CONDITIONAL INFERENCE

BY

JOHN MUSISI SENYONYI-MUBIRU

B.Sc.(Hons.), (NRB)

Thesis submitted for the degree of Doctor of Philosophy in the Department of Statistics, University of Melbourne.

August, 1984.

To my beloved parents

Mr. & Mrs. E.D.K. KAJJA,

who have given unsparingly for me.

CONTENTS.

ACKNOWLEDGEMENTS

First and foremost I wish to express my sincere thanks to my supervisors, Emeritus Professor E.J. Williams and Professor C.C. Heyde for their availability, guidance and encouragement in the course of my research. Various lecturers in the Department of Statistics, University of Melbourne, have given me invaluable advice and help when I most needed it.

I would be lacking in courtesy if I did not mention friends and loved ones who through their prayers and encouragement have made it possible for me to study singlemindedly.

Last but not least, I am grateful to the Australian Government, through the Australian Development Assistance Bureau, for the Post-Graduate Award under which this work was carried out.

CERTIFICATION

Unless otherwise stated, the work presented in this thesis is original; being work done by the author himself. Furthermore, it has not been submitted for any other degree or award of benefit to the author.

I also hereby declare that this thesis is less than 100,000 words in length, exclusive of tables and bibliographies.

JOHN MUSISI SENYONYI-MUBIRU.

ABSTRACT.

Conditional inference is a branch of statistical inference in which observed data is reduced using either sufficient or ancillary statistics. This often simplifies inference about the parameters. In comparison to full likelihood methods, conditional inference theory's performance still needs validating in many areas. Some of these are the concern of this thesis.

While the definition of an ancillary statistic in single parameter models is unequivocal, the presence of accessory (or nuisance) parameters in a model presents problems in defining an ancillary statistic. Statistical literature abounds with definitions of ancillarity in this case. Some of the commonest and most useful of these are discussed and shown to be interrelated. This facilitates the choice of the strongest eligible ancillary in a problem, i.e. that which offers the biggest reduction of the sample space.

The Pitman-Morgan test for variance ratios in bivariate normal populations with unknown correlation coefficient is shown to be a conditional test. We condition on sufficient statistics for the accessory parameters to eliminate them. The test statistic is then derived as an ancillary statistic for the accessory parameters.

Conditional inference procedures are useful in regression problems; in particular we discuss the analysis of concurrent

regressions. Earlier work on a restricted class of concurrent regressions is clarified using conditional methods. A suggested analysis of the general case is also presented.

When a probability model depends on a number of accessory parameters which increases with the sample size, estimation methods based on the full likelihood will often be inconsistent. Using a partial likelihood instead has been suggested. Local maximum partial likelihood estimators are shown to exist, and to be consistent and asymptotically normal under mild conditions. These results also cover conditional and marginal likelihoods, thus considerably strengthening earlier results in this area.

In planning statistical inferences, it is useful to choose a sampling scheme which provides only the essential data to our inferences. Jagers' lemma proposes very general conditions under which maximum likelihood estimation from a subset of the data is identical with that from the full data. However, the lemma is incorrect as given. We show that an additional sufficiency condition repairs the lemma. It is further shown that this lemma cannot be extended to general exponential families.

NOTATION AND SYMBOLS

(i) General

Sections, definitions, examples, equations, propositions and theorems are numbered in the form k.ℓ where k stands for chapter k and ℓ is the numerical sequence of whatever is numbered. eg. Section 1.2 is the 2nd section in chapter 1.

(ii) Symbols

$(\mathbf{X}, \mathbf{A})$	: A measurable space where $\mathbf{A}$ is a σ-algebra on the event space $\mathbf{X}$.
$P(.)$	: Probability measure on a measurable space.
$\Omega_1 \times \Omega_2$	: A cross-product of parameter spaces Ω_1 and Ω_2.
$\mathbb{R}^k$	: k dimensional Euclidean space.
G/G/1	: General arrival, general service, single-server queue.
$N(\mu,\sigma^2)$	: Normal distribution with mean μ and variance σ^2.
$\Gamma(n)$	: Gamma function.
$f'(\theta)$	: First derivative of the function f with respect to θ.
$f''(\theta)$	: Second derivative with respect to θ.
a.s. [u]	: For almost all values of u.
$[A]^T$	: Transpose of matrix A.
$\stackrel{d}{=}$	: "is distributed as"
$\longrightarrow$	: "converges to"
$\xrightarrow{a.s.}$	: "tends almost surely to"
$\xrightarrow{d}$	: "tends in distribution to"
$\sim$	: "asymptotically tends to"

(iii) Abbreviations

ℓog	: Logarithm to base e.
d.f.	: Degrees of freedom.
ANOVA	: Analysis of Variance.
contd.	: Continued.

1. INTRODUCTION.

1.1 Conditional Inference.

Statistical inference procedures which base inference on the conditional distribution given either ancillary statistics or sufficient statistics are what constitute conditional inference. The choice to condition on either the sufficient or ancillary statistics is also very much dependent on the model under consideration. For example, if the distribution model depends on accessory parameters (that is, parameters which are not of interest to our particular problem) in addition to our parameter(s) of interest (we shall call the latter structural parameters), it is helpful to eliminate the accessory parameters. A useful course of action is to condition on sufficient statistics for the accessory parameters. Or it may happen that in the course of estimating a parameter, some 'information' will be lost. Conditioning on ancillary statistics in this case has been suggested. This gives us a measure of the precision of our estimate.

Fisher(1925,1934,1935) laid the foundation for conditional inference by introducing the concepts of 'sufficient statistics' and 'ancillary statistics.' Sufficient statistics are functions of the observed data which contain all the available 'information' in that data about the structural parameter. Ancillary statistics, also functions of the observations, are by themselves uninformative or contain no 'information' about the parameter of interest. Formal definitions for sufficient and ancillary statistics will be introduced in the next section. In the recent past much work has been done, notably that of Andersen(1970,1973), and of Kalbfleisch and

Sprott(1970,1973), showing that the conditional approach often yields identical results to those we get with the full likelihood or even sometimes, more desirable results. Of course there are still many areas in which the effectiveness of conditional inference needs to be investigated. Some of these are the subjects of this thesis. Of particular interest are those distribution models which depend on accessory parameters.

This chapter introduces the general definitions of sufficient and ancillary statistics in section 1.2. We further present a brief overview of the various uses for sufficient and ancillary statistics in inference. In section 1.3, we give an outline of the work covered in this thesis.

1.2 Sufficient and Ancillary Statistics.

We shall consider a random variable X and a family of all possible distributions of X, $\mathfrak{F}$. The p.d.f. of X is denoted by f(x).

Definition 2.1:

A statistic S is sufficient for $\mathfrak{F}$ if the conditional distribution of X given S, namely f(x|s), is the same for all f $\varepsilon\mathfrak{F}$.

If $\mathfrak{F}$ is indexed by some parameter, θ, the definition means that the conditional distribution, f(x|s) is independent of θ. Therefore when S is known, there is no additional 'information' contained in the data, beyond that already contained in S, about θ. It is then logical to make any inference about θ on the basis of S alone. S is called a minimal sufficient statistic for θ if it is a function of any other sufficient statistic for θ.

Definition 2.2:

A statistic U is ancillary for a parameter, θ, if the marginal distribution of U, f(u), is independent of θ.

These definitions are unequivocal in as far as we do not introduce accessory parameters. But it is in this latter situation that conditional inference has proved most relevant, while simultaneously, results obtained in the one parameter case are not automatically applicable to this multiparameter case. As well definitions of both sufficiency and ancillarity are varied in the literature when accessory parameters are present. A detailed discussion of ancillarity definitions is presented in chapter 2.

We may state the problem as follows. Assume that the probability model under consideration is $f(x;\theta,\phi)$, where θ is the structural parameter and ϕ the accessory parameter. If S and U, sufficient and ancillary statistics for ϕ respectively, exist, we can write

$$f_X(x;\theta,\phi) = f_S(s;\theta,\phi).f_{X|S}(x;\theta|s) \qquad (1.1)$$

and

$$f_X(x;\theta,\phi) = f_U(u;\theta).f_{X|U}(x;\theta,\phi|u). \qquad (1.2)$$

Consequently, we may ask

(a) like Dawid(1975), which of the two models, $f_{X|S}(x;\theta|s)$ in (1.1) and $f_U(u;\theta)$ in (1.2), is appropriate for our inference?

(b) assuming we choose $f_{X|S}(x;\theta|s)$, how much 'information' on θ is lost to $f_S(s;\theta,\phi)$?

A related problem is the absence of an explicit measure of information content. Pitman(1979, page 18) gives some examples and interesting remarks on using Fisher's information function as a measure of information showing that it may not actually reflect the information we want.

We should note the lack of general patterns for constructing ancillary statistics in particular. Coupled with this is their possible non-uniqueness. Therefore we often have to choose from equally eligible ancillaries (Basu, 1964).

Notwithstanding, ancillary and sufficient statistics are very useful in statistical inference. Fortunately, in the course of conditioning on an ancillary statistic no information loss is involved since the ancillary statistic is on its own uninformative about the structural parameter. From the outset, Fisher(1935) recognized that the ancillary statistic serves as an index, a measure of the precision of an estimate we may make of a parameter. To put it another way, the ancillary statistic will describe the dimension of the sample space relevant to our problem. An example from Cox and Hinkley (1974) illustrates this.

Example 2.1:

Assume we observe X such that it is equally likely to come from either of the two normal populations, $N(\mu,\sigma_1^2)$ and $N(\mu,\sigma_2^2)$, $\sigma_1^2 \neq \sigma_2^2$, both σ_1^2 and σ_2^2 being known. So inference from X will depend on whether

$\sigma_u^2 = \sigma_1^2$ or σ_2^2. Besides P(U=u) = 1/2, u = 1,2. The joint likelihood of X and U is

$$\frac{1}{2}(2\pi\sigma_u^2)^{-\frac{1}{2}} \exp\{-\frac{(x-\mu)^2}{2\sigma_u^2}\}.$$

Conditioning on U means the population sampled is specified, U being ancillary. Since we know the population, we do not need to consider the population that was not sampled. Therefore the ancillary statistic has specified the dimension of the sample space relevant to our inference.

Sufficient statistics too play major roles in inference, some of which will be evident in the course of this thesis. Only a brief informative overview on their roles is given here. Sufficient statistics are useful in the following:

(i) Sufficient statistics may be used to define the best critical region of a uniformly most powerful test. Lehmann(1959, page 134-136) has shown that, in the case of the exponential family, it is possible to define uniformly most powerful tests whose test function is determined by the sufficient statistic. In fact, on the basis of the sufficient statistic, Fraser(1956) showed that the sign test is uniformly most powerful in a nonparametric location (family) problem.

It is known that the critical region for a uniformly most powerful test is defined by the likelihood ratio. But the partitioning

of the sample space by the likelihood ratio is sufficient (Cox and Hinkley, 1974, page 24). Therefore the sufficient statistic defines the best region by assuming a constant value along the boundary of the critical regions.

It would seem from Neyman and Pearson(1936) that this theory cannot be generalised to all distributions i.e. the existence of a sufficient statistic is no guarantee that a uniformly most powerful test exists.

(ii) In the special case when the underlying distribution family is complete, Lehmann and Scheffé(1950) have shown that any function of a complete sufficient statistic will be the unique minimum variance unbiased estimator of its expectation. When completeness is removed, the estimator may not be unique. A short proof of this is available in Cox and Hinkley(1974, page 258-259). Conversely, if an estimator is minimum variance unbiased estimator for its expected value, it must be a function of the sufficient statistic. In the construction of a minimum variance unbiased estimator therefore, it is advisable to start from a sufficient statistic if it exists. This result is used in chapter 6.

(iii) Sufficient statistics may also be used to derive non-null distributions from null distributions (Madow, 1945; Durbin, 1980). S is sufficient in the model $f(x;\theta)$ if

$$f(x;\theta) = g(s;\theta).h(x|s). \qquad (1.3)$$

We assume the marginal density $g(s;\theta)$ is unknown. Let θ assume some value θ_o and we have

$$f(x;\theta_o) = g(s;\theta_o).h(x|s). \qquad (1.4)$$

From (1.3) and (1.4),

$$f(x;\theta) = f(x;\theta_o).\frac{g(s;\theta)}{g(s;\theta_o)}.$$

This technique will be used in chapter 3 to derive the distribution of a sufficient statistic.

(iv) By far the most common usage of sufficient statistics is in the elimination of accessory parameters. This will be evident in the thesis too.

Let $f(x;\theta,\tau)$be the distribution from which the sample is taken where τ may be a vector of accessory parameters. If the model admits minimal sufficient statistics for the accessory parameters, Andersen(1973) suggests conditioning on the minimal sufficient statistics to eliminate these parameters, and proves that the estimate for θ from the conditional distribution will satisfy a number of desirable asymptotic results. We refer to some of these in chapter 5. So if S is minimal sufficient for τ, we have

$$f(x;\theta,\tau) = f(s;\theta,\tau).f(x;\theta|s).$$

The insistence on using minimal sufficient statistics is to minimise the loss of information about θ through $f(s;\theta,\tau)$, (Andersen,1973, page 42).

Sometimes it happens that the sufficient statistic for τ depends on θ, the structural parameter, as well. An example of this is the sufficient statistic for the variance σ^2 in a normal population with unknown mean μ. This is

$$s^2 = \sum(x_i-\mu)/n$$

and if μ is the structural parameter, we may not condition on s^2 to eliminate σ^2. Kalbfleisch and Sprott(1970) have aptly described the problems involved in so doing.

Basu(1955,1958) showed that under quite general conditions a statistic independent of a sufficient statistic will be ancillary. On the basis of this, Williams(1982) proposed and described how we may construct such an ancillary whenever the sufficient statistic depends on the structural parameter. So if S is sufficient for τ and depends on θ, the conditional distribution function

$$F_C(x;\theta|s) = \frac{\int_{-\infty}^{x} f(u,s;\theta,\tau)du}{\int_{-\infty}^{\infty} f(u,s;\theta,\tau)du}$$

is, irrespective of the value of S, distributed uniformly over (0,1) and is thus independent of S. Any function of F_C can serve as our ancillary for τ and is thus available for inference.

Durbin(1961) presented an interesting use of sufficient statistics in the elimination of accessory parameters to test for goodness-of-fit. Conditioning on sufficient statistics may permit a composite hypothesis to be tested as a simple hypothesis.

1.3 Outline of The Thesis.

As indicated earlier, definitions of ancillarity in the presence of accessory parameters are numerous. In chapter 2, we present these definitions, examine them and discuss their interrelationships. When making statistical inference, this makes it possible to choose from the eligible ancillaries in a particular problem that ancillary statistic which best describes the dimension of our sample space relevant to the inference. Generally, the strongest ancillary statistic will be most useful.

Pitman(1939) and Morgan(1939) devised a test for the ratio of variances in a bivariate normal distribution with unknown correlation coefficient, which is widely used today. Chapter 3 will show that this is a conditional test. (iii) and (iv) in section 1.2 are particularly useful in this derivation. The successful conditional derivation of such an important testing procedure helps to establish conditional inference as a significant approach to problems in statistical inference.

The analysis of concurrent regression lines first discussed in the literature by Tocher(1952) is presented in chapter 4 following the analysis of Williams(1959). It is shown that this analysis is an application of conditional inference procedures. We shall present a generalisation of the analysis to cases which Tocher(1952) and Williams(1959) did not cover. The role of sufficient and ancillary statistics in the analysis is clarified; in fact, these statistics form the basis for the analysis. This suggests that identifying sufficient and ancillary statistics in regression problems may be a

useful initial step, and conditional procedures may then be applied to design an appropriate analysis.

It was pointed out earlier that distribution models with the number of accessory parameters increasing with the sample size, present additional problems which cannot be solved by simply applying the general conditional procedures. The introduction of the 'partial likelihood' by Cox(1975) is an attempt to deal with this problem. Chapter 5 discusses the joint uniform asymptotic normality of the 'normalised' first derivative of the partial likelihood and the 'normalised information' function. It is shown that local maximum partial likelihood estimators exist and are both consistent and asymptotically normal under mild conditions. This result is easily applicable to the conditional and marginal likelihoods under similar general conditions to those for the partial likelihood. The conditions given shall also be shown to generalise results by Andersen(1970, 1973) on the conditional maximum likelihood estimators.

One of the purposes for using conditional inference is the reduction of data using sufficient and ancillary statistics. It is important that any reduction of data provides only the essential data to our inferences. This is not always so. Jagers'(1975) lemma asserted under rather general conditions that the conclusions should be identical. This lemma is not true in its generality and a revised version of it is presented and proved in chapter 6. The revision of the lemma is essentially to include a requirement of sufficiency. We consider an application of the revised version of Jagers' lemma to problems in Branching Processes and thereby show that the revised

version of the lemma is not extendable to the general exponential family.

It is easy to see that although all the chapters fall under the umbrella of conditional inference, there is relative independence across them. Consequently, it is attempted to make notations consistent within each chapter while allowing for some variation in notation between chapters.

2. DEFINITIONS OF ANCILLARITY IN THE PRESENCE OF ACCESSORY PARAMETERS.

2.1 Introduction.

The theory of ancillary statistics was initiated by Fisher(1925) in an attempt to recover some or all lost information in the estimation of unknown parameters with non-sufficient statistics. By definition, the ancillary statistics should, of themselves, contain no information about the parameter of interest since the distribution of an ancillary statistic must be independent of that parameter. As already pointed out in chapter 1, ancillary statistics are helpful in judging the precision of our estimators. When the underlying probability model sampled depends also on other parameters (not of direct interest to our particular study, and hence often called nuisance, incidental, or accessory parameters in the literature), defining an ancillary statistic useful for our inference is not straightforward; many varied definitions have been proposed in the literature. We shall present the major definitions and discuss how they are interrelated to place them in the order of their strengths.

First, we need to make a note on the choice of the word 'accessory' in preference to the more common word 'nuisance.' We may talk of a nuisance parameter only for a particular problem; in general, the word 'nuisance' is a misnomer. An example of sampling from the normal population illustrates this point.

Let X be a random variable such that

$$X \overset{d}{=} N(\mu,\sigma^2)$$

where μ and σ^2 are unknown.

When making inference on μ, the variance σ^2 will be a 'nuisance' parameter whose elimination we may desire. However, if the problem is to make inference on σ^2, then the mean μ is a 'nuisance' parameter. The word 'nuisance' implies uselessness. Yet in each case, neither μ nor σ^2 is unimportant in the model although it may not be relevant to our particular discussion.

Therefore, we shall call parameters which are not of interest in the problem at hand, 'accessory' parameters.

2.2 Definitions.

It has been hinted that the classical definition for an ancillary statistic is "a statistic whose distribution is independent (or free) of the parameter (of interest)." Let X be a random variable with probability density function $f(x;\theta)$, θ being a parameter from the space Ω. This definition implies that U is ancillary for θ if

$$f(x;\theta) = f_U(u).f_C(x;\theta|u)$$

for proper probability density functions f_U and f_C. The subscripts U and C refer to the marginal distribution of U, and the conditional distribution of X given U, respectively. They will be used in like manner throughout this chapter and all factorisations will be taken to be into proper probability density functions.

Now let $\theta = (\eta,\beta)$, where η is the parameter of interest (hereafter called the structural parameter) and β is the accessory parameter; η and β may be vector parameters.

Fraser(1956) proposed a definition for sufficiency in the presence of accessory parameters which is equally suitable for ancillarity. Barndorff-Nielsen(1978, Section 4.4) calls it S-ancillarity.

Definition 2.1: (S-ancillarity)

Let $(\eta,\beta) \in \Omega_1 \times \Omega_2$. Then U is (S-)ancillary for η if

$$f(x;\theta) = f_U(u;\beta).f_C(x;\eta|u).$$

It is clear that U is also sufficient for the accessory parameter β.

