MATEMATISK INSTITUTT Høsten 1983

THE METHOD OF MULTIPLE COMPARISON BY A DELTA-METHOD AND ITS RELATIONSHIP TO THE LIKELIHOOD RATIO TEST. GENERAL THEORY AND APPLICATION TO MULTINOMIAL MODELS

by

Erling Sverdrup

CONTENT

Summary

By the so-called delta-method a test statistic is used which is a function of statistics with known variance-covariance structure. The standard deviation of the statistic is found by linearizing it. Significance is declared if the ratio between the test statistic and its estimated standard deviation surpasses c , where c is the $1-\varepsilon$ fractile of the normal distribution, and ε is asymptotically, the level of significance. Generalizing this method let $\eta = (\eta_1, \ldots, \eta_v)$ be the parameter in the model and H_0 a hypothesis that reduces the number of freely varying parameters to t. "Effects" are functions $f(\eta)$ of η. They are "contrasts" relatively to H_0 if $f(\eta) = 0$ for $\eta \in H_0$. Multiple comparison consists in looking for contrasts which are "present", i.e. for which $f(\eta) > 0$. According to "The delta multiple comparison method" any contrast may be declared present if significance is obtained by using the delta method with the critical point c replaced by $\sqrt{z}$, where z is the $(1-\varepsilon)$-fractile of the chi-square distribution with $w = v-t$ degrees of freedom. It is shown below that then the level still holds (asymptotically). It is also shown that this multiple comparison method is related to the likelihood ratio test for H_0 in a similar manner that the Scheffé method establishes a connection between modified Student testing and Fisher's analysis variance test. However, as discussed in chapter II below, our attitude to the null hypothesis is that we are not interested in the truth or the falsehood of it. The hypothesis is demoted to a tool which on the one hand side is used to impose limitations on the possible comparisons to be undertaken in the statistical analysis. On the other hand the hypothesis defines the degrees of freedom, i.e. the critical point in the comparisons. Thus the hypothesis provides the tie between the comparisons which are desired and the actual shape of the decision criterions, [see Sverdrup (1975), (1977a) and (1977b)].

The general result is given in Theorem 4 in chapter II below.

The general theory is applied to multinomial situations where, of course, likelihood ratio testing may be replaced by chi-square goodness of fit testing.

I. INTRODUCTION. THE BASIC STATISTICAL IDEAS

1. The general idea of contrasts and multiple comparison

The purpose of many statistical analyses is to find important "effects" or "contrasts" concerning the unknown parameters $\eta = (\eta_1, \eta_2, \ldots)$ in the model. Thus we are interested in effects defined as functions $f(\eta)$ of η, and we consider a class $\mathcal{F}$ of such functions f which a priori are assumed to be feasible and interesting. An effect f is present if $f(\eta) > 0$ for the true η. (If $f(\eta) < 0$ then $-f$ is present.) We specify $\mathcal{F}$ by introducing a set H_0 of η. Then for any $f \in \mathcal{F}$, $f(\eta) = 0$ for all $\eta \in H_0$. Hence f is a contrast relatively to H_0. "Multiple comparison methods" aim at discovering effects which are present.

As an illustration, suppose that $\eta_1, \eta_2, \ldots,$ are population means in different classes. We may be interested in comparing the different η_i; i.e. we are interested in effects $f = \eta_i - \eta_j$. This leads to H_0: $\eta_1 = \eta_2 = \eta_3 = \ldots$. If the subscripts "i" of the η_i represent different equidistant points of time, then we may be interested in the curvature, i.e. the escalating effects $\eta_{i+1} - 2\eta_i + \eta_{i-1}$, which leads to H_0: $\eta_i = \alpha + \beta i$ for all i, where α and β are unspecified. In a two-way lay-out with class means η_{ij}, we may for any $i \neq i'$ be interested in knowing if the effect of j on η_{ij} is greater than on $\eta_{i'j}$. Hence H_0: $\eta_{ij} = \alpha_i + \beta_j$ for all i and j, because then $\eta_{ij} - \eta_{ij'} - \eta_{i'j} + \eta_{i'j'}$ would be 0 under H_0.

In general H_0 is what is usually called the null-hypothesis. However, it will appear from the examples, that we have chosen H_0, not because we have any a priori confidence in it or are interested in the truth or falsehood of it, but because we are interested in certain effects which are contrasts relatively to H_0. We may even know in advance that H_0 cannot be true. Hence the term null-state is more appropriate than null-hypothesis.

We shall in this paper give priority to constructing methods for which the probability of stating at least one false contrast is at most ε, paying less attention to the important problem of studying the probability of discovering important contrasts. Thus we, require asymptotically

$$\Pr\left[\bigcup_{f:f(\eta)<0} (\text{stating } f(\eta)>0)\,\middle|\,\eta\right] < \varepsilon \tag{1}$$

(In the present Chapter I, we do not attempt to make the exposition rigorous. This will be done in the subsequent chapters.) Of course, many different multiple comparison methods can be constructed which fulfill this requirement. We shall confine ourselves to consider those methods which are related to the classical tests of H_0 , viz. the likelihood ratio tests or the chi-square goodness of fit tests, in the same manner as the now classical method of Henry Scheffé (1953) is related to the Fisher F-test. We do not claim these methods to be superior to other methods.

It is perhaps correct to state that the classical Karl Pearson test, as we know it today, which to the old generation of statisticians was the very embodiment of statistical testing, has almost never been a two-decision problem. Significance has always meant scrutinizing the data. The same has been true of analysis of variance testing. The progress that was made by Scheffé was to define the last part of the procedure in rigorous terms. The mathematical statisticians of former generations aimed at shedding some light on the randomness involved in handling statistical data, they did not aim at constructing statistical decision functions for the complete statistical treatment.

2. Review of Scheffé's method

To bring out some feature of multiple comparison tests we shall review Scheffé's method. It concerns the linear normal model, like those used in variance analysis or regression analysis; i.e. the observations $(X_1,\ldots,X_n)' = X$ have independent and normally distributed components with unknown variance σ^2 and $\xi = EX = y\beta$ where y is a known $(n\times s)$ matrix of rank s $(< n)$.

We are interested in linear contrasts for $\eta = (\beta_1,\ldots,\beta_r)$; $r < s$. According to Scheffé's multiple comparison rule it is stated that $f'\eta = \Sigma_1^r f_j\beta_j > 0$ if $f'\hat{\eta} > \sqrt{rc}\,\hat{\sigma}_f$, where c is the $(1-\varepsilon)$-fraktile of the Fisher distribution with r and $n-s$ degrees of freedom, and $\hat{\sigma}_f^2 = f'g^{-1}f\hat{\sigma}^2$. $\hat{\beta}_j$ and $\hat{\sigma}^2$ are the usual unbiased estimates of β_j and σ^2, and g is such that $g^{-1}\sigma^2$ is the covariance matrix of $\hat{\eta} = (\hat{\beta}_1,\ldots,\hat{\beta}_r)'$. It is well known that $f'\hat{\eta} > \sqrt{rc}\,\hat{\sigma}_f$ for some $f = (f_1,\ldots,f_r)'$ if and only if $F = \hat{\eta}'g\hat{\eta}/r\hat{\sigma}^2 > c$. It follows immediately from Schwartz inequality in the following form

$$\max_v v'w/\sqrt{v'av} = \sqrt{w'a^{-1}w} \tag{2}$$

where v and w are vectors and a is a symmetric positive definite matrix. We recognize this as the usual Fisher's F-test for testing the null-hypothesis H_0: $\beta_1 =\ldots= \beta_r = 0$. Hence the probability of making a false statement if $\beta_1 =\ldots= \beta_r = 0$, is precisely ε. This result is of little interest in itself, but from this result we easily make the deduction that for any arbitrary η (and β and σ), the probability of making a false statement is at most ε, i.e.

$$\Pr\left[\bigcup_{f:\Sigma_1^r\beta_i f_i<0}\left(\sum_{i=1}^r f_i\hat{\beta}_i>\sqrt{rc}\,\hat{\sigma}_f\right)\right] < \varepsilon$$

regardless of whether the null-hypothesis is true or not.

[Proof: We introduce $\hat{\gamma}_i = \hat{\beta}_i-\beta_i$; $i = 1,2,\ldots,s$; and may write the probability,

$$P = \Pr\left[\bigcup_{f:\Sigma\beta_i f_i<0}\left(\sum_1^r f_i\hat{\gamma}_i+\sum_1^r f_i\beta_i>\sqrt{rc}\,\hat{\sigma}_f\right)\right]$$

However, by leaving out the term $\Sigma f_i\beta_i$, which is < 0, we obtain an expression which is at least as large

$$P < \Pr\left[\bigcup_{f:\Sigma_1^r\beta_i f_i<0}\left(\sum_{i=1}^r f_i\hat{\gamma}_i>\sqrt{rc}\,\hat{\sigma}_f\right)\right]$$

By letting the union go over all f we have

$$P < \Pr\left[\bigcup_f\left(\sum_{i=1}^r f_i\hat{\gamma}_i>\sqrt{rc}\,\hat{\sigma}_f\right)\right]$$

However, the $\hat{\gamma}_i = \hat{\beta}_i - \beta_i$ are least square estimates of the $\gamma_i = \beta_i - \beta_i = 0$ relatively to the "observations" $\tilde{X}_i = X_i - \Sigma_{j=1}^{r} y_{ij}\beta_j$; $i = 1,\ldots,n$; where $E(\tilde{X}_1,\ldots,\tilde{X}_n) = y(0,\ldots,0,\beta_{r+1},\ldots,\beta_s)'$. Hence by what we have stated above, the right hand side of the last relation is equal to ε, and hence $P \leq \varepsilon$.]

Several important points can now be made concerning multiple comparison methods.

(i) Each comparison in a multiple comparison procedure is a modified Student test for the special hypothesis H_f: $f'\eta = 0$, against $f'\eta > 0$, because in that case the rejection takes place if $f'\hat{\eta} > t\hat{\sigma}_f$, where t is the $(1-\varepsilon)$-fractile of the Student distribution with $n-s$ degrees of freedom. Thus we just have to replace the t-fractile in the Student test by $\sqrt{rc}$, where c is the critical point in testing of the nullstate, i.e. the $(1-\varepsilon)$--fractile of the Fisher distribution with r and $n-s$ degrees of freedom.

(ii) The $f'\eta$ are the contrasts relatively to $\eta = 0$. This null-state (null-hypothesis) is used to generate possible effects (contrasts) which we are interested in. The decision space consists of intersections of these effects only, the null-hypothesis or its negation is not subject to decision making, and the level ε has a meaning without refering to the null-hypothesis. It was stated above that the fact that the probability of making a false statement under the null-hypothesis is ε, is of little interest in itself. This is obvious if we are sure in advance that the null-hypothesis can not be true. However, even if the null-hypothesis may be true, it is uninteresting since it is the error of stating that $f'\eta > 0$ for any β for which $f'\eta \leq 0$ we should have safe-guards against. This we have by what we have just proved.

(iii) Nothing can, of course, prevent us from performing the test in the following manner. Ascertain first if $F > c$. If it is not, then drop the whole statistical analysis, since no interesting effects are present. If it is true, then we may look around for effects. Numerical convenience may justify such a procedure, which amounts to "testing" of the "hypothesis". Thus we test without having an hypothesis. Perhaps, therefore, clearance testing would be a more appropriate term. We perform the testing because the decision space includes the possibility of not stating anything, due to the scarcity of the information given by the data. The purpose of the testing is to see if this possibility could be excluded. Significance clears the way for finding contrasts.

(The discussion above also counters the objection sometimes made that after the significance testing has been performed the testing should be conditional, given that $F > c$. The point is that testing is not needed, but may be numerically convenient. You do not want to waste time in looking for effects when no effects are possible.)

(iv). Somebody may perhaps find it peculiar that the construction of the test requires the derivation of the sampling distribution in the null-state, that is under an assumption that cannot be true. Perhaps that is the psychological reason why one has felt compelled to attach credence to the hypothesis. However, the mathematical rational of finding the null-state-distribution should be clear from the derivation above.

On the other hand those who have found it contradictary to test hypotheses which are known to be false and would have been rejected anyhow if the material was large enough, should feel comforted in their predicament.

3. Outline of the general delta multiple comparison method

We shall below develop a method of multiple comparison in the case of parametric models with variables that are not necessarily normal The observations will be assumed to be independent and groupwise identically distributed, i.e. the joint density of the observations X_{ai} ; $i = 1,2,\ldots,n_a$; $a = 1,2,\ldots,s$; is

$$L(X;\eta) = \prod_{a=1}^{s} \prod_{m=1}^{n_a} g_a(X_{am};\eta) \tag{3}$$

where $\eta = (\eta_1,\ldots,\eta_v)$ is an unknown parameter, and g_a is a probability density with respect to a measure μ.

We shall be interested in effects $f(\eta)$ as defined above. We define the class $\mathscr{F}$ of effects by means of a null-state which is such that η may be written

$$H_0: \eta_i = \phi_i(\theta_1,\ldots,\theta_t);\ i = 1,2,\ldots,v>t \tag{4}$$

or briefly $\eta = \phi(\theta)$, where θ varies freely in the t-space. By what we have stated above $\mathscr{F}$ will consist of all or some $f(\cdot)$ for which $f(\eta) = 0$ for all $\eta \in H_0$, i.e. $f(\phi(\theta)) = 0$ for all θ. An effect fulfilling this condition is said to be present if $f(\eta) > 0$ for the true η. A multiple comparison method aims at finding f-s in $\mathcal{F}$ which are present. The f-s are not assumed to be linear, but they must be "smooth" in a sense to be defined later.

Just as Scheffé's method consists in repeated use of modified Student testing, our general multiple comparison method will consist in repeated use of a modified version of the time-honoured delta-method, which in the old days was the handy jackknife to be used in almost every possible practical statistical situation where judgements of uncertainties were deemed necessary.

In our present context the method is constructed roughly in the following manner. Let η^* be the maximum likelihood estimator of η. We consider the estimate $f(\eta^*)$ of $f(\eta)$ and find it natural to state that $f(\eta) > 0$ if $f(\eta^*) > c_f$ where $c_f > 0$. To determine c_f we need the variance of $f(\eta^*)$. This is obtained by linearizing $f(\eta^*)$,

$$f(\eta^*) = f(\eta)+\sum_{j=1}^{v}(\eta_j^*-\eta_j)f_j(\eta) \tag{5}$$

This is justified if all n_a are large, since η^* is a consistent estimate of η. Now, the asymptotic covariance matrix of η^* is $\frac{1}{n}(\lambda^{(n)}(\eta))^{-1}$ where $\lambda^{(n)}(\eta)$ is the famous information matrix

$$\lambda^{(n)}(\eta) = -\left(\sum_{a=1}^{s}\frac{n_a}{n}E\frac{\partial^2 \log g_a(X_{am};\eta)}{\partial\eta_i\,\partial\eta_j}\right)_{i,j=1,\ldots,v} \tag{6}$$

Hence, by (5) and (6), the asymptotic variance of $f(\eta^*)$ is

$$\frac{1}{n}\sigma_f^2(\eta) = \frac{1}{n}\sum_{i,j=1}^{v} f_i(\eta)f_j(\eta)((\lambda^{(n)}(\eta))^{-1})_{ij} \tag{7}$$

Thus the delta method for testing H_f: $f(\eta) = 0$ against $f(\eta) > 0$ would consist in rejecting H_f if $f(\eta^*) > k\,\sigma_f(\eta^*)/\sqrt{n}$, where k is the $(1-\varepsilon)$-fractile of the normal distribution. Now, if this method is to be used repeatedly for different $f \in \mathcal{F}$, then we replace k by $\sqrt{z}$ where z is the $(1-\varepsilon)$-fractile of the chi-square distribution with $v-t$ degrees of freedom. Thus we state that $f(\eta) > 0$ for any $f \in \mathcal{F}$ for which

$$f(\eta^*) > \sqrt{z}\ \sigma_f(\eta^*)/\sqrt{n} \tag{8}$$

where $\sigma_f(\eta)$ is given by (6) and (7). Then we shall prove that asymptotically the probability of stating at least one false effect is at most ε, hence (1) holds in the limit. (See theorem 4 (iii) below)

4. The relationship to likelihood ratio testing

Consider now the usual likelihood ratio for the hypothesis H_0 given by (4)

$$Q = \frac{L(X;\hat{\eta})}{L(X;\eta^*)} \tag{9}$$

where $\hat{\eta} = \phi(\hat{\theta})$ and $\hat{\theta}$ are the maximum likelihood estimates if H_0 is true. If H_0 is rejected when

$$-2\log Q > z \tag{10}$$

then it is wellknown that the testing has level ε, asymptotically. We shall prove below that (10) is asymptotically in probability a necessary condition for the existence of a significant effect f, i.e. (9) holds for at least one $f \in \mathcal{F}$. (Theorem 4, (iv)). Hence we may use (10) as a clearance test. If $-2\log Q < z$, then in the limit the probability is 0 of discovering significant effects. In special cases the significance of the likelihood ratio (10) will be proved to be asymptotically necessary and sufficient condition for having a significant effect.

5. A comment on simultaneous confidence intervals

Closely related to (8) is the construction of simultaneous confidence interval for the different functions f, viz.

$$\Pr\{\bigcap_{f\in\mathcal{F}}[|f(\eta)-f(\eta^*)|<\sqrt{z}\ \sigma_f/\sqrt{n}]\} > 1-\varepsilon \tag{11}$$

asymptotically for large n. It appears that simultaneous confidence intervals are seldom used in practical statistical work. However, they are often used to describe a test procedure for multiple comparisons. As such it is a misnomer, compared to the original meaning of confidence intervals which are meant to be terminal decisions. Thus in the case of a one-way lay-out in analysis of variance with 5 classes, 6 observations in each class, $\varepsilon = 0.05$, observed class means 375, 470, 367, 296, 363 and estimated variance $\hat{\sigma}^2 = 1662.5$, we have as a simultaneous confidence interval for all $\sum_1^5 f_i\xi_i$, where $\xi_1,\ldots,\xi_5$ are the population means, and $\sum f_i = 0$,

$$|\sum_1^5 f_i\xi_i - 375f_1 - 470f_2 - 367f_3 - 296f_4 - 363f_5| < 55.3\sqrt{\sum_1^5 f_i^2} \qquad (12)$$

As a <u>terminal</u> decision (12) would usually say very little. However (12) describes a <u>method</u> of <u>testing</u> interesting contrasts. Thus we may read out directly from (12) that (e.g.) the second mean is significantly greater than the other means, that the fourth mean is significantly less than the first and that there are no other significant differences between means. Trivially, however, this could be seen from (8). (On the other hand, Hotelling-Working's confidence bands for regression lines are cases in point for simultaneous confidence intervals (Sverdrup (1976).)

