}{\text{Maximizing}} \qquad Q = \sum_{i=1}^{r} \sum_{j=1}^{c} (n_{ij} - \lambda\pi_{ij})\ln(\pi_{ij})$$

subject to the particular set of constraints associated with the problem under consideration. λ is a hyperparameter of the parent population which may be estimated from the sample by the methods given in Section 2.5, or may be defined as having a specific value based on other considerations.

The set of constraints does not have to form a set of linearly independent conditions, though for simplicity this is to be preferred. There may be numerical methods which require all conditions to be linearly independent. The constraints must be consistent, that is, the set of constraints must be such that a solution does exist. If the constraints are "only just" consistent in the sense of Good (1963), estimates still exist, but some estimates may be zero.

It is possible to extend the ML/E method to other log-likelihood functions. For example, if both rows and columns are fixed and used to terminate sampling, the log-likelihood for an r x c contingencey table would be

$$K + \sum_i \sum_j n_{ij}\ln(\pi_{ij}) - \sum_i n_{i\cdot}\ln(\pi_{i\cdot}) - \sum_j n_{\cdot j}\ln(\pi_{\cdot j})$$

2.4

for constant K. If $\lambda \sum_i \sum_j \pi_{ij} \ln(\pi_{ij})$ were then subtracted from the log-likelihood function, we would have the objective function for ML/E estimation.

2.5 Estimation of λ^* (1)

The multidimensional integrals given in equation 2.2.2 reduce for the binomial distribution to univariate integrals which are then relatively easy to calculate using known numerical techniques, for example, Romberg Integration as described in Gerald (1970). The double integrals associated with the trinomial distribution become more difficult to manipulate but are still manageable.

Increasing the dimensionality of the integrals further makes the problem much more complex. Methods for multiple integration have been developed, but are often not applicable to particular situations. As an example, we may consider the Centroid Method of Numerical Integration presented by Good and Gaskins (1969, 1971), one of the simpler methods. Unfortunately as the number of categories increases, the storage requirements get exceedingly large. Good and Gaskins (1969) mention that for single precision accuracy, the one-point centroid method using four terms requires approximately $30000 + 8p^5(p - 1)^{-1}$ bytes of storage where p is the dimensionality of the integration. For a 20-category Multinomial this would require 1.4 million bytes, which is a considerable percentage of the available core storage on most current computers.

---

(1) Instead of using the ordinary notation for contingency tables, in this section we will use the vector notation given in Section 1.1.

2.5

Simple approximations to $\prod_i \pi_i^{-\lambda\pi_i}$ and $\prod_i \pi_i^{n_i-\lambda\pi_i}$ apparently are not sufficiently accurate for evaluating the integrals.

An application of Schwarz's Inequality for integrals appeared promising. The inequality states that

$$[\int_a^b f(x)g(x)dx]^2 \leq \int_a^b [f(x)]^2dx \int_a^b [g(x)]^2dx. \qquad (2.5.1)$$

Applying the inequality to the numerator of equation 2.2.2 yields

$$\left[\int_{t-1}\cdots\int_{\substack{\sum_{i=1}^{t-1}\pi_i \leq 1}} \prod_{i=1}^{t} \pi_i^{n_i-\lambda\pi_i}\, d\pi_1\cdots d\pi_{t-1}\right]^2 \leq$$

$$\left[\int_{t-1}\cdots\int_{\substack{\sum_{i=1}^{t-1}\pi_i \leq 1}} \prod_{i=1}^{t-1} \pi_i^{-2\lambda\pi_i}\, d\pi_i\cdots d\pi_{t-1}\right]\left[\int_{t-1}\cdots\int_{\substack{\sum_{i=1}^{t-1}\pi_i \leq 1}} \pi_t^{2n_t-2\lambda\pi_t} \prod_{i=1}^{t-1} \pi_i^{2n_i}\, d\pi_1\cdots d\pi_{t-1}\right] \qquad (2.5.2)$$

By Dirichlet's Multiple Integral (see Gradshteyn and Ryshik (1965), eqn. 4.635.2, p. 621), the second integral on the right becomes

$$\frac{\prod_{i=1}^{t-1}(2n_i)!}{\Gamma(2n_\cdot - 2n_t + t - 1)} \int_0^1 (1-x)^{2n_t-2\lambda(1-x)}\, x^{2n_\cdot-2n_t+t-2}\, dx\,. \qquad (2.5.3)$$

2.5

Now

$$\int_0^1 (1-x)^{2n_t-2\lambda(1-x)}\, x^{2n_\cdot-2n_t+t-2}\, dx$$

$$= \int_0^1 x^{2n_t-2\lambda x}\, (1-x)^{2n_\cdot-2n_t+t-2}\, dx$$

$$= \int_0^1 x^{2n_t} \exp(-2\lambda x \ln x)\, (1-x)^{2n_\cdot-2n_t+t-2}\, dx$$

$$= \int_0^1 x^{2n_t} \sum_{m=0}^{\infty} (-2\lambda)^m \frac{(x\ln x)^m}{m!}\, (1-x)^{2n_\cdot-2n_t+t-2}\, dx$$

$$= \sum_{m=0}^{\infty} \frac{(-2\lambda)^m}{m!} \int_0^1 x^{2n_t+m} (\ln x)^m\, (1-x)^{2n_\cdot-2n_t+t-2}\, dx$$

$$= \sum_{m=0}^{\infty} \frac{(-2\lambda)^m}{m!} (-1)^m\, m! \sum_{k=0}^{2n_\cdot-2n_t+t-2} \binom{2n_\cdot-2n_t+t-2}{k} \frac{(-1)^k}{(2n_t+m+k+1)^{m+1}}$$

$$= \sum_{m=0}^{\infty} (2\lambda)^m \sum_{k=0}^{2n_\cdot-2n_t+t-2} \binom{2n_\cdot-2n_t+t-2}{k} \frac{(-1)^k}{(2n_t+m+k+1)^{m+1}} \quad . \qquad (2.5.4)$$

Note that in the above derivation the one-dimensional integral may be found in Gradshteyn and Ryshik (1965, eqn. 16, p. 551).

Combining equations 2.5.3 and 2.5.4 gives

$$\prod_{i=1}^{t-1} \frac{(2n_i)!}{\Gamma(2n_\cdot-2n_t+t-1)} \sum_{m=0}^{\infty} (2\lambda)^m \sum_{k=0}^{2n_\cdot-2n_t+t-2} \binom{2n_\cdot-2n_t+t-2}{k} \frac{(-1)^k}{(2n_t+m+k+1)^{m+1}} \quad . \qquad (2.5.5)$$

2.5

In this derivation, π_t was given some importance by letting

$$\pi_t = 1 - \sum_{i=1}^{t-1} \pi_i \ .$$

However any π_j may be expressed in this manner, so, to restore the symmetry, we must average equation 2.5.5 over all j, and obtain

$$t^{-1} \sum_{j=1}^{t} \frac{\prod_{i\neq j} (2n_i)!}{\Gamma(2n_\cdot - 2n_j + t - 1)} \sum_{m=0}^{\infty} (2\lambda)^m \sum_{k=0}^{2n_\cdot - 2n_j + t - 2} \binom{2n_\cdot - 2n_j + t - 2}{k} \frac{(-1)^k}{(2n_j + m + k + 1)^{m+1}} \ . \tag{2.5.6}$$

To obtain a similar result for the denominator in equation 2.2.2, let $n_i = 0$ for all i. Then equation 2.5.6 reduces to

$$\frac{1}{\Gamma(t-1)} \sum_{m=0}^{\infty} (2\lambda)^m \sum_{k=0}^{t-2} \binom{t-2}{k} \frac{(-1)^k}{(m+k+1)^{m+1}} \ . \tag{2.5.7}$$

By using the ratio of the two inequalities, we should have a reasonable approximation to the square of the posterior, that is,

$$[P((n_i)|\lambda)]^2 \approx \frac{\Gamma(t-1)}{t} \times$$

$$\frac{\sum_{j=i}^{t} \frac{\prod_{i\neq j} (2n_i)!}{\Gamma(2n_\cdot - 2n_j + t - 1)} \sum_{m=0}^{\infty} (2\lambda)^m \sum_{k=0}^{2n_\cdot - 2n_j + t - 2} \binom{2n_\cdot - 2n_j + t - 2}{k} \frac{(-1)^k}{(2n_t + m + k + 1)^{m+1}}}{\sum_{m=0}^{\infty} (2\lambda)^m \sum_{k=0}^{t-2} \binom{t-2}{k} \frac{(-1)^k}{(m+k+1)^{m+1}}} \tag{2.5.8}$$

Therefore maximizing equation 2.5.8 with respect to λ should be approximately equivalent to maximizing equation 2.2.2 with respect to λ. This was checked against the exact integral for the binomial case and equation 2.5.8 was found to be a reasonable approximation.

However there is one difficulty with this method. Note that the numerator in equation 2.5.8 has a triple summation with Binomial coefficients in the terms. For a problem with either a large sample size or a large number of categories, this method also becomes impractical.

The last method to be considered here evolved from the resemblance between the estimator for π_i given in Good (1965) and the estimator for ML/E in the multinomial case derived in section 3.1. Good found that the Type II Maximum Likelihood estimation procedure using a symmetric Dirichlet prior leads to

$$p_i = \frac{n_i + k^*}{n_\cdot + tk^*} \tag{2.5.9}$$

where k^* maximizes

$$\frac{n_\cdot!\ \Gamma(tk)\ \prod_{i=1}^{t} \Gamma(n_i+k)}{\prod_{i=1}^{t} n_i!\ (\Gamma(k))^t\ \Gamma(n_\cdot+tk)} \tag{2.5.10}$$

In section 3.1, we derive

$$p_i = \frac{n_i - \lambda^* p_i \ln p_i}{n_\cdot - \lambda^* \sum_{i=1}^{t} p_i \ln p_i} \tag{2.5.11}$$

2.5

as the ML/E estimator for the multinomial case. As a first approximation let $k^* = -\lambda^* \sum p_i \ln p_i$ and $tk^* = -\lambda^* \sum_{i=1}^{t} p_i \ln p_i$. Then

$$\lambda^* = - tk^* / \left(\sum_{i=1}^{t} p_i \ln p_i \right) . \qquad (2.5.12)$$

Checks against the exact integral for the binomial using this estimator were within 5%. Comparisons against competitive methods in section 4.5 demonstrate its adequacy for estimation. Equation 2.5.12 will be used as the form for λ^* in the later discussions of the ML/E method.

A comment on the estimation should be made. All of the developments are based on estimating λ^* by treating the $\{n_i\}$ as if a multinomial sampling scheme had generated it. However the actual sampling model may be quite different. The question then arises as to the effect of the sampling model on the value of λ^*. For similar situations, Zellner (1971) prefers using priors which are independent of sample size, but dependent on the sampling procedure. Good (1976a) states his view that "in principle" priors should not depend on the sample procedure.

In this case, the true (but unknown) λ is related to the roughness in the population and is not affected by the sample. The estimate λ^* should therefore measure only the roughness of the $\{n_i\}$ independent of how it is obtained. We therefore follow Good in ignoring the sampling procedure when estimating λ with λ^*.

CHAPTER 3

Estimation Using ML/E

In this chapter we analyze the algorithms to be used for ML/E estimation. Algorithms are presented for the two cases for which the estimation procedure can be reduced to a simple iterative scheme. For the more general case, two methods for calculating the probability estimates are presented. The first is a more complicated iterative scheme analogous to, though more complicated than, the IPFP. We call this new iterative scheme the Iterative Maximum Likelihood/Entropy Scaling, abbreviated as IML/ES. The second method is a general non-linear optimization procedure adapted to this situation.

3.1 Sample Size Fixed

Estimation when $\sum_i \sum_j \pi_{ij} = 1$ is the only constraint occurs when the contingency table is constructed through multinomial or Poisson sampling and there is no marginal information which the sampling procedure must use.

Once λ^* has been calculated, we may obtain the probability estimates by maximizing

$$Q = \sum_i \sum_j (n_{ij} - \lambda^* \pi_{ij}) \ln(\pi_{ij}) - \alpha\left(\sum_i \sum_j \pi_{ij} - 1\right) . \qquad (3.1.1)$$

Now

$$\frac{\partial Q}{\partial \pi_{ij}} = n_{ij}/\pi_{ij} - \lambda^* - \lambda^* \ln(\pi_{ij}) - \alpha \qquad (3.1.2)$$

and

$$0 = n_{ij}/p_{ij} - \lambda^* - \lambda^* \ln(p_{ij}) - \alpha \cdot \qquad (3.1.3)$$

Rearranging terms and taking summation over i and j yields

$$\alpha = n_{..} - \lambda^* - \lambda^* \sum_i \sum_j p_{ij} \ln(p_{ij}) \qquad (3.1.4)$$

and substituting back into equation 3.1.3 gives

$$p_{ij} = \frac{n_{ij} - \lambda^* p_{ij} \ln(p_{ij})}{n_{..} - \lambda^* \sum_i \sum_j p_{ij} \ln(p_{ij})} \quad \text{for all } i,j. \qquad (3.1.5)$$

Compare this formula to those given in Table 1.5.1 under the "no margins fixed" heading. Note that equation 3.1.5 requires an iterative procedure to calculate the set of probability estimates. By calculating $m_{ij} = n_{..} p_{ij}$ for all i and j, we obtain the smoothed frequencies.

The calculating formula given in equation 3.1.5 can be used instead of the more general and therefore more time-consuming method described in section 3.3.

Good (1965, p. 38) gives an example for which $n_1 = 0$, $n_2 = 2$, $n_3 = 3$, $n_4 = 3$, and $n_5 = 12$. The smoothed frequencies using the method described in the monograph are 0.70, 2.35, 3.18, 3.18, and 10.59, respectively, for $k^* = 0.8531$. The corresponding smoothed frequencies using ML/E and λ^* are $m_1 = 0.04$, $m_2 = 2.42$, $m_3 = 3.38$, $m_4 = 3.38$, and $m_5 = 10.78$. The ratio of the two posterior distributions assuming that the entropy prior is in fact true is 1.413.

3.2 One Margin Fixed

Estimation when one set of marginal totals, say rows, is fixed can occur in a number of situations, for example,

1. sampling until the row totals are equal to a previously specified set of values (product multinomial sampling),
2. adjusting the sample row marginal totals to a set of marginal totals known to be proportional to the population marginal probabilities,
3. adjusting the sample row marginal totals to the marginal totals of another contingency table obtained previously.

For a given value of λ^* we may obtain the probability estimates by maximizing

$$Q = \sum_i \sum_j (n_{ij} - \lambda^* \pi_{ij}) \ln(\pi_{ij}) - \sum_i \alpha_i (\sum_j \pi_{ij} - a_i) \qquad (3.2.1)$$

where $\{\alpha_i\}$ is the set of Lagrangian multipliers, $\{a_i\}$ is the set of margins to be fitted, and $\sum_i a_i = 1$. Following a procedure similar to that in Section 3.1 gives

$$\frac{\partial Q}{\partial \pi_{ij}} = n_{ij}/\pi_{ij} - \lambda^* - \lambda^* \ln(\pi_{ij}) - \alpha_i, \qquad (3.2.2)$$

$$0 = n_{ij}/p_{ij} - \lambda^* - \lambda^* \ln(p_{ij}) - \alpha_i, \qquad (3.2.3)$$

$$\alpha_i p_{ij} = n_{ij} - \lambda^* p_{ij} - \lambda^* p_{ij} \ln(p_{ij}), \qquad (3.2.4)$$

3.2

and remembering that $\sum_j p_{ij} = a_i$,

$$\alpha_i a_i = n_{i\cdot} - \lambda^* a_i - \lambda^* \sum_j p_{ij} \ln(p_{ij}). \qquad (3.2.5$$

Substituting equation 3.2.5 back into 3.2.3 gives

$$p_{ij} = \frac{a_i(n_{ij} - \lambda^* p_{ij} \ln(p_{ij}))}{n_{i\cdot} - \lambda^* \sum_j p_{ij} \ln(p_{ij})} \quad \text{for all } i,j. \qquad (3.2.6)$$

Compare this formula to those given in Table 1.5.1 under the "row margin fixed" heading. We again require an iterative procedure to calculate the probability estimates. Equation 3.2.6 may be used in place of the more general technique described in section 3.3 when the proper conditions hold. Finally the smoothed frequencies may be obtained by calculating $m_{ij} = n_{\cdot\cdot} p_{ij}$.

If the sampling were done using the product multinomial model with say, rows fixed, then the log-likelihood would be

$$K + \sum_i \sum_j n_{ij} \ln(\pi_{ij}) - \sum_i n_{i\cdot} \ln(\pi_{i\cdot}) \quad \text{for constant } K.$$

However $p_{i\cdot} = \sum_j p_{ij} = a_i$, so the estimator still has the form given in equation 3.2.6.

3.3 Two Margins Fixed

Estimation when both margins are fixed can occur in similar situations to those mentioned in Section 3.2. Now, however, two sets of

marginal totals are to be fitted. Note that if both margins are fixed a priori and are used to terminate the sampling process, a significance test for independence should be conducted. If the hypothesis of independence can not be rejected the table can be separated into its component parts. If the hypothesis is rejected we can still use ML/E to estimate the probabilities since the kernel of the log-likelihood function is in the proper form.

