MAXIMUM LIKELIHOOD ESTIMATION OF PARAMETERS USING GROUPED DATA SAMPLES

by

Jacob Van Bowen, Jr.

Thesis submitted to the Graduate Faculty of the
Virginia Polytechnic Institute
in partial fulfillment for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

APPROVED:

Chairman, Dr. Donald R. Jensen

Dr. Boyd Harshbarger

Dr. Richard G. Krutchkoff

Dr. Clyde Y. Kramer

Professor W. Emory Pace

August, 1968

Blacksburg, Virginia

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS

The author wishes to express his appreciation to Dr. Boyd Harshbarger for his assistance in obtaining a fellowship from the National Institutes of Health, grant numbers 5T1GM2-08, 5T1GM2-09, and 5T1GM2-10; and in obtaining a teaching assistantship for the academic year of 1967-1968. The funds received from these sources made possible the author's graduate studies.

Figure 1 was prepared by Benjamin F. Dyer, Jr. This contribution is greatly appreciated by the author. Consultations with Glen H. Lemon and Professor W. Emory Pace were very valuable to the author. The author is indebted to Dr. Richard G. Krutchkoff for reading the manuscript and for making helpful suggestions and corrections. The author is exceedingly grateful to his wife, Carol J. Bowen, who spent many hours at the arduous task of typing the manuscript.

Sincere appreciation is extended to Dr. Donald R. Jensen whose suggestions and comments were invaluable during the preparation of this manuscript. The author wishes to thank Dr. Donald R. Jensen for the constant assistance which he has provided as a teacher and as a major professor during the four years in which the author has pursued his graduate studies at Virginia Polytechnic Institute.

LIST OF TABLES AND FIGURE

Page

CHAPTER I

INTRODUCTION

1.1 Introduction

One of the major problems in statistics is that of estimating parameters. This process can be described as follows. Let X be a random variable defined in some subset Ω of the real line. Assume that the events which are to be considered can be assigned relative frequencies by only one of the distribution functions, say $F(x;\theta_o)$, in the family of distribution functions $\{F(x;\theta):\theta \in \Theta\}$, where Θ is some subset of the real line. In order to estimate θ_o, a sample $X_1, \ldots, X_n$, usually a random sample, is observed and some rule based on the sample is used to obtain an estimator of θ_o, say $\hat{\theta}_o = \delta(X_1, \ldots, X_n)$. It usually is assumed that the space on which δ is defined is the n-dimensional Cartesian product space of Ω, i.e. the sample space $\Omega^{(n)}$.

Suppose that the sampling experiment being conducted is such that if $X = x_o \in S$, $S \subset \Omega$, then x_o is observed and if $X = x_o \notin S$, then x_o cannot be observed. The n-dimensional Cartesian product space of S, $S^{(n)}$, will be called the actual sampling space for the random variable X. If for any reason the actual sampling space $S^{(n)}$ is a proper subset of $\Omega^{(n)}$ such that $Pr\{(X_1, X_2, \ldots, X_n) \in S^{(n)}\} < 1$, we shall say that $S^{(n)}$ constitutes an incomplete sampling space. Samples from this space will be called incomplete samples or incomplete data.

There are many examples of this type of sampling space. One of the most frequent examples is that of a measuring device with limited precision. Such devices usually have a range $R \subset \Omega$ which is partitioned into the intervals $(r_1,r_2]$, $(r_2,r_3]$, ..., $(r_{k-2},r_{k-1}]$. If $x_o \in (r_i,r_{i+1}]$, $0 < i < k-1$, the device registers $\frac{r_i + r_{i+1}}{2}$, which is called the class mark of the interval $(r_i,r_{i+1}]$. If x_o is not in R, then the device registers either $x_o \leq r_1$ or $x_o > r_{k-1}$, whichever the case may be. Another example is a population which is assumed to have a frequency function proportional to a known frequency function $f(x;\theta)$, $x \in \Omega$ and $\theta \in \Theta$, on some subset of Ω, and zero otherwise. This is, of course, a generalization of the classical term "truncation" which usually refers to the special cases in which either one or both "tails" (or extremes) of the domain of $f(x;\theta)$ are excluded from Ω to form the sampling space.

There are some cases in which certain values of X are not observed for other reasons. Some of the classical incomplete data problems will be defined in the Literature Review.

In this thesis we shall be interested primarily in the problem of estimating the parameters of a distribution function when grouped data only are available. The term grouped data will be used to describe the results of a sampling experiment in which the only information recorded is

that m_i observations were found to be less than or equal to r_i, $i = 1, \ldots, k$, where $-\infty < r_1 < r_2 < \ldots < r_{k-1} < r_k = \infty$. The set $\{r_i\}_{i=1}^{k}$ determines an aggregate of intervals

$$I_1 \equiv (-\infty, r_1], \; I_2 \equiv (r_1, r_2], \; \ldots, \; I_k \equiv (r_{k-1}, \infty] \; .$$

These intervals will be called cells. The number of observations in the i^{th} cell, $m_i - m_{i-1}$, will be called the cell frequencies and will be denoted by n_i.

The term "grouped data" can be generalized to apply to multivariate sampling experiments. For example, consider the random vector (X,Y,Z). Assume that samples from the population of vectors (X,Y,Z) are grouped such that the measurements of the x-coordinate are grouped into cells $I_1, \ldots, I_k$ and the measurements of the y-coordinate are grouped into cells $J_1, \ldots, J_s$ and the measurements of the z-coordinate are not grouped. Then a sample of size n would consist of a set of observations

$$\{X_t \in I_{i(t)}, \; Y_t \in J_{j(t)}, \; Z_t = z_t\}_{t=1}^{n} \; ,$$

where i(t) can range from 1 to k and j(t) can range from 1 to s.

In this thesis the method of maximum likelihood estimation will be considered. Other approaches are discussed in the Literature Review, but the method of maximum likelihood seems to be the most fruitful method for reasons

to be developed. We shall apply some existing theorems to the problem of estimating parameters using grouped data and thereby obtain some of the properties of the estimators for large samples.

Unfortunately, when one considers the maximum likelihood method of estimation using grouped data samples, he finds that in most cases he will obtain the estimators only after some iterative procedure has been completed. He will seldom know beforehand whether there is a maximum likelihood estimate or whether, if there is an estimate, it is unique.

There have been several approaches to the problem of finding the properties of the maximum likelihood estimators. Frequently these approaches, which will be discussed in the Literature Review, have been confused by authors and as a result there is frequent use of the phrase "under mild regularity conditions." There is frequent misleading usage of this phrase as well. In this respect we shall attempt to be most discreet in the subsequent development.

1.2 Literature Review

1.2.1 Obtaining the Maximum Likelihood Estimates in Particular Cases

Many of the questions which arise when considering problems of estimation involving grouped data also arise when considering problems of estimation involving other types of incomplete data. A voluminous literature has developed around attempts to answer some of these questions for special cases of incomplete data. In this section an attempt is made to present a somewhat complete review of the literature in which grouped data problems are studied and to examine some of the related works in which other types of incomplete data problems are studied. Before consideration is given to some of this literature, we adopt the following definitions.

If a population has the frequency function $f^*(x) = c\ f(x)$, $c > 1$, on some fixed sub-interval, I, of the domain of the frequency function $f(x)$ and is zero otherwise, then we say that the population has the truncated $f(x)$ frequency function, $f^*(x)$. One-tail truncation is seen to be a special case of truncation. Truncation is a property of a frequency function, or equivalently, of a population. The actual sampling space for the random variable X is $I^{(n)}$.

This concept should be compared with a special class of conditional distributions; see section 2.9 of Wilks (1962).

If X is a random variable defined on the set $\overline{X}$ with frequency function f(x), and if

$$\Pr\{X \in S \subset \overline{X}\} = p, \quad 0 < p < 1,$$

then the conditional frequency function of X given that $X \in S$ is $f^*(x) = p^{-1}f(x)$ if $x \in S$ and zero otherwise. If S is the set of all points less than $\alpha \in \overline{X}$, then $f^*(x)$ is the frequency function of X truncated in the right tail at α. Therefore, truncated distributions can be considered to be special cases of conditional distributions.

If the sampling space is such that the values of the random variable are recorded when they are in the interval $I_2 \equiv (r_1, r_2]$, but only the number of occurrences in $I_1 \equiv (-\infty, r_1]$ and the number of occurrences in $I_3 \equiv (r_2, \infty]$ are recorded, then we call this type of sampling Type I censoring. One-tail Type I censoring is the special case for which either $r_2 = \infty$ or $r_1 = -\infty$. Type I censoring is a property of the sampling procedure and should not be confused with Type II censoring which was studied by Gupta (1952).

Hald (1949) obtained the maximum likelihood estimators of the parameters of the univariate normal distribution in which "all record is omitted of observations below a given value."* He also obtained these estimators for the case "in

---

* See page 119 of Hald (1949).

which the frequency of observations below a given value is recorded but the individual values of these observations are not specified."* His estimators must be found by solving relatively complicated equations. Four tables were given to aid in their solution. The first of these cases was treated by R.A. Fisher (1931). The second case was treated by W.L. Stevens in the Appendix, entitled "The Truncated Normal Distribution", of an article by C.I. Bliss (1937).

Hald made a distinction between these two cases. He used the term "truncated" to denote the first case and the term "censored" to denote the second case. The second case now is called Type I censoring to distinguish it from Type II censoring.

Tukey (1949) established that if X is a random variable whose distribution depends on the parameter θ and if $\hat{\theta}$ is a sufficient statistic for θ, then $\hat{\theta}$, the same function of the observations from any truncated distribution of X, is also sufficient for θ. Smith (1957) extended this result and showed that the further property of $\hat{\theta}$ being a minimal sufficient statistic is retained under truncation.

Cohen (1950) presented a consolidated treatment of maximum likelihood estimation of the unknown parameters of a univariate normal distribution in the cases of one- or

---

* See page 119 of Hald (1949).

two-tail truncation and one- or two-tail Type I censoring. Cohen (1957) later presented aid charts and an iterative procedure which simplify the solution of the maximum likelihood equations for these cases. In other articles, (1949) and (1955), Cohen has considered further the problem of estimating parameters of a truncated normal distribution.

Birnbaum (1950) considered the following problem. Assume that $(X_1, X_2, \ldots, X_p, Y_1, \ldots, Y_q)$ is a random vector having the (p+q)-variate normal distribution, where $Y_1, \ldots, Y_q$ are the scores on tests for admission to an educational institution and $X_1, \ldots, X_p$ are the scores on various aptitude tests. He treated the following question. If the decision to admit an applicant is based on the statistic

$$\sum_{i=1}^{q} a_i Y_i$$

being greater than or less than some "cutting" value c, then how can one select $a_1, \ldots, a_q$ and c in order to obtain certain desirable characteristics of the subsequent distribution of $(X_1, \ldots, X_p)$? Birnbaum (1950) considered only the special case where X and Y have the bivariate normal distribution and he examined the properties of the marginal distribution of X given that $Y > c$, where c is a constant. He showed that the variance of this marginal distribution is less than the variance of the marginal distribution of X if the correlation between X and Y is not zero.

Des Raj (1953) considered the problem of estimating the parameters of the bivariate normal distribution when one of the random variables is restricted to lie between two known values, c_1 and c_2. He showed that the method of maximum likelihood and the method of moments give the same results.

Swamy (1959-1960) used the method of maximum likelihood to estimate the mean and variance of the normal distribution when grouped data only are available. He also considered the method of maximum likelihood to estimate the mean and variance of the normal distribution which is truncated in either one or both tails when grouped data only are available. He claimed that the work of Huzurbazar (1947-1949) and of Kulldorff (1958a, 1958b) substantiate his statement that the maximum likelihood estimators can be shown to be consistent and unique in all of these cases. The author can neither deduce this from the work of Huzurbazar (1947-1949) and of Kulldorff (1958a, 1958b), nor can he deduce this from any other source. Conditions which insure the consistency of the maximum likelihood estimators of the mean and variance obtained from a grouped data sample from a normal distribution will be established in Chapter III.

Doss (1962) showed that the maximum likelihood estimators of the parameters of the univariate normal distribution are unique in the four cases of one-tail truncation, two-tail truncation, one-tail Type I censoring,

and two-tail Type I censoring. He considered "∞" to be a unique solution.

Gjeddebaek (1949) was among the first to consider the problem of estimating the mean and variance of the univariate normal distribution when the only information available is the number of observations falling into the intervals

$$I_1 \equiv (-\infty, r_1], \ldots, I_k \equiv (r_{k-1}, \infty].$$

He obtained the maximum likelihood equations which must be solved and he provided two tables to aid in their solution. Numerical examples were given in cases where $k = 3$ and $k = 4$.

Gjeddebaek claimed to have obtained the covariance matrix of the estimators for large sample sizes and he cited Cramér (1946)* as a reference in which justification for this claim might be found. In Chapter III it will be shown that the matrix which he obtained is the covariance matrix, although this result cannot be deduced from the development in Cramér (1946).

Gjeddebaek (1956) further considered grouped data in the special case of the univariate normal distribution with known variance for which the intervals are equally spaced, i.e.

---

* Gjeddebaek and Kulldorff cite Cramér (1945). Mathematical Methods of Statistics by Harald Cramér was first published in Sweden in 1945 and was first printed in the United States in 1946. We shall use 1946 as this text's reference date.

$$I_1 \equiv (-\infty, r_1], \; I_k \equiv (r_1+(k-2)h, \infty]$$

and if $1 < i < k$, then

$$I_i \equiv (r_1+(i-2)h, \; r_1+(i-1)h],$$

where $h > 0$. He studied the effect which the location of the true mean μ_o might have on the asymptotic efficiency of the maximum likelihood estimators of μ_o and the variance σ_o^2. His general conclusions are that if $h < 2\sigma_o$ and no point in $(\mu_o - 3\sigma_o, \mu_o + 3\sigma_o)$ is in I_1 or I_k, then the asymptotic efficiencies of the maximum likelihood estimators for the mean and variance, where both parameters are unknown or not, are not critically changed by the relative location of μ_o.

The term asymptotic efficiency as used here refers to the ratio of two expectations. If θ is a parameter to be estimated and L_c and L_g are the likelihood functions for the complete and grouped data cases respectively, then the asymptotic efficiency of the maximum likelihood estimator for θ in the grouped data case, relative to that of complete data, is defined to be

$$\left[-E \frac{\partial^2 \log L_g}{\partial \theta^2}\right]\left[-E \frac{\partial^2 \log L_c}{\partial \theta^2}\right]^{-1},$$

where the two expectations are taken with respect to all of the random variables in the samples defining L_g and L_c, respectively. This definition differs from a classical definition of asymptotic efficiency; see section 12.3.3 of Wilks (1962).

In a third paper Gjeddebaek (1957) considered using the midpoints of the intervals as quasi observations. Upon using the sample average of these quasi observations to obtain a simple estimator of the mean when the variance is known, he found that if the intervals are equally spaced and of length less than twice the standard deviation, and if the sample size is less than 100, then this simple estimator is almost as good as the maximum likelihood estimator. He noted, however, that his simple estimator is not consistent. Gjeddebaek (1959a) used Sheppard's correction formula to obtain a simple estimator of the variance using the midpoints of equally spaced intervals as quasi observations. He found this estimator to be almost as good as the maximum likelihood estimator if the (equal) lengths of the intervals are less than twice the standard deviation and if the sample size is less than 100. The use of Sheppard's correction formula is discussed later in this section when the work of Tallis (1967) is considered.

In subsequent papers Gjeddebaek (1959b, 1961) further considered the three-group case, the approximate distribution of the simple estimators based on quasi observations, and approximate tests of hypotheses about the mean in the three-group case.

In all of the work done by Gjeddebaek (1957) through (1961), restrictions on the range of the true value of μ

eliminate the occurrence of large numbers of observations in the half lines I_1 and I_k. For those simple estimators which utilize the midpoints of the intervals, the cell frequencies n_1 and n_k must be ignored.

Several authors have considered the problem of finding estimators which are simpler to obtain than the maximum likelihood estimators.

Khatri (1962) developed a procedure to approximate the maximum likelihood estimators obtained from a grouped data sample from a distribution which has the density function $f(x;\underline{\theta})$. The extreme cells, I_1 and I_k, are omitted in the process, and the lengths of the intervals $I_2, \ldots, I_{k-1}$ are assumed to be sufficiently small for

$$p_i(\underline{\theta}) = (r_i - r_{i-1})f(\tfrac{1}{2}(r_i + r_{i-1});\underline{\theta})$$

to be a good approximation of $\Pr\{X \in I_i;\underline{\theta}\}$. The function

$$\sum_{i=2}^{k-1} (\log (\tfrac{n_i}{n}) - \log p_i(\underline{\theta}))^2$$

is minimized with respect to the several parameters in $\underline{\theta}$. If additional regularity conditions are satisfied by the density function, then the procedure seems to provide quite a good approximation to the maximum likelihood estimators if n_1 and n_k are relatively small.

Tallis (1967) considered the general case of a univariate distribution with k equally spaced intervals of

length h. He assumed that h is not too large. He defined x_i to be the midpoint of the i^{th} interval and n_i to be the observed frequency in the i^{th} interval. If $\delta(X_1, \ldots, X_n)$ is the maximum likelihood estimator for a parameter in the complete data case, then the estimator he proposed is $\delta(y_1, \ldots, y_n) + g$, where y_i is the midpoint of the interval into which X_i falls and g is a rather complicated function depending on several derivatives of the logarithm of the frequency function. This estimator is a modification of that obtained by Lindley (1949).

In the case of the univariate normal distribution with unknown mean μ and unknown variance σ^2 the following estimators were obtained:

$$\hat{\mu} = \sum_{i=2}^{k-1} \frac{n_i x_i}{n} \quad , \qquad \hat{\sigma}^2 = \sum_{i=2}^{k-1} \frac{n_i (x_i - \hat{\mu})^2}{n} - \frac{h^2}{12} \quad .$$

The two half-lines are ignored, i.e. I_1 and I_k. These estimators are identical to those obtained using Sheppard's correction formula as considered by Gjeddebaek (1957). Tallis showed that the approximate variances of the estimators are

$$\sigma^2\{n\left(1 - \frac{h^2}{12\sigma^2}\right)\}^{-1} \quad \text{and} \quad 2\sigma^4\{n\left(1 - \frac{h^2}{6\sigma^2}\right)\}^{-1}$$

respectively for $\hat{\mu}$ and $\hat{\sigma}^2$. He also obtained similar estimators in the bivariate normal case which are identical

to those which would be obtained using Sheppard's correction formulas. Tallis extended his results to the multivariate case with several parameters and unequal interval lengths.

In most of the problems involving incomplete data one cannot obtain the distribution of the maximum likelihood estimators. In lieu of more definitive results, many authors consider the information matrix in attempts to describe the asymptotic properties of the maximum likelihood estimators. We define the information matrix as follows. Let H be the likelihood function for a sample of size n from a distribution with parameters $(\theta_1, \ldots, \theta_q)$. Then the matrix

$$[I_{ij}] = \left[E\left(\frac{1}{H}\frac{\partial H}{\partial \theta_i}\frac{1}{H}\frac{\partial H}{\partial \theta_j}\right)\right],$$

which is a q x q matrix, will be called the information matrix for a sample of size n; see section 17.39 of Kendall and Stuart (1961). The expectation is taken with respect to all of the random variables in the sample at the true value of $(\theta_1, \ldots, \theta_q)$. In some cases $[I_{ij}]$ is the inverse of the covariance matrix of the maximum likelihood estimators for large sample sizes, and therefore provides some of the large sample properties of the maximum likelihood estimators. In other cases this matrix has little or no meaning at all. Rao (1965, section 5c.2) gives a discussion of the limitations and interpretation of the information matrix. In the cases we shall study, the information matrix will be assumed to

have limited meaning unless proven otherwise.

There are several numerical methods which might be used to obtain the maximum likelihood estimators from a grouped data sample. The likelihood equations obtained from a grouped data sample are usually non-linear equations, and three methods which might be used to obtain the root (or roots) of the likelihood equations are the Newton-Raphson method, the method of successive approximations, and the method of false positions; see Fox (1963). It is usually very difficult to prove that the sequences defined by these methods will converge to the desired root.

Several authors have considered different numerical methods for solving the likelihood equations in particular cases involving grouped data. Among these authors are Kale (1961, 1962, 1966), Hartley (1958), and Hughes (1962). Barnett (1966), in the context of another class of problems, is one of the few authors who have considered the fact that in many cases the likelihood equations do not have a unique solution.

The method suggested by Hartley (1958) and later extended by Hughes (1962) is very similar to the method of successive approximations. An outline of the method studied by Hughes (1962), which will be referred to as Hughes' method, is given in the Appendix. This method involves the use of quadrature formulae to reduce the likelihood equations,

for a large class of incomplete data problems, to a form similar to that which would be obtained from a complete data sample. Then an iteration procedure is used. A major development in the work of Hughes (1962) is a theorem which establishes sufficient conditions for the convergence of the sequence of iterates. In the Appendix it is shown that these sufficient conditions are not satisfied in the case of a grouped data sample from a normal distribution.

One other method of solution, which is called the method of scoring, is suggested in many cases by various authors; see, for example, Kulldorff (1958a, 1958b, 1961) and section 5g in Rao (1965). This method might be described as an approximation of the Newton-Raphson method. Assume that $L(\theta)$ is the natural logarithm of the likelihood function for a sample of size n. If some root of $\frac{dL(\theta)}{d\theta} = 0$ is sought, then the application of the Newton-Raphson method would define a sequence of numbers $\theta_1, \theta_2, \ldots,$ from a starting value θ_1 by the following relation:

$$\theta_{j+1} = \theta_j - \left[\frac{dL(\theta)}{d\theta} \cdot \left(\frac{d^2L(\theta)}{d\theta^2}\right)^{-1}\right]_{\theta=\theta_j} .$$

The method of scoring defines a similar sequence, but the assumption is made that $\frac{d^2L(\theta)}{d\theta^2}$ is approximately equal to its expected value, $E\,\frac{d^2L(\theta)}{d\theta^2} = -I(\theta)$, where $I(\theta)$ is the information matrix (a scalar in this case) for a sample of

size n; see section 5g in Rao (1965). The method of scoring defines the sequence:

$$\theta'_{j+1} = \theta'_j + \left[\frac{dL(\theta)}{d\theta}\cdot\left(I(\theta)\right)^{-1}\right]_{\theta=\theta'_j} .$$

Sufficient conditions for the convergence of such a sequence to a root of the likelihood equation can be found in Scarborough (1930). The convergence properties of the numerical methods discussed here are usually ignored by their proponents.

1.2.2 Properties of the Maximum Likelihood Estimators

Among those authors who have contributed significantly to the theory of maximum likelihood estimation are Huzurbazar (1947-1949), Wald (1943, 1949), Cramér (1946), and Rao (1965). Each adopted a somewhat different approach to the problem, and these will be described in logical order. We now introduce the notation essential to this task.

Consider three different estimators which might be called maximum likelihood estimators. Let $f(x;\theta)$, $\theta \in \Theta$, be the frequency function of the random variable X, where the set $S = \{x:f(x;\theta) \neq 0\}$ does not depend on θ. Assume that $f(x;\theta)$ is differentiable with respect to θ for all $\theta \in \Theta$. The log-likelihood function for a random sample

$$X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n \text{ is } L(\theta) = \log \prod_{i=1}^{n} f(x_i;\theta).$$

The equation $\frac{\partial L(\theta)}{\partial \theta} = 0$ will be called the likelihood equation for θ. The following have been referred to as maximum likelihood estimators:

(i) That value of θ, $\hat{\theta} = \theta(X_1, \ldots, X_n)$, which maximizes $L(\theta)$ usually is defined to be the maximum likelihood estimator of θ.

(ii) The root of the likelihood equation, $\tilde{\theta}$, sometimes is called the maximum likelihood estimator of θ.

(iii) Let $\{\theta_\alpha\}$, for α in some index set A, be the set of all roots of the likelihood equation. Then that value θ^*_α of θ which is such that $L(\theta^*_\alpha) \geq L(\theta_\alpha)$ for each $\alpha \in A$ sometimes is called the maximum likelihood estimator for θ.

There is no reason for the estimator $\hat{\theta}$ in (i) to be unique, to exist, or to be in Θ. The same can be said for $\tilde{\theta}$ and for θ^*_α.

We shall employ the definitions offered by Kulldorff (1957) and by Rao (1965). If $L(\hat{\theta}) \geq L(\theta)$ for every $\theta \in \Theta$, we shall call $\hat{\theta}$ a maximum likelihood estimator of θ in the strict sense, SMLE. If $\tilde{\theta}$ is a root of the likelihood equation which effectively depends on the sample $(X_1, \ldots, X_n)$, then we shall call $\tilde{\theta}$ a maximum likelihood estimator of θ in the loose sense, LMLE. If θ^* is a root of the likelihood equation and if $L(\theta^*)$ is greater than or equal to $L(\theta)$ at any other root of the likelihood equation, then we shall call θ^* a restricted maximum likelihood estimator of θ, RMLE. Obviously, none of these estimators need exist or be unique and they certainly need not be equal. We adopt the convention that none of these estimators will be called unique unless it is in Θ. In many cases it is necessary to

refer to a sequence of estimators θ_1, θ_2, ... for which the probability that θ_n exists is $p_n < 1$. Kulldorff (1957) has shown that if $p_n \to 1$ as $n \to \infty$, then one can consider the asymptotic properties of the estimator θ_n by ignoring the fact that the sequence θ_1, θ_2, ... is a sequence of estimators, some of which might not exist.

We adopt the following definition. If X is a random variable with frequency function $f(x;\underline{\theta})$, where $\underline{\theta}$ is a $p \times 1$ vector of parameters, then the sequence $\{\hat{\underline{\theta}}_n\}$ will be said to be asymptotically efficient for $\underline{\theta}_o$, the true value of $\underline{\theta}$, if $\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}_o)$ has as its limiting distribution the p-variate normal distribution with mean vector $\underline{0}$ and covariance matrix $I^{-1}(\underline{\theta}_o)$, where the ij^{th} element of $I(\underline{\theta}_o)$ is

$$\left(E_X\left[\frac{\partial \log f(X;\underline{\theta}_o)}{\partial \theta_i}\cdot\frac{\partial \log f(X;\underline{\theta}_o)}{\partial \theta_j}\right]\right) .$$

Kulldorff (1957) has referred to this property as asymptotic efficiency in the strict sense. This definition will be needed in the discussion which follows.

Cramér (1946 pp. 500-504) has proved a theorem which states that if certain conditions are satisfied by the frequency function $f(x;\theta)$, $\theta \in \Theta$, then some root of the likelihood equation is asymptotically efficient for θ_o, the true value of θ. The conditions follow.

1. $\frac{\partial \log f(x;\theta)}{\partial \theta}$ exists for every $\theta \in \Theta$ and for almost all x.

2. $\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}$ and $\frac{\partial^3 \log f(x;\theta)}{\partial \theta^3}$ exist for every $\theta \in \Theta$ and for almost all x.

3. $\int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial \theta} dx = 0$ and $\int_{-\infty}^{\infty} \frac{\partial^2 f(x;\theta)}{\partial \theta^2} dx = 0$ for $\theta \in \Theta$.

4. $-\infty < \int_{-\infty}^{\infty} \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta) dx < 0$ for every $\theta \in \Theta$.

5. There exists a function of x alone, H(x), such that for every $\theta \in \Theta$

$$\left|\frac{\partial^3 \log f(x;\theta)}{\partial \theta^3}\right| < H(x) \quad \text{and}$$

$$\int_{-\infty}^{\infty} H(x)\, f(x;\theta_0)\, dx < \infty .$$

It is condition five which is not satisfied in the case of grouped data from a normal distribution with either the mean or variance unknown or in the case of grouped data from other distributions; see Kulldorff (1961). Kulldorff (1957) has obtained a slight generalization of Cramér's theorem. Kulldorff noted that condition five is not satisfied in many cases in which the maximum likelihood estimator is known to be asymptotically efficient. One of those cases is the

maximum likelihood estimator for the variance σ^2 of the normal distribution when the mean μ is known. The expression

$$\frac{\partial^3 \log f(x;\mu,\sigma^2)}{\partial(\sigma^2)^3} = -\frac{1}{\sigma^6} + \frac{3(x-\mu)^2}{\sigma^8}$$

obviously is not bounded by a function of x alone since it is not bounded as $\sigma^2 \to 0$.

Kulldorff proved that if all of Cramér's conditions are satisfied except number five, and if there exists a positive function $g(\theta)$ which is twice differentiable for every $\theta \in \Theta$, and if there exists a function of x alone, say $H(x)$, such that for every $\theta \in \Theta$,

$$\left|\frac{\partial^2}{\partial\theta^2}\left[g(\theta)\,\frac{\partial \log f(x;\theta)}{\partial\theta}\right]\right| < H(x) \text{ and } \int_{-\infty}^{\infty} H(x)\, f(x;\theta_0)dx < \infty ,$$

then the conclusions of Cramér's theorem remain valid. Kulldorff also developed another condition to replace condition five which he found to be more useful in some cases; see Kulldorff (1958a).

Kulldorff (1958a, 1958b) used these extensions of Cramér's theorem to prove that in the case of grouped data from a normal distribution with one parameter unknown, the LMLE for the mean when the variance is known and the LMLE for the variance when the mean is known are asymptotically efficient. He did not consider the case where both parameters are unknown, nor would it seem that Cramér's

theorem can be extended easily to the two-parameter case.

It must be made clear what conclusions can be drawn from Cramér's theorem, or Kulldorff's extension. One can conclude only that under the regularity conditions there is some solution of the likelihood equation with probability approaching one, and that some solution possesses the property of being asymptotically efficient. We cannot conclude that a sequence of solutions, each of which maximizes the likelihood function, is a sequence possessing this property. This leads one to inquire whether or not there might be more than one consistent solution.

This question leads to the work done by Huzurbazar (1947-1949). He showed that under Cramér's regularity conditions there can be only one consistent solution of the likelihood equation and that the likelihood function has a relative maximum at that solution with probability one as the sample size increases without bound. The solutions of the maximum likelihood equation frequently can be written in closed form by expressing the solutions as zeros of a function of the parameter and the data, but in many cases no explicit mathematical expressions can be found for these solutions. If one could obtain explicit solutions of the likelihood equation, he could verify which one was the consistent solution. If explicit solutions cannot be found, one could choose the root nearest any consistent estimator.

Huzurbazar (1947-1949) and (1955) stated that this approach was considered by R.A. Fisher (1925) and that the approach is called "successive approximations to the maximum likelihood solution starting with an inefficient estimator." This concept will be utilized in Chapter II to obtain starting values for iterative procedures. Huzurbazar also considered distributions of random variables whose ranges depend on the parameter of the distribution.

If we combine the results of Cramér's theorem and the conclusions obtained by Huzurbazar (1947-1949), we still cannot insure that a solution of the likelihood equation, at which the likelihood function is maximum, is necessarily the consistent solution.

This dilemma was resolved by Wald (1949), although he was not particularly concerned with the solutions of the likelihood equations. He proved that under the following conditions the $p \times 1$ vector $\hat{\underline{\theta}}$, a SMLE, if it exists, is consistent for $\underline{\theta}_0$, the true value of $\underline{\theta} \in \Theta$. The symbols $|\underline{\theta}_1 - \underline{\theta}_2|$ represent the Euclidean distance from $\underline{\theta}_1$ to $\underline{\theta}_2$. The conditions are:

1. The distribution function $F(x;\underline{\theta})$ is either discrete for all $\underline{\theta} \in \Theta$ or $F(x;\underline{\theta})$ is absolutely continuous for all $\underline{\theta} \in \Theta$.

2. For sufficiently small $\rho > 0$ and for sufficiently large $r > 0$

$$\int_{-\infty}^{\infty} \log f^*(x;\underline{\theta},\rho)\,dF(x;\underline{\theta}_o) \text{ and } \int_{-\infty}^{\infty} \log \varphi^*(x,r)\,dF(x;\underline{\theta}_o)$$

are finite, where

$$f^*(x;\underline{\theta},\rho) = \max\left\{\sup_{\overline{\underline{\theta}}:|\underline{\theta}-\overline{\underline{\theta}}| \le \rho} f(x;\overline{\underline{\theta}}),\ 1\right\},\ \underline{\theta} \in \Theta \text{ and}$$

$$\varphi^*(x,r) = \max\left\{\sup_{|\underline{\theta}-\underline{0}| > r} f(x;\underline{\theta}),\ 1\right\}.$$

3. If $\lim_{i\to\infty} \underline{\theta}_i = \underline{\theta}$, then $\lim_{i\to\infty} f(x;\underline{\theta}_i) = f(x;\underline{\theta})$ for all x except perhaps on a set which may depend on $\underline{\theta}$ (but not on the sequence $\{\underline{\theta}_i\}$) and whose measure is zero according to the probability distribution function corresponding to the true parameter point $\underline{\theta}_o$.

4. If $\underline{\theta}^*$ is not equal to $\underline{\theta}_o$, then $F(x;\underline{\theta}^*) \neq F(x;\underline{\theta}_o)$ for at least one value of x.

5. If $\lim_{i\to\infty} |\underline{\theta}_i - \underline{0}| = \infty$, then $\lim_{i\to\infty} f(x;\underline{\theta}_i) = 0$ for any x except perhaps on a fixed set (independent of the sequence $\{\underline{\theta}_i\}$) whose measure is zero according to the probability distribution function corresponding to $\underline{\theta}_o$.

6. $\int_{-\infty}^{\infty} |\log f(x;\underline{\theta}_o)|\,dF(x;\underline{\theta}_o) < \infty$.

7. The parameter space Θ is a closed subset of the p-dimensional Cartesian space.

8. The function $\sup_{\underline{\theta}^*:|\underline{\theta}-\underline{\theta}^*| \leq \rho} f(x;\underline{\theta}^*)$ is a Lebesgue-measurable* function of x for any $\underline{\theta} \in \Theta$ and for any $\rho > 0$.