6. <u>The multinomial model</u>

Potentially there are many possible applications of the general theory just outlined. We shall apply it to the case of s multinomial trial sequences. In the a-th sequence; $a = 1,2,\ldots,s$; there are n_a trials, each of which must result in one of r_a mutually exclusive events

$$A_{a1}, \ldots, A_{ar_a} \qquad (13)$$

with probabilities

$$\pi_{a1},\ldots,\pi_{ar_a} ; \sum_j \pi_{aj} = 1 \qquad (14)$$

All the $n = \Sigma n_a$ trials are independent. To see that we have a special case of (3) we introduce $X_{ai} = (Y_{ai1},\ldots,Y_{air_a})$, where $Y_{aij} = 1$ and all other $Y_{aij'}$, $j' \neq j$, are 0 if A_{aj} occurs in the i-th trial in a-th multinomial sequence. Then we have for g_a in (3),

$$g_a(X_{am};\eta) = \prod_{j=1}^{r_a} \pi_{aj}(\eta)^{Y_{amj}} \tag{15}$$

where the

$$\pi_{aj} = \pi_{aj}(\eta);\ \eta = (\eta_1,\ldots,\eta_v) \tag{16}$$

are specified functions of unspecified parameters η_j. In general (16) may impose a priori restrictions on the multinomial model. However, the classical case when the π_{aj} vary in the interval $(0,1)$ only subject to (14) is included as a special case. It will be refered to as the framework model (also called the "saturated" model).

Introducing (15) into (3) we get,

$$L(X;\eta) = \prod_{a=1}^{s} \prod_{j=1}^{r_a} \pi_{aj}^{N_{aj}} \tag{17}$$

where

$$N_{aj} = \sum_{i=1}^{n_a} Y_{aij}\ ;\quad j = 1,2,\ldots,r_a;\ a = 1,2,\ldots,s; \tag{18}$$

is the number of times A_{aj} occurs in trial sequence a.

The results concerning the delta multiple comparison method sketched above can now be applied directly to the multinomial model with a priori restrictions defined by (15) and (16). For the purpose of applications, the effects $f(\eta)$ will also be expressed as $F(\pi)$; i.e. in terms of π instead of η. The results obtained are essentially the same as those obtained by Goldstein (1981) by a different approach.

Of course, in the multinomial situations the likelihood ratios could be replaced by the chi-square goodness of fit statistics. In the of the framework model these situations have been treated before by the present author [(1975) and (1977b)]. However, the derivation in these papers was different from the one we shall now present. It was based on the special properties of the multinomial models and the chi-square goodness of fit tests. It led to the stronger result of a purely algebraic relationship between the appropriate goodness of fit statistic and the multiple comparison rule. (Of course, the level ε still holds asymptotically in probability.) This treatment is repeated in Section III, 4 below in a somewhat modified form.

7. Restrictive multinomial models

Traditionally, only framework multinomial models have been used for categorical data. The possibility of using restrictive models and testing by means of differences between goodness of fit statistics was pointed out by Neyman (1949). Of course, restrictive models account for much higher efficiency of the tests, and that is true also in the case of multiple comparisons. Hence the methods developed in this paper contribute to more efficient readings of tables of categorical data. Anybody who has tried to analyse categorical population data by means of classical homogeneity tests, independence tests, etc., will sooner or later feel disappointed. The tests do not react to effects which seem intuitively obvious. The reason for this is that "intuition", subconsiously and correctly, operates with smooth functions, i.e. restrictive models. Thus the mortality rate is well known to be a smooth function of age, a fact that is disregarded e.g. in homogeneity testing. (See e.g. Cramer (1945), p.449.) (The use of restrictive models also raises the important problem of robustness in multinomial trials. This problem has been treated by Goldstein (1981). It will not be treated here.)

An aspect of non-restrictive models is to create distrust of multiple comparison procedures, which are claimed to be overcautious. Hence one resorts to ordinary testing of a null hypothesis H_f (see discussion after equation (7)), resulting in a significance which may appear reasonable, but which is really not justified by the overcautious a priori attitude. What seems to happen is that two errors, overcautious model and too daring test, roughly cancel each other and give a "satisfactory" result. If now one of the errors are removed (viz. by using the multiple comparison procedure) then the other error (overcautious model) will stand out in its glaring absurdity. Hence the use of adequate restrictive models is almost imperative in connection with multiple comparison procedures.

8. Statistics collected by official Central statistical bureaus

The analysis of categorical observations discussed in the present paper really represents a very basic problem about the kind of statistics that are published in large quantities by government statistical bureaus. The tables often present data that fall in ones lap as a result of government activities, or they present results of observations collected by means of questionaires to study industry, trade, social welfare, education etc. The purpose of collecting such data may be multifarious or even diffuse, but nobody would deny that they may contain important and unexpected informations which may be revealed by "snooping" around in the tables. Of course, the reading of such tables is a challenge to the statistical inference theory. The fact that statistical inference theory has been succesful mostly in cases of carefully planned statistical experiments with a clear purpose, should not induce statisticians to believe that inference theory is meant only for such situations. That would be to put the cart before the horse. Such statistical experiments with a finite, and preferable low, number of possible decisions could often more efficiently be handled by other methods than the multiple comparison methods, e.g. by adjustment of the level of significance by means of Bonferroni's inequalities. This is clearly impossible in the case of an infinite number of possible decisions. The methological problem faced with when reading tables of statistics of the kind just mentioned are not easy. We may discover interesting features and want to test if they are real. We cannot apply the method which would be adequate if we had suspected the relationship in advance. Hence we have to adopt the soul searching attitude of defining the state of our mind before we looked at the data. Some may object to such a procedure. However, it is good to be reminded that statistical inference concerning (e.g.) official statistics is as subjective as just that. On the other hand to discard such data altogether, as being of no concern to the mathematical statistician, would be a too easy way out of the difficulties.

It should also be mentioned that the published statistics from official statistical bureaus often imply a choice of statistical decision functions (methods), viz. crude statistical grouping, in order to expose some, but not all, interesting features. Often the original observations are unavailable, which precludes the use of the methods of the present paper, or any other efficient method to expose further interesting features.

II. GENERAL THEORY

1. The likelihood. Assumptions and definitions

We assume the a priori model and the null-state given by equations (3) and (4) in I.3.

For convenience we shall also assume that we may, after a transfermation in the parameter space, write

$$\eta = (\nu_1,\ldots,\nu_w,\theta_1,\ldots,\theta_t) = (\nu,\theta)$$

$(w = v-t)$, such that the null-state (4) takes the form

$$H_0\colon\ \nu_1=\nu_2=\ldots=\nu_w=0 \tag{19}$$

Hence after the transformation the ϕ_i in (4) have the form

$$\phi_i(\theta) = 0,\quad i = 1,\ldots,w,\quad \phi_i(\theta) = \theta_{i-w}\ ,\ i = w+1,\ldots,v \tag{20}$$

Such a "reduced" formulation is usually easily constructed from the original formulation (4). Consider e.g. the case of a two-way layout in a multinomial situation with cell probabilities π_{ij} and null-state of independence $\pi_{ij} = \pi_{i+}\cdot\pi_{+j}$; where $\pi_{i+} = \Sigma_j\pi_{ij}$ and $\pi_{+j} = \Sigma_i\pi_{ij}$; $i = 1,\ldots,r$; $j = 1,\ldots,s$. By reparametrization we use the descriptive form $\eta_{ij} = \log \pi_{ij} = \mu+\alpha_i+\beta_j+\nu_{ij}$, where $\alpha_1 = \beta_1 = \nu_{1i} = \nu_{j1} = 0$, $\mu = -\log \Sigma_{i,j}\, e^{\alpha_i+\beta_j+\nu_{ij}}$ (since $\Sigma\pi_{ij}=1$). Hence the null-state is $\nu_{ij} = 0$, and there is a one-to-one correspondence between $\eta = (\eta_{11},\ldots,\eta_{rs})$ and $\tilde\eta = (\nu,\theta) = (\nu_{22},\ldots,\nu_{rs};\alpha_2,\ldots,\alpha_r,\beta_2,\ldots,\beta_s)$. $\tilde\eta$ is a reduced formulation and $v = rs-1, t=r+s-1$. In general we may from the general formulation (3) and (4) obtain a reduced formulation in the following manner. We may transform η to $\tilde\eta$ by a one-to-one transformation $\eta = T(\tilde\eta)$, $\tilde\eta = T^{-1}(\eta)$, which is smooth; i.e. both T and T^{-1} have continuous first-order derivatives. Then $g_a(X;\eta)$ is transformed into $\tilde g_a(X;\tilde\eta) = g_a(X;T(\tilde\eta))$, $L(X,\eta)$ into $\tilde L(X,\tilde\eta) = L(X;T(\tilde\eta))$ and the null-state $\eta = \phi(\theta)$ into $\tilde\eta = T^{-1}\phi(\theta) = \tilde\phi(\theta)$.

Among all possible T we choose one which is such that $T_i(0,\ldots,0,\tilde{\eta}_{w+1},\ldots,\tilde{\eta}_v) = \phi_i(\tilde{\eta}_{w+1},\ldots,\tilde{\eta}_v)$, where the ϕ_i are those occurring in the general formulation (4). We now write $(\tilde{\eta}_1,\ldots,\tilde{\eta}_w) = \nu = (\nu_1,\ldots,\nu_w)$ and $(\tilde{\eta}_{w+1},\ldots,\tilde{\eta}_v) = \theta = (\theta_1,\ldots,\theta_t)$, $t = v-w$. Then $\eta = \phi(\theta)$ is equivalent to $\nu_1 = \cdots \nu_w = 0$ and $\tilde{\phi}$ has the special form (20).

Convenience may dictate using reduced formulation (ν,θ) in the course of mathematical derivations.

Sometimes the reduced formulation arises naturally as part of the a priori modelling. The purpose of the statistical analysis may make it natural to focus attention on some indexparameters $\nu_i = T_i(\eta)$; $i = 1,2,\ldots$ $w < v$. Thus we are interested in contrasts $f(\nu) = f(\nu_1,\ldots,\nu_w)$ relatively to the null-state $\nu = 0$. We assume that we may add t functions $\theta_{i-w} = T_i(\eta)$; $i = w+1,\ldots,v$ such that $\tilde{\eta} = (\nu,\theta) = T(\eta) = (T_1(\eta),\ldots,T_v(\eta))$ is a one-to-one smooth transformation. Since $\eta = T^{-1}(\tilde{\eta})$, we may write the null-state $\eta = T^{-1}(0,\theta) = \phi(\theta)$. Contrasts $f(\nu)$ which depends on η only through ν, will be called <u>focalized</u>. Thus in the case of a three-way lay-out in analysis of variance of observations X_{ijk}, with $EX_{ijk} = \xi_{ijk}$, and no three factor interaction, we may take interest in main effects $\nu_{i++} = \bar{\xi}_{i++}-\bar{\xi}, \nu_{+j+} = \bar{\xi}_{+j+}-\bar{\xi}, \nu_{++k} = \bar{\xi}_{++k}-\bar{\xi}$ only. These are the components of ν. We add two-factor interactions $\theta_{ij+} = \bar{\xi}_{ij+}-\bar{\xi}_{i++}-\bar{\xi}_{+j+}-\bar{\xi}_{++k}+\bar{\xi}$ etc. and $\bar{\xi}$ to obtain a one-to-one correspondence between ξ_{ijk} and (ν,θ), where θ has $\bar{\xi}$ and the two-factor interactions as components. The null-state is $\xi_{ijk} = \xi+\theta_{ij+}+\theta_{j+k}+\theta_{+jki}$, which corresponds to the null-state $\nu_{i++} = 0$, $\nu_{+j+} = 0$, $\nu_{++j} = 0$ in the reduced formulation.

Returning to the general theory the maximum likelihood estimate a priori η^* is defined to satisfy

$$\frac{\partial \log L(x;\eta^*)}{\partial \eta_j} = \sum_{a=1}^{s} \sum_{m=1}^{n_a} \frac{\partial \log g_a(x_{am};\eta^*)}{\partial \eta_j} = 0, \quad j=1,\ldots,v \qquad (21)$$

(regardless of whether η^* maximizes (3) or not).

Similarly the maximum likelihood estimate $\hat{\eta} = \phi(\hat{\theta})$ is defined to satisfy

$$\frac{\partial \log L(X;\phi(\hat{\theta}))}{\partial \theta_j} = 0 \; ; \quad j = 1,\dots,t \tag{22'}$$

The estimate η^* and the likelihood ratio Q given by (9) have well-konwn asymptotic properties under general regularity conditions.

For our practical purpose we shall give these regularity conditions a simple and easily checkable form. The simplification consists in assuming the consistency of the estimates η^* and $\hat{\eta}$, since this assumption is usually easily verified and we do not want to dwell to much on the philosophical problem of why we get consistent estimates. Furthermore if this assumption is made, then the additional regularity conditions made below, also have an easily checkable form. The regularity conditions usually made in the litterature are intended to secure consistency, asymptotic normality, proper maximization, optimality, etc. It is difficult to sort out the condition that need checking after consistency has been verified. Theorems 1 and 2 remedy this state of affairs. The conventional assumptions are usually disregarded (see Bishop et. al. (1974) p. 69).

In our asymptotic consideration we shall let all $n_a \to \infty$ in such a manner that $n_a/n \to c_a > 0$.

Regularity conditions about our model and null-state. The mode of convergence.

Assumption A. η varies a priori in an open subset of the v-space The second order derivatives

$$\frac{\partial^2 \log g_a(X_{am};\eta)}{\partial\eta_i \, \partial\eta_j} \tag{23}$$

exist and are continuous functions of η, uniformly in X_{am}. First- and second-order differentiation of $\int g_a d\mu$ may be taken under the sign of integration. The matrix $\lambda(\eta) = \Sigma_{a=1}^{s} c_a \lambda_a(\eta)$ where

$$\lambda_a(\eta) = (E \frac{\partial \log g_a(X_{am};\eta)}{\partial \eta_i} \cdot \frac{\partial \log g_a(X_{am};\eta)}{\partial \eta_j})_{i,j=1,2,\ldots,v} \tag{24}$$

is non-singular. The Jacobian $v \times t$ matrix $D\phi = (\frac{\partial \phi_i}{\partial \theta_j})$ has maximum rank and is a continuous function of θ.

Remark: Concerning the continuity of (23) as a function of η, uniformly, it was perhaps restrictive. However in the multinomial situation it is trivial because X_{am} assumes only a finite number of values. It also holds in the multivariable normal case and the multiple decrement survival models. The assumption of maximal rank of $D\phi$ secures the existence of a reduced formulation, see (33) below.

Below we shall partly let η be constant and partly vary $\eta = \eta^{(n)}$ as the number of observations n goes to infinity. Hence for each n the n observations $X = (X_{am})$ have density $L(x;\eta^{(n)})$ where $L(x;\eta)$ is given by (3). The need for considering changing $\eta = \eta^{(n)}$ arises from the fact that if the number of observations is large it is desirable to consider η close to the null-state. The asymptotic properties of a reasonable inference procedure are often trivial for η fixed $\notin H_0$.

In the assumptions B_1, B_2, C_1, C_2 below it is understood that $\eta^{(n)}$ may vary and converges to some η.

Assumption B_1. The sets of X for which the likelihood equations (21) have unique solutions have probabilities that go to 1 as $n \to \infty$. (Outside these sets η^* may be defined in any manner.)

Assumption B_2. The same assumption and convention are made about equations (22) and $\hat{\eta}$.

Assumption C_1. η^* is a consistent estimate of η, i.e.

$$\text{plim } \eta^* = \eta \quad \text{for all} \quad \eta$$

(note that this is equivalent to $P[|\eta^*-\eta|>\varepsilon)\cap A_n] \to 0$ where the A_n are the sets in assumption B_1).

Assumption C_2. $\hat{\eta}$ is a consistent estimate of $\eta \in H_0$ in the sense that

$$\text{plim}\ \hat{\eta} = \eta \quad \text{for all} \quad \eta \in H_0$$

(Note that we only require that $\eta \in H_0$, not that $\eta^{(n)} \in H_0$.)

For sundry purposes below we need relationships between the blocks of

$$\lambda = \begin{pmatrix} \Lambda_{11}, & \Lambda_{12} \\ \Lambda_{21}, & \Lambda_{22} \end{pmatrix}, \qquad \sigma = \begin{pmatrix} \Sigma_{11}, & \Sigma_{12} \\ \Sigma_{21}, & \Sigma_{22} \end{pmatrix} \tag{25}$$

where λ refers to a reduced formulation, $\sigma = \lambda^{-1}$ and Λ_{11} , Σ_{11} are of order $(w \times w)$, $w = v-t$. By block multiplication of $\lambda\sigma = I$, we get

$$\Sigma_{11} = (\Lambda_{11}-\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})^{-1}, \quad \Lambda_{22}^{-1}\Lambda_{21} = -\Sigma_{21}\Sigma_{11}^{-1} \tag{26}$$

Let us now consider the speed of convergence of $\eta^{(n)} \notin H_0$ to some $\eta \in H_0$. The convergence may be at least as fast as $1/\sqrt{n}$ goes to 0 when $n \to \infty$, i.e.

$$\Delta^{(n)} = \sqrt{n}\ (\eta^{(n)}-\eta) \to \Delta(\text{say}) \tag{27}$$

This property is invariant with respect to a smooth transformation $\eta = T(\tilde{\eta})$. This is seen by a Taylor expansion $\eta_i^{(n)}-\eta_i = DT_i(\tilde{\eta}')(\tilde{\eta}^{(n)}-\tilde{\eta})$, where $\tilde{\eta}'$ is between $\tilde{\eta}^{(n)}$ and $\tilde{\eta}$. [As a principle of notation we write $(\frac{\partial g}{\partial \eta_1},\ldots,\frac{\partial g}{\partial \eta_v}) = Dg(\eta)$ for any function $g(\eta)$.] In the case of reduced formulation $\tilde{\eta} = (\nu,\theta)$, (27) implies that $\sqrt{n}\ \nu^{(n)} \to \Delta^I(\text{say})$.

We shall say that $\eta^{(n)} \to \eta \in H_0$ more slowly than $1/\sqrt{n} \to 0$ if for some reduced formulation (ν,θ) we have that $\eta^{(n)} = (\nu^{(n)},\theta^{(n)})$ is such that

$$\varepsilon_n \nu^{(n)} \to \text{some}\ \Delta^I \neq 0, \text{ where } \varepsilon_n \to \infty,\ \varepsilon_n/\sqrt{n} \to 0 \tag{28}$$

Obviously this property is invariant with respect to smooth transformations.