The problem is now one of maximizing

$$Q = \sum_i \sum_j (n_{ij} - \lambda^* \pi_{ij}) \ln(\pi_{ij}) - \sum_i \mu_i (\sum_j \pi_{ij} - a_i) - \sum_j \nu_j (\sum_i \pi_{ij} - b_j). \tag{3.3.1}$$

The partial derivatives with respect to the π_{ij} are

$$\frac{\partial Q}{\partial \pi_{ij}} = n_{ij}/\pi_{ij} - \lambda^* - \lambda^* \ln(\pi_{ij}) - \mu_i - \nu_j, \tag{3.3.2}$$

therefore

$$0 = n_{ij}/p_{ij} - \lambda^* - \lambda^* \ln(p_{ij}) - \mu_i - \nu_j, \tag{3.3.3}$$

and

$$p_{ij} = \frac{n_{ij} - \lambda^* p_{ij}(1 + \ln(p_{ij}))}{(\mu_i + \nu_j)} \quad \text{for all } i,j. \tag{3.3.4}$$

Letting $\mu_i = n_{..}\alpha_i$ and $\nu_j = n_{..}\beta_j$ gives

$$p_{ij} = \frac{n_{ij} - \lambda^* p_{ij}(1 + \ln(p_{ij}))}{n_{..}(\alpha_i + \beta_j)} \quad \text{for all } i,j. \tag{3.3.5}$$

Compare this equation to those in Table 1.5.1 under the "two margins fixed" heading. The corresponding equation to those in Table 1.5.2 is

$$n_{ij}/p_{ij} - n_{ic}/p_{ic} - n_{rj}/p_{rj} + n_{rc}/p_{rc} - \lambda^* \ln(p_{ij}p_{rc}/p_{ic}p_{rj}) = 0 \tag{3.3.6}$$

Equations 3.3.4 require not only some kind of iterative scheme, but are also functions of the Lagrangian multipliers α_i and β_j which cannot be removed by substitution. We therefore require a more general algorithm for solution of the problem than was given for the cases with sample size and one margin fixed. We derive such a method below.

Rewriting equation 3.3.3 gives

$$0 = n_{ij} - \lambda^* p_{ij} - \lambda^* p_{ij} \ln(p_{ij}) - \mu_i p_{ij} - \nu_j p_{ij} \, . \tag{3.3.7}$$

Summation over j gives

$$0 = n_{i\cdot} - \lambda^* a_i - \lambda^* \sum_j p_{ij} \ln(p_{ij}) - \mu_i a_i - \sum_j \nu_j p_{ij}$$

where $\sum_j p_{ij} = a_i$ or

$$\mu_i = (n_{i\cdot} - \lambda^* a_i - \lambda^* \sum_j p_{ij} \ln(p_{ij}) - \sum_j \nu_j p_{ij})/a_i \, . \tag{3.3.8}$$

Similarly, summation over i gives

$$\nu_j = (n_{\cdot j} - \lambda^* b_j - \lambda^* \sum_i p_{ij} \ln(p_{ij}) - \sum_i \mu_i p_{ij})/b_j \ , \qquad (3.3.9)$$

where $\sum_i p_{ij} = b_j$. Equations (3.3.8) and (3.3.9) form r + c equations; however $\sum_i a_i = \sum_j b_j = 1$, so only r + c - 1 are linearly independent.

Therefore equations (3.3.7)-(3.3.9) form a system of rc + r + c - 1 non-linear equations in as many unknowns. We can solve this set of equations using the following algorithm, termed the Iterative Maximum Likelihood/Entropy Scaling (IML/ES) procedure:

1. Set $\nu_c = 0$ and $k = 0$,
2. Calculate initial probability estimates so that $p_{ij}^{(k)} \neq 0$ for all i and j, e.g.,

   $$p_{ij}^{(k)} = (n_{ij} + (rc)^{-1})/(n_{\cdot\cdot} + 1) \quad \text{for all i and j,}$$

3. Calculate

   $$\mu_i^{(k)} = n_{ic}/p_{ic}^{(k)} - \lambda^* - \lambda^* \ln(p_{ic}^{(k)}) \quad \text{for } i = 1,.., r,$$

4. Calculate

   $$\nu_j^{(k+1)} = (n_{\cdot j} - \lambda^* b_j - \lambda^* \sum_i p_{ij}^{(k)} \ln(p_{ij}^{(k)}) - \sum_i \mu_i^{(k)} p_{ij}^{(k)})/b_j$$

   $$j = 1,.., c-1,$$

3.3

5. Calculate

$$\mu_i^{(k+1)} = (n_{i\cdot} - \lambda^* a_i - \lambda^* \sum_j p_{ij}^{(k)} \ln(p_{ij}^{(k)}) - \sum_j \nu_j^{(k+1)} p_{ij}^{(k)})/a_i \qquad i = 1,.., r,$$

6. Calculate

$$p_{ij}^{(k+1)} = (n_{ij} - \lambda^* p_{ij}^{(k)} \ln(p_{ij}^{(k)}))/(\lambda^* + \mu_i^{(k+1)} + \nu_j^{(k+1)}) \qquad i = 1,.., r \quad j = 1,.., c,$$

7. If $p_{ij}^{(k+1)}$ is sufficiently close to $p_{ij}^{(k)}$ for all i and j, stop. Otherwise set k = k + 1 and go to step 4.

This algorithm was used to calculate the ML/E estimates for the observed table, Table 3.3.1(a). For $\lambda^* = 50$ and the fitted margins in the table, the ML/E estimates are given in Table 3.3.1(b).

In addition to the IML/ES procedure, a different method from the area of non-linear programming was considered. This procedure is the Generalized Reduced Gradient (GRG) method of Lasdon et al (1973, 1975a, 1975b, 1975c). The GRG method treats a general non-linear optimization problem in the same manner as the Simplex Method of linear programming treats linear optimization problems. An initial feasible point is used to describe a set of basic and non-basic variables which in some neighborhood of the feasible point can be used to optimize the objective function for a specified range of the basic variables. GRG then solves the original problem by solving a sequence of reduced problems using unconstrained minimization algorithms. After a reduced problem is solved, GRG examines the basic variables to

discover any violations of inequality constraints. If such a violation occurs, this variable is removed from the basis and a new variable not on a boundary is placed in the basis. After this change of basis, there is now a new reduced problem to be solved. This operation continues until one of a set of specified stopping rules terminates the iteration in the neighborhood of the optimal solution.

Ireland and Kullback (1968) give an observed contingency table shown in Table 3.3.2(a) and describe the MDI smoothed frequencies for given population row and column margin totals given in Table 3.3.2(b). The corresponding ML/E estimates obtained by the GRG method use $\lambda^* = 5.599$ calculated from the data and have the same values as those given in Table 3.3.2(b). This similarity is not surprising since $\lambda^*/19175$ is small, where 19175 is the sample size. For such a large sample we would not expect a large effect due to λ^*; in fact, the asymptotic properties of ML/E should become evident in such a situation.

If we wished to smooth the data keeping the margins fixed at their sample values, we would obtain Table 3.3.2(a) as the ML/E estimates with $\lambda^* = 5.599$ and the MDI estimates, that is, we would not change the observed frequencies. To be a bit more precise, the probability estimates are not identical, but the differences are not large enough to affect the values of Table 3.3.2(b). If we had fixed λ^* at 100 prior to taking the sample and wished to obtain the ML/E estimates for fixed margins, Table 3.3.2(c) would result. If the log-linear method were applied to this data with both margins fixed, Table 3.3.2(d) would result. Since the log-likelihood ratio is very large

TABLE 3.3.1

Artificial Sample

(a) Observed Frequencies

					Totals
	5	75	53	26	159
	7	8	13	6	34
	0	1	4	2	7
Totals	12	84	70	34	200

(b) ML/E Estimates with Fitted Margins and $\lambda^* = 50$

					Totals
	8.59	77.54	44.58	25.29	156.0
	6.66	8.47	9.71	5.16	30.0
	0.75	3.99	5.71	3.55	14.0
Totals	16.0	90.0	60.0	34.0	200.0

TABLE 3.3.2

Ireland and Kullback Data

(a) Observed Frequencies

					Totals
	783	7426	4709	2145	15063
	517	928	622	703	2770
	207	373	337	425	1342
Totals	1507	8727	5668	3273	19175

(b) Smoothed Frequencies with Fitted Margins

					Totals
	771	7504	4709	2044	15028
	529	974	646	695	2844
	201	371	332	399	1303
Totals	1501	8849	5687	3138	19175

(c) Smoothed Frequencies with $\lambda^* = 100$

					Totals
	785	7423	4708	2147	15063
	515	930	623	702	2770
	207	374	337	424	1342
Totals	1507	8727	5668	3273	19175

(d) Independence Frequencies

					Totals
	1184	6856	4452	2571	15063
	218	1260	819	473	2770
	105	611	397	229	1342
Totals	1507	8727	5668	3273	19175

for this particular model, interaction must be present, and Table 3.3.2(a) would have to be chosen as the set of log-linear estimates, though not in the Bayesian form of the log-linear model of Good (1956). Note that Table 3.3.2(d) is the table of independence frequencies.

3.4 Extensions to Multidimensional Contingency Tables and other Constraints

The IML/ES procedure has been extended to multidimensional tables with varying constraints. The present form of IML/ES takes approximately the same amount of CPU time as GRG. For example, in a 4 x 2 x 3 x 3 contingency table with the main margins fixed, IML/ES required 264 CPU seconds and GRG required 248 CPU seconds on an IBM 370/158 to arrive at the same solution.

Extensions beyond the r x c contingency table with fixed margins do not require different methods for estimation; the only differences occur in description of the problem.

The use of the GRG procedure to obtain ML/E estimates requires the development of a subroutine describing the objective function and the constraints being applied to the system. Because of the generality of the procedure, the constraints can take on many forms other than the standard fixed margins. Examples are:

1. fix the frequency of specific cells such as fixed or structural zeros,
2. fix a specified sum of frequencies not in the same row or column,

3. fix a ratio of two cell frequencies, or
4. adjust a sample multidimensional table to a specified face or a higher dimensional "margin".

Not all the examples given above appear to have applications in any reasonable situation, but this versatility may become important in estimation under unusual circumstances.

The IML/ES procedure mentioned above is restricted to fixing or fitting the ordinary margins or faces of a contingency table, so it would not be able to obtain estimates in examples 2 or 3.

When first analyzing a contingency table, it is wise to test for independence of certain margins and faces. If the structure of the table is such that specific hypotheses of independence can not be rejected, the table can be separated into smaller tables which then simplifies the estimation of the probabilities.

Chapter 5 describes the application of ML/E to a particular multidimensional contingency table using both the GRG method and the IML/ES method with varying constraints on the table.

CHAPTER 4

Properties of ML/E

This chapter presents some of the properties of the ML/E method of estimation. We first consider certain characteristics of the ML/E method under specified conditions. The relationship between ML/E and other Pseudo-Bayes methods is described. The asymptotic and small-sample properties of ML/E and other methods are then calculated and compared. Finally recommendations for the use of ML/E in different situations are given.

4.1 Uniqueness of Estimates

Now that we have a method for calculating λ^* and the set $\{p_i\}$ (using the vector notation of section 1.0), it is appropriate to question the uniqueness of these estimates.

Theorem 4.1.1 Assume a sample $\{n_i\}$ is arranged in a contingency table of two or more dimensions and a value for λ^* is calculated. Then the set of probability estimates $\{p_i\}$ found using the ML/E estimation procedure is unique for constraints which are linear in the probabilities.

Proof:

Let $Q = \sum_i (n_i - \lambda^*\pi_i)\ln(\pi_i)$ be the expression whose extreme values are sought when the variables π_i are restricted by a certain number of side conditions $g_1(\underline{\pi}) = 0, .., g_m(\underline{\pi}) = 0$. Form

$$\phi(\pi_1,..,\pi_n) = Q - \sum_i \mu_i g_i(\underline{\pi}) \text{ where } \{\mu_i\} \text{ is a set of constants.}$$

4.1

Consider the set of equations

$$\frac{\partial \phi(\pi_1,..,\pi_n)}{\partial \pi_i} = 0 \qquad i = 1,..,n \tag{4.1.1}$$

$$g_k(\pi_1,..,\pi_n) = 0 \qquad k = 1,..,m. \tag{4.1.2}$$

Lagrange determined that if the point $(\pi_1,..,\pi_n)$ is a solution to the first n equations 4.1.1, then it will also solve the set 4.1.2. Assume that the constraints are not "only just" consistent in the sense of Good (1963) so that no p_j is forced to equal zero. Since these constraints are defined to be linear in the probabilities,

$$\frac{\partial^2 Q}{\partial \pi_i \partial \pi_j} = \frac{\partial^2 \phi}{\partial \pi_i \partial \pi_j} = -n_i/\pi_i^2 - \lambda^*/\pi_i \qquad i = j$$

$$= 0 \qquad \text{otherwise} \tag{4.1.3}$$

Equations 4.1.3 are strictly negative for fixed $\lambda^* > 0$, $n_i \geq 0$, and $0 < \pi_i < 1$, so the local maximum found by the use of Lagrangian multipliers and Theorem 7.9 in Apostol (1964, p. 152) is also the global maximum in the region of feasible solutions by Bard (1974, p. 49).

Levin and Reeds (1977) give a proof of the uniqueness of k*. A shorter, but not quite complete, proof was given by Good (1975a). Since $\lambda^* = -tk^*/(\sum_i p_i \ln(p_i))$, the value of λ^* is unique.

Therefore the solution set $\{p_i\}$ of probability estimates is unique for a sample and set of constraints. Note that when the constraints are "only just" consistent, we can ignore the estimates which

are forced to zero so that the Hessian matrix does not involve these particular estimates. Once the nonzero estimates are calculated, we can replace the zero estimates in their proper positions.

4.2 Properties Under Special Conditions

The ML/E method has certain properties under special conditions which are considered here. First the relation of ML/E to Maximum Likelihood and Maximum Entropy will be discussed.

Property 1 When $\lambda^* = 0$, the Maximum Likelihood/Entropy estimator is equal to the Maximum Likelihood estimator.

Property 2 When the total sample size is zero, the Maximum Likelihood/Entropy estimator is equal to the Maximum Entropy estimator.

Properties 1 and 2 are obvious from the definitions of the estimation procedures.

Property 3 For finite sample size, the Maximum Likelihood/Entropy estimator approaches the Maximum Entropy estimator as λ^* approaches infinity.

Maximizing $Q = \sum_i (n_i - \lambda^* \pi_i)\ln(\pi_i)$ subject to some constraints is the same as maximizing $Q' = \sum_i (n_i/\lambda^* - \pi_i)\ln(\pi_i)$ subject to the same constraints. As $\lambda^* \to \infty$, $Q' \to -\sum_i \pi_i \ln(\pi_i)$.

Property 4 Define

$$X^2 = (t/n_\cdot) \sum_i (n_i - n_\cdot/t)^2.$$

If $X^2 < t - 1$, the ML/E estimator becomes the Maximum Entropy estimator.

This property follows from Property 3 combined with the fact that $k^* = \infty$ when $X^2 < t - 1$.

Property 5 If the constraints on the system are "only just" consistent in the sense of Good (1963), the probability estimates for cells with zero frequencies may also be zero.

Property 5 is a direct result of the definition of "only just" consistent. The example in Section 1.3 is an example of a system which has constraints that are "only just" consistent.

4.3 Pseudo-Bayes Estimation and ML/E

Philosophically the ML/E method and the Pseudo-Bayes methods of Good, Fienberg, and Holland are very close, the difference being in the form of the prior. Of course, all these methods shift the probability estimates away from the Maximum Likelihood estimates. The question arises as to how these shifts are produced.

The Pseudo-Bayes methods use an estimator of the form

$$p_i^* = (n_i + k)/(n_\cdot + tk) \quad \text{for all } i \tag{4.3.1}$$

in the Multinomial case. This can be rewritten as

$$p_i^* = [n_\cdot/(n_\cdot + tk)]\hat{p}_i + k/(n_\cdot + tk) \quad \text{for all } i \tag{4.3.2}$$

where $\hat{p}_i$ is the ML estimator. Therefore p^* and $\hat{p}$ are related linearly for fixed $n_\cdot$, t, and k.

On the other hand

4.3

$$(p_{ML/E})_i = (n_i - \lambda^*(p_{ML/E})_i \ln[(p_{ML/E})_i])/(n_\cdot - \lambda^* \Sigma_i (p_{ML/E})_i \ln[(p_{ML/E})_i]) \quad \text{for all } i \qquad (4.3.3)$$

$$= (n_\cdot/\phi)\hat{p}_i - \lambda^*(p_{ML/E})_i \ln[(p_{ML/E})_i])/\phi \quad \text{for all } i \qquad (4.3.4)$$

where $\phi = n_\cdot - \lambda^* \Sigma_i (p_{ML/E})_i \ln[(p_{ML/E})_i]$, and $p_{ML/E}$ is not a linear function of $\hat{p}$.