At the end of the paper he remarked that if Θ is any abstract space, then assumptions 3, 5, and 7 can be replaced by:

9. It is possible to introduce a distance $\delta(\underline{\theta}_1,\underline{\theta}_2)$ in Θ such that

(i) The distance $\delta(\underline{\theta}_1,\underline{\theta}_2)$ makes Θ a metric space,

(ii) Condition 3 holds with the δ metric,

(iii) If $\underline{\theta}^*$ is any fixed point in Θ and $\lim_{i\to\infty} \delta(\underline{\theta}_i,\underline{\theta}^*) = \infty$, then $\lim_{i\to\infty} f(x;\underline{\theta}_i) = 0$ for any x, and

(iv) Any closed and bounded subset of Θ is compact.

These conditions are somewhat involved, but the phrase "mild regularity conditions" is not a very informative

---

* Wald (1949) did not specify that this function should be Lebesgue-measurable, but when condition 8 was used the property of Lebesgue-measurability was required.

substitute for the conditions.

We note that only the consistency of $\hat{\underline{\theta}}$, the SMLE, is proven. The limiting distribution of $\sqrt{n}(\hat{\underline{\theta}} - \underline{\theta}_o)$ is not determined. Neither is it shown that a SMLE is a root of the likelihood equations for large sample sizes. However if, in addition, one assumes Cramér's regularity conditions and uses the results of Huzurbazar concerning the uniqueness of a consistent solution, one sees that a solution of the likelihood equation which makes the likelihood function a maximum is necessarily a consistent solution or possibly the consistent solution. The argument is clear and will not require a formal theorem.

Rao (1957, 1965)* considered the special case of a multinomial distribution with the k cell probabilities $P_1(\underline{\theta}), P_2(\underline{\theta}), \ldots, P_k(\underline{\theta})$, where $\underline{\theta}$ is a $p \times 1$ vector of parameters such that $\underline{\theta} \in \Theta$. He proved that if the four conditions to be stated hold, then with probability tending to unity as $n \to \infty$, the SMLE, $\hat{\underline{\theta}}$, is a root of the likelihood equations, and in particular it is the RMLE, $\underline{\theta}_n^*$. Further, if $\underline{\theta}_o$ is the true value of $\underline{\theta}$, then $\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}_o)$ and $\sqrt{n}(\underline{\theta}_n^* - \underline{\theta}_o)$ both have as their limiting distribution the p-variate normal

---

* The four publications related to the topic of estimating parameters of multinomial distributions (1957, 1958, 1961a, 1961b) are summarized in Rao (1965).

distribution with mean vector $\underline{0}$ and covariance matrix nI_o^{-1}, where I_o is the information matrix for a sample of size n. The conditions are:

1. $P_i(\underline{\theta}) \neq 0 \quad i = 1, \ldots, k \ , \ \underline{\theta} \in \Theta.$

2. $\frac{\partial P_i(\underline{\theta})}{\partial \theta_j}$ is continuous for $\underline{\theta} \in \Theta$

   $i = 1, \ldots, k \quad \text{and} \quad j = 1, \ldots, p \ .$

3. The information matrix is non-singular.

4. For every $\delta > 0$ there exists an $\epsilon > 0$ such that

$$\inf_{|\underline{\theta}-\underline{\theta}_o| \geq \delta} \sum_{i=1}^{k} P_i(\underline{\theta}_o) \log \left(\frac{P_i(\underline{\theta}_o)}{P_i(\underline{\theta})}\right) \geq \epsilon \ , \ \underline{\theta}_o \text{ in } \Theta.$$

This theorem will be the vehicle for determining the asymptotic properties of the RMLE for the parameters of the normal distribution having unknown mean and variance when grouped data only are available. It also will be used to determine the properties of the RMLE for the five parameters of the bivariate normal distribution when grouped data only are available.

1.3 Discussion of the Problem

In Chapter II the problem of estimating the parameters of a univariate normal density function using grouped data samples is considered. The method of maximum likelihood is employed. Kulldorff (1958a, 1958b) suggested that the method of scoring might be used to solve the likelihood equations, and Hughes (1962) gave examples in which Hughes' method of solution seemed to be adequate. Neither author established that the solution of the likelihood equations which are considered in Chapter II always can be obtained by either method of solution. Therefore the solution of the likelihood equations is studied.

An iterative method of solution is proposed which is similar to the method developed by Hughes (1962). The iterates are shown to converge to the unique solution of the likelihood equation when only one parameter is unknown, and the method is modified to reduce the number of iterations required. An iterative method of solution is proposed for the case in which both parameters are unknown. The iterates are shown to have several favorable properties. The author cannot establish sufficient conditions which would insure the existence of a unique, simultaneous solution of the likelihood equations. Therefore numerical results are

presented which define a region in the parameter space in which there can be only one relative maximum of the likelihood function, and a procedure is developed which can be employed to locate multiple roots if they exist. Easily computed starting values for the iterative procedures, which themselves are consistent estimators, are obtained.

In Chapter II no effort is made to extend the development to other types of incomplete data.

In Chapter III the problem of describing some of the asymptotic properties of maximum likelihood estimators obtained from grouped data samples is studied. The principal objective is to establish the asymptotic efficiency of the restricted maximum likelihood estimators for the mean and variance of a normal density function when the data are grouped and both parameters are unknown. This is an extension of the work of Kulldorff (1958a, 1958b). His results concerning the asymptotic properties of the maximum likelihood estimators when only one parameter is unknown can be shown to be special cases of the theorems in Chapter III. Theorems are developed which are more general than those necessary to establish asymptotic properties in the case of the normal distribution. All of the theorems developed in Chapter III are based on the theorem in section 1.2.2 which was proved by Rao (1965).

No effort is made to establish asymptotic properties of

maximum likelihood estimators in any case other than that of grouped data.

The iterative method of solving likelihood equations which was developed by Hughes (1962) is studied in the Appendix. Hughes established sufficient conditions which insure that the iterates defined by Hughes' method of solution converge to a solution of the likelihood equations. In the Appendix it is shown that relatively few of the distributions commonly encountered satisfy those conditions, and among those distributions which do satisfy the conditions, most have closed parameter spaces.

CHAPTER II

OBTAINING THE ROOTS OF THE LIKELIHOOD EQUATIONS FOR THE MEAN AND VARIANCE OF A NORMAL DISTRIBUTION FROM GROUPED DATA SAMPLES

2.1 Introduction

In this chapter preliminary theorems are developed which are used to prove that the method of successive approximations is an appropriate numerical method to employ when seeking the root of the likelihood equation for the mean or the variance of a normal distribution when the sample is a grouped data sample. The method of successive approximations is modified in order to reduce the number of iterations and in many cases the convergence is sufficiently rapid for the solution to be obtained by hand. Starting values for the iterative procedure are obtained which are themselves consistent estimators.

Numerical results are presented which provide a region in the parameter space in which there can be at most one joint solution of the likelihood equations for the mean and variance of a normal distribution when the sample is a grouped data sample; and if there is a solution in this region, then it is a point at which the likelihood function

assumes a relative maximum. Numerical results also indicate that one should expect only one joint solution of the likelihood equations provided certain conditions are satisfied by the sample. A method is developed which can be used to locate other solutions if there are any. Comparisons are made among the numerical methods which are considered and examples are provided.

2.2 Preliminary Development

2.2.1 Definitions and Notation

Let X be a random variable having the frequency function $f(x;\underline{\theta})$ and the probability distribution function $F(x;\underline{\theta})$, where $\underline{\theta}$ is a $p \times 1$ vector of parameters taking on values in some subspace Θ of Euclidean p-space. Let $r_1, r_2, \ldots, r_{k-1}$ be a set of k-1 known constants such that $-\infty = r_0 < r_1 < r_2 < \ldots < r_{k-1} < r_k = \infty$ and let $I_i = (r_{i-1}, r_i]$, $i = 1, 2, \ldots, k$. Define $P_i(\underline{\theta}) = \Pr\{X \in I_i;\underline{\theta}\}$ and $c_i = P_i^{-1}(\underline{\theta})$ when $P_i(\underline{\theta}) \neq 0$. Then the frequency function of X given that X is in the interval I_i is $c_i f(x;\underline{\theta})$ when $x \in I_i$ and zero otherwise. We define the functional $E_i\varphi(X)$ to be equal to $\int_{I_i} \varphi(x) c_i f(x;\underline{\theta})dx$ if $F(x;\underline{\theta})$ is an absolutely continuous probability distribution function in I_i. If $F(x;\underline{\theta})$ is not absolutely continuous in I_i, we use the Stieltjes integral to define $E_i\varphi(X) = \int_{I_i} \varphi(x) c_i dF(x;\underline{\theta})$.

In the special case when $\varphi(X) = X$ let $\mu_i = E_i(X)$, and when $\varphi(X) = (X - \mu_i)^2$ let $\sigma_i^2 = E_i(X - \mu_i)^2$. The following definition will be used frequently:

$$V_i(\varphi(X)) = E_i[\varphi(X)]^2 - [E_i\varphi(X)]^2 .$$

There will arise situations in which we wish to consider one or both of the end points of the interval I_i as a variable and in these cases we shall use the following expressions interchangeably:

$$\mu_i = \mu_{(r_{i-1},r_i]} \quad \text{and} \quad \sigma_i^2 = \sigma^2_{(r_{i-1},r_i]} .$$

The function $H(\underline{\theta})$ will denote the joint frequency function of the random variables comprising a sample and will be called the likelihood function for the sample. As usual, $H(\underline{\theta})$ is considered to be a function of $\underline{\theta}$. When the grouped data case is considered it is assumed tacitly that $P_i(\underline{\theta}) \neq 0$, $i = 1, \ldots, k$. Therefore if the grouped data sample is of size n with the cell frequencies $n_1, n_2, \ldots, n_k$, such that $\sum_{i=1}^{k} n_i = n$, then the likelihood function for the sample is

$$H(\underline{\theta}) = n! \prod_{i=1}^{k} \{(n_i!)^{-1} P_i^{n_i}(\underline{\theta})\}. \tag{2.0}$$

Usually it will be more convenient to consider the natural logarithm of $H(\underline{\theta})$. Therefore we define the log-likelihood function as follows:

$$L(\underline{\theta}) = \log H(\underline{\theta}) .$$

For the grouped data case

$$L(\underline{\theta}) = C + \sum_{i=1}^{k} n_i \log P_i(\underline{\theta}),$$

where $C = \log(n!) - \sum_{i=1}^{k} \log(n_i!)$ is observed not to depend on $\underline{\theta}$.

We shall employ the following notation. If $\varphi(\underline{x})$ is a scalar function of the variables $[x_1, x_2, \ldots, x_\nu] = \underline{x}'$, then the vector of partial derivatives

$$\left[\frac{\partial\varphi(\underline{x})}{\partial x_1}, \frac{\partial\varphi(\underline{x})}{\partial x_2}, \ldots, \frac{\partial\varphi(\underline{x})}{\partial x_\nu}\right]$$

will be denoted by $\frac{\partial\varphi(\underline{x})}{\partial\underline{x}'}$ and the transpose of this row vector will be denoted by $\frac{\partial\varphi(\underline{x})}{\partial\underline{x}}$. If $\underline{\varphi}'(\underline{x}) = [\varphi_1(\underline{x}), \varphi_2(\underline{x}), \ldots, \varphi_r(\underline{x})]$ is a row vector of functions of the variables $[x_1, x_2, \ldots, x_\nu] = \underline{x}'$, then the matrix of partial derivatives having as its i^{th} column $\frac{\partial\varphi_i(\underline{x})}{\partial\underline{x}}$, $i = 1, \ldots, r$, will be denoted by $\frac{\partial\underline{\varphi}'(\underline{x})}{\partial\underline{x}}$. If $\varphi(\underline{x})$ is a scalar function of the variables $[x_1, x_2, \ldots, x_\nu] = \underline{x}'$, then the $\nu \times \nu$ matrix having $\frac{\partial^2\varphi(\underline{x})}{\partial x_i \partial x_j}$ as its ij^{th} entry will be denoted by $\frac{\partial^2\varphi(\underline{x})}{\partial\underline{x}\partial\underline{x}'}$ or by $\frac{\partial}{\partial\underline{x}}\left[\frac{\partial\varphi(\underline{x})}{\partial\underline{x}'}\right]$.

2.2.2 Preliminary Theorems

In order to establish some of the properties of the numerical methods which will be used to solve the likelihood equations considered in subsequent sections, we need to verify a property of the univariate normal distribution. Let X be a random variable having the normal density function

$$f(x;\underline{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < X < \infty \qquad \text{and}$$

$$\underline{\theta} \in \Theta \equiv \{(\mu,\sigma^2): -\infty < \mu < \infty \text{ and } 0 < \sigma^2 < \infty\} .$$

The property which is needed is :

$$E_i(X - \mu_i)^2 = \sigma_i^2 \leq \sigma^2 .$$

Bowen (1966) considered the problem of showing that this relation is satisfied by the normal density function. Although he was successful in proving that $\sigma_i^2 \leq \sigma^2$ only when zero is not in I_i, some of the theorems developed in that thesis will be useful in establishing the property.

Let I_i be a representative interval with $r_{i-1} = a$ and $r_i = t$. Assume that a is fixed, $-\infty \leq a$, and that t is a variable greater than a. In Theorem 5, section 1.3 of Bowen (1966) it was shown that if X is a random variable which has an absolutely continuous probability distribution function

$F(x)$, then $\frac{d\sigma^2_{(a,t]}}{dt} = \frac{\varphi'(t)}{\varphi(t)}\left[(t - \mu_{(a,t]})^2 - \sigma^2_{(a,t]}\right]$, where $\varphi(t) = F(t) - F(a)$ and $\varphi'(t) = \frac{d\varphi(t)}{dt}$. In that same section it was shown that for any absolutely continuous probability distribution function

$$\sigma^2_{(a,t]} \leq \max\left[(t - \mu_{(a,t]})^2, (\mu_{(a,t]} - a)^2\right].$$

Using this result, we now state and prove the following lemma.

Lemma 1

If $F(x)$ is an absolutely continuous probability distribution function such that $\frac{dF(x)}{dx} = f(x)$ is a monotone decreasing function of x when $x \in (a,b]$, then $\frac{d\sigma^2_{(a,t]}}{dt} \geq 0$ for each $t \in (a,b]$ such that $F(t) - F(a) = \varphi(t) \neq 0$.

(See Theorem C of the Appendix in Bowen (1966).)

Proof: Since $\frac{d\sigma^2_{(a,t]}}{dt} = \frac{f(t)}{\varphi(t)}\left[(t - \mu_{(a,t]})^2 - \sigma^2_{(a,t]}\right]$

and since $\sigma^2_{(a,t]} \leq \max\left[(t - \mu_{(a,t]})^2, (\mu_{(a,t]} - a)^2\right]$ it follows that if $(t - \mu_{(a,t]})^2 \geq (\mu_{(a,t]} - a)^2$, then $\frac{d\sigma^2_{(a,t]}}{dt} \geq 0$. Since $f(x)$ is monotone decreasing on $(a,b]$,

$$(t - \mu_{(a,t]}) \geq (\mu_{(a,t]} - a).$$

Hence $\frac{d\sigma^2_{(a,t]}}{dt} \geq 0$ when $t \leq b$.

In order to show that $\sigma^2_{(a,b]} \leq \sigma^2$ for an arbitrary normal density function, it will suffice to prove the result for the standard normal density function $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, $-\infty < x < \infty$, in which case it must be shown that $\sigma^2_{(a,b]} \leq 1$ for arbitrary intervals (a,b].

Clark (1957) showed for this case that

$$\mu_{(a,t]} = \frac{f(a) - f(t)}{\varphi(t)}$$

and

$$\sigma^2_{(a,t]} = 1 + \frac{af(a) - tf(t)}{\varphi(t)} - \mu^2_{(a,t]} ,$$

where $a < t$ and $\varphi(t) = \int_a^t f(x)dx$. Gordon (1941) and Birnbaum (1942) showed that

$$a < \frac{f(a)}{\varphi(\infty)} , \qquad (2.1)$$

where $\varphi(\infty) = \int_a^\infty f(x)dx$, for all $a \in (-\infty,\infty)$. A much simpler proof of this relation can be exhibited if one considers the function $\varphi(\infty) - \frac{f(a)}{a}$. This function has limit $-\infty$ as a approaches zero from the right and limit 0 as $a \to \infty$. The first derivative of this function with respect to a is $\frac{f(a)}{a^2}$,

which is positive. Thus $a < \frac{f(a)}{\varphi(\infty)}$ for $a > 0$. The inequality is obvious if $a \leq 0$. We now establish an important inequality relating $\sigma^2_{(a,b]}$ and σ^2 for the normal distribution.

Theorem 1

If X is a random variable having the density function $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, $-\infty < x < \infty$, then $\sigma^2_{(a,t]} \leq 1$, the equality holding if and only if $a = -\infty$ and $t = \infty$.

Proof:

Case I. If $a \leq 0 \leq t$, $a < t$, then

$\sigma^2_{(a,t]} = 1 - \mu^2_{(a,t]} + \frac{af(a) - tf(t)}{\varphi(t)} \leq 1$ since both of the terms $-\mu^2_{(a,t]}$ and $\frac{af(a) - tf(t)}{\varphi(t)}$ are non-positive. The only way for both of the terms to be zero is for $a = -t$, making $\mu^2_{(a,t]} = 0$, and for $af(a) = tf(t) = 0$. The latter relation holds only if $a = -\infty$ and $t = \infty$ since $uf(u) \to 0$ only if $u \to -\infty$ or if $u \to \infty$ or if $u = 0$.

Case II. The only other cases are $0 < a < t$ and $a < t < 0$. By symmetry we need consider only one of these cases. Assume that $0 < a < t$. Since $f(x)$ is monotone decreasing on $[0,\infty)$, we use Lemma 1 to establish that

$\sigma^2_{(a,t]} \leq \sigma^2_{(a,\infty)}$. Therefore, if we can show that $\sigma^2_{(a,\infty)} < 1$, then the proof is complete. Since $tf(t) \to 0$ as $t \to \infty$, we have

$$\sigma^2_{(a,\infty)} = 1 - \mu^2_{(a,\infty)} + \frac{af(a)}{\varphi(\infty)}$$

$$= 1 - \left(\frac{f(a)}{\varphi(\infty)}\right)^2 + \frac{af(a)}{\varphi(\infty)}$$

$$= 1 + \frac{f(a)}{\varphi(\infty)}\left(a - \frac{f(a)}{\varphi(\infty)}\right).$$

By inequality (2.1) we know that $a - \frac{f(a)}{\varphi(\infty)} < 0$. Therefore $\sigma^2_{(a,\infty)} < 1$, and it follows that $\sigma^2_{(a,t]} \leq 1$.

From this theorem we conclude that if X has a univariate normal distribution with mean μ and variance σ^2, then $\sigma_i^2 \leq \sigma^2$, the equality holding if and only if $I_i \equiv (-\infty,\infty)$.

2.2.3 Finding the Zeros of a Particular Class of Functions

There are many numerical methods which one might employ to obtain the zeros of a function. We shall consider two methods known as the method of successive approximations and the Newton-Raphson method. A discussion of these and other methods can be found in most numerical analysis texts such as Scarborough (1930). One of the early papers in which the properties of the method of successive approximations were considered is by Ford (1925).

We shall restrict our attention to the class of functions $\Psi(X)$ such that $\Psi(X) = \varphi(X) - X$ for some function $\varphi(X)$. Let $\varphi'(X) = \frac{d\varphi(X)}{dX}$ exist and take on only values greater than zero and less than or equal to δ, where $\delta < 1$, when X is in some interval (a,b).

The method of successive approximations is a procedure which defines a sequence of numbers $X_1, X_2, \ldots$ from the starting value $X_1 \in (a,b)$. The sequence is defined by the recursive formula $X_n = \varphi(X_{n-1})$. The Newton-Raphson procedure defines a similar sequence $Y_1, Y_2, \ldots$ using the recursive relation $Y_n = Y_{n-1} - \frac{\Psi(Y_{n-1})}{\Psi'(Y_{n-1})}$. It is difficult to compare two such methods which have different advantages. In the

application of these procedures which we shall consider, the function $\frac{d\Psi(X)}{dX} = \Psi'(X)$, equivalently $\varphi'(X) - 1$, will be difficult to evaluate. Therefore it will take more effort to make each iteration with the Newton-Raphson method than with the method of successive approximations.

The following theorem will establish some of the properties of the method of successive approximations. The proof of each proposition can be found in Ford (1925).

Theorem 2

If $\Psi(X) = \varphi(X) - X$ and if $\varphi'(X)$ is such that $0 < |\varphi'(X)| \leq \delta < 1$ in the interval (a,b) and if Y is a zero of $\Psi(X)$ in (a,b), then:

1. If $X_1 \in (a,b)$ then the sequence $\{X_n\}$ defined by the relation $X_n = \varphi(X_{n-1})$ converges to Y.
2. The sequence is a monotonic sequence if $\varphi'(X) > 0$ in (a,b).
3. There can be at most one root $Y \in (a,b)$.
4. $|X_{n+1} - Y| < \delta^n |X_1 - Y|$.
5. $|X_n - Y| < \frac{\delta}{1 - \delta} |X_n - X_{n-1}|$.
6. $|X_n - Y| < \delta |X_n - X_{n-1}|$ if $\varphi'(X) < 0$ in (a,b).

We see that conclusions four and five indicate that if δ is relatively small, then the convergence is quite rapid. If

we consider the Newton-Raphson method, $X_n = X_{n-1} - \frac{\Psi(X_{n-1})}{\Psi'(X_{n-1})}$,

equivalently $X_n = X_{n-1} - \frac{X_{n-1}}{1 - \varphi'(X_{n-1})} + \frac{\varphi(X_{n-1})}{1 - \varphi'(X_{n-1})}$, when $|\varphi'(X_{n-1})| \leq \delta$ is relatively small we conclude that the methods are very similar, since $1 + \delta > |1 - \varphi'(X_{n-1})| > 1 - \delta$.

If δ is relatively large and $\varphi'(X) > 0$ in (a,b), then the method of successive approximations will provide a sequence which converges monotonically to Y, but the convergence will tend to be less rapid than when δ is relatively small.

We have noted that $\varphi'(X)$ is a rather complicated function in the cases which we shall consider. However, in these cases it will be known that $\varphi'(X) > 0$ in (a,b). Without computing this function at each iteration how might we take advantage of the monotonicity of the sequence formed by the method of successive approximations? One way to do this is simply to move a little further on each iteration than would be suggested by the successive approximations method, i.e. take $X_n^* = \varphi(X_{n-1}^*) + \lambda\{\varphi(X_{n-1}^*) - X_{n-1}^*\}$, $\lambda \geq 0$. If $\lambda = 0$ the sequence $\{X_n^*\}$ is the monotonic sequence formed by the successive approximations method. If we subtract X_{n-1}^* from each side of this equation, we obtain

$$X_n^* - X_{n-1}^* = (1 + \lambda)\{\varphi(X_{n-1}^*) - X_{n-1}^*\} .$$

This is seen to be the Newton-Raphson method with $\varphi'(X_{n-1})$ taken to be $\frac{\lambda}{1+\lambda}$. Judicious choice of λ would have the effect of compensating for a "slowly" converging monotonic sequence when $\varphi'(X) > 0$ in (a,b).

We shall not undertake a comparison of the relative merits of these numerical methods, but we shall use some of them in subsequent sections of this chapter.

2.2.4 Properties of the Likelihood Function for a Grouped Data Sample from a Normal Distribution

Let us recall the definitions of a strict and a restricted maximum likelihood estimate. If $H(\underline{\theta})$, $\underline{\theta} \in \Theta$, is the likelihood function for the sample, then $\hat{\underline{\theta}}$ is a strict maximum likelihood estimate of $\underline{\theta}$ if $H(\hat{\underline{\theta}}) \geq H(\underline{\theta})$ for all $\underline{\theta} \in \Theta$ and $\underline{\theta}^*$ is a restricted maximum likelihood estimate of $\underline{\theta}$ if $\underline{\theta}^*$ satisfies the likelihood equations and if $H(\underline{\theta}^*) \geq H(\underline{\theta}_\alpha)$ for all $\underline{\theta}_\alpha$ satisfying the likelihood equations. In this section we establish sufficient conditions for the set of strict maximum likelihood estimates to be identical to the set of restricted maximum likelihood estimates for the case of a grouped data sample from a normal distribution.

The likelihood function is considered to be a mathematical function of the variables in the vector $\underline{\theta}$, given as fixed the values of the random variables observed in the sample. The properties of this function, such as the existence of a root of $\frac{\partial H(\underline{\theta})}{\partial \underline{\theta}} = \underline{0}$, are studied apart from any statistical interpretation of the likelihood function.

Before proving the next theorem, several properties of the likelihood function for a grouped data sample from a normal distribution should be examined.

Let us assume that:

1. The random variable X has the frequency function

$$f(x;\underline{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} , \quad -\infty < x < \infty ,$$

where $\underline{\theta}'$ is the vector $[\mu,\sigma^2]$ and

$\underline{\theta} \in \Theta \equiv \{(\mu,\sigma^2): -\infty < \mu < \infty \text{ and } 0 < \sigma^2 < \infty\}$.

2. The cells

$I_1 \equiv (-\infty,r_1]$, $I_2 \equiv (r_1,r_2]$, ..., $I_k \equiv (r_{k-1},\infty]$,

and their corresponding probabilities

$P_i(\underline{\theta}) = Pr\{X \in I_i;\underline{\theta}\}$, are determined by the set of known constants $-\infty < r_1 < r_2 < \ldots < r_{k-1} < \infty$.

3. The sample consists of k cell frequencies

$n_1, n_2, \ldots, n_k$, where $n = \sum_{i=1}^{k} n_i$ is the fixed sample size.

The likelihood function for the sample is

$H(\underline{\theta}) = C \prod_{i=1}^{k} P_i^{n_i}(\underline{\theta})$, where C is a constant with respect to $\underline{\theta}$; see equation (2.0). It is observed that $H(\underline{\theta}) \not\equiv 0$ in Θ and that both first partial derivatives of $P_i(\underline{\theta})$ with respect to the elements in $\underline{\theta}$, i = 1, ..., k, are continuous in Θ. It

follows that both first partial derivatives of $H(\underline{\theta})$ are continuous in Θ.

In the proof of Theorem 3 we shall use the result that if $H(\underline{\theta}) \to 0$ on the boundary of Θ, $\partial\Theta$, then there exists a point $\underline{\theta}^* \in \Theta$ such that $\frac{\partial H(\underline{\theta})}{\partial \underline{\theta}} = \underline{0}$ at $\underline{\theta}^*$ and $H(\underline{\theta}^*) \geq H(\underline{\theta})$ for all $\underline{\theta} \in \Theta$. This result follows from a minor extension of Theorem 1 in Chapter 4 of Widder (1961). In the present context, that theorem states that if the first partial derivatives of $H(\underline{\theta})$ are continuous in the compact set $S \subset \Theta$ and if there exists a point $\underline{\theta}_1 \in S$ such that $H(\underline{\theta}_1) > H(\underline{\theta})$ for all $\underline{\theta}$ on the boundary of S, then there exists a point $\underline{\theta}^* \in S$ such that $\frac{\partial H(\underline{\theta})}{\partial \underline{\theta}} = \underline{0}$ at $\underline{\theta}^*$ and $H(\underline{\theta}^*) \geq H(\underline{\theta})$ for all $\underline{\theta} \in S$. Since $H(\underline{\theta}) \not\equiv 0$ in Θ, there exists a point $\underline{\theta}_1$ in Θ such that $H(\underline{\theta}_1) > \epsilon$ for some $\epsilon > 0$. If $H(\underline{\theta}) \to 0$ on $\partial\Theta$, then there exists a compact set, say
$S_{n_o} = \{(\mu,\sigma^2): -n_o \leq \mu \leq n_o \text{ and } \frac{1}{n_o} \leq \sigma^2 \leq n_o\}$, such that $H(\underline{\theta}) < \epsilon$ for every $\underline{\theta}$ in Θ which is not in S_{n_o}. Using Theorem 1 in Chapter 4 of Widder (1961), it follows that there exists a point $\underline{\theta}^*$ in S_{n_o} such that $\frac{\partial H(\underline{\theta})}{\partial \underline{\theta}} = \underline{0}$ at $\underline{\theta}^*$ and $H(\underline{\theta}^*) \geq H(\underline{\theta})$ for all $\underline{\theta}$ in S_{n_o}. It follows that $H(\underline{\theta}^*) \geq H(\underline{\theta})$ for all $\underline{\theta}$ in Θ.

Using this result we prove the following theorem.

Theorem 3

If the grouped data sample satisfies assumptions 1-3 stated previously, then the likelihood function, $H(\underline{\theta}) = C \prod_{i=1}^{k} P_i^{n_i}(\underline{\theta})$, assumes its absolute maximum value at some solution of the likelihood equations, $\frac{\partial H(\underline{\theta})}{\partial \mu} = 0$ and $\frac{\partial H(\underline{\theta})}{\partial (\sigma^2)} = 0$, if the following condition is satisfied:

(a) The cell frequencies are such that $n_i + n_j \neq n$ if either $j = i + 1$ or $i = 1$ and $j = k$.

Proof: If we show that $H(\underline{\theta}) \to 0$ on $\partial\Theta$, then the application of the extension of Theorem 1 in Chapter 4 of Widder (1961) completes the proof. We now show that $H(\underline{\theta}) \to 0$ on $\partial\Theta$.

Note that condition (a) implies that at least two cell frequencies are greater than zero. Let n_i and n_j be two cell frequencies which are greater than zero, $1 \leq i < j \leq k$. Then $H(\underline{\theta}) = CQP_i^{n_i}(\underline{\theta})P_j^{n_j}(\underline{\theta})$, where C is a constant with respect to $\underline{\theta}$ and $Q = \prod_{\substack{\nu=1 \\ \nu \neq i,j}}^{k} P_\nu^{n_\nu}(\underline{\theta})$. Three cases are considered; $\mu \to \pm\infty$ and $\sigma^2 \to s^2 \in [0,\infty)$, $\sigma^2 \to 0$ and $\mu \to m \in (-\infty,\infty)$, and $\sigma^2 \to \infty$

while $\mu \to m \in [-\infty,\infty]$. These limit points exhaust $\partial\Theta$.

Case 1. Consider $H(\underline{\theta})$ as $\mu \to \infty$. For any finite σ^2 $\lim_{\mu\to\infty} P_k(\underline{\theta}) = 1$. Since $\sum_{i=1}^{k} P_i(\underline{\theta}) = 1$ we conclude that $P_i^{n_i}(\underline{\theta})$ has limit zero since $i < j \leq k$. Therefore $H(\underline{\theta}) \to 0$. As $\mu \to -\infty$ for any finite σ^2, $P_1(\underline{\theta}) \to 1$ and hence $P_j^{n_j}(\underline{\theta})$, and therefore $H(\underline{\theta})$, has limit zero.

Case 2a. Assume that $\sigma^2 \to 0$ and $\mu \to m \in I_{i_o}$ such that $m \neq r_{i_o}$. It follows that $P_{i_o}(\underline{\theta}) \to 1$ and hence either $P_i^{n_i}(\underline{\theta})$ or $P_j^{n_j}(\underline{\theta})$ has limit zero. Therefore $H(\underline{\theta}) \to 0$.

Case 2b. Assume that $\sigma^2 \to 0$ and that $\mu \to r_{i_o}$. Then $P_{i_o}(\underline{\theta}) + P_{i_o+1}(\underline{\theta})$ has limit one. Since $n_i + n_j < n$ if $j = i + 1$, then either $P_i(\underline{\theta})$ or $P_j(\underline{\theta})$ has limit zero or there is some other $P_\nu^{n_\nu}(\underline{\theta})$, $n_\nu > 0$, which has limit zero. Hence $H(\underline{\theta})$ has limit zero.

Case 3. If $\sigma^2 \to \infty$ and $\mu \to m \in [-\infty,\infty]$, then any $P_\nu(\underline{\theta})$ corresponding to an interval of finite length, $I_2, I_3, \ldots, I_{k-1}$, has zero as its limit.

Since $n_i + n_j \neq n$ if $i = 1$ and $j = k$, there is some $P_\nu^{n_\nu}(\underline{\theta})$, $n_\nu > 0$ and $1 < \nu < k$, with limit zero. Therefore $H(\underline{\theta}) \to 0$.

Since $H(\underline{\theta})$ has limit zero on the boundary of Θ, it follows that $H(\underline{\theta})$ assumes its absolute maximum at some solution of the likelihood equations.

The necessary conditions for $H(\underline{\theta})$ to assume its absolute maximum value in Θ will not be discussed here, for these conditions involve excessively intricate arguments. Since the sufficient conditions established in Theorem 3 would be satisfied in all but the most impractical situations, we can use these sufficient conditions in most cases to establish that $H(\underline{\theta})$ assumes its absolute maximum value at some root of the likelihood equations.

We conclude from Theorem 3 that if all of the grouped data do not lie in adjacent cells and if all of the grouped data do not lie only in the two half lines then:

1. A restricted maximum likelihood estimate is also a strict maximum likelihood estimate.
2. If there is a unique solution of the likelihood equations, then the restricted, the strict, and the loose maximum likelihood estimates are unique and equal. (Having defined the loose maximum likelihood estimate to be any solution of the likelihood

equations, we see that if there is a unique solution, then the RMLE and the LMLE are identical.)

It should be noted that the assumption that $f(x;\underline{\theta})$ is the normal frequency function is not used explicitly in the proof of Theorem 3. This assumption is used only to imply that μ and σ^2 are independent location and scale parameters, respectively, and that the $P_i(\underline{\theta})$ have continuous first partial derivatives with respect to μ and with respect to σ^2 in Θ. Therefore, Theorem 3 is applicable to a much larger class of frequency functions than is specified in the hypothesis of the theorem.

