

Statistical Memoirs  
Mathematical Institute  
University of Oslo

No 1.  
March 1971.

TESTING STATISTICAL HYPOTHESIS.  
THE GENERAL NEYMAN-PEARSON THEORY

by

Erling Sverdrup

---

This is a revised version of the paper "Statistical Inference Problems for Darrois-Koopman Classes of Distributions" (Preliminary issue). Oslo, October 1969.

## Preface

With the publication of "On the problem of the most efficient tests of statistical hypotheses" by J. Neyman and E.S. Pearson in 1933 (Philosophical Transactions, series A, vol. 231) a new era emerged in the development of the science of statistical inference. The importance of this paper lies partly in the creation of new concepts, like the "power function", partly in the discovery that for important model situations it was possible to derive statistical methods with stipulated optimal properties which were quite obviously acceptable.

The "model situations" treated by Neyman and Pearson were such as to require certain partial differential equations to be fulfilled by the probability densities. It was later realized that these requirements were essentially equivalent to assuming the Darmois - Koopman exponential family of distributions".

Based on this assumption the modern Neyman-Pearson theory was developed by many authors in the 1950-s, resulting both in a simplification and generalization of the original presentation. It is this theory which is presented in the present memoir. The main result, which is essentially the result obtained by Neyman and Pearson in 1933, is given as theorem II.D.2 (page 38).

The memoir is an adaption of the notes of lectures given at this University at regular intervals since the beginning of the 1950-s, of course with many major alterations, in particular in the 1950-s when new results were steadily forthcoming.

March 1971,

Erling Sverdrup.

## Contents

<u>I. Some special families of distributions</u>	Page
A. Factorized families	1
B. The Darmois-Koopman exponential family of distributions	5
C. The non-parametric family of distributions	16
D. The existence of a conditional probability measure	19
 <u>II. Testing by Darmois-Koopman families of distributions</u>	
A. Neyman-Pearson's lemma	21
B. Conditional testing	23
C. Unbiased and similar tests	27
D. Unbiased one-sided tests. Neyman-Pearson's fundamental theorem	33
E. The general Student Hypothesis	45
F. Performance unbiased three-decision tests	52
G. Unbiased two-sided tests	58
H. Testing of non-regular Darmois-Koopmen classes of distribution	67
 <u>III. Power optimum tests in non-parametric situations</u>	69
 <u>IV. Estimation in connection with Darmois-Koopman classes</u>	
A. Some mathematical results	75
B. The information matrix	77
C. Fisher-consistent estimates	79
D. The maximum likelihood estimator	81

I. Some special families of distributions.

A. Factorized families.

Let  $X$  be a random variable in a sample space  $\mathcal{X}$ ;  $X \in \mathcal{X}$ . A sigmafield  $\mathcal{A}$  is defined in  $\mathcal{X}$ , and  $\mathcal{P}$  is a family of probability measures (distributions)  $P$  over  $\mathcal{A}$ .

Suppose that  $\mathcal{P}$  is dominated by a sigmafinite measure  $\mu$ ; i.e. each  $P \in \mathcal{P}$  is absolutely continuous with respect to  $\mu$ , hence  $dP = f_P(x)d\mu$ . If there exists a real measurable function  $h(x)$  from  $\mathcal{X}$ , a measurable function  $Y(x)$  (statistic) from  $\mathcal{X}$  to a space  $\mathcal{Y}$  with sigmafield  $\mathcal{B}$ , and for each  $P \in \mathcal{P}$  a measurable real function  $g_P(y)$  from  $\mathcal{Y}$  such that

$$f_P(x) = g_P(Y(x))h(x) \quad (1)$$

then  $\mathcal{P}$  is said to be factorized. [We shall agree to call a real function  $h$  from a space  $(\mathcal{X}, \mathcal{A})$  measurable if  $h^{-1}(\text{Borelfield}) \subset \mathcal{A}$ ]. Examples of factorized families will be given in sections B and C below.

There are of course many presentations

(1) if  $\mathcal{P}$  is factorized, since a factor  $H(Y(x))$  could be transferred from  $g_P$  to  $h$ . Obviously we may take  $h(x) \geq 0$ , and will do so below.

We now have

Theorem I.A.1. If  $\mathcal{P}$  is factorized, there exists a probability measure  $\pi$  and a  $g_P$  such that

$$dP = g_P(Y(x)) d\tilde{\mu} \quad (2)$$

Proof: It can be proved that we can always choose  $h$  integrable  $(\mu)$ . Since the proof of this is rather tricky and since this is almost always easily verified in special cases, we shall assume this to be true. Since  $P(X) = 1$ , we have  $\int h(x) d\mu > 0$ , and we may then take the integral to be 1. Then  $d\tilde{\mu} = h(x) d\mu$  defines  $\tilde{\mu}$  as a probability measure, and by the chain rule for the Radon-Nikodym derivative we get (2).