Both types of methods are functions of the roughness of the sample through their respective hyperparameters, k and λ^*. For comparison purposes assume both hyperparameters are fixed. Now p* is a linear function of $\hat{p}$. However $p_{ML/E}$ is still a function of the roughness of the sample through the denominator of equation 4.3.4. We must therefore include another condition on the sample, e.g., all frequencies but the first are equal. With this condition, $n_\cdot = 20$, and $t = 5$, the possible samples are: [(0,1),(5,4)], [(4,5)], [(8,1),(3,4)], [(12,1),(2,4)], [(16,1),(1,4)], and [(20,1),(0,4)], where (x,y) means that there are y cells with frequency x in a sample.

If we fix $\lambda^* = 4$ we have the curve labeled $p_{ML/E}$ in Figure 4.3.1. The line labeled p* corresponds to Pseudo-Bayes estimates when k = 1. In addition lines for $\lambda = k = 0$ and $\lambda = k = \infty$ are included in Figure 4.3.1 to show the effects on the probability estimates when the hyperparameters are at their bounds. Note that $p_{ML/E} = p^* = \hat{p}$ when $\lambda = k = 0$.

4.4 BAN Estimators and Asymptotic Properties

Neyman (1949) defines a class of estimators termed best asymptotically normal (BAN) estimators whose asymptotic properties are the

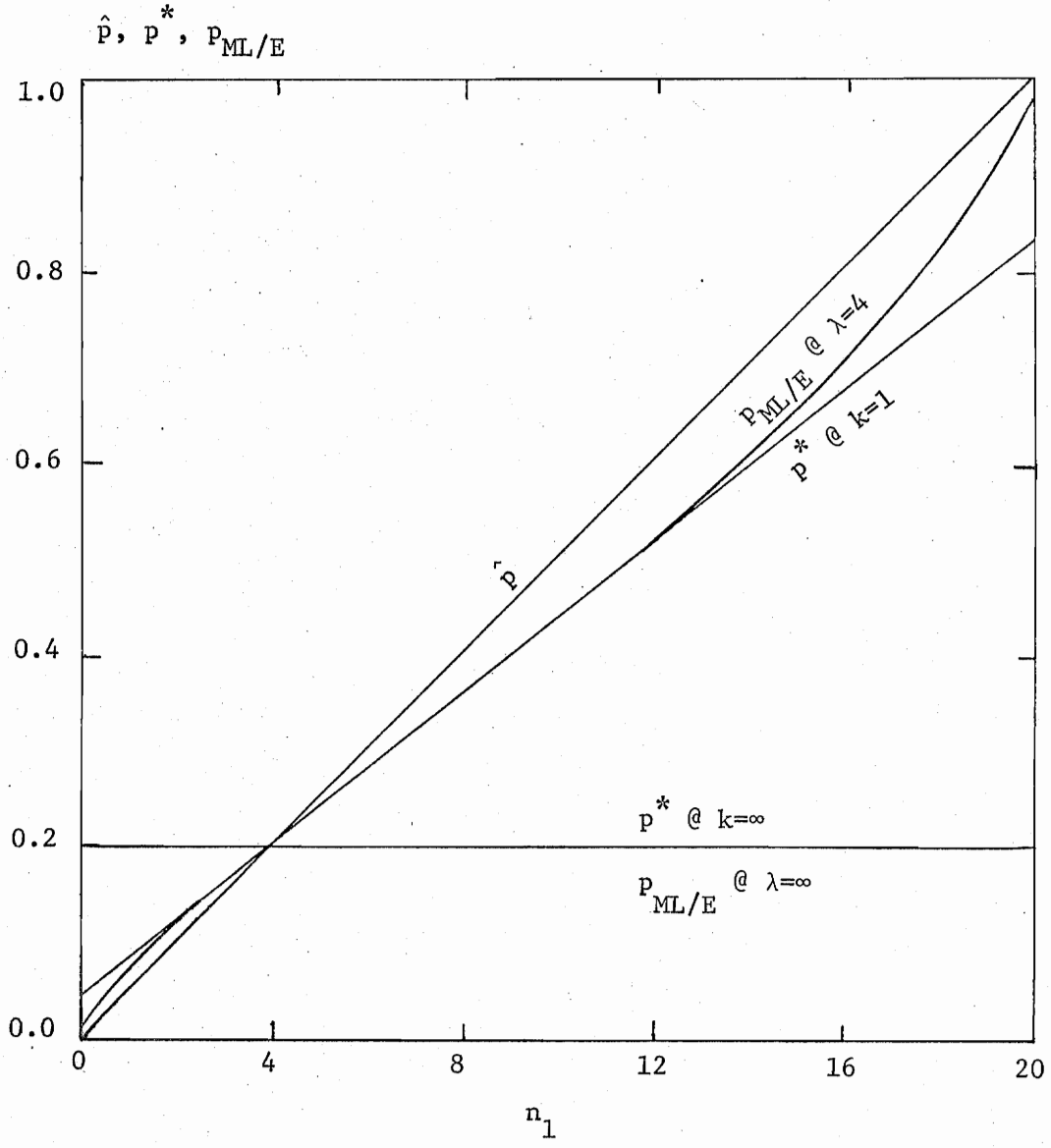

Figure 4.3.1 Comparison of $\hat{p}$, p^*, and $p_{ML/E}$ for $n_\bullet = 20$ and $t = 5$.

same as those of the Maximum Likelihood estimators. That is, if $\hat{\Theta}_i$ is a BAN estimator of a parameter Θ_i, it has the following properties:

1) $\hat{\Theta}_i$ is consistent for Θ_i

2) $\hat{\Theta}_i$ approaches in distribution $N(\Theta_i, \sigma_i/n^{\frac{1}{2}})$ for some constant σ_i and 3) for any other sequence of estimators satisfying properties 1 and 2 with constants σ_i^*, $\sigma_i^* \geq \sigma_i$. BAN estimators are sometimes termed regular BAN (RBAN).

Neyman proved that certain estimation procedures lead to BAN estimators. In particular he showed that minimizing

$$(X')^2 = \sum_i \sum_j \frac{(n_{ij} - n_{..}\pi_{ij})^2}{n_{ij}}$$

for $n_{ij} > 0$ leads to BAN estimators for the population probabilities.

Analysis of Neyman's procedure shows that the proof depends on the form of the $(X')^2$ statistic only through the first and second derivatives of $(X')^2$ with respect to the π_{ij}'s evaluated at $n_{ij} = n_{..}\pi_{ij}$. Therefore if we have an estimation procedure dependent on some statistic S such that

$$1) \quad \left.\frac{\partial S}{\partial \pi_{ij}}\right|_{n_{ij}=n_{..}\pi_{ij}} = \left.\frac{\partial (X')^2}{\partial \pi_{ij}}\right|_{n_{ij}=n_{..}\pi_{ij}} = 0 \qquad (4.4.1)$$

$$\text{and} \quad 2) \quad \left.\frac{\partial^2 S}{\partial^2 \pi_{ij}}\right|_{n_{ij}=n_{..}\pi_{ij}} = \left.\frac{\partial^2 (X')^2}{\partial \pi_{ij}^2}\right|_{n_{ij}=n_{..}\pi_{ij}} = 2n_{..}/\pi_{ij} \qquad (4.4.2)$$

with mixed second partial derivatives equal to zero, then use of Theorem 5 in Neyman (1949) or Lemma 1 in Taylor (1953) shows that the estimators are BAN estimators.

Theorem 4.4.1 The Maximum Likelihood/Entropy Estimation procedure leads to BAN estimators when λ^* is finite.

Proof:

Assume $\lambda^* < \infty$.

Define $S = (-2n_{..}/(n_{..} + \lambda))(\sum_i \sum_j (n_{ij} - \lambda\pi_{ij})\ln(\pi_{ij}) - \underline{u}'\underline{c})$ where $\underline{u}$ is a vector of Lagrangian multipliers and $\underline{c}$ is a vector of constraints linear in the π_{ij}'s. If we minimize S, we can take the first derivative of S with respect to π_{ij} and by setting this equal to zero, we have condition 1 (equation 4.4.1). We also have that

$$\left.\frac{\partial^2 S}{\partial \pi_{ij}^2}\right|_{n_{ij}=n_{..}\pi_{ij}} = (-2n_{..}/(n_{..} + \lambda))(-n_{ij}/\pi_{ij}^2 - \lambda/\pi_{ij})\Big|_{n_{ij}=n_{..}\pi_{ij}}$$

$$= (2n_{..}/(n_{..} + \lambda))(n_{..}\pi_{ij}/\pi_{ij}^2 + \lambda/\pi_{ij}) = 2n_{..}/\pi_{ij} .$$

We therefore have agreement with both the conditions given in equations 4.4.1 and 4.4.2 and have that the estimators obtained using S are BAN estimators. Finally it is clear that minimizing S is equivalent to maximizing $Q = \sum_i \sum_j (n_{ij} - \lambda\pi_{ij})\ln(\pi_{ij})$ with respect to the π_{ij}'s and subject to the same constraints as S. Thus the Maximum Likelihood/Entropy procedure also leads to BAN estimators when λ^* is finite, verifying a conjecture by Good (1969b).

When λ^* is infinite, the first derivative of S no longer exists, so ML/E is not BAN for all λ^*.

4.5 Special Asymptotics

Fienberg and Holland (1973) derive a method for asymptotic comparisons of probability estimates of sparse multinomials which they then use to demonstrate the asymptotic properties of four methods of estimation, namely:

1) Maximum Likelihood
2) Add ½ pseudo count to each cell
3) Minimax
4) Pseudo Bayes.

This comparison can be seen in Figure 4.5.1 and Table 4.5.1. The basis of the comparison is defined by Fienberg and Holland as $D = (N/t) \int (\pi(x) - 1/t)^2 \, dx$ (in their notation) where $\pi(x)$ is a continuous approximation to π, $w_0 = N/(N + t/2)$, and N/t is held constant as t approaches infinity.

They mention that it is difficult to calculate the expected risk in general for their proposed estimator (Pseudo-Bayes) because the estimator is dependent on the sample. The ML/E method has not only this difficulty, but in addition, has the characteristic that there is no closed analytical form of the estimator. Therefore no comparison of this type is possible for the general case. However it is possible to estimate the form of the expected risk.

4.5

Setting D = 0 implies $\pi_{ij} = (rc)^{-1}$ for all i,j. Define $X^2 = \sum_i \sum_j (n_{ij} - n_{..}\pi_{ij})^2/(n_{..}\pi_{ij})$. At D = 0 we have

$$X^2 = (rc/n_{..}) \sum_i \sum_j (n_{ij} - n_{..}/(rc))^2$$

$$= (rc/n_{..}) \sum_i \sum_j (n_{ij})^2 - n_{..}$$

and

$$EX^2 = (rc/n_{..}) \sum_i \sum_j E(n_{ij})^2 - n_{..}$$

$$= (rc/n_{..}) \sum_i \sum_j [(n_{..}/(rc))(1 - 1/(rc)) + n_{..}^2/(rc)^2] - n_{..}$$

$$= (rc/n_{..})(n_{..})(1 - 1/(rc)) + n_{..} - n_{..}$$

$$= rc - 1.$$

Good (1975a) showed that when $X^2 = rc - 1$, k^* must be ∞. Thus we have $\lambda^* = \infty$ by equation 2.5.12. In Section 4.2 we showed that $\lambda^* = \infty$ implies $p_{ij} = (rc)^{-1}$, and therefore the expected risk at D = 0 is also equal to zero. The point at D = 0 in Figure 4.5.1 is exact. The dotted line from that point is an estimate of the expected risk for ML/E based on the fact that it must asymptote to the same function as the Pseudo-Bayes estimator of Fienberg and Holland.

4.6 Small Sample Properties

Sections 4.4 and 4.5 show the asymptotic properties of ML/E estimation, but give no indication of what sample size is necessary for these properties to be meaningful. In this section we show that

TABLE 4.5.1

Leading Term in Expansion of Risk Function
for Four Estimators of $\underline{\pi}$

Estimator	Leading Term
Maximum Likelihood $(\hat{p})$	1
Minimax (p_M)	$1 - 2/N^{\frac{1}{2}}$
Goodman, Jeffreys (p')	$w_0^2 + (1 - w_0)^2 D$
Pseudo-Bayes (p^*)	$(D^2 + 3D + 1)/(D + 2)^2$

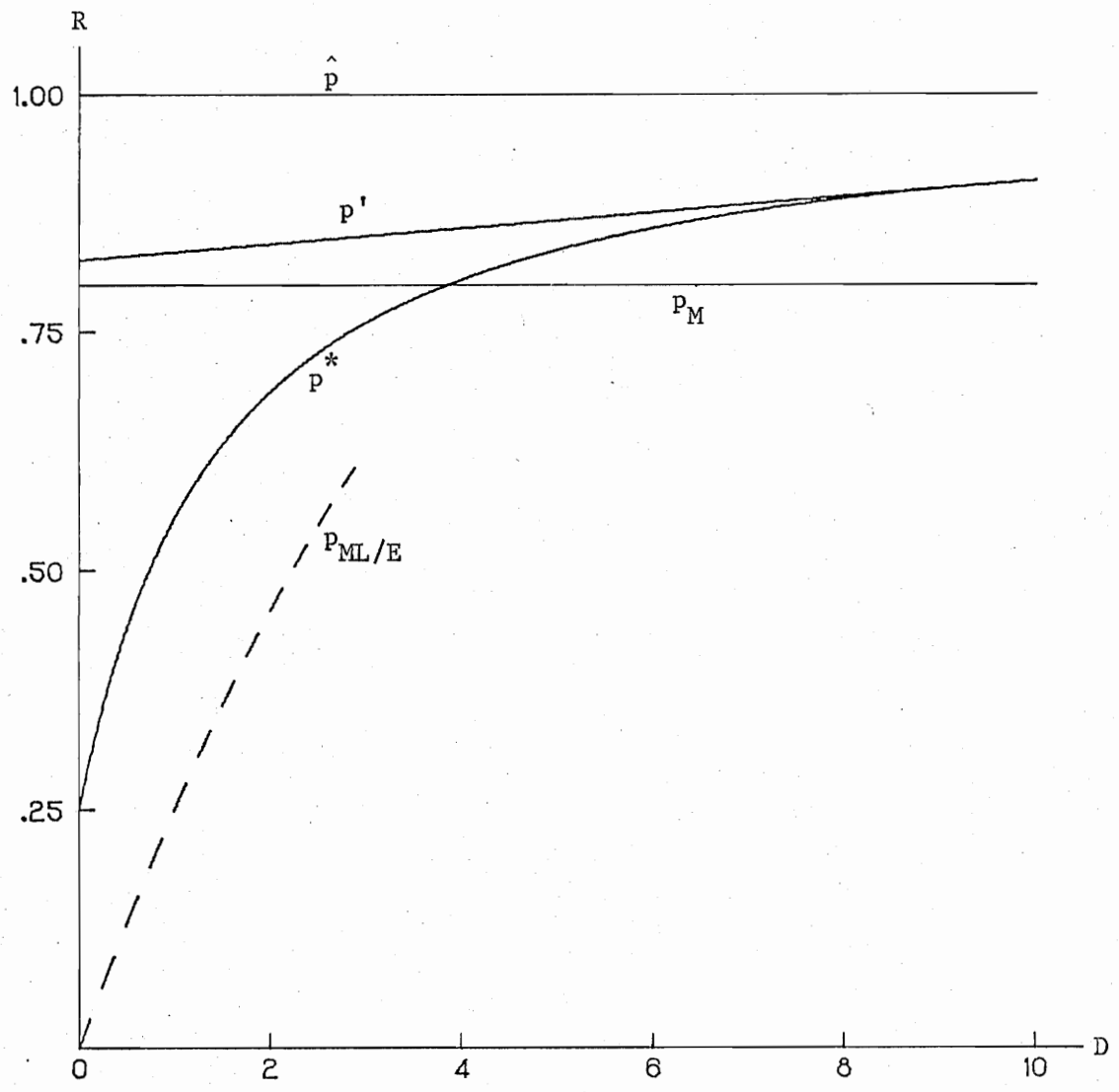

Figure 4.5.1 Leading terms of risk functions for five estimators of π with $t = 20$ and $n_\cdot = 100$ (See Table 4.5.1 for notation).

these properties manifest themselves in binomial and trinomial problems with $n_\cdot = 15$. All comparisons are based on the specific way in which λ is determined.

A criterion must be established for the purpose of comparing different methods. The measure to be used will be the expected risk defined by

$$R = n_\cdot E(||\underline{\pi} - \underline{p}||^2) = n_\cdot E[\sum_i (\pi_i - p_i)^2]$$

where the summation is taken over all cell categories.

Other risk functions could have been considered as well. For instance

$$R = n_\cdot E \sum_i (\pi_i - p_i)^2/\pi_i \quad \text{or} \quad R = n_\cdot E \sum_i \pi_i(1 - p_i)/[p_i(1 - \pi_i)]$$

would give comparisons based on proportional error rather than absolute error. This difference would be important when analyzing a procedure independent of other competing techniques. However the squared error risk function is easier to manipulate and is sufficient for the comparisons to be done here.

We first examine expected risk in the Binomial problem for the following estimators:

1. p_M , Minimax
2. $\hat{p}$, Maximum Likelihood
3. p^* , Pseudo-Bayes

and 4. $p_{ML/E}$, Maximum Likelihood/Entropy.