In section 2.3 we shall consider three cases, μ unknown and σ^2 known, σ^2 unknown and μ known, and the case where both μ and σ^2 are unknown. If σ^2 is known to be equal to $\sigma_o^2 \in (0,\infty)$, then we write $H(\underline{\theta}) = H(\mu)$ with the understanding that $\sigma^2 = \sigma_o^2 \in (0,\infty)$. Similarly, if μ is known to be equal to $\mu_o \in (-\infty,\infty)$, then we write $H(\underline{\theta}) = H(\sigma^2)$ with the understanding that $\mu = \mu_o \in (-\infty,\infty)$. Accordingly, all of the related functions of $\underline{\theta}$ such as $L(\underline{\theta}) = \log H(\underline{\theta})$ and $P_i(\underline{\theta})$ will have either μ or σ^2 as their only arguments. Before considering these special cases in which one parameter is known, we should determine the conditions for which $H(\mu)$ and $H(\sigma^2)$ assume their absolute maxima at the roots of the

likelihood equations $\frac{dH(\mu)}{d\mu} = 0$ and $\frac{dH(\sigma^2)}{d(\sigma^2)} = 0$, respectively.

The following two corollaries to Theorem 3 will establish the necessary and sufficient conditions for $H(\mu)$ to assume its absolute maximum at the solution of $\frac{dH(\mu)}{d\mu} = 0$ for every given $\sigma_o^2 \in (0,\infty)$, and for $H(\sigma^2)$ to assume its absolute maximum at the solution of $\frac{dH(\sigma^2)}{d(\sigma^2)} = 0$ for every given $\mu_o \in (-\infty,\infty)$. We would stress that we are considering the conditions which will insure that the likelihood functions attain their respective absolute maxima at the unique solution of the respective likelihood equation regardless of the known value of the other parameter.

Corollary 1

The condition that $n_1 \neq n$ and $n_k \neq n$ is implied by condition (a) of Theorem 3 and for every fixed $\sigma_o^2 \in (0,\infty)$, the likelihood function $H(\mu)$ attains its absolute maximum at the unique solution of $\frac{dH(\mu)}{d\mu} = 0$ if and only if $n_1 \neq n$ and $n_k \neq n$.

Proof: Condition (a) of Theorem 3 implies that $n_1 \neq n$ and that $n_k \neq n$. In Theorem 1 of Kulldorff (1958a) it is shown that $H(\mu)$ attains its absolute maximum if and only if $n_1 \neq n$ and $n_k \neq n$. The condition that $n_1 \neq n$ and $n_k \neq n$ does not depend on the value of $\sigma_o^2 \in (0,\infty)$, thus completing the

proof.

A very simple proof that $H(\mu)$ does not attain its absolute maximum at a solution of $\frac{dH(\mu)}{d\mu} = 0$ when $n_1 = n$ or $n_k = n$ can be exhibited as follows. If $n_1 = n$, then $H(\mu) = CP_1^n(\mu)$ and $P_1(\mu)$ is a strictly monotone decreasing function of $\mu \in (-\infty,\infty)$. Therefore, $\frac{dH(\mu)}{d\mu} < 0$ for each $\mu \in (-\infty,\infty)$. If $n_k = n$, then $\frac{dH(\mu)}{d\mu} > 0$ for each $\mu \in (-\infty,\infty)$.

In Corollary 2 we establish that the conditions of Theorem 3 are necessary and sufficient to establish that for every given $\mu_o \in (-\infty,\infty)$ there is a unique maximum of $H(\sigma^2)$ at the unique solution of $\frac{dH(\sigma^2)}{d\sigma^2} = 0$. Kulldorff (1958b) proved that under assumptions 1-3 given previously there is a unique solution of $\frac{dH(\sigma^2)}{d\sigma^2} = 0$ at which $H(\sigma^2)$ attains its absolute maximum for a fixed $\mu_o \in (-\infty,\infty)$ if and only if:

either

(i) $n_1 + n_k < n$ and $n_i > 0$ for some i satisfying $\mu_o < r_{i-1}$ or $\mu_o > r_i$

or

(ii) $n_1 + n_k = n$, $0 < n_1 < n$, and $n_1(r_1 - \mu_o) > n_k(r_{k-1} - \mu_o)$.

Corollary 2

For every fixed $\mu_o \in (-\infty,\infty)$, $H(\sigma^2)$ attains its absolute

maximum at the unique solution of $\frac{dH(\sigma^2)}{d\sigma^2} = 0$ if and only if condition (a) of Theorem 3 is satisfied.

Proof: It must be shown that either condition (i) or condition (ii) is satisfied for every fixed $\mu_o \in (-\infty,\infty)$ if and only if condition (a) of Theorem 3 is satisfied.

Condition (ii) is not satisfied for arbitrary $\mu_o \in (-\infty,\infty)$. Condition (i) holds for every $\mu_o \in (-\infty,\infty)$ if and only if

(i') $n_1 + n_k < n$

and

(ii') there is no n_i such that $(n_i + n_{i+1}) = n$.

Conditions (i') and (ii') are equivalent to condition (a) of Theorem 3.

In section 2.2.5 a relationship which exists between the information matrix associated with the complete data case and the information matrix associated with the grouped data case is established.

2.2.5 Derivatives of the Log-likelihood Function in the General Case of Grouped Data

In this section the derivatives of the log-likelihood function for grouped data samples are considered. These derivatives are related to the derivatives of the frequency function underlying the grouped data sample. These relations among derivatives provide the basis for determining the relationship which exists between the information matrix associated with the complete data case and the information matrix associated with the general grouped data case, provided certain conditions are satisfied by the underlying frequency function. Often it is necessary to determine whether certain information matrices are singular. The theorems in this section will be useful in answering these and other related questions.

We have used $I_i = (r_{i-1}, r_i]$, $i = 1, \ldots, k$ and $-\infty = r_0 < r_1 < \ldots < r_{k-1} < r_k = \infty$, to denote the cells defining a grouped data sample. In this section we wish to consider a large class of frequency functions, some of which can be identically zero on some fixed interval (a,b). Since we wish to avoid prolix discussion of ranges of integration, the following convention will be adopted in this section. If $f(x;\underline{\theta})$ is the frequency function of some random variable X,

where $\underline{\theta}$ is a p x 1 vector of parameters defined in some subspace Θ of Euclidean p-space, then $\ln(f(x;\underline{\theta}))$ will be used to denote that function which is equal to the natural logarithm of $f(x;\underline{\theta})$, $\log(f(x;\underline{\theta}))$, when $f(x;\underline{\theta}) > 0$, and zero otherwise. This convention allows one to use $\frac{\partial f(x;\underline{\theta})}{\partial \underline{\theta}}$ and $f(x;\underline{\theta})\left[\frac{\partial \ln(f(x;\underline{\theta}))}{\partial \underline{\theta}}\right]$ interchangeably in the arguments which follow. If W is some function which is positive, then log(W) or log W will be used to denote the natural logarithm of W.

The following assumptions define the class of underlying frequency functions which are to be considered:

(i) The random variable X has the probability distribution function $F(x;\underline{\theta})$ which is either absolutely continuous or discrete, where $\underline{\theta}$ is a p x 1 vector of functionally independent parameters defined in some subspace Θ of Euclidean p-space. Let $f(x;\underline{\theta})$ be the frequency function defined by $F(x;\underline{\theta})$.

(ii) The subset of the x axis for which $f(x;\underline{\theta}) = 0$ does not depend on $\underline{\theta}$.

(iii) If $F(x;\underline{\theta})$ is absolutely continuous with respect to Lebesgue-measure, then:

$$\text{(a)}\quad \frac{\partial}{\partial \underline{\theta}} \int_{-\infty}^{\infty} f(x;\underline{\theta})\,dx = \int_{-\infty}^{\infty} \frac{\partial f(x;\underline{\theta})}{\partial \underline{\theta}}\,dx$$

$$= E_X\left[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}}\right] = \underline{0}$$

for every $\underline{\theta} \in \Theta$ and

$$(b) \quad \frac{\partial^2}{\partial \underline{\theta} \partial \underline{\theta}'} \int_{-\infty}^{\infty} f(x;\underline{\theta})dx = \int_{-\infty}^{\infty} \frac{\partial^2 f(x;\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}'} dx$$

$$= [0]$$

for every $\underline{\theta} \in \Theta$.

(iv) If $F(x;\underline{\theta})$ is the probability distribution function of the discrete random variable X, then we assume that $f(x;\underline{\theta}) > 0$ for every x in the domain of definition of $f(x;\underline{\theta})$ and that properties (a) and (b) in (iii) hold with the proper interpretation of the integral.

All of the theorems to be developed will apply to discrete random variables. Each theorem will be stated in the context of absolutely continuous probability distribution functions, and only the interpretation of the integral will need modification in order for the theorems to remain valid if the probability distribution function is that of a discrete random variable satisfying conditions (i), (ii), and (iv).

We assume that a grouped data sample of size n, from a distribution in the class of probability distribution functions defined by (i) - (iv) preceding, has the following properties:

(v) The cells $I_i \equiv (r_{i-1}, r_i]$, $i = 1, \ldots, k$, are determined by the known constants r_i, $i = 1, \ldots, k-1$, where $-\infty = r_0 < r_1 < r_2 < \ldots < r_{k-1} < r_k = \infty$.

(vi) The underlying distribution function having the properties (i) - (iv) defines the set of functions $P_i(\underline{\theta}) = \Pr\{X \in I_i; \underline{\theta}\} \neq 0$, $i = 1, \ldots, k$.

(vii) The random sample of size n from the underlying distribution gives rise to the cell frequencies $n_1, n_2, \ldots, n_k$, where $\sum_{i=1}^{k} n_i = n$.

The definitions which follow are used often in the sequel when consideration is given to the derivatives of the log-likelihood functions for grouped data samples satisfying conditions (i) - (vii) stated previously. Let these conditions be satisfied and let M be an r x s matrix of scalar functions of X, say $m_{\alpha\beta}(X)$ where $\alpha = 1, \ldots, r$ and $\beta = 1, \ldots, s$. We define

$$E_i m_{\alpha\beta}(X) = P_i^{-1}(\underline{\theta}) \int_{I_i} m_{\alpha\beta}(x) f(x;\underline{\theta}) dx \ , \ i = 1, \ldots, k \ .$$

A matrix of such expected values, $\left[E_i m_{\alpha\beta}(X)\right]$, will be denoted by $E_i[m_{\alpha\beta}(X)]$ or by $E_i M$. If $\underline{\Psi}(X)$ is a p x 1 vector of scalar functions of X, then

$$V_i[\underline{\Psi}(X)] = E_i[\underline{\Psi}(X)][\underline{\Psi}(X)]' - E_i[\underline{\Psi}(X)]E_i[\underline{\Psi}(X)]',$$

where $[\underline{\Psi}(X)]'$ denotes the transpose of $\underline{\Psi}(X)$.

In some cases it will be desired to perform an operation on a function of X and $\underline{\theta}$ and then to evaluate the result at the true value of $\underline{\theta}$. We designate the true value of $\underline{\theta}$ by $\underline{\theta}_o$. If $\underline{\theta}_o$ appears in the argument of a function of X and $\underline{\theta}$ which has some operator immediately preceding it, we shall mean that the operation should be performed and that the result be evaluated at $\underline{\theta}_o$.

If a random sample, $X_1, \ldots, X_n$, is drawn from a distribution satisfying conditions (i) - (iv) given previously, then the likelihood function for this complete data sample is $H_c(\underline{\theta}) = \prod_{j=1}^{n} f(X_j;\underline{\theta})$ and the log-likelihood function is $L_c(\underline{\theta}) = \sum_{j=1}^{n} \ln(f(X_j;\underline{\theta}))$ and the information matrix is $I_c = E_{\underline{X}}\left[\frac{\partial L_c(\underline{\theta}_o)}{\partial \underline{\theta}} \frac{\partial L_c(\underline{\theta}_o)}{\partial \underline{\theta}'}\right]$. The subscript $\underline{X}$ is affixed to the expectation operator to indicate that the expectation is with respect to the joint distribution of $X_1, \ldots, X_n$. If a random sample of size n is drawn from a distribution satisfying conditions (i) - (iv) and if this sample gives rise to the grouped data sample, $n_1, n_2, \ldots, n_k$, satisfying

(v) - (vii), then the likelihood function for this grouped data sample is $H_g(\underline{\theta}) = C \prod_{i=1}^{k} P_i^{n_i}(\underline{\theta})$, where C is a constant with respect to $\underline{\theta}$; see equation (2.0). The log-likelihood function is $L_g(\underline{\theta}) = \log H_g(\underline{\theta})$ and the information matrix is $I_g = E_{\underline{X}}\left[\frac{\partial L_g(\underline{\theta}_o)}{\partial \underline{\theta}} \frac{\partial L_g(\underline{\theta}_o)}{\partial \underline{\theta}'}\right]$, where the subscript $\underline{X}$ affixed to the expectation operator indicates that the expectation is with respect to the joint distribution of $n_1, n_2, \ldots, n_k$, which, of course, is multinomial. By the definition of $P_i(\underline{\theta})$ we know that $E_{\underline{X}}(n_i)\big|_{\underline{\theta}_o} = nP_i(\underline{\theta}_o)$, $i = 1, \ldots, k$.

It is known that the information matrix, I_c, associated with a complete data sample of size n can be written

$$I_c = -E_{\underline{X}}\left[\frac{\partial^2 L_c(\underline{\theta}_o)}{\partial \underline{\theta} \partial \underline{\theta}'}\right] \tag{2.2}$$

if conditions (i) - (iv) are satisfied; see sections 17.14 and 18.24 in Kendall and Stuart (1961) for proof. The same argument also establishes that

$$I_g = -E_{\underline{X}}\left[\frac{\partial^2 L_g(\underline{\theta}_o)}{\partial \underline{\theta} \partial \underline{\theta}'}\right] . \tag{2.3}$$

In Theorem 4 a general relationship is established which simplifies the differentiation of the log-likelihood function

for a grouped data sample.

Theorem 4

If a grouped data sample satisfies conditions (i) - (vii),

then $\frac{\partial L_g(\underline{\theta})}{\partial \underline{\theta}} = \sum_{i=1}^{k} n_i E_i\left[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}}\right]$.

Proof: $$\frac{\partial L_g(\underline{\theta})}{\partial \underline{\theta}} = \sum_{i=1}^{k} n_i\left[\frac{\partial \log(P_i(\underline{\theta}))}{\partial \underline{\theta}}\right]$$

$$= \sum_{i=1}^{k} n_i P_i^{-1}(\underline{\theta})\left[\int_{I_i} \frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}} f(X;\underline{\theta})dX\right]$$

$$= \sum_{i=1}^{k} n_i E_i\left[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}}\right] ,$$

thus completing the proof.

We note that

$$E_{\underline{X}}\left[\frac{1}{n}\left[\frac{\partial L_g(\underline{\theta}_o)}{\partial \underline{\theta}}\right]\right] = \sum_{i=1}^{k}\left(E_{\underline{X}}\left(\frac{n_i}{n}\right)\Big|_{\underline{\theta}_o}\right)E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}}\right]$$

$$= \sum_{i=1}^{k} P_i(\underline{\theta}_o)E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}}\right]$$

$$= \sum_{i=1}^{k} \int_{I_i}\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}}\right]f(X;\underline{\theta}_o)dX$$

$$= E_{\underline{X}}\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}}\right] = \underline{0} .$$

The last equality follows from assumption (a) in (iii) given previously. Therefore:

$$E_{\underline{X}}\left[\frac{\partial L_g(\underline{\theta}_o)}{\partial \underline{\theta}}\right] = \underline{0} . \qquad (2.4)$$

In Theorem 5 the matrix of second partial derivatives of the log-likelihood function for a grouped data sample is expressed in a form which will be useful in the proof of Theorem 6.

Theorem 5

If a grouped data sample satisfies conditions (i) - (vii), then

$$\frac{\partial^2 L_g(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}'} = \sum_{i=1}^{k} n_i V_i\left[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}}\right] + \sum_{i=1}^{k} n_i E_i\left[\frac{\partial^2 \ln(f(X;\underline{\theta}))}{\partial \underline{\theta} \partial \underline{\theta}'}\right].$$

Proof: Using Theorem 4 and recalling that $E_i\varphi(X) = P_i^{-1}(\underline{\theta})\int_{I_i} \varphi(X) f(X;\underline{\theta})dX$ is a function of $\underline{\theta}$, we have:

$$\frac{\partial^2 L_g(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}'} = \frac{\partial}{\partial \underline{\theta}}\left[\frac{\partial L_g(\underline{\theta})}{\partial \underline{\theta}'}\right]$$

$$= \frac{\partial}{\partial \underline{\theta}}\left[\sum_{i=1}^{k} n_i E_i\left[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}'}\right]\right]$$

$$= \sum_{i=1}^{k} n_i \Big(P_i^{-2}(\underline{\theta})\Big(P_i^2(\underline{\theta}) E_i\left[\frac{\partial^2 \ln(f(X;\underline{\theta}))}{\partial \underline{\theta} \partial \underline{\theta}'}\right] +$$

$$P_i^2(\underline{\theta}) E_i\left[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}}\right]\left[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}'}\right]\Big) -$$

$$E_i\left[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}}\right] E_i\left[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}'}\right]\Big)$$

$$= \sum_{i=1}^{k} n_i \Big(E_i\Big[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}}\Big]\Big[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}'}\Big] -$$

$$E_i\Big[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}}\Big] E_i\Big[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}'}\Big]\Big) +$$

$$\sum_{i=1}^{k}{}' n_i E_i\Big[\frac{\partial^2 \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}\partial \underline{\theta}'}\Big] .$$

Therefore,

$$\frac{\partial^2 L_g(\underline{\theta})}{\partial \underline{\theta}\partial \underline{\theta}'} = \sum_{i=1}^{k} n_i V_i\Big[\frac{\partial \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}}\Big] + \sum_{i=1}^{k} n_i E_i\Big[\frac{\partial^2 \ln(f(X;\underline{\theta}))}{\partial \underline{\theta}\partial \underline{\theta}'}\Big] ,$$

thus completing the proof.

We note that

$$nE_{\underline{X}} \sum_{i=1}^{k} \frac{n_i}{n} E_i\Big[\frac{\partial^2 \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}\partial \underline{\theta}'}\Big] = n \sum_{i=1}^{k} P_i(\underline{\theta}_o) E_i\Big[\frac{\partial^2 \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}\partial \underline{\theta}'}\Big]$$

$$= E_{\underline{X}}\Big[\frac{\partial^2 L_c(\underline{\theta}_o)}{\partial \underline{\theta}\partial \underline{\theta}'}\Big] = -I_c ,$$

by equation (2.2). Therefore,

$$E_{\underline{X}} \sum_{i=1}^{k} n_i E_i\Big[\frac{\partial^2 \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}\partial \underline{\theta}'}\Big] = -I_c . \qquad (2.5)$$

There is a relation between the information matrix for a grouped data sample of size n and that for a complete data sample of size n. That relation is developed for the general case of grouped data samples in the theorem which follows.

Theorem 6

If a grouped data sample satisfies conditions (i) - (vii), then $I_g = I_c - n \sum_{i=1}^{k} P_i(\underline{\theta}_o) V_i \left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}} \right]$.

Proof: By equation (2.3), $-I_g = E_{\underline{X}} \left[\frac{\partial^2 L_g(\underline{\theta}_o)}{\partial \underline{\theta} \partial \underline{\theta}'} \right]$. Using Theorem 5 and equation (2.5) we have

$-I_g = E_{\underline{X}} \sum_{i=1}^{k} n_i V_i \left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}} \right] - I_c$. Therefore,

$I_g = I_c - n \sum_{i=1}^{k} P_i(\underline{\theta}_o) V_i \left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}} \right]$, thus completing the proof.

Note that

$$n \sum_{i=1}^{k} P_i(\underline{\theta}_o) V_i \left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}} \right]$$

$$= n \sum_{i=1}^{k} P_i(\underline{\theta}_o) E_i \left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}} \right] \left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}'} \right] -$$

$$n \sum_{i=1}^{k} P_i(\underline{\theta}_o) E_i \left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}} \right] E_i \left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}'} \right]$$

$$= E_{\underline{X}}\left[\frac{\partial L_c(\underline{\theta}_o)}{\partial \underline{\theta}} \frac{\partial L_c(\underline{\theta}_o)}{\partial \underline{\theta}'}\right] -$$

$$n \sum_{i=1}^{k} P_i(\underline{\theta}_o) E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}}\right] E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}'}\right]$$

$$= I_c - n \sum_{i=1}^{k} P_i(\underline{\theta}_o) E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}}\right] E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}'}\right] .$$

Therefore:

$$n \sum_{i=1}^{k} P_i(\underline{\theta}_o) V_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}}\right]$$

$$= I_c - n \sum_{i=1}^{k} P_i(\underline{\theta}_o) E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}}\right] E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}'}\right]. \quad (2.6)$$

An expression which frequently simplifies the computation of the information matrix for a grouped data sample of size n is established in Theorem 7 which follows.

Theorem 7

If a grouped data sample satisfies conditions (i) - (vii), then

$$I_g = n \sum_{i=1}^{k} P_i(\underline{\theta}_o) E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}}\right] E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}'}\right] .$$

Proof: Using Theorem 6 and equation (2.6)

$$I_g = I_c - \left(I_c - n \sum_{i=1}^{k} P_i(\underline{\theta}_o) E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}}\right] E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}'}\right]\right)$$

$$= n \sum_{i=1}^{k} P_i(\underline{\theta}_o) E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}}\right] E_i\left[\frac{\partial \ln(f(X;\underline{\theta}_o))}{\partial \underline{\theta}'}\right] .$$

In the section which follows, the problem of obtaining the maximum likelihood estimates for the mean and variance of the normal distribution from grouped data samples is considered. Three specific cases are treated: estimating the mean when the variance is known, estimating the variance when the mean is known, and estimating the mean and variance when both are unknown.

2.3 Obtaining the Maximum Likelihood Estimates of the Mean and Variance of a Normal Distribution from a Grouped Data Sample

2.3.1 Introduction and Definitions

Essentially three methods are available for solving the likelihood equations when the mean or the variance of a normal distribution, or both, are unknown and grouped data only are available. The first method is to employ the tables presented by Gjeddebaek (1949) to evaluate the various functions found in the likelihood equations. Excessive interpolation and approximation are required in the use of these tables. Kulldorff (1958a, 1958b, 1961) employed the method of scoring to solve the likelihood equations. The evaluation of the information matrix is very laborious and it would seem that the method of scoring should be considered only when electronic computer facilities are to be used. The third method is the method of successive approximations. Hughes' (1962) procedure utilizes quadrature formulae in the iterations required by the method of successive approximations, and thus would require the facilities of an electronic computer in most cases.

When employing iterative procedures one must obtain a starting value in some region containing the solution being sought. The only known starting values which are relatively simple to obtain for the problem at hand are those simple estimators found by using the midpoints of the intervals as quasi observations in the sense of Gjeddebaek (1957) and others.

In this section the method of successive approximations is modified in order to find the solutions of the likelihood equations for unknown parameters of a normal distribution using a grouped data sample. A method for obtaining consistent, easily computed estimators of the parameters is presented. By using the modified method of successive approximations with a "good" starting value, it is possible to obtain the solutions of the likelihood equations using only standard normal tables such as Table 26.1, "Normal Probability Function and Derivatives", in Abramowitz and Stegun (1964). If there are more than six non-zero cell frequencies, then the work required by this method of solution indicates that electronic computing facilities are more desirable.

The following definitions and notation will be useful in this section and in the remaining chapters.

A sample $n_1, n_2, \ldots, n_k$ will be said to have the

property $G_kN_n(\mu,\sigma^2)$ or to be a $G_kN_n(\mu,\sigma^2)$-sample if the following conditions are satisfied:

(i) The probability density function of the random variable X is

$$f(x;\underline{\theta}) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad -\infty < x < \infty ,$$

where $\underline{\theta}' = [\mu,\sigma^2]$ and

$\underline{\theta} \in \Theta \equiv \{(\mu,\sigma^2): -\infty < \mu < \infty \text{ and } 0 < \sigma^2 < \infty\}$.

(ii) The set of intervals $I_i \equiv (r_{i-1},r_i]$, $i = 1, \ldots, k$, are defined by the known constants $r_1, r_2, \ldots, r_{k-1}$, where $-\infty = r_o < r_1 < r_2 < \ldots < r_{k-1} < r_k = \infty$.

(iii) The joint distribution of the random vector of cell frequencies, $[n_1, n_2, \ldots, n_k]$, is

$$\frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} P_i^{n_i}(\underline{\theta}) \text{ ; where } n_i \geq 0,\ i = 1, \ldots, k ,$$

$\sum_{i=1}^{k} n_i = n$ and $P_i(\underline{\theta}) = \Pr\{X \in I_i;\underline{\theta}\}$, $i = 1, \ldots, k$.

If one of the parameters, μ or σ^2, is known, then that parameter will be omitted in the argument of functions associated with the sample having the property $G_kN_n(\mu,\sigma^2)$.

For example, if σ^2 is known to be equal to σ_o^2, then the log-likelihood function for the sample having the property $G_k N_n(\mu,\sigma_o^2)$ will be written:

$$L(\mu) = C + \sum_{i=1}^{k} n_i \log P_i(\mu),$$

where C is observed not to be a function of μ. The subscript g is omitted from the log-likelihood function in section 2.3. In this section it will be understood that the likelihood functions which are considered are those for grouped data samples from a normal distribution.

2.3.2 Estimating the Mean When the Variance Is Known

Consider a sample with the property $G_k N_n(\mu,\sigma^2)$, where $\sigma_o^2 \in (0,\infty)$ is the known variance. Using Theorem 4 in section 2.2.5 and the relation $\frac{d}{d\mu}\big(\log(f(X;\mu))\big) = \frac{X-\mu}{\sigma_o^2}$ (see condition (i) in the preceding section) it follows that $\sum_{i=1}^{k} n_i E_i\left(\frac{X-\mu}{\sigma_o^2}\right) = 0$ is the likelihood equation for μ. It follows from Theorem 5 in section 2.2.5 that

$$\frac{d^2L(\mu)}{d\mu^2} = \sum_{i=1}^{k} n_i V_i\left(\frac{X-\mu}{\sigma_o^2}\right) - \frac{n}{\sigma_o^2}\,, \tag{2.7}$$

since $\frac{d}{d\mu}\left(\frac{X-\mu}{\sigma_o^2}\right) = -\frac{1}{\sigma_o^2}$. If $E_i(X)$ is designated by μ_i as in section 2.2.1, then the solution of

$$\sum_{i=1}^{k} \frac{n_i}{n}\mu_i - \mu = 0 \tag{2.8}$$

is sought. In order to consider the method of successive approximations we define the following functions of the generic variable m:

$$\Psi(m) = \left[\sum_{i=1}^{k} \frac{n_i}{n}\mu_i - \mu\right]_{|\mu=m}$$

$$\varphi(m) = \left[\sum_{i=1}^{k} \frac{n_i}{n}\mu_i\right]_{|\mu=m}.$$

Using the expression for $\frac{d^2L(\mu)}{d\mu^2}$ in equation (2.7) we find that

$$\psi'(m) = \sum_{i=1}^{k} \frac{n_i}{n} V_i\left(\frac{X-m}{\sigma_o}\right) - 1$$

$$= \sum_{i=1}^{k} \frac{n_i}{n} V_i\left(\frac{X}{\sigma_o}\right) - 1$$

by the invariance of $V_i(z)$ to change in location. Therefore

$$\varphi'(m) = \sum_{i=1}^{k} \frac{n_i}{n} V_i\left(\frac{X}{\sigma_o}\right) , \qquad (2.9)$$

where one should bear in mind that both μ_i and $V_i(X)$ are functions of m.

If Hughes' method is used as outlined in the Appendix, then $\varphi(m)$ is approximated by using some quadrature formula to approximate each μ_i. Each of the intervals I_i is partitioned into q_i sub-intervals of equal length. The j^{th} sub-interval of I_i is designated by $(x_{ij}, x_{i(j+1)}]$. Constants a_{ij} are determined by the particular method of quadrature employed (see Kunz (1957), section 7.12), such that

$\sum_{j=1}^{q_i+1} a_{ij} = r_i - r_{i-1}$. Then

$$\tilde{\varphi}(m) = \sum_{i=1}^{k} \frac{n_i}{n} \left[\sum_{j=1}^{q_i+1} a_{ij}\, x_{ij}\, f(x_{ij};m) \right] \left[\sum_{j=1}^{q_i+1} a_{ij}\, f(x_{ij};m) \right]^{-1}$$

is used to approximate $\varphi(m)$. It has been pointed out in a previous section that the convergence of the sequence defined by Hughes' method cannot be established using Hughes' convergence theorem in the problem at hand.

Let $\hat{\mu}$ be the unique solution of the likelihood equation $\frac{dL(\mu)}{d\mu} = 0$. Then by Theorem 2 it follows that the sequence $\{m_j\}$, defined by $m_j = \varphi(m_{j-1})$, converges to $\hat{\mu}$ for any starting value $m_o \in (-\infty,\infty)$ provided that for any interval $[a,b]$, $-\infty < a < b < \infty$, there exists a δ such that $\delta < 1$ and $\varphi'(m) \leq \delta$ when $m \in [a,b]$, where δ may depend on a and b.

The lemma which follows will be used in the proof of Theorem 8.

Lemma 2

If $[a,b]$ is any interval such that $-\infty < a < b < \infty$, then there exists a $\delta < 1$, which may depend on a and b, such that $\varphi'(m) \leq \delta$ when $m \in [a,b]$.

Proof: Since $\frac{d^2L(\mu)}{d\mu^2}$ is a continuous function of μ (see equation (2.7)), the functions $V_i\left(\frac{X}{\sigma_o}\right)$, $i = 1, \ldots, k$, are continuous functions of μ. Further, $V_i\left(\frac{X}{\sigma_o}\right)$ is a positive, continuous function of m and by Theorem 1 is bounded less than 1 for $m \in [a,b]$, $i = 1, \ldots, k$. Therefore, by Theorem 1 it follows that $\sup_{m \in [a,b]} V_i\left(\frac{X}{\sigma_o}\right) = \delta_i$, $i = 1, \ldots, k$, exist

and that each $\delta_i < 1$. Let the maximum of the δ_i be δ. Then

$$\varphi'(m) = \sum_{i=1}^{k} \frac{n_i}{n} V_i\left(\frac{X}{\sigma_o}\right) \leq \sum_{i=1}^{k} \frac{n_i}{n} \delta_i \leq \delta < 1 ,$$

when $m \in [a,b]$.

Recall that in Corollary 1 of section 2.2.4 it was shown that there exists a unique solution of the likelihood equation for μ or, equivalently, of $\Psi(m) = 0$ (see equation (2.8)) if and only if $n_1 \neq n$ and $n_k \neq n$. Using Lemma 2 we prove the following theorem.

Theorem 8

If $n_1 \neq n$ and $n_k \neq n$ in a sample with the property $G_k N_n(\mu,\sigma_o^2)$, then the sequence $\{m_j\}$ defined by the recursive relation $m_j = \varphi(m_{j-1})$, where m_o is any starting value in $(-\infty,\infty)$, converges monotonically to the unique root of the likelihood equation and the likelihood function assumes its absolute maximum value at this root.

Proof: It follows from Corollary 1 in section 2.2.4 that there is a unique solution of the likelihood equation and that the likelihood function attains its absolute maximum at the solution of the likelihood equation. It follows from Lemma 2 that for any interval $[a,b]$, $-\infty < a < b < \infty$, there exists a δ, depending on a and b, such that $\delta < 1$ and

$\varphi'(m) \leq \delta$ when $m \in [a,b]$. Therefore the conditions of Theorem 2 are satisfied. Since $\varphi'(m) > 0$, it follows from conclusion 2 of Theorem 2 that the sequence $\{m_j\}$ is a monotonic sequence converging to the unique solution of the likelihood equation for μ.

Now that it is known that the sequence $\{m_j\}$, defined by the method of successive approximations, is a monotonic sequence which converges to $\hat{\mu}$, one should be able to reduce the number of iterations required if he uses the sequence $\{m_j^*\}$ defined by $m_j^* = \varphi(m_{j-1}^*) + \lambda(\varphi(m_{j-1}^*) - m_{j-1}^*)$, for some $\lambda > 0$. Applying Theorem 2 to the function

$$\begin{aligned}(1+\lambda)\Psi(m) &= [(1+\lambda)\varphi(m) - \lambda m] - m \\ &= [\varphi(m) + \lambda(\varphi(m) - m)] - m ,\end{aligned}$$

we have that the sequence $\{m_j^*\}$ converges to $\hat{\mu}$ if

$$|(1+\lambda)\varphi'(m) - \lambda| \leq \delta_1 < 1 .$$

Since $0 < \varphi'(m) \leq \delta < 1$ for $m \in [a,b]$, it follows that $\{m_j^*\}$ converges to $\hat{\mu}$, provided $n_1 \neq n$ and $n_k \neq n$, for any λ such that $0 \leq \lambda \leq \delta_2 < 1$. Obviously there is no optimal selection of λ for all situations. In the discussion following Theorem 2 in section 2.2.3 it was found that the sequence $\{m_j^*\}$ is the same sequence which would be obtained if one used $\frac{\lambda}{1+\lambda}$ in place of $\varphi'(m_{j-1}^*)$ in the Newton-Raphson method. In

the introduction the method of scoring was shown to be the Newton-Raphson method with, in this case, $1 - \varphi'(m_{j-1})$ approximated by

$$E_{\underline{X}}(1 - \varphi'(m_{j-1})).$$

Tables 3 and 4 presented by Kulldorff (1958a) indicate that

$$E_{\underline{X}}(1 - \varphi'(m_{j-1}))$$

is roughly 0.9 when m_{j-1} is near $\hat{\mu}$. Therefore if λ is chosen such that $\frac{\lambda}{1+\lambda} \approx 1 - 0.9$, i.e. $\lambda \approx 0.1$, the number of iterations required will be reduced in many cases. The method of obtaining the sequence $\{m_j^*\}$, such that $m_j^* = \varphi(m_{j-1}^*) + \frac{1}{10}(\varphi(m_{j-1}^*) - m_{j-1}^*)$, will be called the modified method of successive approximations. After a procedure for obtaining m_o is developed, an example will be given in which the execution of the iterations is discussed.