(27) and (28) are concerned with the speed of convergence in the direction of the ν-coordinates (index parameters). In studying the asymptotic power of our statistical methods we shall also be concerned with the orthogonal (precipitous) speed of convergence to H_0. For this purpose we measure the distance $\rho(\eta^{(1)},\eta)$ between $\eta^{(1)}$ and η given by

$$\rho^2(\eta^{(1)},\eta) = (\eta^{(1)}-\eta)'\lambda(\eta)(\eta^{(1)}-\eta) \tag{29}$$

where $\lambda(\eta)$ is the information matrix. In order to find the distance from an arbitrary $\eta^{(1)}$ to H_0 we minimize (29) when $\eta = \phi(\theta)\in H_0$ with respect to θ, and obtain a minimizing $\bar{\theta}$ given by

$$-2\sum_{i,j}\frac{\partial\phi_i(\bar{\theta})}{\partial\theta_k}(\eta_j^{(1)}-\phi_j(\bar{\theta}))\lambda_{ij}(\phi(\bar{\theta}))+\sum_{i,j}(\eta_i^{(1)}-\phi_i(\bar{\theta}))\frac{\partial\lambda_{ij}(\phi(\bar{\theta}))}{\partial\theta_k}(\eta_j^{(1)}-\phi_j(\bar{\theta})) = 0$$

for $k = 1,2,\ldots,t$. Then $\bar{\eta} = \phi(\bar{\theta})$ is the footpoint of $\eta^{(1)}$ and the squared distance from $\eta^{(1)}$ to H_0 is obtained by replacing η by $\bar{\eta}$ in (29). We now assume that

$$\Delta^{(n)} = \sqrt{n}(\eta^{(n)}-\phi(\bar{\theta}^{(n)}) \rightarrow \Delta \tag{30}$$

in such a manner that for each $\eta^{(n)}$, $\phi(\bar{\theta}^{(n)})$ is its footpoint. Thus it is the speed in the orthogonal direction that is at least as fast as $1/\sqrt{n} \rightarrow 0$.

Now we replace $\eta^{(1)}$, $\bar{\theta}$ by $\eta^{(n)}$, $\bar{\theta}^{(n)}$ in the above equations for $\bar{\theta}$ and then multiply the equations by $\sqrt{n}$. Going to the limit we then get

$$(D\phi)'\ \lambda\ \Delta = 0 \tag{31}$$

That this property is invariant with respect to a smooth transformation $\eta = T(\tilde{\eta})$ is seen in the following manner.
By (6) and (24) we have

$$\sigma(\eta)^{-1} = \lambda(\eta) = \sum_{a=1}^{s} c_a E[D\log g_a(X_{am},\eta)]\,[D\log g_a(X_{am},\eta)]' \tag{32}$$

and $\tilde{\lambda}(\tilde{\eta})$ is obtained by replacing $g_a(\cdot,\eta)$ by $\tilde{g}(\cdot,\tilde{\eta}) = g(\cdot,T(\tilde{\eta}))$.

By the chain rule for differentiation we have $D \log g_a(x,\tilde{\eta})' = DT(\tilde{\eta})'D \log g_a(x;\eta)'$. By substitution in $\tilde{\lambda}(\tilde{\eta})$ we then get $\tilde{\lambda}(\tilde{\eta}) = DT(\tilde{\eta})'\lambda(\eta)DT(\tilde{\eta})$. From $\phi(\theta) = T(\tilde{\phi}(\theta))$ we have $D\phi(\theta) = DT(\tilde{\eta})D\tilde{\phi}(\theta)$. Furthermore using again the Taylor expansion $\sqrt{n}(T_i(\tilde{\eta}^{(n)})-T_i(\tilde{\phi}(\theta^{(n)})) = \sqrt{n}\, DT_i(\tilde{\eta}')\ (\tilde{\eta}^{(n)}-\tilde{\phi}(\bar{\theta}^{(n)}))$, we get $\Delta = DT(\tilde{\eta})\tilde{\Delta}$, where $\tilde{\Delta} = \lim \tilde{\Delta}^{(n)}$ and $\tilde{\Delta}^{(n)}$ is defined by replacing $\Delta^{(n)},\eta^{(n)},\phi$ by $\tilde{\Delta}^{(n)},\tilde{\eta}^{(n)},\tilde{\phi}$ in (30). Combining we then get $(D\tilde{\phi})'\tilde{\lambda}\tilde{\Delta} = (D\phi)'\lambda\Delta$, which proves the invariance property of the restriction (31) on Δ.

Let us now consider what (31) looks like in the reduced formulation. By assumption A, $D\phi$ has rank t. Hence we may arrange that the submatrix of $D\phi$ consisting of the t last rows is non--singular. We now use the following transformation from η to the reduced form (ν,θ),

$$\begin{aligned} \eta_j &= \phi_j(\theta)+\nu_j \ ; \quad j = 1,\dots,w \\ \eta_j &= \phi_j(\theta) \qquad ; \quad j = w+1,\dots,v \end{aligned} \tag{33}$$

Then (31) becomes

$$\tilde{\Lambda}_{21}\,\Delta^{I} + \tilde{\Lambda}_{22}\,\Delta^{II} = 0$$

where the $\tilde{\Lambda}_{ij}$ are defined as in (25) and $\Delta' = (\Delta^{I'},\Delta^{II'})$.
Hence (see (26))

$$\Delta^{II} = -\tilde{\Lambda}_{22}^{-1}\,\tilde{\Lambda}_{21}\,\Delta^{I} = \tilde{\Sigma}_{21}\,\tilde{\Sigma}_{11}^{-1}\,\Delta^{I} \tag{34}$$

(Thus Δ^{II} depends on Δ^{I} similarly to the regression of $\hat{\theta}-\theta^*$ on ν^*, see equation (53) below.)

2. Properties of the maximum likelihood estimates and the likelihood ratio.

Theorem 1. Under the regularity conditions A, B_1 and C_1 of II.1 the maximum likelihood estimates (21) are asymptotically normal with mean η and covariance matrix

$$\frac{1}{n}\,\lambda(\eta)^{-1} = \frac{1}{n}\,\sigma(\eta), \text{ where } \lambda(\eta) = \sum_{a=1}^{s} c_a \lambda_a(\eta) \tag{35}$$

and $\lambda_a(\eta)$ is given by (24) or

$$\lambda_a(\eta) = -\left(E \frac{\partial^2 \log g_a(X,\eta)}{\partial\eta_i \, \partial\eta_j} \right)_{i,j=1,\ldots,v} \tag{36}$$

More precisely if $\eta^{(n)} \to \eta$, then $\sqrt{n}(\eta^* - \eta^{(n)})$ converges in distribution to the multivariable normal $(0,\lambda(\eta)^{-1})$. Hence if $\eta^{(n)} = \eta+\Delta^{(n)}/\sqrt{n}$, where $\Delta^{(n)} \to \Delta$, then $\sqrt{n}(\eta^*-\eta)$ converges in distribution to the normal $(\Delta,\lambda(\eta)^{-1})$.

<u>Theorem 2</u>. Assume the regularity conditions A, B_1, B_2, C_1, C_2 and let $\eta = T(\tilde{\eta}) = T(\nu,\theta)$ be a smooth transformation to a reduced formulation (see (19) and (29)). Consider the following statistics

$$Z_1 = -2 \log \frac{L(X,\hat{\eta})}{L(X,\eta^*)} \tag{37}$$

$$Z_0 = n(\hat{\eta}-\eta^*)'\sigma(\eta^*)^{-1}(\hat{\eta}-\eta^*) \tag{38}$$

$$Z_0' = n\nu^{*\prime}\,\tilde{\Sigma}(\tilde{\eta}^*)^{-1}\,\nu^* \tag{39}$$

where $\tilde{\Sigma}(\tilde{\eta}) = \tilde{\Sigma}_{11}(\tilde{\eta})$ is given by (25).

<u>a</u> Let $\eta^{(n)} \to \eta \in H_0$ at least as fast as $1/\sqrt{n} \to 0$. Then Z_1, Z_0, Z_0' are equivalent to each other up to limit in probability measure and converge in distribution to the eccentric chi-square distribution with $w = v-t$ degrees of freedom and eccentricity

$$\kappa = \Delta^{I\prime}\tilde{\Sigma}(\tilde{\eta})^{-1}\,\Delta^{I} \tag{40}$$

where Δ^I is given by $\tilde{\Delta}' = (\Delta^{I\prime},\Delta^{II\prime})$ and $\tilde{\Delta}$ by (27).
If $\eta^{(n)} \to H_0$ with an orthogonal speed at least as fast as $\frac{1}{\sqrt{n}} \to 0$, see (30) and (31), then κ is given by

$$\kappa = \Delta'\sigma(\eta)^{-1}\Delta \tag{41}$$

This gives us the asymptotic power of the criterion $Z_1 > z$.

Remark: (41) shows that the "asymptotic" eccentricity is $n(\eta-\phi(\bar{\theta}))'\sigma(\bar{\eta})^{-1}(\eta-\phi(\bar{\theta}))$. The analogy with the normal variance- and regression analyses is obvious.

b (Assumptions B_2, C are not needed here.) If $\eta^{(n)} \to \eta \in H_0$ more slowly than $1/\sqrt{n}$ (see (28)), then $Pr(Z_1<z)$, $Pr(Z_0<z)$, $Pr(Z_0'<z) \to 0$. Hence Z_1, Z_0, Z_0' do not converge in distribution and the asymptotic power is 1, i.e. the likelihood ratio test is what may be called locally consistent. The test of H_0 based on Z_0' is also consistent, i.e. the power goes to 1 if $\eta^{(n)} \to \eta \notin H_0$.

To prove the theorems we need the following lemmas.

Lemma 1. $X_1,\dots,X_n$ are independent observations of X and Z_n; $n = 1,2,\dots$; a sequence of random variables such that plim $Z_n = \zeta$ (non-random). Furthermore $F(X,\zeta)$ is a continuous function of ζ, uniformly in X and $EF(X,\zeta) = \mu$ exists for all ζ. Then

$$\bar{F}_n(Z_n) = \frac{1}{n}\sum_{m=1}^{n} F(X_m,Z_n) \xrightarrow{p} \mu$$

(It is also true with probability 1.)

Lemma 2. Let X have density $g(x;\eta)$ with respect to a measure μ, and let $V_1(X,\eta),\dots,V_v(X,\eta)$ be v functions such that

$$EV_i(X,\eta) = 0, \ \mathrm{cov}(V_i(X,\eta),V_j(X,\eta)) = \lambda_{ij}(\eta)$$

where $\lambda(\eta) = (\lambda_{ij}(\eta))$ is non-singular. $\lambda_{ij}(\eta)$, $V_i(x,\eta)$ and $g(x,\eta)$ are continuous functions of η. For each n let X_{mn}; $m = 1,\dots,n$; be independent with common density $g(x;\eta^{(n)})$, where $\eta^{(n)} \to \eta$. Then the vector

$$\left(\frac{1}{\sqrt{n}}\sum_{m=1}^{n} V_i(X_{mn};\eta^{(n)})\right)_{i=1,2,\dots,v}$$

converges in distribution to the multivariable normal $(0,\lambda(\eta))$.

The proofs of the two lemmas are given below in II.3.

Proof of Theorem 1. We assume first that $\eta^{(n)} = \eta$. We replace the double subscripts (a,m) in (3), (6) etc. by single letters, which may be m, going from 1 to n. Thus we write X_m in place of X_{am}. We also write, somewhat inconsistently, $g_m(x,\eta)$, λ_m and c_m; $m = 1,\ldots,n$; in place of $g_a(x;\eta)$, λ_a, and c_a; $a = 1,2,\ldots,s$. Thus the g_m, λ_m, c_m are independent of m within sections $S_1,\ldots,S_s$ of $(1,\ldots,n)$ of length $n_1,\ldots,n_s$ respectively ($S_1 = (1,\ldots,n_1)$, etc.)

We introduce

$$V_{im}(\eta) = \frac{\partial \log g_m(X_m,\eta)}{\partial \eta_i} \; ; \; W_{ijm}(\eta) = \frac{\partial^2 \log g_m(X_m;\eta)}{\partial \eta_i \, \partial \eta_j} \tag{42}$$

and we have by differentiation of $\int g_m d\mu = 1$,

$$\begin{aligned} E\, V_{im}(\eta) = 0,\ \lambda_{ijm} &= \operatorname{cov}(V_{im}(\eta), V_{jm}(\eta)) = \\ &= E\, V_{im}(\eta)\, V_{jm}(\eta) = -\, E\, W_{ijm}(\eta) \end{aligned} \tag{43}$$

We also introduce,

$$\bar{V}_i(\eta) = \frac{1}{n} \textstyle\sum_m V_{im}(\eta),\ \bar{W}_{ij}(\eta) = \frac{1}{n} \sum_m W_{ijm}(\eta) \tag{44}$$

Note that

$$\bar{V}_i(\eta) = \sum_{a=1}^{s} \frac{c_a^{(n)}}{n_a} \sum_{m\in S_a} V_{im}(\eta),\ \bar{W}_{ij}(\eta) = \sum_{a=1}^{s} \frac{c_a^{(n)}}{n_a} \sum_{m\in S_a} W_{ijm}(\eta) \tag{45}$$

where $c_a^{(n)} \to c_a$. Obviously,

$$\operatorname{plim} \bar{V}_i(\eta) = 0,\ \operatorname{plim} \bar{W}_{ij}(\eta) = -\textstyle\sum c_m \lambda_{ijm} = -\lambda_{ij} \tag{46}$$

The likelihood equations (21) may be written

$$\sum_{m=1}^{n} V_{im}(\eta^*) = 0 \; ; \quad i = 1,\cdots,v \tag{47}$$

from which we get

$$0 = \bar{V}_i(\eta^*) = \bar{V}_i(\eta) + \sum_{j=1}^{v} (\eta_j^* - \eta_j)\bar{W}_{ij}(\eta') \tag{48}$$

where η' is between η and η^* (componentwise).

Equations (47) and (48) are true except on a set with the probability that goes to 0. Since $\operatorname{plim} \eta^* = \eta$, then $\operatorname{plim} \eta' = \eta$. Applying Lemma 1 within each of the s subsequences of X_m (see (45)), we get

$$\operatorname{plim} \bar{W}_{ij}(\eta') = -\lambda_{ij} \tag{49}$$

From (48) we get

$$\bar{W}(\eta^* - \eta) = -\bar{V}(\eta)$$

where $\bar{W}$ is the matrix $(\bar{W}_{ij}(\eta'))$. Hence

$$\sqrt{n}(\eta^* - \eta) = -\bar{W}^{-1} \sqrt{n}\, \bar{V}(\eta) \tag{50}$$

except on a set the probability of which goes to 0.

By Laplace theorem it follows that $\sqrt{n}\, \bar{V}(\eta)$ converges in distribution to the multivariable normal with mean 0 and covariance matrix λ. Hence by (49) and (50), $\sqrt{n}(\eta^* - \eta)$ converges in distribution to the multivariable normal with mean 0 and covariance matrix $\lambda^{-1}\lambda\lambda^{-1} = \lambda^{-1}$. (The last reasoning is easily expanded upon in full details, taking into account that (47) is true and $\bar{W}$ nonsingular only on a set, the probability of which goes to 1.) Hence the Theorem 1 is proved if $\eta^{(n)} = \eta$. That the same is true if $\eta^{(n)}$ varies with n such that $\eta^{(n)} \to \eta$ follows immediately from Lemma 2.

<u>Proof of Theorem 2</u>: By a Taylor expansion using (21), we get

$$\log L(X;\hat{\eta}) = \log L(X,\eta^*) + \frac{n}{2} \sum_{i,j=1}^{v} (\hat{\eta}_i-\eta_i^*)(\hat{\eta}_j-\eta_j^*)\bar{W}_{ij}(\eta')$$

where η' is between $\hat{\eta}$ and η^*. Hence

$$Z_1 = -n(\hat{\eta}-\eta^*)'\bar{W}(\eta')(\hat{\eta}-\eta^*) \tag{51}$$

We now prove the results in <u>Theorem 2a</u>. Consider first the case of <u>reduced formulation</u>, where $\tilde{\eta} = (\nu,\theta)$. From (3), (22), (42), (44) we have $\bar{V}_i(0,\hat{\theta}) = 0$; $i = w+1,\dots,v$. Replacing η by $(0,\hat{\theta})$ in (48) we then get that,

$$\sum_{j=1}^{w} \bar{W}_{ij}(\eta'')\sqrt{n}\nu_j^* + \sum_{j=w+1}^{v} \bar{W}_{ij}(\eta'')\sqrt{n}(\theta_{j-w}^*-\hat{\theta}_{j-w}) = 0 \tag{52}$$

except on a set the probability of which goes to 0. Here η'' is between (ν^*,θ^*) and $(0,\hat{\theta})$. By (52) we easily express $\sqrt{n}(\theta^*-\hat{\theta})$ and $\sqrt{n}(\eta^*-\hat{\eta})$ by means of $\sqrt{n}\nu^*$ and $W_{ij}(\eta'')$.

By assumption C_2 plim $\eta'' = \eta$. Thus by Theorem 1 and Lemma 1, $\sqrt{n}(\theta^*-\hat{\theta})$ and $\sqrt{n}(\eta^*-\hat{\eta})$ converge in distribution. We get from (52) and (43),

$$\sqrt{n}(\hat{\theta}-\theta^*) = \Lambda_{22}^{-1}\Lambda_{21}\sqrt{n}\,\nu^* \tag{53}$$

$$Z_1 = n(\hat{\eta}-\eta^*)'\lambda(\eta)(\hat{\eta}-\eta^*) \tag{54}$$

up to equivalence in probability limit. (We leave out the tildas here and below.) The asymptotic distribution of $\sqrt{n}(\theta^*-\hat{\theta})$ follows from (53) and the second equation (26).