In many situations which we shall consider the family  $\mathcal{S}$  will be homogeneous, i.e. any two measures in  $\mathcal{S}$  are absolutely continuous with respect to each other. Then the proof of theorem 1 is simple without making use of the fact that  $h$  may be chosen integrable. Indeed, in that case we may choose as  $\tilde{\mu}$  any measure in  $\mathcal{S}$ . Taking  $\bar{f}_P(x) = dP/d\tilde{\mu}$ , we then have

$$dP = \bar{f}_P(x) d\tilde{\mu} = \bar{f}_P(x) g_{\tilde{\mu}}(Y(x)) h(x) d\mu$$

On the other hand

$$dP = g_P(Y(x)) h(x) d\mu$$

Combining, we get, since  $g_{\tilde{\mu}}(Y(x)) h(x) > 0$ , a.e.  $(\tilde{\mu})$ ,

$$\bar{f}_P(x) = \frac{g_P(Y(x))}{g_{\tilde{\mu}}(Y(x))} = \bar{g}_P(Y(x)) \quad \text{a.e. } (\tilde{\mu})$$

Hence  $dP = \bar{g}_P(Y(x))d\tilde{\eta}$ , which proves theorem 1. Furthermore, choosing  $\bar{h} = \frac{d\tilde{\eta}}{d\mu}$ , we get  $dP = \bar{g}_P(Y(x))\bar{h}(x)d\mu$ , so it is obvious that in the homogeneous case,  $h$  in (1) could be chosen integrable.

In the homogeneous case we might take  $\tilde{\eta} \in \mathcal{P}$ . Note, however, that in the general case this may not be possible. (Consider, for example, the class of all uniform distributions over  $(0, \tau)$ , with varying  $\tau$ ).

From (2) the sampling distribution  $PY^{-1}$  of  $Y(X)$  is easily found to be given by

$$dPY^{-1} = g_P(y)d\tilde{Y}^{-1} \quad (3)$$

(It is assumed that the reader is familiar with Appendix D: "A more rigorous treatment of some fundamental statistical concepts" (§ 1-2) in Erling Sverdrup: Laws and chance variations, vol.II, p.292.)

We shall now prove

Theorem I.A.2. If the family  $\mathcal{P}$  of distributions is factorized with respect to a statistic  $Y(X)$  (i.e.  $dP$  is given by (1)), then  $Y(X)$  is sufficient for  $\mathcal{P}$ ; i.e. the conditional distribution of  $X$  given  $Y(X)$  is independent of  $P \in \mathcal{P}$  (or rather, "could be chosen independent of  $P$ ").

Proof: We shall use theorem 1 in the proof, assuming that  $h(x)$  in (1) can be chosen integrable. Let  $f(X)$  have finite expectation for any  $P \in \mathcal{P}$ . Then, we have for the

expectation with respect to  $P$

$$\begin{aligned} E_P f(X) &= E_P E_P(f(X)|Y) = \\ &= \int E_P(f(X)|y) dP_Y^{-1} \end{aligned}$$

On the other hand we have by (2) and (3)

$$\begin{aligned} E_P f(X) &= \int f(x) dP = \int f(x) g_P(Y(x)) d\tilde{\mu} = \\ &= E_{\tilde{\mu}}[f(X)g_P(Y(X))] = E_{\tilde{\mu}}[g_P(Y(X))E_{\tilde{\mu}}(f(X)|Y)] = \\ &= \int g_P(y)E_{\tilde{\mu}}(f(X)|y) d\tilde{\mu}^{-1} = \\ &= \int E_{\tilde{\mu}}(f(X)|y) dP_Y^{-1} \end{aligned}$$

Hence, combining the two expressions for  $E_P f$ , we get

$$\int [E_P(f|y) - E_{\tilde{\mu}}(f|y)] dP_Y^{-1} = 0.$$

Now, let  $I_B(y)$  be the indicator function for a set  $B \in \mathcal{B}$ . Take  $f(x) = g(x)I_B(Y(x))$ . Then  $E_P(f|Y) = I_B(Y)E_P(g|Y)$ . Hence, substituting in the integral,

$$\int_B [E_P(g|y) - E_{\tilde{\mu}}(g|y)] dP_Y^{-1} = 0.$$

Since this is true for all  $B \in \mathcal{B}$ , we get

$$E_P(g|y) = E_{\tilde{\mu}}(g|y)$$

a.e.  $(PY^{-1})$  . We now choose in particular  $g(x) = I_A(x)$  in which case we get

$$P(A|y) = \pi(A|y)$$

a.e.  $(PY^{-1})$  . This proves theorem 2. (The obvious requirements of measurabilities of functions and sets will not be explicitly stated here, or below.)