We now consider the problem of obtaining a starting value m_o. The starting value proposed in this section has two important properties; it is a consistent estimator of the true mean, μ_o, and it can be obtained easily.

Consider a $G_2N_n(\mu,\sigma_o^2)$-sample defined by I_1, I_2, n_1, and n_2, where $n_1 \neq 0$ and $n_2 \neq 0$. Using equation (2.8) we see

that the likelihood equation is

$$\frac{n_1}{n} P_1^{-1}(\mu)\int_{I_1} xf(x;\mu)dx + \frac{n_2}{n} P_2^{-1}(\mu)\int_{I_2} xf(x;\mu)dx = \mu \ . \qquad (2.10)$$

There is a unique solution, $\hat{\mu}$, of this equation; see Corollary 1 to Theorem 3. Note that if $P_1(\mu) = \frac{n_1}{n}$, then the left side of equation (2.10) is $\int_{-\infty}^{\infty} xf(x;\mu)dx$. Therefore, the unique solution of equation (2.10) is the unique solution of the equation $P_1(\mu) = \frac{n_1}{n}$. Since $P_1(\mu) = \Pr\{X \leq r_1;\mu\}$, the unique solution of $P_1(\mu) = \frac{n_1}{n}$ is the unique solution of $F\left(\frac{r_1 - \mu}{\sigma_o}\right) = \frac{n_1}{n}$, where $F(t)$ is the standard normal probability distribution function. Therefore, the solution can be written $\hat{\mu} = r_1 - \sigma_o t$, where $F(t) = \frac{n_1}{n}$.

The ease with which the maximum likelihood estimate of the mean is obtained from a $G_2N_n(\mu,\sigma_o^2)$-sample suggests that one might group further the observations in a $G_kN_n(\mu,\sigma_o^2)$-sample to form a $G_2N_n(\mu,\sigma_o^2)$-sample. For example, if a $G_kN_n(\mu,\sigma_o^2)$-sample is defined by $I_1, \ldots, I_k$ and $n_1, \ldots, n_k$, then one might form a $G_2N_n(\mu,\sigma_o^2)$-sample as follows: let $I_1^* = I_1 \cup I_2$, $I_2^* = I_3 \cup \ldots \cup I_k$, $n_1^* = n_1 + n_2$, and $n_2^* = n_3 + \ldots + n_k$.

Suppose that we desire a starting value, m_o, in the case of a $G_k N_n(\mu,\sigma_o^2)$-sample. Let $I_{1j} \equiv I_1 \cup \ldots \cup I_j$, $I_{2j} \equiv I_{j+1} \cup \ldots \cup I_k$, $n_{1j} = n_1 + \ldots + n_j$, and $n_{2j} = n - n_{1j}$, for $j \in J$, where J is the set of values of j ranging from 1 to k-1 excluding those values of j for which either $n_{1j} = 0$ or $n_{2j} = 0$. For each fixed $j \in J$, I_{1j}, I_{2j}, n_{1j}, and n_{2j} define a $G_2 N_n(\mu,\sigma_o^2)$-sample and a unique estimate M_j which satisfies an equation similar to equation (2.10). The set of possible estimates, M_j, is $\{M_j = r_j - \sigma_o t_j : j \in J \text{ and } F(t_j) = \frac{n_{1j}}{n}\}$. For any fixed j_o, $0 < j_o < k$, M_{j_o} is a consistent estimator of the true mean μ_o. This result follows from the special case, $k = 2$, of Theorem 3 proved by Kulldorff (1958a).

There should be some optimal way to select a single $j_o \in J$ to obtain an M_{j_o} for use as a starting value m_o. Kulldorff (1958a) showed that the large sample variance of $\hat{\mu}$ obtained from a $G_2 N_n(\mu,\sigma_o^2)$-sample is minimum, $\frac{1}{.6366}\left(\frac{\sigma_o^2}{n}\right)$, if $I_1 \equiv (-\infty,\mu_o]$ and $I_2 \equiv (\mu_o,\infty]$. Kulldorff (1958a) indicated that if the mean is unknown, then it is of little value to know that I_1 should be $(-\infty,\mu_o]$. In our case, however, we can choose the j which is such that $\frac{n_1 + \ldots + n_j}{n}$ is nearest 0.5.

Therefore let m_o be such that if $|\frac{n_1 + \ldots + n_j}{n} - \frac{1}{2}|$ is smaller than $|\frac{n_1 + \ldots + n_{j'}}{n} - \frac{1}{2}|$ for each $j' \neq j$, then

$$F(\frac{r_j - m_o}{\sigma_o}) = \frac{n_1 + \ldots + n_j}{n} . \qquad (2.11)$$

In case of a tie, i.e.

$|\frac{n_1 + \ldots + n_j}{n} - \frac{1}{2}| = |\frac{n_1 + \ldots + n_{j+1}}{n} - \frac{1}{2}|$, let m_o be the average of M_j and M_{j+1} where

$F(\frac{r_j - M_j}{\sigma_o}) = \frac{n_1 + \ldots + n_j}{n}$ and

$F(\frac{r_{j+1} - M_{j+1}}{\sigma_o}) = \frac{n_1 + \ldots + n_{j+1}}{n}$. If the r_j selected is reasonably near μ_o, then m_o is a reasonably good estimator of μ_o, but, even more important for our purposes, it is a reasonably good approximation to the solution of the likelihood equation for the $G_k N_n(\mu, \sigma_o^2)$-sample.

Now we consider the execution of the iterations for the method of successive approximations. Let $t_i = \frac{r_i - m}{\sigma_o}$ and $I_i^* = (t_{i-1}, t_i]$. Recall that we seek the root of $\varphi(m) - m = 0$, where

$\varphi(m) = \sum_{i=1}^{k} \frac{n_i}{n} \left[\int_{I_i} f(x;m,\sigma_o^2)dx\right]^{-1} \int_{I_i} xf(x;m,\sigma_o^2)dx$. Letting

$t = \frac{x - m}{\sigma_o}$, we have

$$\varphi(m) = \sum_{i=1}^{k} \frac{n_i}{n} \left[\int_{I_i^*} dF(t)\right]^{-1} \int_{I_i^*} (m + \sigma_o t)dF(t)$$

$$= m + \sigma_o \sum_{i=1}^{k} \frac{n_i}{n} \left[\frac{f(t_{i-1}) - f(t_i)}{F(t_i) - F(t_{i-1})}\right]$$

so that

$$\varphi(m) = m - \sigma_o \sum_{i=1}^{k} \frac{n_i}{n} \left[\frac{f(t_i) - f(t_{i-1})}{F(t_i) - F(t_{i-1})}\right] , \qquad (2.12)$$

where now $F(t)$ and $f(t)$ are the standard normal distribution and density functions, respectively.

Therefore, using conclusion 1 of Theorem 2, the method of successive approximations provides the sequence $\{m_j\}$, where $m_{j+1} = \varphi(m_j)$. Using equation (2.12) we obtain

$$m_{j+1} = m_j - \sigma_o \sum_{i=1}^{k} \frac{n_i}{n} \left[\frac{f(t_i) - f(t_{i-1})}{F(t_i) - F(t_{i-1})}\right], \qquad (2.13)$$

and $t_i = \frac{r_i - m_j}{\sigma_o}$. Let $f(t_i) - f(t_{i-1}) = \Delta_i$ and $F(t_i) - F(t_{i-1}) = \delta_i$. The modified method of successive approximations becomes:

1. Let m_o be the starting value defined in equation (2.11) and form the sequence $\{m_j^*\}$ using the relation,

2. $m_{j+1}^* = m_j^* - (1.1)\sigma_o \sum_{i=1}^{k} \frac{n_i \Delta_i}{n\delta_i}$, where Δ_i and δ_i are evaluated at m_j^*.

Note that it is not necessary to introduce error in the evaluation of $\tilde{\varphi}(m)$ as in Hughes' method. The numerator of each μ_i in equation (2.8) can be obtained in terms of Δ_i and only the denominator of each μ_i must be approximated.

In the following example a $G_4N_{10}(50,100)$-sample is used to obtain the estimate of the true mean $\mu_o = 50$. The sample is composed of the first ten numbers in Table 1, "Table of Random Normal Numbers with Mean Equal to 50 and Variance Equal to 100", of the Appendix in Li (1964). The groups considered are $I_1 \equiv (-\infty,45]$, $I_2 \equiv (45,55]$, $I_3 \equiv (55,60]$, and $I_4 \equiv (60,\infty]$ and the data are 46, 53, 58, 60, 60, 49, 59, 48, 46, 78. This sample defines the cell frequencies: $n_1 = 0$, $n_2 = 5$, $n_3 = 4$, and $n_4 = 1$. Using $m_o = 55$, $F(\frac{55-m_o}{10}) = \frac{1}{2}$, will not exhibit the advantage of using the modified successive approximations method because m_o is very close to the solution $\hat{\mu}$. Instead the starting value is taken to be 50. The solution for $\hat{\mu}$ is given in

Table I. Since the values of the t_i are rounded to the nearest one-hundredth in order to facilitate the use of Table 26.1 in Abramowitz and Stegun (1964), the solution $\hat{\mu}$ is rounded off to the nearest one-tenth.

Only two iterations are required to obtain $\hat{\mu}$ using the modified method of successive approximations with $m_o = 50$. Therefore Table I is arranged in two sections, one for each step in the iterative procedure. The values m_1 and m_2 in the table indicate the values which would be obtained from m_o and m_1^*, respectively, using the method of successive approximations instead of the modified procedure which produced m_1^* and m_2^*.

Since the average of the complete data sample, 55.7, is greater than 50.0, the solution $\hat{\mu}$ might be expected to be larger than the true mean $\mu_o = 50.0$ also. If $m_o = 55.0$ is used as a starting value, it is very close to the solution $\hat{\mu} = 54.8$ and only one iteration is required.

Many numerical examples were considered, some of which will be discussed in section 2.3.4.2, and it was found that the method of successive approximations required roughly thirty per cent more iterations than the modified method of successive approximations to obtain $\hat{\mu}$ with the same precision.

The modified method of successive approximations was

TABLE I: The Modified Method of Successive Approximations

Solution of the Likelihood Equation for μ

$m_o = 50$

i	t_i	$f(t_i)$	$\frac{n_i}{n}\Delta_i$	$F(t_i)$	δ_i
4	$+\infty$	.00000		1.00000	
			0.1(-.24197)		.15866
3	1.0	.24197		0.84134	
			0.4(-.11010)		.14988
2	0.5	.35207		0.69146	
			0.5(+.00000)		.38292
1	-0.5	.35207		0.30854	
			0.0(+.35207)		.30854
0	$-\infty$	.00000		0.00000	

$\sum_{i=1}^{4} \frac{n_i \Delta_i}{n\delta_i} = -0.4463 \quad \therefore m_1^* = 50 - (1.1)(10)(-0.4463) = 54.91$

Note that $m_1 = 50 - 10(-0.4463) = 54.46$.

$m_1^* = 54.91$

i	t_i	$f(t_i)$	$\frac{n_i}{n}\Delta_i$	$F(t_i)$	δ_i
4	$+\infty$	.00000		1.00000	
			0.1(-.35029)		.30503
3	.51	.35029		0.69497	
			0.4(-.04863)		.19098
2	.01	.39892		0.50399	
			0.5(+.15453)		.34290
1	-.99	.24439		0.16109	
			0.0(+.24439)		.16109
0	$-\infty$	.00000		0.00000	

$\sum_{i=1}^{4} \frac{n_i \Delta_i}{n\delta_i} = .0086 \quad \therefore m_2^* = 54.91 - (1.1)(10)(.0086) = 54.815$

$\hat{\mu} = 54.8$ Note that $m_2 = 54.91 - 10(.0086) = 54.824$.

compared to the method of scoring used in the example presented on page 90 in Kulldorff (1961). The modified method of successive approximations solution for $\hat{\mu}$ in that example required one more iteration than did the method of scoring.

2.3.3 Estimating the Variance When the Mean Is Known

In this section the problem of estimating the variance of a normal distribution using grouped data is considered. We assume that the mean, μ_o, is known. It is shown that the method of successive approximations defines a sequence $\{s_j^2\}$ which converges to the unique solution, $\hat{\sigma}^2$, of the likelihood equation for a $G_k N_n(\mu_o,\sigma^2)$-sample, for any starting value $s_o^2 \in (0,\infty)$, provided there is a unique solution.

It further is shown that Hughes' method of solution is the application of quadrature formulae in the iterations required by the successive approximations method. It has been pointed out that the conditions of Hughes' convergence theorem are not satisfied by grouped data samples from a normal distribution; see section 1.2.1 in the Literature Review.

A consistent estimator for the true variance σ_o^2 is found which can be used for the starting value, s_o^2, in any iterative method of solution. The method of successive approximations again is modified for the purposes of solving the particular equation under consideration.

In order to avoid cases where $\hat{\sigma}^2$ is not unique, it is assumed tacitly in this section that the cell frequencies are such that $n_i + n_{i+1} \neq n$, $i = 1, 2, \ldots, k-1$ and such that

$n_1 + n_k \neq n$. This assumption is sufficient to insure that there is a unique solution $\hat{\sigma}^2$ for any $\mu_o \in (-\infty,\infty)$; see Corollary 2 at the end of section 2.2.4 .

At the end of this section an example is provided in which a table of the normal cumulative distribution function and its derivatives is used to obtain the maximum likelihood estimate of the variance of a normal distribution, using a grouped data sample when the mean is known.

Consider a sample with the property $G_k N_n(\mu_o,\sigma^2)$, where $\mu_o \in (-\infty,\infty)$ is the known mean. Using Theorem 4 in section 2.2.5 and the relation

$\frac{d\log(f(x;\sigma^2))}{d(\sigma^2)} = -\frac{1}{2\sigma^2} + \frac{1}{2}(x-\mu_o)^2/\sigma^4$, it follows that

$\sum_{i=1}^{k} n_i E_i\left(-\frac{1}{2\sigma^2} + \frac{1}{2}(X-\mu_o)^2/\sigma^4\right) = 0$ is the likelihood equation for σ^2. This equation can be written as follows:

$$\sigma^2 \sum_{i=1}^{k} \frac{n_i}{n} E_i\left(\frac{X-\mu_o}{\sigma}\right)^2 - \sigma^2 = 0 \ . \qquad (2.14)$$

We define the following functions of the generic variable s^2:

$$\Psi(s^2) = \varphi(s^2) - s^2 \ , \qquad (2.15)$$

where

$$\varphi(s^2) = \left[\sigma^2 \sum_{i=1}^{k} \frac{n_i}{n} E_i\left(\frac{X-\mu_o}{\sigma}\right)^2\right]_{\sigma^2=s^2} \tag{2.16}$$

$$= \left[\sum_{i=1}^{k} \frac{n_i}{n} E_i(X-\mu_o)^2\right]_{\sigma^2=s^2} . \tag{2.17}$$

The derivative of $\varphi(s^2)$ with respect to s^2 is computed as follows:

$$\varphi'(s^2) = \sum_{i=1}^{k} \frac{n_i}{n} P_i^{-2}(s^2)\left(P_i(s^2)\int_{I_i} (X-\mu_o)^2\left[\frac{d}{d(s^2)}f(X;s^2)\right]dX - \right.$$

$$\left. \int_{I_i} (X-\mu_o)^2 f(X;s^2)dX \int_{I_i} \left[\frac{d}{d(s^2)}f(X;s^2)\right]dX\right) .$$

After simplification of this expression, it is found that

$$\varphi'(s^2) = \frac{1}{2}\sum_{i=1}^{k} \frac{n_i}{n} V_i\left(\frac{X-\mu_o}{s}\right)^2 . \tag{2.18}$$

Recall that in order to apply Theorem 2 of section 2.2.3 it must be established that $\varphi'(s^2) \leq \delta < 1$, since it is obvious that $\varphi'(s^2) > 0$. It can be shown that $\varphi'(s^2)$ is not bounded less than one for all $s^2 > 0$.

If Hughes' method were used as in the Appendix, then each $E_i(X-\mu_o)^2$ in equation (2.17) would be approximated using some quadrature formula. The quadrature of each of the terms $E_i(X-\mu_o)^2$ would be performed in exactly the same way that the

quadrature of each μ_i was performed in section 2.3.2 . The only difference is that the integrand in the numerator is now $(X-\mu_o)^2 f(X;s^2)$ instead of $Xf(X;s^2)$. Therefore $\tilde{\varphi}(s^2)$ would be used to approximate $\varphi(s^2)$, where

$$\tilde{\varphi}(s^2) = \sum_{i=1}^{k} \frac{n_i}{n} \left(\frac{\sum_j a_{ij}(x_{ij}-\mu_o)^2 f(x_{ij};s^2)}{\sum_j a_{ij} f(x_{ij};s^2)} \right) , \qquad (2.19)$$

where the summation over j is from $j = 1$ to $j = q_i + 1$.

Kulldorff (1958b) has shown that the second derivative of the log-likelihood function, $\Psi'(s^2)$, is negative at any solution $\hat{\sigma}^2$ and since this second derivative is continuous in $(0,\infty)$, there exists some interval $[a,b]$ containing $\hat{\sigma}^2$ in which some starting value s_o^2 will provide a sequence which converges to $\hat{\sigma}^2$. The condition that the second derivative of the log-likelihood function be negative throughout a closed interval containing s_o^2 and $\hat{\sigma}^2$ depends upon the choice of s_o^2 and is not established easily. The question of how near s_o^2 must be to $\hat{\sigma}^2$ in order to enjoy convergence is answered in the theorems which are presented below.

It will be convenient to express $\varphi(s^2)$ and $\varphi'(s^2)$ in forms similar to those forms considered first by Gjeddebaek (1949) and later by Kulldorff (1958a, 1958b). In order to write these functions in the desired forms it will be necessary to use the following transformations and

definitions:

Let 1. $t = \frac{X-\mu_o}{s}$,

2. $t_i = \frac{r_i - \mu_o}{s}$,

3. $I_i^* \equiv (t_{i-1}, t_i]$,

4. $\frac{dF(u)}{du} = f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$, $-\infty < u < \infty$, and

5. $Z_{ji} = \frac{t_{i-1}^j f(t_{i-1}) - t_i^j f(t_i)}{F(t_i) - F(t_{i-1})}$.

Using definition 5, Clark's (1957) result used in Lemma 1 of section 2.2.2 can be stated as follows:

$$E_i\left(\frac{X-\mu_o}{s}\right) = Z_{0i}$$

$$V_i\left(\frac{X-\mu_o}{s}\right) = 1 + Z_{1i} - Z_{0i}^2 \qquad (2.20)$$

and hence

$$E_i\left(\frac{X-\mu_o}{s}\right)^2 = 1 + Z_{1i} \ . \qquad (2.21)$$

Using equations (2.14), (2.15), and (2.16) we can write:

$$\Psi(s^2) = s^2 \sum_{i=1}^{k} \frac{n_i}{n}\{1 + Z_{1i}\} - s^2$$

$$= s^2 \sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} \qquad (2.22)$$

and

$$\varphi(s^2) = s^2 + s^2 \sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} \ . \qquad (2.23)$$

Using the relation $\frac{d}{dt}f(t) = - tf(t)$ and integration by parts, it follows that

$$\int_{I_i^*} t^k f(t)dt = -\int_{I_i^*} t^{k-1}\{-tf(t)\}dt$$

$$= -\left[t^{k-1}f(t)\Big|_{I_i^*} - \int_{I_i^*}(k-1)t^{k-2}f(t)dt\right]$$

$$= [F(t_i) - F(t_{i-1})][Z_{k-1,i}] +$$

$$(k-1)\int_{I_i^*} t^{k-2}f(t)dt \ .$$

Therefore,

$$\int_{I_i^*} t^k f(t)dt =$$

$$[F(t_i) - F(t_{i-1})][Z_{k-1,i}] + (k-1)\int_{I_i^*} t^{k-2}f(t)dt \ . \qquad (2.24)$$

The expression $V_i\left(\frac{X-\mu_o}{s}\right)^2$ can be written as follows:

$$V_i\left(\frac{X-\mu_o}{s}\right)^2 = E_i\left(\frac{X-\mu_o}{s}\right)^4 - \left(E_i\left(\frac{X-\mu_o}{s}\right)^2\right)^2 \ . \qquad (2.25)$$

Equation (2.24) is used first with k = 4, and then with k = 2

to evaluate the terms of (2.25); the first result is

$$E_i\left(\frac{X-\mu_o}{s}\right)^4 = Z_{3i} + 3\{1 + Z_{1i}\} \ . \qquad (2.26)$$

Applying equations (2.26) and (2.21) to equation (2.25), it follows that

$$V_i\left(\frac{X-\mu_o}{s}\right)^2 = Z_{3i} + 3\{1 + Z_{1i}\} - \{1 + Z_{1i}\}^2 \ .$$

Therefore, we can write equation (2.18) as follows:

$$\varphi'(s^2) = \frac{1}{2} \sum_{i=1}^{k} \frac{n_i}{n}[Z_{3i} + 3\{1 + Z_{1i}\} - \{1 + Z_{1i}\}^2]$$

$$= \frac{1}{2} \sum_{i=1}^{k} \frac{n_i}{n}[Z_{3i} + 2 + Z_{1i} - Z_{1i}^2] \ . \qquad (2.27)$$

Kulldorff (1958b) showed that if $\Psi(s^2)$ (see equation (2.22)) is equal to zero, then

$$\sum_{i=1}^{k} \frac{n_i}{n}(Z_{3i} + Z_{1i} - Z_{1i}^2) = \sum_{i=1}^{k} \frac{n_i}{n}(Z_{3i} - Z_{1i}^2) \ .$$

He also showed that

$$\sum_{i=1}^{k} \frac{n_i}{n}(Z_{3i} - Z_{1i}^2) < 0 \qquad (2.28)$$

for every $s^2 \in (0,\infty)$. Since $\sum_{i=1}^{k} \frac{n_i}{n}(Z_{3i} - Z_{1i}^2)$ is a continuous function of s^2 in $(0,\infty)$, it follows that for $s^2 \in [s_1^2, s_2^2] \subset (0,\infty)$ the function

$\sum_{i=1}^{k} \frac{n_i}{n}(Z_{3i} - Z_{1i}^2)$ is bounded less than or equal to some number less than zero. Kulldorff (1958b) also showed that $\sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} \to \infty$ as $s^2 \to 0$ and that $\sum_{i=1}^{k} \frac{n_i}{n} Z_{1i}$ is less than zero for some $s^2 > \hat{\sigma}^2$. Using these results, it follows that $\Psi(s^2)$ (see equation (2.22)) is a continuous function which is positive for sufficiently small s^2 and is negative for sufficiently large s^2. Since the conditions of Corollary 2 in section 2.2.4 have been assumed, there is a unique zero, $\hat{\sigma}^2$, of $\Psi(s^2)$ and therefore, $\Psi(s^2)$ is negative if $s^2 < \hat{\sigma}^2$ and $\Psi(s^2)$ is positive if $s^2 > \hat{\sigma}^2$. Therefore,

$$\sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} \quad \text{has the same sign as} \quad \hat{\sigma}^2 - s^2. \tag{2.29}$$

Lemma 3

If there is a unique solution, $\hat{\sigma}^2$, of the likelihood equation for σ^2 in a $G_k N_n(\mu_o, \sigma^2)$-sample, then if $s_o^2 \geq \hat{\sigma}^2$ the sequence $\{s_j^2\}$, defined by $s_j^2 = \varphi(s_{j-1}^2)$, is a monotonic decreasing sequence which converges to $\hat{\sigma}^2$.

Proof: Using statement (2.29) and the assumption that $s_o^2 \geq \hat{\sigma}^2$, it follows that

$$\sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} \leq 0 \quad \text{in} \quad [\hat{\sigma}^2, s_o^2] .$$

By the continuity of

$$\sum_{i=1}^{k} \frac{n_i}{n}(Z_{3i} - Z_{1i}^2)$$

it is deduced from equation (2.28) that

$$\sum_{i=1}^{k} \frac{n_i}{n}(Z_{3i} + Z_{1i} - Z_{1i}^2) \leq -\delta_o < 0 \quad \text{in} \quad [\hat{\sigma}^2, s_o^2] \ .$$

This relation implies $\varphi'(s^2) \leq \delta^* < 1$ in $[\hat{\sigma}^2, s_o^2]$; see equation (2.27). The conditions of Theorem 2 in section 2.2.3 are satisfied and it follows that the sequence $\{s_j^2\}$, defined by $s_j^2 = \varphi(s_{j-1}^2)$, is a monotonic, decreasing sequence converging to $\hat{\sigma}^2$.

The relation

$$s_j^2 = s_{j-1}^2 + s_{j-1}^2 \sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} \qquad (2.30)$$

is the computational form of $s_j^2 = \varphi(s_{j-1}^2)$. The functions Z_{1i} can be computed from tables which provide $F(t)$ and $\frac{d^2F(t)}{dt^2}$, where $F(t)$ is the probability distribution function of a standard normal random variable (see Table 26.1 in Abramowitz and Stegun (1964)).

Lemma 4

If there is a unique solution, $\hat{\sigma}^2$, of the likelihood equation for σ^2 in a $G_k N_n(\mu_o, \sigma^2)$-sample, then if $s_o^2 \leq \hat{\sigma}^2$

the sequence $\{s_j^2\}$, defined by $s_j^2 = \varphi(s_{j-1}^2)$ or equivalently $s_j^2 = s_{j-1}^2 + s_{j-1}^2 \sum_{i=1}^{k} \frac{n_i}{n} Z_{1i}$, is either a monotonic increasing sequence converging to $\hat{\sigma}^2$ or there is some j_o such that $s_j^2 \geq s_{j-1}^2$, $j = 1, 2, \ldots, j_o$ and the sequence $\{s_{j_o+\nu}^2\}$ is a monotonic decreasing sequence converging to $\hat{\sigma}^2$ as $\nu \to \infty$.

Proof: (i) Since $s_o^2 \leq \hat{\sigma}^2$ it follows from statement (2.29) that $\sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} > 0$ in $[s_o^2, \hat{\sigma}^2)$. If any $s_{j_o}^2 > \hat{\sigma}^2$, then applying Lemma 3 $s_{j_o+\nu}^2$, $\nu = 0, 1, \ldots$, forms a monotonic decreasing sequence converging to $\hat{\sigma}^2$.

(ii) If there is no $s_j^2 > \hat{\sigma}^2$, then the sequence $\{s_j^2\}$ is comprised of numbers between s_o^2 and $\hat{\sigma}^2$. This follows from the relation $s_j^2 = s_{j-1}^2 + s_{j-1}^2 \sum_{i=1}^{k} \frac{n_i}{n} Z_{1i}$ and the fact that $\sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} > 0$ in $[s_o^2, \hat{\sigma}^2)$.

Therefore, since the sequence $\{s_j^2\}$ is a monotone increasing sequence in $[s_o^2, \hat{\sigma}^2)$ there is some convergent subsequence

$\{s_\nu^2\}$, $\nu = j_1, j_2, \ldots,$. Suppose that $\lim_{\nu\to\infty} s_\nu^2 = \tilde{\sigma}^2 < \hat{\sigma}^2$.

Then $\sum_{i=1}^{k} \frac{n_i}{n} Z_{1i}$, being a continuous function, would be zero at $\tilde{\sigma}^2$. But we assumed that there was but one solution, $\hat{\sigma}^2$. Therefore, by contradiction, $\tilde{\sigma}^2 = \hat{\sigma}^2$. Therefore, $s_j^2 \to \hat{\sigma}^2$ as $j \to \infty$.

The following theorem provides the conditions which, if satisfied by the $G_k N_n(\mu_o,\sigma^2)$-sample, insure that the sequence $\{s_j^2\}$ converges to the unique solution of the likelihood equation for σ^2.

Theorem 9

Let μ_o be the known mean of the underlying normal density function giving rise to the $G_k N_n(\mu_o,\sigma^2)$-sample in which $n_1 + n_k \neq n$ and at least two non-adjacent cell frequencies are non-zero, i.e. $n_i + n_{i+1} \neq n$ for $i = 1, \ldots, k-1$. Then there is a unique solution, $\hat{\sigma}^2$, of the likelihood equation for σ^2 and the sequence $\{s_j^2\}$, defined by

$$s_j^2 = s_{j-1}^2 + s_{j-1}^2 \sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} ,$$

converges to $\hat{\sigma}^2$ for any starting value $s_o^2 \in (0,\infty)$. If the starting value is greater than $\hat{\sigma}^2$, then $\{s_j^2\}$ is a monotonic decreasing sequence. If s_o^2 is less than $\hat{\sigma}^2$, then $\{s_j^2\}$ either

is monotone increasing or is monotone increasing for $j \leq j_o$ and monotone decreasing for $j \geq j_o$.

Proof: Since $n_1 + n_k \neq n$ and since there are at least two non-zero, non-adjacent cell frequencies, there exists a unique solution $\hat{\sigma}^2 \in (0,\infty)$; see Corollary 2 at the end of section 2.2.4. Therefore, Lemma 3 and Lemma 4 are applicable and establish the proof.

We now consider the problem of obtaining a starting value s_o^2. In a way similar to that in which a starting value, m_o, was found for solving the likelihood equation for μ in the preceding section, a starting value, s_o^2, can be found for solving the likelihood equation for σ^2 when the mean is known. An equation similar to equation (2.11) is solved for s_o^2:

$$F\left(\frac{r_j - \mu_o}{s}\right) = \frac{n_1 + \ldots + n_j}{n} . \qquad (2.31)$$

In Table 1 of Kulldorff (1958b) it is seen that if only two intervals $(0,r]$ and $(r,\infty]$ are to be used to obtain a $G_2N_n(\mu_o,\sigma^2)$-sample, then $r = \mu_o + 1.575\sigma_o$ or $r = \mu_o - 1.575\sigma_o$, where σ_o is the true variance, is optimal in the sense that the maximum likelihood estimator, using $r = \mu_o + 1.575\sigma_o$ or $r = \mu_o - 1.575\sigma_o$, has minimum variance over choice of r for large sample sizes. Kulldorff (1958b) proved that the maximum likelihood estimator, $\hat{\sigma}$, in a

$G_2N_n(\mu_o,\sigma^2)$-sample is consistent provided $r \neq \mu_o$. In Table 3 of Kulldorff (1958b) it is found that, asymptotically, the maximum likelihood estimator for σ_o in a $G_2N_n(\mu_o,\sigma^2)$-sample, where $r = \mu_o - 1.575\sigma_o$ or $r = \mu_o + 1.575\sigma_o$, has variance $(.3042)^{-1}V$, where V is the variance of the maximum likelihood estimator for σ_o in a complete data sample of size n.

Therefore, the optimal choice of r_j with which to solve equation (2.31) might be $r_{j_o} \in J$, chosen such that either

$$\left|\frac{n_1 + \ldots + n_{j_o}}{n} - .06\right| \quad \text{or} \quad \left|\frac{n_1 + \ldots + n_{j_o}}{n} - .94\right|$$

is smaller than both of the terms

$$\left|\frac{n_1 + \ldots + n_j}{n} - .06\right| \quad \text{and} \quad \left|\frac{n_1 + \ldots + n_j}{n} - .94\right|$$

for each $j \in J$, $j \neq j_o$, where the probability that a standard normal random variable is less than -1.575 is approximately .06 . The set J is defined as it was in the preceding section and, as in that section, if there is a tie one may average the several possible estimates which might be obtained by solving equation (2.31) using the several possible r_{j_o}'s.

This method of finding a starting value s_o^2 is more complicated than that of finding m_o and the estimator does

have poor asymptotic efficiency relative to that of complete data (see section 1.2.1), but s_o^2 is consistent and would seem to be very useful in problems in which large sample sizes are available. Using this starting value and the modified method of successive approximations to be developed, one can obtain the solution of the likelihood equation in a manner almost identical to that used for solving the likelihood equation for μ.

Recall equation (2.23),

$$\varphi(s^2) = s^2 + s^2 \sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} .$$

The modification suggested here is to choose some λ, $0 \leq \lambda \leq \delta < 1$, with which to define the sequence $\{s_j^{*2}\}$, where

$$s_j^{*2} = \varphi(s_{j-1}^{*2}) + \lambda(\varphi(s_{j-1}^{*2}) - s_{j-1}^{*2})$$

or

$$s_j^{*2} = s_{j-1}^{*2} + (1+\lambda)s_{j-1}^{*2} \sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} .$$

Recall that in the preceding section the equation $\frac{\lambda}{1+\lambda} = 1 - .9$ was solved for λ and the value $\lambda = 0.1$ was found to be satisfactory. In the present case the asymptotic efficiency, relative to that of complete data (see section 1.2.1), is

roughly .8 ; see Table 3 in Kulldorff (1958b). Therefore, $\frac{\lambda}{1+\lambda} = 1 - .8$, $\lambda = .25$ might be satisfactory. Of course there is no best λ for all cases and $\lambda = .2$ is selected for our present purposes. Note that if only three or four intervals define a sample, the value $\lambda = .3$ might be better than $\lambda = .2$, and if $k \geq 10$, then $\lambda = .1$ would be preferable to $\lambda = .2$. These conclusions follow an examination of Table 3 in Kulldorff (1958b).