Example 2.1:

Let $\{Y_{ijk}\}$, i=1,...,m; j=1,...,n; k=1,...,ℓ; mnℓ=N be a set of random variables with finite mean such that

$$E(Y_{ijk}) = \mu+\alpha_i+\beta_j.$$

Then under the usual assumptions (i.e. $\sum\alpha_i = 0 = \sum\beta_j$),

$$E(\bar{y}_{i..} - \bar{y}_{...}) = \alpha_i.$$

If we assume Y_{ijk} is from a normal population with $Var(Y_{ijk}) = 1$, then $u_i = \bar{y}_{i..} - \bar{y}_{...}$ has the probability density function

$$f_U(u_i;\alpha_i) = c.\exp\{-\frac{N}{2m}\sum (u_i-\alpha_i)^2\},$$

where c is a constant. Therefore if $Y = (Y_{111},...,Y_{mn\ell})$,

$$f_Y(y;\mu,\alpha_i,\beta_j) = f_U(u_i;\alpha_i).f_C(y;\mu,\beta_j|u_i),\ i=1,\ldots,m;\ j=1,\ldots,n.$$

U_i is ancillary for β_j by definition 2.1.

Although this definition has the advantage of separating the parameters completely, and also follows directly from the classical definition of ancillarity in the single parameter case, it lacks the wide applicability we need. In most of the examples we give later, it is clear that definition 2.1 is not satisfied. We need to define an ancillary statistic with the following properties:

(i) It must be strong in the sense that it facilitates effective separation of the parameters, and its distribution should be functionally independent of the structural parameter. Definition 2.1 has this property.
(ii) It must be widely relevant in application.

Although property (ii) may be desirable, it can only be satisfied to the loss (or at least undesirable weakening) of property (i). The following are among the major definitions in the literature which attempt to relax definition 2.1 somewhat to permit wider applicability.

The next definition is due to Andersen(1970,1973, page 99).

Definition 2.2:

U is 'weakly ancillary' for η in the presence of the accessory parameter β if given any values η_o and β_o of η and β

respectively, we can find a differentiable function $\psi(\eta)$ such that $\psi(\eta_o)=\beta_o$ (i.e. β_o is expressible as a function of η_o) and

$$f_U(u;\eta_o,\beta_o) = f_U(u;\eta_o,\psi(\eta_o))$$

$$= f_U(u;\eta,\psi(\eta))$$

a.s. [u] for all η.

The intuitive idea in this definition is that the family of marginal distributions of u does not vary with η, i.e. this family is independent of η. Therefore no inference about η can be drawn based on the marginal distribution of u when β is unknown.

Example 2.2 (Liang,1983):

Let X_1 and X_2 be independent normally distributed random variables such that

$$X_1 \overset{d}{=} N(\eta+\beta,1),$$

and

$$X_2 \overset{d}{=} N(\beta,1).$$

Define $T=X_1+X_2$. Clearly T is normally distributed with mean $\eta+2\beta$ and variance 2. Then for any $(\eta_o,\beta_o) \in \mathbb{R}^2$ if we choose $\psi(\eta) = -\eta+\eta_o+\beta_o$, T is seen to be weakly ancillary for η. We note of course that T is also sufficient for β the accessory parameter. But since the distribution

of T depends on η as well, it does not satisfy the ancillarity of definition 2.1 (Note that in this example T is also weakly ancillary for β).

When $f(x;\theta)$ is from the Darmois-Koopman-Pitman class of the exponential family, i.e.

$$f(x;\theta) = c(\eta,\beta)\exp\{s(x,\eta) + u(x)\beta\}, \qquad (2.1)$$

we know (Lehmann, 1959, page 52) that the marginal density of U is

$$f_U(u;\eta,\beta) = c(\eta,\beta)\exp(u\beta)\,\gamma(u,\eta),$$

where $\gamma(u,\eta)$ is the integral of $\exp\{s(x,\eta)\}$ over all x such that U(x)=u. Andersen(1970) proved the following lemma which characterises weak ancillarity in this case.

Lemma 2.1:

The statistic U in (2.1) is weakly ancillary if and only if

$$\log \gamma(u,\eta) = a(\eta)u + b(\eta) + d(u),$$

where a and b are functions of η only.

We give an example to illustrate the value of this lemma.

Example 2.3 (Andersen,1970):

Let X_i and Y_i be independent Poisson variates with means $\exp(\eta+\beta_i)$ and $\exp(\beta_i)$ respectively (i=1,...,n). Define $T_i=X_i+Y_i$. It is

easy to see that the conditional distribution of (X_i, Y_i) given T_i is free of β_i. Furthermore, the marginal distribution of T_i is

$$f_T(t_i;\eta,\beta_i) = \frac{e^{-e^{\beta_i}(1+e^{\eta})} \; e^{\beta_i t} \; (1+e^{\eta})^{t_i}}{t_i!}.$$

Then

$$\log \gamma(t_i,\eta) = t_i \log(1+e^{\eta}) - \log(t_i!).$$

Therefore t_i is weakly ancillary for η. Moreover a simple investigation of the mean of t_i shows that the choice of $\psi_i(\eta)$ is

$$\psi_i(\eta) = -\log(1+e^{\eta}).$$

These examples show that inference about η should be drawn from the conditional distribution given the weak ancillary U. Moreover U is sufficient for the accessory parameter so that the conditional distribution is free of β and thus available for inference.

Our next definition is due to Cox(1958) (also in Cox and Hinkley, 1974, page 31-32). This is an extension of the definition of ancillarity for a single parameter model, contained in the same paper. Let S=(T,U) be minimal sufficient in the one parameter probability model. Then if U is distributed free of the structural parameter, U is said to be ancillary for the parameter. It is also required that U is of maximum dimension i.e. a maximal ancillary. Such U may also be

called 'S-contained' or 'internal' because it is a function of the (minimal) sufficient statistic.

Clearly, according to Basu's(1955, theorem 2) theorem, we cannot have an S-contained ancillary if S is (boundedly) complete since then ancillary statistics are independent of S. This point is further clarified in Lehmann(1981, sections 2 and 4).

Definition 2.3 (Cox,1958):

Let $S=(T,V,U)$ be minimal sufficient for $\theta=(\eta,\beta)$ so that

(i) There exist functions $m(T,\eta)$ such that the conditional distribution $f(m;\eta|u)$ is independent of β. Furthermore, T,V and U are of the maximum dimension permitting this independence.

(ii) Any of the following conditions holds with respect to the density of U, $f_U(u;\eta,\beta)$:

(a) $f_U(u;\eta,\beta)$ is free of η, i.e. $f_U(u;\eta,\beta) = f_U(u;\beta)$.

(b) Given any pair of values η_1, η_2 and any u, the ratio $f_U(u;\eta_1,\beta)/f_U(u;\eta_2,\beta)$ runs through all the positive values as β varies.

(c) If $f_V(v;\eta,\beta)$ and $f_U(u;\eta,\beta)$ are the marginal densities of V and U respectively, then given values η_1, η_2 of η, there will exist admissible values β_1, β_2 of β such that

$$\frac{f_V(v;\eta_1,\beta)}{f_V(v;\eta_2,\beta)} = \frac{f_V(v;\eta,\beta_1)}{f_V(v;\eta,\beta_2)}$$

and

$$\frac{f_U(u;\eta_1,\beta)}{f_U(u;\eta_2,\beta)} = \frac{f_U(u;\eta,\beta_1)}{f_U(u;\eta,\beta_2)},$$

i.e. any intended distinction between values of η can be regarded as a distinction between values of β. Hence U and V provide no direct information on η.

The U thus defined is called an (S-contained) ancillary statistic for η, and inference about η should be based on $f(m;\eta|u)$.

From this definition U should be a maximal ancillary. This is the main purpose of the requirement on dimension in condition (i). Condition (ii) ensures the uninformativeness with which we shall have much to do in the discussion in section 2.3 as we compare this definition with others. V is the residual statistic when S is partitioned into T and U and need not actually exist.

Example 2.4 (Cox and Hinkley, 1974, page 32-33):

Let $Y_1,\ldots,Y_n$ be an independent normally distributed sample with

$$Y_j \overset{d}{\sim} N(\gamma+\beta X_j, \sigma^2),$$

and X has probability density function f(x). Then

$S=(\hat{\gamma},\hat{\beta},SS_{res},\sum X_j,\sum X_j^2)$ is minimal sufficient where $SS_{res}=\sum(Y_j-\hat{\gamma}-\hat{\beta}X_j)^2$, and $\hat{\gamma}$ and $\hat{\beta}$ are the maximum likelihood estimators for γ and β.

If we let $T=(\hat{\gamma},\hat{\beta})$, $V=SS_{res}$ and $U=(\sum X_j,\sum X_j^2)$, U will be an S-contained ancillary for (γ,β). U is distributed free of all parameters so that condition (ii) is trivially satisfied.

Example 2.5 (Cox,1958):

Let X and Y be two independent Poisson distributed random variables with means λ_1 and λ_2 respectively. Define $\eta=\lambda_2/\lambda_1$ so that $\beta\eta=\lambda_2$ (or $\beta=\lambda_1$). The joint likelihood of X and Y is

$$\frac{e^{-\beta}\beta^x}{x!}\,\frac{e^{-\beta\eta}(\beta\eta)^y}{y!} = \frac{e^{-\beta(1+\eta)}\{\beta(1+\eta)\}^s}{s!}\,\frac{s!}{t!(s-t)!}\left(\frac{1}{1+\eta}\right)^t\left(\frac{\eta}{1+\eta}\right)^{s-t}$$

where $S=X+Y$, $T=X$. S is an S-contained ancillary for η since we may redefine the parameters so that $\beta^*=\beta(1+\eta)$. Clearly, we can choose some β_o for any given η_o such that $\beta(1+\eta_o)=\beta_o(1+\eta)$; thus leaving the likelihood ratio unaltered. Condition (ii)(c) is thereby satisfied.

Sprott(1975) proposed definitions for marginal and conditional sufficiency. But as we shall make clear, these can be adopted as definitions of ancillarity. Motivation for these definitions is the property of the likelihood ratio that it is minimal sufficient. Therefore, defining a statistic that is 'in some sense' independent of the likelihood ratio obtains for us another definition of ancillarity in the presence of accessory parameters.

Definition 2.4 (Sprott,1975):

Let $X=(X_1,\ldots,X_n)$ be summarised in (U,V) whose distribution factorises as follows:

$$f(u,v;\eta,\beta) = f_U(u;\beta).f_C(v;\eta,\beta|u);$$

and

either

(a) given any value β_o of β, $f_C(v;\eta,\beta|u)/f_C(v;\eta,\beta_o|u)$ is independent of v,

or

(b) we can find a function $m(v,\eta,\beta)$ such that

(i) given any values m_o, v_o, v_1, η_o of m, v, η respectively we can find β_o, β_1 such that

$$m(v_o,\eta_o,\beta_o) = m(v_1,\eta_o,\beta_1) = m_o$$

and the distribution of m is independent of β,

(ii) $f_C(v;\eta,\beta|u)/f_C(v;\eta,\beta_o|u)$ is a function of β, β_o and m only.

Then U is ancillary for η.

We note that condition (b)(i) implies that two different values of v can be made to yield the same value of m by an appropriate choice of β.

Clearly, even without conditions (a) or (b), U is already ancillary for η since U is distributed free of η. However, these conditions enable the conditional distribution given u to be

uninformative toward β thus facilitating useful separation of the parameters. This does not of course mean that U is maximal ancillary. For, if the underlying probability model is complete, say, by Basu's(1955) theorem we would expect any ancillary statistic to be independent of the likelihood ratio.

Further, from Sprott(1975) we have

Definition 2.5:

Assume $X=(X_1,\ldots,X_n)$ is summarised in (U,V) such that

$$f(u,v;\eta,\beta) = f_U(u;\eta,\beta).f_C(v;\eta|u);$$

and

either

(a) given any value η_o of η, $f_U(u;\eta,\beta)/f_U(u;\eta_o,\beta)$ is independent of u,

or

(b) we can find a function $n(u,\eta,\beta)$ such that

(i) given any n_o, u_o, u_1 and η_o values of n,u and η respectively, there exist β_o, β_1 such that

$$n(u_o,\eta_o,\beta_o) = n(u_1,\eta_o,\beta_1) = n_o$$

and the distribution of n is independent of β,

(ii) $f_U(u;\eta,\beta)/f_U(u;\eta_o,\beta)$ is a function of η, η_o and n only.

Then U is ancillary for η.

In our later discussion, we shall find definition 2.5 more useful than definition 2.4. From definition 2.5, U is functionally independent of the likelihood ratio in (a) which therefore achieves our purpose. In (b) however, part (i)implies that U can be made to yield the same value of n by a convenient choice of β i.e. U cannot be used to distinguish between values of η when β is unknown. In fact if $g(n;\theta)$ is the probability density function of n, $g(n_o;\eta)/g(n_1;\eta)$ is independent of U since some choice of β will satisfy the values of n_o and n_1 irrespective of the value of U. It follows by (ii) that the likelihood ratio for η on η_o is essentially independent of U since it depends on n, η and η_o only.

Example 2.6 (Godambe,1980):

Let X be a gamma random variable such that X_1 and X_2 are observations on X. The joint likelihood is given by

$$L(\eta,\beta;x_1,x_2) = \frac{\beta^{2\eta}\, e^{-\beta(x_1+x_2)}\, (x_1x_2)^{\eta-1}}{\{\Gamma(\eta)\}^2},$$

$(\eta,\beta) \in (0,\infty)\times(0,\infty)$. Then (X_1X_2, X_1+X_2) is minimal sufficient for (η,β). Let $U=X_1+X_2$. The probability density function of U is

$$f_U(u;\eta,\beta) = \frac{\beta^{2\eta}\, e^{-\beta u}\, u^{(2\eta-1)}}{\Gamma(2\eta)}, \quad u>0,$$

and the likelihood ratio is

$$\frac{f_U(u;\eta_1,\beta)}{f_U(u;\eta_2,\beta)} = \frac{(\beta u)^{2(\eta_1-\eta_2)}\, \Gamma(2\eta_2)}{\Gamma(2\eta_1)}.$$

By choosing $n=\beta u$, we see that given any values of u and η_1, some β exists appropriately chosen to satisfy any value of n, i.e. $\beta=n/u$. The probability density function of n is

$$\frac{1}{\beta} f_U(\frac{u}{\beta};\eta,\beta) = \frac{e^{-u}\, u^{2\eta-1}}{\Gamma(2\eta)}, \quad u>0,$$

which is independent of β. So (b)(i) is satisfied. The likelihood ratio above can be written as

$$\frac{n^{2(\eta_1-\eta_2)}\,\Gamma(2\eta_2)}{\Gamma(2\eta_1)},$$

a function of η_1,η_2 and n only as required in (b)(ii). Thus U is ancillary. We shall later show that U is not ancillary by definition 2.3.

We now present yet another definition of ancillarity due to Godambe(1980).

Definition 2.6:

U is ancillary for η 'ignoring β' if

(i) The conditional distribution of X given U depends on $\theta(=(\eta,\beta))$ only through η, i.e.

$$f(x;\eta,\beta) = f_U(u;\eta,\beta).f_C(x;\eta|u).$$

(ii) The class of marginal distributions of U

$$H = \{f_U(u;\eta,\beta) \mid (\eta,\beta)\varepsilon\Omega\}$$

is complete for each fixed η.

Example 2.7 (Godambe,1980):

Let $X_o,\ldots,X_n$ be independent and identically distributed Poisson random variables such that

$$f(x_i;\lambda) = \frac{e^{-\theta_2\theta_1^i}(\theta_2\theta_1^i)^{x_i}}{x_i!}, \qquad i=0,1,\ldots,n$$

and $\lambda=(\theta_1,\theta_2)\ \varepsilon\ (0,\infty)\times(0,\infty)$. If we write

$$\phi = \theta_2\left(\frac{1-\theta_1^{n+1}}{1-\theta_1}\right),$$

it is easy to see that the marginal distribution of $U=\sum X_i$ is

$$f_U(u;\theta_1,\theta_2) = \frac{e^{-\phi}\,\phi^u}{u!},$$

so that the distribution of X given U is

$$f_C(x;\theta_1|u) = \frac{u!(1-\theta_1)^u\,\theta_1^{\sum ix_i}}{(1-\theta_1^{n+1})^u\,\Pi x_i!}$$

which is free of θ_2.

A second example of this type of ancillarity is example 2.6 given above with $U=X_1+X_2$ ancillary.

Lehmann and Scheffé(1950, examples 3.5 and 3.8) have shown the completeness of U in both these examples. So by Godambe's definition, U is ancillary in both cases.

We make the following remarks on definition 2.6:

a) Since in condition (i) the accessory parameter β is eliminated upon conditioning on U, U is sufficient for β. Basu(1955,1958) and Lehmann(1981) clarify the role of completeness in (ii), showing that a complete sufficient statistic separates out ancillary information by making the ancillary part of the data independent of the sufficient statistic. Therefore in the definition completeness of U rids it of any ancillary information about β.

(b) It is not clear what the phrase 'ignoring β' means. The definition therefore remains vague in this phrase.

(c) Unlike the earlier definitions this one is based on the inability to extract information from the distribution of the ancillary statistic. However, if the structural parameter is inextricably mixed up with the accessory parameter, this is not conclusive evidence toward lack of information or even ancillarity. This demands further justification which is lacking in Godambe's definition.

Since this definition is based on 'unavailability of information' we shall see that it is closely related to the definition we now proceed to present, due to Johansen(1976):

First of all we introduce the concept of 'easily available information' on which Johansen based this definition. (U,T) is said to contain 'easily available information' about η if Borel functions v and w exist such that

$$f_U(u;\eta,\beta) = f_1(u;\eta,\beta|v).f_2(v;\eta|w).f_3(w;\eta,\beta|t).f_4(t;\eta,\beta).$$

Definition 2.7:

A statistic U is ancillary for η if (U,1) contains no 'easily available information' on η; i.e. if we cannot find some Borel functions v and w such that

$$f_U(u;\eta,\beta) = f_1(u;\eta,\beta|v).f_2(v;\eta|w).f_3(w;\eta,\beta).$$

As justification for this definition, Johansen(1976) proved the following theorem.

Theorem 2.1:

Let the family $F=\{f(x;\eta,\beta),\ x\varepsilon\mathcal{X}\ |(\eta,\beta)\varepsilon\Omega_1\times\Omega_2\}$ of distributions of X be complete for each fixed $\eta=\eta_o$ and let there exist some U such that

$$f(x;\eta,\beta) = f_U(u;\eta,\beta).f_C(x;\eta|u), \qquad (2.2)$$

i.e. U is sufficient for the accessory parameter β. Then if there exist Borel functions a(u) and b(u) such that

$$f_U(u;\eta,\beta) = f_1(u;\eta,\beta|a(u)).f_2(a(u);\eta|b(u)).f_3(b(u);\eta,\beta),$$

the distribution $f_2(a(u);\eta|b(u))$ is a one-point measure.

We may rephrase this result as follows: if the family F be complete for fixed $\eta=\eta_o$ and equation (2.2) above is true, no non-trivial Borel functions of U, a(u) and b(u) exist such that (U,1) contains 'easily available information' about η, i.e. U is ancillary according to definition 2.7. This forms a bridgehead with definition 2.6 and this result is formalised in proposition 2.7. Theorem 2.1 therefore serves to clarify the basis of definition 2.6 as well.

Examples 2.6 and 2.7 illustrate this definition.

We note the following consequent results from theorem 2.1.

(i) The case when the family F is complete for fixed $\eta=\eta_o$ provides us with a requirement on this definition. Then the sufficiency of U for β implies ancillarity of U for η. When F is not complete, the definition may be weak in application as we cannot always check for the 'non-existence' of functions a(u) and b(u).

(ii) This theorem is also significant for other definitions. We recall that the purpose in all definitions is that inference be based on the conditional model $f_C(x;\eta|u)$ in (2.2) and we want $f_U(u;\eta,\beta)$ to be independent of η in some way. Under the conditions in theorem 2.1, no further useful reduction of the distributions of U is possible. Therefore for complete families, this definition (or equivalently definition 2.6) will be at least as relevant in inference as any of

the others. For, whatever conditions we impose on $f_U(u;\eta,\beta)$-which is the essence of the earlier definitions-it will remain the simplest reduction in η possible in the distributions of U.