### B. The Darmois-Koopman exponential family of distributions.

This is a special type of a factorized family of distributions relatively to an  $s$ -dimensional statistic  $Y(x) = (Y_1(x), \dots, Y_s(x))$  . A member of the family  $\mathcal{P}$  is given by

$$dP_\tau = A(\tau) e^{\sum_{j=1}^s \tau_j Y_j(x)} h(x) d\mu, \quad (1)$$

and the family is generated by varying  $\tau = (\tau_1, \dots, \tau_s)$  in a set  $\Omega$  giving a one-to-one correspondence between  $\Omega$  and  $\mathcal{P}$  .

It is easily seen that we may assume, without impairing generality, that (i),  $Y_1(x), \dots, Y_s(x)$  are linearly independent, (ii), there are at least  $s$  linearly independent vectors  $\tau \in \Omega$ , (iii),  $A(\tau) \geq 0$  . Then  $h(x) \geq 0$  and since  $P(X) = 1$  , we must have  $A(\tau) > 0$  .

Example 1.  $X = (X_1, X_2, \dots, X_n)$  has independent components which are normal  $(\xi, \sigma)$ . We then get

$$dP_{\gamma} = (2\pi)^{-\frac{n}{2}} \sigma^{-n} e^{-\frac{n\xi^2}{2\sigma^2}} e^{-\frac{1}{2\sigma^2} \sum x_j^2 + \frac{\xi}{\sigma^2} \sum x_j} dx_1 \dots dx_n \quad (2)$$

Hence we have a D.-K. exponential family (1), with

$$Y(x) = (Y_1(x), Y_2(x)) = (\sum x_j^2, \sum x_j), \quad \gamma = (\gamma_1, \gamma_2) = \left(-\frac{1}{2\sigma^2}, \frac{\xi}{\sigma^2}\right),$$

$$A(\gamma) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} e^{-\frac{n\xi^2}{2\sigma^2}}, \quad d\mu = dx_1 \dots dx_n, \quad h = 1.$$

If  $\sigma$  is fixed = 1, then we have a D.-K. exponential family relatively to  $Y = \sum x_i$ . In that case

$$h = e^{-\frac{1}{2} \sum x_j^2}$$

All families of distributions in the case of linear-normal models, i.e. all models in regression analysis and variance analysis; both those with fixed and those with random effects; are D.-K. families of distributions. This is left to the reader to verify.

Example 2.  $X = (X_1, X_2, \dots, X_n)$  has independent components which are gamma-distributed with density (for  $x > 0$ )

$$\frac{x^{\beta-1}}{\Gamma(\beta)} e^{-\lambda x}$$

where  $\alpha, \beta > 0$ . It is easily seen that the family of distributions of  $X = (X_1, \dots, X_n)$  is D.-K. exponential relatively to the statistic

$$Y(x) = (\sum x_j, \sum \log x_j)$$

Example\_3. The components of  $X = (X_1, \dots, X_n)$  form a Bernoulli trial sequence, i.e. they are independent and  $\Pr(X_j = 1) = 1 - \Pr(X_j = 0) = p$ . Then

$$\Pr(\cap (X_j = x_j)) = p^{\sum x_j} (1-p)^{n - \sum x_j} = (1-p)^n e^{\sum x_j \log \frac{p}{1-p}}$$

if all  $x_j$  are either 0 or 1. Thus the probability measure  $P$  is given by (1) with  $Y = \sum x_j$ ,  $\bar{\gamma} = \log \frac{p}{1-p}$ ,  $A(\bar{\gamma}) = (1-p)^n$  and  $\mu(C) = \text{number of points } x = (x_1, \dots, x_n)$  in  $C$  for which all components are 0 or 1. Thus

$$dP = (1-p)^n e^{\sum x_j \log \frac{p}{1-p}} d\mu$$

(Note that when writing  $dP = f d\mu$  it is always understood to mean  $P(C) = \int_C f d\mu$ . Choose in particular  $C = \{x\}$  = the set consisting of the one point  $x$ .)

All ordinary models in multinomial trials, e.g. when testing independence or homogeneity, are families of D.-K. exponential types. This is left to the reader to verify.

Example 4. The components of  $X = (X_1, \dots, X_n)$  are independent and Poisson distributed with common mean  $\lambda$ . It is easily seen that we have a D.-K. class of distributions with  $Y = \sum x_j$  and  $\tau = \log \lambda$ .