In the example which follows, the same grouped data sample of ten observations is used as that used for estimating the mean when the variance was known in the preceding section. Recall that $\mu_o = 50$ and $I_1 \equiv (-\infty,45]$, $I_2 \equiv (45,55]$, $I_3 \equiv (55,60]$, $I_4 \equiv (60,\infty]$, $n_1 = 0$, $n_2 = 5$, $n_3 = 4$, and $n_4 = 1$.

The first step is to obtain s_o. Since $(n_1+n_2+n_3)/n = .9$, the starting value s_o is found by solving the equation $F\left(\frac{60 - 50}{s}\right) = .9$, i.e. $s_o = 7.407$. Since the transformation $t_i = \frac{r_i - \mu_o}{s}$ is required, it is somewhat more convenient to define

$$s_j^* = s_{j-1}^*\left(1 + 1.2 \sum_{i=1}^{k} \frac{n_i}{n} z_{1i}\right)^{\frac{1}{2}} .$$

This is equivalent to the modified method of successive

approximations

$$s_j^{*2} = s_{j-1}^{*2}\left(1 + 1.2 \sum_{i=1}^{k} \frac{n_i}{n} Z_{1i}\right) ,$$

$\lambda = .2$. The notation which is used in this example is identical to that used in the example in section 2.3.2, except that instead of using the symbol $\Delta_i = f(t_i) - f(t_{i-1})$ in the evaluation of the Z_{0i} we use $\Delta_i' = f'(t_i) - f'(t_{i-1})$ to evaluate the Z_{1i}. Values of $f'(t)$ are found in Table 26.1 of Abramowitz and Stegun (1964), and the term $\sum_{i=1}^{k} \frac{n_i}{n} Z_{1i}$ is evaluated by computing $\sum_{i=1}^{k} \frac{n_i \Delta_i'}{n\delta_i}$.

The values, s_j, in Table II represent the value one would obtain from s_{j-1}^{*} using the method of successive approximations. The solution, $\hat{\sigma}$, has limited precision since the values of t_i are rounded off to the nearest one-hundredth in order to use Table 26.1 in Abramowitz and Stegun (1964). Therefore, one significant digit is reported as the approximate solution of the likelihood equation. The actual solution is $\hat{\sigma} \approx 6.48$. This would indicate that after the first iteration the values of t_i should be rounded off to the nearest one-thousandth. Then interpolation of tabular values could be used in subsequent iterations to obtain two significant digits for $\hat{\sigma}$.

TABLE II: The Modified Method of Successive Approximations Solution of the Likelihood Equation for σ

$s_0 = 7.407$

i	t_i	$f'(t_i)$	$\frac{n_i}{n}\Delta_i'$	$F(t_i)$	δ_i
4	$+\infty$	.00000		1.00000	
			.00000		.08691
3	1.36	-.21519		0.91309	
			.00004		.16134
2	0.68	-.21528		0.75175	
			-.17224		.50350
1	-0.68	.21528		0.24825	
			.02153		.24825
0	$-\infty$	.00000		0.00000	

$$\sum_{i=1}^{4} \frac{n_i \Delta_i'}{n\delta_i} = -0.2550875$$

$$\therefore s_1 = 7.407\sqrt{1-0.2550875} = 6.393$$

$$s_1^* = 7.407\sqrt{1-1.2(0.2550875)} = 6.170$$

$s_1^* = 6.170$	$s_1 = 6.393$
$s_2^* = 6.17(.988892) = 6.101$	$s_2 = 6.113$
$s_3^* = 6.101(.999038) = 6.095$	$s_3 = 6.096$

$\therefore \hat{\sigma} \approx 6.$ and $\hat{\sigma}^2 \approx 40.$

The estimate of σ_o which is obtained in this example seems to be much worse than the estimate of μ_o which was obtained in the example in section 2.3.2 . This is not surprising since the asymptotic efficiency of $\hat{\sigma}$ relative to that of complete data (see section 1.2.1) is generally much smaller than that for $\hat{\mu}$ if k is less than ten. See Table 8.3 and Table 9.3 in Kulldorff (1961) in order to compare the asymptotic efficiencies of these estimators relative to those of complete data.

Many numerical examples were considered, some of which will be discussed in section 2.3.4.2 . It was found that the method of successive approximations required roughly seventy-five per cent more iterations than the modified method of successive approximations to obtain $\hat{\sigma}$ with the same precision.

The modified method of successive approximations was compared to the method of scoring used in the example given on page 97 in Kulldorff (1961). The modified method of successive approximations for $\hat{\sigma}$ in that example required two more iterations than did the method of scoring.

All of the improvements made in the interest of decreasing the number of iterations for solution without the use of an electronic computer can be used when employing Hughes' method of solution. It should be noted that the

numerator of the functions Z_{1i}, which would be approximated using quadrature formulae if Hughes' method were employed, can be computed as they were in Table II. Only $F(t_i) - F(t_{i-1})$ must be approximated using quadrature formulae; see equation (2.19).

2.3.4 Estimating the Mean and Variance

2.3.4.1 Preliminary Results

In the two preceding sections the properties of the method of successive approximations were examined in the two cases of estimating the unknown mean and the unknown variance using a $G_k N_n(\mu,\sigma^2)$-sample. In this section both parameters are assumed to be unknown. In the two cases previously considered, sufficient conditions were established for a unique solution of the likelihood equations. Great effort was made to establish some set of sufficient conditions for the existence of a unique, simultaneous solution of the two likelihood equations. No success was achieved in this venture, but some numerical results are presented which indicate that if a solution satisfies certain conditions, then it is a relative maximum.

In section 2.3.4.2 a theorem is presented which establishes that a particular iterative procedure developed in that section defines a sequence of iterates which converge to the unique solution of the likelihood equations if there is a unique solution. A procedure is developed which can be used to locate multiple roots of the likelihood equations if there is more than one root.

The reader should refer to Barnett (1966) to acquaint

himself with the favorable properties of the numerical method of false positions when multiple roots exist. Rather strong evidence presented in section 2.3.4.2 suggests that there is a unique solution if the conditions of Theorem 3 are satisfied, and therefore, we shall not consider the method of false positions.

The method of successive approximations was generalized by Ford (1925) to accommodate two equations in two variables. The sufficient conditions given by Ford (1925) for the convergence of the sequence formed by this method are not satisfied by the likelihood equations unless the starting value, (m_o, s_o^2), is sufficiently close to a solution. The method of successive approximations is very similar to Hughes' method of solution, but as it is shown in the Appendix, Hughes' sufficient conditions for convergence are not satisfied either.

The equations which must be solved are:

$$\sum_{i=1}^{k} \frac{n_i}{n} \mu_i - \mu = 0 \qquad (2.8)$$

$$\sigma^2 \sum_{i=1}^{k} \frac{n_i}{n} E_i\left(\frac{X-\mu}{\sigma}\right)^2 - \sigma^2 = 0 \qquad (2.14)$$

as given in the preceding two sections. These equations are quite complicated and all of the questions concerning their

solution will not be answered in this thesis. The objective in section 2.3.4 is to set forth as many explicit results as possible and to acknowledge the problems which are not resolved.

The first question to be answered is whether or not there is a maximum of the likelihood function and, if there is, then when should one look to the solutions of the likelihood equations for the purpose of locating it (or them)? It is assumed tacitly in this section that the conditions of Theorem 3 in section 2.2.4 are satisfied. These conditions insure that the likelihood function,

$$H(\underline{\theta}) = C \prod_{i=1}^{k} P_i^{n_i}(\underline{\theta}) ,$$

has limit zero on the boundary of $\Theta = \{(\mu,\sigma) = (\theta_1,\theta_2): -\infty < \mu < \infty \text{ and } 0 < \sigma < \infty\}$. Therefore, one is justified in seeking the maximum of $H(\underline{\theta})$ by considering the simultaneous solutions of equations (2.8) and (2.14). The use of $\theta_1 = \mu$ and $\theta_2 = \sigma$ in this section provides several simplifications and does not alter the structure of the likelihood equations.

Recalling equations (2.12) and (2.13) in section 2.2.2 and equations (2.23) and (2.30) in section 2.3.3, we define the following functions:

$$M(m,s^2) = m + s \sum_{i=1}^{k} \frac{n_i}{n} Z_{0i}$$

$$S(m,s^2) = s^2 + s^2 \sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} ,$$

where the argument of Z_{ji} is now $t_i = \frac{r_i - m}{s}$. See definitions 1-5 in section 2.3.3 .

The method of successive approximations, generalized by Ford (1925), is as follows: Let (m_o,s_o^2) be a starting value. Then the sequence $\{(m_j,s_j^2)\}$ is obtained by the recursive relation

$$(m_j,s_j^2) = (M(m_{j-1},s_{j-1}^2),S(m_{j-1},s_{j-1}^2)). \qquad (2.32)$$

This method of solution is identical to Hughes' method if the quadrature formulae which were used to evaluate $\phi(m)$ and $\tilde{\phi}(s^2)$ (see equation (2.19)) in the two preceding sections are used to approximate $M(m,s^2)$ and $S(m,s^2)$, respectively, in equation (2.32). The author cannot establish any specific properties of the successive approximations method of solution except those suggested by the numerical study which is presented in section 2.3.4.2 . Examples are found in Hughes (1962) in which Hughes' method seems to be quite good. If more than one solution of the equations exist, then these two methods may, or may not, have favorable properties.

Let us assume that $\underline{\theta} = \underline{Q}$ is obtained, by some method, as a solution of $\frac{\partial L(\underline{\theta})}{\partial \underline{\theta}} = \underline{0}$. From the preceding two sections we know that at $\underline{Q}$:

$$\sum_{i=1}^{k} \frac{n_i}{n} Z_{0i} = \sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} = 0. \qquad (2.33)$$

Gjeddebaek (1949) obtained the following matrix:

$$\frac{\partial^2 L(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}'} =$$

$$\sum_{i=1}^{k} \frac{n_i}{s} \begin{bmatrix} -Z_{0i}^2 + Z_{1i} & -(Z_{0i} - Z_{2i}) - Z_{0i}Z_{1i} \\ -(Z_{0i} - Z_{2i}) - Z_{0i}Z_{1i} & Z_{3i} - 2Z_{1i} - Z_{1i}^2 \end{bmatrix}.$$

At any solution of $\frac{\partial L(\underline{\theta})}{\partial \underline{\theta}} = \underline{0}$, using the relations in (2.33), this matrix reduces to:

$$\frac{\partial^2 L(\underline{Q})}{\partial \underline{\theta} \partial \underline{\theta}'} = \frac{n}{s} \sum_{i=1}^{k} \frac{n_i}{n} \begin{bmatrix} -Z_{0i}^2 & Z_{2i} - Z_{0i}Z_{1i} \\ Z_{2i} - Z_{0i}Z_{1i} & Z_{3i} - Z_{1i}^2 \end{bmatrix}. \qquad (2.34)$$

Let the k matrices in this sum be designated by $Z(i)$, $i = 1, \ldots, k$. If $\frac{\partial^2 L(\underline{Q})}{\partial \underline{\theta} \partial \underline{\theta}'}$ is negative definite, then $\underline{Q}$ is a relative maximum of $L(\underline{\theta})$; see Theorem 4 in section 4.2 of Chapter 4 in Widder (1961). In Lemma 2 it was shown that $\sum_{i=1}^{k} \frac{n_i}{n} V_i\left(\frac{X}{\sigma_o}\right) < 1$. Using this result and equation (2.20), it

follows that $\sum_{i=1}^{k} - \frac{n_i}{n} Z_{0i}^2 < 0$ at any solution $\underline{Q}$. Inequality (2.28) indicates that $\sum_{i=1}^{k} \frac{n_i}{n}(Z_{3i} - Z_{1i}^2) < 0$ at any solution $\underline{Q}$. Therefore, no solution can be a relative minimum (see Widder (1961), loc. cit.). If the determinant of $\frac{n}{s}\sum_{i=1}^{k} \frac{n_i}{n} Z(i)$ is less than zero at $\underline{Q}$, then $\underline{Q}$ is a saddle-point; see Theorem 3 in section 3.1 of Chapter 4 in Widder (1961).

If each Z(i) is negative definite, then $\frac{n}{s}\sum_{i=1}^{k} \frac{n_i}{n} Z(i)$ is negative definite. This follows from the fact that if $\underline{X}'Z(i)\underline{X} < 0$ for every non-null (2 x 1)-vector $\underline{X}$, i = 1, ..., k, then $\underline{X}'\left[\frac{n}{s}\sum_{i=1}^{k} \frac{n_i}{n} Z(i)\right]\underline{X} = \frac{n}{s}\sum_{i=1}^{k} \frac{n_i}{n}(\underline{X}'Z(i)\underline{X}) < 0$. In section 2.3.4.2 it will be shown that if $\frac{\partial^2 L(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}'}$ is negative definite at every solution of the likelihood equations in a region, say R, contained in the parameter space, then any solution in R is a relative maximum.

In order to determine whether the matrices Z(i) are negative definite at any solution of the likelihood equations, the determinant of Z(i), say D(i), was computed for a large

number of intervals $(t_{i-1},t_i]$. If D(i) is greater than zero, then the matrix Z(i) is negative definite. The IBM 7040 computer at the Virginia Polytechnic Institute Computing Center was used to compute D(i) for those intervals having length a multiple of one-tenth from $t_{i-1} = -3.5$ to $t_i = 3.5$ for the standard normal distribution. Those pairs of values t_{i-1} and t_i for which D(i) might not be positive were sought.

Examination of each of the Z_{ji} in Z(i) indicates that D(i) is symmetric with respect to the values t_{i-1} and t_i, i.e. D(i) for (-3,-2] has the same value as D(i) for (2,3], and that D(i) is zero where $t_{i-1} = -t_i$. It can be shown that D(i) is positive where $t_{i-1} = -\infty$ and $t_i < 0$.

The accompanying Table III indicates the values t_{i-1} and t_i^* for which D(i) is negative when the interval $(t_{i-1},t_i]$ is such that $t_i^* \leq t_i < -t_{i-1}$ and is positive for each interval $(t_{i-1},t_i]$ which is such that $t_{i-1} < t_i < t_i^*$.

From Table III it is seen that if $(\hat{\mu},\hat{\sigma})$ is a solution of the likelihood equations for a $G_k N_n(\mu,\sigma^2)$-sample, then if

1. $\hat{\mu} \in I_i$, $1 < i < k$
2. the length of the longest interval,

   $I_2, I_3, \ldots, I_{k-1}$, is less than $3.3\hat{\sigma}$,

then each Z(i) is negative definite and therefore $\frac{\partial^2 L(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}'}$ is

TABLE III: The Intervals for Which the Determinant of Z(i) Becomes Negative

t_{i-1}	t_i^*	Span in Units of σ
-3.5	.5	4.0σ
-3.4	.5	3.9σ
-3.3	.6	3.9σ
-3.2	.6	3.8σ
-3.1	.7	3.8σ
-3.0	.7	3.7σ
-2.9	.8	3.7σ
-2.8	.8	3.6σ
-2.7	.9	3.6σ
-2.6	.9	3.5σ
-2.5	1.0	3.5σ
-2.4	1.1	3.5σ
-2.3	1.1	3.4σ
-2.2	1.2	3.4σ
-2.1	1.3	3.4σ
-2.0	1.4	3.4σ
-1.9	1.5	3.4σ
-1.8	1.5	3.3σ
-1.7	1.6	3.3σ

negative definite at $(\hat{\mu},\hat{\sigma})$. Therefore $(\hat{\mu},\hat{\sigma})$ is a point at which the likelihood function has a relative maximum. Thus, Table III can be used in some cases to determine when a solution might be a relative maximum. The region defined by conditions 1 and 2 will be referred to as R;

$$R = \{(\mu,\sigma): r_1 < \mu < r_{k-1} \text{ and } \max_{1 < i < k} (r_i - r_{i-1}) < 3.3\sigma\}.$$

Conditions 1 and 2 are conservative since each D(i) is required to be positive and since $\hat{\mu}$ is required to be in an interval of finite length. (If $\hat{\mu} \approx r_1 - .5\hat{\sigma}$, then D(1) would be very near zero. The smallest computed value of D(i), from t_{i-1} to t_i, was -.02 .) Some component factors, Z_{ji}, of D(i) were checked in Table 1 in Clark (1957) and were found to be accurate to at least three decimal places.

2.3.4.2 The Iterative Procedure

Before the iterative procedure is discussed, several further properties of the method of successive approximations should be examined. These properties will enable us to establish that the iterative procedure to be presented will provide a sequence of iterates which guarantee that the likelihood function is increased with each iteration. Barnett (1966) stressed that the criterion of convergence is not the only favorable property of an iterative procedure if one seeks the restricted maxima of the likelihood function. If there are multiple roots, and one seeks the value of the restricted maximum likelihood estimator, then one would not be satisfied if the numerical method employed located a solution at which the likelihood function is smaller than it was at the starting value. For this reason a numerical method which insures that the likelihood function increases with each iteration is very desirable.

We must examine the behavior of the log-likelihood function when each iteration is performed. Using the notation $M(m,s^2)$ and $S(m,s^2)$ defined previously, we examine the behavior of the log-likelihood function after each successive iteration on one variable holding the other variable fixed.

Lemma 5

If the conditions of Theorem 3 are satisfied by the $G_k N_n(\mu,\sigma^2)$-sample, then for each fixed value of $s^2 = \sigma_o^2 \in (0,\infty)$ the inequality $L(m_j,\sigma_o^2) > L(m_{j-1},\sigma_o^2)$ is satisfied provided m_{j-1} is not the solution of $\frac{\partial L(\mu,\sigma_o^2)}{\partial \mu} = 0$, where $m_j = M(m_{j-1},\sigma_o^2)$ and $m_o \in (-\infty,\infty)$.

Proof: 1. The only way m_j can equal m_{j-1} is for $\sum_{i=1}^{k} \frac{n_i}{n} Z_{oi} = 0$; see equation (2.13). This occurs only at the solution of $\frac{\partial L(\mu,\sigma_o^2)}{\partial \mu} = 0$. Therefore, by Theorem 8, the sequence $\{m_j\}$ is a strictly monotonic sequence converging to $\hat{\mu}$, the solution of $\frac{\partial L(\mu,\sigma_o^2)}{\partial \mu} = 0$.

2. Since $L(\mu,\sigma_o^2) = 0$ has a unique solution, it is strictly increasing on any interval $(m_o,\hat{\mu})$ and strictly decreasing on any interval $(\hat{\mu},m_o)$.

It follows that $\{L(m_j,\sigma_o^2)\}$ is a strictly increasing sequence for all $m_j \neq \hat{\mu}$.

A similar lemma can be developed for the iterations

performed on s^2, but slight alterations must be made after examination of Lemma 4 in section 2.3.3 . The author cannot establish that $L(\mu_o, s_j^2) > L(\mu_o, s_{j-1}^2)$ unless s_{j-1}^2 is sufficiently large, but the following lemma establishes results which are adequate in subsequent arguments.

Lemma 6

If the conditions of Theorem 3 are satisfied by the $G_k N_n(\mu, \sigma^2)$-sample, then for each fixed value of $m = \mu_o \in (-\infty, \infty)$ the sequence $\{L(\mu_o, s_j^2)\}$ is a strictly increasing sequence for all $j > j_o$, $j_o < \infty$, provided no s_j^2 is the solution of $\frac{\partial L(\mu_o, \sigma^2)}{\partial \sigma^2} = 0$, where $s_j^2 = S(\mu_o, s_{j-1}^2)$ and $s_o^2 \in (0, \infty)$ and j_o is that defined in Lemma 4.

Proof: 1. The only way s_j^2 can equal s_{j-1}^2 is for $\sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} = 0$; see equation (2.30). This occurs only at the unique solution, $\hat{\sigma}^2$, of $\frac{\partial L(\mu_o, \sigma^2)}{\partial \sigma^2} = 0$. Therefore, by Lemma 4 in section 2.3.3, $\{s_j^2\}$ is a strictly monotonic sequence for $j > j_o$ converging to $\hat{\sigma}^2$.

2. Since $L(\mu_o, \sigma^2) = 0$ has a unique solution, it is strictly increasing on any interval

$(s_o^2, \hat{\sigma}^2)$ and is strictly decreasing on any interval $(\hat{\sigma}^2, s_o^2)$. It follows that $L(\mu_o, s_j^2)$ is a strictly increasing sequence for all $j > j_o$ and $s_j^2 \neq \hat{\sigma}^2$.

The following notation will be useful in defining the method of alternating successive approximations. Let s_{j-1}^2 be the j^{th} value in the sequence $\{s_j^2\}$ defined by the recursive relation $s_j^2 = S(m, s_{j-1}^2)$. Denote $s_{j-1+\nu}^2$ by $S^\nu(m, s_{j-1}^2)$, i.e. the iterate obtained from s_{j-1}^2 by performing the operation, $S(m, s_{j-1}^2)$, ν times while holding m fixed. For example, if s_2^2 is given, then, using $\nu = 3$, s_5^2 would be obtained as follows:

$$s_3^2 = S(m, s_2^2)$$

$$s_4^2 = S(m, s_3^2) = S^2(m, s_2^2)$$

$$s_5^2 = S(m, s_4^2) = S^3(m, s_2^2) \ .$$

The method of alternating successive approximations will be defined so that the likelihood function increases with each iteration. In order to do this a rather peculiar iteration will be performed in order to obtain the sequence

$\{(m_j,s_j^2)\}$. The iterate (m_j,s_j^2) will be obtained from (m_{j-1},s_{j-1}^2) by first performing the operation, $M(m_{j-1},s_{j-1}^2)$, to obtain m_j. Then the operation, $S^{\nu}(m_j,s_{j-1}^2)$, is performed to obtain s_j^2, where ν is sufficiently large to insure that $L(m_j,s_j^2) > L(m_j,s_{j-1}^2)$.

The method of alternating successive approximations is defined as follows:

1. Let (m_o,s_o^2) be a starting value in Θ.

2. Let $m_j = M(m_{j-1},s_{j-1}^2)$.

3. Let $s_j^2 = S^{J_{(j-1)}+1}(m_j,s_{j-1}^2)$, where

   (a) $J_{(j-1)} = 0$ if

   $s_{j-1}^2 < S(m_j,s_{j-1}^2) < S^2(m_j,s_{j-1}^2)$

   or if $s_{j-1}^2 > S(m_j,s_{j-1}^2)$.

   (b) If $s_{j-1}^2 < S(m_j,s_{j-1}^2) > S^2(m_j,s_{j-1}^2)$, then $J_{(j-1)}$ is the number of iterations required for

   $L(m_j,S^{J_{(j-1)}}(m_j,s_{j-1}^2)) > L(m_j,s_{j-1}^2)$ and

   $L(m_j,S^{\nu}(m_j,s_{j-1}^2)) \leq L(m_j,s_{j-1}^2)$ for each

   $\nu < J_{(j-1)}$.

Recall from Lemma 6 that if $S(m_j,s^2_{j-1}) < s^2_{j-1}$, then $S^n(m_j,s^2_{j-1})$ converges monotonically to $\hat{s}^2_j$ as $n \to \infty$, where $\hat{s}^2_j$ is the unique solution of $\frac{\partial L(m_j,s^2)}{\partial s^2} = 0$. But if $S(m_j,s^2_{j-1}) > s^2_{j-1}$, then $S(m_j,s^2_{j-1})$ might be on either side of $\hat{s}^2_j$. The definition of s^2_j in step 3 insures that the log-likelihood function has increased and that the value $\hat{s}^2_j$, where $\frac{\partial L(m_j,\hat{s}^2_j)}{\partial s^2} = 0$, is not between $S^{J(j-1)}(m_j,s^2_{j-1})$ and $S^{J(j-1)+1}(m_j,s^2_{j-1})$. If we knew that $L(m_j,S(m_j,s^2_{j-1})) > L(m_j,s^2_{j-1})$, then it would not be necessary to execute the $J_{(j-1)}$ additional iterations to insure that s^2_j is such that $L(m_j,s^2_j) > L(m_j,s^2_{j-1})$. Examination of Lemma 5 indicates that these properties also are enjoyed by the first iteration $m_j = M(m_{j-1},s^2_{j-1})$.

In the theorem which follows it is shown that the sequence $\{L(m_j,s^2_j)\}$, defined by the method of alternating successive approximations, converges to some $L > L(m_o,s^2_o)$.

Theorem 10

Let the conditions of Theorem 3 be satisfied by a $G_kN_n(\mu,\sigma^2)$-sample and let (m_o,s^2_o) be a starting value which

is not a solution of the likelihood equations. The method of alternating successive approximations defines a sequence $\{(m_j,s_j^2)\}$ which is such that the sequence $\{L(m_j,s_j^2)\}$ is a monotonic sequence which converges to $L > L(m_o,s_o^2)$.

Proof: It follows from the definition of the sequence $\{(m_j,s_j^2)\}$ that $[L(m_j,s_{j-1}^2) - L(m_{j-1},s_{j-1}^2)] > 0$ and that $[L(m_j,s_j^2) - L(m_j,s_{j-1}^2)] > 0$. Therefore the sum of these two positive terms, $[L(m_j,s_j^2) - L(m_{j-1},s_{j-1}^2)]$, is positive, and it follows that the sequence $\{L(m_j,s_j^2)\}$ is a monotonic increasing sequence. Since $L(m,s^2)$ is bounded less than n! (see equation (2.0)), then $L(m_j,s_j^2) \to L$ for some L, $L(m_o,s_o^2) < L < n!$.

In Theorem 11 we shall use the relations:

$$m_j - m_{j-1} = \left(\frac{s_{j-1}^2}{n}\right)\frac{\partial L(m_{j-1},s_{j-1}^2)}{\partial m}$$

$$S(m_{j-1},s_{j-1}^2) - s_{j-1}^2 = \left(\frac{2s_{j-1}^4}{n}\right)\frac{\partial L(m_{j-1},s_{j-1}^2)}{\partial s^2} \quad .$$

These follow an examination of equations (2.8) and (2.14) and of the definitions of $M(m,s^2)$ and $S(m,s^2)$. Recall that $s_j^2 = S^{J(j-1)+1}(m_j,s_{j-1}^2)$. In Theorem 11 we shall define

$S^{J(j-1)}(m_j,s^2_{j-1})$ to be u^2_j, i.e. $s^2_j = S(m_{j-1},u^2_j)$. Note that the inequalities

$L(m_j,s^2_j) - L(m_j,s^2_{j-1}) \geq L(m_j,s^2_j) - L(m_j,u^2_j) \geq 0$ follow from the definition of $J_{(j-1)}$.

Theorem 11

If the conditions of Theorem 3 are satisfied by a $G_kN_n(\mu,\sigma^2)$-sample and if there is a unique solution, $\hat{\underline{\theta}}$, of the likelihood equations, then the method of alternating successive approximations defines a sequence $\{(m_j,s^2_j)\}$ which converges to $\hat{\underline{\theta}}$ for any starting value $(m_o,s^2_o) \in \Theta$.

Proof: The difference $[L(m_j,s^2_j) - L(m_{j-1},s^2_{j-1})]$ can be expressed as the sum of two positive terms:

$$[L(m_j,s^2_{j-1}) - L(m_{j-1},s^2_{j-1})] + [L(m_j,s^2_j) - L(m_j,s^2_{j-1})].$$

It follows from Theorem 10 that both of these positive terms must have limit zero. Since

$[L(m_j,s^2_j) - L(m_j,s^2_{j-1})] \geq [L(m_j,s^2_j) - L(m_j,u^2_j)]$,

$u^2_j = S^{J(j-1)}(m_j,s^2_{j-1})$, it follows that

$[L(m_j,s^2_j) - L(m_j,u^2_j)] \to 0$. Using the mean value theorem (see Theorem 2 in Chapter 1 of Widder (1961)), we have:

$$[L(m_j,s^2_j) - L(m_j,u^2_j)] = (s^2_j - u^2_j)\frac{\partial L(m_j,\xi_j)}{\partial s^2} \to 0 ,$$

where ξ_j is between s_j^2 and u_j^2. By the definition of the sequence $\{(m_j,s_j^2)\}$ we know that $\frac{\partial L(m_j,s^2)}{\partial s^2}$ has no zero between s_j^2 and u_j^2. Therefore, if $(s_j^2 - u_j^2)$ fails to have limit zero, then $[L(m_j,s_j^2) - L(m_j,u_j^2)]$ would not have limit zero. Therefore, $s_j^2 - u_j^2 \to 0$, but $(s_j^2 - u_j^2) = \left(\frac{2u_j^4}{n}\right)\frac{\partial L(m_j,u_j^2)}{\partial s^2}$ also has limit zero. If $u_j^2 \to 0$, then $L(m_j,u_j^2) \to -\infty$ by Theorem 3. Since $L(m_j,u_j^2) \geq L(m_o,s_o^2)$, then u_j^2 cannot have limit zero. Therefore, $\frac{\partial L(m_j,u_j^2)}{\partial s^2}$ has limit zero. Since $(s_j^2 - u_j^2) \to 0$ and since $\frac{\partial L(m,s^2)}{\partial s^2}$ is a continuous function, it follows that $\frac{\partial L(m_j,s_j^2)}{\partial s^2}$ also has limit zero.

Using a similar argument, $[L(m_j,s_{j-1}^2) - L(m_{j-1},s_{j-1}^2)] = (m_j - m_{j-1})\frac{\partial L(\xi_j,s_{j-1}^2)}{\partial m} \to 0$, we conclude that $\frac{\partial L(m_j,s_j^2)}{\partial m} \to 0$. Since both partial derivatives have limit zero and since $\hat{\underline{\theta}}$ is the unique point at which $\frac{\partial L(\underline{\theta})}{\partial \underline{\theta}} = \underline{0}$, it follows that $(m_j,s_j^2) \to \hat{\underline{\theta}}'$.

The method of alternating successive approximations is

of little practical value in its present form. In practice the rate of convergence can be increased if $J_{(j-1)}$ is taken to be zero. The rate of convergence is also increased if the modified methods of successive approximations developed in sections 2.3.2 and 2.3.3 are used to obtain subsequent iterates in the method of alternating successive approximations in lieu of the successive approximations methods. If these modifications are made, then there is no guarantee that the solution which might be obtained is a point at which the likelihood function is larger than it was at the starting value, unless there is a unique solution.

Before further consideration is given to this method of solution, a criterion for choosing the starting value (m_o, s_o^2) should be established.

When one of the parameters is known, it was suggested in sections 2.3.2 and 2.3.3 that the k intervals be combined to form two intervals. Then one simple equation is solved which maximizes the likelihood function for the $G_2N_n(\mu,\sigma^2)$-sample. Suppose that one wishes to find the maximum of the likelihood function for a $G_3N_n(\mu,\sigma^2)$-sample satisfying the conditions of Theorem 3. The likelihood function is

$$H(\underline{\theta}) = C \prod_{i=1}^{3} P_i^{n_i}(\underline{\theta}) ;$$

see Theorem 3. It is easy to see that $\hat{\underline{\theta}}$, such that

$P_1(\hat{\underline{\theta}}) = \frac{n_1}{n}$ and $P_1(\hat{\underline{\theta}}) + P_2(\hat{\underline{\theta}}) = \frac{n_1 + n_2}{n}$, satisfies the likelihood equations (2.8) and (2.14) given previously. It is not obvious that this value of $\underline{\theta}$ also maximizes the likelihood function. Closer examination indicates that

$\max_{\pi_i} \prod_{i=1}^{3} \pi_i^{n_i}$, $0 < \pi_i < 1$ and $\sum_{i=1}^{3} \pi_i = 1$, also occurs at

$\pi_i = \frac{n_i}{n}$, and therefore the maximum of $H(\underline{\theta})$ over the set of π_i restricted to the parametric relation $\pi_i = P_i(\underline{\theta})$ can be no larger. It follows that the value of $\hat{\underline{\theta}}$ for which $P_i(\hat{\underline{\theta}}) = \frac{n_i}{n}$ does maximize $H(\underline{\theta})$. This argument also serves to prove that a strict maximum likelihood estimator of $\underline{\theta}$ in a $G_3N_n(\mu,\sigma^2)$-sample exists and is the solution of the equations $P_1(\underline{\theta}) = \frac{n_1}{n}$ and $P_1(\underline{\theta}) + P_2(\underline{\theta}) = \frac{n_1 + n_2}{n}$, provided that the conditions of Theorem 3 are satisfied. The consistency of $\hat{\underline{\theta}}$ obtained from a $G_3N_n(\mu,\sigma^2)$-sample is established in Chapter III.

If there are more than three cells, then one way to obtain a consistent estimator of $\underline{\theta}$ is to combine adjacent cells and form a $G_3N_n(\mu,\sigma^2)$-sample. As before, there should be some optimal way of doing this.

The objective is to obtain an estimator, from some $G_3N_n(\mu,\sigma^2)$-sample, for (μ,σ^2) which can be used for a

starting value (m_o, s_o^2) which is close to a value $(\hat{\mu}, \hat{\sigma}^2)$ at which the likelihood function for the $G_k N_n(\mu, \sigma^2)$-sample attains its absolute maximum. Using Table 1 in Kulldorff (1963), we find that the product of the variances, for large sample sizes, of $\hat{\mu}$ and $\hat{\sigma}$ obtained from a $G_3 N_n(\mu, \sigma^2)$-sample is minimized if $I_1 \equiv (-\infty, \mu - 2.221\sigma]$,

$I_2 \equiv (\mu - 2.221\sigma, \mu + 2.221\sigma]$ and $I_3 \equiv (\mu + 2.221\sigma, \infty]$.