Barndorff-Nielsen(1973) defined 'M-ancillarity' based on the notion of 'universality.'

Let $\mathbf{P}_\theta$ be a family of probability measures indexed with $\theta\varepsilon\Omega$. $\mathbf{P}_\theta$ is said to be universal if given any x´ from the sample space $\mathbf{X}$, we may choose $\theta_o\varepsilon\Omega$ such that

$$P(x';\theta_o) \geq P(x;\theta_o)$$

for all x, and for all $P\varepsilon\ \mathbf{P}_\theta$.

Example 2.8:

Let X be a binomial random variable with probability function

$$b(x;n,p) = \binom{n}{x} p^x(1-p)^{n-x}, \quad x=0,1,\ldots,n$$

and n is fixed. Let n=5, say and p_o be the value of p at which b(x;n,p) is maximum.

If x=2, $p_o=2/5$,

if x=1, $p_o=1/5$, etc.

This family too is universal.

Definition 2.8:

A statistic U is M-ancillary for $\eta=\eta(\theta)$ if

(i) $f(x;\theta) = f_U(u;\theta).f_C(x;\eta|u)$,

where f_U is the marginal density of U, and

(ii) For each fixed $\eta=\eta_o$, the family

$$G = \{f_U(u;\theta)|\eta(\theta)=\eta_o\}$$

is universal.

Barndorff-Nielsen argues that since we can make any value of U the mode by an appropriate choice of η, then U is uninformative about η and hence ancillary for η. It is therefore the condition of universality in (ii) of the definition that is considered definitive of uninformativeness in M-ancillarity. This would mean that X in example 2.8 is uninformative about p. This is not altogether true.

Johansen(1977) shows that M-ancillarity may be unreasonable with the following example.

Example 2.9:

Let X and Y be independent variables with

$$P(X=1) = p,\ P(X=0) = q;\ p+q=1$$

$$P(Y=-1) = a,\ P(Y=0) = q,\ P(Y=1) = p-a.$$

Let $\eta(a,p) = p$ and

$$\Omega = \{(a,p)|0 \leq a \leq p,\ 1/2 \leq p \leq 2/3\}.$$

Then Y=-1 is the mode for a=p, Y=0 is the mode for a=p/2 and Y=1 is the mode for a=0. Therefore Y is universal.

Moreover (X,Y) is minimal sufficient for (a,p) such that

$$f(x,y;a,p) = f_X(x;p).f_Y(y;a,p)$$

$$= f_{X|Y}(x;p|y).f_Y(y;a,p)$$

since X and Y are independent. Thus Y is sufficient for a, the accessory parameter. If we choose U(X,Y)=Y, U is M-ancillary for p. Thus inference about p should be conditional on U i.e. based on X alone.

But the distribution of Y^2 depends on p alone so that X and Y^2 are independent and identically distributed random variables. Furthermore, (X,Y^2) will contain twice as much information on p as X; while conditioning on U utilises only half of that information. So M-ancillarity may yield unreasonable results. In the light of this, our discussion that now follows will not further consider M-ancillarity.

2.3 Relationships Between The Definitions.

Since the concept and purpose of ancillarity are clear, i.e. that the distribution of an ancillary statistic is 'in some sense' independent of the structural parameter and inference is made conditional on the ancillary statistic, there are necessarily similarities in the way these objectives are achieved in the definitions. The work in this section explores these similarities. Proposition 2.7 has previously been proved in Gordon(1981); the rest are original.

In definition 2.3 it is condition (ii) which is meant to define an effective ancillary statistic in inference. Condition (i) ensures that the ancillary statistic is S-contained and maximal. Therefore only condition (ii) shall be used for the discussion in this section.

Proposition 2.1:

Definition 2.1 implies definition 2.2.

Proof:

From definition 2.1, U is ancillary for η if

$$f_X(x;\eta,\beta) = f_U(u;\beta).f_C(x;\eta|u).$$

If we take any values η_o and β_o of η and β, and define a function $\psi(\eta)$ such that $\psi(\eta)=\beta_o$ for all η, clearly

$$f_U(u;\beta_o) = f_U(u;\psi(\eta)),$$

since the marginal distribution of U is free of η. So U is 'weakly ancillary' for η.

Similarly condition (ii)(a) of definition 2.3 will imply definition 2.2.

Proposition 2.2

Definition 2.2 implies condition (ii)(c) of definition 2.3.

Proof:

Let U be weakly ancillary. Then given η_o, β_o, we can find a differentiable function $\psi'(\eta)$ with $\psi'(\eta_o)=\beta_o$ such that

$$f_U(u;\eta_o,\beta_o) = f_U(u;\eta,\psi'(\eta))$$

a.s. [u]. As well, given η_1, β_o, we can still find a differentiable function $\psi^*(\eta)$ with $\psi^*(\eta_1)=\beta_o$ such that

$$f_U(u;\eta_1,\beta_o) = f_U(u;\eta,\psi^*(\eta))$$

a.s. [u]. So

$$\frac{f_U(u;\eta_o,\beta_o)}{f_U(u;\eta_1,\beta_o)} = \frac{f_U(u;\eta,\psi'(\eta))}{f_U(u;\eta,\psi^*(\eta))}.$$

Since this is possible for any choice of $\beta=\beta_o$, we may write

$$\frac{f_U(u;\eta_o,\beta)}{f_U(u;\eta_1,\beta)} = \frac{f_U(u;\eta,\psi_1(\eta))}{f_U(u;\eta,\psi_2(\eta))}$$

$$= \frac{f_U(u;\eta,\beta_1)}{f_U(u;\eta,\beta_2)}$$

which is condition (ii)(c) of definition 2.3.

As a corollary, definition 2.2 will imply definition 2.3 if U is an S-contained weak ancillary. Furthermore, we can conclude from these two propositions that definition 2.1 implies condition (ii)(c) of definition 2.3; the direct proof to this conclusion is trivial.

Proposition 2.3:

Definition 2.2 implies definition 2.5.

Proof:

Let U be weakly ancillary for η in the model $f(x;\eta,\beta)$. Then for any values η_o and β_o of η and β, a differentiable function $\psi'(\eta)$ exists such that $\psi'(\eta_o)=\beta_o$ and

$$f_U(u;\eta_o,\beta_o) = f_U(u;\eta,\psi'(\eta))$$

a.s. [u]; i.e.

$$f_U(u;\eta_o,\psi'(\eta_o)) = f_U(u;\eta,\psi'(\eta))$$

a.s. [u]. This is true for any β_o such that a differentiable function satisfying this equality exists. Since β_o can be chosen arbitrarily, we write β instead of β_o to allow for variation. Thus for any β value, a differentiable function $\psi^*(\eta)$ for which $\psi^*(\eta_o)=\beta$ will exist such that

$$f_U(u;\eta_o,\beta) = f_U(u;\eta,\psi^*(\eta)),$$

a.s. [u]. η is allowed to vary in the right hand side of the equality so that $\psi^*(\eta)=\beta$ at some η value (one such value is η_o but there could be others). Then

$$\frac{f_U(u;\eta_o,\beta)}{f_U(u;\eta,\beta)} = 1$$

independent of U. Hence U is ancillary according to definition 2.5.

From propositions 2.1 and 2.3, we can conclude that Fraser's definition 2.1 implies definition 2.5 too; the direct proof to this is also trivial.

Proposition 2.4;

Condition (ii)(a) of definition 2.3 implies definition 2.5.

Proof;

This proposition follows easily from a similar result that definition 2.1 implies definition 2.5, noted above since in condition (ii)(a) of definition 2.3 we have

$$f_U(u;\eta,\beta) = f_U(u;\beta),$$

as for definition 2.1.

To give further clarification on the difference between condition (ii)(c) of definition 2.3 and definition 2.5, we give two examples both quoted earlier. They show that neither definition need imply the other.

Example 2.5 (contd.);

The probability function for the ancillary statistic was found to be

$$f_U(u;\eta,\beta) = \frac{e^{-\beta(1+\eta)}\ \{\beta(1+\eta)\}^u}{u!},$$

and hence the likelihood ratio of η_1 to η_2 is

$$\frac{f_U(u;\eta_1,\beta)}{f_U(u;\eta_2,\beta)} = e^{-\beta(\eta_1-\eta_2)} \left(\frac{1+\eta_1}{1+\eta_2}\right)u.$$

This evidently satisfies neither condition (a) nor (b) of definition 2.5, although it satisfies condition (ii)(c) in definition 2.3.

Example 2.6 (contd.):

This example was quoted to illustrate definition 2.5. The likelihood ratio for $\eta_1:\eta_2$ is

$$\frac{f_U(u;\eta_1,\beta)}{f_U(u;\eta_2,\beta)} = (\beta u)^{2(\eta_1-\eta_2)} \frac{\Gamma(2\eta_2)}{\Gamma(2\eta_1)};$$

while for $\beta_1:\beta_2$ it is

$$\frac{f_U(u;\eta,\beta_1)}{f_U(u;\eta,\beta_2)} = \left(\frac{\beta_1}{\beta_2}\right)^{2\eta} e^{-u(\beta_1-\beta_2)}.$$

Condition (ii)(c) in definition 2.3 requires that these ratios be equal for some β_1 and β_2. Clearly this will not always be possible. Therefore definition 2.3 is not satisfied.

It is interesting to note however that this latter example satisfies condition (ii)(b) of definition 2.3. The relationship between this condition and definition 2.5 is as follows:

Proposition 2.5:

Condition (ii)(b) of definition 2.3 implies definition 2.5.

Proof:

By condition (ii)(b) of definition 2.3, $f_U(u;\eta_1,\beta)/f_U(u;\eta_2,\beta)$ runs through all the positive values as β varies. This is irrespective of the value of U. Thus this likelihood ratio is independent of U and definition 2.5 is satisfied.

Proposition 2.6:

Definition 2.1 implies definition 2.7.
(Similarly condition (ii)(a) of definition 2.3 will imply definition 2.7).

Proof:

If U does not satisfy definition 2.7, there must exist v and w such that

$$f_U(u;\eta,\beta) = f_1(u;\eta,\beta|v).f_2(v;\eta|w).f_3(w;\eta,\beta).$$

However, in definition 2.1, $f_U(u;\eta,\beta)=f_U(u;\beta)$ so that such v and w cannot exist. This proves the proposition.

Definitions 2.6 and 2.7 are similarly based on the concept of completeness and therefore on the concept of 'unavailable information.' The relationship between them was formalised by Gordon(1981) in the following proposition.

Proposition 2.7:

The ancillarity of definition 2.6 implies that of definition 2.7.

Proof:

It suffices to show that if the class $H=\{f_U(u;\eta,\beta)|(\eta,\beta)\epsilon\Omega\}$ of marginal distributions of U is complete for each fixed η, then (U,1) contains no 'easily available information' on η, thus satisfying definition 2.7.

Let g and h be functions of U such that

$$f_1(g;\eta,\beta|h) = f_1(g;\eta|h).$$

According to theorem 2.1, this distribution is a one-point measure. Thus no such g(u) and h(g(u)) exist for the equality above to hold in a non-trivial way. So (U,1) contains no 'easily available information' about η, and the proposition is proved.

2.4 Conclusion.

We have established some interrelationships between the various definitions of ancillarity in the presence of accessory parameters. In particular, it is instructive to classify together the definitions of Fraser(1956), Andersen(1970,1973), Cox(1958) and Sprott(1975) while the definitions of Godambe(1980) and Johansen(1976) are related through their requirement of 'unavailable information' achieved through completeness.

Therefore we compare the first four definitions. On the basis of their relative strengths, it is clear that definition 2.1 is the most restrictive while condition (ii)(c) of definition 2.3 and definition 2.5 appear to be the least restrictive. It is difficult to say which definition will be most useful in practice. However, in each

problem it seems advisable that the strongest eligible definition be applied as that would appear to offer the biggest reduction in the sample space; a fact illustrative of its strength.

Note too that the difference between condition (ii)(a) of definition 2.3 and definition 2.1 lies in the fact that the latter ancillary statistic is also sufficient for β.

On the other hand, definition 2.6 is stronger than definition 2.7. The usefulness of these two definitions in inference may be in doubt as may be deduced from the comments in section 2.2. It was pointed out to me however, that definition 2.1 will imply definition 2.6 in case of the exponential families since they are complete. It also follows that condition (ii)(a) in definition 2.3 will not imply definition 2.6 since we cannot have S-contained ancillaries in complete families. Nevertheless, theorem 2.1 gives an indication of how much reduction by ancillarity we can achieve in complete families.

The following table is given to assist in summarising the results of this chapter. A tick means that the definition in the row implies the definition in the corresponding column.

DEFINITION	2.1	2.3 C (ii) (a)	2.2	2.3 C (ii) (b)	2.3 C (ii) (c)	2.5	2.6	2.7
2.1		✓	✓		✓	✓	✓*	✓
2.3 C (ii) (a)			✓		✓	✓		✓
2.2					✓	✓		
2.3 C (ii) (b)						✓		
2.3 C (ii) (c)								
2.5								
2.6								✓
2.7								

where C ≡ Condition. *This holds for complete families.

3. THE PITMAN-MORGAN TEST AS A CONDITIONAL TEST.

3.1 Introduction.

Finney(1938) developed a significance test for the ratio of variances in a bivariate normal distribution when the correlation coefficient is known. Nevertheless his adaptation to the case when the correlation coefficient is unknown, and hence can only be estimated, is inadequate. Pitman(1939) and Morgan(1939), however, developed a suitable test criterion for the latter case. This is what we call the Pitman-Morgan test for variance ratios in a bivariate normal distribution.

In our work, the correlation coefficient, ρ, is an accessory parameter. By conditioning on a sufficient statistic for ρ to eliminate it and applying conditional inference procedures developed in Williams(1982), we show that the Pitman-Morgan test is, in fact, a conditional test. Successful application to such an important test criterion, lends credibility to conditional inference procedures as being able to give a significant alternative approach to problems in statistical inference. In this problem, sampling conditionally yields identical results to unconditional sampling. Here the test criterion is derived as a test for independence between sufficient and ancillary statistics. Furthermore, it is often possible to transform tests so that we test for such independence, with identical results.

3.2 The Pitman and Morgan Approaches.

Let X and Y be correlated variables from a normal distribution whose joint probability density function is given by

$$\frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}} \exp\{-\frac{1}{2(1-\rho^2)}[(\frac{x-\mu_1}{\sigma_1})^2-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+(\frac{y-\mu_2}{\sigma_2})^2]\},$$

where μ_1 and σ_1^2 are the mean and variance for X and similarly μ_2 and σ_2^2 are the mean and variance for Y. ρ is the correlation coefficient between X and Y.

Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be pairs of observations from this population. We need to define the sample means

$$\bar{X} = \sum X_i/n, \quad \bar{Y} = \sum Y_i/n$$

and the sample variances

$$S_1^2 = \sum(X_i-\bar{X})^2/n, \; S_2^2 = \sum(Y_i-\bar{Y})^2/n$$

and the sample correlation coefficient

$$r = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\{[\sum(X_i-\bar{X})^2][\sum(Y_i-\bar{Y})^2]\}^{\frac{1}{2}}}.$$

Then the joint probability density function for the n observations is

$$\{\frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}}\}^n \exp\{-\frac{n}{2(1-\rho^2)}[(\frac{\bar{x}-\mu_1}{\sigma_1})^2-2\rho\frac{(\bar{x}-\mu_1)(\bar{y}-\mu_2)}{\sigma_1\sigma_2}+(\frac{\bar{y}-\mu_2}{\sigma_2})^2$$

$$+\frac{s_1^2}{\sigma_1^2}-\frac{2\rho r s_1 s_2}{\sigma_1\sigma_2}+\frac{s_2^2}{\sigma_2^2}]\}, \qquad (3.1)$$

$-\infty < \mu_1, \mu_2 < \infty$, $0 < \sigma_1, \sigma_2 < \infty$, $-1 \leq \rho \leq 1$.

If we put

$$U_i = X_i/\sigma_1 + Y_i/\sigma_2,\ V_i = X_i/\sigma_1 - Y_i/\sigma_2,\ i = 1,\ldots,n;$$

we know that the correlation between U_i and V_i is zero. We can write their sample correlation coefficient as

$$R = \frac{\sum(U_i-\bar{U})(V_i-\bar{V})}{\{\sum(U_i-\bar{U})^2.\sum(V_i-\bar{V})^2\}^{\frac{1}{2}}}$$

$$= \frac{s_1^2/\sigma_1^2 - s_2^2/\sigma_2^2}{\{(s_1^2/\sigma_1^2 + s_2^2/\sigma_2^2)^2 - 4r^2 s_1^2 s_2^2/\sigma_1^2\sigma_2^2\}^{\frac{1}{2}}}.$$

We put $\omega = s_1^2/s_2^2$ and $\psi = \sigma_1^2/\sigma_2^2$. Then

$$R = \frac{\omega-\psi}{\{(\omega+\psi)^2 - 4r^2\omega\psi\}^{\frac{1}{2}}}.$$

From this form of R, Pitman(1939) was able to write down a more appropriate test criterion for ψ since the distribution of R is known. This is

$$t = \frac{R(n-2)^{\frac{1}{2}}}{(1-R^2)^{\frac{1}{2}}}$$

$$= \frac{(\omega-\psi)(n-2)^{\frac{1}{2}}}{\{4(1-r^2)\omega\psi\}^{\frac{1}{2}}}$$

which has a 'Student's' t distribution with n-2 degrees of freedom.

Morgan(1939) used the likelihood ratio method to arrive at the same criterion. For if we assume that the null hypothesis $H_o:\sigma_1^2=\sigma_2^2=\sigma^2$ is true, the joint probability function for the n pairs of observations from (3.1) is

$$\{\frac{1}{2\pi\sigma^2(1-\rho^2)^{\frac{1}{2}}}\}^2\exp\{-\frac{n}{2(1-\rho^2)}[(\frac{\bar{x}-\mu_1}{\sigma})^2 - 2\rho\frac{(\bar{x}-\mu_1)(\bar{y}-\mu_2)}{\sigma^2} + (\frac{\bar{y}-\mu_2}{\sigma})^2$$

$$+ \frac{s_1^2}{\sigma^2} - \frac{2\rho r s_1 s_2}{\sigma^2} + \frac{s_2^2}{\sigma^2}]\}.$$

Hence the likelihood ratio is

$$\lambda = \{ 1- \frac{(s_1^2-s_2^2)^2}{(s_1^2 + s_2^2)^2 - 4r^2 s_1^2 s_2^2} \}^{n/2}$$

$$= \{1-R^2\}^{n/2}.$$

Thus λ depends on R alone which led Morgan(1939) to the test criterion

$$t = \frac{R(n-2)^{\frac{1}{2}}}{(1-R^2)^{\frac{1}{2}}}$$

as for Pitman(1939).

The t thus derived, is the Pitman-Morgan test statistic. H_o: $\sigma_1 = \sigma_2$ is rejected if |t| falls outside the desired probability level.

3.3 The Conditional Approach.

In both approaches above, it was necessary to derive a test either free of or independent of the accessory parameter ρ. This is inevitable for any adequate test criterion to be derived. Therefore in the conditional approach, accessory parameters are eliminated through conditioning on their respective sufficient statistics. The test statistic is derived using methods suggested by Williams(1982) to overcome the difficulty of conditioning on sufficient statistics which depend on the parameter of interest (i.e. on the structural parameter).