Thus many of the classes of distributions commonly considered in statistics are of the D.-K. type. Examples of classes of distributions which are not of D.-K. are classes of rectangular distributions with unknown endpoints, hypergeometric distribution with unknown population size. Of course the non-parametric situations are examples where the classes of distributions are not D.-K.

We now return to the general theory. It is seen that if  $P_\tau(A) = 0$ , then  $h(x) = 0$  a.e. for  $x \in A$ , hence  $P_{\tau'}(A) = 0$ . Thus  $\mathcal{P}$  is homogeneous. We may assume that  $\tau = 0 \in \Omega$ , since this could be obtained in any case by changing origin for  $\tau$ . Furthermore, since  $A(\tau) > 0$  we may assume  $A(0) = 1$ , since this can always be obtained by transferring a constant factor from  $A$  to  $h$ . Hence we have  $dP_0 = h d\mu$  and by the chain rule for Randon-Nikodym derivatives

$$dP_\tau = A(\tau) e^{\sum_{j=1}^S \tau_j Y_j(x)} dP_0 \quad (2)$$

The sampling distribution of  $Y(X)$  is given by

$$dP_\tau Y^{-1} = A(\tau) e^{\sum_{j=1}^S \tau_j Y_j} dP_0 Y^{-1} \quad (3)$$

Of course we have

$$[A(\tilde{z})]^{-1} = \int_{e^{\sum_{j=1}^s \tilde{z}_j Y_j(x)}} dP_0 = \int_{e^{\sum_{j=1}^s \tilde{z}_j Y_j}} dP_0 Y^{-1} \quad (4)$$

It has been tacitly assumed, of course, that the integrals in (4) converge. Thus  $\bar{\Omega}$  is a subset of the region  $\bar{\Omega}$  of all  $\tilde{z}$  for which (4) converges, i.e. (1) defines a probability measure. We now have

Theorem I.B.1: The set  $\bar{\Omega}$  of all  $\tilde{z}$  for which (1) defines a probability measure  $P_{\tilde{z}}$  is convex.

Proof: Let  $\tilde{z}', \tilde{z}'' \in \bar{\Omega}$ . Then we have for  $\tilde{z} = c\tilde{z}' + (1-c)\tilde{z}''$  ( $1 > c > 0$ ),

$$\begin{aligned} [A(\tilde{z})]^{-1} &= \int_{e^{\sum Y_j (c\tilde{z}'_j + (1-c)\tilde{z}''_j)}} dP_0 Y^{-1} = \\ &= \int (e^{\sum Y_j \tilde{z}'_j})^c (e^{\sum Y_j \tilde{z}''_j})^{1-c} dP_0 Y^{-1} \leq \\ &= \left( \int e^{\sum Y_j \tilde{z}'_j} dP_0 Y^{-1} \right)^c \left( \int e^{\sum Y_j \tilde{z}''_j} dP_0 Y^{-1} \right)^{1-c} \end{aligned}$$

by Hölder-Jensen's inequality. Hence  $\tilde{z} \in \bar{\Omega}$ , Q.E.D.

[Hölder-Jensen's inequality: If  $f_1(x), f_2(x) \geq 0$ ,  $0 < c < 1$ ,  $\mu$  a measure, then

$$\int f_1^c f_2^{1-c} d\mu \leq \left( \int f_1 d\mu \right)^c \left( \int f_2 d\mu \right)^{1-c} \quad (5)$$

Proof: The case where  $\int f_i d\mu = 0$  for  $i = 1$  or  $2$  is trivial. Hence we assume  $\int f_i d\mu > 0$  and introduce

$g_i = f_i / \int f_i d\mu$ . The above inequality then takes the form:

$$\int g_1^c g_2^{1-c} d\mu \leq 1; \int g_i d\mu = 1. \quad (5)'$$

Consider now the function  $F(t) = t^c$ . It is concave for  $t > 0$ . Hence its tangent for  $t = 1$  is wholly above the curve, i.e.

$$t^c \leq ct + (1-c)$$

Inserting  $t = g_1/g_2$ , we get

$$g_1^c g_2^{1-c} \leq cg_1 + (1-c)g_2$$

By integration we now get (5)'.]