If these are the cells, then the variances of the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}$ obtained from a large, grouped data sample are approximately 1.5 and 1.8 times as large, respectively, as the variances of these estimators obtained from a complete data sample. The probability that a normal random variable is less than 2.22σ below its mean is approximately .013 . Therefore, the $G_3 N_n(\mu, \sigma^2)$-sample might be selected optimally as follows: choose as the right end point of the interval, I_i, that value r_i for which $\left|\frac{n_1 + \ldots + n_i}{n} - .013\right|$ is nearest zero to form $I_1^* \equiv (-\infty, r_i]$, $n_1 + \ldots + n_i \neq 0$; and choose as the right end point of the interval, I_j, that value r_j for which $\left|\frac{n_1 + \ldots + n_j}{n} - .987\right|$ is nearest zero to form $I_2^* \equiv (r_i, r_j]$, $n_1 + \ldots + n_j \neq n$. Then solve the equations:

$$F\left(\frac{r_i - m_o}{s_o}\right) = \frac{n_1 + \ldots + n_i}{n} \qquad (2.35)$$

$$F\left(\frac{r_j - m_o}{s_o}\right) = \frac{n_1 + \ldots + n_j}{n} , \qquad (2.36)$$

where F(t) is the standard normal probability distribution function.

Now that we have an approximation to a point at which the likelihood function assumes its absolute maximum, we must select an iterative procedure with desirable properties. Since it has not been shown that the likelihood function for a $G_k N_n(\mu,\sigma^2)$-sample satisfying the conditions of Theorem 3 has a unique maximum, then one desirable property is that the iterative procedure provide a sequence which converges to one of the restricted maximum likelihood estimates.

In the forthcoming development the following assumptions are made:

1. The log-likelihood function $L(\underline{\theta})$, where now $\underline{\theta}' = [\mu,\sigma]$, is the log-likelihood function for a $G_k N_n(\mu,\sigma^2)$-sample satisfying the conditions of Theorem 3.
2. The numerical results presented in Table III establish that $\frac{\partial^2 L(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}'}$ is a negative definite matrix at any solution of $\frac{\partial L(\underline{\theta})}{\partial \underline{\theta}} = \underline{0}$ which is in R, where

$$R = \{(\mu,\sigma): r_1 \leq \mu \leq r_{k-1},\ \max_{1<i<k}(r_i - r_{i-1}) \leq 3.3\sigma\}.$$

Using the property of the uniqueness of the maxima (see Theorems 8 and 9) the following definitions are used:

3. The unique maximum of $L(\mu,\sigma_o)$ is given by $L(\hat{\mu},\sigma_o)$, where $\mu = g(\sigma)$ and $\hat{\mu} = g(\sigma_o)$ for every $\sigma_o \in (0,\infty)$.
4. The unique maximum of $L(\mu_o,\sigma)$ is given by $L(\mu_o,\hat{\sigma})$, where $\sigma = h(\mu)$ and $\hat{\sigma} = h(\mu_o)$ for every $\mu_o \in (-\infty,\infty)$.

The differentiability of $h(\mu)$ and $g(\sigma)$ can be established at each $\underline{\theta} \in \Theta$ for which $\frac{\partial^2 L(\underline{\theta})}{\partial\mu\partial\sigma} \neq 0$; see Theorem 14, [Implicit Functions Theorem], in Chapter 1, section 12.1 of Widder (1961):

$$h'(\mu) = -\frac{\partial^2 L(\underline{\theta})}{\partial\mu^2}\left(\frac{\partial^2 L(\underline{\theta})}{\partial\mu\partial\sigma}\right)^{-1} \tag{2.37}$$

and

$$g'(\sigma) = -\frac{\partial^2 L(\underline{\theta})}{\partial\sigma^2}\left(\frac{\partial^2 L(\underline{\theta})}{\partial\mu\partial\sigma}\right)^{-1} \tag{2.38}$$

By the definition of $g(\sigma)$ and $h(\mu)$ it follows that $\frac{\partial L(\underline{\theta})}{\partial\underline{\theta}} = \underline{0}$ only when the two arcs $\mu = g(\sigma)$ and $\sigma = h(\mu)$ intersect in the (μ,σ)-plane.

If $\hat{\underline{\theta}}' = [\hat{\mu},\hat{\sigma}] \in R$ is a solution of $\frac{\partial L(\underline{\theta})}{\partial\underline{\theta}} = \underline{0}$, then using Taylor's Theorem (see section 9.2, Chapter 1 of Widder

(1961)),

$$L(\mu,\sigma) - L(\hat{\mu},\hat{\sigma}) = [\underline{\theta} - \hat{\underline{\theta}}]' \frac{\partial L(\hat{\underline{\theta}})}{\partial \underline{\theta}} + \frac{1}{2}[\underline{\theta} - \hat{\underline{\theta}}]' \frac{\partial^2 L(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}'}\Big|_{\underline{\xi}} [\underline{\theta} - \hat{\underline{\theta}}] ,$$

where $\xi_1 = \hat{\theta}_1 + \lambda(\theta_1 - \hat{\theta}_1)$ and $\xi_2 = \hat{\theta}_2 + \lambda(\theta_2 - \hat{\theta}_2)$ and $0 < \lambda < 1$. Since $\frac{\partial L(\hat{\underline{\theta}})}{\partial \underline{\theta}} = \underline{0}$, it follows that if $(\mu,\sigma) \in R$, then $L(\mu,\sigma) - L(\hat{\mu},\hat{\sigma}) < 0$ for every $\underline{\theta}$ in R, where $\underline{\theta} \neq \hat{\underline{\theta}}$, since $\frac{\partial^2 L(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}'}\Big|_{\underline{\xi}}$, $\underline{\xi} \in R$, is a negative definite matrix by assumption 2. Hence $L(\mu,\sigma) < L(\hat{\mu},\hat{\sigma})$ for every $\underline{\theta} \neq \hat{\underline{\theta}}$ in R. Therefore if there is a solution $(\hat{\mu},\hat{\sigma})$ in R, then it is the unique solution in R and is the point at which $L(\underline{\theta})$ assumes its absolute maximum for $\underline{\theta} \in R$.

Ford (1925) proved that if the starting value, in this case (m_o,s_o), is sufficiently close to a solution $(\hat{\mu},\hat{\sigma}) \in R$, then the method of successive approximations (or Hughes' method) provides a sequence which converges to $(\hat{\mu},\hat{\sigma})$. This method would seem to be appropriate for obtaining the solution $(\hat{\mu},\hat{\sigma}) \in R$, if indeed there is one. If the configuration of the cell frequencies is such that it is rather obvious that any solution in the complement of R would be a ridiculous estimate of $\underline{\theta}$, then one might be satisfied to accept the solution $(\hat{\mu},\hat{\sigma}) \in R$. But if $(\hat{\mu},\hat{\sigma})$ is near the boundary of R one would want to know whether or not

there might be a solution $(\mu^*,\sigma^*) \notin R$ such that $L(\mu^*,\sigma^*) > L(\hat{\mu},\hat{\sigma})$. The sequence $\{(m_j,s_j^2)\}$ defined by any method of solution might converge to some value (μ^*,σ^*) which is not in R. If this happened one would ask whether or not (μ^*,σ^*) might be a value of the restricted maximum likelihood estimator. In general, one wishes to locate all of the solutions of the likelihood equations and to select a solution at which the likelihood function is maximum, i.e. one seeks a value of the restricted maximum likelihood estimator.

The procedure immediately following was found to be quite helpful in answering these questions. Reference to Figure 1 will clarify some aspects of the discussion.

Suppose that $(\hat{\mu},\hat{\sigma})$ is a solution of the likelihood equations and that one wishes to examine the likelihood equations to see whether or not there is another root in the region $R^* = \{(\mu,\sigma):|\hat{\mu} - \mu| \leq \Delta_1, |\hat{\sigma} - \sigma| \leq \Delta_2\}$. Let $m_1 < m_2 < \ldots < m_q$ be q equally spaced points in $\{\mu:|\hat{\mu} - \mu| \leq \Delta_1\}$ and let $s_1 < s_2 < \ldots < s_p$ be p equally spaced points in $\{\sigma:|\hat{\sigma} - \sigma| \leq \Delta_2\}$. Using the two modified methods of successive approximations developed in sections 2.3.2 and 2.3.3, obtain $\hat{\sigma}_1, \ldots, \hat{\sigma}_q$ on the arc given by $\sigma = h(\mu)$ which are such that $L(m_i,\hat{\sigma}_i) > L(m_i,s)$,

for $s \neq \hat{\sigma}_i$, $i = 1, \ldots, q$, and $\hat{\mu}_1, \ldots, \hat{\mu}_p$ on the arc given by $\mu = g(\sigma)$ which are such that $L(\hat{\mu}_i, s_i) > L(m, s_i)$, for $m \neq \hat{\mu}_i$, $i = 1, \ldots, p$. The $\hat{\mu}_i$, as used here, are the points at which $L(m, s_i)$ is maximum. They should not be confused with the μ_i defined in section 2.3.2 . These points, $\{(m_i, \hat{\sigma}_i)$, $i = 1, \ldots, q$, and $(\hat{\mu}_i, s_i)$, $i = 1, \ldots, p\}$, and $(\hat{\mu}, \hat{\sigma})$ provide a rough graph of $\mu = g(\sigma)$ and $\sigma = h(\mu)$. If these two curves intersect at some point (μ^*, σ^*) other than at $(\hat{\mu}, \hat{\sigma})$, then (μ^*, σ^*) is another solution of $\frac{\partial L(\underline{\theta})}{\partial \underline{\theta}} = \underline{0}$. Then $L(\hat{\mu}, \hat{\sigma})$ and $L(\mu^*, \sigma^*)$ can be compared to ascertain which is the larger.

If $(\hat{\mu}, \hat{\sigma})$ is in the region R, then there can be no other intersection of $\mu = g(\sigma)$ and $\sigma = h(\mu)$ in R. Therefore, a more appropriate selection of $m_1, \ldots, m_q$ and $s_1, \ldots, s_p$ can be made, which are not points on $\mu = g(\sigma)$ or $\sigma = h(\mu)$ in R.

The following statements can be made concerning the intersections of $\mu = g(\sigma)$ and $\sigma = h(\mu)$, say (μ^*, σ^*);

1. If $\frac{\partial^2 L(\mu^*, \sigma^*)}{\partial \mu \partial \sigma} = 0$, then $\frac{\partial^2 L(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}'}$ is a diagonal matrix and therefore it is negative definite. Hence (μ^*, σ^*) is a relative maximum.
2. If $\frac{\partial^2 L(\mu^*, \sigma^*)}{\partial \mu \partial \sigma} \neq 0$, then (see equations (2.37) and

(2.38))

$$h'(\mu) = -\frac{\partial^2 L(\underline{\theta})}{\partial \mu^2}\left(\frac{\partial^2 L(\underline{\theta})}{\partial \mu \partial \sigma}\right)^{-1}$$

$$g'(\sigma) = -\frac{\partial^2 L(\underline{\theta})}{\partial \sigma^2}\left(\frac{\partial^2 L(\underline{\theta})}{\partial \mu \partial \sigma}\right)^{-1}$$

and therefore

(a) $h'(\mu^*)g'(\sigma^*) > 1$ if and only if $\frac{\partial^2 L(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}'}$ is negative definite

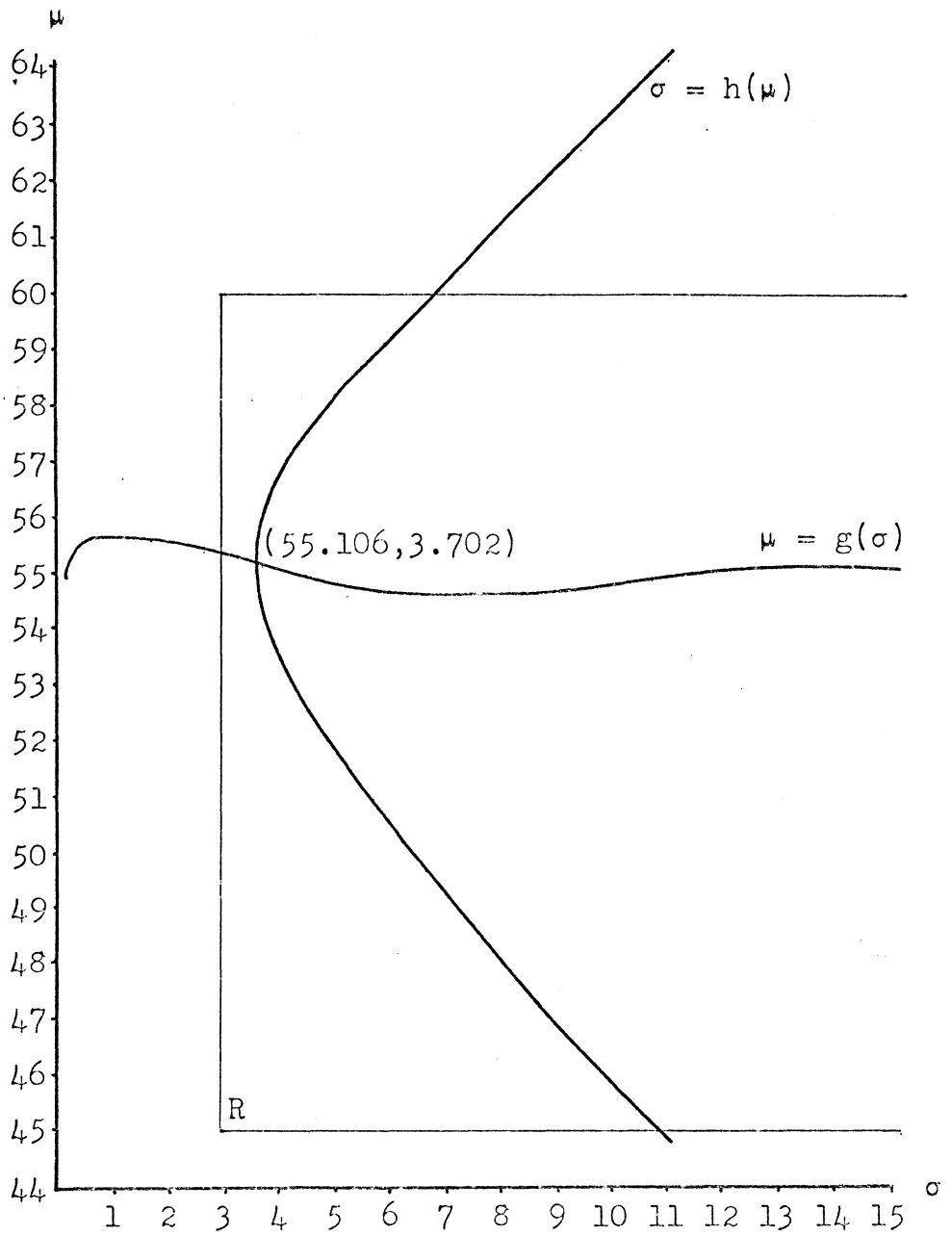
(b) $0 < h'(\mu^*)g'(\sigma^*) < 1$ if and only if $\frac{\partial^2 L(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}'}$ is positive definite

(c) if $h'(\mu^*)g'(\sigma^*) = 1$, then $\left|\frac{\partial^2 L(\underline{\theta})}{\partial \underline{\theta} \partial \underline{\theta}'}\right| = 0$.

In the following example the same sample which was used in sections 2.3.2 and 2.3.3 is assumed to have been drawn from a normal distribution with unknown mean and variance. Recall that the $G_4N_{10}(50,100)$-sample is:

$r_1 = 45$, $r_2 = 55$, $r_3 = 60$, $n_1 = 0$, $n_2 = 5$, $n_3 = 4$, $n_4 = 1$. Using the method for obtaining (m_o, s_o) developed earlier in this section, we find the solution of the equations $F\left(\frac{m_o - 55}{s_o}\right) = \frac{5}{10}$ and $F\left(\frac{m_o - 60}{s_o}\right) = \frac{9}{10}$. The starting value $(m_o, s_o) = (55.0, 3.91)$ is obtained as the solution of these equations.

FIGURE 1: The Solution of the Likelihood Equations for μ and σ



The solution of the likelihood equations is $(\hat{\mu},\hat{\sigma}) = (55.106,3.702)$. This solution was obtained using all of the methods described in the discussion which follows this example. The two curves $\mu = g(\sigma)$ and $\sigma = h(\mu)$ are sketched in Figure 1 and R is the set $\{(\mu,\sigma): 45 \leq \mu \leq 60,\ 10 \leq 3.3\sigma\}$.

Several numerical methods of solution were compared and a search was made for a $G_k N_n(\mu,\sigma^2)$-sample satisfying the conditions of Theorem 3 which might provide two solutions of $\frac{\partial L(\underline{\theta})}{\partial \underline{\theta}} = \underline{0}$. A description of some of these numerical methods follows:

1. Successive approximations, method one:

$$(m_{j+1},s^2_{j+1}) = (M(m_j,s^2_j),S(m_j,s^2_j)),$$

2. Modified successive approximations, method two:

$$m_{j+1} = m_j + 1.3s_j \sum_{i=1}^{k} \frac{n_i}{n} Z_{0i}$$

$$s^2_{j+1} = s^2_j + 1.4s^2_j \sum_{i=1}^{k} \frac{n_i}{n} Z_{1i},$$

3. Modified alternating successive approximations, method three:

$$m_{j+1} = m_j + 1.1s_j \sum_{i=1}^{k} \frac{n_i}{n} Z_{0i}$$

$$s^2_{j+1} = s^2_j + 1.2 s^2_j \sum_{i=1}^{k} \frac{n_i}{n} Z_{1i} ,$$

where the Z_{1i} are evaluated at (m_{j+1}, s^2_j).

Method one is the method of successive approximations. Recall that this method is very similar to Hughes' method in this case.

Method two is very similar to method one, but the coefficients 1.3 and 1.4 were selected following an examination of Table 1 in Kulldorff (1963) in the same way in which the λ's were selected in sections 2.3.2 and 2.3.3, respectively, in order to reduce the number of iterations required for solution.

Method three is similar to the alternating successive approximations method. In all of the examples which were studied the method of successive approximations for obtaining $\hat{\sigma}^2$ never produced a sequence $\{s^2_j\}$ which was not a monotonic sequence converging to the solution $\hat{\sigma}^2$. This would indicate that $J_{(j-1)} = 0$; see Lemma 4 and Lemma 6 and the definition of the method of alternating successive approximations. The coefficients 1.1 and 1.2 which appear in the equations defining method three are the same as those used for modification of the method of successive approximations in sections 2.3.2 and 2.3.3, respectively.

A large number of numerical examples were solved using the IBM 7040 computer at the Virginia Polytechnic Institute Computing Center. Some of the comparisons of the methods of successive approximations and the modified methods of successive approximations considered in sections 2.3.2 and 2.3.3 were made earlier in those sections.

General observations concerning the three methods of solution described in this section follow:

A. Method one required roughly thirty-five per cent more iterations than methods two and three.

B. Methods two and three required about the same number of iterations. Method two appears to be the most appropriate method of solution if the iterations are to be executed using a high speed computer. This method never failed to converge.

C. It was shown that the method of alternating successive approximations provides a convergent sequence if there is a unique solution. If we assume that $J_{(j-1)} = 0$, then method three is essentially the same method of solution as the alternating successive approximations method, except that the number of iterations required for solution was reduced by roughly thirty per cent in those cases considered. Method three might be the

best method if the solution must be obtained by hand. The same operations which were performed in Table I and Table II can be used to facilitate these iterations.

A search was conducted to find a sample which might provide two distinct solutions of the likelihood equations. Three cases were considered, $k = 3$, $k = 4$, and $k = 5$. Cell frequencies were selected in such a way as to satisfy the conditions of Theorem 3, but which might represent extremely unusual sample configurations, for example, $n_1 = 4$, $n_3 = 1$, and $n_5 = 5$. The arcs $\mu = g(\sigma)$ and $\sigma = h(\mu)$ were drawn. In no case were two solutions found.

The numerical results presented in the preceding section and those presented here would seem to indicate that one should expect only one joint solution of the likelihood equations when both parameters are unknown. The method of locating multiple roots can be used to verify this for any given $G_k N_n(\mu,\sigma^2)$-sample. It was found that the general characteristics of the two arcs in Figure 1 seemed to be common to all sample configurations, except when over one-half of the observations fell in I_1 or I_k.

CHAPTER III

ASYMPTOTIC PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS OBTAINED FROM GROUPED DATA SAMPLES

3.1 Introduction and Definitions

In this chapter theorems are developed which provide sufficient conditions to insure the asymptotic efficiency of maximum likelihood estimators obtained from grouped data samples. Recall that if $\underline{X}$ is a vector of random variables with frequency function $f(\underline{x};\underline{\theta})$, where $\underline{\theta}$ is a q x 1 vector of parameters, then the sequence $\{\hat{\underline{\theta}}_n\}$ has been defined to be asymptotically efficient for $\underline{\theta}_o$, the true value of $\underline{\theta}$, if $\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}_o)$ has as its limiting distribution the q-variate normal distribution with mean vector $\underline{0}$ and covariance matrix $\left[E_{\underline{X}}\left[\frac{\partial \log(f(\underline{X};\underline{\theta}_o))}{\partial \underline{\theta}}\right]\left[\frac{\partial \log(f(\underline{X};\underline{\theta}_o))}{\partial \underline{\theta}'}\right]\right]^{-1}$. This covariance matrix is the inverse of the information matrix for a sample of size one.

In the Introduction several maximum likelihood estimators were defined. The present objective is to find sufficient conditions for the strict maximum likelihood estimator $\hat{\underline{\theta}}_n$ and for the restricted maximum likelihood estimator $\underline{\theta}_n^*$ to be asymptotically efficient. The asymptotic efficiency of the maximum likelihood estimators of μ and σ^2 from a $G_k N_n(\mu,\sigma^2)$-sample is established. Recall that

Kulldorff, in Theorem 3 (1958a) and in Theorem 3 (1958b), showed only that the maximum likelihood estimators, when they exist for μ given σ^2 and for σ^2 given μ, are asymptotically efficient.

In the development which follows it will be necessary to restrict our attention to a sub-class of parametric families of distribution functions. In general, if $\underline{X}$ is a p x 1 vector of random variables defined in a subset $\underline{\mathcal{X}}$ of Euclidean p-space, E_p, and if $\underline{\theta}$ is a q x 1 vector of variables defined in a subspace Θ of Euclidean q-space, E_q, then the class of distribution functions $F(\underline{x};\underline{\theta})$ such that $\theta \in \Theta$ is called a parametric family of distribution functions. We shall refer to the class of distribution functions $\{F(\underline{x};\underline{\theta}): \underline{\theta} \in \Theta\}$ as a continuous, parametric family of distribution functions if and only if the following conditions are satisfied:

1. There exists a non-void class of s distinct points, $Q = \{(\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_s): s \leq q,\ 0 < F(\underline{x}_i;\underline{\theta}) < 1$ for each $\underline{\theta} \in \Theta$, $i = 1,\ldots,s\}$, such that if $(\underline{x}_1,\ldots,\underline{x}_s) \in Q$ and $\underline{\theta}_1,\underline{\theta}_2 \in \Theta$ and $F(\underline{x}_i;\underline{\theta}_1) = F(\underline{x}_i;\underline{\theta}_2)$, $i = 1,\ldots,s$, then $\underline{\theta}_1 = \underline{\theta}_2$.
2. The function $F(\underline{x};\underline{\theta})$ is continuous in $\underline{\theta}$ for each $\underline{\theta} \in \Theta$.

In this definition, continuity is with respect to the elements of $\underline{\theta}$, and $F(\underline{x};\underline{\theta})$ may be discrete, continuous, or absolutely continuous with respect to the elements of $\underline{x}$.

It will be desirable to extend the definition of a $G_kN_n(\mu,\sigma^2)$-sample. If all of the conditions necessary to constitute a sample with the property $G_kN_n(\mu,\sigma^2)$ are satisfied except that the underlying random variable has some probability distribution function in the family $D = \{F(x;\underline{\theta}):\underline{\theta} \in \Theta\}$, where $\Theta \subset E_q$, then the sample will be said to have the property $G_kD_n(\underline{\theta})$.

Suppose that the underlying random variable $\underline{X}$ is a $p \times 1$ vector of random variables defined in $\underline{\mathcal{X}} \subset E_p$, and suppose that the probability distribution function of $\underline{X}$ is in the family $D = \{F(\underline{x};\underline{\theta}):\underline{\theta} \in \Theta\}$, where $\Theta \subset E_q$. If each of the x_i-component axes is partitioned into k_i intervals

$$I_{ij} = (r_{i,j-1}, r_{i,j}],\ r_{i0} = -\infty < r_{i1} < \ldots < r_{ik_i} = \infty ,$$

$i = 1, \ldots, p$, then a random sample of size n, for which each component of $\underline{X}$ is grouped according to the component intervals, provides cell frequencies for each of the hyper-rectangles, i.e. for each of the sets which might be formed by a Cartesian product of the intervals $I_{1j(1)}, I_{2j(2)}, \ldots, I_{pj(p)}$, where $1 \leq j(i) \leq k_i$, $i = 1, \ldots, p$. Denote a particular hyper-rectangle by $R_{\underline{\nu}}$,

where $\underline{\nu}' = [j(1), j(2), \ldots, j(p)]$, and let $\Pr\{\underline{X} \in R_{\underline{\nu}}; \underline{\theta}\} = P_{\underline{\nu}}(\underline{\theta})$ and let $n_{\underline{\nu}}$ be the number of grouped observations in $R_{\underline{\nu}}$, where $\sum_{i,j} n_{\underline{\nu}} = n$. The sample will be said to have the property $G_{\underline{\nu}}D_n(\underline{\theta})$. The log-likelihood function for a $G_{\underline{\nu}}D_n(\underline{\theta})$-sample is

$$L(\underline{\theta}) = C + \sum_{i,j} n_{\underline{\nu}} \log\left(P_{\underline{\nu}}(\underline{\theta})\right),$$

where C is not a function of $\underline{\theta}$, and where the sum is understood to be performed for $j = 1, \ldots, k_i$ and $i = 1, \ldots, p$.

3.2 Preliminary Theorems

If the parameter space Θ is contained in E_q, then, in general, the closure of Θ must be considered to be a subset of the q-dimensional Cartesian product space of the extended real numbers, $\bar{E}_q$; see page 54 of Hewitt and Stromberg (1965). Denote the closure of Θ by $\bar{\Theta} = \Theta \cup \partial\Theta$. The following theorems are needed in subsequent developments:

A. If $\{\underline{\theta}_n\}$ is a sequence in Θ, then there exists some subsequence $\{\underline{\theta}_{n_k}\}$ of $\{\underline{\theta}_n\}$ which has limit $\underline{\beta} \in \bar{\Theta}$. (See Hewitt and Stromberg (1965), Theorems 6.37 and 6.43 .)

B. If $\{\underline{\theta}_n\}$ is a sequence in $\bar{\Theta}$ such that no subsequence $\{\underline{\theta}_{n_k}\}$ has limit other than $\underline{\alpha} \in \Theta$, then $\{\underline{\theta}_n\}$ has limit $\underline{\alpha} \in \Theta$. This conclusion follows from Theorem A above. Since there is at least one subsequence with limit $\underline{\beta} \in \bar{\Theta}$ and since no subsequence has limit other than $\underline{\alpha}$, then the sequence $\{\underline{\theta}_n\}$ has limit $\underline{\alpha}$.

Theorem 1

Let the following conditions be satisfied:

1. $\underline{X}$ is a p x 1 vector of random variables defined in $\underline{\mathcal{X}} \subset E_p$,

2. $\underline{\theta}$ is a q x 1 vector of parameters defined in $\Theta \subset E_q$,

3. $\{F(\underline{x};\underline{\theta}):\underline{\theta} \in \Theta\}$ is a continuous, parametric family of distribution functions,

4. $(\underline{x}_1, \ldots, \underline{x}_s) \in Q$ and $\{\underline{\theta}_n\}$ is a sequence in Θ,

5. $F(\underline{x}_i;\underline{\theta}_n) \to F(\underline{x}_i;\underline{\alpha})$, $i = 1, \ldots, s$ and $\underline{\alpha} \in \Theta$.

Then $\underline{\theta}_n \to \underline{\alpha}$.

Proof: Since $\{\underline{\theta}_n\}$ is a sequence in Θ, it follows from Theorem A that there exists a subsequence $\{\underline{\theta}_{n_k}\}$ such that $\underline{\theta}_{n_k} \to \underline{\beta} \in \bar{\Theta}$. Since $\{F(\underline{x};\underline{\theta}):\underline{\theta} \in \Theta\}$ is a continuous, parametric family of distribution functions, $F(\underline{x};\underline{\theta})$ is continuous in the elements of $\underline{\theta}$ and $F(\underline{x}_i;\underline{\theta}_{n_k}) \to F(\underline{x}_i;\underline{\beta})$, $i = 1, \ldots, s$, by the definition of continuity. But, $\{F(\underline{x}_i;\underline{\theta}_{n_k})\}$ is a subsequence of $\{F(\underline{x}_i;\underline{\theta}_n)\}$ and hence $F(\underline{x}_i;\underline{\theta}_{n_k}) \to F(\underline{x}_i;\underline{\alpha})$ for $i = 1, \ldots, s$. Using property 1 of the definition of a continuous, parametric family of distribution functions, it follows that since $F(\underline{x}_i;\underline{\beta}) = F(\underline{x}_i;\underline{\alpha})$, $i = 1, \ldots, s$, then $\underline{\alpha} = \underline{\beta}$. Thus any subsequence of $\{\underline{\theta}_n\}$ has limit $\underline{\alpha}$. Using Theorem B, we have that $\underline{\theta}_n \to \underline{\alpha}$.

Rao (1957) proved the following theorem which is given here as Lemma 1.

Lemma 1

If $a_1, a_2, \ldots, a_k$ and $b_1, b_2, \ldots, b_k$ are positive

constants such that $\sum_{i=1}^{k} a_i = \sum_{i=1}^{k} b_i$, then

$$\sum_{i=1}^{k} a_i \log\left(\frac{a_i}{b_i}\right) \geq 0$$

and the equality holds if and only if $a_i = b_i$, $i = 1, \ldots, k$.

The expectation operator, E, will bear no subscript in subsequent arguments. It will be understood that the expectation should be taken with respect to the joint distribution of the random variables in the sample. The information matrix for a sample of size one will be denoted by J_o. When $\frac{\partial^2 L(\underline{\theta})}{\partial \theta_i \partial \theta_j}$ is written, it is assumed that it exists and is finite for each $\underline{\theta} \in \Theta$.

Rao (1961, 1965) used Lemma 1 to prove the following theorem which will be used frequently in subsequent developments in this chapter.

Theorem 2

If $P_i(\underline{\theta})$, $i = 1, \ldots, k < \infty$, are the k cell probabilities of a multinomial distribution, where $\underline{\theta}$ is a q x 1 vector of parameters defined in $\Theta \subset E_q$ and n is the sample size, then if:

1. $P_i(\underline{\theta}) \neq 0$ for each $\underline{\theta} \in \Theta$, $i = 1, \ldots, k$,
2. $E\left[-\frac{\partial^2 L(\underline{\theta}_o)}{\partial \underline{\theta} \partial \underline{\theta}'}\right] = nJ_o$ is non-singular, where $\underline{\theta}_o$ is the true value of $\underline{\theta}$ in Θ and $L(\underline{\theta})$ is the logarithm of

the likelihood function for a sample of size n,*

3. For every $\delta > 0$ there exists an $\epsilon > 0$ such that

$$\inf_{|\underline{\theta} - \underline{\theta}_o| \geq \delta} S(\underline{\theta}_o, \underline{\theta}) \geq \epsilon \ , \ \underline{\theta}_o \text{ in } \Theta \ , \text{ where}$$

$$S(\underline{\theta}_o, \underline{\theta}) = \sum_{i=1}^{k} P_i(\underline{\theta}_o) \log\left(\frac{P_i(\underline{\theta}_o)}{P_i(\underline{\theta})}\right) \text{ and } |\underline{\theta} - \underline{\theta}_o| \text{ is the}$$

distance from $\underline{\theta}$ to $\underline{\theta}_o$ in E_q ,

then with probability one as $n \to \infty$ the value $\hat{\underline{\theta}}_n$ at which the likelihood function assumes its unique absolute maximum is a solution of the likelihood equations and $\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}_o)$ has the multivariate normal distribution with mean vector $\underline{0}$ and covariance matrix J_o^{-1} as its limiting distribution.**

Using the conclusion that, with probability one as $n \to \infty$, the likelihood function assumes its unique absolute maximum at $\hat{\underline{\theta}}_n$, a solution of the likelihood equations, one

---

* The conditions of this theorem can be relaxed; see Note 2 on page 298 of Rao (1965).

** In order to establish that Theorem 2 is actually proven by Rao, the following references should be made in Rao (1965).

1. See Note 1 on page 298.
2. See section 5c.2, particularly statement (iii) on page 286.
3. Statement (iii), page 286, together with the relation 5e.2.5 on page 296 establishes that J_o^{-1} is the appropriate covariance matrix for the limiting distribution.

can conclude that $\sqrt{n}(\underline{\theta}^*_n - \underline{\theta}_o)$ has the same limiting distribution as $\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}_o)$. Hence, both the strict and restricted maximum likelihood estimators would be asymptotically efficient.

The following observations should be made. In order to establish that

$$\inf_{|\underline{\theta} - \underline{\theta}_o| \geq \delta} S(\underline{\theta}_o, \underline{\theta}) \geq \epsilon$$

one must show only that if $\{\underline{\theta}_m\}$ is a sequence in $\bar{\Theta}$ and

$$\text{if } S(\underline{\theta}_o, \underline{\theta}_m) \to 0 \text{ , then } |\underline{\theta}_m - \underline{\theta}_o| \to 0 \text{ ;} \tag{3.1}$$

see the paragraph following assumption 1.1 on page 295 of Rao (1965). This follows from consideration of Lemma 1. It also follows from Lemma 1 that

$$\text{if } S(\underline{\theta}_o, \underline{\theta}_m) \to 0, \text{ then } P_i(\underline{\theta}_m) \to P_i(\underline{\theta}_o),\ i = 1, \ldots, k. \tag{3.2}$$

If one is concerned with estimating the parameters of a particular distribution using a grouped data sample, then conditions one and two of the preceding theorem must be examined for that particular distribution. In Theorems 3 and 4 following, sufficient conditions are established which, if satisfied by the cells defining the grouped data sample, will insure that condition 3 of Theorem 2 is satisfied.