Replacing $\lambda(\eta)$ by $\lambda(\eta^*)$ it is seen that we have proved that $\text{plim}(Z_1-Z_0) = 0$. By block partioning, (54) may be written

$$Z_1 = n[\nu^*{}'\Lambda_{11}\nu^*-2\nu^*{}'\Lambda_{12}(\hat{\theta}-\theta^*)+(\hat{\theta}-\theta^*)'\Lambda_{22}(\hat{\theta}-\theta^*)] \tag{55}$$

Introducing (53) we get

$$Z_1 = n\,\nu^*{}'(\Lambda_{11}-\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})\nu^*$$

Hence from (26) we get

$$Z_1 = n\,\nu^{*\prime}\,\Sigma_{11}(\eta)^{-1}\nu^{*} \tag{56}$$

up to equivalence in probability limit. Replacing $\Sigma_{11}(\eta)$ by $\Sigma_{11}(\eta^*)$ we obtain that $\mathrm{plim}(Z_1-Z_0') = 0$. On the other hand it follows immediately from Theorem 1 that the right hand side of (56) converge in distribution to the chi-square distribution with w degrees of freedom and eccentricity given by (40). To prove (41) in the case when $\eta^{(n)} \to H_0$ with an orthogonal speed at least as fast as $1/\sqrt{n} \to 0$, we consider the right hand side of (41) in the case of reduced formulation

$$\Delta'\lambda\,\Delta = (\Delta^{I'},\Delta^{II'})\lambda\begin{pmatrix}\Delta^{I}\\ \Delta^{II}\end{pmatrix}$$

Introducing $\Delta^{II} = -\Lambda_{22}^{-1}\,\Lambda_{21}\,\Delta^{I}$ (see (34)) and λ given by (25) we obtain (40). Hence $\Delta'\lambda\,\Delta = \kappa$ in the case of reduced formulation.

Consider now <u>Theorem 2a in the case of a general formulation</u>. We then use the special transformation $\eta = T(\tilde{\eta})$ given by (33). Z_0' will refer to this transformation. Trivially Z_1 given by (37) is invariant with respect to any smooth transformation $\eta = T(\tilde{\eta})$. (See (21) and (22).) Hence $\mathrm{plim}(Z_1-Z_0') = 0$.

Consider now Z_0. We have by Taylor $\eta_i^*-\hat{\eta}_i = DT_i(\tilde{\eta}^{(i)})(\tilde{\eta}^*-\hat{\tilde{\eta}})$. Let $\overline{DT}$ denote the matrix consisting of the rows $DT_i(\tilde{\eta}^{(i)})$, $i = 1,2,\ldots,v$. Then

$$\sqrt{n}(\eta^*-\hat{\eta}) = \overline{DT}\,\sqrt{n}(\tilde{\eta}^*-\hat{\tilde{\eta}}) \tag{57}$$

However, by the remarks after equation (52) $\sqrt{n}(\tilde{\eta}^*-\hat{\tilde{\eta}}) = \sqrt{n}(\nu^*,\theta^*-\hat{\theta})$ converges in distribution. Hence the same is true of $\sqrt{n}(\eta^*-\hat{\eta})$. Thus the derivation leading to (54) does not depend upon the assumption about reduced formulation, and the expression for Z_1 given by (54) is true in the case of a general formulation. Hence $\mathrm{plim}(Z_1-Z_0) = 0$ still holds.

Z_0 has the same limit in distribution as Z_1 and Z_0'. The eccentricity κ given by (41) is now easily seen to be invariant with respect to a smooth transformation (see the invariance considerations after (31)). Hence 2a is proved.

We shall prove the assertions in 2b, viz. that Z_0' , Z_0 , Z_1 diverge in probability to infinity when $\eta^{(n)} \to \eta \epsilon H_0$ more slowly than $1/\sqrt{n}$ goes to 0 . We then prove first that Z_0' , Z_0 , Z_1 may be expressed in a form

$$Z = n \, \nu^{\star\,\prime} \, M \, \nu^{\star} \qquad (58)$$

except on a set the probability of which goes to 0, where plim M $= \tilde{\Sigma}_{11}(\tilde{\eta})^{-1}$. If $Z = Z_0'$, this is obvious with $M = \tilde{\Sigma}_{11}(\tilde{\eta})^{-1}$. For $Z = Z_1$ or Z_0 , (58) is meant to be true when Z_1 and Z_0 refer to a general formulation. To prove (58) when $Z = Z_1$ is defined by (37), we observe first that (37) is invariant with respect to smooth transformation (see (21)). Hence it is enough to prove (58) in the case of reduced formulation. For this purpose we partition $\bar{W}(\eta) = (\bar{W}_{ij}(\eta))$ in blocks

$$\bar{W} = \begin{pmatrix} \bar{W}_{(11)} \, , & \bar{W}_{(12)} \\ \bar{W}_{(21)} \, , & \bar{W}_{(22)} \end{pmatrix}$$

similarly to σ and λ in (25). By block multiplication in the right hand side of (51) we obtain similarly to (55),

$$Z_1 = n[\nu^{\star\prime}\bar{W}_{(11)}(\tilde{\eta}')\nu^{\star} - 2\nu^{\star\prime}\bar{W}_{(12)}(\tilde{\eta}')(\hat{\theta}-\theta^{\star}) + (\hat{\theta}-\theta^{\star})'\bar{W}_{(22)}(\tilde{\eta}')(\hat{\theta}-\theta^{\star})] \qquad (59)$$

On the other hand we get from (52)

$$\sqrt{n}(\hat{\theta}-\theta^{\star}) = \bar{W}_{(22)}(\tilde{\eta}'')^{-1}\bar{W}_{(21)}(\tilde{\eta}'')\sqrt{n}\,\nu^{\star}$$

Introducing in (59) we obtain (58), where M is a function of the $\bar{W}_{(ij)}(\tilde{\eta}')$ and $\bar{W}_{(ij)}(\tilde{\eta}'')$. (It is noteworthy that, contrary to (56), (58) is true for $Z = Z_1$ on a set, the probability of which goes to 0, and (58) does not require that $\sqrt{n}\,\nu^{\star}$ converges in distribution.) Now, since plim $\bar{W}_{(ij)}(\tilde{\eta}') =$ plim $\bar{W}_{(ij)}(\tilde{\eta}'') = -\Lambda_{ij}$; i,j = 1,2 ; then the algebra from (55) to (56) shows that plim $M = \Sigma_{11}(\tilde{\eta})^{-1}$. Hence our assertion about (57) in the case when $Z = Z_1$ is proved.

To prove (58) in the case of $Z = Z_0$, we start from Z_0 given by (38) in the case of a general formulation. We study the effect on

Z_0 of a smooth transformation $\eta = T(\tilde{\eta})$. By (58) and (32), and the equations given after (32) and (27) we get for (38)

$$Z_0 = n(\tilde{\eta}^* - \hat{\tilde{\eta}})'(\overline{DT})'[DT(\tilde{\eta}^*)']^{-1}(\lambda(\tilde{\eta}^*))DT(\tilde{\eta}^*)^{-1}\overline{DT}(\tilde{\eta}^* - \hat{\tilde{\eta}})$$

Now $\operatorname{plim} DT(\tilde{\eta}^*)^{-1}\overline{DT} = DT(\tilde{\eta})(DT(\tilde{\eta}))^{-1} = I$. Hence

$$Z_0 = n(\tilde{\eta}^* - \hat{\tilde{\eta}})'m(\tilde{\eta}^* - \hat{\tilde{\eta}}) \tag{60}$$

where $\operatorname{plim} m = \tilde{\lambda}(\tilde{\eta})$.

Taking now $\tilde{\eta}$ to be a reduced formulation parameter, we now treat (60) similarly to (51), see (59), and obtain (58) where now M is a function of the $\bar{W}_{ij}(\tilde{\eta}'')$ and the 4 blocks $m_{(ij)}$; $i=1,2$; $j=1,2$; of m. Since $\operatorname{plim} m_{(ij)} = \operatorname{plim} \bar{W}_{ij}(\tilde{\eta}'') = -\Lambda_{ij}(\tilde{\eta})$ then $\operatorname{plim} M = \Sigma_{11}(\tilde{\eta})^{-1}$.

We can now show that $\operatorname{plim} Z = \infty$, where Z is given by (58). We introduce

$$Z(\nu) = n(\nu^* - \nu)'M(\nu^* - \nu)$$

which is ≥ 0 (and converges in distribution if $\nu = \nu^{(n)}$, by Theorem 1). We write (58),

$$Z = Z(\nu^{(n)}) + n\,\nu^{(n)'}M\nu^{(n)} + 2n(\nu^* - \nu^{(n)})'M\nu^{(n)}$$

which may also be written,

$$Z = Z(\nu^{(n)}) + (A_n + W_n\varepsilon_n/\sqrt{n})n/\varepsilon_n^2$$

where ε_n is given by (28), $\varepsilon_n\nu^{(n)} \to \Delta^I$,

$$A_n = \varepsilon_n\nu^{(n)'}M\,\varepsilon_n\nu^{(n)}, \quad W_n = 2\sqrt{n}(\nu^* - \nu^{(n)})'M\,\varepsilon_n\nu^{(n)}$$

Obviously $\operatorname{plim} A_n = \Delta^{I'}\Sigma(\tilde{\eta})^{-1}\Delta^I = A$ (say) > 0. W_n converges, by Theorem 1, to a normal distribution. Hence $\operatorname{plim}(A_n+W_n\varepsilon_n/\sqrt{n}) = A$. Since $n/\varepsilon_n^2 \to \infty$ it follows from the last expression for Z that $\operatorname{plim} Z = \infty$.

Consider now the case when $\eta^{(n)} \to \eta \notin H_0$, i.e. $\nu^{(n)} \to \nu \neq 0$. Then

$$\frac{Z_0'}{n} = \nu^{\star\prime}\Sigma(\eta^\star)^{-1}\nu^\star \underset{P}{\to} \nu'\,\Sigma(\tilde{\eta})^{-1}\nu > 0$$

and hence $\operatorname{plim} Z_0' = \infty$. Theorem 2 is proved.

3. Proof of the lemmas

Proof of Lemma 1 : To any $\varepsilon > 0$ and ζ there exists a δ such that $|\zeta'-\zeta| < \delta$ implies $|F(X;\zeta')-F(X;\zeta)| < \frac{\varepsilon}{2}$ for all X. Now $|\bar{F}_n(Z_n)-\mu| > \varepsilon$ implies either $|\frac{1}{n}\Sigma(F(X_m;Z_n)-F(X_m;\zeta))| > \frac{\varepsilon}{2}$ or $|\frac{1}{n}\Sigma_{m=1}^n(F(X_m,\zeta)-\mu)| > \frac{\varepsilon}{2}$. But the first possibility implies $|F(X_m,Z_n)-F(X_m,\zeta)| > \frac{\varepsilon}{2}$ for some $m = 1,2,\ldots,n$, which again implies $|Z_n-\zeta| > \delta$. Hence

$$\Pr(|\bar{F}_n(Z_n)-\mu|>\varepsilon) < \Pr(|Z_n-\zeta|>\delta)+\Pr(|\tfrac{1}{n}\Sigma F(X_m,\zeta)-\mu)|)>\tfrac{\varepsilon}{2}).$$

But the second term on the right hand side goes to 0 by Khintchine's theorem and the first term on the right hand side goes to 0 by the assumption.

Proof of Lemma 2: It is well known that it is enough to prove that

$$s_n = \sum_{i=1}^{v} t_i \frac{1}{\sqrt{n}} \sum_{m=1}^{n} V_i(X_{mn},\eta^{(n)})$$

converges in distribution to the normal with mean 0 and variance $t'\lambda t$, for any t. We introduce

$$T_n(x) = \sum_{i=1}^{v} t_i V_i(x,\eta^{(n)}), \text{ hence } s_n = \frac{1}{\sqrt{n}}\sum_{m=1}^{n} T_n(X_{nm})$$

We have,

$$E\,T_n(X_{nm}) = 0,\ \sigma_n^2 = \operatorname{var} s_n = \operatorname{var} T_n(X_{nm}) = \Sigma t_i t_j \lambda_{ij}(\eta^{(n)})$$

We shall now apply the following general proposition (see Loeve 1977, p.307).

Suppose that for each n ; $Y_{n1},\cdots,Y_{nn}$; $n = 1,2\ldots$ ad. inf. are independent, $E\,Y_{nm} = 0$, $\mathrm{var}\,Y_{nm} = \sigma^2_{nm} < \infty$, $\Sigma^n_{m=1}\sigma^2_{nm} = 1$ and for every $a > 0$, the Lindeberg assumption

$$\tau_n = \sum_{m=1}^{n} \int_{|y|>a} y^2 dG_{nm}(y) \to 0$$

as $n \to \infty$, where $G_{nm}(y) = \Pr(Y_{nm}<y)$. Then $S_n = \Sigma^n_{m=1}Y_{nm}$ converges in distribution to the normal (0,1).

We can apply this proposition to

$$Y_{nm} = \frac{T_n(X_{nm})}{\sqrt{n\ \mathrm{var}\ s_n}}$$

We only have to check the Lindeberg assumption, the other assumptions in the proposition are trivially true. We have

$$\tau_n = \sum_{m=1}^{n} \int_{|T_n(x)|>\sqrt{n}\sigma_n a} \frac{(T_n(x))^2}{n\,\sigma_n^2}\, g(x;\eta^{(n)})d\mu$$

$$= \int_{(T_n(x))^2 > a^2 n\sigma^2_n} \left(\frac{T_n(x)}{\sigma_n}\right)^2 g(x;\eta^{(n)})d\mu = \int_{M_n} f_n(x)d\mu$$

where we have introduced f_n and M_n. It is seen that $\int f_n(x)d\mu = 1$ and

$$f_n(x) \to \left(\frac{T(x)}{\sigma}\right)^2 g(x;\eta) = f(x)$$

where

$$T(x) = \sum t_i V_i(x;\eta),\ \sigma^2 = \mathrm{var}\ T(x) = \Sigma t_i t_j \lambda_{ij}(\eta) = \lim \sigma^2_n$$

and $\int f(x)d\mu = 1$. Then by Scheffé $f_n(x) \to f(x)$ in the mean (μ) and

$$|\tau_n - \int_{M_n} f(x)d\mu| < \int_{M_n} |f_n(x)-f(x)|d\mu < \int |f_n(x)-f(x)|d\mu \to 0$$

Hence τ_n has the same limit as

$$\tau'_n = \int_{M_n} f(x)d\mu$$

But since $T_n(x)$ has finite limit $T(x)$, and $\sigma_n^2 \to \sigma^2$, it follows that no x is contained in M_n for sufficiently large n. Hence $M_n \to$ empty set and $\tau_n' \to 0$. The Lemma 2 is proved.

4. The case of Darmois Koopman exponential class

Most application of the general theory which we are about to develop are aimed at situations where for each of the observations X_{am} the class of densities is a Darmois-Koopman exponential class,

$$g_a(x;\eta) = e^{\tau_{a0}(\eta)+\Sigma_{k=1}^{r_a} \tau_{ak}(\eta) Z_{ak}(x)} h_a(x) \qquad (61)$$

The class need not be regular, i.e. the set of all

$(\tau_{a1}(\eta),\cdots,\tau_{ar_a}(\eta))$ under variation of η need not be open in the r_a-space. Thus the situations with life testing, possibly with transfers and several causes of decrements, fall under the theory.

We shall use the notation

$$\bar{Z}_{ak} = \frac{1}{n_a} \sum_{m=1}^{n_a} Z_{ak}(X_{am}) \qquad (62)$$

We have for the likelihood,

$$L(X,\eta) = [\prod_{a=1}^{s} e^{c_a^{(n)}(\tau_{a0}(\eta)+ \sum_{k=1}^{r_a} \tau_{ak}(\eta)\bar{Z}_{ak})}]^n \prod_{a,m} h_a(X_{am}) \qquad (63)$$

The maximum likelihood equations a priori for the estimate η^* are

$$\sum_{a=1}^{s} c_a^{(n)}[\tau_{aoj}(\eta^*)+ \sum_{k=1}^{r_a} \tau_{akj}(\eta^*)\bar{Z}_{ak}] = 0 \qquad (64)$$

$j = 1,2,\ldots v$, where

$$\tau_{aki} = \frac{\partial\ \tau_{ak}}{\partial\ \eta_i} \qquad (65)$$

The maximum likelihood equations in the null-state ($\eta=\phi(\theta)$) for the estimate $\hat{\eta} = \phi(\hat{\theta})$ are

$$\sum_{a=1}^{s} c_a^{(n)} \sum_{j=1}^{v} [\tau_{aoj}(\phi(\hat{\theta}))+ \sum_{k=1}^{r_a} \tau_{akj}(\phi(\hat{\theta}))\bar{Z}_{ak}]\frac{\partial\phi_j(\hat{\theta})}{\partial\theta_h} = 0 \qquad (66)$$

$h = 1,\ldots,t.$

We shall also need

$$V_{jam} = \frac{\partial \log g_a(X_{am},\eta)}{\partial \eta_j} = \tau_{ao\,j}(\eta) + \sum_{k=1}^{r_a} \tau_{akj}(\eta) Z_{ak}(X_{am})$$

Hence from (43)

$$0 = \tau_{ao\,j} + \sum_{k=1}^{r_a} \tau_{akj}(\eta)\zeta_{ak}(\eta) \tag{67}$$

where

$$\zeta_{ak} = E\, Z_{ak}(X_{am}) \tag{68}$$

We get

$$V_{jam} = \frac{\partial \log g_a(X_{am},\eta)}{\partial \eta_j} = \sum_{k=1}^{r_a} \tau_{akj}[Z_{ak}(X_{am}) - \zeta_{ak}(\eta)] \tag{69}$$

Furthermore,

$$W_{ijam} = \frac{\partial^2 \log g_a(X_{am},\eta)}{\partial \eta_i\, \partial \eta_j} = \tau_{aoij} + \sum_{k=1}^{r_a} \tau_{akij}\, Z_{ak}(X_{am}) \tag{70}$$

where

$$\tau_{akij} = \frac{\partial \tau_{ak}}{\partial \eta_i \partial \eta_j} \tag{71}$$

The information matrix λ_a may be found from (69) or (70) (see (24) and (43)).

In the special case we are now considering, it is possible to replace the assumptions A, B_1, B_2, C_1, C_2 in II·1 by the following simpler assumptions.

A'. η varies a priori in an open set of the v-space. In the null-state $\eta = \phi(\theta)$, θ varies in an open set in the t-space and ϕ has continuous first order derivatives. The second order derivatives τ_{akij} (see (71)) are continuous functions of η. The continuity of W_{ijam} with respect to η is uniform with respect to Z. The matrix $\lambda = \Sigma\, c_a \lambda_a$ is non-singular. For each a, there is an open subset of vectors $t = (t_1,\dots,t_{r_a})$ in the r_a-space; $a = 1,\dots,s$; for which

$$\int e^{\Sigma t_k Z_{ak}(x)} h_a(x)\, d\mu$$

exists and which contains the range of $(\tau_{a1}(\eta),\dots,\tau_{ar_a}(\eta))$ under variation of η. The Jacobian $v{\times}t$ matrix $D\phi$ has rank $t < v$.