Let  $f(X)$  be any statistic, the expectation of which exists for all  $\tau \in \bar{\Omega}$ . We shall consider the function

$$\beta(\tau) = E_{\tau} f(X) = \int f(x) dP_{\tau} = A(\tau) \int f(x) e^{\sum \tau_j Y_j(x)} dP_0 \quad (6)$$

which may also be written

$$\beta(\tau) = E_{\tau} E_{\tau} [f(X) | Y] = E_{\tau} g(Y) = A(\tau) \int g(y) e^{\sum \tau_j y_j} dP_0 Y^{-1} \quad (7)$$

where  $g(Y) = E_{\tau} [f(X) | Y]$  by theorem I.A.2 is independent of  $\tau$ .

In particular if  $f(X)$  equals a testfunction  $\delta(X)$  then  $\beta(\tau)$  is the powerfunction.

We now denote by  $\bar{\Omega}^c$  the set of all vectors

$\tau = \zeta + i\sigma = (\zeta_1 + i\sigma_1, \dots, \zeta_s + i\sigma_s)$  with complex components for which  $\zeta = (\zeta_1, \dots, \zeta_s) \in \bar{\Omega}$ . It is seen that the integral expression in (6) and (7)

$$B(\tau) = E_0 f(X) e^{\sum_j \tau_j Y_j(X)} = E_0 g(Y) e^{\sum_j \tau_j Y_j} \quad (8)$$

exists for all  $\tau \in \bar{\Omega}^c$ . In particular this is true for  $f = 1$ , identically. Hence  $A(\tau)$  can be defined by (4) and  $\beta(\tau)$  by (7) for all  $\tau \in \bar{\Omega}^c$ . We shall prove

Theorem I.B.2.  $B(\tau)$  given by (8) is an analytic function of each component of  $\tau$  for all inner points  $\tilde{\tau}$  of  $\bar{\Omega}^c$ . Its derivatives of all orders can be found by interchanging derivation and integration. If, in addition,  $A(\tau)$  is finite then  $\beta(\tau) = A(\tau)B(\tau)$  is analytic and

$$\begin{aligned} \frac{\partial \beta(\tau)}{\partial \tau_r} &= \frac{\partial A(\tau)}{\partial \tau_r} \int g(y) e^{\sum_j \tau_j Y_j} dP_0 Y^{-1} + \\ &+ A(\tau) \int g(y) y_r e^{\sum_j \tau_j Y_j} dP_0 Y^{-1} \end{aligned} \quad (9)$$

Proof: Consider  $B(\tau) = B(\tilde{\tau}_r)$  as a function of the component  $\tilde{\tau}_r$  of  $\tau$ . We shall prove that  $B$  is an analytic function of  $\tilde{\tau}_r$ . We have

$$\frac{B(\tilde{\tau}_r + z) - B(\tilde{\tau}_r)}{z} = \int g(y) \frac{e^{z Y_r} - 1}{z} e^{\sum_j \tau_j Y_j} dP_0 Y^{-1} \quad (10)$$

We now have with  $z = u + iv$

$$\left| \frac{e^{zy_r} - 1}{z} \right| \leq \text{const.} \cdot |y_r| e^{|u| \cdot |y_r|} \quad (11)$$

$$\begin{aligned} & \left[ \text{This is seen as follows: } \frac{e^{az} - 1}{z} = \right. \\ & = (e^{ua} \cdot \cos va - 1 + i e^{ua} \cdot \sin va) / z = \\ & = \frac{e^{ua} - 1}{u} \frac{u}{z} \cos va + \frac{\cos va - 1}{v} \frac{v}{z} + i \frac{\sin va}{v} e^{ua} \end{aligned}$$

Obviously  $\left| \frac{u}{z} \right|$  and  $\left| \frac{v}{z} \right| \leq 1$ . By the mean value theorem for derivatives

$$\left| \frac{e^{ua} - 1}{u} \right| = |ae^{u_1 a}| \leq |a| e^{|u| \cdot |a|} \quad \text{where } |u_1| \leq |u|$$

$$\left| \frac{\cos va - 1}{v} \right| = |a \sin v_1 a| \leq |a| \quad \text{where } |v_1| \leq |v|$$

$$\left| \frac{\sin va}{a} \right| \leq |a|$$

Hence

$$\left| \frac{e^{az} - 1}{z} \right| \leq |a| e^{|u| \cdot |a|} + |a| + |a| e^{|u| \cdot |a|} \leq 3|a| e^{|u| \cdot |a|}$$

which is the same as (11).]