For brevity, we put

$$T_{11} = \sum X_i^2 \ , \ T_{22} = \sum Y_i^2 \ , \ T_{12} = \sum X_i Y_i \ ,$$

where

$$\text{Var}(X) = \sigma_1^2/(1-\rho^2)$$

$$\text{Var}(Y) = \sigma_2^2/(1-\rho^2).$$

We shall assume (without loss of generality) that X and Y have zero means. (We may note in passing that if the means of X and Y were μ_1 and μ_2 (both nonzero), they could be eliminated either by conditioning on their sufficient statistics, $\bar{X}$ and $\bar{Y}$, or by writing

$X' = X-\mu_1$ and $Y' = Y-\mu_2$ if μ_1 and μ_2 are known. X' and Y' are pivotal quantities ancillary for μ_1 and μ_2 respectively). The bivariate normal density

$$\frac{(1-\rho^2)^{\frac{1}{2}}}{2\pi\sigma_1\sigma_2} \exp\{-\frac{1}{2}(\frac{x_i^2}{\sigma_1^2} - 2\rho\frac{x_i y_i}{\sigma_1\sigma_2} + \frac{y_i^2}{\sigma_2^2})\}$$

becomes (for the joint probability function of T_{11}, T_{22} and T_{12})

$$f(t_{11},t_{22},t_{12};\rho) =$$

$$k(1-\rho^2)^{n/2} (t_{11}t_{22}-t_{12}^2)^{(n-3)/2} \exp\{-\frac{1}{2}(\frac{t_{11}}{\sigma_1^2} - 2\rho\frac{t_{12}}{\sigma_1\sigma_2} + \frac{t_{22}}{\sigma_2^2})\}, \qquad (3.2)$$

k being a constant. Clearly, t_{12} is sufficient for ρ. We apply Madow's(1945) method for densities of sufficient statistics to derive the marginal distribution of t_{12}. For then

$$f(t_{11},t_{22},t_{12};\rho) = f_{T_{12}}(t_{12};\rho).f_C(t_{11},t_{22}|t_{12}),$$

the second factor on the right hand side being free of ρ upon conditioning on $T_{12} = t_{12}$. Let ρ have a specific value, $\rho_o = 0$, say. Then

$$\frac{f(t_{11},t_{22},t_{12};0)}{f(t_{11},t_{22},t_{12};\rho)} = \frac{f(t_{12};0)}{f(t_{12};\rho)}. \qquad (3.3)$$

The marginal density of T_{12} is, from (3.2) and (3.3),

$$f(t_{12};\rho) = \frac{f(t_{11},t_{22},t_{12};\rho)}{f(t_{11},t_{22},t_{12};0)}.f(t_{12};0)$$

$$= (1-\rho^2)^{n/2} e^{\rho t_{12}/\sigma_1\sigma_2}.f(t_{12};0). \qquad (3.4)$$

But

$$T_{12} = \sum_1^n X_i Y_i \sim X_1\sqrt{\sum Y_i^2}.$$

So for $\rho = 0$, T_{12} is distributed as a product of a normal and an independent χ_n variate, and thus has the Bessel distribution given by

$$\frac{(t_{12}/\sigma_1\sigma_2)^{(n-1)/2} K_{(n-1)/2}(t_{12}/\sigma_1\sigma_2)}{\sigma_1\sigma_2\, 2^{(n-1)/2}\, \Gamma(\tfrac{1}{2})\Gamma(n/2)},$$

where $K_{(n-1)/2}(t_{12}/\sigma_1\sigma_2)$ is a Bessel function. Therefore the marginal density of T_{12} in (3.4) is

$$f(t_{12};\rho) = \frac{(1-\rho^2)^{n/2} e^{\rho t_{12}/\sigma_1\sigma_2} (t_{12}/\sigma_1\sigma_2)^{(n-1)/2} K_{(n-2)/2}(t_{12}/\sigma_1\sigma_2)}{\sigma_1\sigma_2\, 2^{(n-1)/2}\, \Gamma(\tfrac{1}{2})\, \Gamma(n/2)},$$

(3.5)

a non-central density function. The conditional density of (T_{11}, T_{22}) given $T_{12} = t_{12}$ is, using (3.2) and (3.5),

$$f(t_{11},t_{22}|t_{12}) = c\,\frac{(t_{11}t_{22}-t_{12}^2)^{(n-3)/2}\; e^{-\frac{1}{2}(t_{11}/\sigma_1^2 + t_{22}/\sigma_2^2)}}{(t_{12}/\sigma_1\sigma_2)^{(n-1)/2}\; K_{(n-1)/2}(t_{12}/\sigma_1\sigma_2)},$$

(3.6)

where

$$c = k\sigma_1\sigma_2\; 2^{(n-1)/2}.\Gamma(\tfrac{1}{2}).\Gamma(n/2).$$

The conditional density in (3.6) can be written as

$$c\,\frac{(t_{11}t_{22}-t_{12}^2)^{(n-3)/2}\; e^{-\frac{1}{2\sigma_1^2}(t_{11} + t_{22}\sigma_1^2/\sigma_2^2)}}{(t_{12}/\sigma_1\sigma_2)^{(n-1)/2}\;.\; K_{(n-1)/2}(t_{12}/\sigma_1\sigma_2)}. \quad (3.7)$$

We reparametrize with $\psi = \sigma_1^2/\sigma_2^2$ and $\phi = 1/\sigma_1^2$. Clearly $S = T_{11} + \psi T_{22}$ is sufficient for the accessory parameter ϕ while S also depends on the structural parameter ψ. This means that if we attempt to condition on S to eliminate ϕ, the differential element in the conditional distribution given S will depend on ψ. For the difficulties involved in this, we refer to Kalbfleisch and Sprott(1970). We shall apply the procedures developed in Williams(1982) to overcome this problem. Using these procedures, an ancillary statistic for the accessory parameter ϕ shall be derived as a statistic independent of S. This ancillary statistic then becomes the basis for inference about ϕ.

We obtain the 'marginal' density of S. Write $T = T_{11}$ so that $T_{22} = (S-T)/\psi$. It is easy to see from (3.7) that

$$h(s,t|t_{12}) = c'(t(s-t) - \psi t_{12}^2)^{(n-3)/2},$$

where

$$c' = \frac{ce^{-\phi s/2}}{\psi^{(n-5)/2}(t_{12}/\sigma_1\sigma_2)^{(n-1)/2}K_{(n-1)/2}(t_{12}/\sigma_1\sigma_2)}.$$

Since this is a density,

$$t(s-t)-t_{12}^2\psi > 0$$

and so

$$|t - s/2| < (s^2/4 - \psi t_{12}^2)^{1/2}.$$

The 'marginal' density of S is therefore

$$h_1(s) = c' \int_{s/2 - (s^2/4 - \psi t_{12}^2)^{1/2}}^{s/2 + (s^2/4 - \psi t_{12}^2)^{1/2}} (ts - t^2 - \psi t_{12}^2)^{(n-3)/2}\, dt$$

$$= 2c' \int_0^{(s^2/4 - \psi t_{12}^2)^{1/2}} \{(s/2+t)(s/2-t) - \psi t_{12}^2\}^{(n-3)/2}\, dt$$

$$= 2c' \int_0^{(s^2/4 - \psi t_{12}^2)^{1/2}} (s^2/4 - t^2 - \psi t_{12}^2)^{(n-3)/2}\, dt.$$

If we put

$$U = \frac{T}{(S^2/4 - \psi T_{12}^2)^{\frac{1}{2}}},$$

then

$$h_1(s) = 2c' \int_0^1 (s^2/4-\psi t_{12}^2)^{(n-3)/2} (1-u^2)^{(n-3)/2} (s^2/4-\psi t_{12}^2)^{\frac{1}{2}} du,$$

and letting $Z = U^2$, it is easy to show that the 'marginal' density of S simplifies to

$$h_1(s) = c' (s^2/4 - \psi t_{12}^2)^{(n-2)/2}.\beta(\tfrac{1}{2},(n-1)/2)$$

where β is the beta function. Hence the conditional density of T given T_{12} and S is

$$h_2(t|t_{12},s) = \frac{(t(s-t) - \psi t_{12}^2)^{(n-3)/2}}{(s^2/4 - \psi t_{12}^2)^{(n-2)/2}.\beta(\tfrac{1}{2},(n-1)/2)}.$$

Putting

$$V = \frac{T - S/2}{(S^2/4 - \psi T_{12}^2)^{\frac{1}{2}}},$$

it is easy to see that

$$h_2(v|t_{12},s) = \frac{(1-v^2)^{(n-3)/2}}{\beta(\tfrac{1}{2},(n-1)/2)}.$$

As t runs from 0 to t_o, v goes from $-(s/2)/(s^2/4 - \psi t_{12}^2)^{\frac{1}{2}}$ to $(t_o - s/2)/(s^2/4 - \psi t_{12}^2)^{\frac{1}{2}}$ so that the conditional distribution function of V given T_{12} and S varies with $(t - s/2)/(s^2/4 - \psi t_{12}^2)^{\frac{1}{2}}$ alone. Therefore $(t - s/2)/(s^2/4 - \psi t_{12}^2)^{\frac{1}{2}}$ (or a function of it) is independent of T_{12} and S, sufficient statistics for the accessory parameters, ρ and ϕ, and consequently is independent of the accessory parameters. Williams(1982) therefore suggests that inference about ψ be based on such a statistic. It is easy to see that

$$\frac{t - s/2}{(s^2/4 - \psi t_{12}^2)^{\frac{1}{2}}} = \frac{t_{11}/t_{22} - \psi}{\{4\psi t_{11}/t_{22}(1 - t_{12}^2/t_{11}t_{22}) + (t_{11}/t_{22} - \psi)^2\}^{\frac{1}{2}}}.$$

In our earlier notation, $\omega = t_{11}/t_{22}$ and $r^2 = t_{12}^2/t_{11}t_{22}$. Therefore this statistic simplifies to

$$\frac{(\omega - \psi)/(4\omega\psi(1 - r^2))^{\frac{1}{2}}}{\sqrt{\{1 + (\omega - \psi)/(4\omega\psi(1 - r^2))^{\frac{1}{2}}\}}}$$

so that the conditional distribution function is a function of the term

$$\frac{\omega - \psi}{(4\omega\psi(1 - r^2))^{\frac{1}{2}}}.$$

Therefore, inference on the variance ratio ψ is based on either

$$t = \frac{(n-2)^{\frac{1}{2}}(\omega - \psi)}{[4\omega\psi(1 - r^2)]^{\frac{1}{2}}}$$

which has a Students' t-distribution with n-2 degrees of freedom (Cramér, 1945, page 400-401) or alternatively on

$$F = \frac{(n-2)(\omega - \psi)^2}{[4\omega\psi(1-r^2)]},$$

which has F-distribution with 1 and n-2 degrees of freedom. These are the Pitman-Morgan test statistics. The null hypothesis on ψ is rejected if the t (or F) value exceeds the required level of significance obtained from their respective tables.

3.4 Conclusion.

The Pitman-Morgan test is therefore a conditional test and the conditional approach is an appropriate alternative approach to deriving the test for variance ratios in a bivariate normal distribution. A comparison, particularly with Morgan's(1939) likelihood ratio approach, is interesting. The likelihood ratio, being minimal sufficient, summarises all the available information in the data about the parameter of interest. Since the resultant test statistics in both our cases are identical, we conclude that 'no information' is lost to the 'marginal' distribution of S through conditioning, in our approach. Nevertheless, the 'marginal' distribution of S depends on ψ.

We have used the word 'marginal' loosely in reference to

the density $h_1(S)$. More correctly, $h_1(S)$ is the conditional distribution of S given $T_{12} = t_{12}$.

Williams(1982) cited the Pitman-Morgan test as an example of inferences which cannot be derived based on a test for independence between sufficient statistics (for the accessory parameters) and ancillary statistics(for the accessory parameters). In our work T_{12} and S are sufficient for the accessory parameters ρ and ϕ respectively while the resultant statistic $(t - s/2)/(s^2/4 - \phi t_{12}^2)^{\frac{1}{2}}$ is ancillary for both ρ and ϕ by virtue of being independent of T_{12} and S. Thus the work in this chapter affirms Williams'(1982) conditional procedures in application and proves that contrary to his comments regarding the Pitman-Morgan test, this test can be derived as a test for independence between sufficient statistics and ancillary statistics.

The conditional inference theory in which we test for independence between a sufficient statistic for the accessory parameter and a statistic independent of the sufficient statistic (and therefore ancillary) has developed from Basu's(1955) theorem. Basu's(1955) theorem established the equivalence in general of an ancillary statistic for a parameter β, say, and a statistic independent of a sufficient statistic for β.

4. THE ANALYSIS OF CONCURRENT REGRESSIONS.

4.1 Introduction.

The theory of linear regression concerns the prediction of a random variable Y using information obtained by observing another independent (or concomitant) variable X, where X and Y are linearly related. Consider several sets of data observed on such (X,Y) under varying conditions. Sometimes it happens that the resultant lines have different slopes but are concurrent, i.e. all the straight lines pass through some common fixed point, (ξ,η) say. The point (ξ,η) is called the point of concurrence while we refer to the lines as representing concurrent regressions.

The analysis of concurrent regression lines first received attention in Tocher(1952) and Williams(1953). They developed test procedures and methods for constructing confidence limits for ξ and η, the abscissa and ordinate of concurrence respectively. That analysis received further clarification in Williams(1959, page 137-149). Following the analysis in Williams(1959), we show that their work is an application of conditional procedures.

The analysis in Williams(1959) deals with a special case in which there are equal subsample sizes (for each set); and for which the dependent variable Y is observed for the same values of the independent variable X in all the sets of data, but under some varied conditions. Sections 4.3 and 4.4 describe and discuss the analysis of concurrent regressions without these restrictions; this general case has previously received no attention.

We shall now introduce some notation relevant to our discussion. Assume there are m sets of data, each set with n_i observations. The observations shall be denoted with (X_{ij},Y_{ij}), $j=1,\ldots,n_i$; $i=1,\ldots,m$; where Y_{ij} shall be assumed to come from a normal population such that

$$E(Y_{ij}) = \eta + \beta_i(X_{ij}-\xi), \quad Var(Y_{ij}) = \sigma^2.$$

Thus Y_{ij} is the j^{th} value of Y in the i^{th} set, corresponding to X_{ij}, the j^{th} value of X in the i^{th} set. $\bar{Y}_{i.}$ and $\bar{X}_{i.}$ shall denote the mean values of the Y's and X's in the i^{th} group. We further define:

$$p_i = \sum_j y_{ij}(x_{ij} - \bar{x}_{i.})$$

$$t_i = \sum (x_{ij} - \bar{x}_{i.})^2$$

$$b_i = \frac{\sum_j y_{ij}(x_{ij} - \bar{x}_{i.})}{\sum_j (x_{ij} - \bar{x}_{i.})^2}$$

(b_i is the least squares estimate for β_i in the i^{th} set of values). In respect to the concurrent regressions, the corresponding definitions are

$$p_i^* = \sum y_{ij}(x_{ij} - \xi)$$

$$t_i^* = \sum (x_{ij} - \xi)^2.$$

The definition for b_i^*, the regression coefficient for the i^{th} concurrent line shall be presented later.

In the specific case of Williams(1959), $n_i=n$ and $x_{ij}=x_j$ for all i=1,...,m. Consequently, for this case we shall write p_i' in place of p_i^*, t in place of t_i, t´ in place of t_i^* and b_i' in place of b_i^*. For brevity we also put

$$J = n\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$K = \frac{n}{t}\sum_i \bar{y}_{i.}(p_i-\bar{p})$$

and

$$L = \frac{n}{t^2}\sum_i (p_i-\bar{p})^2.$$

4.2 Conditional Procedures In The Analysis of Williams(1959).

Given several sets of data on (X,Y) whose concurrence we wish to investigate, an appropriate procedure for analysis is:

(i) to test for difference in the slopes of the regression lines;

(ii) where the slopes do not differ significantly, the lines are assumed to be parallel. Then we test for difference between lines, i.e. whether the distance between the lines is zero at any fixed value of X;

(iii) if in (i) the slopes differ significantly, we test for the concurrence of the regression lines. We require a test for the departure from concurrence. This is essentially the analysis discussed in this chapter.

We identify two possible situations in (iii):

(a) when ξ, the abscissa of concurrence is known but η is unknown (we shall write $\xi=\xi_o$ in this case).

(b) when both ξ and η are unknown.

(a) Since ξ_o is known, η is the structural parameter while β_i is incidental (we use 'incidental' instead of 'accessory' to describe the β_i's because the β_i's increase in number with the sets of data). Thus the test for departure from concurrence is equivalent to a test for departure of the regression lines from η at $x_j=\xi_o$; and this test procedure needs to be independent of the incidental parameters, β_i, i=1,...,m.

It is easy to see that if the regression lines are concurrent, the point of concurrence (ξ_o,η) must lie on the mean regression line for all the sets of data. This line is

$$Y = \bar{y}_{..} + \bar{b}(x_j - \bar{x}_{.}).$$

Putting $x_j=\xi_o$, obtains for us an estimate of the ordinate of concurrence, η, as

$$y_c = \bar{y}_{..} + \bar{b}(\xi_o - \bar{x}_{.}).$$

But obviously y_c is the mean for the quantities

$$y_{ic} = \bar{y}_{i.} + b_i(\xi_o - \bar{x}_{.})$$

$$= \bar{y}_{i.} + \frac{p_i}{t} (\xi_o - \bar{x}_{.}), \quad i=1,\ldots,m;$$

and since ξ_o is known, it is easy to show that

$$E(y_{ic}) = \eta.$$

y_{ic} is evidently sufficient for η but ancillary for β_i. The departure of lines from η at $x_j=\xi_o$ is reflected in the variations of y_{ic}'s from y_c; and this test is free of the incidental parameters, β_i, $i=1,\ldots,m$. Since

$$Var(y_{ic}) = \sigma^2 \left(\frac{t+n(\xi_o-\bar{x})^2}{nt}\right)$$

$$= \frac{\sigma^2 t'_o}{nt},$$

where $t'_o = \sum(x_j-\xi_o)^2$, the sum of squares for departure from concurrence is

$$\frac{nt}{t'_o} \sum_i (y_{ic} - y_c)^2$$

$$= \frac{nt}{t'_o} \{ \sum(\bar{y}_{i.}-\bar{y}_{..})^2 + \frac{(\xi_o-\bar{x}_{.})^2}{t^2} \sum(p_i-\bar{p})^2 + 2\frac{(\xi_o-\bar{x}_{.})}{t} \sum\bar{y}_{i.}(p_i-\bar{p})\}$$

$$= \frac{t}{t'_o} \{ J + 2(\xi_o-\bar{x}_{.})K + (\xi_o-\bar{x}_{.})^2 L \} \qquad (4.1)$$

on m-1 degrees of freedom.

Only if this departure is not significant, may we carry out an analysis of the concurrent regressions. Then the sources for variation in the concurrent regressions will be the difference among the regression lines, the mean regression and the ordinate of concurrence.

Assuming concurrence, variations (or differences) in concurrent regressions are reflected in the differences between slopes. The parameters of interest will now be the β_i's (η is accessory). Let us define

$$b_i' = \frac{\sum_j (y_{ij} - y_c)(x_j - \xi_o)}{\sum (x_j - \xi_o)^2}$$

$$= \frac{\{ p_i' + n y_c (\xi_o - \bar{x}) \}}{t_o'}. \qquad (4.2)$$

Clearly

$$E(b_i') = \frac{1}{t_o'} \{ \eta \Sigma (x_j - \xi_o) + \beta_i \Sigma (x_j - \xi_o)^2 - \eta \Sigma (x_j - \xi_o) \}$$

$$= \beta_i.$$

b_i' is a sufficient estimate for β_i and consequently the test for difference in concurrent regressions is based on the sufficient statistics, b_i'. But from (4.2), b_i' and p_i' are statistically equivalent so that this test may very well be based on

$$p_i' = p_i + n \bar{y}_{i.} (\bar{x}_{.} - \xi_o).$$

Since

$$\mathrm{Var}(p_i') = \sigma^2 t_o',$$

the appropriate sum of squares for difference in concurrent regressions is

$$\frac{1}{t_o'} \sum_i (p_i' - \bar{p}')^2$$

$$= \frac{1}{t_o'} \{ \sum (p_i - \bar{p})^2 + 2n(\bar{x}_{\cdot} - \xi_o) \sum (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})(p_i - \bar{p}) + n^2 (\bar{x}_{\cdot} - \xi_o)^2 \sum (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \}$$

$$= \frac{1}{t_o'} \{ \frac{t^2 L}{n} + 2t(\bar{x}_{\cdot} - \xi_o) K + n(\bar{x}_{\cdot} - \xi_o)^2 J \}, \qquad (4.3)$$

on m-1 degrees of freedom.