Theorem 3

If X is a random variable defined in $\underline{X} \subset E_1$, and $\underline{\theta}$ is a q x 1 vector of parameters defined in $\Theta \subset E_q$ and if:

1. $D = \{F(x,\underline{\theta}) : \underline{\theta} \in \Theta\}$ is a continuous, parametric family of distribution functions, one of which, say $F(x;\underline{\theta}_o)$, is the probability distribution function of X,
2. the $G_k D_n(\underline{\theta})$-sample is defined by
   $I_i = (r_{i-1}, r_i]$, i = 1, ..., k such that $\sum_{i=1}^{k} P_i(\underline{\theta}) = 1$
   and there is a subset of $\{r_0, r_1, \ldots, r_k\}$,
   say $(\rho_1, \rho_2, \ldots, \rho_s)$, in Q for each $\underline{\theta} \in \Theta$,
3. $P_i(\underline{\theta}) \neq 0$, i = 1, ..., k for each $\underline{\theta} \in \Theta$,
4. $E\left[-\frac{\partial^2 L(\underline{\theta}_o)}{\partial \underline{\theta} \partial \underline{\theta}'}\right] = nJ_o$ is not singular,

then the strict and restricted maximum likelihood estimators are asymptotically efficient.

Proof: Conditions 3 and 4 satisfy the first two conditions of Theorem 2. Using observation (3.1) we must show that if $S(\underline{\theta}_o, \underline{\theta}_m) \to 0$, then $|\underline{\theta}_m - \underline{\theta}_o| \to 0$ in order to satisfy condition 3 of Theorem 2. If $S(\underline{\theta}_o, \underline{\theta}_m) \to 0$, then by observation (3.2) it follows that $P_i(\underline{\theta}_m) \to P_i(\underline{\theta}_o)$, i = 1, ..., k and hence $F(\rho_i;\underline{\theta}_m) \to F(\rho_i;\underline{\theta}_o)$, i = 1, ..., s.

Using Theorem 1 we see that $\underline{\theta}_m \to \underline{\theta}_o$ and hence $|\underline{\theta}_m - \underline{\theta}_o| \to 0$. Therefore, the conditions of Theorem 2 are satisfied and the strict and restricted maximum likelihood estimators are asymptotically efficient.

This theorem is extended easily to the multivariate case.

Theorem 4

If $\underline{X}$ is a p x 1 vector of random variables defined in E_p and $\underline{\theta}$ is a q x 1 vector of parameters defined in $\Theta \subset E_q$ and if:

1. $D = \{F(\underline{x}_i;\underline{\theta}) : \underline{\theta} \in \Theta\}$ is a continuous, parametric family of distribution functions, one of which, say $F(\underline{x};\underline{\theta}_o)$, is the probability distribution function of $\underline{X}$,
2. The $G_{\underline{\nu}}D_n(\underline{\theta})$-sample is defined by the points $\underline{r}_0, \underline{r}_1, \ldots, \underline{r}_k$, such that $\sum_{\underline{\nu}\in V} P_{\underline{\nu}}(\underline{\theta}) = 1$, where V is the set of all $\underline{\nu}$ defining the cells, and a subset of $\{\underline{r}_0, \underline{r}_1, \ldots, \underline{r}_k\}$, say $(\underline{\rho}_1, \underline{\rho}_2, \ldots, \underline{\rho}_s)$ is in Q for each $\underline{\theta} \in \Theta$,
3. $P_{\underline{\nu}}(\underline{\theta}) \neq 0$, $\underline{\nu} \in V$ for each $\underline{\theta} \in \Theta$,
4. $E\left[-\frac{\partial^2 L(\underline{\theta}_o)}{\partial\underline{\theta}\partial\underline{\theta}'}\right] = nJ_o$ is not singular,

then the strict and restricted maximum likelihood estimators

are asymptotically efficient.

Proof: Conditions 3 and 4 satisfy the first two conditions of Theorem 2. Therefore, we must show that if $S(\underline{\theta}_o, \underline{\theta}_m) \to 0$, then $|\underline{\theta}_m - \underline{\theta}_o| \to 0$. Using observation (3.2) we see that $P_{\underline{\nu}}(\underline{\theta}_m) \to P_{\underline{\nu}}(\underline{\theta}_o)$ for each $\underline{\nu} \in V$ and therefore $F(\underline{\rho}_i; \underline{\theta}_m) \to F(\underline{\rho}_i; \underline{\theta}_o)$, $i = 1, \ldots, s$. It follows from Theorem 1 that $|\underline{\theta}_m - \underline{\theta}_o| \to 0$. Hence the strict and restricted maximum likelihood estimators are asymptotically efficient.

The characteristics of the class of points Q, if it exists, will determine conditions which must be satisfied by the cells defining the grouped data sample in order for the restricted maximum likelihood estimator to be asymptotically efficient. We are primarily interested in the class Q only for the normal distribution, but a much wider class of parametric families of distribution functions has similar characteristics. The two theorems which follow establish some classes of distribution functions for which Q can be characterized. One must keep in mind that even after some class of points in Q is determined, it must still be shown that conditions 3 and 4 of either Theorem 3 or Theorem 4 are satisfied before the asymptotic efficiency of the strict and restricted maximum likelihood estimators is guaranteed.

Theorem 5

If X is a random variable defined in $\overline{\underline{X}} \subset E_1$ and if θ is a parameter defined in $\Theta \subset E_1$, then $D \equiv \{F(x;\theta):\theta \in \Theta\}$ is a continuous parametric family if the following conditions are satisfied.

1. For each $x \in \overline{\underline{X}}$, $\frac{\partial F(x;\theta)}{\partial \theta}$ is a continuous function of θ at each $\theta \in \Theta$,
2. If $\rho \in \overline{\underline{X}}$ and $0 < F(\rho;\theta) < 1$ for every $\theta \in \Theta$, then $|\frac{dF(\rho;\theta)}{d\theta}| > 0$ for every $\theta \in \Theta$.

Proof: Since the derivative of $F(\rho;\theta)$ is not zero and since $F(\rho;\theta)$ is a continuous function of θ, it follows that $F(\rho;\theta)$ is a strictly monotonic function of θ and hence if $F(\rho;\theta_1) = F(\rho;\theta_2)$, then $\theta_1 = \theta_2$. Therefore, the set Q contains any point $\rho \in \overline{\underline{X}}$ provided $0 < F(\rho;\theta) < 1$ for every $\theta \in \Theta$. The family D satisfies the conditions of a continuous, parametric family since condition 1 obviously implies that $F(x;\theta)$ is continuous in θ for each $\theta \in \Theta$.

Theorem 6

Under the same conditions stated in Theorem 5, the strict and restricted maximum likelihood estimators of θ obtained from a $G_k D_n(\theta)$-sample are asymptotically efficient provided $P_1(\theta), \ldots, P_k(\theta)$, $k \geq 2$, are all different from zero and $\sum_{i=1}^{k} P_i(\theta) = 1$ for every $\theta \in \Theta$.

Proof: Using Theorem 5 it follows that D is a continuous parametric family of distribution functions. Since $P_i(\theta) \neq 0$, $i = 1, \ldots, k$, $k \geq 2$, conditions 1-3 of Theorem 3 are satisfied. Note that $P_1(\theta) = F(r_1;\theta)$ and, by condition 2 of Theorem 5, $\frac{dF(r_1;\theta)}{d\theta} \neq 0$ for any $\theta \in \Theta$ and in particular $\frac{dF(r_1;\theta_o)}{d\theta} \neq 0$. All which remains to be shown is that condition 4 of Theorem 3 is satisfied, i.e. we must show that $J_o \neq 0$. But J_o is

$$J_o = \sum_{i=1}^{k} P_i^{-1}(\theta_o)\left(\frac{dP_i(\theta_o)}{d\theta}\right)^2$$

(see Rao (1965) page 295, Assumption 3). Since $\frac{dP_1(\theta_o)}{d\theta} \neq 0$, it follows that $J_o \neq 0$ and by Theorem 3 the strict and restricted maximum likelihood estimators are asymptotically efficient.

In many cases the large sample properties of the restricted maximum likelihood estimator obtained from grouped data samples can be verified using Theorem 6. The following three discrete distributions provide examples. Let

$$f(u;\theta) = \frac{e^{-\theta}\theta^u}{u!}, \quad u = 0, 1, \ldots,$$

$$f(v;\theta) = \theta(1 - \theta)^v, \quad v = 0, 1, \ldots, \text{ and}$$

$$f(w;N,\theta) = \binom{N}{w}\theta^w(1 - \theta)^{N-w}, \quad w = 0, 1, \ldots, N.$$

Using relations already enumerated by other authors (see Hughes (1962)) we find that:

$$\frac{dF(u;\theta)}{d\theta} = -f(u;\theta) < 0 \text{ for } 0 < \theta < \infty,$$

$$\frac{dF(v;\theta)}{d\theta} = (1 - \theta)^{-1}(v + 1)[1 - F(v;\theta)] > 0 \text{ for } 0 < \theta < 1, \text{ and}$$

$$\frac{dF(w;N,\theta)}{d\theta} = -Nf(w;N-1,\theta) < 0 \text{ for } 0 < \theta < 1 \text{ and } w < N.$$

Therefore, by Theorem 6, if one has a grouped data sample from one of these three distributions with at least two cells, such that $P_1(\theta)$, $P_2(\theta)$, ..., $P_k(\theta)$ are not zero (and sum to one) for every θ in the appropriate parameter space, then the strict and restricted maximum likelihood estimators are asymptotically efficient.

There are many other one-parameter distributions which satisfy the conditions of Theorem 6; however, it is the multiparameter cases which have received little development. As we have seen, the normal distribution with both parameters unknown was not considered by Kulldorff, who established the asymptotic properties of the maximum likelihood estimators from grouped data samples when only one parameter is unknown.

In the subsequent developments in this chapter our attention is focused on parametric families of absolutely continuous probability distribution functions. The parameters will be limited to first and second moments of the distributions.

The following lemma is stated without proof.

Lemma 2

If $F(x)$ is the probability distribution function of the random variable X, defined in $\overline{X} \subset E_1$, and if $F(x)$ is a strictly increasing function of $x \in \overline{X}$, then the probability distribution function of $Y = \frac{X - a}{b}$, where $-\infty < a < \infty$ and $0 < b < \infty$, say $F_1(y)$, is a strictly increasing function of $y \in \overline{Y} \subset E_1$, where $\overline{Y} = \{Y : Y = \frac{X-a}{b} \text{ and } X \in \overline{X}\}$.

In the theorem which follows the vector $\underline{\theta}' = [\theta_1, \theta_2]$ is defined to be $[EX, E(X - \theta_1)^2] = [\mu, \sigma^2]$ and Θ is defined to be the set $\{\underline{\theta} : a < \theta_1 < b, 0 < \theta_2 < \infty\}$, where $a < b$ and a and b are points in $E_1^* \equiv [-\infty, \infty]$. The symbols θ_1 and μ and the symbols θ_2 and σ^2 will be used interchangeably.

Theorem 7

Let $D \equiv \{F(x;\underline{\theta}) : \underline{\theta} \in \Theta\}$ be a parametric family of distribution functions of the random variable X defined in $\overline{X} \in E_1$. If, for each $\underline{\theta} \in \Theta$, $F(x;\underline{\theta})$ is continuous at $\underline{\theta}$ and is a strictly increasing function of $x \in \overline{X}$, then D is a continuous parametric family of distribution functions.

Proof: Suppose that $\underline{\theta}_1$ and $\underline{\theta}_2$ are points in Θ and that ρ_1 and ρ_2 are two distinct points in $\overline{X}$ such that $F(\rho_1;\underline{\theta})$ and $F(\rho_2;\underline{\theta})$ are greater than zero and less than one for every

$\underline{\theta} \in \Theta$. Define the random variable $Y = (X - EX)\left(E(X-E(X))^2\right)^{-\frac{1}{2}}$. Let Y have the probability distribution function $F_1(y)$. It follows that if $F(\rho_1;\underline{\theta}_1) = F(\rho_1;\underline{\theta}_2)$, then $F_1\left(\frac{\rho_1 - \mu_1}{\sigma_1}\right) = F_1\left(\frac{\rho_1 - \mu_2}{\sigma_2}\right)$ and if $F(\rho_2;\underline{\theta}_1) = F(\rho_2;\underline{\theta}_2)$, then $F_1\left(\frac{\rho_2 - \mu_1}{\sigma_1}\right) = F_1\left(\frac{\rho_2 - \mu_2}{\sigma_2}\right)$. Applying Lemma 2 to obtain the strict monotonicity of $F_1(y)$, it follows that $\frac{\rho_1 - \mu_1}{\sigma_1} = \frac{\rho_1 - \mu_2}{\sigma_2}$ and $\frac{\rho_2 - \mu_1}{\sigma_1} = \frac{\rho_2 - \mu_2}{\sigma_2}$, and therefore $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$. By assumption, $F(x;\underline{\theta})$ is continuous at each $\underline{\theta} \in \Theta$ and therefore D is a continuous, parametric family of distribution functions. The set Q consists of any pair of distinct points ρ_1, ρ_2 in $\underline{X}$ such that $F(\rho_1;\underline{\theta})$ and $F(\rho_2;\underline{\theta})$ are greater than zero and less than one for every $\underline{\theta} \in \Theta$.

The general multivariate extension of Theorem 7 is far more complicated than Theorem 7. The following corollary is presented as a very weak multivariate extension of Theorem 7. In the corollary the parameter space Θ is the set of $\underline{\theta}' = [\underline{\mu}', \underline{\sigma}^{2\prime}]$, where $E\underline{X} = \underline{\mu}$ and $\underline{\sigma}^2 = [E(X_1 - \mu_1)^2, \ldots, E(X_p - \mu_p)^2]'$, and $0 < E(X_i - \mu_i)^2 < \infty$.

Corollary 1

Let $D = \{F(\underline{x};\underline{\theta}):\underline{\theta} \in \Theta\}$ be a parametric family of distribution functions of the random variable $\underline{X}$ defined in $\overline{\underline{X}} \subset E_p$. If, for each $\underline{\theta} \in \Theta$, $F(\underline{x};\underline{\theta})$ is continuous at $\underline{\theta}$ and is a strictly increasing function of each x_i, $i = 1, \ldots, p$, then D is a continuous parametric family of distribution functions.

Proof: We must exhibit a collection of points $\underline{\rho}_i$, $i = 1, \ldots, 2p$, in Q. Let $\underline{\rho}_1' = [a, \infty, \ldots, \infty]$ and $\underline{\rho}_2' = [b, \infty, \ldots, \infty]$, $a \neq b$, be two points such that $F(a, \infty, \ldots, \infty)$ and $F(b, \infty, \ldots, \infty)$ are not equal and both are between zero and one for all $\underline{\theta} \in \Theta$. Let $\underline{\theta}^*$ and $\underline{\theta}^{**}$ be two points in Θ. If $F(\underline{\rho}_1;\underline{\theta}^*) = F(\underline{\rho}_1;\underline{\theta}^{**})$ and $F(\underline{\rho}_2;\underline{\theta}^*) = F(\underline{\rho}_2;\underline{\theta}^{**})$, then using Theorem 7 it follows that $\mu_1^* = \mu_1^{**}$ and $\sigma_1^{2*} = \sigma_1^{2**}$. Similar pairs of points, i.e. having only one finite coordinate, establish that $\underline{\theta}^* = \underline{\theta}^{**}$. The total number of points required is 2p.

In the next section it will be shown that the strict and restricted maximum likelihood estimators obtained from a $G_k N_n(\mu,\sigma^2)$-sample are asymptotically efficient, using Theorem 7 to satisfy the conditions of Theorem 3. It also will be shown that the strict and restricted maximum likelihood estimators in the case of grouped data samples from a bivariate normal distribution are asymptotically

efficient.

3.3 Asymptotic Properties of the Strict and Restricted Maximum Likelihood Estimators in the Case of a Normal Distribution

In Theorem 8 we shall prove that the strict and restricted maximum likelihood estimators for the mean and variance of a normal distribution obtained from a $G_k N_n(\mu,\sigma^2)$-sample, $k \geq 3$, are asymptotically efficient. In order to satisfy the conditions of Theorem 3 we must show that the information matrix for a sample of size one, J_o, is non-singular. The following lemma will be used in Theorem 8 to establish that J_o is not singular provided $k \geq 3$. The matrix J_o can be computed using Theorem 7 in section 2.2.5, but using equation (2.34) and the fact that $\sum_{i=1}^{k} P_i(\underline{\theta}_o) Z_{ji} = 0$, $j = 0, 1, \ldots,$ (see definition 5 in section 2.3.3), it follows that $J_o = c \sum_{i=1}^{k} P_i(\underline{\theta}_o) \begin{bmatrix} Z_{0i} \\ Z_{1i} \end{bmatrix} [Z_{0i}, Z_{1i}]$, where c is some non-zero constant. We shall prove that $\sum_{i=1}^{k} P_i(\underline{\theta}_o) \begin{bmatrix} Z_{0i} \\ Z_{1i} \end{bmatrix} [Z_{0i}, Z_{1i}]$ is not singular provided $k \geq 3$.

Lemma 3

If $k \geq 3$, then $\sum_{i=1}^{k} P_i(\underline{\theta}_o) \begin{bmatrix} Z_{0i} \\ Z_{1i} \end{bmatrix} [Z_{0i}, Z_{1i}]$ is not singular.

Proof: It is sufficient to show that the set of vectors $\{P_i^{\frac{1}{2}}(\underline{\theta}_o)\begin{bmatrix} Z_{0i} \\ Z_{1i} \end{bmatrix}\}_{i=1}^{k}$ span the space E_2. This can be done by showing that there are at least two non-null vectors, $[Z_{0i}, Z_{1i}]$ and $[Z_{0j}, Z_{1j}]$, which are such that one is not a scalar multiple of the other. The following two cases are considered.

Case 1. $k = 3$ and $[Z_{02}, Z_{12}] = [0,0]$. If $Z_{02} = 0$, then $f(t_1) = f(t_2)$ and $-t_1 = t_2 > 0$. Note that $\begin{bmatrix} Z_{01} \\ Z_{11} \end{bmatrix} = \begin{bmatrix} -f(t_1) \\ -t_1 f(t_1) \end{bmatrix}$ and $\begin{bmatrix} Z_{03} \\ Z_{13} \end{bmatrix} = \begin{bmatrix} f(t_2) \\ t_2 f(t_2) \end{bmatrix}$. Since $-t_1 = t_2$, it is obvious that the second vector is not a scalar multiple of the other.

Case 2. $k \geq 3$ and $Z_{02} \neq 0$. (Note that the case where $Z_{0(k-1)}$ is not zero would be accompanied by the same argument using $Z_{i(k-1)}$ and Z_{ik}, $i = 0, 1$.) Since $Z_{02} \neq 0$, then $t_1 \neq t_2$. Suppose that $\begin{bmatrix} Z_{01} \\ Z_{11} \end{bmatrix} = c\begin{bmatrix} Z_{02} \\ Z_{12} \end{bmatrix}$, for some c. Then:

$$\begin{bmatrix} -f(t_1) \\ -t_1 f(t_1) \end{bmatrix} = c\begin{bmatrix} f(t_1) - f(t_2) \\ t_1 f(t_1) - t_2 f(t_2) \end{bmatrix} \text{ and}$$

$$(1 + c) f(t_1)\begin{bmatrix} 1 \\ t_1 \end{bmatrix} = f(t_2)\begin{bmatrix} 1 \\ t_2 \end{bmatrix}.$$

Since $t_1 \neq t_2$ there is no c for which this equality holds. If $Z_{02} = 0$, then either Case 1 is applicable or $Z_{i(k-1)} \neq 0$ and, as noted, the proof in Case 2 would suffice using t_{k-1} and t_{k-2}.

Theorem 8

The strict and restricted maximum likelihood estimators obtained from a $G_kN_n(\mu,\sigma^2)$-sample, $k \geq 3$, are asymptotically efficient.

Proof: From Theorem 7 it follows that the parametric family of univariate normal probability distribution functions is a continuous, parametric family with Q consisting of any two distinct points in $(-\infty,\infty)$. Since $k \geq 3$, conditions 1 and 2 of Theorem 3 are satisfied. By the definition of a $G_kN_n(\mu,\sigma^2)$-sample (see section 2.3.1) the length of each interval is greater than zero and therefore $P_i(\underline{\theta}) \neq 0$, $i = 1, \ldots, k$ and condition 3 of Theorem 3 is satisfied. The conclusion of this theorem follows from Theorem 3 if it can be shown that J_o is non-singular when $k \geq 3$. Applying Lemma 3 it follows that J_o is not singular. Therefore, by Theorem 3, the strict and restricted maximum likelihood estimators obtained from a $G_kN_n(\mu,\sigma^2)$-sample, $k \geq 3$, are asymptotically efficient.

Corollary 1 could be used to prove that the strict and

restricted maximum likelihood estimators of the means and variances of a bivariate normal distribution obtained from a grouped data sample, $k_1 \geq 3$, $k_2 \geq 3$, when ρ is known are asymptotically efficient. The more realistic case is that of a grouped data sample from a bivariate normal distribution with all five parameters unknown. Establishing that the information matrix J_o is non-singular is similar to, but much more intricate than, the development in Lemma 3 and therefore it will be assumed in the following theorem. One might expect that if at least three intervals are required on each axis to establish that the maximum likelihood estimators for the four parameters μ_1, μ_2, σ_1^2, and σ_2^2 are asymptotically efficient, then either k_1 or k_2 might have to be greater than three in the case of five unknown parameters. The need for this condition is established in Theorem 9.

In Theorem 9 it is assumed that $\underline{X}' = [X_1, X_2]$ has the bivariate normal density function $f(\underline{x};\underline{\theta})$, where $\underline{\theta}$ is the vector containing the elements of the two matrices $E\underline{X} = \underline{\mu} = [\mu_1, \mu_2]'$ and $E(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})' = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$, $\sigma_{12} = \sigma_{21}$. The grouped data sample is defined by $I_{ij} = (r_{i,j-1}, r_{i,j}]$, $j = 1, \ldots, k_i$, $i = 1, 2$ and $\min(k_1, k_2) \geq 3$ and $\max(k_1, k_2) > 3$, (see section 3.1). Since a bivariate distribution is being considered, let the cell

frequency of $I_{1i} \times I_{2j} \equiv R_{ij}$ be n_{ij} and let $\Pr\{\underline{X} \in R_{ij};\underline{\theta}\} = P_{ij}(\underline{\theta})$. If the sequence $\{\underline{\theta}_m\}$ is considered, then each element of $\underline{\theta}_m$ will be designated by adding the subscript m to each element, e.g. $\mu_{1,m}$ and $\sigma_{12,m}$. If $\underline{\theta}_o$ is a particular value of $\underline{\theta}$ in

$$\Theta \equiv \{(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\sigma_{12}): -\infty<\mu_1,\mu_2<\infty,\ 0<\sigma_1^2,\sigma_2^2<\infty,\ |\sigma_{12}|<\sigma_1\sigma_2\},$$

then the subscript o is added to each element of $\underline{\theta}_o$, e.g. $\mu_{2,o}$. The parameter ρ will be used interchangeably with $\sigma_{12}/\sigma_1\sigma_2$.

The lemma which follows will be needed in the proof of Theorem 9.

Lemma 4

If $\underline{X}$ has the bivariate normal density function

$$f(\underline{x};\underline{\theta}) = (2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})^{-1}\exp\left(-\frac{1}{2(1-\rho^2)}\left[x_1^{*2} - 2\rho x_1^* x_2^* + x_2^{*2}\right]\right),$$

where $x_i^* = \frac{x_i - \mu_i}{\sigma_i}$, $i = 1, 2$, and $\rho^2 < 1$, then

1. $W = \frac{X_1^* - \rho X_2^*}{\sqrt{1-\rho^2}}$ and X_2^* are distributed independently with density functions $f(w)$ and $f(x_2^*)$, where

   $$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}, \quad -\infty < t < \infty \quad \text{and}$$

2. $\Pr\{X_1 < r_1 \text{ and } r_2 < X_2 < r_3\} = \Pr\{r_2 < X_2 < r_3\}Q,$

where $Q = \int_{-\infty}^{\lambda} f(w)dw$, $\lambda = \dfrac{r_1^* - \min(\rho r_2^*, \rho r_3^*)}{\sqrt{1-\rho^2}}$,

$r_1^* = \dfrac{r_1 - \mu_1}{\sigma_1}$, $r_2^* = \dfrac{r_2 - \mu_2}{\sigma_2}$, and $r_3^* = \dfrac{r_3 - \mu_2}{\sigma_2}$.

Proof: We first complete the square in the exponent of $f(\underline{x};\underline{\theta})$:

$$\frac{-1}{2(1-\rho^2)}[x_1^{*2} - 2\rho x_1^* x_2^* + x_2^{*2}] = \frac{-1}{2(1-\rho^2)}[(x_1^* - \rho x_2^*)^2 + (1-\rho^2)x_2^{*2}] .$$

If we let $W = \dfrac{X_1^* - \rho X_2^*}{\sqrt{1-\rho^2}}$, then $dx_1^* = (\sqrt{1-\rho^2})dw$ and therefore

$$f(w, x_2^*) = f(w)f(x_2^*), \quad -\infty < x_2^* < \infty \text{ and } -\infty < w \leq \frac{x_1^* - \rho x_2^*}{\sqrt{1-\rho^2}} .$$

Thus, conclusion 1 is established.

If $r_2^* < X_2^* < r_3^*$, then $-\infty < W < \dfrac{X_1^* - \min(\rho r_2^*, \rho r_3^*)}{\sqrt{1-\rho^2}}$.

From this we conclude that

$$\int_{-\infty}^{r_1} \int_{r_2}^{r_3} f(\underline{x};\underline{\theta})dx_1 dx_2 = \int_{-\infty}^{\lambda} f(w)dw \int_{r_2^*}^{r_3^*} f(t)dt$$ and therefore

conclusion 2 is established.

Theorem 9

Assuming that the information matrix J_o is non-singular, the strict and restricted maximum likelihood estimators of the parameters of the bivariate normal density function from a grouped data sample, $\min(k_1, k_2) \geq 3$ and $\max(k_1, k_2) > 3$,

are asymptotically efficient.

Proof: Since the area of each R_{ij}, defined previously, is not zero, it follows that $P_{ij}(\underline{\theta}_o) \neq 0$ and it was assumed that $|J_o| \neq 0$. Therefore conditions 1 and 2 of Theorem 2 are satisfied. In this case

$$S(\underline{\theta}_o, \underline{\theta}) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} P_{ij}(\underline{\theta}_o) \log\left(\frac{P_{ij}(\underline{\theta}_o)}{P_{ij}(\underline{\theta})}\right) .$$

Using statement (3.1) and Lemma 1, it must be shown that if $P_{ij}(\underline{\theta}_m) \to P_{ij}(\underline{\theta}_o)$, $i = 1, \ldots, k_1$ and $j = 1, \ldots, k_2$, then $|\underline{\theta}_m - \underline{\theta}_o| \to 0$. If $P_{ij}(\underline{\theta}_m) \to P_{ij}(\underline{\theta}_o)$, then

$$\sum_{j=1}^{k_2} P_{ij}(\underline{\theta}_m) \to \sum_{j=1}^{k_2} P_{ij}(\underline{\theta}_o).$$

Therefore

$$\sum_{j=1}^{k_2} P_{1j}(\underline{\theta}_m) \to \sum_{j=1}^{k_2} P_{1j}(\underline{\theta}_o) = \Pr\{X_1 \leq r_{1,1}\}$$

and

$$\sum_{j=1}^{k_2} P_{2j}(\underline{\theta}_m) \to \sum_{j=1}^{k_2} P_{2j}(\underline{\theta}_o) = \Pr\{X_1 \leq r_{1,2}\} .$$

By Theorem 7 it follows that $\mu_{1,m} \to \mu_{1,o}$ and $\sigma^2_{1,m} \to \sigma^2_{1,o}$. The same argument, reversing the roles of i and j, indicates that $\mu_{2,m} \to \mu_{2,o}$ and $\sigma^2_{2,m} \to \sigma^2_{2,o}$. The theorem is completed

if it can be shown that $\sigma_{12,m} \to \sigma_{12,o}$, or equivalently that $\rho_m \to \rho_o$. Let $f(\underline{x};\rho)$ denote the bivariate normal density function of the random variable $\underline{X}$, where $E\underline{X} = \underline{\mu}_o$ and $E(X_1 - \mu_{1,o})^2 = \sigma^2_{1,o}$ and $E(X_2 - \mu_{2,o})^2 = \sigma^2_{2,o}$ and where the correlation coefficient is ρ. Let

$$X^*_i = \frac{X_i - \mu_{i,o}}{\sigma_{i,o}} \text{ and } r^*_{i,j} = \frac{r_{i,j} - \mu_{i,o}}{\sigma_{i,o}}, \; j = 1, \ldots, k_i,$$

$i = 1, 2$. Two assumptions are chosen for the argument which follows. Similar arguments establish the theorem for the other cases. Assume that:

1. $\rho_o \geq 0$
2. $k_2 > 3$.

Since $P_{ij}(\underline{\theta}_m) \to P_{ij}(\underline{\theta}_o)$, $i = 1, \ldots, k_1$, $j = 1, \ldots, k_2$, then applying Lemma 4 it follows that:

$$\int_{-\infty}^{\lambda_{i,j,m}} f(w)dw \int_{r^*_{2,j}}^{r^*_{2,j+1}} f(t)dt \to \int_{-\infty}^{\lambda_{i,j,o}} f(w)dw \int_{r^*_{2,j}}^{r^*_{2,j+1}} f(t)dt,$$

$i = 1, \ldots, k_1$ and $j = 1, \ldots, k_2$, where

$$\lambda_{i,j,m} = \frac{r^*_{1,i} - \min(\rho_m r^*_{2,j}, \rho_m r^*_{2,j+1})}{\sqrt{1-\rho^2_m}}$$

and

$$\lambda_{i,j,o} = \frac{r^*_{1,i} - \min(\rho_o r^*_{2,j}, \rho_o r^*_{2,j+1})}{\sqrt{1-\rho_o^2}} \quad .$$

Since the integral of f(w) is a strictly increasing function of its upper limit, it follows that

$$\lambda_{i,j,m} \to \lambda_{i,j,o} \text{ , } i = 1, 2 \text{ and } j = 1, 2.$$

Consider $$\lambda_{1,j,m} - \lambda_{2,j,m} = \frac{r^*_{1,1} - r^*_{1,2}}{\sqrt{1-\rho_m^2}}$$

and

$$\lambda_{1,j,o} - \lambda_{2,j,o} = \frac{r^*_{1,1} - r^*_{1,2}}{\sqrt{1-\rho_o^2}} \quad .$$

Since $(\lambda_{1,j,m} - \lambda_{2,j,m}) \to (\lambda_{1,j,o} - \lambda_{2,j,o})$, it follows that $|\rho_m| \to \rho_o$.

If $\rho_o = 0$, then $\rho_m \to \rho_o$ and the conclusion of the theorem follows without using the fact that $k_2 > 3$; compare this result with Corollary 1 to Theorem 7.

If $\rho_o > 0$, then, since $\sqrt{1-\rho_m^2} \to \sqrt{1-\rho_o^2}$ and since

$r^*_{2,j} < r^*_{2,j+1}$,

$\min(\rho_m r^*_{2,j}, \rho_m r^*_{2,j+1}) \to \min(\rho_o r^*_{2,j}, \rho_o r^*_{2,j+1}) = \rho_o r^*_{2,j}$,

$j = 1, \ldots, k_2-1$.

Since $|\rho_m| \to \rho_o$, either $\rho_m \to \rho_o$ or there is a subsequence, $\{\rho_s\}$, of negative terms in the sequence $\{\rho_m\}$ such that $\rho_s \to -\rho_o$. If each $\rho_s < 0$ and if $\{\rho_s\}$ is a subsequence of $\{\rho_m\}$, then $\{\min(\rho_s r^*_{2,j}, \rho_s r^*_{2,j+1})\} = \{\rho_s r^*_{2,j+1}\}$, $j = 0, \ldots, k-2$, is a subsequence of (and has the same limit as) $\{\min(\rho_m r^*_{2,j}, \rho_m r^*_{2,j+1})\}$. Therefore, $(-\rho_o) r^*_{2,j+1} = \rho_o r^*_{2,j}$, $j = 1, \ldots, k_2-2$. If $k_2 = 3$ and $r^*_{2,1} = -r^*_{2,2}$, then there might be a subsequence such that $\rho_s \to -\rho_o$, but since $k_2 > 3$ $r^*_{2,j+1}$ cannot be equal to $-r^*_{2,j}$, $j = 1, 2, \ldots, k_2-2$. Therefore $\rho_m \to \rho_o$. This completes the proof.

See the Appendix for an iterative procedure with which the maximum likelihood estimators of these five parameters might be obtained. An iterative procedure is also developed which might be used to obtain the maximum likelihood estimators of these five parameters if one of the variates were grouped and the other were not grouped.

CHAPTER IV

SUMMARY

In this thesis the method of maximum likelihood is employed to estimate the parameters of probability distribution functions when grouped data only are available. A grouped data sample of size n is a sample in which the only information available is that there are m_i observations less than or equal to r_i, where $-\infty < r_0 < r_1 < \ldots < r_k = \infty$. The intervals $(r_{i-1}, r_i]$ are called cells and the cell frequencies are denoted by $n_i = m_i - m_{i-1}$, where $\sum_{i=1}^{k} n_i = n$.

In Chapter II grouped data samples from univariate normal distributions are considered. The method of maximum likelihood is used to estimate the parameters of the normal density function when grouped data only are available.

The first case which is considered is that in which only the mean of the normal distribution is unknown. The necessary and sufficient conditions for the existence of a unique solution of the likelihood equation are that $n_1 \neq n$ and $n_k \neq n$.