For any  $\varepsilon > 0$  we can always choose  $K$  such that  $|y_r| \leq K e^{\frac{\varepsilon}{2}} |y_r|$ . With  $|u| \leq |z| \leq \frac{\varepsilon}{2}$  we then have

$$\left| \frac{e^{zy_r} - 1}{z} \right| \leq \text{const.} \cdot e^{\frac{\varepsilon}{2}} |y_r| \quad (12)$$

Thus the integrand in (10) has modulus

$$\leq \text{const.} \cdot |g(y)| e^{\frac{\varepsilon}{2}} |y_r| + \sum_j \rho_j y_j$$

where  $\rho_j$  = the real part of  $\gamma_j$ . But since  $\gamma$  is an inner

point of  $\bar{\Omega}^c$  the integral of  $|g(y)| e^{\sum p_j y_j} dP_0 Y^{-1}$  converges if  $\int_{\mathcal{R}}$  is replaced by  $\int_{\mathcal{R} \pm \varepsilon}$  for  $\varepsilon$  sufficiently small. By the Lebesgue dominated convergence theorem we then have

$$\frac{\partial B(\tau)}{\partial \tau_r} = \int y_r g(y) e^{\sum \tau_j y_j} dP_0 Y^{-1} \quad (13)$$

Hence we have proved analyticity. Since the proof holds for any  $P_0 Y^{-1}$ -integrable  $g(y)$ , we have in particular for  $g \equiv 1$  that  $A(\tau)^{-1}$  is analytic. Hence if  $A(\tau)^{-1} \neq 0$ , we have that  $\beta(\tau) = A(\tau)B(\tau)$  is analytic. (9) follows from (13). The statement about derivatives of higher order is proved similarly, Q.E.D.

A family  $\mathfrak{P}$  of distributions is said to be complete if for any  $f$

$$\int f dP = 0 \quad \text{for all } P \in \mathfrak{P} \quad (14)$$

implies that  $f = 0$  a.e. ( $\mathfrak{P}$ ).

A Darmois-Koopman family of distributions  $\mathfrak{P} = \{P_\tau\}_{\tau \in \Omega}$  is said to be regular if  $\Omega$  contains inner points. Obviously we can, and will, in that case take  $\tau = 0 \in \Omega$  as an inner point.

Theorem I.B.3: In a regular Darmois-Koopman family of distributions the class of sampling distributions for the sufficient statistic  $Y$  is complete.

Proof: By assumption  $\mathfrak{P} = \{P_\tau\}_{\tau \in \Omega}$  contains  $\tau = 0$  as an inner point. We shall prove that

$$\int g(y) dP_{\tau} Y^{-1} = 0$$

for all  $\tau \in \Omega$  implies  $g = 0$  a.e.  $(P_0 Y^{-1})$ . Using (3) this equation can be written (since  $A(\tau) > 0$ ):

$$\int g(y) e^{\sum \tau_j y_j} dP_0 Y^{-1} = 0 \quad (15)$$

Now, the left hand side and the right hand side of (15) can be considered as functions of  $\tau_1, \dots, \tau_s$ , which are both analytic when  $\tau$  is an inner point of  $\bar{\Omega}^c$ . They are identical for all  $\tau$  when each  $\tau_r$  is real and is contained in an open interval  $(-a, +a)$ . By a famous theorem about analytic functions we then have that the identity (15) is true for all  $\tau = \rho + i\epsilon$  satisfying the condition  $-a < \rho_r < a$ . Hence we have in particular for  $\tau = it$  ( $t$  real)

$$\int g(y) e^{i \sum t_j y_j} dP_0 Y^{-1} = 0$$

We write  $g = g^+ - g^-$ , splitting  $g$  in its positive and negative parts. Hence

$$\int g^+(y) e^{i \sum t_j y_j} dP_0 Y^{-1} = \int g^-(y) e^{i \sum t_j y_j} dP_0 Y^{-1} \quad (16)$$

and in particular

$$K = \int g^+ dP_0 Y^{-1} = \int g^- dP_0 Y^{-1} \quad (17)$$

If  $K = 0$ , then  $g^+ = g^- = 0$  a.e. and everything is proved. Assume now  $K > 0$ . Then  $\rho^+$  and  $\rho^-$  defined by

$$\rho^+(B) = \int_B \frac{g^+}{K} dP_0 Y^{-1}, \quad \rho^-(B) = \int_B \frac{g^-}{K} dP_0 Y^{-1}$$

are probability measures since

$$\int \frac{g^+}{K} dP_0 Y^{-1} = \int \frac{g^-}{K} dP_0 Y^{-1} = 1$$

Dividing by  $K$ , we can then write (16) by the chain rule for Radon-Nikodym derivatives

$$\int e^{i \sum t_j Y_j} d\rho^+ = \int e^{i \sum t_j Y_j} d\rho^- \quad (18)$$

Hence by the inversion theorem for characteristic functions,  $\rho^+(B) = \rho^-(B)$  for all  $B \in \mathcal{B}$ , i.e.

$$\int_B \frac{g^+}{K} dP_0 Y^{-1} = \int_B \frac{g^-}{K} dP_0 Y^{-1}$$

for all  $B$ . Thus  $g^+ = g^-$  a.e., and  $g = 0$  (contradicting that  $K > 0$ ), Q.E.D.