The expected risk of p_M as given by Fienberg and Holland (1973)

is $(\sqrt{n_\cdot}/(1 + \sqrt{n_\cdot}))^2(1 - t^{-1}) = 0.3158$ when $t = 2$ and $n_\cdot = 15$.

The expected risk for Maximum Likelihood is

$$\begin{aligned} R &= n_\cdot \, E[\textstyle\sum_i (\pi_i - n_i/n_\cdot)^2] \\ &= n_\cdot \textstyle\sum_i E(\pi_i^2 - 2\pi_i n_i/n_\cdot + n_i^2/n_\cdot^2) \\ &= n_\cdot \textstyle\sum_i (\pi_i^2 - 2\pi_i^2 + \pi_i(1 - \pi_i)/n_\cdot + \pi_i^2) \\ &= \textstyle\sum_i (\pi_i - \pi_i^2) \\ &= 1 - \textstyle\sum_i \pi_i^2 \, . \end{aligned}$$

The risks for p^* and $p_{ML/E}$ cannot be derived so easily since the expected values of these estimates do not have closed-form expressions. Instead we use

$$\begin{aligned} R &= n_\cdot \, E[(\pi_1 - p_1)^2 + (\pi_2 - p_2)^2] \\ &= n_\cdot \, E[(\pi_1 - p_1)^2 + (1 - \pi_1 - 1 + p_1)^2] \\ &= 2n_\cdot \, E[(\pi_1 - p_1)^2] \\ &= 2n_\cdot \textstyle\sum_{n_1=0}^{n_\cdot} B(n_\cdot, n_1, \pi_1)(\pi_1 - p_1)^2 \end{aligned}$$

where $B(n_\cdot, n_1, \pi_1)$ is the Binomial probability of cell frequency n_1 occurring in a sample of size $n_\cdot$ with cell probability π_1.

Figure 4.6.1 shows the risk functions (times $n_\cdot = 15$) for the four different estimators. Note that the estimators $p_{ML/E}$ and p^*

have larger risks than $\hat{p}$ towards the boundaries and smaller risks near the center. The risks for both $p_{ML/E}$ and p^* approach 0 as π_1 approaches either 0 or 1, so the behavior of these estimators is satisfactory. However the ratio of the risk of p_M to any of the others near the boundary approaches ∞, and for some purposes this property may be unsatisfactory. p_M is not considered further.

It is worthwhile considering Figure 4.6.1 with respect to the average expected risk, that is,

$$\sum_{j=0}^{m} R(j/m)f(j/m)$$

where R(j/m) is the expected risk at j/m and f(j/m) is the weight one places on each probability with $\sum_{j=0}^{m} f(j/m) = 1$. Of course the probabilities are continuous in the range [0,1], but this discrete function can give an indication of how our weightings affect the average expected risk. For instance if we were to believe that all the probabilities were equally likely, we would place $f(j/m) = (m+1)^{-1}$ for all j. If we felt that the middle probabilities were more likely to occur, we would place higher weights on them and correspondingly lower weights on the boundaries. Table 4.6.1 gives some weights for three cases with m = 10 and the corresponding values of the average expected risk for each estimation procedure.

Note that as the weights are concentrated closer to the center, the risk for ML/E first increases slightly but then decreases considerably. The increase at the beginning is due to the increased weights

placed on those probabilities with the "ears" on the risk function in Figure 4.6.1. As the sample size increases these ears move towards the risk function of Maximum Likelihood.

Figure 4.6.2 gives a two-dimensional representation of the situation for the trinomial case. The curves are contours of constant ratio of the risks associated with $\hat{p}$ and $p_{ML/E}$. When this ratio is less than 1, $p_{ML/E}$ is superior to $\hat{p}$. This figure shows that $p_{ML/E}$ is considerably better than $\hat{p}$ over a large region near the center and boundaries. This property is not surprising given the knowledge that ML/E has a tendency to smooth estimates towards the center. However ML/E does well over a large region of the probability space including a large portion of the boundaries.

Figure 4.6.3 gives the same comparison for $p_{ML/E}$ and p^* with $p_{ML/E}$ superior to p^* for values less than 1. Again we have improvement near the center and a poorer showing further away, though now the differences are not as convincing as in the previous figure. The similarity between the risks of the two estimators $p_{ML/E}$ and p^* in the binomial case makes this relation an expected result.

Figure 4.6.3 shows an important characteristic of $p_{ML/E}$. Note that the portion of the probability space near the edges has risk ratio less than unity. This property is presumably caused by the superiority of ML/E in the binomial situation near the equiprobable case. These edges apparently approximate this condition. It therefore appears reasonable to suggest the use of ML/E when there is more than one non-zero probabilities, but less than the total possible number of

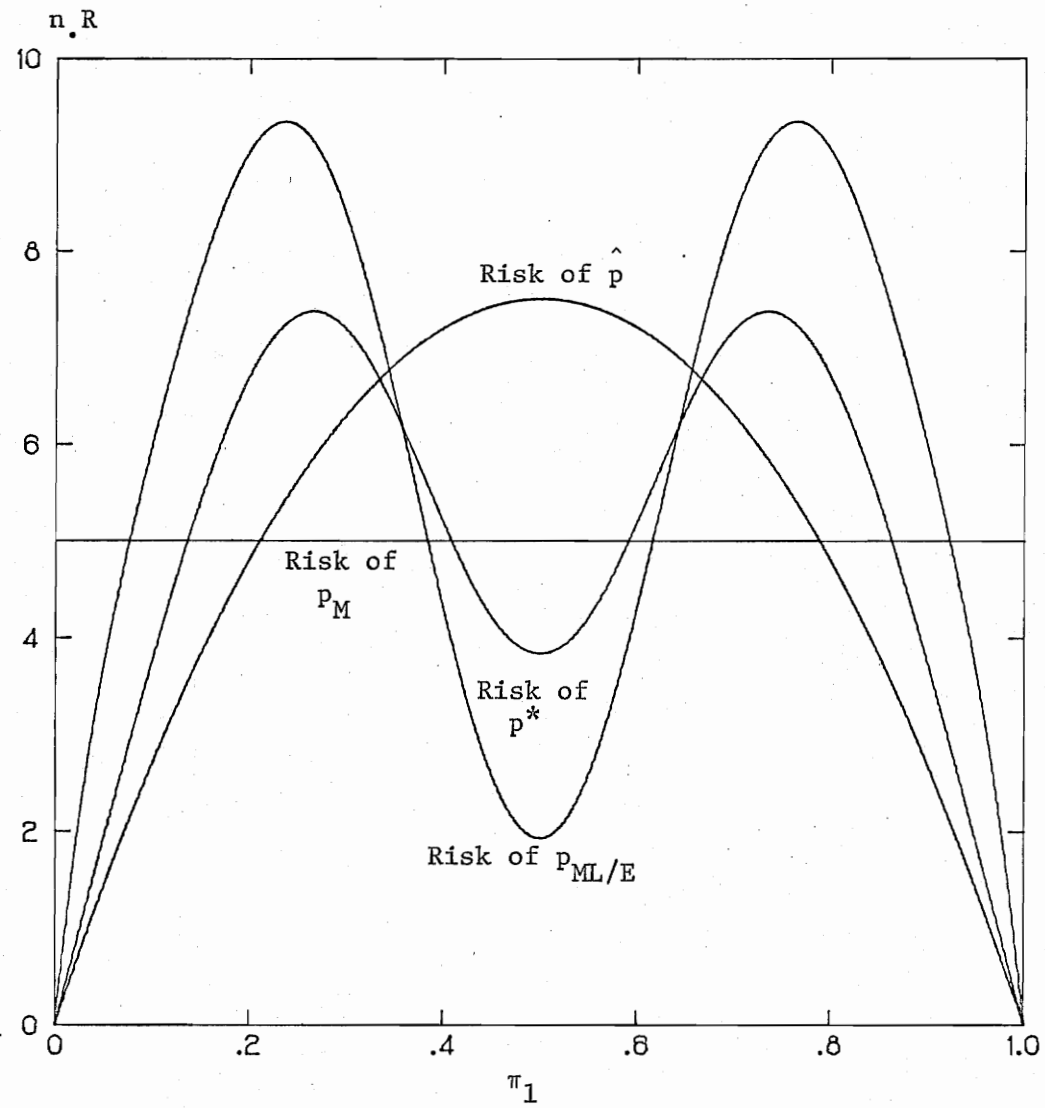

Figure 4.6.1 Comparison of Expected Risks in Binomial Case for $n_. = 15$.

TABLE 4.6.1

Average Expected Risks for Three Weight Functions

Probability	Weights		
	A	B	C
0.0	0.09	0.03	0.01
0.1	0.09	0.05	0.02
0.2	0.09	0.07	0.04
0.3	0.09	0.10	0.10
0.4	0.09	0.15	0.18
0.5	0.09	0.20	0.30
0.6	0.09	0.15	0.18
0.7	0.09	0.10	0.10
0.8	0.09	0.07	0.04
0.9	0.09	0.05	0.02
1.0	0.09	0.03	0.01

	Risk (x 15)		
Estimator	A	B	C
p_M	4.738	4.738	4.738
$\hat{p}$	4.501	5.862	6.594
p^*	4.501	5.076	5.143
$p_{ML/E}$	5.215	5.219	4.769

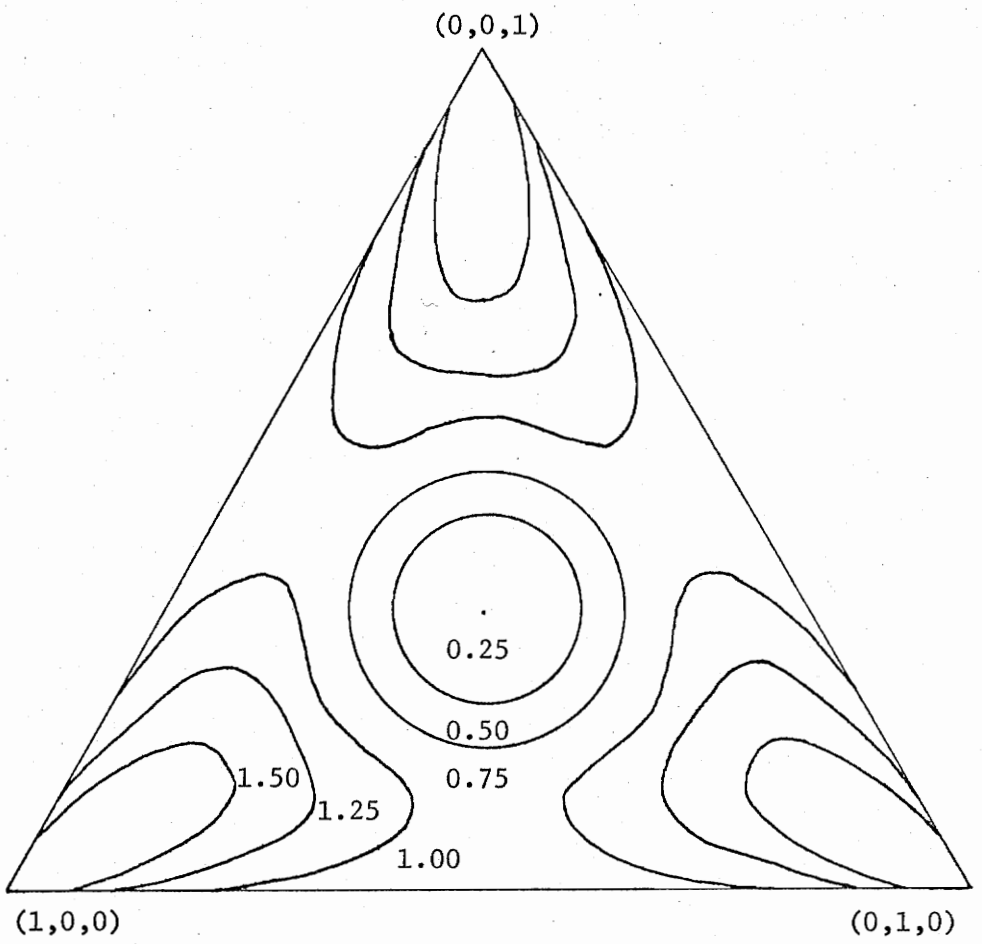

Figure 4.6.2 Contours of constant risk ratio ($p_{ML/E}$ over $\hat{p}$) for $n_{\cdot} = 15$ and $t = 3$.

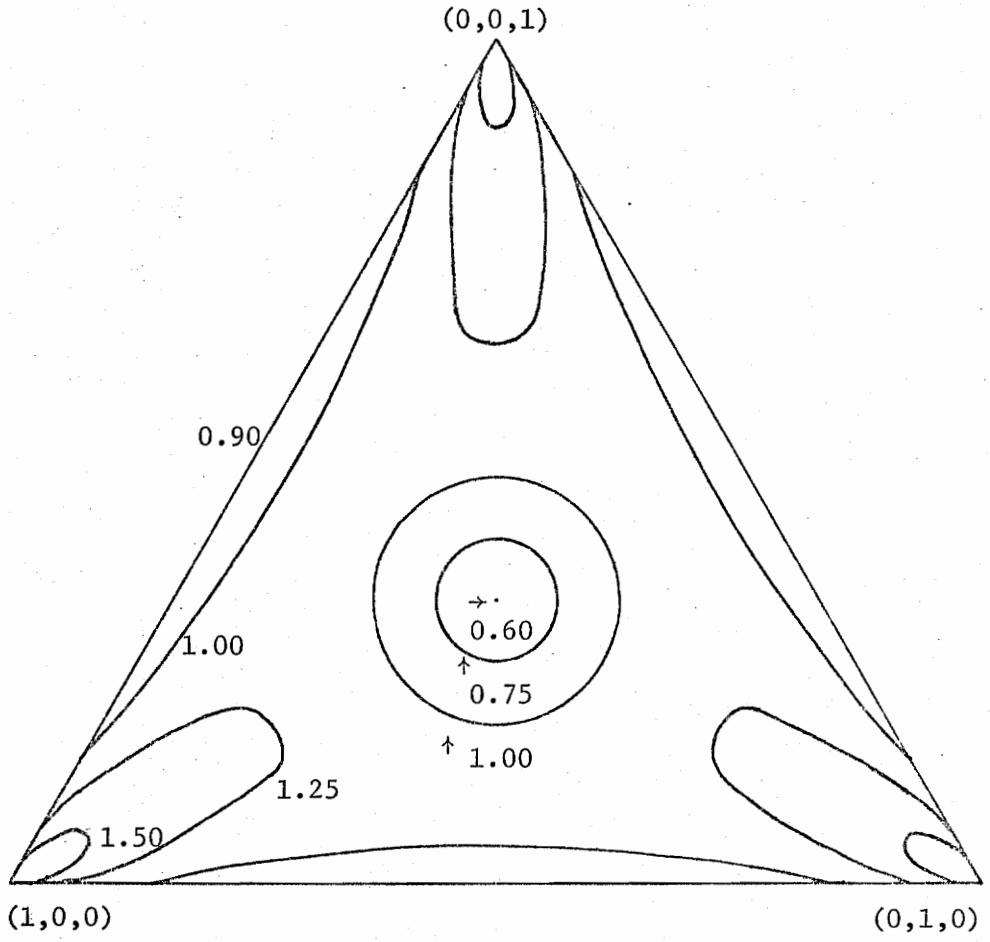

Figure 4.6.3 Contours of constant risk ratio ($p_{ML/E}$ over p^*) for $n_. = 15$ and $t = 3$.

non-zero probabilities, i.e., ML/E is preferable if one is on an edge but not on a vertex. We conjecture that this property can be generalized to more dimensions.

4.7 Recommendations

The Pseudo-Bayes estimators of Good, Fienberg, and Holland; and the ML/E estimator have similar characteristics in that all provide:

1. estimates superior to the Maximum Likelihood estimates when interest is solely in the cell probabilities

and 2. methods for removing zeros in observed contingency tables when these zeros could hamper additional analyses.

For these purposes, the ML/E estimator is preferred because of its slight edge in risk (ref. Section 4.6) and the simplicity of calculation when there are no constraints on the margins (ref. Section 3.1). The Pseudo-Bayes methods are also possible choices; selection being dependent upon which estimation philosophy the user believes to be the most appropriate for his particular situation.

When constraints must be considered, the problem becomes more difficult. It is possible to use the Pseudo-Bayes estimators to remove zeros and then use another method to fit the margins. Unfortunately no distribution theory on this type of approach has been done. The asymptotic distribution of ML/E is known, but the calculations of the ML/E estimates when margin constraints are present quickly become complex. Currently both methods used to calculate ML/E estimates are rather time-consuming. Making the Iterative ML/E Scaling algorithm more efficient in terms of CPU time would help alleviate this problem.

4.7

Significance tests for independence should be performed so that the contingency table may be separated into smaller components when possible.

In addition the set of constraints may be such that zero probability estimates occur, i.e., when the constraints are "only just" consistent. This situation may arise when the sample margins are fixed; it is less common when the margins are fitted to a set of margins from another source. When the constraints are "only just" consistent, one must take care to formulate the estimation procedure so that problems with zero estimates do not occur in the Hessian matrix.