B'. The set S_n (resp. S_{no}) of X for which the likelihood equations (64) (resp. (66)) have a unique solution η^* (resp. $\hat{\theta}$) has a probability which goes to 1 for any η (resp. θ). Outside this set η^* (resp. $\hat{\theta}$) will be assumed to be defined such that it depends upon $\bar{Z} = (\bar{Z}_{11},\ldots,\bar{Z}_{sr_s})$ only.

C'. $\hat{\eta} = \phi(\hat{\theta}) \to \eta$ in probability if $\eta^{(n)}$ goes to $\eta \in H_0$, where the $\eta^{(n)}$ need not be in H_0.

Remarks. The assumptions in A' secure that the assumptions in A in II·1 are fulfilled. The uniformity assumption in A follows from A' the expression for W_{ijam} given by (70). The permissibility of integrating under the sign of integration in A follows from the assumptions connected with (72) above. B' is a repetition of B of II·1. It will follow from Theorem 3 (iii) that except for what is said in C' the assumptions C_1 and C_2 are not needed. To prove the non-singularity of λ may sometimes cause some difficulties. However Theorem 3 (ii) shows that the non-singularity of λ may follow in the course of checking B'.

Note that the null-state has the same mathematical structure as the a priori state. It is only in the a priori state to replace η by θ, $\tau_{ah}(\cdot)$ by $\tau_{ah}(\phi(\cdot))$ and $\zeta_{ah}(\cdot)$ by $\zeta_{ah}(\phi(\cdot))$. Hence no separate treatment of the null-state is needed. Note also that it follows from (32) and the chain rule for differentiation applied to $\log g_a(X_{am},\phi(\theta))$, that the information matrix in the mull-state is

$$(D\phi)'\,\lambda(D\phi)$$

which is non-singular if and only if λ is non-singular.

To avoid trivial complications we shall modify (63), (64) and (66) by replacing $c_a^{(n)}$ by $c_a = \lim c_a^{(n)}$.

Consider now the likelihood equations (64) and replace $(\bar{Z}_{ak},\eta^*)$ by (ζ_{ak},η). We then obtain

$$G_j(\eta,\zeta) = \sum_{a=1}^{s} c_a[\tau_{ao\,j}(\eta) + \sum_{h=1}^{r_a} \tau_{akj}(\eta)\zeta_{ah}] = 0 \qquad (73)$$

It follows from (67) that these equations have a solution $\eta = A(\zeta)$ for ζ in the range σ of $\zeta(\eta) = (\zeta_{11}(\eta), \ldots)$. Now the Jacobian of the system (73) is

$$\frac{DG}{D\eta} = \left(\sum_{a=1}^{s} c_a[\tau_{ao\,ij}(\eta) + \sum_{k=1}^{r_a} \tau_{akij}(\eta)\zeta_{ak}]\right)_{ij=1,\ldots,v.} \qquad (74)$$

However, by (70) and (43) this matrix equals $-\lambda$.

Hence the solution of (73) is unique if and only if λ is non--singular. We may now take $A(\cdot)$ to be the unique solution of (64) for $\bar{Z}$ on the union $\sigma \cup \bar{Z}(S_n)$ of the ranges of $\zeta(\eta)$ and $\bar{Z}(X)$. $\eta^* = A(\bar{Z})$ will now be consistent by B' and since the var Z_{ak} are continuous functions of $\eta^{(n)}$. Hence they are bounded on compact sets and $\text{plim}(\bar{Z}_{ak} - \zeta_{ak}(\eta^{(n)})) = 0$ by Chebychev's inequality. Thus $\text{plim}\,\bar{Z} = \zeta$ and $\text{plim}\,A(\bar{Z}) = A(\zeta) = \eta$. It is seen from A', B', in particular the assumption about ϕ, that we have an analogous null--state estimate $\hat{\theta} = A_0(\bar{Z})$ of (66) for $\bar{Z} \in \sigma \cup \bar{Z}(S_{no})$. We define $\hat{\eta} = B(\bar{Z}) = \phi(A_0(\bar{Z}))$. We can now sum up what we have said in the following theorem.

Theorem 3. Let the classes of g_a ; $a = 1,\ldots,s$; be Darmois-Koopmann (61). Then we have, (i) Under the assumptions A', B', there is a unique solution $\eta^* = A(\bar{Z})$ of (64) defined for any $\bar{Z}$ on the range of either $\bar{Z}$ or ζ, which depends on the observations X only through $\bar{Z}$, and not on n, and which is Fisher-consistent, i.e. $A(\zeta) = \eta$. (ii) Under the assumptions in A', but without the non-singularity of λ, then λ is non-singular if and only if the solution of the maximum likelihood equations is unique for $\bar{Z}$ in the range of ζ. (iii) The assumptions A', B', C' secure that the assumptions A, B_1 , B_2 , C_1 , C_2 of II·1 are true with consistent estimates in C_1 , C_2 defined by $\eta^* = A(\bar{Z})$, $\hat{\eta} = B(\bar{Z})$.

The crucial assumption B' of Theorem 3 is often easy to verify. Thus in the case of a simple survival model, let $X_1,\ldots,X_n$ be the independent survival periods for n persons (articles) observed for one year. The force of mortality is η. Hence $Pr(X_i=1) = e^{-\eta}$, $Pr(x<X_i < x+dx) = \eta e^{-\eta x}$ if $0 < x < 1$. Thus the density (with respect to an apropriate measure) may be written

$$g_a(x;\eta) = \eta^D e^{-x\eta} \quad ; \quad 0 \leq x \leq 1 \tag{75}$$

where D is 1 or 0 according as $x < 1$ or $x = 1$.
Hence the likelihood equals

$$L(X,\eta) = e^{n(\bar{D}\log\eta - \bar{X}\eta)}$$

where $\bar{D} = \Sigma D_i/n$ is the rate of mortality and $\bar{X}$ is the average survival time. The maximum likelihood equation is

$$\frac{\bar{D}}{\eta^*} - \bar{X} = 0$$

which has the solution $\eta^* = \bar{D}/\bar{X}$ if and only if both $\bar{D}$ and $\bar{X} > 0$. Thus the probability of no solution is

$$\leq Pr(\bar{X}=0)+Pr(\bar{D}=0) = 0+e^{-\eta n} \to 0.$$

Hence B' is fulfilled.

In the fundamental Theorem 4 below we have to make an assumption about the consistency of the likelihood ratio which differs from the result in Theorem 2b. Hence the Lemma 3 is usefull.

<u>Lemma 3</u>. Let the class of g_a ; $a = 1,\ldots,s$ be Darmois-Koopman (61). Make the assumptions of Theorem 3 and assume also that η is identified in the model (i.e. $\tau(\eta) \neq \tau(\eta')$ implies $\eta \neq \eta'$), and that $\operatorname{plim} B(\bar{Z}) \neq \eta$ when $\eta^{(n)} \to \eta \notin H_0$. Then the likelihood ratio is consistent, i.e. $\operatorname{plim} Z_1 = \infty$.

<u>Proof</u>: We have for Z_1

$$Z_1 = 2n \sum_{a=1}^{s} c_a^{(n)}[\tau_{a0}(\eta^*)-\tau_{a0}(\hat{\eta})+\sum_{k=1}^{r_a}(\tau_{ak}(\eta^*)-\tau_{ak}(\hat{\eta}))\bar{Z}_{ak}] = 2n\, R_n$$

where

$$\operatorname{plim} R_n = \sum_{a=1}^{s} c_a[\tau_{a0}(\eta)-\tau_{a0}(B(\zeta))+\sum_{1}^{r_a}\zeta_{ak}(\eta)(\tau_{ak}(\eta)-\tau_{ak}(B(\zeta))] = R$$

since obviously $\operatorname{plim} B(\bar{Z}) = B(\zeta)$ as $\eta^{(n)} \to$ any η, (see Remark before Theorem 3).

Consider now the "true" log-likelihood

$$\log L_0(\eta,\zeta) = n\sum_{a=1}^{s} c_a^{(n)}(\tau_{ao}(\eta)+\sum_k \tau_{ak}\zeta_{ak})+\text{const.}$$

obtained by replacing $\bar{Z}$ by ζ in $\log L(X,\eta)$ (see (63)). By the remark after (74) this expression is maximized by $\eta=A(\zeta)$. Hence $R \geq 0$.

By the assumption about identifiability of η, the true likelihood is not maximized both by $A(\zeta)$ and $B(\zeta)$. Hence $R > 0$ and $Z_1 = 2n\,R_n \to \infty$, i.e. $L(X;\hat{\eta})/L(X;\eta^*) \to 0$, which proves the lemma.

5. <u>The contrast analysis</u>

We return to the general situation in I.3 (see equations (3) and (4)) and recall that $f(\eta)$ is a contrast if $f(\eta) = 0$ for $\eta \in H_0$, i.e. if $f(\phi(\theta)) = 0$ for all θ. We shall also assume a smoothness property, and use the following regularity assumption to be added to assumptions A, B_1, B_2, C_1 and C_2 of II.1.

<u>Assumption D</u>. The class of contrasts $\mathcal{F}$ is such that (<u>i</u>), for any $f \in \mathcal{F}$, $f(\phi(\theta)) = 0$ for all θ; (<u>ii</u>), the first order derivatives $f_j = \frac{\partial f}{\partial \eta_j}$ are equicontinuous for variation of $f \in \mathcal{F}$. Such a class is called regular.

We shall in Lemmas 4 and 5 below consider reduced formulation (see II.1), where $\eta = (\nu_1,\ldots,\nu_w;\theta_1,\ldots,\theta_t) = (\nu,\theta)$ and the null-state is $H_0 : \nu=0$. Then $f(\phi(\theta)) = 0$ in D becomes $f(0,\theta) = 0$.

We shall operate with three types of regular contrasts. (<u>1</u>) <u>General contrast</u> classes, satisfying assumptions D only, (<u>2</u>) <u>Focalized</u> ("målrettet") <u>contrast</u> classes, where all f depend on the interest parameter ν only, $f(\eta) = f(\nu)$, not on the nuisance parameter θ (see II.1), (<u>3</u>) <u>Linear</u> contrasts where each f is a linear function of η, hence $f(\eta) = f_0 + \Sigma f_j\,\nu_j + \Sigma f_{w+j}\,\theta_j$. But since $f(0,\theta) = 0$, we obtain $f(\eta) = \Sigma_{j=1}^{w} f_j\,\nu_j$. Thus any linear contrast is focalized.

Obviously we have for any η^* and η ,

$$f(\eta^*) = f(\eta)+\sum_{j=1}^{w}(\nu_j^*-\nu_j)f_j(\eta')+\sum_{j=1}^{t}(\theta_j^*-\theta_j)f_{w+j}(\eta') \tag{76}$$

where η' is between η^* and η , componentwise.

Now, we have from $f(0,\theta) = 0$, that $f_j(0,\theta) = 0$; $j = w+1,\dots,v$. Hence if η is close to H_0 , then the last sum in (76) is small. We have more precisely,

Lemma 4. Assume A, B_1 , C_1 , D , let $\eta^{(\eta)} \to \eta \in H_0$, and let η^* be the maximum likelihood estimate of η. Then for any general contrast, we have in a reduced formulation, with $\eta = (\nu,\theta)$,

$$f(\eta^*) = f(\eta^{(n)})+\sum_{j=1}^{w}(\nu_j^*-\nu_j^{(n)})f_j(\eta^*)+ A_f \tag{77}$$

where $\text{plim}\ \sqrt{n}A_f = 0$ uniformly in f. (77) is also true if $\eta^{(n)} \to$ any η and f is focalized. If in addition $\nu^{(n)} = \nu+\Delta^{(n)}/\sqrt{n}$, where $\Delta^{(n)}$ converges, then

$$f(\eta^*) = f(\eta)+\sum_{j=1}^{w}(\nu_j^*-\nu_j)f_j(\eta^*)+A_f \tag{78}$$

where still $\text{plim}\ \sqrt{n}A_f = 0$ uniformly in f.

Proof: Comparing (77) and (76) we have

$$\sqrt{n}\ |A_f| < Q\sum_{j=1}^{v}|\sqrt{n}(\eta_j^*-\eta_j^{(n)})| \tag{79}$$

where

$$Q = \sup_{f,i,j}[|f_i(\eta')-f_i(\eta^*)|,|f_{w+j}(\eta')|]$$

The sum in (79) converges in distribution by Theorem 1. If $\lim \eta^{(n)} = (0,\theta)$, then $\text{plim}\ \eta' = (0,\theta)$ and we also have $f_j(0,\theta) = 0$; $j = w+1,\dots,v$. It then follows from the equicontinuity of f_i that to any $\varepsilon > 0$ there exists a δ such that $|\eta_j^{(n)}-\eta_j^*| < \delta$ for all j imply $Q < \varepsilon$. Hence

$$\Pr(Q>\varepsilon) < \sum_{j=1}^{v}\Pr(|\eta_j^*-\eta_j^{(n)}|>\delta)$$

and the first assertion follows. The second assertion follows in a similar, but simpler manner. To prove the assertion connected with (78), we observe that (79) is true with $\eta^{(n)}$ replaced by η in (79), and make use of the last statement in Theorem 1.
From Lemma 4 we get,

Lemma 5. Assume A, B_1, C_1, D. Under the different limit conditions concerning $\eta^{(n)}$ in Lemma 4, $\sqrt{n}(f(\eta^*)-f(\eta^{(n)}))$, resp. $\sqrt{n}(f(\eta^*)-f(\eta))$ converges in distribution to the normal with mean 0, resp. $\Sigma_1^w \Delta_j f_j(\eta)$, and variance.

$$\sigma_f^2(\eta) = \sum_{i,j=1}^{w} f_i(\eta) f_j(\eta) \sigma_{ij}(\eta) \tag{80}$$

where $\eta \in H_0$, and $(\Delta_1,\ldots,\Delta_w)' = \Delta = \lim \Delta^{(n)}$.

Now, it is natural to state that $f(\eta) > 0$ if $f(\eta^*)$ is sufficiently large. With a reduced formulation, we decide to state that $f(\eta) > 0$ whenever

$$f(\eta^*) > \sqrt{z}\, \sigma_f(\eta^*)/\sqrt{n} = \sqrt{z}\, \sigma_f^*/\sqrt{n} \tag{81}$$

where z is the $(1-\varepsilon)$-fractile of the chi-square distribution with $w = v-t$ degrees of freedom.

With a general formulation we decide to state that $f(\eta) > 0$, when

$$f(\eta^*) > \sqrt{z}\, \rho_f(\eta^*)/\sqrt{n} \tag{82}$$

where

$$\rho_f^2(\eta) = \sum_{i,j=1}^{v} f_i(\eta) f_j(\eta) \sigma_{ij}(\eta) \tag{83}$$

Alternatively, instead of using a priori estimated standard deviations we might use null-state estimated standard deviations $\hat{\sigma}_f = \sigma_f(\hat{\eta})$, $\hat{\rho}_f = \rho_f(\hat{\eta})$.

These are our delta-multiple comparison tests.

We shall now state and prove what we consider the fundamental properties of this procedure.

It seems as if focalized contrasts will commonly occur in practical applications. We shall, however, also consider general contrasts.

<u>Theorem 4</u>. The vector of observations X has density $L(x;\eta^{(n)})$, where $L(x;\eta)$ is given by I.3.(3). We make the assumptions A, B_1, B_2, C_1, C_2 of II.1 and D of II.5. A, B_1, B_2, C_1, C_2 may be replaced by assumptions A', B', C' in the case when the classes of g_a; $a = 1,\dots,s$; are Darmois-Koopman (see (61) and Theorem 3). See also Lemma 3 for the consistency of the likelihood ratio assumed in (ii) and (iv) below.

Let $\eta^{(n)} \to \eta$.

Then the delta-multiple comparison tests (81) and (82) with a priori estimated variances, have the following properties (i)-(v).

(<u>i</u>) Let $\mathcal{F}$ be the linear contrast class for the null-state, consisting of all linear f. Assume that the reduced formulation test (81) is used. Then we state that $f(\eta) > 0$ for at least one f, if and only if

$$Z_0' = n\, \nu^{*\,\prime} \Sigma(\eta^*)^{-1} \nu^* > z \tag{84}$$

(see Theorem 2). If η is consistent with the null state $(\eta \in H_0)$, then the probability of stating a significant contrast (falsely) converges to ε.

(<u>ii</u>) Let $\mathcal{F}$ be as in (i). Then the probability of having a significant likelihood ratio $(Z_1 > z)$ without having a significant contrast, goes to 0. Vice-versa if the likelihood ratio is consistent (see Lemma 3) the probability of having a significant contrast without having a significant likelihood ratio $(Z_1 < z)$ goes to 0.

(<u>iii</u>) Let $\mathcal{F}$ be any regular contrast class and assume $\eta^{(n)} \to \eta \in H_0$; alternatively $\mathcal{F}$ may be focalized and η need not be in H_0. Then the probability of falsely stating $f(\eta) > 0$ for some $f \in \mathcal{F}$ for which $f(\eta) \leq 0$ is asymptotically $\leq \varepsilon$, more precisely,

$$\limsup \Pr\{ \bigcup_{f: f(\eta^{(n)}) < 0} [f(\eta^*) > \sqrt{z}\, \sigma_f(\eta^*)/\sqrt{n}] \} \leq \varepsilon \tag{85}$$

A simultaneous confidence interval for all contrasts in $\mathcal{F}$ follows from

$$\liminf \Pr[\bigcap_{f \in \mathcal{F}} \{\sqrt{n}(f(\eta^*) - f(\eta^{(n)}) < z\, \sigma_f(\eta^*)\}] \geq 1-\varepsilon \tag{86}$$

In (85) and (86), $\sigma_f(\eta)$ is given by (80) or may be replaced by $\rho_f(\eta)$ given by (83).

(iv) Assume that $\mathcal{F}$ is focalized. If η need not be in H_0, assume also that the likelihood ratio is consistent. Then the probability of having a significant contrast without having a significant likelihood ratio goes to 0.

Remark: We might say that significant likelihood ratio is "a necessary condition in probability limit" for having a significant contrast.

(v) Assume $\mathcal{F}$ to be the class of all linear contrasts for null-state and let $\Delta^{(n)} = \sqrt{n}(\eta^{(n)} - \phi(\theta)) \to \Delta$. Then the probability of stating a significant contrast approaches $\Pr(Z_w(\kappa) > z)$, where $Z_w(\kappa)$ is chi-square distributed with eccentricity given by (40) in Theorem 2. If Δ is defined by the orthogonal speed of descent to H_0 (see (30) and (31)) then κ is also given by (41).