The mean regression sum of squares (for concurrent regressions) will then be

$$\sum_i \frac{\bar{p}'^2}{t_o'} \qquad (4.4)$$

on 1 degree of freedom; and for the ordinate of concurrence,

$$\frac{y_c^2 mnt}{t_o'}$$

on 1 degree of freedom. This latter sum of squares is for departure of y_c from 0. However, in most of the problems we would require the sum of squares for departure of y_c from some hypothesized value η_o of η. This is

$$\frac{(y_c-\eta)^2 mnt}{t'_o}. \qquad (4.5)$$

The analysis may be summarised as follows:

ANOVA Table 4.1

Source	d.f.	Sum of Squares
Mean Regression	1	(4.4)
Ordinate of Concurrence	1	(4.5)
Difference in Concurrent Regressions	m-1	(4.3)
Departure from Concurrence	m-1	(4.1)
Total Variation due to Regression	2m	$\frac{1}{t'_o}(m\bar{p}'^2+mnt(y_c-\eta)^2) + J + \frac{tL}{n}$
Residual	m(n-2)	by subtraction
TOTAL	mn	$\sum\sum_{ij} y_{ij}^2$

The analysis in table 4.1 is evidently based on $\bar{y}_{i.}$ and p_i, statistically independent variables. These in turn are transformed to y_{ic} (ancillary statistic for β_i) and p_i' (sufficient statistic for β_i). Moreover,

$$\text{Cov}(y_{ic},p_i') = \text{Cov}(p_i',\bar{y}_{i.}) + (\xi_o-\bar{x})\text{Cov}(p_i',b_i)$$

$$= \sigma^2(\bar{x}_{.}-\xi_o) + (\xi_o-\bar{x}_{.})\sigma^2 = 0. \qquad (4.6)$$

(b) If both ξ and η are unknown, we test for an additional hypothesis, $H_o:\xi=\xi_o$. When H_o is not true, p_i' and y_{ic} will no longer be sufficient and ancillary respectively for β_i. In fact, if we put $\delta=\xi-\xi_o$,

$$E(p_i') = n\eta(\bar{x}_{.}-\xi) + \beta_i(t' + n\delta(\bar{x}_{.}-\xi))$$

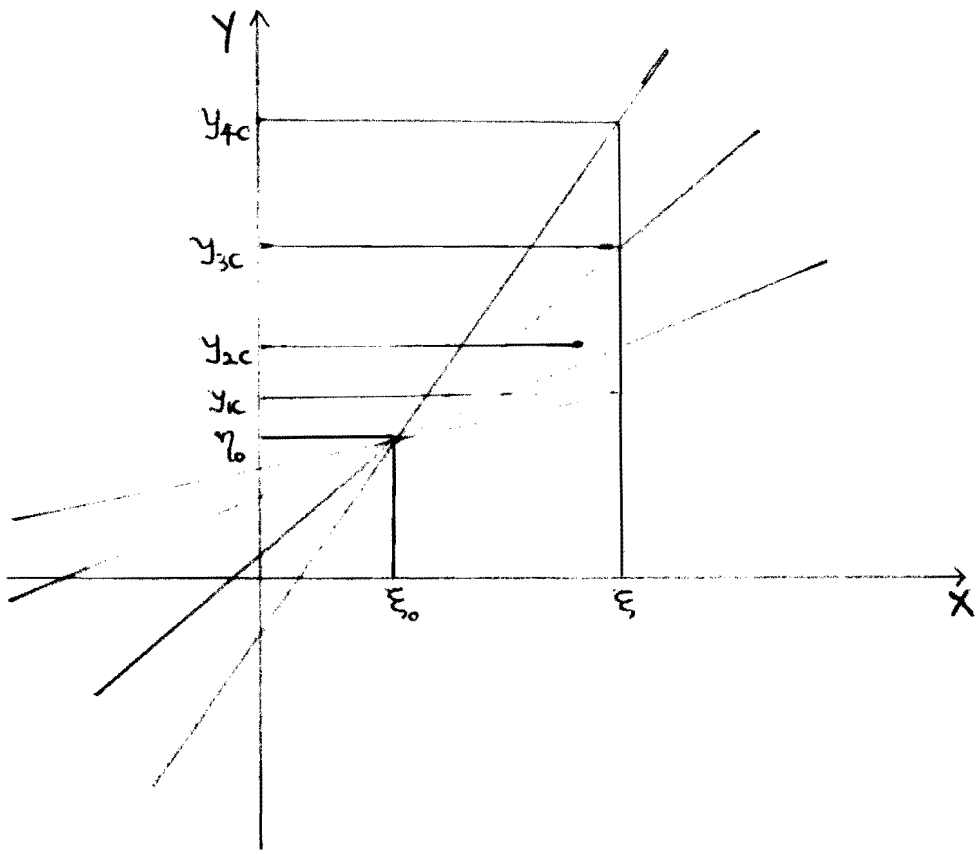
and

$$E(y_{ic}) = \eta + \beta_i\delta$$

so that y_{ic} varies with β_i. Under H_o, p_i' and y_{ic} will be sufficient and ancillary respectively, for β_i similar to what we had in case (a). By comparison with the discussion on linear functional relationships in Williams(1976), we know that the sample correlation between p_i' and y_{ic} will be centrally distributed under H_o and free of β_i. Otherwise $E(y_{ic})$ depends on β_i so that the sample correlation has a non-central distribution which depends on β_i.

Diagram 1 may clarify this. It shows how y_{ic} varies with β_i and consequently the sample correlation coefficient is non-centrally distributed when H_o is false.

Diagram 4.1:



The test for $H_o:\xi=\xi_o$ may therefore be designed as a test for the centrality (or non-centrality) of the distribution of the sample correlation between p_i' and y_{ic}. Equivalently, we may test for the significance of the regression of y_{ic} on p_i'. In this problem the β_i's are incidental parameters.

Under H_o, p_i' is sufficient for β_i. But when H_o is not true, conditioning on p_i' does not eliminate β_i. (In the same way y_{ic} will

not be ancillary for β_i). Thus the regression of y_{ic} on p_i' will be significant, i.e. the regression sum of squares will be non-centrally distributed since $E(y_{ic}|p_i')$ depends on β_i (Williams,1976; Rao,1973, page 264-265). This sum of squares may be appropriately called the sum of squares for the abscissa of concurrence, and is given by

$$\frac{nt\{ \sum y_{ic}(p_i'-\bar{p}') \}^2}{t' \sum (p_i'-\bar{p}')^2}$$

$$= \frac{nt\{ (\xi-\bar{x}_.)^2K - (\xi-\bar{x}_.)(\frac{t}{n}L - J) - \frac{t}{n}K \}^2}{t'\{ n(\xi-\bar{x}_.)^2J - 2(\xi-\bar{x}_.)tK + \frac{t^2}{n}L \}}. \qquad (4.7)$$

Clearly, the regression sum of squares in (4.7) will be 0 when ξ is replaced with its maximum likelihood estimator. Therefore an estimate of ξ is obtained by equating (4.7) to 0, i.e.

$$\sum y_{ic} (p_i'-\bar{p}') = 0. \qquad (4.8)$$

Letting x_c be the estimate for ξ from (4.8), we write $z=x_c-\bar{x}_.$. From (4.7), equation (4.8) simplifies to

$$z^2K - z(\frac{tL}{n} - J) - \frac{t}{n}K = 0$$

so that

$$x_c = \bar{x}_. + \frac{(\frac{tL}{n} - J) \pm \{(J + \frac{t}{n}L)^2 - 4\frac{t}{n}(JL - K^2)\}^{1/2}}{2K}. \qquad (4.9)$$

One root from (4.9) will maximise, and the other minimise the sum of squares for departure from concurrence (for unknown ξ_o); we require the latter. This sum of squares may be obtained as a difference from the ANOVA table 4.2, and is

$$\frac{\frac{t}{n}(JL - K^2)}{\frac{1}{t'}\sum (p_i' - \bar{p}')^2}. \qquad (4.10)$$

Since the numerator is free of ξ, we need x_c that maximises the denominator, i.e. the sum of squares for difference in concurrent regressions (as indicated in (4.3)).

The analysis is as follows:

ANOVA Table 4.2

Source	d.f.	Sum of Squares
Ordinate of Concurrence	1	$\frac{mnt}{t'}(y_c - \eta)^2$
Abscissa of Concurrence	1	(4.7)
Departure from Concurrence	m-1	(4.10)
TOTAL	m	$\frac{nt}{t'}\sum (y_{ic} - \eta)^2$

The test statistic for $H_o: \xi = \xi_o$ is (4.7)/(4.10).

From Williams(1973), we may conclude that the variance of x_c (the estimate from (4.9) and (4.10)) is asymptotically the reciprocal of the second derivative (with respect to ξ) of this test statistic at ξ_o. This is

$$\frac{t_o'^2 \,(JL - K^2)}{4(m-2)n\,\{n(\xi_o - \bar{x}_.)K - tL + nJ\}^2}$$

Derivations in this specific case require constant weights. When we allow the subsample sizes, n_i, to vary from set to set, we need varying weights as well. This is the general case discussed in the following section.

4.3 The General Case.

The analysis in section 4.2 placed restrictions on n_i and the independent variable, X. Over the various sets of data, we now allow for:

(i) variation in the subsample sizes, i.e. $j=1,\ldots,n_i$;

(ii) the values of the independent variable X to vary from set to set.

The observations are (X_{ij},Y_{ij}); $j=1,\ldots,n_i$; $i=1,\ldots,m$.

We do not discuss the case when ξ_o is known since the analysis in section 4.2 (a) applies to it. The only variation from (4.1) and (4.3) is that in this case the weights will vary with i. The weights are still chosen as inversely proportional to the respective variances.

Assume ξ_o is unknown. We redefine some of the terms.

Let

$$y^*_{ic} = \bar{y}_{i.} + b_i(\xi - \bar{x}_{i.}),\ i=1,\ldots,m. \qquad (4.11)$$

As before,

$$E(y^*_{ic}) = \eta + \beta_i\delta \qquad (4.12)$$

so that $E(y^*_{ic})=\eta$ and y^*_{ic} is ancillary for β_i, when $H_o:\xi=\xi_o$ is true. Furthermore,

$$\mathrm{Var}(y^*_{ic}) = \frac{\sigma^2}{n_i} + (\xi-\bar{x}_{i.})^2\frac{\sigma^2}{t_i}$$

$$= \frac{\sigma^2\, t^*_i}{n_i t_i}$$

(t_i and t^*_i are as defined in section 4.1). Since $\mathrm{Var}(y^*_{ic})$ varies with i, we define the weighted mean of the y^*_{ic}'s as

$$y^*_c = \sum w_i y^*_{ic},$$

$\sum w_i=1$. We choose w_i to be inversely proportional to $\mathrm{Var}(y^*_{ic})$ (Hoel, 1971, page 128-129 and page 195-196); i.e.

$$w_i = \frac{n_i t_i}{t^*_i} \Big/ \sum \frac{n_i t_i}{t^*_i}\ .$$

The regression coefficient for the i^{th} concurrent regression line is defined as

$$b_i^* = \frac{\sum_j (y_{ij}-y_c^*)(x_{ij}-\xi)}{\sum_j (x_{ij}-\xi)^2}$$

$$= \frac{1}{t_i^*} \{ p_i^* + n_i y_c^* (\xi - \bar{x}_{i.}) \}. \qquad (4.13)$$

Since

$$E(p_i^*) = \sum (x_{ij}-\xi)(\eta + \beta_i(\bar{x}_{i.}-\xi_o))$$

$$= \beta_i t_i^* + n_i \eta (\bar{x}_{i.}-\xi) + \beta_i n_i \delta (\bar{x}_{i.}-\xi), \qquad (4.14)$$

then

$$E(b_i^*) = \beta_i + \frac{\delta}{t_i^*} (n_i \beta_i (\bar{x}_{i.}-\xi) + n_i (\xi - \bar{x}_{i.}) \sum w_i \beta_i)$$

$$= \beta_i$$

and b_i^* is sufficient for β_i, if H_o:δ=0 is true.

The problem then is to design an appropriate test statistic for H_o:δ=0. As we show in the following section, the choice of suitable weights is the problem we face here.

4.4 The Analysis of The General Case.

The structural parameters in the problem are ξ and η, while the β_i's are incidental. If H_o is true, y^*_{ic} will be ancillary for β_i, while b^*_i is sufficient for β_i.

A comparison with table 4.2 enables us to write the sum of squares for the ordinate of concurrence as

$$\sum w_i(y^*_c-\eta)^2. \qquad (4.15)$$

It is clear from (4.13) that in general p^*_i will not be equivalent to b^*_i, being a different linear function of b^*_i in each group. Moreover, $Cov(b^*_i,y^*_{ic})\neq 0$. For,

$$Cov(b^*_i,y^*_{ic}) = \frac{1}{t^*_i}\{ Cov(p^*_i,y^*_{ic}) + n_i(\xi-\bar{x}_{i.})Cov(y^*_{ic},y^*_c) \}.$$

From (4.11),

$$Cov(p^*_i,y^*_{ic}) = Cov(p^*_i,\bar{y}_{i.}) + (\xi-\bar{x}_{i.})Cov(p^*_i,b_i)$$

$$= \sigma^2(\bar{x}_{i.}-\xi) + (\xi-\bar{x}_{i.})\sigma^2$$

$$= 0.$$

Also

$$Cov(y^*_{ic},y^*_c) = w_i\frac{\sigma^2 t^*_i}{n_i t_i}.$$

Therefore

$$\mathrm{Cov}(b_i^*, y_{ic}^*) = \frac{\sigma^2 w_i(\xi - \bar{x}_{i.})}{t_i}.$$

So it is misleading to design a test for $H_o:\xi=\xi_o$ on the assumption that ξ_o annihilates $\mathrm{Cov}(b_i^*, y_{ic}^*)$, although it seems reasonable that b_i^* and y_{ic}^* should be independent (or uncorrelated) at (ξ_o, η_o) and correlated elsewhere (diagram 4.1).

However, we know from (4.12) that $E(y_{ic}^*)$ depends on β_i so that the arguments in section 4.2 (b) apply here; i.e. although $\mathrm{Corr}(p_i^*, y_{ic}^*)=0$, the sample correlation between p_i^* and y_{ic}^* will be non-centrally distributed, and depends on β_i when H_o is not true. The test for the non-centrality of this distribution (i.e. the test for $H_o:\xi=\xi_o$) is modelled as a test for significance of the regression of y_{ic}^* on p_i^*. The regression sum of squares will be non-centrally distributed (and hence the regression is significant) if the incidental parameters, the β_i's, are not eliminated from the conditional model (conditioned on p_i^*). In turn, this shows that p_i^* is not adequately sufficient for β_i or $H_o:\xi=\xi_o$ is false (Williams, 1976).

The regression sum of squares, also called the sum of squares for abscissa of concurrence, is

$$\frac{\{\sum k_i y_{ic}^* (p_i^* - \bar{p}^*)\}^2}{\sum k_i (p_i^* - \bar{p}^*)^2} \qquad (4.16)$$

where

$$\bar{p}^* = \sum c_i p_i^*,$$

$$c_i = \frac{1}{t_i^* \sum(\frac{1}{t_i^*})}.$$

The choice of the weights, k_i, will be discussed later.

It follows, by difference, in table 4.3 below, that the sum of squares for departure from concurrence will be

$$\frac{[\sum w_i(y_{ic}^*-y_c^*)^2][\sum k_i(p_i^*-\bar{p}^*)^2] - [\sum k_i y_{ic}^*(p_i^*-\bar{p}^*)]^2}{\sum k_i(p_i^*-\bar{p}^*)^2}. \qquad (4.17)$$

The analysis is as follows:

ANOVA Table 4.3

Source	d.f.	Sum of Squares
Ordinate of Concurrence	1	(4.15)
Abscissa of Concurrence	1	(4.16)
Departure from Concurrence	m-2	(4.17)
TOTAL	m	$\sum w_i(y_{ic}^*-\eta)^2$

The sum of squares in (4.16) attains the value zero when ξ is the maximum likelihood estimator. Therefore we estimate ξ_o from

$$\sum k_i y^*_{ic}(p^*_i - \bar{p}^*) = 0. \qquad (4.18)$$

This simplifies to

$$\sum k_i(\bar{y}_{i.} - b_i\bar{x}_{i.} + \xi b_i)(p_i - \sum c_i p_i + n\bar{y}_{i.}\bar{x}_{i.} - \sum c_i n_i \bar{y}_{i.}\bar{x}_{i.} - \xi(n_i\bar{y}_{i.} - \sum c_i n_i \bar{y}_{i.})) = 0.$$

If we put

$$r_i = \bar{y}_{i.} - b_i\bar{x}_{i.}$$

$$s_i = p_i - \sum c_i p_i + n_i\bar{y}_{i.}\bar{x}_{i.} - \sum c_i n_i \bar{y}_{i.}\bar{x}_{i.}$$

$$t_i = n_i\bar{y}_{i.} - \sum c_i n_i \bar{y}_{i.},$$

(4.18) may be written as

$$\sum k_i(\xi^2 b_i t_i - \xi(r_i t_i + b_i s_i) + r_i s_i) = 0,$$

and the estimate is

$$x_c = \frac{\sum k_i(r_i t_i + b_i s_i) \pm \{[\sum k_i(r_i t_i + b_i s_i)]^2 - 4[\sum k_i b_i t_i][\sum k_i r_i s_i]\}^{\frac{1}{2}}}{2\sum k_i b_i t_i}.$$

There are two roots to equation (4.18). But at $\xi=\xi_o$, the sum of

squares for departure from concurrence (in (4.17)) will be minimum. Hence we take as estimate for ξ the x_c which minimises (4.17).

Attention shall now be given to the choice of k_i. It seemed appropriate to choose weights, k_i, to maximise the correlation between y^*_{ic} and $k_i p^*_i$. However, it became evident that such flexibility in choosing the weights permits k_i to be chosen making the correlation between y^*_{ic} and $k_i p^*_i$, unity.

A practical and reasonable approach is to work with the ratios p^*_i/t^*_i ($=q^*_i$, say). From (4.14) it is easy to show that

$$E(q^*_i) = \beta_i + \frac{n_i\eta(\bar{x}_{i.} - \xi) + \beta_i n_i \delta(\bar{x}_{i.} - \xi)}{t^*_i};$$

and so q^*_i will not be too greatly affected by the spread in the x_{ij} values, while simultaneously it reflects the variations in the slopes of the concurrent regression lines.

In place of the regression sum of squares in (4.16), we now have the regression of y^*_{ic} on q^*_i with w_i as the weights. This sum of squares is

$$\frac{\{\sum w_i(y^*_{ic} - y^*_c)q^*_i\}^2}{\sum w_i(q^*_i - \bar{q}^*)^2}$$

where

$$\bar{q}^* = \sum w_i q^*_i.$$

Consequently, the sum of squares for departure from concurrence (i.e. in place of (4.17)), shall be

$$\frac{\{\sum w_i(y^*_{ic}-y^*_c)^2\}\{\sum w_i(q^*_i-\bar{q}^*)^2\} - \{\sum w_i(y^*_{ic}-y^*_c)q^*_i\}^2}{\sum w_i(q_i-\bar{q}^*)^2}.$$

The remaining sums of squares in table 4.3 remain unchanged.

5. AN ASYMPTOTIC PROPERTY OF THE PARTIAL LIKELIHOOD.

5.1 Introduction.

Neyman and Scott(1948) pointed out that when the number of accessory parameters in the distribution of the population sampled increases to infinity with the sample size (though the structural parameters be finite), the method of maximum likelihood will yield inconsistent estimators. Nevertheless the likelihood theory remains so relevant and useful in statistical inference that it is tempting to consider modifications which allow an application to this problem. Much discussion of this question is available in the literature, notably that of Andersen(1970,1973) and Kalbfleisch and Sprott(1970) among others. Cox(1975), in a bid to reduce dimensionality, has introduced the 'partial likelihood.' It is desirable that the partial likelihood, which would then be the basis for our inference, be free of the accessory parameters.