CHAPTER 5

Estimation of Probabilities in a Multidimensional Contingency Table - Food Selection by Beavers

In this chapter we examine a multidimensional contingency table obtained by sampling and compare the estimates of cell frequencies using the methods of log-linear models and ML/E.

5.1 Description of the Table and Log-Linear Estimates

The data in Table 5.1.1 were collected by Jenkins (1975) in a study of food selection by beavers. He examined the interactions among the variables genus, diameter, site, and selection (for gnawing), by calculating the expected frequencies under a variety of models and comparing them to the observed frequencies. Ultimately he wished to find the simplest model which adequately explained the data.

The models were fitted using the log-linear analysis described in Fienberg (1970c) and then were evaluated using the likelihood-ratio statistic as the measure of goodness of fit with the level of significance set at 5%. Let "genus" be variable 1; "choice", variable 2; "diameter", variable 3; and "site", variable 4. A model is specified as follows: [(1), (2), (3), (4)], or [(134), (12), (23)]; where the model [(1), (2), (3), (4)] means that each of the variables is independent of all the others. Similarly, the model [(134), (12), (23)] means that variables 1, 3, and 4 are jointly dependent, variables 1 and 2 are dependent, and variables 2 and 3 are dependent. The model chosen by Jenkins to explain the data was [(123), (14), (34)], that is, choice (2) depends jointly on genus (1) and diameter (3), genus (1)

depends on site (4), and diameter (3) depends on site (4). The expected frequencies for this model are given in Table 5.1.2 and should be compared with the data in Table 5.1.1.

In addition, the estimates for the model [(1), (2), (3), (4)] are given in Table 5.1.3 and will be used for additional comparisons. Note that these estimates are a very poor fit of the data and that this model would be rejected using the likelihood-ratio statistic at a 5% significance level.

5.2 ML/E Estimates

The data in Table 5.1.1 were also examined using the ML/E procedure. Note that there are zero expected frequencies in Table 5.1.2 which indicates that the constraints are "only just" consistent. The ML/E estimates using [(123), (14), (34)] with $\lambda^* = 567$ are given in Table 5.2.1 and should be compared with the log-linear estimates in Table 5.1.2 and the original data in Table 5.1.1.

The estimates in Table 5.2.2 are the values obtained using the model [(1), (2), (3), (4)] and $\lambda^* = 567$. Though we have a much simpler model than that used in Table 5.2.1, we have still retained some of the structure of the table. Table 5.2.2 is a much better fit to the original data than Table 5.1.3, though worse than Table 5.2.1 or Table 5.1.2.

Now assume that we were given the margins in Table 5.2.3 obtained from a much larger sample, so that these margins are more likely to be close to the true population margins than those obtained from the

TABLE 5.1.1

Data from Study of Food Selection by Beavers

Diameter (cm.)	Genus: Birch gnawed down	Birch not gnawed down	Maple gnawed down	Maple not gnawed down	Oak gnawed down	Oak not gnawed down	Pine gnawed down	Pine not gnawed down	Site
2.5-6.2	0	0	10	4	0	0	0	1	
6.2-11.3	11	7	0	9	1	2	0	1	W13
>11.3	11	14	0	12	1	7	0	5	
2.5-6.2	0	1	19	15	1	2	2	2	
6.2-11.3	2	9	1	20	5	8	0	4	West Shore
>11.3	2	2	0	31	4	36	0	14	
2.5-6.2	4	1	10	16	1	5	17	48	
6.2-11.3	30	18	2	9	7	10	1	36	E1 + East Shore + E11
>11.3	2	2	0	1	4	33	0	34	

TABLE 5.1.2

Log-Linear Estimates for Model [(123), (14), (34)]

Diameter (cm.)	Genus: Birch gnawed down	Birch not gnawed down	Maple gnawed down	Maple not gnawed down	Oak gnawed down	Oak not gnawed down	Pine gnawed down	Pine not gnawed down	Site
2.5-6.2	0.81	0.40	6.64	5.95	0.05	0.17	0.27	0.71	
6.2-11.3	11.96	9.45	0.55	6.96	0.46	0.70	0.02	0.90	W13
>11.3	9.26	11.12	0	14.90	1.02	8.60	0	5.10	
2.5-6.2	0.34	0.17	18.80	16.87	0.37	1.31	1.13	3.02	
6.2-11.3	5.72	4.52	1.78	22.55	3.92	6.02	0.11	4.38	West Shore
>11.3	2.39	2.86	0	26.00	4.70	39.68	0	13.36	
2.5-6.2	2.86	1.43	13.56	12.17	1.58	5.52	17.61	47.27	
6.2-11.3	25.32	20.02	0.67	8.49	8.63	13.27	0.87	35.72	E1 + East Shore + E11
>11.3	3.35	4.02	0	3.10	3.28	27.72	0	34.53	

TABLE 5.1.3

Log-linear Estimates for Model [(1), (2), (3), (4)]

Diameter (cm.)	Genus: Birch gnawed down	Birch not gnawed down	Maple gnawed down	Maple not gnawed down	Oak gnawed down	Oak not gnawed down	Pine gnawed down	Pine not gnawed down	Site
2.5-6.2	1.44	4.07	1.97	5.58	1.57	4.46	2.04	5.79	
6.2-11.3	1.75	4.94	2.39	6.77	1.91	5.41	2.48	7.03	W13
>11.3	1.94	5.50	2.66	7.54	2.13	6.03	2.77	7.83	
2.5-6.2	2.70	7.63	3.69	10.46	2.95	8.35	3.83	10.85	
6.2-11.3	3.27	9.26	4.48	12.70	3.58	10.14	4.65	13.18	West Shore
>11.3	3.64	10.32	5.00	14.14	3.99	11.30	5.18	14.68	
2.5-6.2	4.36	12.34	5.97	16.91	4.77	13.51	6.20	17.55	
6.2-11.3	5.29	14.98	7.25	20.53	5.79	16.40	7.52	21.30	E1 + East Shore + E11
>11.3	5.89	16.68	8.08	22.87	6.45	18.26	8.38	23.73	

sample in Table 5.1.1. Using ML/E on Table 5.1.1 with $\lambda^* = 567$ and the margins from Table 5.2.3, we would obtain Table 5.2.4 as the ML/E estimates for the fitted margins.

In addition if we had determined that the zero frequencies in cells (1111), (1211), (3111), and (3211) in Table 5.1.1 were in fact fixed zeros, then the ML/E estimates for Table 5.1.1 using $\lambda^* = 567$ and the margins from Table 5.2.3 would be those estimates given in Table 5.2.5.

TABLE 5.2.1

ML/E Estimates with $\lambda^* = 567$ for Model [(123), (14), (34)]

Diameter (cm.)	Genus: Birch gnawed down	Birch not gnawed down	Maple gnawed down	Maple not gnawed down	Oak gnawed down	Oak not gnawed down	Pine gnawed down	Pine not gnawed down	Site
2.5-6.2	1.37	0.29	9.15	5.66	0.04	0.12	0.18	1.14	
6.2-11.3	11.65	8.61	0.20	7.49	0.74	1.33	0.01	0.97	W13
>11.3	10.09	12.54	0	12.50	1.01	7.77	0	4.70	
2.5-6.2	0.38	0.53	18.49	15.68	0.70	1.76	1.68	2.79	
6.2-11.3	3.65	6.74	1.44	21.16	4.53	7.13	0.04	4.30	West Shore
>11.3	2.23	2.47	0	29.18	4.31	37.57	0	13.20	
2.5-6.2	5.23	1.19	11.29	13.67	1.26	5.11	17.14	47.05	
6.2-11.3	26.30	18.62	1.36	9.35	7.74	11.54	0.95	35.72	E1 + East Shore + E11
>11.3	2.67	2.99	0	2.34	3.68	30.67	0	33.64	

TABLE 5.2.2

ML/E Estimates with $\lambda^* = 567$ for Model [(1), (2), (3), (4)]

Diameter (cm.)	Genus: Birch gnawed down	Birch not gnawed down	Maple gnawed down	Maple not gnawed down	Oak gnawed down	Oak not gnawed down	Pine gnawed down	Pine not gnawed down	Site
2.5-6.2	0.74	1.83	5.63	5.16	0.73	1.82	0.99	3.32	
6.2-11.3	5.75	6.35	1.10	8.20	1.58	3.64	1.14	3.71	W13
>11.3	6.06	9.98	1.28	10.26	1.74	6.82	1.32	6.85	
2.5-6.2	1.32	4.18	10.49	13.26	2.11	4.92	3.27	6.13	
6.2-11.3	2.98	9.65	2.82	16.52	4.55	9.10	2.05	8.28	West Shore
>11.3	3.27	6.13	2.30	22.60	4.40	22.23	2.38	15.06	
2.5-6.2	4.85	6.23	8.71	17.32	2.97	9.13	11.92	31.97	
6.2-11.3	16.00	17.30	4.82	14.65	6.82	13.07	4.18	28.74	E1 + East Shore + E11
>11.3	4.47	8.91	3.70	10.18	5.72	25.65	3.84	29.85	

TABLE 5.2.3

Margins to be Fitted for Table 5.2.4

Margin	Variable	Values
1	Genus	125, 150, 130, 162
2	Choice	150, 417
3	Diameter	160, 195, 212
4	Site	100, 185, 282

TABLE 5.2.4

ML/E Estimates with $\lambda^* = 567$ for Fitted Margins in Table 5.2.3

Diameter (cm.)	Genus: Birch gnawed down	Birch not gnawed down	Maple gnawed down	Maple not gnawed down	Oak gnawed down	Oak not gnawed down	Pine gnawed down	Pine not gnawed down	Site
2.5-6.2	0.93	2.28	5.62	5.07	0.86	2.09	1.07	3.49	
6.2-11.3	6.23	7.03	1.08	8.04	1.74	3.95	1.22	3.87	W13
>11.3	6.48	10.69	1.20	9.86	1.86	7.10	1.36	6.88	
2.5-6.2	1.66	4.96	10.42	13.00	2.34	5.39	3.43	6.39	
6.2-11.3	3.41	10.71	2.75	16.13	4.86	9.68	2.18	8.51	West Shore
>11.3	3.63	6.87	2.13	21.69	4.60	22.81	2.42	15.01	
2.5-6.2	5.16	6.72	8.17	15.95	3.03	9.15	11.69	31.13	
6.2-11.3	16.54	17.95	4.33	13.18	6.87	13.01	3.99	27.74	E1 + East
>11.3	4.64	9.11	3.04	8.34	5.61	25.05	3.45	28.17	Shore + E11

TABLE 5.2.5

ML/E Estimates with $\lambda^* = 567$ for Fitted Margins in Table 5.2.4 and Four Fixed Zeros

Diameter (cm.)	Genus: Birch gnawed down	Birch not gnawed down	Maple gnawed down	Maple not gnawed down	Oak gnawed down	Oak not gnawed down	Pine gnawed down	Pine not gnawed down	Site
2.5-6.2	0	0	6.20	5.74	0	0	1.13	4.28	
6.2-11.3	6.26	7.35	1.26	8.49	1.92	4.31	1.00	4.77	W13
>11.3	6.75	11.10	1.50	10.23	2.15	7.51	1.18	6.87	
2.5-6.2	1.90	5.79	10.61	12.64	2.76	6.36	3.42	8.03	
6.2-11.3	3.27	10.16	2.66	15.12	4.82	9.65	2.00	9.85	West Shore
>11.3	3.47	7.84	1.43	20.14	4.54	21.33	2.07	15.14	
2.5-6.2	5.65	7.73	8.53	15.82	3.29	9.27	11.78	29.07	
6.2-11.3	15.17	18.00	4.41	13.59	6.89	14.03	4.22	25.80	E1 + East Shore + E11
>11.3	4.57	9.99	3.42	8.21	5.63	25.54	4.14	27.25	

BIBLIOGRAPHY

(Part I)

[1] Apostol, T.M., *Mathematical Analysis*, Addison-Wesley, Reading, Mass., 1964.

[2] Arnold, B.C., "Some Examples of Minimum Variance Unbiased Estimates," *Amer. Statist.*, 26(1972), 34-36.

[3] Bard, Y., *Nonlinear Parameter Estimation*, Academic Press, New York, 1974.

[4] Bartlett, M.S., "Contingency Table Interactions," *J. Roy. Statist. Soc., Suppl.*, 2(1935), 248-252.

[5] Berkson, J., "Maximum Likelihood and Minimum χ^2 Estimates of the Logistic Function," *J. Amer. Statist. Assoc.*, 50(1955), 130-162.

[6] Birch, M.W., "Maximum Likelihood in Three-Way Contingency Tables," *J. Roy. Statist. Soc., B*, 25(1963), 220-233.

[7] Bishop, Y.M.M., *Multi-dimensional Contingency Tables: Cell Estimates*, Ph.D. Dissertation, Harvard University, 1967.

[8] Bishop, Y.M.M. and Mosteller, F., "Smoothed Contingency-Table Analysis," Chapter IV-3 in the National Halothane Study, Edited by J.P. Bunker et al, 237-286, National Institutes of Health, U.S. Government Printing Office, Washington, D.C., 1969.

[9] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W., *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, Mass., 1975.

[10] Brown, D.T., "A Note on Approximations to Discrete Probability Distributions," *Information and Control*, 2(1959), 386-392.

[11] Caussinus, H., "Contribution à l'Analyse Statistique des Tableaux de Corrélation," *Ann. Fac. Sci. Univ. Tolouse*, 29(1965), 77-182.

[12] Chew, V., "Point Estimation of the Parameter of the Binomial Distribution," *Amer. Statist.*, 25(1971), 47-50.

[13] Darroch, J.N., "Interactions in Multi-Factor Contingency Tables," *J. Roy. Statist. Soc., B*, 24(1962), 251-263.

[14] Darroch, J.N. and Ratcliff, D., "Generalized Iterative Scaling for Log-Linear Models," *Ann. Math. Statist.*, 43(1972), 1470-1480.

[15] Deming, W.E., *Statistical Adjustment of Data*, Wiley, London, 1943.

[16] Deming, W.E. and Stephan, F.F., "On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known," *Ann. Math. Statist.*, 11(1940), 427-444.

[17] El-Badry, M.A. and Stephan, F.F., "On Adjusting Sample Tabulations To Census Counts," *J. Amer. Statist. Assoc.*, 50(1955), 738-762.

[18] Farebrother, R.W., "Modified Estimators for the Binomial Parameter," *Amer. Statist.*, 31(1977), 98.

[19] Fienberg, S.E., "The Geometry of an r x c Contingency Table," *Ann. Math. Statist.*, 39(1968), 1186-1190.

[20] Fienberg, S.E., "An Iterative Procedure for Estimation in Contingency Tables." *Ann. Math. Statist.*, 41(1970a), 907-917.

[21] Fienberg, S.E., "Quasi-Independence and Maximum Likelihood Estimation in Incomplete Contingency Tables," *J. Amer. Statist. Assoc.*, 65(1970b), 1610-1616.

[22] Fienberg, S.E., "The Analysis of Multidimensional Contingency Tables," *Ecology*, 51(1970c), 419-433.

[23] Fienberg, S.E. and Holland, P.W., "Methods for Eliminating Zero Counts in Contingency Tables," *Random Counts in Scientific Work*, Edited by G.P. Patil, 233-260, Penn. State Univ. Press, University Park, Penn., 1970.

[24] Fienberg, S.E. and Holland, P.W., "On the Choice of Flattening Constants for Estimating Multinomial Probabilities," *J. Multivariate Anal.*, 2(1972), 127-134.

[25] Fienberg, S.E. and Holland, P.W., "Simultaneous Estimation of Multinomial Cell Probabilities," *J. Amer. Statist. Assoc.*, 68(1973), 683-691.

[26] Fisher, R.A., "On the Mathematical Foundations of Theoretical Statistics," *Phil. Trans. Roy. Soc. London A*, 222(1922), 309-368.

[27] Fisher, R.A., *Statistical Methods for Research Workers, Fifth Edition*, Oliver and Boyd, London, 1934.

[28] Gart, J.J., "Approximate Confidence Limits for Relative Risks," *J. Roy. Statist. Soc., B,* 24(1962), 454-463.

[29] Gerald, C.F., *Applied Numerical Analysis,* Addison-Wesley, Menlo Park, Calif., 1970.

[30] Gokhale, D.V., "An Iterative Procedure for Analysing Log-Linear Models," *Biometrics,* 27(1971), 681-687.

[31] Good, I.J., "Rational Decisions," *J. Roy. Statist. Soc., B,* 14(1952), 107-114.

[32] Good, I.J., "On the Estimation of Small Frequencies in Contingency Tables," *J. Roy. Statist. Soc., B,* 18(1956), 113-124.

[33] Good, I.J., "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables," *Ann. Math. Statist.* 34(1963), 911-934.

[34] Good, I.J., *The Estimation of Probabilities: An Essay on Modern Bayesian Methods,* MIT Press, Cambridge, Mass., 1965.

[35] Good, I.J., "A Bayesian Significance Test for Multinomial Distributions," *J. Roy. Statist. Soc., B,* 29(1967), 399-431.

[36] Good, I.J., "Statistics of Language," in Encyclopedia of Linguistics, *Information and Control,* 567-581, Pergamon Press, 1969a.