From this theorem it follows that two test functions  $\delta_1(x) = \Delta_1(Y(x))$  and  $\delta_2(x) = \Delta_2(Y(x))$  which are based on the sufficient statistic  $Y$  and have the same power function must be equal a.e.  $\Delta_1(y) = \Delta_2(y)$ , a.e.

It also follows that if

$$E_{\gamma} m(Y) = \mu(\gamma) \quad (19)$$

then  $m(Y(X))$  is the only unbiased estimator for  $\mu(\tau)$  based on the sufficient statistic. Furthermore it is a Markov estimator, i.e. if  $M(X)$  is any other unbiased estimator for  $\mu(\tau)$ , then  $\text{var } M(X) \geq \text{var } m(Y)$  for all  $\tau$ . These are immediate consequences of the results in Erling Sverdrup: Laws and chance variations. Vol. II. Ch. III 2.2, where also examples are given of estimands  $\mu(\tau)$  connected with Darmais-Koopman families of distributions.

### C. The non-parametric family of distributions.

We shall be primarily concerned with Darmais-Koopman families of distributions. However, in order to throw light on the general nature of some of the principles used we shall occasionally consider other classes of distributions.

Let  $\mathcal{S}$  be the class of distributions of  $X = (X_1, \dots, X_n)$ , where the components are independent with the same density  $f$ , with respect to the Lebesgue measure, and  $f$  is any among a class of densities  $\Omega$ ,  $\int f dx = 1$ . For the time being let  $\Omega$  be the class of all densities. The density of  $X$  is  $f(x_1) \cdot f(x_2) \cdots f(x_n)$ . The density of the order statistic  $Y(X) = (Y_1(X), \dots, Y_n(X))$ ; where  $Y_1(X) \leq Y_2(X) \leq \dots \leq Y_n(X)$  are  $X_1, \dots, X_n$  arranged in non-decreasing sequence; is then,

$$n! f(y_1) \cdots f(y_n) \quad (y_1 \leq \dots \leq y_n) \quad (1)$$

Theorem I.C.1. In the non-parametric family of distributions the class of distributions of the order statistic is complete.

Proof: We have to prove that  $E_f g(Y) = 0$  for all densities  $f$ , implies that  $g(y) = 0$  a.e. Consider then

$$E_f g(Y) = n! \int_{y_1 \leq y_2 \leq \dots \leq y_n} g(y_1, \dots, y_n) f(y_1) \dots f(y_n) dy_1 \dots dy_n = 0 \quad (2)$$

Let now  $h(x) = g(Y(x))$ . We see that  $h(x)$  may be written in this form if and only if  $h(x)$  is symmetric in  $x_1, \dots, x_n$ ; i.e.  $h(x_1, \dots, x_n) = h(x_{i_1}, \dots, x_{i_n})$  for all permutations  $i_1, \dots, i_n$  of  $1, 2, \dots, n$ . Hence (2) may be written

$$Eh(X) = \int \dots \int h(x_1, \dots, x_n) f(x_1) \dots f(x_n) dx_1 \dots dx_n = 0 \quad (3)$$

for all  $f$ . We thus have to prove that if (3) is true and  $h$  is symmetric, then  $h = 0$  a.e.

We now insert a particular  $f$  in (3). Let  $I_1, \dots, I_n$  be  $n$  arbitrary non-overlapping finite intervals with lengths  $L_1, \dots, L_n$  respectively, and let  $p_1, \dots, p_n$  be arbitrary non-negative numbers such that  $\sum p_j = 1$ . Then we define

$$\begin{aligned} f(x) &= p_j / L_j \quad \text{if } x \in I_j; \quad j = 1, 2, \dots, n \\ f(x) &= 0 \quad \text{otherwise} \end{aligned} \quad (4)$$

(Note the peculiarity that the density  $f$  of any single  $X_j$  depends on  $n$ .) We see that  $\int f(x) dx = 1$  and we find

$$Eh(X) = \sum_{i_1, \dots, i_n=1}^n p_{i_1} \dots p_{i_n} Q(i_1, \dots, i_n) = 0 \quad (5)$$

for all  $p_1, \dots, p_n$  and  $I_1, \dots, I_n$ , where

$$Q(i_1, \dots, i_n) = \frac{1}{L_{i_1} \dots L_{i_n}} \int_{I_{i_1}} \dots \int_{I_{i_n}} h(x_1, \dots, x_n) dx_1 \dots dx_n \quad (6)$$