Let Y, the observed random variable, be such that it can be transformed into a sequence

$$(X_1, S_1, X_2, S_2, \ldots, X_m, S_m).$$

We may then rewrite the full likelihood as

$$\prod_{j=1}^{m} f(x_j | x^{(j-1)}, s^{(j-1)}; \theta) . \prod_{j=1}^{m} f(s_j | x^{(j)}, s^{(j-1)}; \theta),$$

where

$$X^{(j)} = (X_1,\ldots,X_j),\ S^{(j)} = (S_1,\ldots,S_j),$$

and $\theta \in \Theta$, an open subset in R^k, is the vector of parameters. The product

$$L_\theta = \prod_1^{p\ m} f(s_j|x^{(j)},s^{(j-1)};\theta)$$

is called a 'partial likelihood.' As aforementioned, it is useful if this likelihood depends only on the structural parameters and not on the accessory parameters. Furthermore, marginal and conditional likelihoods are special cases of the partial likelihood. The partial likelihood will be identical with the marginal likelihood if the sequences of X's and S's are independent; and it is identical with the conditional likelihood if and only if S_j is independent of $(x_{j+1},x_{j+2},\ldots)$.

Cox(1975) further points out, without proof, that certain useful asymptotic properties of consistency and asymptotic normality should hold for the partial likelihood. The following work is a contribution toward formalising these results. It is an extension of Sweeting's(1980) results on the maximum likelihood estimator. It will also be shown that these results are applicable to the marginal and conditional likelihoods. In particular, work in this chapter aims to establish the consistency of maximum likelihood estimators from these likelihoods under very general conditions.

5.2 Definitions And Conditions.

We define P_θ^t, a probability measure on the measurable space

$(\mathbf{X}_t, \mathbf{A}_t)$, t being either discrete or continuous, such that P_θ^t is (absolutely) continuous with respect to λ_t, a σ-finite measure. It is assumed that the density

$$f_t(\theta) = \frac{dP_\theta^t}{d\lambda_t}$$

has second-order partial derivatives for all $\theta \in \Theta$.

Let the corresponding logarithm of the partial likelihood be

$$\ell_t^p = \sum_{j=1}^{t} \log f_t^p(s_j | x^{(j)}, s^{(j-1)}; \theta). \qquad (5.1)$$

The superscript 'p' will refer to terms derived from the partial likelihood wherever it appears.

Moreover, we shall write

$$U_t^p = \sum_{j=1}^{t} \frac{d}{d\theta} \log f_t^p(s_j | x^{(j)}, s^{(j-1)}; \theta). \qquad (5.2)$$

The symbols $\xrightarrow[u]{}$ and $\Longrightarrow_u$ shall respectively stand for uniform convergence and uniform weak convergence in compact subsets of Θ.

Definition 5.1:

A sequence $\{g_n(s)\}$ converges uniformly on a set E if and only if given $\varepsilon>0$, we can find n_o such that

$$|g_n(s) - g(s)| < \varepsilon$$

for all $n \geq n_o$ and for all $s \in E$.

Definition 5.2:

Let $P_{n,s}$ and P_s, $n \geq 1$, be probability measures on Borel subsets of a metric space, which depend on s, and C be the space of real bounded uniformly continuous functions. Then

$$P_{n,s} \overset{u}{\Longrightarrow} P_s$$

in s if and only if

$$\int u dP_{n,s} \longrightarrow \int u dP_s,$$

uniformly in s for all $u \in C$.

Let Γ be the matrix $(\theta_1,\ldots,\theta_k)$ with $\theta_i \in \Theta$, $i = 1,\ldots,k$. We define the norm of a matrix A denoted by $|A|$, as

$$|A| = (\mathrm{tr}A^TA)^{1/2}.$$

Define the information matrix

$$I_t^p(\theta) = -\ell_t^p{}''(\theta), \qquad (5.3)$$

i.e. minus the second derivative of (5.1); and assume the following conditions hold:

C1: Some nonrandom square matrices $A_t(\theta)$, continuous in θ, exist with

$$\{A_t(\theta)\}^{-1} \xrightarrow[u]{} 0$$

and such that

$$W_t(\theta) \equiv \{A_t(\theta)\}^{-1} \; I_t^p(\theta) \; [\{A_t(\theta)\}^{-1}]^T \xrightarrow[u]{} W(\theta),$$

where $P(W(\theta) > 0) = 1$.

C2: For all $c > 0$,

$$\text{i) Sup } |\{A_t(\theta)\}^{-1} A_t(\theta') - I_k| \xrightarrow[u]{} 0,$$

the sup being taken over the set

$$|\{A_t(\theta)\}^T (\theta'-\theta)| \leqslant c,$$

and I_k is the identity matrix.

$$\text{ii) Sup } |\{A_t(\theta)\}^{-1}[I_t^p(\Gamma) - I_t^p(\theta)][\{A_t(\theta)\}^{-1}]^T| \xrightarrow[u]{} 0$$

in probability, where the sup is taken over the set

$$|\{A_t(\theta)\}^T(\theta_i-\theta)| \leqslant c, \; 1 \leqslant i \leqslant k.$$

C3: The partial probability measure $P_\theta^{p,t}$, is (absolutely) continuous with respect to λ_t.

5.3 The Main Results.

In the proofs of the results, it is assumed that

second-order partial derivatives of f_t^p exist and are continuous a.s. for each $\theta \varepsilon \Theta$. In addition assumptions C1 - C3 will be assumed to hold.

Define

$$X_t^p(\theta) = \{A_t(\theta)\}^{-1} U_t^p(\theta). \tag{5.4}$$

Then

Theorem 5.1:

$$[X_t^p(\theta) , W_t(\theta)] \xrightarrow[u]{} [\{W(\theta)\}^{\frac{1}{2}} Z , W(\theta)]$$

where $Z \overset{d}{=} N(0, I_k)$, independent of $W(\theta)$.

Proof:

Wherever our reference is clear, the fixed argument θ is dropped. Thus we write W for $W(\theta)$.

Let $s \varepsilon \mathbb{R}$ and $\psi_t = \theta_t + \{A_t^{-1}\}^T s$, $\theta_t \longrightarrow \theta$ as $t \longrightarrow \infty$.

Since $A_t^{-1} \xrightarrow[u]{} 0$ by C1, some t_o exists such that wherever $t > t_o$, $\psi_t \longrightarrow \theta$ and $\psi_t \varepsilon \Theta$.

Using the Taylor expansion series about θ_t,

$$\ell_t^p(\psi_t) = \ell_t^p(\theta_t) + (\psi_t - \theta_t)^T \ell_t^{p\prime}(\theta_t) + \frac{1}{2}(\psi_t - \theta_t)^T \ell_t^{p\prime\prime}(\phi_t)(\psi_t - \theta_t) \tag{5.5}$$

where

$$\phi_t = \alpha_t \theta_t + (1-\alpha_t)\psi_t, \quad 0 < \alpha_t < 1.$$

Putting

$$V_t = A_t^{-1} \; I_t^p(\phi_t) \; \{A_t^{-1}\}^T \tag{5.6}$$

and taking exponentials in (5.5), we have

$$f_t^p(\psi_t) = f_t^p(\theta_t)\exp\{(\psi_t-\theta_t)^T \ell_t^{p\prime}(\theta_t) + \tfrac{1}{2}(\psi_t-\theta_t)^T \ell_t^{p\prime\prime}(\phi_t)(\psi_t-\theta_t)\}. \tag{5.7}$$

But from the definition of ψ_t above,

$$(\psi_t-\theta_t)^T \; \ell_t^{p\prime}(\theta_t) = s^T A_t^{-1} \; \ell_t^{p\prime}(\theta_t)$$

$$= s^T X_t^p,$$

by (5.1), (5.2) and (5.4); and similarly by (5.3) and (5.6),

$$\tfrac{1}{2}(\psi_t-\theta_t)^T \ell_t^{p\prime\prime}(\phi_t)(\psi_t-\theta_t) = -(\tfrac{1}{2})s^T A_t^{-1} \; I_t^p(\phi_t)\{A_t^{-1}\}^T s$$

$$= -(\tfrac{1}{2}) \; s^T V_t s.$$

Hence (5.7) may be written as

$$f_t^p(\psi_t) = f_t^p(\theta_t) \; \exp\{s^T X_t^p - (\tfrac{1}{2})s^T V_t s\}.$$

Therefore we have

$$\exp\{\tfrac{1}{2}s^T V_t s\} \; f_t^p(\psi_t) = f_t^p(\theta_t)\exp\{s^T X_t^p\}. \tag{5.8}$$

But we know

$$\text{i) } \{A_t(\theta')\}^{-1} \; I_t^p(\Gamma) \; [\{A_t(\theta')\}^{-1}]^T \xrightarrow[u]{} W(\theta)$$

by C1 and C2.

ii) $g_n(s) \xrightarrow[u]{} g(s)$ in s if and only if $g_n(s_n) \longrightarrow g(s)$ for every sequence $\{s_n\}$ such that $s_n \longrightarrow s$.

iii) Sweeting(1980) has proved that the distribution G_θ of $W(\theta)$ so constructed is continuous in θ. His lemma 3 is applicable here with our assumptions since no conditions unique to the full likelihood are used in its proof.

So $V_t \Longrightarrow W$ under either $\{\theta_t\}$ or $\{\psi_t\}$. Given $0<\varepsilon<1$, we may choose K such that

$$P(|W| \geqslant K) \leqslant \varepsilon$$

and

$$P(|W| = K) = 0.$$

Furthermore, since $V_t \Longrightarrow W$, and $\{|X| < K\}$ is a G_θ-continuity set,

$$P_{\theta_t}^{p,t}(|V_t| < K) \longrightarrow Pr(|W| < K). \qquad (5.9)$$

We define $\{Q^{p,t}\}$ as the distributions $\{P_{\theta_t}^{p,t}\}$ conditional on $\{|V_t| < K\}$ i.e. $Q^{p,t}$ has the density

$$q_t^p = \begin{cases} \dfrac{f_t^p(\theta_t)}{P_{\theta_t}^{p,t}(|V_t| < K)}, & |V_t| < K \\ 0 & \text{otherwise.} \end{cases}$$

Let U be a bounded function on the space of all k×k matrices, M_k, continuous on $|A| < K$ such that $U(A)=0$ for $|A| \geq K$. $E_t^{*,p}$ shall denote the expectation under $Q^{p,t}$.

Multiplying (5.8) above by $U(V_t)$ and integrating with respect to λ_t over $(|V_t| < K)$, which is permissible because of C3, we have by (5.8) and (5.9),

$$E_t^{*,p}(U(V_t)\exp\{s^T X_t^p\}) = \frac{E(U(V_t)\exp\{\tfrac{1}{2}s^T V_t s\})}{P_{\theta_t}^{p,t}(|V_t| < K)}$$

$$\longrightarrow \frac{E(U(W)\exp\{\tfrac{1}{2}s^T W s\})}{P(|W| < K)}$$

$$= E^*(U(W)\exp\{\tfrac{1}{2}s^T W s\}), \qquad (5.10)$$

where E^* is the expectation conditional on $|W| < K$. This holds since $U(W)\exp\{\frac{1}{2}s^T Ws\}$ is a bounded G_θ-continuous function.

But for any $Z \overset{d}{=} N(0, I_k)$ independent of W,

$$E^*(U(W)\exp\{s^T W^{1/2} Z\}) = E^*(E[U(W)\exp\{s^T W^{1/2} Z\}|W])$$

$$= E^{*}(U(W)E[\exp\{s^{T}W^{1/2}Z\}|W])$$

$$= E^{*}(U(W)\exp\{\tfrac{1}{2}s^{T}Ws\}); \qquad (5.11)$$

since when $Z \overset{d}{=} N(0,I_k)$, then

$$E[\exp\{\alpha^{T}Z\}] = \exp\{\tfrac{1}{2}\alpha^{T}\alpha\}.$$

Using the uniqueness of bilateral Laplace transforms and the weak compactness theorem, we have from (5.10) and (5.11)

$$(X_t^p \ , \ V_t) \Longrightarrow (W^{1/2}Z \ , \ W)I_{(|W| \ < \ K)}$$

with respect to the family $(Q^{p,t})$ of distributions. 'I' here is the indicator function.

But ε was arbitrary so that unconditionally

$$(X_t^p \ , \ V_t) \Longrightarrow (W^{1/2}Z \ , \ W)$$

and

$$(X_t^p \ , \ W_t) \Longrightarrow (W^{1/2}Z \ , \ W)$$

since $V_t - W_t \longrightarrow 0$, from the definitions of W_t, V_t and ϕ_t.

This will be true for all $\theta_t \longrightarrow \theta$. As well by lemma 3 in Sweeting(1980), the distribution of $(W^{1/2}Z \ , \ W)$ is continuous in θ. Therefore

$$(X_t^p \ , \ W_t) \underset{u}{\Longrightarrow} (W^{1/2}Z \ , \ W),$$

and the theorem is proved.

From this theorem, we may get an insight into the asymptotic joint distribution of the maximum likelihood estimator $\hat{\theta}_t^p$ from the partial likelihood, and $W_t(\theta)$. What is required is to relate $\hat{\theta}_t^p$ to $X_t^p(\theta)$.

We define

$$Y_t^p(\theta) = [A_t(\theta)]^T(\hat{\theta}_t^p - \theta).$$

Before stating the theorem, we define what we mean by uniform stochastic boundedness (u.s.b.) as used in Sweeting(1980).

Definition 5.3:

A family $(T_t(\theta))$ of $\mathcal{A}_t$-measurable functions is u.s.b. if given any $\varepsilon>0$ and a compact set K in Θ, there will exist some c and t_o such that

$$P_\theta[|T_t(\theta)| > c] < \varepsilon$$

for all $t>t_o$ and $\theta \varepsilon K$.

Sweeting(1980, Lemma 4) has shown that for a similarly defined $Y_t(\theta)$ from the full likelihood, a local maximum likelihood estimator exists so that $Y_t(\theta)$ is u.s.b. The proof uses no restrictions unique to the full likelihood. So we conclude that some local maximum likelihood estimator from the partial likelihood exists such that $Y_t^p(\theta)$ is u.s.b.

Theorem 5.2:

There exists a local maximum $\hat{\theta}_t^p$ of $\ell_t^p(\theta)$ with probability tending to 1 such that

$$X_t^p(\theta) - W_t(\theta)Y_t^p(\theta) \xrightarrow[u]{} 0$$

in probability.

Proof:

Again, we shall drop the fixed argument θ wherever our reference is clear. Define

$$G_t = (\hat{\theta}_t^p < \infty)$$

since the existence of $\hat{\theta}_t^p$ is already known. By the Taylor expansion series on G_t, we have from (5.2) and (5.3)

$$U_t^p(\theta) = I_t^p(\Gamma)\ (\hat{\theta}_t^p - \theta),$$

for some $\Gamma = (\theta_1,\ldots,\theta_k)$. We know

$$Y_t^p = A_t^T(\hat{\theta}_t^p - \theta)$$

so that

$$X_t^p = A_t^{-1}U_t^p$$

$$= A_t^{-1}\ I_t^p(\hat{\theta}_t^p - \theta)$$

$$= A_t^{-1} \; I_t^p \; [A_t^{-1}]^T \; A_t^T(\hat{\theta}_t^p - \theta)$$

$$= W_t(\theta, \hat{\theta}_t^p) Y_t^p$$

where

$$W_t(\theta, \hat{\theta}_t^p) = [A_t(\theta)]^{-1} \; I_t^p(\Gamma) \; [[A_t(\theta)]^{-1}]^T.$$

Moreover $[A_t(\theta)]^T(\theta_i - \theta)$, $1 \leqslant i \leqslant k$ are u.s.b. Therefore from condition C2(ii), we have

$$W_t(\theta, \hat{\theta}_t^p) - W_t(\theta) \xrightarrow[u]{} 0.$$

But Y_t^p is u.s.b. Therefore the theorem holds on the set G_t.

Now we need to show that $P_\theta^{p,t}(G_t) \xrightarrow[u]{} 1$ for the theorem to hold in general. But this follows from Sweeting's(1980, Lemma 4) result that $\hat{\theta}_t^p$ exists and is therefore finite.

Therefore the theorem is proved.

5.4 Application To Conditional And Marginal Likelihoods.

This section discusses an extension of the above results to marginal and conditional likelihoods (Kalbfleisch and Sprott, 1973; Andersen, 1970, 1973). A comparison with Andersen's(1970, 1973) asymptotic results is also presented.

Andersen(1970, 1973) discussed the problem of making inference from distribution models which depend on many accessory parameters. It is suggested that the accessory parameters $\tau_1, \ldots, \tau_n$ be

eliminated by conditioning on their respective minimal sufficient statistics, $t_1,\ldots,t_n$. Thus inference is based on the conditional likelihood

$$f(y_1,\ldots,y_n|t_1,\ldots,t_n;\theta),$$

where θ is the structural parameter.

Kalbfleisch and Sprott(1973) give the marginal likelihood based on the ancillary for the accessory parameter as $f(a;\theta)$, a being ancillary for the accessory parameter, τ.

The preceeding theory is applicable to the marginal and conditional likelihoods. In place of C3, we shall use the more specialised:

C4: The conditional (or marginal) probability measure $P_\theta^{c,t}$ (or $P_\theta^{m,t}$) is (absolutely) continuous with respect to λ_t.

Under the regularity assumptions and conditions C1, C2 and C4, the results and proofs follow identically. As indicated previously the conditional and marginal likelihoods are special cases of the partial likelihood.

The asymptotic results thus proved considerably strengthen those derived by Andersen(1970). We shall note the following significant differences in the imposed conditions:

(i) The maximum likelihood estimator from the conditional likelihood is called the "conditional maximum likelihood estimator" and Andersen requires that this should be unique (assumption 1.2). The

existence of a local maximum is sufficient for our discussion.

(ii) Assumption 1.3 in Andersen(1971) requires continuity in τ (the accessory parameter) of the mean and variance of the log conditional likelihood ratio for small variations from the true value of θ (denoted θ_o) to θ_o. Furthermore, this variance must be finite for all τ. In our proof, no conditions are placed on this log likelihood ratio.

(iii) While Andersen(1970) places regularity requirements on the third derivative of the log likelihood (assumption 1.4), no restrictions whatsoever are imposed on the third derivative in our case. In fact regularity assumptions on the first and second derivatives suffice in our proof.

(iv) Some of Andersen's(1970) continuity conditions in assumption 1.5 are comparable to our continuity requirements in C2. We require continuity of $A_t(\theta)$ and $I_t^c(\theta)$ in θ, where $I_t^c(\theta)$ is minus the second derivative of the conditional log likelihood. Note that θ may be a vector parameter with some accessory parameters. It is easy to see that since similar conditions to Sweeting's(1980) apply in the conditional likelihood situation, $A_t(\theta)$ will often be taken as $\{E[I_t^c(\theta)]\}^{1/2}$, when it exists. For the purpose of our proof it is not even necessary that $E[I_t^c(\theta)]$ exist. Therefore whenever $A_t(\theta)$ is chosen as $\{E[I_t^c(\theta)]\}^{1/2}$ our condition C2 will be more general than Andersen's(1970) assumption 1.5. In fact Andersen further assumes that $E[I_t^c(\theta)]$ is positive in the accessory parameter and the unconditional distribution is continuous in the accessory parameter as well.

(v) In the proof for asymptotic normality of the conditional maximum likelihood estimator, Andersen requires that the sequence of accessory parameters $\tau_1, \tau_2, \ldots$ should be bounded. We impose no such

requirement.

(vi) Andersen's(1970) proof is restricted to a discrete indexing parameter t while we allow for a continuous indexing parameter as well.

(vii) Although not formalised into an assumption, it is clear that Andersen's(1970) proofs assume independence among the random variables $Y_1, Y_2, \ldots$. Our results here permit dependence in $Y_1, Y_2, \ldots$.

Clearly then, conditions C1, C2 and C4 in the present proofs in addition to the regularity assumptions we make, relax Andersen's assumptions considerably so that the asymptotic results are proved in a much more general setting.