[37] Good, I.J., "Review of C.T. Ireland and S. Kullback, Contingency Tables with Given Marginals,' *Biometrika,* 55(1968), 179-188" in *Mathematical Reviews,* 37(1969b), Rev #4903, 906-907.

[38] Good, I.J., "Scientific Induction and Exponential-Entropy Distributions," *Amer. Statist.,* 26(1972), 45.

[39] Good, I.J., "The Bayes Factor Against Equiprobability of a Multinomial Population Assuming a Symmetric Dirichlet Prior," *Ann. Statist.,* 3(1975a), 246-250.

[40] Good, I.J., Review of Arnold Zellner "An Introduction to Bayesian Inference in Econometrics" in *Technometrics* 17(1975b), 137-138.

[41] Good, I.J., "Amendment to the Review of Arnold Zellner 'An Introduction to Bayesian Inference in Econometrics' in *Technometrics* 17(1975), 137-138; in *Technometrics,* 18(1976a), 123.

[42] Good, I.J., "On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables," *Ann. Statist.,* 4(1976b), 1159-1189.

[43] Good, I.J. and Crook, J.F., "The Bayes/Non-Bayes Compromise and the Multinomial Distribution," *J. Amer. Statist. Assoc.*, 69(1974), 711-720.

[44] Good, I.J. and Gaskins, R.A., "Centroid Method of Integration," *Nature*, 222(1969), 697-698.

[45] Good, I.J. and Gaskins, R.A., "The Centroid Method of Numerical Integration," *Numer. Math.* 16(1971), 343-359.

[46] Good, I.J. and Gaskins, R.A., "Global Nonparametric Estimation of Probability Densities," *Virginia J. of Science*, 23(1972), 171-193.

[47] Goodman, L.A., "Some Alternatives to Ecological Correlation," *Amer. J. of Sociology*, 64(1959), 610-625.

[48] Goodman, L.A., "On Plackett's Test for Contingency Table Interactions," *J. Roy. Statist. Soc., B*, 25(1963), 179-188.

[49] Goodman, L.A., "Interactions in Multidimensional Contingency Tables," *Ann. Math. Statist.*, 35(1964), 632-646.

[50] Goodman, L.A., "The Analysis of Cross-Classified Data: Independence, Quasi-Independence, and Interactions in Contingency Tables With or Without Missing Entries," *J. Amer. Statist. Assoc.*, 63(1968), 1091-1131.

[51] Gower, J.C., "Simulating Multidimensional Arrays in One Dimension," *Applied Statistics*, 17(1968), 180-185.

[52] Gradshteyn, I.S. and Ryzhik, I.W., *Table of Integrals, Series, and Products*, Academic Press, New York, 1965.

[53] Haberman, S.J., "Log-Linear Fit for Contingency Tables," *Applied Statistics*, 21(1972), 218-225.

[54] Haberman, S.J., *The Analysis of Frequency Data*, Univ. of Chicago Press, Chicago, 1974.

[55] Haldane, J.B.S., "A Problem in the Significance of Small Numbers," *Biometrika*, 42(1955), 266-267.

[56] Huber, C. and Lellouch, J., "Estimation dans les Tableaux de Contingence a un Grand Nombre d'Entrées," *Int. Stat. Rev.* 42(1974), 193-203.

[57] Imrey, P.B. and Koch, G.G., "Linear Models Analysis of Incomplete Multivariate Categorical Data," Institute of Statistics Mimeo Series No. 820, Dept. of Biostatistics, Univ. of North Carolina at Chapel Hill, 1972.

[58] Ireland, C.T. and Kullback, S., "Contingency Tables with Given Marginals," *Biometrika*, 55(1968), 179-188.

[59] Jeffreys, H., *Theory of Probability, 3rd Ed.*, Clarendon Press, Oxford, 1961.

[60] Jenkins, S.H., "Food Selection by Beavers, A Multidimensional Contingency Table Analysis," *Oecologia* (Berl.), 21(1975), 157-173.

[61] Johnson, W.E., Appendix (Ed. by R.B. Braithwaite) to "Probability: Deductive and Inductive Problems," *Mind*, 41(1932), 421-423.

[62] Kendall, M.G. and Stuart, A., *The Advanced Theory of Statistics, Vol. II, 3rd Ed.*, Hafner, New York, 1973.

[63] Koch, G.G., Imrey, P.B., and Reinfurt, D.W., "Linear Model Analysis of Categorical Data with Incomplete Response Vectors," *Biometrics*, 28(1972), 663-692.

[64] Lancaster, H.O., *The Chi-Squared Distribution*, Wiley, New York, 1969.

[65] Laplace, P.S., "Mémoire sur la Probabilité des Causes par les Événements," *Mém. de. l'Acad. R. de. Sci. Paris*, 6(1774), 621-656.

[66] Lasdon, L.S., Fox, R.L., and Ratner, M.W., "Nonlinear Optimization Using the Generalized Reduced Gradient Method," Tech. Mem. No. 325, School of Engrg., Case Western Reserve Univ., Cleveland, 1973.

[67] Lasdon, L.S., Waren, A.D., Jain, A., and Ratner, M.W., "Design and Testing of a Generalized Reduced Gradient Code for Nonlinear Optimization," Tech. Mem. No. 353, School of Management, Case Western Reserve Univ., 1975a.

[68] Lasdon, L.S., Waren, A.D., Ratner, M.W., and Jain, A., "GRG System Documentation," Tech. Mem. CIS-75-01, Comp. and Infor. Sci. Depart., Cleveland S. Univ., Cleveland, 1975b.

[69] Ladson, L.S., Waren, A.D., Ratner, M.W., and Jain, A., "GRG User's Guide," Tech. Mem. CIS-75-02, Comp. and Infor. Sci. Depart., Cleveland S. Univ., Cleveland, 1975c.

[70] Leonard, T., "Bayesian Methods for Binomial Data," *Biometrika*, 59(1972), 581-589.

[71] Leonard, T., "Bayesian Estimation Methods for Two-Way Contingency Tables," Univ. of Warwick, (1973), 1-37.

[72] Levin, B. and Reeds, J., "Compound Multinomial Likelihood Functions are Unimodal: Proof of a Conjecture of I.J. Good," *Ann. Statist.* 5(1977), 79-87.

[73] Lewis, P.M., II, "Approximating Probability Distributions to Reduce Storage Requirements," *Information and Control*, 2(1959), 214-225.

[74] Lidstone, G.J., "Note on the General Case of the Bayes-Laplace Formula for Inductive or a Posteriori Probabilities," *Trans. Fac. Actuar.*, 8(1920), 182-192.

[75] Lindley, D.V., "The Bayesian Analysis of Contingency Tables," *Ann. Math. Statist.*, 35(1964), 1622-1643.

[76] Makeham, W.M., "On the Theory of Inverse Probabilities," *J. Inst. Actuar.*, 29(1891), 242-251.

[77] Neyman, J., "Contribution to the Theory of the χ^2 Test," *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability*, Edited by J. Neyman, 239-273, Univ. of California Press, Berkeley, 1949.

[78] Neyman, J. and Pearson, E.S., "Further Notes on the χ^2 Distribution," *Biometrika*, 22, (1930-1931), 298-305.

[79] Neyman, J. and Pearson, E.S., "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Phil. Trans. Roy. Soc. London A*, 231(1933), 289-337.

[80] Pearson, K., "On a Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that It Can be Reasonably Supposed to Have Arisen from Random Sampling," *Phil. Mag.*, (5), 50(1900), 157-172.

[81] Perks, W., "Some Observations on Inverse Probability Including a New Indifference Rule," *J. Inst. Actuar.*, 73(1947), 285-312.

[82] Perks, W., "Contribution to the Discussion of G.A. Barnard 'The Bayesian Controversy in Statistical Inference'," *J. Inst. Actuar.*, 93(1967), 264-268.

[83] Plackett, R.L., "A Note on Interactions in Contingency Tables," *J. Roy. Statist. Soc., B*, 24(1962), 162-166.

[84] Quade, D. and Salama, I.A., "A Note on Minimum Chi-Square Statistics in Contingency Tables," *Biometrics*, 31(1975), 953-956.

[85] Rudolph, G.J., "A Quasi-Multinomial Type of Contingency Table," *S. Afr. Statist. J.*, 1(1967), 59-65.

[86] Savage, I.R., "Incomplete Contingency Tables: Condition for the Existence of Unique MLE," in *Mathematics and Statistics Essays in Honor of Harold Bergström*, edited by P. Jogars and L. Råde, 87-99. Göteborg, Sweden, Chalmers Institute of Technology, 1973.

[87] Schafer, R.E., "Modified Estimators for the Binomial Parameter," *Amer. Statist.*, 30(1976), 98-100.

[88] Smith, J.H., "Estimation of Linear Functions of Cell Proportions," *Ann. Math. Statist.*, 18(1947), 231-254.

[89] Stabler, E.L., "On Mr. Makeham's Theory of Inverse Probabilities," *J. Inst. Actuar.*, 30(1892), 239-244.

[90] Steinhaus, H., "The Problem of Estimation," *Ann. Math. Statist.*, 28(1957), 633-648.

[91] Stephan, F.F., "An Iterative Method of Adjusting Sample Frequency Tables When Expected Marginal Totals are Known," *Ann. Math. Statist.*, 13(1942), 166-178.

[92] Stevens, W.L., "The Distribution of Entries in a Contingency Table with Fixed Marginal Totals," *Ann. of Eugenics*, 8(1938), 238-244.

[93] Sutherland, M., Holland, P.W., and Fienberg, S.E., "Combining Bayes and Frequency Approaches to Estimate a Multinomial Parameter," in *Studies in Bayesian Econometrics and Statistics*, edited by S.E. Fienberg and A. Zellner, 585-617, Amsterdam, North Holland, 1974.

[94] Taylor, W.F., "Distance Functions and Regular Best Asymptotically Normal Estimates," *Ann. Math. Statist.*, 24(1953), 85-92.

[95] Trybula, S., "Some Problems of Simultaneous Minimax Estimation," *Ann. Math. Statist.*, 29(1958), 245-253.

[96] Wilks, S.S., "The Likelihood Test of Independence in Contingency Tables," *Ann. Math. Statist.*, 6(1935), 190-196.

[97] Wilks, S.S., "The Large Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses," *Ann. Math. Statist.*, 9(1938), 60-62.

[98] Yates, F., "Contingency Tables Involving Small Numbers and the χ^2 Test," *J. Roy. Statist. Soc., Suppl.*, 1(1934), 217-235.

[99] Young, K.H. and Young, L.Y., "Estimation of Regressions Involving Logarithmic Transformation of Zero Values in the Dependent Variable," *Amer. Statist.*, 29(1975), 118-120.

[100] Young, K.H. and Young, L.Y., "Empty Cells in Contingency Tables," *Amer. Statist.*, 30(1976), 101.

[101] Zellner, A., *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York, 1971.

PART II

Methods for Calculation of the Kolmogorov-Smirnov One-Sample Statistic

CHAPTER 1

Introduction to the Kolmogorov-Smirnov One-Sample Statistic

Lex X be a random variable with continuous probability distribution function

$$F_X(x) = \text{Prob}(X \leq x).$$

Let a random sample of N observations be taken from this distribution, arranged as order statistics $X_1, X_2, \ldots, X_N$, where $X_1 \leq X_2 \leq \ldots \leq X_N$. Let

$$S_N(x) = \begin{cases} 0 & \text{for } x < X_1 \\ i/N & \text{for } X_i \leq x < X_{i+1} \qquad (i = 1, 2, \ldots, N-1) \\ 1 & \text{for } X_N \leq x, \end{cases}$$

the "empirical distribution function".

Kolmogorov (1933) introduced the statistic

$$D_N = \sup_x |S_N(x) - F_X(x)|.$$

He obtained recursion formulae for

$$F_N(z) = \text{Prob}(D_N < zN^{-\frac{1}{2}}), \tag{1.1}$$

and showed that this probability is asymptotically equal to $K_0(z)$

defined by

$$K_0(z) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2z^2}. \qquad (1.2)$$

For a proof of this limiting distribution, a convenient reference is Durbin (1973).

Several calculations of (1.1) have been made using various methods. Smirnov (1948) tabulated the limiting distribution (1.2). Birnbaum (1952)[1] used Kolmogorov's recursion formulae to tabulate (1.1) for $N = 1(1)100$ and $c = 1(1)15$ where $c = zN^{\frac{1}{2}}$. Li-Chien (1956) derived an approximation for (1.1) which Korolyuk (1960)[2] wrote in the form:

$$F_N(z) \sim \sum_{r=0}^{\infty} N^{-\frac{1}{2}r} K_r(z), \qquad (1.3)$$

where

$$K_1(z) = (-4z/3) \sum_{k=1}^{\infty} (-1)^k k^2 e^{-2k^2z^2}, \qquad (1.4)$$

$$K_2(z) = (-1/9) \sum_{k=1}^{\infty} (-1)^k [f_1 - 4(f_1 + 3)k^2z^2 + 8k^4z^4] e^{-2k^2z^2}, \qquad (1.5)$$

and

$$K_3(z) = (2z/27) \sum_{k=1}^{\infty} (-1)^k k^2 [f_2/5 - 4(f_2 + 45)k^2z^2/15 + 8k^4z^4] e^{-2k^2z^2}, \qquad (1.6)$$

---

[1]The expression for $K_0(z)$ in Birnbaum (1952) contains a misprint: z/N instead of $z/N^{\frac{1}{2}}$.

[2]The expression for $K_3(z)$ in Korolyuk (1960) should be multiplied by k^2.

where

$$f_1 = k^2 - \tfrac{1}{2}[1 - (-1)^k],\ f_2 = 5k^2 + 22 - 15[1 - (-1)^k]/2.$$

Table 1.1 shows the contributions of $K_i(z)$ $(i = 0, 1, 2, 3)$ for various values of z. Terms beyond K_3 were not necessary since the coefficient $N^{-\frac{1}{2}r}$ forces rapid convergence to zero for these terms.

Kolmogorov (1933) mentioned that if z is very small (1.2) converges slowly and can be replaced by the asymptotic formula

$$(2\pi)^{\frac{1}{2}}z^{-1}\exp(-\pi^2 z^{-2}/8),$$

and Feller (1948), using Jacobi's formula ((3.3) below), gave the more complete formula

$$K_0(z) = (2\pi)^{\frac{1}{2}}z^{-1}\sum_{k=1}^{\infty}\exp[-(2k-1)^2\pi^2 z^{-2}/8]. \qquad (1.7)$$

Durbin (1973) transformed the recursion formulae of Kolmogorov into a method for calculating (1.1) exactly using matrix multiplication.

Here we transform the series for $K_1(z)$, $K_2(z)$, and $K_3(z)$ into series analogous to (1.7) so that the results are useful when z is small, that is, when the evaluations of left-hand tail-area probabilities are required. We also make comparisons of the various formulae in terms of speed of calculation and show the usable ranges of the approximation formulae.

Left-hand tail-area probabilities for D_N are required for testing whether a fit is "too good to be true". Such a test is useful in the following situation.

TABLE 1.1

Contributions of K_0, K_1, K_2, and K_3
for Various Values of z in (1.3)

z	K_0	K_1	K_2	K_3
0.5	.03605	.10660	.01561	-.09091
1.0	.73000	.17866	-.06089	.05245
1.5	.97778	.02222	.01666	-.01074
2.0	.99933	.00089	.00298	.00346
2.5	.99999	.00001	.00010	-.00015

Suppose that we are estimating a probability density function by the method of "penalized likelihood", that is, by the maximization of an expression of the form $L - \beta\Phi$, where L denotes the log-likelihood pertaining to a putative density function f, and Φ is some functional of f called the roughness of f. This method of density estimation is nonparametric in the usual sense that it does not constrain the density function to belong to any specific family, so that β is best described as a "hyperparameter", especially as it can be regarded as a parameter in a prior when the method is interpreted in a Bayesian manner, that is, when the interpretation is that we are to maximize the posterior density of f in function space, where the prior is proportional to $\exp(-\beta\Phi)$. If we took $\beta = 0$, the method would imply that the integrated form of f was the empirical distribution function, f would be unreasonably rough, and the Kolmogorov-Smirnov statistic would vanish. Thus we can detect whether the hyperparameter β is too small by reference to the left or lower tail of the Kolmogorov-Smirnov distribution, and whether it is too large by the right or upper tail of this statistic, or of some other statistic, such as X^2 ("chi-squared"). This technique for selecting the hyperparameter was used in relation to some real (histogram) data from high-energy scattering experiments, with $N = 25752$ (Good (1971), Good and Gaskins (1971, 1972, 1974)). For example, when $z = 0.2804$ we have a lower tail probability $P = 0.000001535$. It may not be possible to interpret this as a true probability in the application, because the hypothesis for f has been chosen in terms of the histogram, that is, it has been fitted to the data. It may

be better to describe P as a "measure of overagreement", that is, as a quantitative indication that the corresponding value of β was too small. There might be a better quantitative indication but we do not know of one. (A value of β was then found, which was suggested for other reasons, that gave a lower tail-area probability of 0.10 to the Kolmogorov-Smirnov statistic and an upper tail-area probability to X^2 of 0.215. This value of β is acceptable.)