Remark: It should be noted that this result could be used to study the performance of the multiple comparison method with respect to certain comparisons in a subset $\mathcal{F}' \subset \mathcal{F}$, which corresponds to a less restrictive "hypothesis" $H_0' \supset H_0$.

(vi) The statements in (i)-(v) still hold if null-state estimated standard deviations $\sigma_f(\hat{\eta})$ and $\rho_f(\hat{\eta})$ are used in place of $\sigma_f(\eta^*)$, $\rho_f(\eta^*)$, provided $\Sigma(\eta^*)$ is replaced by $\Sigma(\hat{\eta})$ in (84) and it is everywhere assumed that $\eta^{(n)} \to \eta \in H_0$.

Proof. We consider first the case of reduced formulation.

Proof of (i). We use Schwartz inequality (2). Then with $h = (f_1,\ldots,f_w)'$,

$$\max_h h'\nu^* / \tfrac{1}{\sqrt{n}}\, \sigma_f(\eta^*) = \max_h h'\nu^* / \tfrac{1}{\sqrt{n}} \sqrt{h'\Sigma(\eta^*)h} = \sqrt{n}\sqrt{\nu^{*\prime}\Sigma(\eta^*)^{-1}\nu^*} = \sqrt{Z_0'}$$

by (84) and (80). Hence $h'\nu^* > \sqrt{z}\,\sigma_f(\eta^*)/\sqrt{n}$ for some f if and only if $Z_0' > z$.

This proves the first statement in (i). The second statement follow from the fact that if $\eta \in H_0$, then by Theorem 1, Z_0' has a distribution which converges to the central chi-square distribution - with w degrees of freedom.

Proof of (ii). We note that from the first part of (i), that the probability of having a significant contrast without having a significant liklihood ratio and the probability of a significant likelihood ratio without having a significan contrast are respectively

$$\Pr[(Z_0'> z)\cap(Z_1<z)]\ ,\ \Pr[(Z_0'<z)\cap(Z_1> z)] \tag{87}$$

Consider the second probability. The event $(Z_0'<z)\cap(Z_1>z)$ may (trivially), occur either when also $Y = Z_1-Z_0'>\varepsilon$ or when also $Y < \varepsilon$, $\varepsilon > 0$. The second event implies $z-\varepsilon < Z_0' < z$. Hence the last probability in (87) is

$$< P(Y>\varepsilon) + \Pr(z-\varepsilon< Z_0'<z)$$

where the first term $P(Y>\varepsilon)$ goes to 0 by Theorem 2a if $\sqrt{n}\,\nu^{(n)} \to 0$. The second term goes to $\Gamma(z)-\Gamma(z-\varepsilon)$, by Theorem 2a, where Γ is a cumulative chi-square distribution. Letting $\varepsilon \to 0$ we obtain that the second term in (87) goes to 0. If $\nu^{(n)} \to \nu\neq 0$ or $\nu^{(n)} \to 0$ more slowly than $1/\sqrt{n}$, then we make use of the trivial fact that the second term in (87) $<\Pr(Z_0'<z)$, which goes to 0 by Theorem 2b.

That the first probability in (87) goes to 0 is proved in a similar manner, making use of the assumed consistency of Z_1.

Proof of (iii). We denote the union in (85) by $S(\eta^{(n)})$ and have by (77), writing $\sigma_f(\eta^*) = \sigma_f^*$,

$$S(\eta^{(n)}) = \bigcup_{f:f(\eta^{(n)})\leq 0} [\, f(\eta^{(n)})+\sum_{j=1}^{w} (\nu_j^{*}-\nu_j^{(n)})f_j(\eta^{*})+A_f \geq \sqrt{z}\ \sigma_f^{*}/\sqrt{n}] \tag{88}$$

Since $f(\eta^{(n)}) \leq 0$ this union is a subset of

$$\bigcup_{f:f(\eta^{(n)})\leq 0} [\, \sum_{j=1}^{w} (\nu_j^{*}-\nu_j^{(n)})f_j(\eta^{*})+A_f \geq \sqrt{z}\ \sigma_f^{*}/\sqrt{n}]$$

If we take the union over all f, we get a set which is at least as wide, hence

$$S(\eta^{(n)}) \subset \bigcup_{f\in} [\, \sum_{j=1}^{w} (\nu_j^{*}-\nu_j^{(n)})f_j(\eta^{*})+A_f \geq \sqrt{z}\ \sigma_f^{*}/\sqrt{n}] \tag{89}$$

The first term in the bracket in (89) is by Schwartz inequality (2)

$$\leq \sigma_f^{*}\sqrt{Z(\nu^{(n)})}/\sqrt{n} \tag{90}$$

where $Z(\nu) = n(\nu^{*}-\nu)'\Sigma(\eta^{*})^{-1}(\nu^{*}-\nu)$. Combining (89) and (90) we get

$$S(\eta^{(n)}) \subset \bigcup_{f} [\sqrt{Z(\nu^{(n)})} + B_f \geq \sqrt{z}] = T(\eta^{(n)}) \tag{91}$$

where

$$B_f = \sqrt{n}\ A_f/\sigma_f^{*} \to 0 \tag{92}$$

in probability, uniformly in f by Lemma 4 and (80).

Let $0 < a < \sqrt{z}$. Then $T(\eta)-[\sqrt{Z(\nu)}\geq\sqrt{z}-a] = T(\eta)\cap(\sqrt{Z(\nu)}<\sqrt{z}-a)$ implies $B_f > a$ for some f. Hence

$$S(\eta^{(n)}) \subset T(\eta^{(n)}) \subset (\sqrt{Z}(\nu^{(n)})\geq\sqrt{z}-a)\cup \bigcup_{f} (B_f > a) \tag{93}$$

and

$$\limsup \Pr(S(\eta^{(n)})) \leq 1-\Gamma((\sqrt{z}-a)^2)+ \lim \Pr(\bigcup_{f}(B_f > a)) \tag{94}$$

where Γ is a chi-square distribution by Theorem 1.
However, by the uniform convergence of B_f to 0, the last term is 0. Hence by letting $a \to 0$ we get ε on the right side of (94) and the first statement in (iii) is proved.

To prove the second statement in (iii) we consider the compliment of the intersection in (86), which by (77) is identical with the right hand side of (89). Hence we may proceed from (89) to (94), from which (86) follows.

Proof of (iv). We shall prove that

$$\Pr\{[\bigcup_f (f(\eta^*)>\sqrt{z}\sigma_f)] \cap (Z_1<z)\} \tag{95}$$

goes to 0. This probability is $\leq \Pr(Z_1<z)$ which goes to 0 by the consistency assumption if $\nu^{(n)} \to \nu \neq 0$, or by Theorem 2b if $\nu^{(n)} \to 0$ more slowly than $1/\sqrt{n}$. (Hence we need not the assumption that $\mathcal{F}$ is focalized in these cases.) Assume now that $\mathcal{F}$ is focalized and $\nu^{(n)} = \frac{1}{\sqrt{n}} \Delta^{(n)}$, where $\Delta^{(n)} \to \Delta$. Then we make use of (78) to obtain that (95) may be written,

$$\Pr\{[\bigcup_f (\sqrt{n} \sum_{j=1}^{w} \nu_j^* f_j(\eta^*)+\sqrt{n}\, A_f>\sqrt{z}\, \sigma_f^*)] \cap (Z_1<z)\} \tag{96}$$

Proceeding as in the development from (89) to (92) we obtain as in (94), that (95) is

$$\leq \Pr\{\bigcup_f[\sqrt{Z_0'}+B_f>\sqrt{z}] \cap (Z_1<z)\}$$

By (91) and (93) this probability is

$$\leq \Pr[(\sqrt{Z_0'} > \sqrt{z}-a)\cap(Z_1<z)] + \Pr(\bigcup_f (B_f > a))$$

where the first term goes to $\Gamma_w(z;\kappa) - \Gamma_w((\sqrt{z}-a)^2;\kappa)$ (Theorem 2) and the last term goes to 0 by (92). Hence (iv) follows.

Now (v) follows trivially from (i).

To prove (vi) we note first that all the results in Theorem 2 are true if $\sigma(\eta^*)$ and $\Sigma(\eta^*)$ are replaced by $\sigma(\hat{\eta})$ and $\Sigma(\hat{\eta})$ and $\eta^{(n)} \to \eta \in H_0$.

Going through the proof above we notice that the derivation following (87) still holds good with $\sigma_f(\eta^*)$ and $\Sigma(\eta^*)$ replaced by $\sigma_f(\hat{\eta})$ and $\Sigma(\hat{\eta})$. Furthermore $B_f=\sqrt{n}A_f/\sigma_f(\hat{\eta}) \to 0$ uniformly in f (see (92)) since we have now assumed that $\eta^{(n)} \to \eta \in H_0$ and can use the assumption C_2. Also Z_0' and $Z(\nu^{(n)})$ with $\Sigma(\eta^*)$ replaced by $\Sigma(\hat{\eta})$ still have the convergence in distribution properties used in connection with (94) and the derivation after (96). This proves (vi).

We now turn to the proof of the results in Theorem 4 in the case of <u>a general formulation</u>. We shall first prove that $\rho_f^2(\eta)$ is invariant under smooth transformations $\eta = T(\tilde{\eta})$. We then write (83)

$$\rho_f^2(\eta) = Df(\eta)\sigma(\eta)\ Df(\eta)' \tag{97}$$

using the same principle of notations as in the proof of Theorem 2. From the invariance consideration after equation (32) we obtained $\tilde{\lambda}(\tilde{\eta}) = DT(\tilde{\eta})'\lambda(\eta)DT(\tilde{\eta})$. Now $\tilde{f}(\tilde{\eta}) = f(T(\tilde{\eta}))$. By the chain rule for differentiation we also have $D\tilde{f}(\tilde{\eta})' = DT(\tilde{\eta})'Df(\eta)'$. Thus we get for $\tilde{\rho}_f^2(\tilde{\eta}) = D\tilde{f}(\tilde{\eta})'\ \tilde{\lambda}(\tilde{\eta})^{-1}D\tilde{f}(\tilde{\eta})$ that

$$\begin{aligned}\tilde{\rho}_f^2(\tilde{\eta}) &= Df(\eta)\ DT(\tilde{\eta})DT(\tilde{\eta})^{-1}\ \lambda(\eta)^{-1}(DT(\tilde{\eta})')^{-1}\ DT(\tilde{\eta})'Df(\eta)' = \\ &= Df(\eta)\lambda(\eta)^{-1}\ Df(\eta)' = \rho_f^2(\eta)\end{aligned}$$

Hence $\rho_f^2(\eta)$ is invariant.

Let now $\tilde{\eta}$ be the parametre in a reduced formulation. Then (80) should be written

$$\sigma_f^2(\tilde{\eta}) = \sum_{i,j}^{w} \tilde{f}_i(\tilde{\eta})\tilde{f}_j(\tilde{\eta})\tilde{\sigma}_{ij}(\tilde{\eta}) \tag{98}$$

On the other hand, if either $\eta \in H_0$ or the $\tilde{f} \in \tilde{\mathcal{F}}$ are focalized (in the reduced formulation), then

$$\rho_f^2(\eta) = \tilde{\rho}_f^2(\tilde{\eta}) = \sigma_f^2(\tilde{\eta})$$

since $\tilde{f}_i(\tilde{\eta}) = 0$ for $i > w$. From the equicontinuity of the f_i it follows that

$$\lim[\rho_f^2(\eta^{(n)}) - \sigma_f^2(\tilde{\eta}^{(n)})] = 0 \tag{99}$$

uniformly for $\tilde{f} \in \tilde{\mathcal{F}}$. Since this is true for any sequence $\eta^{(n)} \to \eta$ and $\text{plim}\ \eta^* = \eta$ according to assumption C_1, we also have

$$\text{plim}(\rho_f^2(\eta^*)-\sigma_f^{*2}) = 0 \tag{100}$$

uniformly in f.

Now consider (iii) in the theorem. We observe that by the invariance of $\rho_f^2(\eta)$ and the obvious invariance of $\tilde{f}(\tilde{\eta}^*)$, we may let the criterion in (85) $\sqrt{n}f(\eta^*) > \sqrt{z}\,\rho_f(\eta^*)$ refer to the reduced formulation. This may now be written,

$$f(\eta^*)+\sqrt{z}(\sigma_f^*-\rho_f^*)/\sqrt{n} > \sqrt{z}\,\sigma_f^*/\sqrt{n}$$

where $\rho_f^* = \rho_f(\eta^*)$. Hence in the proof of (iii) we may just replace A_f by $\bar{A}_f = \sqrt{z}(\sigma_f^*-\rho_f^*)/\sqrt{n} + A_f$ where by (100) $\sqrt{n}\,\bar{A}_f$ goes to 0 in probability uniformly for f in $\mathcal{F}$. Hence the proof goes through as before with A_f replaced by $\bar{A}_f$. The same argument applies to the proof of (iv). The other results in the theorem now follow from the limit properties of Z_0', Z_0, Z_1, given in Theorem 2.

For some applications of the theory of the present Chapter II the following lemma is convenient.

Lemma 6. We make the assumptions A, B_1, B_2, C_1, C_2, of II.1, and consider reduced formulation $\eta = (\nu,\theta)$. Let $F(\eta)$ be any function of η with continuous first order derivatives. η^* and $\hat{\eta} = (0,\hat{\theta})$ are maximum likelihood estimates a priori and in the null-state respectively. If $\eta^{(n)} = (\nu^{(n)},\theta^{(n)}) \to \eta = (\nu,\theta)$ there exist sequences $\{Q_{1n}(\eta)\},\ldots,\{Q_{wn}(\eta)\}$ converging in probability, such that

$$\sqrt{n}\,[F(\eta^*)-F(\hat{\eta})] - \sqrt{n}\sum_{j=1}^{w} Q_{jn}(\eta)\nu_j^{(n)}$$

converges in distribution to the normal with mean 0.

Proof. We have for some η' between η^* and $\hat{\eta}$,

$$\sqrt{n}[F(\overset{*}{\eta})-F(\hat{\eta})] = \sqrt{n}\sum_{i=1}^{v} F_i(\eta')(\eta_i^*-\hat{\eta}_i) =$$

$$= \sqrt{n}\sum_{1}^{w} F_i(\eta')\nu_i^* + \sqrt{n}\sum_{i=1}^{t} F_{w+i}(\eta')(\theta_i^*-\hat{\theta}_i)$$

where $F_i(\eta) = \dfrac{\partial F(\eta)}{\partial \eta_i}$. Introducing (53) from II.2, and writing $\nu_i^* = (\nu_i^*-\nu_i^{(n)})+\nu_i^{(n)}$ we obtain the lemma from Theorem 1.

III. APPLICATION OF THE GENERAL THEORY TO STATISTICAL ANALYSIS OF CATEGORICAL OBSERVATIONS

1. Assumptions

The situation is as described in I.6-7. There are s independent sequencies of multinomial trials leading to the likelihood (17)

$$L(X;\eta) = \prod_{a=1}^{s} \prod_{j=1}^{r_a} \pi_{aj}^{N_{aj}}$$

where the $\pi_{aj} = \pi_{aj}(\eta)$; $j=1,2,\dots,r_a$; are the probabilities of the r_a outcomes A_{aj}; $j=1,2,\dots,r_a$; $a=1,2,\dots,s$. The N_{aj} are the frequencies of the A_{aj}; $j=1,2,\dots,r_a$; $a=1,2,\dots,s$. We aim at constructing a multiple comparison method relatively to a null-state $\eta_i = \phi_i(\theta_1,\dots,\theta_t)$; $i=1,2,\dots,v$.

We still consider asymptotic results under increasing $n_a = \Sigma_{j=1}^{r_a} N_{aj}$ such that $\lim n_a/n = c_a > 0$ where $n = \Sigma n_a$. The likelihood is $L(X,\eta^{(n)})$ where $\eta^{(n)}$ converges to some η as described in II.1.

Obviously the present situation falls under the general set-up of the Darmois-Koopman classes of distributions described in II.4. The regularity conditions A', B', C' now simplify to

<u>A"</u>. η varies in an open set of the v-space. In the null--state $\eta = \phi(\theta)$, where θ varies in an open set in the t-space. The second order derivatives of the functions $\pi_{aj}(\eta)$ and $\phi(\theta)$ exist and are continuous. $0<\pi_{aj}(\eta) < 1$ for any η. The Jacobian $D\phi$ has rank t. The matrix $\lambda = \Sigma c_a \lambda_a$, where

$$\lambda_a = - \left(\sum_{j=1}^{r_a} \pi_{aj} \frac{\partial^2 \log \pi_{aj}}{\partial \eta_k \, \partial \eta_\ell} \right)_{k,\ell=1,\dots,v} \tag{101}$$

is non singular.

<u>Remark</u>. By Theorem 3 (ii) of II.4 a necessary and sufficient condition for λ to be nonsingular is that the equations

$$\sum_{a=1}^{s} c_a \sum_{j=1}^{r_a} \pi_{aj} \frac{\partial \log \pi_{aj}(\eta)}{\partial \eta_k} = 0; \; k=1,\dots,v \tag{102}$$

(compare (103)) have a unique solution in η.

B" The set of all X for which the maximum likelihood equations a priori

$$\sum_{a=1}^{s} \sum_{j=1}^{r_a} N_{aj} \frac{\partial \log \pi_{aj}(\eta^*)}{\partial \eta_k} = 0; \; k=1,2,\dots,v \tag{103}$$

have a unique solution η^*, has a probability that goes to 1 as $n \to \infty$. The same is assumed about the maximum likelihood equations in the null-state obtained from (103) by replacing η^* by $\hat{\eta} = \phi(\hat{\theta})$, $\partial\eta_k$ by $\partial\theta_k$ and letting $k = 1,\dots,t$. Outside the sets $\eta^* = A(q)$, $\hat{\eta} = B(q)$ are defined in any manner depending upon the $q_{aj} = N_{aj}/n_a$ alone.

The assumption B" is obviously the critical one.