5.5 Examples.

In this section we give some illustrative examples of the partial likelihood. A further example to illustrate (vii) of section 5.4 is also presented.

Example 5.1 (Basawa and Prabhu,1981): An example in queuing theory.

In a one server queuing system partly observed till n customers have departed, assume the service times of the customers are independent and identically distributed, and independent of the interarrival times; furthermore, either the interarrival time distributions behave erratically or are unobservable.

Let F and G be distribution functions in a G/G/1 queue with probability density functions f and g, depending on parameters θ and ϕ respectively i.e. F is the distribution function for interarrival

times depending on θ while G is for the service times depending on ϕ. $\{u_k,\ k\geq 1\}$ and $\{v_k,\ k\geq 1\}$ are the observed interarrival and service times. D_n is the n^{th} departure epoch so that during $(0,D_n]$, we observe N_A interarrival times $(u_1,\ldots,u_{N_A})$.

The full likelihood for this model given by Basawa and Prabhu (1981) is

$$L_n(f,g) = \{\prod_1^{N_A} f(u_j;\theta_j)\}\ \{1 - F(x_n;\theta_1,\ldots,\theta_j)\}.\prod_1^n g(v_j;\phi)$$

where

$$X_n = X_n(D_n) = D_n - \sum_1^{N_A} u_j.$$

The term $\{1-F(x_n;\theta_1,\ldots,\theta_j)\}$ is the contribution of the incomplete arrival interval when sampling is terminated at D_n.

Then $\prod g(v_j;\phi)$ is a <u>partial likelihood</u> based on the set of service times, V, in the sequence $\{u_j,\ v_j\}$. Clearly this is free of the accessory parameters $\theta_1,\ldots,\theta_j$ and hence available for inference on the structural parameter ϕ. When the requirement for independence of the arrival times is removed, the partial likelihood is still available for inference.

Basawa and Prabhu(1981) further showed that the asymptotic properties applicable to the full likelihood also apply to $\prod g(v_j;\phi)$ (which is essentially the full likelihood for the service times). Of course our results hold for this likelihood.

Example 5.2:

Dawid(1975) raised questions about the inconsistencies in inference caused by defining sufficient and ancillary statistics in the presence of accessory parameters. For data Y let $f(Y;\omega)$ be the probability density function in question with $\omega=(\theta,\phi)$ where ϕ is an accessory parameter. Assume U and V exist such that

$$f(Y;\omega) = f(U;\theta)f(Y|U;\phi) \tag{5.12}$$

and

$$f(Y;\omega) = f(Y|V;\theta)f(V;\phi); \tag{5.13}$$

i.e. U and V are respectively S-sufficient and S-ancillary for θ.

Then which of $f(U;\theta)$ and $f(Y|V;\theta)$ should be the basis for our inference on θ? To resolve this problem, Dawid(1975) introduced the concept of 'likelinesses.' We assume that we can write

$$f(Y;\omega) = A(Y;\theta).B(Y;\phi).$$

From equations (5.12) and (5.13), this separation is already possible. Both $f(U;\theta)$ and $f(Y|V;\theta)$ are, as functions of θ, proportional to $A(Y;\theta)$. Any function proportional to $A(Y;\theta)$ in θ is called a 'likeliness' for θ.

If we assume the observable variable Y can be transformed to

$$(X_1,S_1,\ldots,X_m,S_m),$$

the corresponding partial likelihood from this transformation is

$$\prod_{1}^{m} f(S_j|X^{(j)},S^{(j-1)};\theta),$$

assuming it does not depend on ϕ at all; i.e. the full likelihood partitions as follows:

$$\prod f(Y_j;\omega) = \prod f(X_j|X^{(j-1)},S^{(j-1)};\phi) \prod f(S_j|X^{(j)},S^{(j-1)};\theta).$$

Whenever such partitioning is possible, the partial likelihood is a 'likeliness' for θ and as valid a basis for inference as the marginal and conditional likelihoods.

It was pointed out earlier that Andersen's results are restricted to an independent sequence of random variables $Y_1,Y_2,\ldots$. Examples in which $Y_1,Y_2,\ldots$ is a dependent sequence, exist such that our conditions widen the scope of application of these results. The following examples illustrate this point.

We consider conditional exponential families discussed by Heyde and Feigin(1975) in their discussion of Markov processes.

Let $f(X_i|X_{i-1};\theta)$ be the conditional probability density function of X_i given X_{i-1} in a time homogeneous Markov process and define

$$L_n(\theta) = \prod_{i=1}^{n} f(X_i|X_{i-1};\theta).$$

Then if $\hat{\theta}_n$ is the conditional maximum likelihood estimator from $L_n(\theta)$, and $I_n(\theta)$ is the conditional information given as

$$I_n(\theta) = \sum_{k=1}^{n} E\{ \left(\frac{d\ell og\ L_k(\theta)}{d\theta} - \frac{d\ell og\ L_{k-1}(\theta)}{d\theta}\right)^2 | \mathcal{F}_{k-1} \}$$

with $\mathcal{F}_k$ as the σ-field generated by $X_1,\ldots,X_k$, $k \geq 1$, then the conditional exponential family is characterised by the equation

$$\frac{d\ell og\ L_n(\theta)}{d\theta} = I_n(\theta)(\hat{\theta}_n - \theta),$$

for all $n \geq 1$.

In fact the property

$$\frac{d\ell og\ f(x_i|x_{i-1};\theta)}{d\theta} = \phi(\theta)H(x_{i-1})[m(x_i,x_{i-1}) - \theta]$$

for some ϕ, a function of θ alone, and H, a function of the X_i's alone, defines the conditional exponential families. $m(x_i,x_{i-1})$ is the root of $(d/d\theta)f(x_i|x_{i-1};\theta) = 0$.

If $I_n(\theta) \longrightarrow \infty$ as $n \longrightarrow \infty$, the logical choice for $A_t(\theta)$ in our conditions C1 and C2 is $\{I_n(\theta)\}^{\frac{1}{2}}$. In fact Heyde and Feigin show that

$$I_n(\theta) = \sum \phi(\theta)H(x_{i-1})$$

a.s. Therefore for this choice of $A_t(\theta)$, we require that $\sum H(x_{i-1})$ diverges a.s. This will certainly be so in a wide range of conditional exponential distributions.

Example 5.3:

One example of these families is the family of power series distributions for the offspring distribution in a Galton-Watson Branching process. If $Z_0,\ldots,Z_n$ are the successive generation sizes, a power series distribution is of the form

$$p_j = P(Z_1=j|Z_0=1)$$

$$= \frac{a_j\lambda^j}{f(\lambda)} \quad , \quad j=0,1,\ldots; \ \lambda>0$$

where $a_j \geq 0$ and $f(\lambda) = \sum a_j\lambda^j$.

It is easy to see that the offspring mean and variance from this distribution are respectively

$$\mu = \frac{\lambda f'(\lambda)}{f(\lambda)} \quad , \quad \sigma^2 = \{\frac{d\ell og\lambda}{d\mu}\}^{-1}.$$

Then

$$f(Z_i|Z_{i-1}) = \frac{\lambda^{Z_j}}{\{f(\lambda)\}^{Z_{j-1}}} \sum_{i_1+\ldots+i_{Z_{j-1}} = Z_j} a_{i_1}\ldots a_{i_{Z_{j-1}}} .$$

It is easily checked that the property

$$\frac{d}{d\mu} f(Z_i|Z_{i-1}) = \sigma^{-2}(Z_i-\mu Z_{i-1})f(Z_i|Z_{i-1})$$

characterises the power series family.

Example 5.4:

A second example of the conditional exponential families is from the estimation of parameter θ in a first-order autoregression

$$X_i = \theta X_{i-1} + \varepsilon_i,$$

where ε_i are independent and identically distributed normal random variables with mean zero and variance σ^2, and ε_i is independent of X_{i-1}. If $g(X)$ is the density function of ε_i, clearly

$$f(x_i|x_{i-1};\theta) = g(x_i-\theta x_{i-1}),$$

and hence

$$\frac{d\log f(x_i|x_{i-1};\theta)}{d\theta} = -x_{i-1}\frac{g'(x_i-\theta x_{i-1})}{g(x_i-\theta x_{i-1})}.$$

Heyde and Feigin(1975) simplify this equation to

$$\frac{d}{d\theta}\log f(x_i|x_{i-1};\theta) = c\, x_{i-1}^2\left(\frac{x_i}{x_{i-1}} - \theta\right) \qquad (5.18)$$

where

$$c = -E\left\{\frac{g(\varepsilon_1)g''(\varepsilon_1) - [g'(\varepsilon_1)]^2}{g^2(\varepsilon_1)}\right\}.$$

(5.18) characterises this family of distributions.

6. ON JAGERS' LEMMA ON THE MAXIMUM LIKELIHOOD ESTIMATORS FROM SUBSETS OF THE DATA.

6.1 Introduction.

When planning statistical inferences, it is important to be able to choose a sampling scheme which provides no more than the essential data for the inferences. Jagers(1975) proposed a lemma (2.13.2) under very general conditions to the effect that a maximum likelihood estimator which depends on only a subset of the data collected will be unchanged if only the relevant subset is collected. This lemma is incorrect as it stands. We shall show that Jagers' lemma is retrievable, subject to an additional sufficiency condition, but does not extend to general exponential families.

The theory of Branching Processes will be used to illustrate the lemma. We shall investigate the effect of basing inference about the reproduction mean on either the generation sizes or on the more detailed information of the number of offspring for different individuals in any generation. The conclusions drawn about the reproduction mean are identical for the exponential family in canonical form for an offspring distribution but not for general exponential families.

Jagers' lemma 2.13.2 (Jagers,1975):

"Let $(p_{\theta,\sigma},\ \theta\varepsilon\Theta,\ \sigma\varepsilon\Sigma)$ be a class of densities with respect to some σ-finite measure μ on $(\mathcal{X}, \mathcal{A})$. Let T be some statistic on this space (i.e. simply a function on $\mathcal{X}$). If there is a maximum

likelihood estimator $\hat{\theta}$ of θ based on observations in $\mathcal{X}$ and $\hat{\theta}=g_0T$ for some g, then g is a maximum likelihood estimator of θ based on observation of $T(X)$, $X\varepsilon\mathcal{X}$, the density of T being with respect to μT^{-1}."

Feigin(1977) put forth the following counter example to this lemma.

Example 6.1:

Let $X_1,\ldots,X_n$ be observations from the normal distribution, $N(\mu,\sigma^2)$. Then the sample variance

$$s^2 = \sum_{1}^{n} (x_i - \bar{x})^2/n$$

is the maximum likelihood estimator of σ^2 based on the full sample. However, if only s^2 is known, the maximum likelihood estimator of σ^2 is $(n/(n-1))s^2$ which is different. Jagers' lemma states that these should be the same.

In the following section it is shown that the lemma can be repaired under additional sufficiency conditions.

6.2 Modified Version of Jagers' Lemma.

There are two important points ignored in Jagers' lemma:

(a) It is assumed (but certainly not stated) that $\hat{\theta}$ will be free of any (accessory) parameter, σ, say (assuming σ is unknown). If $\hat{\theta}$ depends on σ, then σ too has to be estimated resulting in some information loss. This is comparable to the case discussed in Williams(1982).

(b) Jagers does not clarify whether the differential element $\mu T^{-1}(dy)$ is free of the (structural) parameter, θ. This requirement is necessary; thus we shall restrict T to being sufficient for θ, (or that the partial likelihood if used shall be free of the accessory parameter), i.e.

$$p_X(x;\theta,\sigma) = p_T(t;\theta,\sigma).p_{X|T}(x|t;\sigma)$$

and

$$p_X(x;\hat{\theta},\sigma) = p_T(t;\hat{\theta},\sigma).p_{X|T}(x|t;\sigma).$$

The conditional distribution given T should be free of θ. Then the maximum likelihood estimator, $\hat{\theta}$, is not influenced by this term.

Lemma (modified):

Let $\{P_{\theta,\sigma},\ \theta\varepsilon\Theta,\ \sigma\varepsilon\Sigma\}$ be a class of densities with respect to a σ-finite measure μ on $(\mathbf{X},\mathbf{A})$. Let T be a sufficient statistic for θ on this space and $\hat{\theta}$, the maximum likelihood estimator for θ based on the full sample in $\mathbf{X}$ and free of σ (or if it depends on σ, then we require that σ be known) such that $\hat{\theta}=g\circ T$ for some g. Then g is a maximum likelihood estimator for θ based on the observation of T(X), $X\varepsilon\mathbf{X}$.

Proof:

$P_{\theta,\sigma}$ is the probability measure on $\mathbf{X}$ with μ-density $p_{\theta,\sigma}$; $P_{\theta,\sigma}T^{-1}$ is the probability measure on $T(\mathbf{X})$ with μT^{-1}-density $p^T_{\theta,\sigma}$. Take any B in $T(\mathbf{X})$, the range space of T and let $A=T^{-1}(B)$.

Then

$$\int_B p^T_{\theta,\sigma}(y)\ \mu T^{-1}(dy) = \int_A p_{\theta,\sigma}(x)\ \mu(dx),$$

$$\leq \int_A p_{g\circ T(x),\sigma}(x)\ \mu(dx),$$

$$= \int_B p^T_{g(y),\sigma}(y)\ \mu T^{-1}(dy).$$

This last equality is now valid because the new conditions imposed ensure that the differential element is effectively free of θ. Besides, any loss of information due to estimation of σ in $\hat{\theta}$ is avoided. This completes the proof.

Note that if in the counter-example we had $s^2=\sum(x_i-\mu_o)^2/n$, μ_o known, then this lemma would be satisfied.

An objection may be raised that the requirement of the sufficiency of T is too strong. Of course the problem is further complicated by the presence of an accessory parameter. There are weaker definitions of sufficiency (in the presence of an accessory parameter) that have been suggested. However, since in most of those definitions the conditional distribution given T will not be free of θ, such 'sufficiency' will be unsuitable for this lemma. Alternatively we could have appealed to asymptotic sufficiency whereby asymptotically the conditional distribution given T is free of θ. s^2 in example 6.1 is asymptotically sufficient for σ^2 and thus serves to show that such relaxation of the requirement of sufficiency in the lemma will not do.

6.3 Application of the Modified Version of Jagers' Lemma.

In this section we illustrate the lemma with some important examples.

Example 6.2 (Application in Branching Processes):

The results of Harris(1948) and Feigin(1977) on the maximum likelihood estimator of the reproduction mean in Branching Processes serve to illustrate this lemma.

Let Z_j, $j=0,1,\ldots,n$ be the j^{th} generation size, and z_{jr}, $j=0,1,\ldots,n-1$; $r=0,1,\ldots$ be the number of members in the j^{th} generation with r offspring. Harris(1948), using z_{jr}, and Feigin-(1977), using Z_j, show that the (nonparametric) maximum likelihood estimator of the offspring mean μ is given by

$$\hat{\mu} = \frac{Y_n - 1}{Y_{n-1}}, \qquad Y_n = 1 + Z_1 + \ldots + Z_n,$$

regardless of the form of the offspring distribution.

In the case of the power series offspring distribution, $\sum Z_j$ is actually sufficient for μ and thus the result is easy to interpret to this parametric family. For then

$$P(Z_1 = j \mid Z_0 = 1) = \frac{a_j \lambda^j}{f(\lambda)}, \qquad j=0,1,\ldots$$

where $f(\lambda)=\sum a_j \lambda^j$. Clearly $\mu=\lambda f'(\lambda)/f(\lambda)$. Hence

$$P(Z_j \mid Z_{j-1}) = \sum_{\substack{i_1+\ldots+i_{Z_{j-1}} \\ = Z_j}} a_{i_1} \frac{\lambda^{i_1}}{f(\lambda)} \cdots a_{i_{Z_{j-1}}} \frac{\lambda^{i_{Z_{j-1}}}}{f(\lambda)}$$

$$= \frac{\lambda^{Z_j}}{\{f(\lambda)\}^{Z_{j-1}}} \sum a_{i_1} \cdots a_{i_{Z_{j-1}}}$$

and the likelihood based on $Z_0,\ldots,Z_n$ is

$$\prod_{j=1}^{n} P(Z_j|Z_{j-1}) = A(Z_0,\ldots,Z_n)\times\frac{\lambda^{\sum Z_j}}{(f(\lambda))^{\sum Z_{j-1}}} \qquad (6.1)$$

which is the factorisation on the basis of the sufficient statistic, $\sum Z_j$.

Similarly, the likelihood based on z_{jr} is

$$\frac{Z_0!}{\prod_{r=0}^{\infty} z_{0r}!}\prod_{r=0}^{\infty}\left(\frac{\lambda^r a_r}{f(\lambda)}\right)^{z_{0r}} \cdots \frac{Z_{n-1}!}{\prod_{r=0}^{\infty} z_{n-1,r}!}\prod_{r=0}^{\infty}\left(\frac{\lambda^r a_r}{f(\lambda)}\right)^{z_{n-1,r}}$$

$$= B(z_{jr};\ r=0,1,\ldots;\ j=0,1,\ldots,n-1)\times\frac{\lambda^{\sum Z_j}}{(f(\lambda))^{\sum Z_{j-1}}},$$

a similar factorisation to (6.1) with the same term that involves λ. The sufficient statistic is easily seen to determine that term; thus illustrating the modified version of Jagers' lemma.

For a general form of the offspring distribution, and in the parametric context, it is more difficult to illustrate the lemma with the results of Harris and Feigin. For then

$$P(Z_1=j|Z_0=1) = f(j;\lambda)$$

$$= \sum_{i=0}^{\infty}\frac{\lambda^i}{i!}f^{(i)}(j;0)$$

by Taylor's expansion; where $f^{(i)} = (d/d\lambda)^i f$.

On the basis of Z_j, the likelihood is

$$\prod_{j=1}^{n} P(Z_j|Z_{j-1}) = \prod_{j=1}^{n} \sum_{\substack{k_1+\ldots+k_{Z_{j-1}} \\ =Z_j}} \sum_{i=0}^{\infty} \prod_{\ell=1}^{Z_{j-1}} \frac{\lambda^i}{i!} f^{(i)}(k_\ell;0)$$

$$= \prod_{j=1}^{n} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} \sum_{\substack{k_1+\ldots+k_{Z_{j-1}} \\ =Z_j}} \prod_{\ell=1}^{Z_{j-1}} f^{(i)}(k_\ell;0),$$

$$= \prod_{j=1}^{n} \sum_{i=0}^{\infty} \lambda^i A(i,Z_0,Z_1,\ldots,Z_j).$$

If we base our inference on z_{jr}, the likelihood in general is

$$\prod_{j=0}^{n-1} \frac{Z_j!}{\prod_{r=0}^{\infty} z_{jr}!} \prod_{r=0}^{\infty} (f(r;\lambda))^{z_{jr}}$$

$$= \prod_{j=0}^{n-1} \frac{Z_j!}{\prod_{r=0}^{\infty} z_{jr}!} \prod_{r=0}^{\infty} \left\{ \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} f^{(i)}(r;0) \right\}^{z_{jr}}.$$

Specific knowledge of either the form of distribution $f(j;\lambda)$ or at least the family from which it is derived is necessary for us to be more conclusive about the result since otherwise we do not know the sufficient statistic.

Jagers' lemma was proved on the assumption that the underlying probability model is parametric. Feigin(1977) and Harris'(1948) results were proved for nonparametric offspring distributions showing that in that case, the maximum likelihood estimators for μ from either $Z_0,\ldots,Z_n$ or from the more detailed information of z_{jr}, $j=0,1,\ldots,n-1$; $r=0,1,2,\ldots$, are identical. Although Feigin(1977) and Harris(1948) place no restriction on the distribution model used, it will be evident from this chapter that such generality is not automatically applicable to the parametric distribution models.