Another use of the lower tail of the Kolmogorov statistic occurs when a test due to Ajne is used for testing whether a number of points are distributed randomly on a circle, a question that has application to geophysics, to bubble-chamber experiments, and to the migration of birds. Durbin (1973, p. 39) points out that the upper tail of Ajne's distribution can be expressed in terms of the lower tail of the Kolmogorov-Smirnov distribution.

CHAPTER 2

Expressions Involving Differentiation

Equations (1.2), (1.4), (1.5) and (1.6) can be expressed in terms containing derivatives with respect to z of (1.2) and of

$$K_0^*(z) = 1 + 2\sum_{k=1}^{\infty} e^{-2k^2z^2} . \tag{2.1}$$

We write

$$\Delta_0(z) = K_0(z),\ \Delta_0^*(z) = K_0^*(z) \tag{2.2}$$

and

$$\Delta_n(z) = \frac{d^n}{dz^n} K_0(z),\ \Delta_n^* = \frac{d^n}{dz^n} K_0^*(z) \quad (n = 1, 2, \ldots) . \tag{2.3}$$

Then

$$\Delta_1(z) = -\sum_{k=1}^{\infty} (-1)^k\ 8k^2ze^{-2k^2z^2} , \tag{2.4}$$

$$\Delta_1^*(z) = -\sum_{k=1}^{\infty} 8k^2ze^{-2k^2z^2} , \tag{2.5}$$

$$\Delta_2(z) = \sum_{k=1}^{\infty} (-1)^k[32k^4z^2 - 8k^2]\ e^{-2k^2z^2}$$

$$= \sum_{k=1}^{\infty} (-1)^k\ 32k^4z^2e^{-2k^2z^2} + z^{-1}\Delta_1(z) , \tag{2.6}$$

$$\Delta_2^*(z) = \sum_{k=1}^{\infty} [32k^4z^2 - 8k^2]\, e^{-2k^2z^2}$$

$$= \sum_{k=1}^{\infty} 32k^4z^2e^{-2k^2z^2} + z^{-1}\Delta_1^*(z) \ , \qquad (2.7)$$

and

$$\Delta_3(z) = \sum_{k=1}^{\infty} (-1)^k[-128k^6z^3 + 32k^4z + 64k^4z]\, e^{-2k^2z^2}$$

$$= \sum_{k=1}^{\infty} (-1)^k[-128k^6z^3 + 96k^4z]\, e^{-2k^2z^2}$$

$$= -\sum_{k=1}^{\infty} (-1)^k\, 128k^6z^3e^{-2k^2z^2} + 3z^{-1}\Delta_2(z) - 3z^{-2}\Delta_1(z) \ . \qquad (2.8)$$

From equation (1.4), we have

$$K_1(z) = (-4z/3) \sum_{k=1}^{\infty} (-1)^k\, k^2e^{-2k^2z^2}$$

$$= \Delta_1(z)/6 \ . \qquad (2.9)$$

From equation (1.5), we have

$$K_2(z) = (-1/9) \sum_{k=1}^{\infty} (-1)^k[f_1 - 4(f_1 + 3)k^2z^2 + 8k^4z^4]\, e^{-2k^2z^2}$$

$$= (-1/9) \sum_{k=1}^{\infty} (-1)^k[k^2 - \tfrac{1}{2}\{1 - (-1)^k\}$$

$$- 4(k^2 - \tfrac{1}{2}\{1 - (-1)^k\} + 3)k^2z^2 + 8k^4z^4]\, e^{-2k^2z^2}$$

$$= (-1/9) \sum_{k=1}^{\infty} (-1)^k[-\tfrac{1}{2} + k^2 - 10k^2z^2 + 8k^4z^4 - 4k^4z^2]\, e^{-2k^2z^2}$$

$$- (1/9) \sum_{k=1}^{\infty} [\tfrac{1}{2} - 2k^2z^2]\, e^{-2k^2z^2}$$

$$= (-1/9)[-\Delta_0(z)/4 - z^{-1}\Delta_1(z)/8 + 10z\Delta_1(z)/8$$

$$+ \{z^2(\Delta_2(z) - z^{-1}\Delta_1(z))\}/4 + (-\Delta_2(z) + z^{-1}\Delta_1(z))/8$$

$$+ \Delta_0^*(z)/4 + z\Delta_1^*(z)/4]$$

$$= [2\Delta_0(z) - 8z\Delta_1(z) - (2z^2 - 1)\Delta_2(z) - 2\Delta_0^*(z) - 2z\Delta_1^*(z)]/72 \,. \tag{2.10}$$

From equation (1.6), we have

$$K_3(z) = (2z/27) \sum_{k=1}^{\infty} (-1)^k\, k^2[f_2/5 - 4(f_2 + 45)k^2z^2/15 + 8k^4z^4]\, e^{-2k^2z^2}$$

$$= (2z/27) \sum_{k=1}^{\infty} (-1)^k\, k^2\{k^2 + 22/5 - 3[1 - (-1)^k]/2$$

$$- 4(5k^2 + 22 - 15[1 - (-1)^k]/2 + 45)k^2z^2/15 + 8k^4z^4\}\, e^{-2k^2z^2}$$

$$= (2/27)\, [-29\Delta_1(z)/80 + (32z)^{-1}(\Delta_2(z) - z^{-1}\Delta_1(z))$$

$$- 238(z\Delta_2(z) - \Delta_1(z))/480 - 8z^2(\Delta_3(z) - 3z^{-1}\Delta_2(z) + 3z^{-2}\Delta_1(z))$$

$$/128 + (\Delta_3(z) - 3z^{-1}\Delta_2(z) + 3z^{-2}\Delta_1(z))/96]$$

$$= [-26\Delta_1(z) - 148z\Delta_2(z) - (30z^2 - 5)\Delta_3(z)$$

$$- 60\Delta_1^*(z) - 30z\Delta_2^*(z)]/6480. \qquad (2.11)$$

CHAPTER 3

Relationship to Theta Functions

Definitions of the theta functions $\theta_i(\nu, t)$ may be found, for example, in van der Pol and Bremmer (1959). Only $\theta_2(\nu, t)$ and $\theta_3(\nu, t)$ are of interest here and are given by the equations:

$$\theta_2(\nu, t) = \sum_{k=-\infty}^{\infty} \exp\{-\pi(k-\tfrac{1}{2})^2 t + 2\pi i(k-\tfrac{1}{2})\nu\}$$

$$= t^{-\frac{1}{2}} \sum_{k=-\infty}^{\infty} (-1)^k \exp\{-\pi(k+\nu)^2/t\} \qquad (3.1)$$

$$\theta_3(\nu, t) = \sum_{k=-\infty}^{\infty} \exp\{-\pi k^2 t + 2\pi i k\nu\}$$

$$= t^{-\frac{1}{2}} \sum_{k=-\infty}^{\infty} \exp\{-\pi(k+\nu)^2/t\} \qquad (3.2)$$

In each of (3.1) and (3.2), the first equation provides a definition and the second one depends on Jacobi's transformation of theta functions (for example, Whittaker and Watson (1963, § 21.51)),

$$\sum_{k=-\infty}^{\infty} e^{-k^2\pi/t} \cos 2\pi ka = t^{\frac{1}{2}} e^{-\pi a^2 t} \sum_{k=-\infty}^{\infty} \exp\{-k^2\pi t - 2\pi kat\} \qquad (3.3)$$

We now introduce the definitions:

$$\theta_2 = \theta_2(z) = \theta_2(0, \tfrac{1}{2}\pi z^{-2}) \qquad (3.4)$$

$$\theta_3 = \theta_3(z) = \theta_3(0, \tfrac{1}{2}\pi z^{-2}) \qquad (3.5)$$

$$\theta_2^{(n)} = \theta_2^{(n)}(z) = \frac{d^n}{dz^n}\ \theta_2 \tag{3.6}$$

$$\theta_3^{(n)} = \theta_3^{(n)}(z) = \frac{d^n}{dz^n}\ \theta_3\ . \tag{3.7}$$

Letting $\nu = 0$ and $t = \frac{1}{2}\pi z^{-2}$ in (3.1) gives

$$\Theta_2(0, \tfrac{1}{2}\pi z^{-2}) = (\tfrac{1}{2}\pi)^{-\frac{1}{2}}\ z \sum_{k=-\infty}^{\infty} (-1)^k\ e^{-2k^2z^2} \tag{3.8}$$

Combining (3.8) and (1.2) yields

$$K_0(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}}\ z^{-1}\Theta_2(0, \tfrac{1}{2}\pi z^{-2})$$

$$= (\tfrac{1}{2}\pi)^{\frac{1}{2}}\ z^{-1}\theta_2\ . \tag{3.9}$$

Using the same substitutions in (3.2) and combining with (2.1) yields

$$K_0^*(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}}\ z^{-1}\Theta_3(0, \tfrac{1}{2}\pi z^{-2})$$

$$= (\tfrac{1}{2}\pi)^{\frac{1}{2}}\ z^{-1}\theta_3\ . \tag{3.10}$$

We can now calculate the various derivatives needed for expressing $K_i(z)$; $i = 1, 2, 3$ in terms of the theta functions using equations (3.9) and (3.10). The notation $\Delta_n(z)$; $n = 0, 1, 2, \ldots$ is defined in Chapter 2.

$$\Delta_1(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}} [z^{-1}\theta_2^{(1)} - z^{-2}\theta_2] \tag{3.11}$$

$$\Delta_1^*(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}} [z^{-1}\theta_3^{(1)} - z^{-2}\theta_3] \tag{3.12}$$

$$\Delta_2(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}} [z^{-1}\theta_2^{(2)} - 2z^{-2}\theta_2^{(1)} + 2z^{-3}\theta_2] \tag{3.13}$$

$$\Delta_2^*(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}} [z^{-1}\theta_3^{(2)} - 2z^{-2}\theta_3^{(1)} + 2z^{-3}\theta_3] \tag{3.14}$$

$$\Delta_3(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}} [z^{-1}\theta_2^{(3)} - 3z^{-2}\theta_2^{(2)} + 6z^{-3}\theta_2^{(1)} - 6z^{-4}\theta_2] . \tag{3.15}$$

From (2.9) and (3.11)

$$K_1(z) = \Delta_1(z)/6 = (\tfrac{1}{2}\pi)^{\frac{1}{2}} [z^{-1}\theta_2^{(1)} - z^{-2}\theta_2]/6 . \tag{3.16}$$

From (2.10) and (3.9)-(3.13)

$$\begin{aligned} K_2(z) &= [2\Delta_0(z) - 8z\Delta_1(z) - (2z^2 - 1)\Delta_2(z) \\ &\quad - 2\Delta_0^*(z) - 2z\Delta_1^*(z)]/72 \\ &= (\tfrac{1}{2}\pi)^{\frac{1}{2}} [2z^{-1}\theta_2 - 8z(z^{-1}\theta_2^{(1)} - z^{-2}\theta_2) \\ &\quad - (2z^2 - 1)(z^{-1}\theta_2^{(2)} - 2z^{-2}\theta_2^{(1)} + 2z^{-3}\theta_2) \\ &\quad - 2z^{-1}\theta_3 - 2z(z^{-1}\theta_3^{(1)} - z^{-2}\theta_3)]/72 \end{aligned}$$

$$= (\tfrac{1}{2}\pi)^{\frac{1}{2}} [-(2z - z^{-1})\theta_2^{(2)} - (4 + 2z^{-2})\theta_2^{(1)}$$

$$+ (6z^{-1} + 2z^{-3})\theta_2 - 2\theta_3^{(1)}]/72 \qquad (3.17)$$

From (2.11) and (3.11)-(3.15)

$$K_3(z) = [-26\Delta_1(z) - 148z\Delta_2(z) - (30z^2 - 5)\Delta_3(z)$$

$$- 60\Delta_1^*(z) - 30z\Delta_2^*(z)]/6480$$

$$= (\tfrac{1}{2}\pi)^{\frac{1}{2}} [-26(z^{-1}\theta_2^{(1)} - z^{-2}\theta_2) - 148z(z^{-1}\theta_2^{(2)} - 2z^{-2}\theta_2^{(1)} + 2z^{-3}\theta_2)$$

$$- (30z^2 - 5)(z^{-1}\theta_2^{(3)} - 3z^{-2}\theta_2^{(2)} + 6z^{-3}\theta_2^{(1)} - 6z^{-4}\theta_2)$$

$$- 60(z^{-1}\theta_3^{(1)} - z^{-2}\theta_3) - 30z(z^{-1}\theta_3^{(2)} - 2z^{-2}\theta_3^{(1)} + 2z^{-3}\theta_3)]/6480$$

$$= (\tfrac{1}{2}\pi)^{\frac{1}{2}} [-(30z - 5z^{-1})\theta_2^{(3)} - (58 + 15z^{-2})\theta_2^{(2)}$$

$$+ (90z^{-1} + 30z^{-3})\theta_2^{(1)} - (90z^{-2} + 30z^{-4})\theta_2 - 30\theta_3^{(2)}]/6480 . \qquad (3.18)$$

As far as we know, tables of the second and third derivatives of the theta functions are not yet available. When they become available the formulae of this section may be found convenient.

CHAPTER 4

Expressions of $K_i(z)$ when z is Small

If we give $\underline{a}$ the value $\frac{1}{2}$ and t the value $\frac{1}{2}\pi z^{-2}$ in (3.3) we obtain (1.7) whereas if t is given the same value and $\underline{a}$ the value 1 we obtain

$$K_0^*(z) = \sum_{k=-\infty}^{\infty} e^{-2k^2z^2} = (\tfrac{1}{2}\pi)^{\frac{1}{2}} z^{-1} \exp(-\tfrac{1}{2}\pi^2 z^{-2}) \sum_{k=-\infty}^{\infty} \exp\{-\tfrac{1}{2}k^2\pi^2 z^{-2} - k\pi^2 z^{-2}\}$$

$$= (\tfrac{1}{2}\pi)^{\frac{1}{2}} z^{-1} \sum_{k=-\infty}^{\infty} \exp(-\tfrac{1}{2}k^2\pi^2 z^{-2}) \ . \tag{4.1}$$

We can now calculate the transformed equations for Δ_i; i = 1, 2, 3 and Δ_i^*; i = 1, 2 as in Section 2.

$$\Delta_1(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}} \sum_{k=-\infty}^{\infty} [-z^{-2} + \pi^2 z^{-4}(k + \tfrac{1}{2})^2] \exp\{-\tfrac{1}{2}\pi^2 (k + \tfrac{1}{2})^2 z^{-2}\} \tag{4.2}$$

$$\Delta_1^*(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}} \sum_{k=-\infty}^{\infty} [-z^{-2} + \pi^2 z^{-4} k^2] \exp(-\tfrac{1}{2}\pi^2 k^2 z^{-2}) \tag{4.3}$$

$$\Delta_2(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}} \sum_{k=-\infty}^{\infty} [2z^{-3} - 5\pi^2 z^{-5}(k + \tfrac{1}{2})^2 + \pi^4 z^{-7}(k + \tfrac{1}{2})^4]$$

$$\exp\{-\tfrac{1}{2}\pi^2 (k + \tfrac{1}{2})^2 z^{-2}\} \tag{4.4}$$

$$\Delta_2^*(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}} \sum_{k=-\infty}^{\infty} [2z^{-3} - 5\pi^2 z^{-5} k^2 + \pi^4 z^{-7} k^4] \exp(-\tfrac{1}{2}\pi^2 k^2 z^2) \tag{4.5}$$

$$\Delta_3(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}} \sum_{k=-\infty}^{\infty} [-6z^{-4} + 27\pi^2 z^{-6}(k + \tfrac{1}{2})^2$$

$$- 12\pi^4 z^{-8}(k + \tfrac{1}{2})^4 + \pi^6 z^{10}(k + \tfrac{1}{2})^6] \exp\{-\tfrac{1}{2}\pi^2 (k + \tfrac{1}{2})^2 z^{-2}\} \tag{4.6}$$

From equations (2.9) and (4.2), we have

$$K_1(z) = \Delta_1(z)/6$$

$$= (\tfrac{1}{2}\pi)^{\frac{1}{2}} \sum_{k=-\infty}^{\infty} (6z^4)^{-1} \left[\pi^2(k + \tfrac{1}{2})^2 - z^2\right] \exp\{-\tfrac{1}{2}\pi^2(k + \tfrac{1}{2})^2 z^{-2}\} \quad (4.7)$$

From equations (2.10), (1.7), and (4.1)-(4.4), we have

$$K_2(z) = [2\Delta_0(z) - 8z\Delta_1(z) - (2z^2 - 1)\Delta_2(z) - 2\Delta_0^*(z) - 2z\Delta_1^*(z)]/72$$

$$= (\tfrac{1}{2}\pi)^{\frac{1}{2}}(72)^{-1} \sum_{k=-\infty}^{\infty} [2z^{-1} - 8z(-z^{-2} + \pi^2 z^{-4}(k + \tfrac{1}{2})^2)$$

$$- (2z^2 - 1)(2z^{-3} - 5\pi^2 z^{-5}(k + \tfrac{1}{2})^2 + \pi^4 z^{-7}(k + \tfrac{1}{2})^4)]$$

$$\exp\{-\tfrac{1}{2}\pi^2(k + \tfrac{1}{2})^2 z^{-2}\}$$

$$+ (\tfrac{1}{2}\pi)^{\frac{1}{2}}(72)^{-1} \sum_{k=-\infty}^{\infty} [-2z^{-1} - 2z(-z^{-2} + \pi^2 z^{-4} k^2)]$$

$$\exp(-\tfrac{1}{2}\pi^2 k^2 z^{-2})$$

$$K_2(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}} \sum_{k=-\infty}^{\infty} (72z^7)^{-1}[(6z^6 + 2z^4) + \pi^2(2z^4 - 5z^2)$$

$$(k + \tfrac{1}{2})^2 + \pi^4(1 - 2z^2)(k + \tfrac{1}{2})^4] \exp\{-\tfrac{1}{2}\pi^2(k + \tfrac{1}{2})^2 z^{-2}\}$$

$$- (\tfrac{1}{2}\pi)^{\frac{1}{2}} \sum_{k=-\infty}^{\infty} (36z^3)^{-1}\pi^2 k^2 \exp(-\tfrac{1}{2}\pi^2 k^2 z^{-2}) \quad (4.8)$$