B" is fulfilled for the framework (saturated) model. Because then the maximumlikelihood equations have a unique solution

$$\pi_{aj}(\eta^*) = N_{aj}/n_a \; ; \; n_a = \sum_j N_{aj} \; ; \; a=1,2,\dots,s$$

provided all $N_{aj} > 0$. But the probability that some $N_{aj} = 0$, is

$$< \sum_{a,j} \Pr(N_{aj}=0) = \sum (1-\pi_{aj})^{n_a} \to 0$$

Consider another example, viz. the restrictive loglinear threeway classification model

$$\log \pi_{ijk} = \mu+\gamma_{ij}+\delta_{jk}+\phi_{ki} \tag{104}$$

where the π_{ijk} now denote the probabilities of the factor combinations (i,j,k); $i=1,\dots,I$; $j=1,2,\dots,J$; $k=1,2,\dots,K$. For the sake of identification we set $\gamma_{i1} = \delta_{j1} = \phi_{k1} = 0$. From $\Sigma\pi_{ijk} = 1$ we get

$$e^{\mu} = 1/\sum_{i,j,k} e^{\gamma_{ij}+\delta_{jk}+\phi_{ki}} \tag{105}$$

The likelihood $L(X;\eta)$ is given by

$\log L = n\mu+\Sigma\gamma_{ij}N_{ij+}+\Sigma\delta_{jk}N_{+jk}+\Sigma\phi_{ki}N_{i+k}$ where

$\eta = (\gamma_{12},\dots,\gamma_{IJ};\ \delta_{12},\dots,\delta_{JK};\ \phi_{12},\dots,\phi_{KI})$, N_{ijk} is the number of trials with factor combination (i,j,k), and $N_{i+k} = \Sigma_{j=1}^{J} N_{ijk}$, etc.

In order to verify B" we study maximization in the case when all N_{ij+}, N_{i+k}, $N_{+jk} > 0$, the probability of which goes to 1, by the same reasoning as above. We shall show that L is maximized for precisely one finite $\eta = (\gamma_{12}, \dots)$. Then obviously the likelihood equations have just one solution.

We choose to use the Lagrange device of first maximizing

$$Q(\theta) = \log L - n \Sigma \pi_{ijk}$$

under free variation of $\theta = (\mu, \eta) = (\mu, \gamma_{12}, \dots)$ without the side condition $\Sigma \pi_{ijk} = 1$. It is seen from $Q(\theta^0 + t\theta^1)$; $t(\text{scalar}) \to \pm\infty$; $\theta^1(\text{vector}) \neq 0$; that $Q \to -\infty$ as μ, γ_{ij}, ϕ_{ki} go to $\pm\infty$ in all directions. Hence Q is maximized locally for at least one θ. We also have

$$\frac{d^2}{dt^2} Q(\theta^0 + t\theta^1) = - n \Sigma (\mu^1 + \gamma^1_{ij} + \delta^1_{jk} + \phi^1_{ki})^2 \pi_{ijk}(\theta^0 + t\theta^1)$$

which is < 0 unless $\mu^1 + \gamma^1_{ij} + \delta^1_{jk} + \phi^1_{ki} = 0$ for all i,j,k. However, this reduces to $\mu^1 = 0$ for $i = j = k = 1$; to $\delta^1_{1k} = 0$ for $i = j = 1$ and to $\phi^1_{ki} = 0$ for $j = 1$. Similarly $\gamma^1_{ij} = \delta^1_{jk} = 0$. Hence $\theta^1 = 0$, which is a contradiction. It follows that there is just one local maximum which must be the absolute maximum, and that there are no other stationary points. For this maximum $\pi^*_{ijk} = \pi_{ijk}(\mu^*, \eta^*)$, we then have for any point (π_{ijk}) given by (104)

$$\sum N_{ijk} \log \pi_{ijk} - N \sum \pi_{ijk} < \sum N_{ijk} \log \pi^*_{ijk} - N \sum \pi^*_{ijk} \qquad (106)$$

On the other hand the θ^* must satisfy $\frac{\partial}{\partial \theta_i} Q(\theta^*) = 0$, which lead to

$$n \sum_{k=1}^{K} \pi^*_{ijk} = N_{ij+}, \quad n \sum_{j=1}^{J} \pi^*_{ijk} = N_{i+k}, \quad n \sum_{i=1}^{I} \pi^*_{ijk} = N_{+jk},$$
$$\sum_{i,j,k} \pi^*_{ijk} = 1 \qquad (107)$$

Introducing the last equation (107) into (106) we have for any point (π_{ijk}) for which $\Sigma \pi_{ijk} = 1$, that

$$\sum N_{ijk} \log \pi_{ijk} < \sum N_{ijk} \log \pi^*_{ijk}$$

Thus (107) gives us the unique maximum likelihood estimates of the π_{ijk} if all $N_{ijk} > 0$. Then everything is proved. It now follows that λ is nonsingular. [For a convenient iterative procedure for the solution of the linear equations (107), see Bishop et.al. (1975).]

2. The likelihood ratio and the chi-square goodness of fit statistics

The following statistics are well known to be useful measures of goodness of fit of a nullhypothesis $\nu = 0$.

The likelihood ratio is given by

$$Q = \frac{L(X;\hat{\eta})}{L(X;\eta^*)} = \prod_{a,j} \left(\frac{\hat{\pi}_{aj}}{\pi^*_{aj}}\right)^{N_{aj}} \tag{108}$$

where $\hat{\pi}_{aj} = \pi_{aj}(\hat{\eta})$ and $\pi^*_{aj} = \pi_{aj}(\eta^*)$.

The chi-square goodness of fit difference is

$$Z_H - Z_a \tag{109}$$

where

$$Z_H = \sum_{a,j} \frac{(N_{aj}-n_a\hat{\pi}_{aj})^2}{N_{aj}}, \quad Z_a = \sum \frac{(N_{aj}-n_a\pi^*_{aj})^2}{N_{aj}} \tag{110}$$

In the case of the framework model, we have $Z_a = 0$, since then $\pi^*_{aj} = q_{aj} = N_{aj}/n_a$.

We shall study these statistics.

For convenience we make the explicite assumption that there exist functions

$$p_{aj} = \Pi_{aj}(\rho_1,\ldots,\rho_r, \nu_1,\ldots,\nu_w, \theta_1,\ldots,\theta_t) = \Pi_{aj}(\rho,\nu,\theta), \quad (111)$$

with second order continuous derivatives establishing a one-to-one correspondence between freely varying multinomial $p_{aj} > 0$, $(\Sigma_{j=1}^{r_a} p_{aj} = 1)$, and $\mu = (\rho,\nu,\theta)$, such that $\Pi_{aj}(0,\nu,\theta) = \pi_{aj}(\nu,\theta)$. Furthermore $r+w+t = \Sigma r_a - s = R-s$.

<u>Theorem 5</u>. We also make the assumptions A", B". Z_0 and Z_0' are defined as in Theorem 2 with $\lambda = \Sigma c_a\lambda_a$ (see (101)) and $\tilde{\Sigma}(\tilde{\eta})$ refers to a reduced formulation.

<u>a</u> If $\eta^{(n)} \to H_0$ at least as fast as $1/\sqrt{n}$ goes to 0, then $Z_1 = -2\log Q$, Z_0', Z_0 and $Z_H - Z_a$ are mutually equivalent in probability limit, and they converge in distribution to the chi-square distribution with $w = v-t$ degrees of freedom and eccentricity $\kappa = \Delta^{I'}(\tilde{\Sigma}(\tilde{\eta})^{-1})\Delta^{I}$, where Δ^I is given by $\Delta' = (\Delta^{I'},\Delta^{II'})$.

In particular, if $\eta^{(n)} \to H_0$ with orthogonal speed at least as fast as $1/\sqrt{n}$ goes to 0, i.e. $(D\phi)'\lambda\Delta = 0$ (see (31)), then the eccentricity is also given by $\kappa = \Delta'\sigma(\eta)^{-1}\Delta$ (see (41)).

<u>b</u> If $\eta^{(n)} \to \eta \in H_0$ more slowly than $1/\sqrt{n}$, then $\Pr(Z_1<z)$, $\Pr(Z_0'<z)$, $\Pr(Z_0<z) \to 0$ and the asymptotic power is 1 when using the test statistics Z_1 or Z_0'.

<u>Remark</u>. It also follows from the proof that Z_a and Z_H converge in distribution to the chi-square distribution with R-s-t and R-s-v degrees of freedom respectively. Z_a has eccentricity 0 in the limit.

<u>Proof</u>. The statements in <u>a</u> not involving $Z_H - Z_a$, and the statement in <u>b</u> are direct consequences of Theorem 2. We prove first that $-2\log Q$ and $Z_H - Z_a$ are equivalent in probability limit. We write

$$- 2\log Q = - 2\log Q_H + 2\log Q_a \tag{112}$$

where

$$Q_H = \prod \left[\frac{\hat{\pi}_{aj}}{q_{aj}}\right]^{N_{aj}} , \quad Q_a = \prod \left[\frac{\pi^{*}_{aj}}{q_{aj}}\right]^{N_{aj}} \tag{113}$$

We see that Q_H is the likelihood ratio in the case of the framework model, i.e. where the a priori model is (111). We have

$$- 2\log Q_H = - 2 \Sigma N_{aj} \log(1+R_{aj}) \tag{114}$$

where

$$R_{aj} = \frac{\hat{\pi}_{aj}-q_{aj}}{q_{aj}} \tag{115}$$

We make use of $\log(1+x) = x-\frac{1}{2}x^2K(x)$ where $\lim_{x\to 0} K(x) = 1$, and get

$$\begin{aligned} - 2\log Q_{aj} &= - 2 \Sigma N_{aj} \frac{\hat{\pi}_{aj}-q_{aj}}{q_{aj}} \\ &+ \Sigma N_{aj}\left(\frac{\hat{\pi}_{aj}-q_{aj}}{q_{aj}}\right)^2 K(R_{aj}) \\ &= Z_H + \Sigma \frac{(n_a\hat{\pi}_{aj}-n_a q_{aj})^2}{N_{aj}}(K(q_{aj})-1) \end{aligned} \tag{116}$$

where

$$K(q_{aj}) \to 1$$

in probability.

But by Lemma 6, there exist sequences $\{(Q_{1n}(\ ),\ldots,Q_{nn}(\ )\}$ converging in probability such that

$$\sqrt{n}(q_{aj}-\hat{\pi}_{aj}) - \sqrt{n}\sum_{k=1}^{w} Q_{kn}(\eta)\nu_k^{(n)}$$

converges in distribution. Hence the same is true of $\sqrt{n}(q_{aj}-\hat{\pi}_{aj})$ if

$$\nu_j^{(n)} = \Delta_j^{(n)}/\sqrt{n}$$

Hence by (116)

$$\text{plim}(-2\log Q_H-Z_H) = 0 \tag{117}$$

On the other hand Q_a is the likelihood ratio if the framework model is considered as the a priori model and the real a priori model is considered as the hypothesis. Hence by (117),

$$\text{plim}(-2\log Q_a-Z_a) = 0 \tag{118}$$

Combining (112), (117), (118) we get

$$\text{plim}(-2\log Q-(Z_H-Z_a)) = 0 \tag{119}$$

It now follows from Theorem 2 that $Z_H - Z_a$ converges in distribution as stated in Theorem 5.

3. Contrast analysis of categorical observations

We can now apply the results from the general theory of Chapter II directly to the present situation defined by (17) and the assumptions A", B", C' (see II.4).

It sometimes makes the contrasts more meaningful to express them in terms of the π_{aj} directly instead of η. Hence we use the form

$$F(\pi(\eta)) = F(\pi_{11}(\eta),\ldots,\pi_{sr_s}(\eta)) \tag{120}$$

and apply the general theory to $f(\eta) = F(\pi(\eta))$. We consider a class $\mathcal{F}_\pi$ of functions. In place of assumption D of II.5. we now assume that

D". The class of contrasts $\mathcal{F}_\pi$ is such that (i), for all $F \in \mathcal{F}_\pi$, $F(\pi(\phi(\theta))) = 0$ for all θ, (ii) the first order derivatives $F_{ai} = \frac{\partial F}{\partial \pi_{ai}}$ are equicontinuous for variation of $F \in \mathcal{F}_\pi$. Such a class $\mathcal{F}_\pi$ is called regular.

Applying the general theory in II.5 we can now write out the rule for multiple comparison. The asymptotic variance of $\sqrt{n}(F(\pi^*)-F(\pi^{(n)}))$, where $\pi^{(n)} = \pi(\eta^{(n)})$, is

$$\sigma_F^2(\eta) = \sum_{a,a',i,i'} F_{ai}(\pi)F_{a'i'}(\pi) \sum_{k,k'=1}^{v} \frac{\partial\pi_{a'i'}}{\partial\eta_{k'}} \frac{\partial\pi_{ai}}{\partial\eta_k}(\lambda^{-1})_{k,k'} \tag{121}$$

where $\lambda = \Sigma\, c_a\lambda_a$ (see eq. (101)).

<u>We obtain that the delta multiple comparison rule is now to the effect that we decide to state that $F(\pi(\eta)) > 0$ for any $F \in \mathcal{F}_\pi$ for which</u>

$$F(\pi^*) > \sqrt{z}\,\sigma_F(\eta^*)/\sqrt{n} \tag{122}$$

<u>where z is the $(1-\varepsilon)$-fractile of the chi-square distribution with $v-t$ degrees of freedom,</u>

$$\sigma_F^2(\eta) = \underline{F}'DA(A'A)^{-1}A'D\,\underline{F}\,, \tag{123}$$

$$\underline{F} = (F_{11},\ldots,F_{sr_s})',$$

$$A_{ajk} = \sqrt{c_a\pi_{aj}}\,\frac{\partial \log\pi_{aj}(\eta)}{\partial\eta_k}\,,$$

$$A = (A_{ajk})_{aj,k}$$

<u>is a matrix such that aj enumerates the rows and k the colums;</u>

and D is a diagonal matrix with diagonal elements

$$D_{aj} = \sqrt{\frac{\pi_{aj}}{c_a}}$$

(Thus A and D are of order $R\times R$; $R = \Sigma\, r_a$. It is understood that the subscripts aj of F_{aj}, A_{ajk} and D_{aj} in F, A and D are taken in, say, lexical ordering.)
In the case of the framework model the test (122) takes the form

$$F(q) > \sqrt{z \sum_a \frac{1}{n_a}[\Sigma F^2_{aj}(q)q_{aj} - (\Sigma F_{aj}(q)q_{aj})^2]} \tag{124}$$

where $q_{aj} = N_{aj}/n_a$, and the degrees of freedom for z is $R-s-t$.
Making use of Theorem 4 and Theorem 5 we can now state

Theorem 6. We have a multinomial situation with likelihood (17) and make the assumptions A", B", C' (see II.4) and D (see II.5). Then the results about the multiple comparison rules (81), (82) and the test statistics Z_0, Z_0', $Z_1 = -2\log Q$ (see (108)) in Theorem 4 (i)-(vi) still hold. These results also hold under assumptions A", B", C', D" if the class of contrasts are defined by (120). It is only in Theorem 4 (i)-(vi) to replace (85) and (86) by

$$\limsup \Pr\{\bigcap_{F:F(\pi^{(n)})<0} [F(\eta^*) > \sqrt{z}\,\sigma_F(\eta^*)]\} < \varepsilon \tag{125}$$

$$\liminf \Pr\{\bigcap_{F\in_\pi} [\sqrt{n}(F(\pi^*) - F(\pi^{(n)}) < \sqrt{z}\,\sigma_F(\eta^*)]\} > 1-\varepsilon \tag{126}$$

where σ_F is given by (123).

The test statistic $Z_H - Z_a$ (see (111)) may replace the statistic Z_1 under the assumptions above and also assuming that $\eta^{(n)} \to H_0$ at least as fast as $\frac{1}{\sqrt{n}} \to 0$.

4. Exact relationships in the case of framework models and contrasts linear in the multinomial probabilities

The results in Theorem 4 (i), (ii), and (v) are of little interest in the case of multinomial situations with restrictive models and contrasts given by (120), since it would assume that $F(\pi(\eta))$ is linear in η. However, in the case of framework model, linearity in η means that $F(\pi)$ is linear in π, which holds in some interesting situations (e.g. homogeneity testing). In that case some rather strong results can be proved.
(See Sverdrup (1975) and (1977 b).)

Theorem 7. Assume the framework model and that $\mathscr{F}$ consists of all linear contrasts

$$F(\pi) = \sum_{a,j} F_{aj}\,\pi_{aj} + F_0 \qquad (127)$$

which are contrasts relatively to the null-state

$$\pi_{aj} = \phi_{aj}(\theta) = \phi_{aj}(\theta_1,\dots,\theta_t) \qquad (128)$$

where $t < R\text{-}s$ and the matrix

$$\left(\frac{\partial\phi_{aj}}{\partial\theta_i}\right)_{aj=(11,\dots,sr_s)}$$

is of full rank. (Note that $F(\pi) = \Sigma_{a,j}(F_{aj} + \frac{1}{s}F_0)\pi_{aj}$, hence $\mathscr{F}$ is the class of $F(\pi) = \Sigma\, F_{aj}\pi_{aj}$, i.e. we may put $F_0 = 0$ in (127).) Let

$$\frac{1}{n}\,\sigma_F^2(\pi) = \sum_a \frac{1}{n_a}\Big[\sum_j F_{aj}^2\,\pi_{aj} - \Big(\sum_j F_{aj}\,\pi_{aj}\Big)^2\Big], \qquad (129)$$

z be the $(1-\varepsilon)$-fractile of the chi-square distribution with $R\text{-}s\text{-}t$ degrees of freedom, $q_{aj} = N_{aj}/n_a$.

Two methods of multiple comparisons will be considered.

Method A. The method with null-state estimated variances. Let $\hat{\pi} = \phi(\hat{\theta})$ where $\hat{\theta}$ be the unique solution of the maximum likelihood equations in the null-state, i.e.

$$\sum_{a,j} \frac{N_{aj}}{\phi_{aj}(\hat{\theta})} \frac{\partial\phi_{aj}(\hat{\theta})}{\partial\theta_i} = 0 \; ; \quad i=1,2,\ldots,t \tag{130}$$

if all N_{aj} are > 0.