The second case considered is that in which only the

variance of the normal distribution is unknown. It is shown that if neither $n_1 + n_k = n$ nor $n_i + n_{i+1} = n$, $i = 1, \ldots, k-1$, then there is a unique solution of the likelihood equation. An iterative method of solution is developed which is similar to that used in the case in which only the mean is unknown. The iterative procedure is shown to provide a sequence which converges to the unique root of the likelihood equation if there is a unique root.

The third case which is considered is that case in which both the mean and the variance are unknown. It is shown that if neither $n_1 + n_k = n$ nor $n_i + n_{i+1} = n$, $i = 1, \ldots, k-1$, then there is a joint solution of the likelihood equations at which the likelihood function assumes its absolute maximum. Numerical results are presented which allow the determination of a region in the parameter space in which there can be at most one joint solution of the likelihood equations, and if there is a solution in this region, then it is a point at which the likelihood function has a relative maximum. A search to find some sample configuration which might provide two joint solutions of the likelihood equations indicates that a unique solution should be expected if the conditions for the existence of a solution are satisfied. A procedure is developed which can be used to locate multiple roots if

they might exist.

Several numerical methods of solution of the likelihood equations are studied. One of these methods is shown to provide a sequence which converges to the solution of the likelihood equations if there is a unique solution. This method of solution is found to have about the same rate of convergence as a modified version of the method of successive approximations.

Simply computed, consistent estimators are developed for use as starting values with each of the iterative methods of solution which is considered.

The Appendix contains a discussion of the major theorem found in Hughes (1962), in which he established sufficient conditions for the convergence of Hughes' method of solution. This method is shown to be very similar to the method of successive approximations in each of the three cases discussed in Chapter II. The conditions of Hughes' theorem are shown to be satisfied by a rather limited class of distributions. In the Appendix Hughes' method is extended to the case of a grouped data sample from a bivariate normal distribution and to the case of a sample from a bivariate normal distribution in which only one of the variates is grouped into cells.

In Chapter III the large-sample properties of the

maximum likelihood estimators obtained from grouped data samples are considered. Kulldorff (1958a, 1958b) has presented the conditions which will insure the asymptotic efficiency of the maximum likelihood estimators in the cases where only the mean or the variance is unknown. A q x 1 vector of estimators $\hat{\underline{\theta}}_n$ obtained from a sample of size n is defined to be asymptotically efficient for the true value of the parameter $\underline{\theta}_o$ if $\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}_o)$ has as its limiting distribution the q-variate normal distribution with mean vector $\underline{0}$ and covariance matrix equal to the inverse of the information matrix for a sample of size one.

In Chapter III it is shown that if there are at least three groups (cells) defining the grouped data sample, then the maximum likelihood estimators for the mean and the variance are asymptotically efficient. The theorems which are developed to establish this large-sample property are shown to have a much wider application. Other distributions are considered, such as the Poisson and bivariate normal distributions, and sufficient conditions are established which must be satisfied by the cells defining the sample in order for the maximum likelihood estimators to be asymptotically efficient.

It should be mentioned that there are at least two major areas of research which deserve further study. The first of

these might be called "grouped data inference." Numerical and theoretical studies could be made in order to ascertain the large- and small-sample properties of hypothesis testing, confidence intervals, and prediction using maximum likelihood as well as other methods when grouped data only are available. Since the distribution of the maximum likelihood estimators is usually very difficult or impossible to determine, numerical approaches to describe the properties of the likelihood ratio tests of hypotheses might provide very useful results. There are many other similar problems, particularly in multivariate cases, which apparently have not been studied at all.

The second area which might provide fruitful results is the general theory of maximum likelihood estimation. The present state in which one finds this theory might be described best by saying that there is a group of somewhat unrelated theorems, each with its special application. If the conditions and conclusions of many of these theorems were studied simultaneously, a few theorems might be developed which would provide a more complete, unified treatment of the theory of maximum likelihood estimation. This development would seem to be a great contribution to the present theory of estimation of parameters.

BIBLIOGRAPHY

Abramowitz, Milton, and Stegun, Irene A., editors (1964), Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Washington, D.C., United States Department of Commerce, National Bureau of Standards Applied Mathematics Series • 55.

Barnett, V.D. (1966), "Evaluation of the maximum likelihood estimator where the likelihood equation has multiple roots," Biometrika, Vol. 53, pp. 151-166.

Birnbaum, Z.W. (1942), "An inequality for Mill's ratio," Annals of Mathematical Statistics, Vol. 13, pp. 245-246.

Birnbaum, Z.W. (1950), "Effect of linear truncation on a multinormal population," Annals of Mathematical Statistics, Vol. 21, pp. 272-279.

Bliss, C.I. (1937), "The calculation of the time-mortality curve," Appendix by W.L. Stevens: "The truncated normal distribution," Annals of Applied Biology, Vol. 24, pp. 815-852.

Bowen, Jacob Van, Jr. (1966), Some Properties of Conditional Distributions of a Special Type, M.S. Thesis, Virginia Polytechnic Institute.

Clark, Frank Eugene (1957), "Truncation to meet requirements on means," Journal of the American Statistical Association, Vol. 52, pp. 527-536.

Cohen, A.C., Jr. (1949), "On estimating the mean and standard deviation of truncated normal distributions," Journal of the American Statistical Association, Vol. 44, pp. 518-525.

Cohen, A.C., Jr. (1950), "Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples," Annals of Mathematical Statistics, Vol. 21, pp. 557-569.

Cohen, A.C., Jr. (1955), "Restriction and selection in samples from bivariate normal distributions," Journal of the American Statistical Association, Vol. 50, pp. 884-893.

Cohen, A.C., Jr. (1957), "On the solution of estimating equations for truncated and censored samples from normal populations," Biometrika, Vol. 44, pp. 225-236.

Cramér, Harald (1946), Mathematical Methods of Statistics, Princeton, Princeton University Press.

Des Raj (1953), "On estimating the parameters of bivariate normal populations from doubly and singly, linearly truncated samples," Sankhyā, Vol. 12, pp. 277-290.

Doss, S.A.D.C. (1962), "On uniqueness and maxima of the roots of likelihood equations under truncated and censored sampling from normal populations," Sankhyā, Vol. 24, pp. 355-362.

Fisher, R.A. (1925), "Theory of statistical estimation," Proceedings of the Cambridge Philosophical Society, Vol. 22, pp. 700-725.

Fisher, R.A. (1931), "The truncated normal distribution," British Association for the Advancement of Science, Math. Tables, I, pp. XXXIII-XXXIV.

Ford, L.R. (1925), "The solution of equations by the method of successive approximations," American Mathematical Monthly, Vol. 32, pp. 272-287.

Fox, Augustus H. (1963), Fundamentals of Numerical Analysis, New York, The Ronald Press Company.

Gjeddebaek, N.F. (1949), "Contribution to the study of grouped observations. Application of the method of maximum likelihood in case of normally distributed observations," Skandinavisk Aktuarietidskrift, Vol. 32, pp. 135-159.

Gjeddebaek, N.F. (1956), "Contribution to the study of grouped observations. II. Loss of information caused by grouping of normally distributed observations," Skandinavisk Aktuarietidskrift, Vol. 39, pp. 154-159.

Gjeddebaek, N.F. (1957), "Contribution to the study of grouped observations. III. The distribution of estimates of the mean," Skandinavisk Aktuarietidskrift, Vol. 40, pp. 20-25.

Gjeddebaek, N.F. (1959a), "Contribution to the study of grouped observations. IV. Some comments on simple estimates," Biometrics, Vol. 15, pp. 433-439.

Gjeddebaek, N.F. (1959b), "Contribution to the study of grouped observations. V. Three-class grouping of normal observations," Skandinavisk Aktuarietidskrift, Vol. 42, pp. 194-207.

Gjeddebaek, N.F. (1961), "Contribution to the study of grouped observations. VI.," Skandinavisk Aktuarietidskrift, Vol. 44, pp. 55-73.

Gordon, Robert D. (1941), "Values of Mill's ratio of area to bounding ordinate and of the normal probability integral for large values of the argument," Annals of Mathematical Statistics, Vol. 12, pp. 364-366.

Gupta, A.K. (1952), "Estimation of the mean and standard deviation of a normal population from a censored sample," Biometrika, Vol. 39, pp. 260-273.

Hald, A. (1949), "Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point," Skandinavisk Aktuarietidskrift, Vol. 32, pp. 119-134.

Hartley, H.O. (1958), "Maximum likelihood estimation from incomplete data," Biometrics, Vol. 14, pp. 174-194.

Hewitt, Edwin, and Stromberg, Karl (1965), Real and Abstract Analysis, New York, Springer-Verlag, Inc.

Hughes, Edwin Joseph (1962), Maximum Likelihood Estimation of Distribution Parameters from Incomplete Data, Ph.D. Thesis, Iowa State University.

Huzurbazar, V.S. (1947-1949), "The likelihood equation, consistency and the maxima of the likelihood function," Annals of Eugenics, Vol. 14, pp. 185-200.

Huzurbazar, V.S. (1955), "On successive approximations to the maximum likelihood estimator," Journal of the University of Poona: Science Section, Vol. 12, pp. 96-97.

Kale, B.K. (1961), "On the solution of the likelihood equation by iteration processes," Biometrika, Vol. 48, pp. 452-456.

Kale, B.K. (1962), "On the solution of likelihood equations by iteration processes. The multiparameter case," Biometrika, Vol. 49, pp. 479-486.

Kale, B.K. (1966), "Approximations to the maximum-likelihood estimator using grouped data," Biometrika, Vol. 53, pp. 282-284.

Kendall, Maurice G., and Stuart, Alan (1961), The Advanced Theory of Statistics, Vol. 2, New York, Hafner Publishing Company.

Khatri, C.G. (1962), "A method for estimating approximately the parameters of a certain class of distributions from grouped observations," Annals of the Institute of Statistical Mathematics, Tokyo, Vol. 14, pp. 57-62.

Kulldorff, Gunnar (1957), "On the conditions for consistency and asymptotic efficiency of maximum likelihood estimates," Skandinavisk Aktuarietidskrift, Vol. 40, pp. 129-144.

Kulldorff, Gunnar (1958a), "Maximum likelihood estimation of the mean of a normal random variable when the sample is grouped," Skandinavisk Aktuarietidskrift, Vol. 41, pp. 1-17.

Kulldorff, Gunnar (1958b), "Maximum likelihood estimation of the standard deviation of a normal random variable when the sample is grouped," Skandinavisk Aktuarietidskrift, Vol. 41, pp. 18-36.

Kulldorff, Gunnar (1961), Contributions to the Theory of Estimation from Grouped and Partially Grouped Samples, New York, John Wiley and Sons, Inc.

Kulldorff, Gunnar (1963), "Asymptotically optimum equidistant grouping for the normal distribution," Skandinavisk Aktuarietidskrift, Vol. 46, pp. 157-161.

Kunz, Kaiser S. (1957), Numerical Analysis, New York, McGraw-Hill Book Company, Inc.

Li, Jerome C.R. (1964), Statistical Inference, Vol. I, Ann Arbor, Michigan, Edwards Brothers, Inc.

Lindley, D.V. (1949), "Grouping corrections and maximum likelihood equations," Proceedings of the Cambridge Philosophical Society, Vol. 46, pp. 106-110.

Rao, C.R. (1957), "Maximum likelihood estimation for the multinomial distribution," Sankhyā, Vol. 18, pp. 139-148.

Rao, C.R. (1958), "Maximum likelihood estimation for the multinomial distribution with infinite number of cells," Sankhyā, Vol. 20, pp. 211-218.

Rao, C.R. (1961a), "Asymptotic efficiency and limiting information," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 531-545.

Rao, C.R. (1961b), "A study of large sample test criteria through properties of efficient estimates," Part 1: Tests for goodness of fit and contingency tables., Sankhyā, Vol. 23, pp. 25-40.

Rao, C.R. (1965), Linear Statistical Inference and Its Applications, New York, John Wiley and Sons, Inc.

Sarhan, Ahmed E., and Greenberg, Bernard G. (1962), Contributions to Order Statistics, New York, John Wiley and Sons, Inc.

Scarborough, J.B. (1930), Numerical Mathematical Analysis, Baltimore, The Johns Hopkins Press.

Smith, Walter L. (1957), "A note on truncation and sufficient statistics," Annals of Mathematical Statistics, Vol. 28, pp. 247-252.

Swamy, P.S. (1959-1960), "Estimating the mean and the variance of a normal population from singly and doubly truncated samples of grouped observations," Calcutta Statistical Association Bulletin, Vol. 9, pp. 145-156.

Tallis, G.M. (1967), "Approximate maximum likelihood estimates from grouped data," Technometrics, Vol. 9, pp. 599-606.

Tukey, John W. (1949), "Sufficiency, truncation, and selection," Annals of Mathematical Statistics, Vol. 20, pp. 309-311.

Wald, Abraham (1943), "Tests of statistical hypotheses concerning several parameters when the number of observations is large," Transactions of the American Mathematical Society, Vol. 54, pp. 426-482.

Wald, Abraham (1949), "Note on the consistency of the maximum likelihood estimate," Annals of Mathematical Statistics, Vol. 20, pp. 595-601.

Widder, D.V. (1961), Advanced Calculus (second edition), Englewood Cliffs, N.J., Prentice-Hall, Inc.

Wilks, Samuel S. (1962), Mathematical Statistics, New York, John Wiley and Sons, Inc.

APPENDIX

The nature of the problem considered in this thesis requires that the solution of several equations involving integrals be found. The method of solution presented by Hartley (1958) and extended by Hughes (1962) is a very appropriate method of solution in some cases. The method, as presented here, is oriented in such a way as to be applicable when solving the maximum likelihood equations for grouped data samples from absolutely continuous distribution functions. The method has the very desirable property of being applicable to very general types of incomplete data, but this generality will be omitted in the present development. A unified treatment is found in Hughes (1962).

In Chapter II the method of successive approximations is compared to Hughes' method in the special cases considered there. These comparisons indicate that Hughes' method can be improved in many cases.

The development is presented in three parts; (i) The Iterative Procedure, (ii) The Sufficient Conditions for Convergence, and (iii) Extensions of Hughes' Method.

Before consideration is given to the iterative procedure, the definition of a quadrature formula should be established. Let the value of $\int_a^b \varphi(x)dx$ be desired. If the interval [a,b]

is partitioned into ν-1 intervals of equal length, $[x_i, x_{i+1}]$, where $x_1 = a$, $x_2 = a + \frac{b-a}{\nu-1}$, ..., $x_\nu = b$, then $\int_a^b \varphi(x)dx \approx \sum_{i=1}^{\nu} a_i \varphi(x_i)$, where $a_i > 0$ and $\sum_{i=1}^{\nu} a_i = b - a$. The precision of the approximation depends on ν, the function $\varphi(x)$, and the quadrature rule governing the selection of the a_i. There are many quadrature formulae, such as Simpson's rules and Weddel's rules (see Kunz (1957) and Fox (1963)), which determine the a_i and hence the quadrature formula.

(i) The Iterative Procedure

Let $f(x;\underline{\theta})$ be the density function of the random variable X, where $\underline{\theta}$ is a (q x 1)-vector of parameters defined in Θ, a subset of Euclidean q-space. The following assumptions are made:

1. The grouped data sample consists of k cell frequencies n_i, $\sum_{i=1}^{k} n_i = n$, where the i^{th} cell is $I_i \equiv (r_{i-1}, r_i]$ and $-\infty = r_0 < r_1 < \ldots < r_k = \infty$.

2. The log-likelihood function $L(\underline{\theta}) = C + \sum_{i=1}^{k} n_i \log P_i(\underline{\theta})$, where C is not a function of $\underline{\theta}$ and $P_i(\underline{\theta}) = \int_{I_i} f(x;\underline{\theta})dx$,

can be approximated by applying quadrature formulae to approximate each $P_i(\underline{\theta})$, $i = 1, \ldots, k$.

3. $\frac{\partial f(x;\underline{\theta})}{\partial \theta_i}$, $i = 1, \ldots, q$, can be expressed in terms of $f(x;\underline{\theta})$. (This assumption simply insures that the equations which must be solved to perform the iterations are similar to those maximum likelihood equations which are obtained in the complete data case.)

The q equations which must be solved are:

$$\frac{\partial L(\underline{\theta})}{\partial \underline{\theta}} = \underline{0}. \tag{1}$$

The iterative procedure consists of the following steps. Select some quadrature formula to approximate each $P_i(\underline{\theta}_o)$, where $\underline{\theta}_o$ is a starting value in Θ. Let

$$P_i^*(\underline{\theta}) = \sum_{j=1}^{\nu_i} a_{ij} f(X_{ij}, \underline{\theta}),$$

where

$$X_{i1} = r_{i-1},\ X_{i2} = r_{i-1} + \frac{r_i - r_{i-1}}{\nu_i - 1},\ \ldots,\ X_{i\nu_i} = r_i$$

and $\sum_{j=1}^{\nu_i} a_{ij} = (r_i - r_{i-1})$, for $i = 1, \ldots, k$.

Then $\frac{\partial L(\underline{\theta})}{\partial \underline{\theta}}$ in equation (1) can be approximated by

$$\frac{\partial L^*(\underline{\theta})}{\partial \underline{\theta}} = \frac{\partial}{\partial \underline{\theta}} \sum_{i=1}^{k} n_i \log P_i^*(\underline{\theta})$$

$$= \sum_{i=1}^{k} n_i P_i^{*-1}(\underline{\theta}) \frac{\partial P_i^*(\underline{\theta})}{\partial \underline{\theta}}$$

$$= \sum_{i=1}^{k} n_i P_i^{*-1}(\underline{\theta}) \sum_{j=1}^{\nu_i} a_{ij} \frac{\partial f(X_{ij};\underline{\theta})}{\partial \underline{\theta}} .$$

Define the pseudo frequencies $f'_{X_{ij}} = n_i P_i^{*-1}(\underline{\theta}) a_{ij} f(X_{ij};\underline{\theta})$ and write

$$\frac{\partial L^*(\underline{\theta})}{\partial \underline{\theta}} = \sum_{i=1}^{k} \sum_{j=1}^{\nu_i} f'_{X_{ij}} \frac{\partial \log(f(X_{ij};\underline{\theta}))}{\partial \underline{\theta}} . \qquad (2)$$

Note that if the $f'_{X_{ij}}$ were omitted from the expression on the right, then the expression would be identical to the expression which would be obtained from a complete data sample consisting of the observations $\{X_{ij}\}$, $j = 1, \ldots, \nu_i$ and $i = 1, \ldots, k$. Hughes referred to the $f'_{X_{ij}}$ as proportionally allocated weights.

The iterative procedure is as follows:

1. Use $\underline{\theta}_o$, the starting value, to compute the $f'_{X_{ij}}$.
2. Substitute these weights into equation (2) and then compute $\underline{\theta}_1$, the solution of equation (2).

(Note that $\sum_{i=1}^{k} \sum_{j=1}^{\nu_i} n_i P_i^{*-1}(\underline{\theta}) a_{ij} f(X_{ij};\underline{\theta}) = n$.)

3. Using the value of $\underline{\theta}$ obtained in step 2, repeat step 1.

(ii) The Sufficient Conditions for Convergence

Hughes' method of solution generates a sequence $\underline{\theta}_0, \underline{\theta}_1, \ldots$. A set of sufficient conditions for this sequence to converge to the solution of equation (2) are given in Hughes (1962). One of the conditions of Hughes' convergence theorem is very restrictive and it would seem appropriate to examine this condition to see why it is not satisfied by any of a very large class of distributions.

Hughes' convergence theorem requires that the underlying density function (or more generally frequency function) $f(x;\underline{\theta})$, $\underline{\theta} \in \Theta \subset E_q$, satisfy the following condition. There exists a bounded, convex set $C \subset \Theta$ and a point $\underline{\theta}_c$ in the interior of C such that for any sample $(x_1, \ldots, x_n)$

$$\prod_{i=1}^{n} f(x_i;\underline{\theta}_c) > \prod_{i=1}^{n} f(x_i;\underline{\theta})$$

for every $\underline{\theta} \in \partial C$ and for every $\underline{\theta}$ in the complement of C. This condition eliminates all underlying distributions for which any parameter is defined in an open space and the

(complete data) maximum likelihood estimator for $\underline{\theta}$ is consistent. This assertion is proven in the theorem which follows. See section 1.1 of this thesis for sufficient conditions for consistency of maximum likelihood estimators as developed by Wald (1949).

Theorem 1

If $f(x;\underline{\theta})$ is the frequency function of the random variable X and $\underline{\theta}$ is a q x 1 vector defined in $\Theta \subset E_q$ such that at least one parameter, say θ_1, is defined in an open space and if the maximum likelihood estimator in the strict sense, $\hat{\underline{\theta}}_n$, is consistent for the true value of $\underline{\theta}$, $\underline{\theta}_o \in \Theta$, then the convex set C required in Hughes' convergence theorem does not exist.

Proof: Assume that C is some bounded, convex set in Θ such that the likelihood function $L(\underline{\theta})$ is greater at some point in the interior of C than at any point on ∂C or in the complement of C for every sample $(x_1, \ldots, x_n)$. This implies that the values of the strict maximum likelihood estimator (there could be more than one value of this estimator) are all in the interior of C for every sample $(x_1, \ldots, x_n)$. Suppose that the true value of $\underline{\theta}$, $\underline{\theta}_o$, is not in C. Then the distance from $\underline{\theta}_o$ to any point in C is greater than some $\epsilon > 0$. Therefore $\Pr\{|\hat{\underline{\theta}}_n - \underline{\theta}_o| > \epsilon\} = 1$ for every $n > 0$ and for any value of the strict maximum

likelihood estimator $\underline{\hat{\theta}}_n$. This conclusion contradicts the assumption that the strict maximum likelihood estimator is consistent for $\underline{\theta}_o$.

One concludes from this theorem that if the parameter space is not a closed metric space, then the conditions of Hughes' convergence theorem cannot be satisfied by any distribution which admits consistent maximum likelihood estimators in the complete data case.

Therefore, the normal, exponential, Poisson, and many other distributions do not satisfy Hughes' sufficient conditions which establish the convergence of the sequence defined by steps 1-3 given previously. The failure of many distributions to satisfy the sufficient conditions developed by Hughes does not imply that the method is not a very useful method of solution. There are many examples in which the method is very good. See examples in Hughes (1962).

(iii) Extensions of Hughes' Method

Assume that $f(x,y;\underline{\theta})$ is the density function of the random variables X and Y defined in $\underline{X} \times \underline{Y} \subset E_2$ and that $\underline{\theta}$ is a q x 1 vector of parameters defined in $\Theta \subset E_q$. Assume that the likelihood equations for the complete data sample,

$$\frac{\partial L_c(\underline{\theta})}{\partial \underline{\theta}} = \sum_{i=1}^{n} \frac{\partial \log(f(x_i,y_i;\underline{\theta}))}{\partial \underline{\theta}} = \underline{0},$$

can be solved for $\hat{\underline{\theta}}_n = \hat{\underline{\theta}}((x_1,y_1), \ldots, (x_n,y_n))$. Assume that the x-axis is partitioned into k_1 cells, $I_i \equiv (r_{i-1},r_i]$, $i = 1, \ldots, k_1$, where $-\infty = r_0 < r_1 < \ldots < r_{k_1} = \infty$ and that the y-axis is partitioned into k_2 cells, $J_j \equiv (s_{j-1},s_j]$, $j = 1, \ldots, k_2$, where $-\infty = s_0 < s_1 < \ldots < s_{k_2} = \infty$. Let $R_{ij} \equiv I_i \times J_j$. Then if the observations are grouped into the cells R_{ij} and there are n_{ij} observations recorded in R_{ij}, where $\sum_{i,j} n_{ij} = n$, the log-likelihood function for the grouped data sample is:

$$L(\underline{\theta}) = C + \sum_{i,j} n_{ij} \log P_{ij}(\underline{\theta}),$$

where C is not a function of $\underline{\theta}$ and $P_{ij}(\underline{\theta}) = \iint_{R_{ij}} f(x,y;\underline{\theta})dxdy$.

The range of i is understood to be $1, \ldots, k_1$ and the range of j is understood to be $1, \ldots, k_2$ unless specified otherwise. If a quadrature formula such as Simpson's rule is applied twice to the function $P_{ij}(\underline{\theta})$, then I_i is partitioned into $\nu_{(i)} - 1$ sub-intervals with end points $X_{\alpha(i)}$, where

$$X_{1(i)} = r_{i-1}, \; X_{2(i)} = r_{i-1} + \frac{r_i - r_{i-1}}{\nu_{(i)} - 1}, \; \ldots, \; X_{\nu_{(i)}(i)} = r_i$$

and J_j is partitioned in a similar manner with $Y_{\beta(j)}$ being

the $\nu_{(j)}$ end points of the $\nu_{(j)} - 1$ sub-intervals. The quadrature formula defines the constants $a_{\alpha(i)}$ and $a_{\beta(j)}$ such that

$$\sum_{\alpha(i),\beta(j)} a_{\alpha(i)} a_{\beta(j)} = (r_i - r_{i-1})(s_j - s_{j-1}),$$

where $\alpha(i) = 1, 2, \ldots, \nu_{(i)}$ and $\beta(j) = 1, 2, \ldots, \nu_{(j)}$. Then

$$\iint_{R_{ij}} f(x,y;\underline{\theta})\,dxdy \approx \sum_{\alpha(i),\beta(j)} a_{\alpha(i)} a_{\beta(j)} f(X_{\alpha(i)}, Y_{\beta(j)};\underline{\theta})$$

and letting this expression be $P^*_{ij}(\underline{\theta})$ we have

$$L^*(\underline{\theta}) = \sum_{i,j} n_{ij} \log P^*_{ij}(\underline{\theta}) + C .$$

In the same manner described in part (i) preceding, $\frac{\partial L(\underline{\theta})}{\partial \underline{\theta}} = \underline{0}$ can be reduced to the (approximate) equation:

$$\frac{\partial L^*(\underline{\theta})}{\partial \underline{\theta}} = \sum_{i,j} \sum_{\alpha(i),\beta(j)} f'_{X_{\alpha(i)},Y_{\beta(j)}} \cdot \frac{\partial \log(f(X_{\alpha(i)},Y_{\beta(j)};\underline{\theta}))}{\partial \underline{\theta}} = \underline{0},$$

where

$$f'_{X_{\alpha(i)},Y_{\beta(j)}} = n_{ij} P^{*-1}_{ij}(\underline{\theta}) a_{\alpha(i)} a_{\beta(j)} f(X_{\alpha(i)},Y_{\beta(j)};\underline{\theta}) .$$

Steps 1-3 given in part (i) are then performed.

If this method is applied to the case of a bivariate normal distribution, then the roots of the following equations must be obtained using steps 1-3 given in part (i):

$$\frac{1}{n} \sum_{i} \sum_{\alpha(i)} f'_{X_{\alpha(i)}} X_{\alpha(i)} = \mu_x ,$$

where

$$\mu_x = EX \text{ and } f'_{X_{\alpha(i)}} = \sum_{\beta(j)} f'_{X_{\alpha(i)}, Y_{\beta(j)}}$$

$$\frac{1}{n} \sum_{j} \sum_{\beta(j)} f'_{Y_{\beta(j)}} Y_{\beta(j)} = \mu_y ,$$

where

$$\mu_y = EY \text{ and } f'_{Y_{\beta(j)}} = \sum_{\alpha(i)} f'_{X_{\alpha(i)}, Y_{\beta(j)}}$$

$$\frac{1}{n} \sum_{i} \sum_{\alpha(i)} f'_{X_{\alpha(i)}} (X_{\alpha(i)} - \mu_x)^2 = \sigma_x^2 ,$$

where

$$E(X - \mu_x)^2 = \sigma_x^2$$

$$\frac{1}{n} \sum_{j} \sum_{\beta(j)} f'_{Y_{\beta(j)}} (Y_{\beta(j)} - \mu_y)^2 = \sigma_y^2 ,$$

where

$$E(Y - \mu_y)^2 = \sigma_y^2$$

$$\frac{1}{n} \sum_{i,j} \sum_{\alpha(i),\beta(j)} f'_{X_{\alpha(i)}, Y_{\beta(j)}} (X_{\alpha(i)} - \mu_x)(Y_{\beta(j)} - \mu_y) = \sigma_{xy} ,$$

where

$$E(X - \mu_x)(Y - \mu_y) = \sigma_{xy} .$$

One other case which might be considered is the sample from a bivariate normal distribution, with the unknown parameters μ_x, μ_y, σ_x^2, σ_y^2, and σ_{xy}, which consists of observations of the form $\{X \in I_i, Y = y_{ij}\}$, $j = 1, \ldots, n_i$,

$\sum_{i=1}^{k} n_i = n$, and $i = 1, \ldots, k$, where $I_i \equiv (r_{i-1}, r_i]$ such that $-\infty = r_0 < r_1 < \ldots < r_k = \infty$. The log-likelihood function for the sample is

$$L(\underline{\theta}) = \sum_{i,j} \log \left[\Pr\{X \in I_i | y_{ij}\} f(y_{ij}) \right] ,$$

where $f(y_{ij})$ is the marginal density function of Y evaluated at y_{ij}. The parameter vectors are omitted in the arguments of those functions on the right side of this equation to avoid unnecessary, complicated notation. The log-likelihood function can be reduced to the form:

$$L(\underline{\theta}) = \sum_{i,j} \log(f(y_{ij})) + \sum_{i,j} \log \int_{I_i} f(x|y_{ij})dx .$$

Applying a quadrature formula, defining the constants a_{ijt} such that $\sum_{t=1}^{\nu_{ij}} a_{ijt} = r_i - r_{i-1}$, to $P_{i|j} = \int_{I_i} f(x|y_{ij})dx$, let

$$P^*_{i|j} = \sum_{t=1}^{\nu_{ij}} a_{ijt} f(X_{ijt}|y_{ij}) ,$$

where, as before, the X_{ijt} are the end points of $\nu_{ij} - 1$ subintervals of I_i, when y_{ij} was the observed value of Y such that $X_{ij1} = r_{i-1}$, $X_{ij2} = r_{i-1} + \frac{r_i - r_{i-1}}{\nu_{ij}-1}$, $\ldots$, $X_{ijt} = r_i$.

We obtain the equations:

$$\frac{1}{n}\sum_{i,j} y_{ij} = \hat{\mu}_y \,, \qquad \frac{1}{n}\sum_{i,j}(y_{ij} - \hat{\mu}_y)^2 = \hat{\sigma}_y^2$$

and

$$\frac{1}{n}\sum_{i,j,t} f'_{X_{ijt}} X_{ijt} = \mu_x$$

$$\frac{1}{n}\sum_{i,j,t} f'_{X_{ijt}} (X_{ijt} - \mu_x)^2 = \sigma_x^2$$

$$\frac{1}{n}\sum_{i,j,t} f'_{X_{ijt}} (X_{ijt} - \mu_x)(y_{ij} - \hat{\mu}_y) = \sigma_{xy} \,,$$

where

$$f'_{X_{ijt}} = P^{*-1}_{i|j}\, a_{ijt} f(X_{ijt} | y_{ij}; \mu_y = \hat{\mu}_y,\ \sigma_y^2 = \delta_y^2) \,.$$

Steps 1-3 in part (i) must be applied to the last three of these equations.

In this Appendix we have examined the iterative procedure which was introduced by Hartley (1958) and further developed by Hughes (1962). Extensions of this method were made to accommodate grouped data samples from a bivariate normal distribution in which either one or both variates are grouped.

Hughes' (1962) convergence theorem treating the convergence properties of this method of solution was studied and was shown to have very limited application. This fact does not mean that the method itself is of little value since

modifications of this method were found to be quite useful in Chapter II.

Attention Patron:

The one-page vita has been removed

from the scanned document

MAXIMUM LIKELIHOOD ESTIMATION OF PARAMETERS USING GROUPED DATA SAMPLES

(Jacob Van Bowen, Jr.)

Abstract

The method of maximum likelihood was employed to estimate the parameters of the univariate normal distribution when grouped data only are available. The term grouped data sample was used to describe a sample in which the only information available is that there are m_i observations less than or equal to r_i, $i = 0, 1, \ldots, k$ and $-\infty = r_0 < r_1 < \ldots < r_k = \infty$ and $m_k = n$. Three cases were considered: the variance of the underlying normal distribution known, the mean of the underlying normal distribution known, and the case where both parameters are unknown. The numerical method of successive approximations was studied and was compared to other iterative methods of solution of the likelihood equations. Theorems which establish the convergence of the sequence defined by the method of successive approximations to the unique root of the likelihood equations were developed for the first two cases. In the case where both parameters are unknown, numerical results indicate that a region can be ascertained in which there can be at most one joint solution of the likelihood equations and that, if there is a solution in this region, then it is a point at which the likelihood function has a

relative maximum.

A search was conducted to find a sample which satisfies the conditions insuring the existence of a joint solution of the likelihood equations and which might provide two solutions of the likelihood equations. No sample was found with this property. It was concluded that, except possibly for extremely unusual grouped data sample configurations, one should expect only one solution of the likelihood equations in the case of a normal distribution.

Simply computed, consistent estimators were developed which provide good approximations to the solution of the likelihood equations and which can be used for starting values in any iterative method of solution.

Modifications of the method of successive approximations were developed. These methods of solution define sequences which are rapidly convergent, and which enable the solution to be obtained with the use of tables of the standard normal probability distribution function and its derivatives.

The maximum likelihood estimators obtained when both parameters are unknown were proved to be asymptotically efficient provided there are at least three groups, i.e. $k = 3$. The theorems which were developed to establish this property were shown to have a much wider application. Other distributions such as the Poisson distribution and the

bivariate normal distribution were considered. Sufficient conditions were established for a grouped data sample, from these as well as other distributions, to provide maximum likelihood estimators which are asymptotically efficient.