From equations (2.11) and (4.2)-(4.6), we have

$$K_3(z) = [-26\Delta_1(z) - 148z\Delta_2(z) - (30z^2 - 5)\Delta_3(z)$$

$$- 60\Delta_1^*(z) - 30z\Delta_2^*(z)]/6480$$

$$= (\tfrac{1}{2}\pi)^{\frac{1}{2}}(6480)^{-1} \sum_{k=-\infty}^{\infty} [-26(-z^{-2} + \pi^2 z^{-4}(k + \tfrac{1}{2})^2)$$

$$- 148z(2z^{-3} - 5\pi^2 z^{-5}(k + \tfrac{1}{2})^2 + \pi^4 z^{-7}(k + \tfrac{1}{2})^4)$$

$$- (30z^2 - 5)(-6z^{-4} + 27\pi^2 z^{-6}(k + \tfrac{1}{2})^2 - 12\pi^4 z^{-8}(k + \tfrac{1}{2})^4$$

$$+ \pi^6 z^{10}(k + \tfrac{1}{2})^6] \exp\{-\tfrac{1}{2}\pi^2(k + \tfrac{1}{2})^2 z^{-2}\}$$

$$+ (\tfrac{1}{2}\pi)^{\frac{1}{2}}(6480)^{-1} \sum_{k=-\infty}^{\infty} [-60(-z^{-2} + \pi^2 z^{-4}k^2)$$

$$- 30z(2z^{-3} - 5\pi^2 z^{-5}k^2 + \pi^2 z^{-7}k^2)] \exp(-\tfrac{1}{2}\pi^2 k^2 z^2)$$

$$K_3(z) = (\tfrac{1}{2}\pi)^{\frac{1}{2}} \sum_{k=-\infty}^{\infty} (6480z^{10})^{-1} [-(90z^8 + 30z^6)$$

$$- \pi^2(96z^6 - 135z^4)(k + \tfrac{1}{2})^2 + \pi^4(212z^4 - 60z^2)(k + \tfrac{1}{2})^4$$

$$- \pi^6(30z^2 - 5)(k + \tfrac{1}{2})^6] \exp\{-\tfrac{1}{2}\pi^2(k + \tfrac{1}{2})^2 z^{-2}\}$$

$$+ (\tfrac{1}{2}\pi)^{\frac{1}{2}} \sum_{k=-\infty}^{\infty} (216z^6)^{-1} [-\pi^4 k^4 + 3\pi^2 k^2 z^2] \exp(-\tfrac{1}{2}\pi^2 k^2 z^{-2}) \qquad (4.9)$$

An alternative method for the derivation of these formulae is by the use of the Poisson Summation formula which can be expressed as

$$\sum_{n=-\infty}^{\infty} f^*(n/\lambda) = \lambda \sum_{m=-\infty}^{\infty} f(\lambda m) \tag{4.10}$$

where

$$f^*(t) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i t} \, dx \tag{4.11}$$

Under regularity conditions one of the applications of the Poisson Summation Formula is to prove Jacobi's Transformation. In this work it was used as a check on the derivations of equations (1.7), (4.1) and (4.7)-(4.9). Let $f(x) = e^{-x^2}$. Then by equation (4.11) we have $f^*(t) = \pi^{\frac{1}{2}} e^{-\pi^2 t^2}$. Substitution into equation (4.10) gives

$$\sum_{n=-\infty}^{\infty} \pi^{\frac{1}{2}} e^{-n^2\pi^2/\lambda^2} = \lambda \sum_{m=-\infty}^{\infty} e^{-\lambda^2 m^2} \tag{4.12}$$

This equation is in the form necessary to transform equations (1.4)-(1.6) into (4.7)-(4.9). The equations (4.7)-(4.9) are not necessarily in the form best suited for calculations, especially for extremely small z. In this situation the calculations should be made with care so that computer rounding does not introduce large errors into the final result.

CHAPTER 5

Comparison of Methods and Recommendations

Durbin's method for calculating the exact tail area probability of the Kolmogorov-Smirnov Statistic requires the computation of the Nth power of a matrix of $1 + 2[z\sqrt{N}]$ rows or columns, where $[z\sqrt{N}]$ denotes the integral part of $z\sqrt{N}$. It is therefore easy to carry out on a computer when $z\sqrt{N}$ is small and N is not extremely large. Of course the Nth power of a matrix $\underset{\sim}{M}$ can be obtained by first expressing N in the binary system of notation, and then multiplying together the appropriate selection of the matrices M, M^2, M^4, M^8, ... It is therefore doubtful whether N would ever be large enough in practice to cause trouble when $z\sqrt{N}$ is not large. The method using the recursion formulae is equivalent to the matrix method, and therefore has the same limitation.

The values obtained from Li-Chien's approximation formulae, or our transformation of them, were compared with the exact values from the recursion formulae tabulated by Birnbaum in our Table 1.1. For each combination of c and N, the first entry is the approximation value and the second is the value obtained from Birnbaum's Table 1. Our Table 5.1 shows that the approximation is correct to five decimal places by the time N = 100.

Table 5.2 compares the approximation formulae with the limiting distribution K_0 given by (1.2). In this table the entry for a specific combination of z and N gives the probability determined by the approximation formulae. The value given by K_0 does not depend on N so here there is only one entry for each z.

For comparison of Li-Chien's formulae with the transformed formulae, a computer program was written to count the number of terms in each series necessary to achieve five decimal place accuracy for $K_i(z)$; $i = 0, \ldots, 3$. Table 5.3 summarizes the results. For $z \cong 1.2$, the number of terms needed for each method is the same; but the complexity of the transformed formulae makes their calculation more difficult. The exact point where both methods require the same amount of effort will depend on the computer, but will usually depend primarily on the number of multiplications. A good rule of thumb is to use the transformed formulae when $z \leq 0.8$; otherwise the formulae given by Li-Chien and Korolyuk would be easier to use.

Figure 5.1 shows the recommended range of z and N for each method mentioned in this paper using the criterion of accuracy to five significant figures.

TABLE 5.1

Comparison of Birnbaum's Table 1 with Li-Chien's Approximation or Its Transformation

	N					
c	60		80		100	
1	.00000	.00000	.00000	.00000	.00000	.00000
3	.00306	.00349	.00031	.00033	.00003	.00003
5	.22563	.23242	.10396	.10597	.04678	.04678
7	.63429	.64035	.45369	.45664	.31533	.31533
9	.87593	.87889	.75363	.75557	.62937	.62937
11	.96808	.96916	.91087	.91182	.83504	.83504
13	.99375	.99406	.97374	.97412	.93791	.93791
15	.99907	.99914	.99370	.99382	.98016	.98016

TABLE 5.2

Comparison of the Limiting Distribution K_0 with
Li-Chien's Approximation or Its Transformation

		N				
z	Limiting Distribution	1000	5000	10000	50000	100000
0.2	.50521(-12)	.20908(-11)	.99802(-12)	.82358(-12)	.63126(-12)	.59181(-12)
0.4	.28077(-2)	.33629(-2)	.30507(-2)	.29786(-2)	.28836(-2)	.28613(-2)
0.6	.13572	.14263	.13919	.13792	.13670	.13642
0.8	.45586	.46431	.45967	.45856	.45707	.45672
1.0	.73000	.73559	.73252	.73178	.73080	.73056
1.4	.96032	.96150	.96084	.96069	.96053	.96044
1.8	.99693	.99706	.99699	.99697	.99695	.99694
2.2	.99988	.99988	.99988	.99988	.99988	.99988

TABLE 5.3

Number of Terms Necessary to Obtain
Five Decimal-Place Accuracy

z	Li-Chien's Approximation [(1.2),(1.4),(1.5),(1.6)]	Transformed Approximation [(1.7),(4.7),(4.8),(4.9)]
0.05	64	2
0.1	30	2
0.2	13	2
0.4	7	3
0.6	5	3
0.8	4	3
1.0	3	3
1.2	3	3
1.4	3	3
1.6	3	4
1.8	3	4
2.0	3	4
2.5	3	5

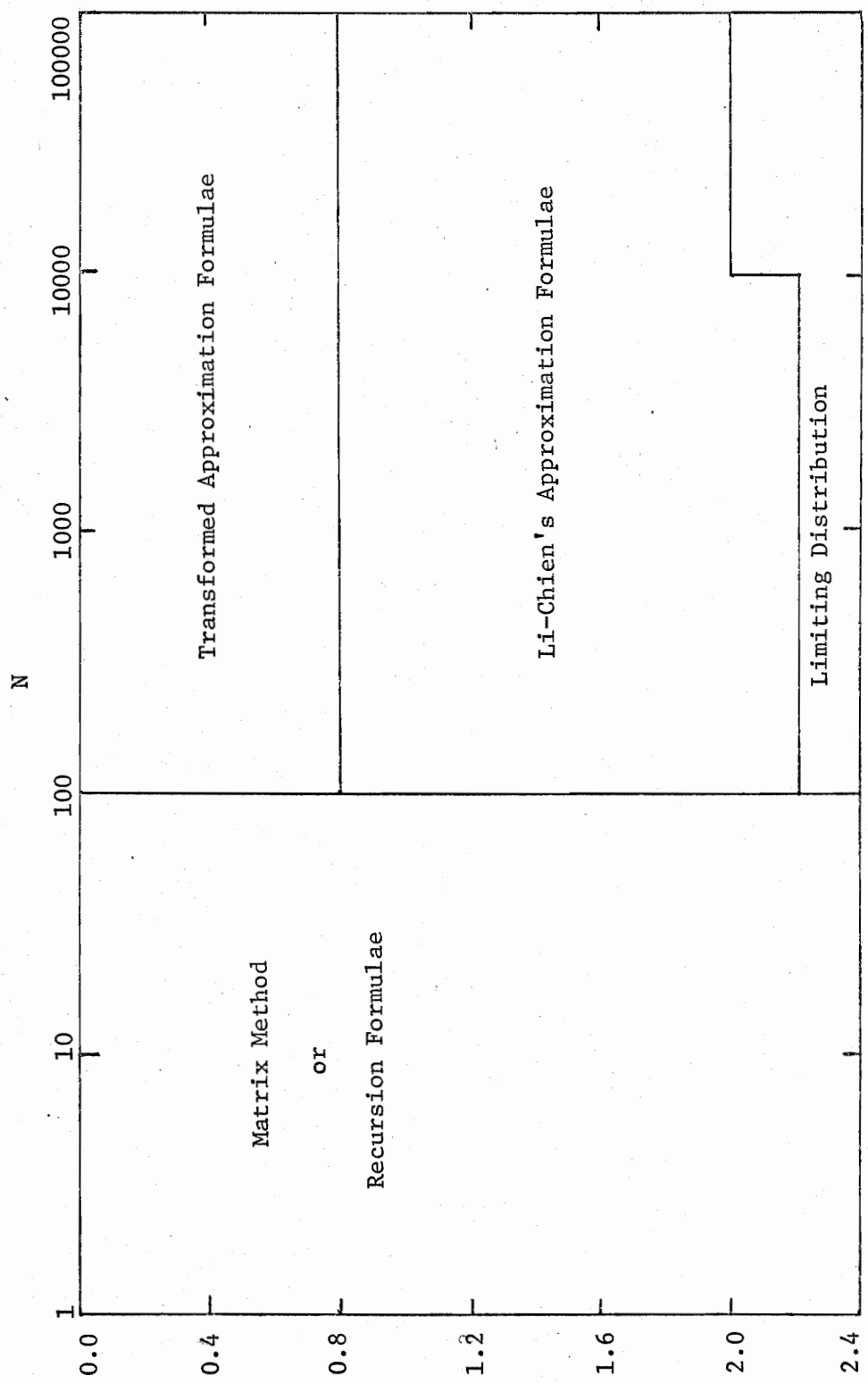

Figure 5.1 Recommended Ranges of Use for the Different Methods.

BIBLIOGRAPHY

(Part II)

[1] Birnbaum, Z.W., "Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size," *J. Amer. Statist. Assoc.*, 47(1952), 425-441.

[2] Durbin, J., *Distribution Theory for Tests Based on the Sample Distribution Function*, Society for Industrial and Applied Mathematics, Philadelphis, 1973.

[3] Feller, W., "On the Kolmogorov-Smirnov limit theorems for Empirical Distributions," *Ann. Math. Statist.*, 19(1948), 177-189.

[4] Good, I.J. "A Nonparametric Roughness Penalty for Probability Densities," *Nature (Physical Science)*, 229(1971), 29-30. (Contains 21 misprints owing to a postal strike but corrected reprints are available).

[5] Good, I.J. and Gaskins, R.A., "Nonparametric Roughness Penalties for Probability Densities," *Biometrika*, 58(1971), 255-277.

[6] Good, I.J. and Gaskins, R.A., "Global Nonparametric Estimation of Probability Densities," *Virginia Journal of Science*, 23(1972), 171-193.

[7] Good, I.J. and Gaskins, R.A., "Bump-hunting by the Penalized-Likelihood Method" (1974), unpublished.

[8] Kolmogorov, A. "Sulla Determinazione Empirica Di Una Legge Di Distribuzione," *Giornale dell'Instituto Italiano degli Attuari*, 4(1933), 83-91.

[9] Korolyuk, V.S., "Asymptotic Analysis of the Distribution of the Maximum Deviation in the Bernoulli Scheme," *Theory of Probability and its Applications*, 4(1960), 339-366.

[10] Li-Chien, C, "On the Exact Distribution of the Statistics of A.N. Kolmogorov and their Asymptotic Expansion," *Acta Mathematica Sinica*, 6(1956), 55-81.

[11] Smirnov, N., "Table for Estimating the Goodness of Fit of Empirical Distributions," *Ann. Math. Statist.*, 19(1948), 279-281.

[12] van der Pol, B. and Bremmer, H., *Operational Calculus Based on the Two-Sided Laplace Integral*, Cambridge Univ. Press, London, 1959.

[13] Whittaker, E.T. and Watson, G.N., *A Course of Modern Analysis, Fourth Edition*, Cambridge University Press, Cambridge, 1963.

VITA

Wolfgang Pelz was born on July 22, 1949, in Stuttgart, Germany to Mr. and Mrs. Heinz G. Pelz. He graduated as valedictorian from Deer Park High School near Cincinnati, Ohio, in June, 1967. He then attended Rose Polytechnic Institute in Terre Haute, Indiana, under a National Merit Scholarship, and graduated with high honors, receiving a Bachelor of Science degree in mathematics. He entered V.P.I. and S.U. in September, 1971, and received a Master of Science degree in statistics in June, 1973. He interrupted his graduate education in September, 1974, to accept a position as Management Scientist with the B.F. Goodrich Company in Akron, Ohio. He took a leave of absence beginning November, 1976, in order to complete the requirements for a Ph.D. degree in statistics and returned to B.F. Goodrich in August, 1977.

He is a member of Phi Kappa Phi scholastic honorary society, Pi Mu Epsilon mathematics honorary society, and the American Statistical Association.

Wolfgang Pelz

TOPICS ON THE ESTIMATION OF SMALL PROBABILITIES

by

Wolfgang Pelz

(ABSTRACT)

In Part I the Maximum Likelihood/Entropy (ML/E) method of estimation of the cell probabilities for multinomial and contingency table problems is derived and discussed. This method is a generalization of the Maximum Likelihood estimator to situations when small probabilities are to be estimated and the standard Maximum Likelihood estimator is inadequate. In addition when no sample exists the technique gives meaningful results by reducing to the method of Maximum Entropy. The ML/E method is based on assuming an entropy prior on the cell probabilities and closely resembles the Pseudo-Bayes methods of Good, Fienberg, and Holland in which Dirichlet priors are assumed. Methods for calculating the ML/E estimates for varying circumstances including multidimensional tables are presented. Comparisons with other estimation methods are made and recommendations for selection of the more appropriate method in particular situations are given.

In Part II we consider the Kolmogorov-Smirnov one-sample statistic. Various methods for calculating the Kolmogorov-Smirnov one-sample statisti have been developed in the literature. A transformation of an approximation method is here derived and some of its properties discussed. The main value of the new formulae is to obtain better convenient approximations in the lower tail than have been possible

using other methods. The formulae are related to the theta functions. The relationships between various methods are given, as well as recommendations for each method of a usable range of the independent variable. An analysis is made of the errors obtained by use of the approximation.