Assume all $N_{aj} > 0$ and that the delta multiple comparison test with criterion

$$F(q) > \sqrt{z}\,\sigma_F(\hat{\pi})/\sqrt{n} \tag{131}$$

is used. Then some contrast is declared present if and only if the ordinary goodness of fit statistic satisfies

$$Z = \sum \frac{(N_{aj}-n_a\hat{\pi}_{aj})^2}{n_a\,\hat{\pi}_{aj}} > z \tag{132}$$

Method B. The method with a priori estimated variances. Suppose that $\hat{\pi} = \phi(\hat{\theta})$, where $\hat{\theta}$ is the unique solution of

$$\sum_{a,j} \frac{n_{aj}^2}{N_a} \phi_{aj}(\hat{\theta}) \frac{\partial\phi_{aj}(\hat{\theta})}{\partial\theta_i} = 0 \; ; \quad i=1,2,\ldots,t \; ; \tag{133}$$

for all $N_{aj} \neq 0$, which are the formal minimizing equations for the modified chi-square goodness of fit

$$\sum_{a,j} \frac{(N_{aj}-n_a\phi(\theta))^2}{N_{aj}} \,,$$

and that the criterion

$$F(q) > \sqrt{z}\,\sigma_F(q)/\sqrt{n} \tag{134}$$

is used. Then some contrast is declared present if and only if the modified chi-square goodness of fit statistic satiefies,

$$Z = \sum_{a,j} \frac{(N_{aj}-n_a\hat{\pi}_{aj})^2}{N_{aj}} > z \tag{135}$$

Of course the asymptotic properties of the methods A and B follow from Theorem 4 and Theorem 6.

Remark: Note that the relations between the multiple comparison rules on the one hand side and the classical goodness of fit tests on the other hand side are purely algebraic. They are strictly true, there are no approximations involved and they are not probability statements. (The assumption that all $N_{aj} > 0$ is of course probabilistic, but with a probability that goes to 1 as $n \to \infty$, $n_a/n \to c_a > 0$.)

Proof. We shall first prove the contention about (130). We introduce

$$Y_{aj} = \frac{N_{aj} - n_a \hat{\pi}_{aj}}{\sqrt{n_a \hat{\pi}_{aj}}}$$

For convenience we replace (a,j) by a single letter such that $1,2,\ldots,n$ represent (a,j) in lexical ordering (say). Hence $N_{aj} = N_i$, $q_{aj} = q_i$, $\pi_{aj} = \pi_i$, $\phi_{aj}(\theta) = \phi_i(\theta)$, $F_{aj} = F_i$ (where $i = i(a,j)$). We also replace n_a by n_i and c_a by c_i (if $i = i(a,j)$). Thus c_i and n_i are constants on sections $S_1, S_2, \ldots, S_s$ of length $r_1, r_2, \ldots, r_s$ respectively, $\Sigma_1^s r_a = R$. Thus we may write

$$Y_{aj} = Y_i = \frac{N_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i}} \tag{136}$$

We have from (129)

$$\frac{1}{n}\sigma_F^2(\pi) = \sum \frac{1}{n_a} \sum_{j,j'} F_{aj} F_{aj'} (\delta_{jj'} \pi_{aj} - \pi_{aj} \pi_{aj'})$$

($\delta_{jj'}$ is the Krönecker δ). This may now be written

$$\frac{1}{n}\sigma_F^2(\pi) = \frac{1}{n} \sum_{a=1}^{s} \sum_{i,i' \in S_a} \frac{n}{n_i} F_i F_{i'} (\delta_{ii'} - \pi_{i'}) \pi_i$$

We introduce

$$h_i = F_i\sqrt{n\hat{\pi}_i/n_i} \tag{137}$$

and get

$$\frac{1}{n}\hat{\sigma}^2_F = \frac{1}{n}\sigma^2_F(\hat{\pi}) = \frac{1}{n}\sum_{a=1}^{s}\sum_{i,i'\in S_a}(\delta_{ii'}-\sqrt{\hat{\pi}_i}\sqrt{\hat{\pi}_{i'}})h_i h_{i'} \tag{138}$$

Now let b denote a matrix of order $R\times s$, the a-th column of which is

$$(0,\ldots,0,\sqrt{\hat{\pi}_{R_a+1}},\ldots,\sqrt{\hat{\pi}_{R_{a+1}}},0,\ldots,0) \tag{139}$$

and starts with $R_a = \Sigma_1^{a-1} r_i$ zeros, $R_1 = 0$. Then obviously,

$$b'\,b = I \tag{140}$$

and (138) may now be written,

$$\frac{1}{n}\hat{\sigma}^2_F = \frac{1}{n}h'(I-bb')h \tag{141}$$

We have from (127) and $F(\phi(\theta)) = 0$, that

$$F_0 + \Sigma\, F_i\phi_i(\theta) = 0 \tag{142}$$

Hence

$$F(q) = \Sigma\, F_i(q_i-\hat{\pi}_i)$$

which by (136) and (137) may be written

$$F(q) = \frac{1}{\sqrt{n}}h'\,Y \tag{143}$$

We also have from (142)

$$\sum_{i=1}^{R} F_i\frac{\partial\phi_i(\theta)}{\partial\theta_j} = 0\ ; \quad j=1,2,\ldots,t \tag{144}$$

for any θ. We introduce

$$B_{ij} = \sqrt{n_i/\hat{\pi}_i n}\,\frac{\partial\phi_i(\hat{\theta})}{\partial\theta_j} \tag{145}$$

It is seen that $B = (B_{ij})$ is the matrix $\left(\frac{\partial\phi_i(\hat{\theta})}{\partial\theta_j}\right)$ multiplied by a diagonal non-singular matrix. Hence by the assumption about $\left(\frac{\partial \phi_i(\hat{\theta})}{\partial \theta_j}\right)$, B has rank t.

We have from (144)

$$h' \; B = 0 \tag{146}$$

From $\sum_{i \in S_a} \phi_i(\theta) = 1$ we get

$$\sum_{i \in S_a} \frac{\partial \phi_i(\theta)}{\partial \theta_j} = 0 \;; \quad a=1,\dots,s; \; j=1,\dots,t \tag{147}$$

Replacing θ by $\hat{\theta}$ we obtain

$$b' \; B = 0 \tag{148}$$

Since B has full rank, the space V_t spanned by the columns of B is a t-dimensional subspace of the R-dimensional space V_R. Let H be a $R \times t$ matrix such that its columns constitute an orthogonal basis for V_t . Then of course $H' \; H = I$ and by (146) and (148) h and all columns of b are perpendicular to V_t. We have

$$h' \; H = 0 \tag{149}$$

$$b' \; H = 0 \tag{150}$$

From (150) it is seen that the matrix $R \times (t+s)$-matrix (H,b) has orthonormal columns. We complete it and obtain an orthogonal matrix

$$K = (G,H,b) \tag{151}$$

of order $R \times R$. G is of order $R \times (R-t-s)$.

Let us now introduce

$$d = K' \; h \tag{152}$$

$$V = K' \; Y \tag{153}$$

If we let F run through $\mathcal{F}$ then $(F_1,\dots,F_R)$ varies only subject to (144). Thus $F_1,\dots,F_R$ varies freely in the (R-t)-dimensional subspace. By (144), (146), and (152) the same is true of h and d. We now have from (143), (152), and (153)

$$\sqrt{n}\ F(q) = d'\ V \tag{154}$$

(149) reduces to

$$0 = h'H = d'K'H = (d_{R-s-t+1},\dots,d_{r-s})$$

Hence

$$d_{R-s-t+1} = \dots = d_{R-s} = 0 \tag{155}$$

From (153)

$$V = \begin{pmatrix} G' \\ H' \\ b' \end{pmatrix} Y$$

But by (136) and (139) the a-th component of $b'Y$ is

$$\sum_{i \in S_a} (N_i - n_i \hat{\pi}_i)/\sqrt{n_i} = 0$$

since n_i is constant if $i \in S_a$. Hence

$$V_{R-s+1} = \dots = V_R = 0 \tag{156}$$

We now express our multiple test (131) in terms of our new variables. From (141) we get

$$\hat{\sigma}_f{}^2 = h'h - h'bb'h = d'd - d'K'bb'Kd$$

But

$$K'\ b = \begin{pmatrix} 0 \\ 0 \\ I \end{pmatrix}$$

which combined with (155) gives

$$\hat{\sigma}_F^2 = \sum_1^{R-s-t} d_i^2 \tag{157}$$

Combining this with (155) and (156) we <u>get for the multiple test criterion (131)</u>

$$\sum_1^{R-s-t} d_i\ V_i > \sqrt{z\ \Sigma_{i=1}^{R-s-t}\ d_i^2} \tag{158}$$

Consider now the goodness of fit test (132), i.e. $Z = \Sigma_{i=1}^{R} Y_i^2 > z$. At this point for the first time we make use of the fact that $\hat{\theta}$ is a maximum likelihood estimate of θ in the null-state, i.e. satisfies (130) which may be written

$$\sum \frac{N_i}{\hat{\pi}_i} \frac{\delta \phi_i(\hat{\theta})}{\delta \theta_j} = 0 \; ; \qquad j=1,2,\ldots,t$$

or, since $\Sigma_{i \in S_a} \phi_i(\theta) = 0$,

$$\sum_{i=1}^{R} \frac{N_i - n_i \hat{\pi}_i}{n_i \hat{\pi}_i} \, n_i \, \frac{\partial \phi_i(\hat{\theta})}{\partial \theta_j} = 0 \tag{159}$$

By (136) and (145) we obtain

$$B' \; Y = 0 \tag{160}$$

Hence we also have $H'Y = 0$ and $H'KV = 0$, i.e.

$$V_{R-s-t+1} = \ldots = V_{R-s} = 0 \tag{161}$$

which combined with (157) and (152) gives for the chi-square test (132)

$$Z = \sum_{i=1}^{R-s-t} V_i^2 > z \tag{162}$$

Now, significance according to the multiple comparison criterion is obtained if and only if there exists a d such that (158) is true; i.e.

$$\max_{d} \sum_{1}^{R-s-t} d_i \, V_i / \sqrt{\Sigma_1^{R-s-t} \, d_i^2} > \sqrt{z} \tag{163}$$

It follows from the remark after (153) that the (R-t)-dimensional vector $(d_1, \ldots, d_{R-s-t}, d_{R-s+1}, \ldots, d_R)$ varies freely in the (R-t) -dimensional space. Hence there are now restrictions on the components and $d_1, \ldots, d_{R-s-t}$ varies freely. It follows then from Schwartz inequality (2) that the left hand side of (163) is equal to $\sqrt{\Sigma_1^{R-s-t} V_i^2} = \sqrt{Z}$ and (163) is identical with (162). We have proved the contention about (130) and (131) in the theorem.

As to the proof of contention about the <u>method B</u>, i.e. about the multiple comparison criterion (134) and the chi-square goodness of fit test (135), we first note that the equations (133) for $\hat{\theta}$ may also be written

$$\sum_{a,j} \frac{N_{aj} - n_a \phi_{aj}(\hat{\theta})}{N_{aj}} \, n_a \, \frac{\partial \phi_{aj}(\hat{\theta})}{\partial \hat{\theta}_i} = 0 \; ; \quad i=1,1,\ldots,t \; , \tag{164}$$

since $S_{i\in S_a} \phi_i(\theta) = 1$. In the proof Y_i defined by (136) should be replaced by

$$Y_i = \frac{N_i - n_i\hat{\pi}_i}{\sqrt{N_i}} \tag{165}$$

and h_i should be defined by

$$h_i = F_i\sqrt{q_i n/n_i} \tag{166}$$

instead of by (137). In the definition of b by (139), $\hat{\pi}_i$ should be replaced by q_i, and the definition of B_{ij} by (145) should be replaced by

$$B_{ij} = \sqrt{n_i/nq_i}\,\frac{\partial\phi(\hat{\theta})}{\partial\theta_i} \tag{167}$$

With these changes the proof of the contention about method B in Theorem 2 follows closely the proof we have already gone through.

5. Homogeneity testing

We use the results in III.4 to consider the special case of homogeneity testing (treated by Goodman (1964)). Then $r_1 = \ldots = r_s = r$ and we choose as a null--state that $\pi_{a1},\ldots,\pi_{ar}$ are indępendent of a. This can be written

$$\phi_{aj} = \theta_j;\ j=1,2,\ldots,r-1,\ \phi_{ar} = 1-\theta_1-\ldots-\theta_{r-1} = \theta_r. \tag{168}$$

Hence $\Sigma_{a,j} F_{aj}\,\pi_{aj}$ is a contrast if and only if

$$0 = \sum_{a,j} F_{aj}\,\theta_j = \sum_{j=1}^{r-1}\theta_j\Big(\sum_{a=1}^{s} F_{aj} - \sum_{a=1}^{s} F_{ar}\Big) + \sum_{a=1}^{s} F_{ar}$$

is true for all $(\theta_1,\ldots,\theta_{r-1})$. It follows that a contrast is characterized by

$$\sum_{a=1}^{s} F_{aj} = 0 \tag{169}$$

for all j. Thus <u>comparisons may consist in comparing the probabilities</u> $\pi_{1j},\ldots\pi_{sj}$ <u>of</u> A_j <u>for any fixed</u> j, i.e. in studying the dependence of $\Pr(A_j)$ on a.

<u>It may also consist in studying the relative degree of dependence of $Pr(A_j)$ on a for different j.</u> Is $\pi_{6j}-\pi_{5j}$ > or < $\pi_{6i}-\pi_{5i}$? Is $\pi_{6j}-2\pi_{5j}+\pi_{4j}$ > or < $\pi_{6i}-2\pi_{5i}+\pi_{4i}$?

The <u>method A</u> with null-state estimated variances would now consist in finding maximum likelihood estimates in the null-state

$$\hat{\phi}_{aj} = \hat{\theta}_j = \sum_a N_{aj}/n = N_j/n = \hat{\pi}_j \text{ (say)} \tag{170}$$

A contrast $\Sigma F_{aj}\pi_{aj}$ is declared > 0 if

$$\sum q_{aj} F_{aj} > \sqrt{z \sum_a [\sum F_{aj}^2 \hat{\pi}_j - (\sum F_{aj}\hat{\pi}_j)^2]/n_a} \tag{171}$$

where z is the $1-\varepsilon$ fractile of the chi-square distribution with $(r-1)(s-1)$ degrees of freedom. (171) is true for some $F = (F_{11},\ldots,F_{sr})$ if

$$Z = n(\sum \frac{N_{aj}^2}{n_a N_j} - 1) > z \tag{172}$$

The <u>method B</u> with a priori estimated variances would be to the effect of declaring $\Sigma F_{aj}\pi_{aj} > 0$ if

$$\sum q_{aj} F_{aj} > \sqrt{z \sum_a [\sum F_{aj}^2 q_{aj} - (\sum F_{aj} q_{aj})^2]/n_a} \tag{173}$$

The minimum modified chi-square estimates are

$$\hat{\phi}_{aj} = \hat{\theta}_j = \bar{\pi}_j / \sum_{i=1}^{r} \bar{\pi}_i \tag{174}$$

where the $\bar{\pi}_j$ are harmonic means of the q_{aj}, i.e.

$$\bar{\pi}_j = n/\sum_a \frac{n_a}{q_{aj}} = n/\sum \frac{n_a^2}{N_{aj}} \tag{175}$$

The modified chi-square statistic is

$$Z = n(-1+1/\Sigma\, \bar{\pi}_j) \tag{176}$$

and this Z is $> z$ if and only if (173) holds for some F. This clearing test $Z > z$ may be written

$$\sum \bar{\pi}_j < (1+ \frac{z}{n})^{-1} \tag{177}$$

(Note that since an harmonic mean is always < an aritmetic mean, we have $\bar{\pi}_j < \hat{\pi}_j$ with equality if and only if all q_{aj} are strictly independent of a. We get $\Sigma\, \bar{\pi}_j < \Sigma\, \hat{\pi}_j = 1$. Hence the heterogeneity of q_{aj} is measured by the degree to which the harmonic means fall short of the arithmetic means.)

REFERENCES

Bishop, Yonne M.M., Fiendberg, Stephen E., and Holland, Paul W. [1974]. Discrete Multivariate Analysis. Cambridge Mass. The MIT Press.

Cramér, H. [1945]. Mathematical Methods of Statistics. Uppsala. Almquist og Wiksell.

Goldstein, Harald [1981]. Robust Inference in Contingency Tables. Oslo, University of Oslo, Institute of Mathematics.

Goodman, L.A. [1964]. Multinomial Confidence Interval for Contrasts among Multinomial Populations. Ann. Math. Stat., vol. 35.

Loève, M., Probability Theory I, [1977]. Berlin Springer Verlag 4th edition.

Neyman J. [1949]. Contribution to the Theory of χ^2 Tests. Proc. of the first Berkeley Symp. of Math. Stat. and Prob. Berkeley, Univ. of California Press.

Reiersøl, O. [1961]. Linear and non-linear Multiple Comparison in Logit Analysis, Biometrika vol. 48.

Scheffé, H. [1953]. A Method of judging all Contrasts in the Analysis of Variance. Biometrica, vol. 40.

Sverdrup, E. [1975]. Multiple Comparisons by Binary and Multiary Observations. Artikler fra Statistisk Sentralbyrå, nr 75, Oslo.

Sverdrup, E. [1976]. Significance Testing in Multiple Statistical Inference. Scand. J. Stat. 3.

Sverdrup, E. [1977a]. The Statistical Analysis of Categorical Observations in the Case of "Large" Homogeneous Subgroups Informal Memo nr. 2 from Statistisk Sentranbyrå, Oslo.

Sverdrup, E. [1977b]. The Logic of Statistical Inference: Significance Testing and Decision Theory. Bull. Int. Stat. Inst. XLVII (1).

"THE METHOD OF MULTIPLE COMPARISON BY A DELTA-METHOD AND ITS RELATIONSHIP TO THE LIKELIHOOD RATIO TEST. GENERAL THEORY AND APPLICATION TO MULTINORMAL MODELS"

by Erling Sverdrup
Statistical Research Report, University of Oslo, No.1 1984

LIST OF CORRECTIONS

p. 31, l.11↑, l.10↑, p.45, l.11↑, l.10↑; "open set → "open convex set"

p. 31, l.3↑; add eq. number "(72)"

p. 32, l.9-10↓; The sentence "The uniformity assumption .... given by (70)" to be deleted.

p. 33, l.7↓; "and only if" to be deleted.

l.3-6↑; The statement in (ii) to be replaced by "(ii). Make the assumption as in A', but without the non-singularity assumption for λ. Then λ is non-singular if the log-likelihood surface $Q(\eta) = \log L(X,\eta)$ has strictly negative curvature in all directions for all η and every $\bar{Z}(X)$ in the joint range of ζ and $\bar{Z}$. Furthermore then $Q(\eta^*) > Q(\eta)$ for $\eta^* \neq \eta$ and $\bar{Z}$ (X) in the same joint range"

p. 34, line 10↓; replace D by $\bar{D}$

line 6↑; replace "3" by "3.(ii)"

p. 38, l.13↑, replace "If η" by "If $\eta^{(n)} = \eta$"

p. 45, l.4↑; delete "and sufficient".