Let p_r be the probability that an individual produces r offspring, where p_r is assumed nonparametric. The joint probability function of the z_{jr}, $j=0,1,\ldots,n$; $r=0,1,2,\ldots$ is

$$\left\{\prod_{j=0}^{n}\frac{Z_j!}{\prod_{r=0}^{\infty}z_{jr}!}\right\}\prod_{r=0}^{\infty}p_r^{\sum_j z_{jr}}. \qquad (6.2)$$

If $(Z_0,\ldots,Z_n)$ were sufficient for p_r, (6.2) would be factorable into a product of a term free of p_r, and another term which depends on the z_{jr}'s only through $(Z_0,\ldots,Z_n)$ and which involves p_r. Obviously (6.2) does not permit this factorisation in general except for the special case when $p_r=p$ for all r. So $(Z_0,\ldots,Z_n)$ is not sufficient for p_r in general. Consequently, our version of Jagers' lemma cannot be used to explain Feigin's(1977) result for nonparametric families.

Example 6.3:

In this example, we look at the application of these results to the exponential family. The exponential family has distributions of the form

$$p_Z(z;\theta) = a(\theta)b(z)\exp(\theta T(z)). \qquad (6.3)$$

The exponent may be more complicated, say $d(\theta)T(z)$. The simpler form however, will do for our discussion.

Keiding(1975) has shown that Jagers' lemma will hold for the case when

$$Z = \sum_k c_k T_k(Z);$$

i.e.

$$p_Z(z;\theta) = a(\theta)\, b(z) e^{\theta z},$$

for then the maximum likelihood estimator for μ is

$$\hat{\mu} = \frac{Z_1 + \ldots + Z_n}{Z_0 + \ldots + Z_{n-1}}$$

as for the results of Harris(1948) and Feigin(1977). Z_j is here, as before, the j^{th} generation size and it is evident that $\sum Z_j$ is sufficient in the model (6.3). This form of p_Z is called the canonical form of the exponential family.

Using (6.3) and since $\sum p_Z(z;\theta) = 1$, it is easy to show that in general

$$E(T(Z)) = -a'(\theta)/a(\theta), \tag{6.4}$$

and

$$Var(T(Z)) = \left(\frac{a'(\theta)}{a(\theta)}\right)^2 - \frac{a''(\theta)}{a(\theta)}. \tag{6.5}$$

We now proceed to discuss an example given in Keiding(1975) for which Jagers' lemma does not hold. Although the example is from the exponential family, it is obviously not of the canonical form. Therefore it serves to show that Jagers' lemma cannot be extended from exponential families in canonical form to apply to all exponential family forms in general.

Example 6.4:

Let the offspring distribution be

$$p_Y(y;\theta) = (y+1)^{-\theta}/\xi(\theta), \tag{6.6}$$

$y=0,1,\ldots$; $1<\theta<\infty$; where

$$\xi(\theta) = \sum_{m=1}^{\infty} m^{-\theta}$$

is the Riemann Zeta function. We may rewrite this distribution as

$$p_Y(y;\theta) = [\xi(\theta)]^{-1} e^{-\theta \log(y+1)}; \qquad (6.7)$$

i.e. $T(Y) = \log(Y+1)$.

Let Y_{ij}, $i=0,1,\ldots,n-1$; $j=1,\ldots,Z_i$ be the number of offspring of the j^{th} member in the i^{th} generation. Clearly, the $(i+1)^{th}$ generation size is

$$Z_{i+1} = \sum_{j=1}^{Z_i} Y_{ij}.$$

It follows that the likelihood function from this distribution is

$$L_n(\theta) = [a(\theta)]^{Z_0+\ldots+Z_{n-1}} \exp\{-\theta \sum_{i=0}^{n-1} \sum_{j=1}^{Z_i} \log(y_{ij}+1)\}, \qquad (6.8)$$

with $a(\theta) = [\xi(\theta)]^{-1}$. Applying the factorisation theorem for sufficient statistics to (6.8), it is evident that $\sum Z_i$ is not sufficient for θ; and consequently, not sufficient for $\mu(\theta)$. Therefore our version of Jagers' lemma is not applicable for this example.

The form of $\mu(\theta)$ for this example can be derived from (6.6) as

$$\mu(\theta) = E(Y_{ij}) = a(\theta) \sum_{m=0}^{\infty} m(m+1)^{-\theta}$$

$$= a(\theta)\{\xi(\theta-1) - \xi(\theta)\}$$

$$= \frac{a(\theta)}{a(\theta-1)} - 1.$$

It follows that since $E(Z_1|Z_0=1)=\mu(\theta)$, then

$$E(Z_{i+1}) = E_{Z_i}(E(Z_{i+1}|Z_i))$$

$$= E(Z_i)\mu(\theta)$$

$$= [\mu(\theta)]^{i+1}.$$

Therefore, since $\mu(\theta)>1$

$$E(\sum_{i=1}^{n} Z_i) = \frac{\mu^{n+1}(\theta)-1}{\mu(\theta)-1}$$

$$\sim \frac{\mu^{n+1}(\theta)}{\mu(\theta)-1}$$

as $n \longrightarrow \infty$.

From (6.4) and (6.7) we can deduce that

$$\nu(\theta) = E[\log(Y_{ij}+1)] = \frac{a'(\theta)}{a(\theta)}$$

$$= a(\theta) \sum_{m=1}^{\infty} m^{-\theta} \log m,$$

and from (6.5) and (6.7),

$$c^2(\theta) = \mathrm{Var}(\log(Y_{ij}+1))$$

$$= a(\theta) \sum_{m=1}^{\infty} m^{-\theta} (\log m)^2 - (a(\theta) \sum_{m=1}^{\infty} m^{-\theta} \log m)^2.$$

We establish the following proposition.

Proposition 6.1:

In the distribution model (6.6),

$$(Z_0+\ldots+Z_{n-1})^{\frac{1}{2}} \left(\frac{\sum\sum \log(Y_{ij}+1)}{Z_0+\ldots+Z_{n-1}} - \nu(\theta)\right) \xrightarrow{d} N(0, c^2(\theta)),$$

conditional on non-extinction.

Proof:

$(\log(Y_{ij}+1))$, $i=0,1,\ldots,n-1$; $j=1,\ldots,Z_i$; is a sequence of independent and identically distributed random variables. Thus we can apply the law of large numbers for random sums to get, conditional on non-extinction,

$$\frac{\sum\sum \log(Y_{ij}+1)}{Z_0+\ldots+Z_{n-1}} \xrightarrow{a.s.} \nu(\theta),$$

since $Z_0+\ldots+Z_{n-1}$ is the number of summations of $\log(Y_{ij}+1)$ in the numerator and $(\mu(\theta))^{-n}(Z_0+\ldots+Z_{n-1})$ converges a.s. to a proper limit law which is a.s. positive.

By the central limit theorem for random sums it follows that, conditional on non-extinction,

$$\frac{(Z_0+\ldots+Z_{n-1})^{\frac{1}{2}}}{v(\theta)}\left(\frac{\sum\sum \log(Y_{ij}+1)}{Z_0+\ldots+Z_{n-1}} - \nu(\theta)\right) \xrightarrow{d} N(0,1).$$

This proves the proposition.

We shall now discuss the derivation of the maximum likelihood estimator, $\hat{\mu}(\theta)$. From (6.8), the maximum likelihood estimator for θ will be the solution to the equation

$$\frac{d\log L_n(\theta)}{d\theta} = (Z_0+\ldots+Z_{n-1})\frac{a'(\theta)}{a(\theta)} - \sum_i \sum_j \log(Y_{ij}+1) = 0,$$

i.e. $\hat{\theta}$ is a solution to

$$\hat{\nu}(\theta) = \frac{a'(\theta)}{a(\theta)} = \frac{\sum\sum \log(Y_{ij}+1)}{Z_0+\ldots+Z_{n-1}}. \qquad (6.9)$$

Therefore in practice, the θ value satisfying equation (6.9) also obtains for us $\hat{\mu}(\theta)$. From the general likelihood (6.8), the right hand side in (6.9) is a function of a sufficient statistic for θ, namely $(Z_0+\ldots+Z_{n-1}, \sum\sum \log(Y_{ij}+1))$. In fact, equation (6.9) gives the maximum likelihood estimator for $\nu(\theta) = E(\log(Y_{ij}+1))$.

As well from (6.8), we know that $\sum\sum \log(Y_{ij}+1)$ is (minimal) sufficient for θ (in the exponential distribution). Lehmann and Scheffé(1950, theorem 5.1) extended the Rao-Blackwell theorem (Rao,1945; Blackwell,1947) proving that V is a minimum variance estimator of its expected value (MVUE) if and only if it is a function of

the (minimal) sufficient statistic. Since $\sum\sum \log(Y_{ij}+1)$ is a function of $(Z_0+\ldots+Z_{n-1}, \sum\sum \log(Y_{ij}+1))$, a sufficient statistic (and it is in fact minimal sufficient) for θ, the above mentioned results suggest that asymptotically, it should derive for us a minimum variance unbiased estimator for $\nu(\theta)$.

We may extend this application of the Lehmann-Scheffé theorem to permit minimum variance unbiased estimation for $\mu(\theta)$. For let V be any unbiased estimate for $\mu(\theta)$; and we write $S=\sum\sum \log(Y_{ij}+1)$. Define $T=E(V|S)$. Then it follows (Cox and Hinkley, 1974, page 258-259) that T is a minimum variance unbiased estimator for $\mu(\theta)$.

A similar result to that in proposition 6.1, and following the same line of proof, is

$$\sigma^{-1}(Z_0+\ldots+Z_{n-1})^{\frac{1}{2}}\left(\frac{Z_1+\ldots+Z_n}{Z_0+\ldots+Z_{n-1}} - \mu(\theta)\right) \xrightarrow{d} N(0,1),$$

since $E(Z_1|Z_0=1)=\mu(\theta)$ and $Var(Z_1|Z_0=1)=\sigma^2(\theta)$. The Lehmann and Scheffé(1950, theorem 5.1) result suggests that $c^2(\theta) \leq \sigma^2(\theta)$. In fact we can establish that this inequality holds and is strict

Proposition 6.2:

In the offspring distribution model (6.6),

$$c^2(\theta) = Var(\log(Y_{ij}+1)) < \sigma^2(\theta) = Var(Z_1|Z_0=1).$$

Proof:

First of all, we find the form of $\sigma^2(\theta)$. We omit the limits in the summations wherever they are clear. We have

$$E(Z_1^2 | Z_0 = 1) = E(Y_{1j}^2) = a(\theta) \sum_{m=0}^{\infty} m^2 (m+1)^{-\theta}$$

$$= a(\theta) \sum_{m=0}^{\infty} (m^2 - 1)(m+1)^{-\theta} + a(\theta) \sum_{m=0}^{\infty} (m+1)^{-\theta}$$

$$= a(\theta)(\zeta(\theta-2) - 2\zeta(\theta-1) + \zeta(\theta)).$$

Therefore,

$$\sigma^2(\theta) = \frac{\zeta(\theta-2)}{\zeta(\theta)} - \left(\frac{\zeta(\theta-1)}{\zeta(\theta)}\right)^2$$

$$= \frac{(\sum m^{-\theta})\sum m^2 m^{-\theta} - (\sum m.m^{-\theta})^2}{(\sum m^{-\theta})^2},$$

and we compare this with

$$c^2(\theta) = \frac{(\sum m^{-\theta})\sum m^{-\theta}(\log m)^2 - (\sum m^{-\theta}\log m)^2}{(\sum m^{-\theta})^2}.$$

The numerator in $\sigma^2(\theta) \cdot c^2(\theta)$ can be written as

$$(\sum m^{-\theta}).\sum m^{-\theta}[m^2 \cdot (\log m)^2] - ([\sum m.m^{-\theta}]^2 - [\sum m^{-\theta}\log m]^2)$$

$$- (\sum m^{-\theta}).(\sum m^{-\theta}(m-\log m)(m+\log m)) - (\sum m^{-\theta}(m-\log m))(\sum m^{-\theta}(m+\log m)),$$

(6.10)

We put

$$T_M = M-\log M$$

$$S_M = M+\log M.$$

Clearly, T_M and S_M are both increasing functions of M. We define the probability function of M as

$$P(M=m) = km^{-\theta}, \quad m=1,2,\ldots,$$

where $k^{-1}=\sum m^{-\theta}$. Note that $E(T_M)$ and $E(S_M)$ exist. If we multiply (6.10) by k^2, it is easy to see that the numerator for $k^2(\sigma^2(\theta)-c^2(\theta))$ becomes

$$E(T_M S_M) - E(T_M)E(S_M) = Cov(T_M, S_M)$$

$$= Var(M) - Var(\log M) > 0,$$

This proves the proposition.

It seems reasonable therefore, that inference about $\mu(\theta)$ in the model (6.7) be carried out on the basis of the minimal sufficient statistic, $\sum\sum\log(Y_{ij}+1)$. Such values of θ as are appropriate in the estimation of $\nu(\theta)$, will be equally appropriate in the estimation of $\mu(\theta)$.

It should be noted that, since $c^2(\theta)<\sigma^2(\theta)$, the advantage in obtaining the information contained in the $\log(Y_{ij}+1)$, $0<i\leq n$, $0<j\leq Z_i$, rather than that in the Z_i, $0\leq i\leq n$, is quite clear. This appears, at face value, to contradict the nonparametric results of Harris(1948) and Feigin(1977) which lead one to expect that there should be no advantage in sampling any more than the Z_i, $0\leq i\leq n$. In fact, there is a subtle difference between the parametric and nonparametric maximum likelihood approaches and they are not equivalent. Many authors have been in error on this point.

BIBLIOGRAPHY.

1. Andersen, E.B. (1970): Asymptotic properties of conditional maximum likelihood estimators. J.R. Statist. Soc. B,32, 283-301.
2. Andersen, E.B. (1971): Correction: Asymptotic properties of conditional maximum likelihood estimators. J.R. Statist. Soc. B,33, 167.
3. Andersen, E.B. (1973): Conditional Inference and Models for Measuring. Mentalhygiejnisk Forlag, Copenhagen.
4. Barndorff-Nielsen, O. (1973): On M-ancillarity. Biometrika 60, 447-455.
5. Barndorff-Nielsen, O. (1978): Information and Exponential Families in Statistical Theory. J. Wiley and Sons, New York.
6. Basawa, I.V. and Prabhu, N.U. (1981): Estimation in single server queues. Nav. Res. Logistics Quarterly 28, 475-487.
7. Basu, D. (1955): On statistics independent of a complete sufficient statistic. Sankhya 15, 377-380.
8. Basu, D. (1958): On statistics independent of sufficient statistics. Sankhya 18, 223-226.
9. Basu, D. (1964): Recovery of ancillary information. Sankhya A,26, 3-16.
10. Blackwell, D. (1947): Conditional expectation and unbiased estimation. Ann. Math. Statist. 18, 105-110.
11. Cox, D.R. (1958): Some problems connected with statistical inference. Ann. Math. Statist. 29, 357-372.
12. Cox, D.R. (1975): Partial likelihood. Biometrika 62, 269-276.

13. Cox, D.R. and Hinkley, D.V. (1974): Theoretical Statistics. Chapman and Hall, London.

14. Cramér, H. (1974): Mathematical Methods of Statistics. Princeton University Press, Princeton.

15. Dawid, A.P. (1975): On the concepts of sufficiency and ancillarity in the presence of nuisance parameters. J.R. Statist. Soc. B,37, 248-255.

16. Durbin, J. (1961): Some methods of constructing exact tests. Biometrika 48, 41-55.

17. Durbin, J. (1980): Approximations for densities of sufficient estimators. Biometrika 67, 311-333.

18. Feigin, P.D. (1977): A note on maximum likelihood estimation for branching processes. Austr. J. Statist. 19, 152-154.

19. Finney, D.J. (1938): The distribution of the ratio of estimates of the two variances in a sample from a normal bivariate population. Biometrika 30, 190-192.

20. Fisher, R.A. (1925): Theory of statistical estimation. Proc. Camb. Phil. Soc. 22,700-725.

21. Fisher, R.A. (1934): Two new properties of maximum likelihood. Proc. Roy. Soc. A,144, 285-307.

22. Fisher, R.A. (1935): The logic of inductive inference. J.R. Statist. Soc. 98, 39-54.

23. Fraser, D.A.S. (1956): Sufficient statistics with nuisance parameters. Ann. Math. Statist. 27, 838-842.

24. Godambe, V.P. (1980): On sufficiency and ancillarity in the presence of a nuisance parameter. Biometrika 67, 155-162.

25. Gordon, I.R. (1981): On Conditioning in Statistical Inference. (Unpublished) M.Sc. Thesis, Dept of Statist., La Trobe University.

26. Harris, T.E. (1948): Branching processes. Ann. Math. Statist. 19, 474-494.

27. Heyde, C.C. and Feigin, P.D. (1975): On efficiency and exponential families in stochastic process estimation. Statistical Distributions in Scientific Work, ed. G.P. Patil, 227-240.

28. Hoel, P.G. (1971): Introduction to Mathematical Statistics. J. Wiley and Sons, New York.

29. Jagers, P. (1975): Branching Processes with Biological Applications. J. Wiley and Sons, London.

30. Johansen, S. (1976): Two notes on conditioning and ancillarity. Preprint 12, Instit. of Math. Statist., University of Copenhagen.

31. Johansen, S. (1977): Some remarks on M-ancillarity. Scand. J. Statist. 4, 181-182.

32. Kalbfleisch, J.D. and Sprott, D.A. (1970): Application of likelihood methods to models involving large numbers of parameters. J.R. Statist.Soc. B,32, 175-192.

33. Kalbfleisch, J.D. and Sprott, D.A. (1973): Marginal and conditional likelihoods. Sankhya 35, 311-328.

34. Keiding, N. (1975): Estimation theory for branching processes. Contributed papers, 40^{th} Int. Statist. Instit. Session, Warszawa, 438-445.

35. Lehmann, E.L. (1959): Testing Statistical Hypotheses. J. Wiley and Sons, New York.

36. Lehmann, E.L. (1981): An interpretation of completeness and Basu's theorem. J. Amer. Statist. Assoc. 76, 335-340.

37. Lehmann, E.L. and Scheffé, H. (1950): Completeness, similar regions, and unbiased estimation. Sankhya 10, 305-340.

38. Liang, K.Y. (1983): On information and ancillarity in the presence of a nuisance parameter. Biometrika 70, 607-612.

39. Madow, W.G. (1945): Note on the distribution of the serial correlation coefficient. Ann. Math. Statist. 16, 308-310.

40. Morgan, W.A. (1939): A test for the significance of the difference between the two variances in a sample from a normal bivariate population. Biometrika 31, 14-19.

41. Neyman, J. and Pearson, E.S. (1936): Sufficient statistics and uniformly most powerful tests of statistical hypotheses. Statist, Res. Mem. 1, 113-137.

42. Neyman, J. and Scott, E.L. (1948): Consistent estimates based on partially consistent observations. Econometrica 16, 1-32.

43. Pitman, E.J.G. (1939): A note on normal correlation. Biometrika 31, 9-12.

44. Pitman, E.J.G. (1979): Some Basic Theory for Statistical Inference. Chapman and Hall, London.

45. Rao, C.R. (1945): Information and the accuracy attainable in the estimation of statistical parameters. Bull. Cal. Math. Soc. 37, 81-91.

46. Rao, C.R. (1973): Linear Statistical Inference and its Applications. J. Wiley and Sons, New York.

47. Sprott, D.A. (1975): Marginal and conditional sufficiency. Biometrika 62, 599-605.

48. Sweeting, T.J. (1980): Uniform asymptotic normality of the maximum likelihood estimator. Ann. Statist. 8, 1375-1381.

49. Tocher, K.D. (1952): On the concurrence of a set of regression lines. Biometrika 39, 109-117.

50. Williams, E.J. (1953): Tests of significance for concurrent regression lines. Biometrika 40, 297-305.

51. Williams, E.J. (1959): Regression Analysis. J. Wiley and Sons, New York.

52. Williams, E.J. (1973): Test of correlation in multivariate analysis. Bull. Int. Statist. Instit. 45, 219-231.

53. Williams, E.J. (1976): The power of some tests of correlation. Perspectives in Probability and Statistics, ed. J. Gani, distributed by Academic Press, London, for the Applied Probability Trust, Sheffield, 105-116.

54. Williams, E.J. (1982): Some classes of conditional inference procedures. Adv. Appl. Prob. 19A, 293-303.