From the symmetry of  $h$  follows the symmetry of  $Q$ . Hence the value of  $Q$  is determined by giving the number of  $i_1, \dots, i_n$  which are equal to  $1, 2, \dots, n$ , respectively. Let this be  $m_1, \dots, m_n$ ;  $\sum m_j = n$ . Thus we can write

$$Q(i_1, \dots, i_n) = K(m_1, \dots, m_n) \quad (7)$$

(5) may now be written

$$\sum_{i_1, \dots, i_n} p_1^{m_1} \dots p_n^{m_n} K(m_1, \dots, m_n) = 0 \quad (8)$$

Let us now collect the terms in (8) leading to the same  $m_1, \dots, m_n$ . Assume that there are  $N(m_1, \dots, m_n)$  different  $(i_1, \dots, i_n)$  leading to  $(m_1, \dots, m_n)$ . (Obviously  $N = \frac{n!}{m_1! \dots m_n!}$ ). Set  $C = KN$ . We then get

$$\sum_{m_1 + \dots + m_n = n} p_1^{m_1} \dots p_n^{m_n} C(m_1, \dots, m_n) = 0 \quad (9)$$

for all  $I_1, \dots, I_n$  and all  $p_1, \dots, p_n$  such that  $\sum_{i=1}^n p_i = 1$ . We assume  $p_n > 0$  and divide (9) by  $p_n^n$ ,

$$\sum_{m_1, \dots, m_n} t_1^{m_1} \dots t_{n-1}^{m_{n-1}} C(m_1, \dots, m_n) = 0 \quad (10)$$

where  $t_j = p_j/p_n$ . Thus (10) is true for all  $t_j \geq 0$ . We now

operate with  $\frac{\partial^{m_1 + \dots + m_{n-1}}}{\partial t_1^{m_1} \dots \partial t_{n-1}^{m_{n-1}}}$  on the identity and obtain

$C(m_1, \dots, m_n) = 0$  after putting all  $t_j = 0$ . Hence, since  $N > 0$ ,

$K(m_1, \dots, m_n) = 0$  and  $Q(i_1, \dots, i_n) = 0$ . Thus we get from (6)

$$\int_{I_{i_1}} \dots \int_{I_{i_n}} h(x_1, \dots, x_n) dx_1 \dots dx_n = 0 \quad (11)$$

Now, it is seen that the class of  $n$ -dimensional intervals of the type  $I_{i_1} \times \dots \times I_{i_n}$  (which is not an arbitrary interval) is a basis for the Borel class. Thus

$$\int_B \dots \int h(x_1, \dots, x_n) dx_1 \dots dx_n = 0 \quad (12)$$

for any Borel set  $B$ . From (12) it follows that

$h(x_1, \dots, x_n) = 0$  a.e.; Q.E.D.

#### D. The existence of a conditional probability measure.

We shall make a remark concerning the existence of a conditional probability measure. (See E. Sverdrup: Laws and Chance Variations, volum II, appendix D.) Given a probability space  $(\mathcal{X}, \mathcal{A}, \mathcal{P})$  and a measurable function  $Y(x)$  from this space to a space  $(\mathcal{Y}, \mathcal{B})$ , we define  $P(A|y) = \Pr(X \in A | Y = y)$  as the measurable function of  $y$  such that

$$P(A \cap Y^{-1}(B)) = \int_B P(A|y) dPY^{-1}; \quad A \in \mathcal{A}; \quad B \in \mathcal{B}.$$

This conditional probability  $P(A|y)$  is known to have "almost" the properties of a probability measure. However, in general it need not be a probability measure and this fact will cause mathematical difficulties in stating and proving some general results (e.g. theorems II.C.2, II.D.1-2, II.G.1). In special cases it is easily verified that  $P(A|y)$  does exist as a proper measure,

and we shall be content with that.

There are several courses to follow: (i) Prove that if  $\mathcal{K}$  is Euclidian then the conditional probability measure does indeed exist. Then we should assume  $\mathcal{K}$  to be Euclidian. (ii) Add the assumption "if  $P(A|y)$  exists as a probability measure" in the theorems below. We shall do neither. Instead we shall (iii) state once and for all that we assume everywhere below that the conditional probability measure does exist.

## II. Testing by Darms-Koopman families of distributions

### A. Neyman-Pearson's lemma.

We shall briefly state and prove a theorem which is known as "Neyman-Pearson's lemma".

Theorem II. A.1. Let  $f_1, f_2, \dots, f_m, f$  be  $m+1$  real-valued functions, integrable with respect to a measure  $\mu$  and let  $c_1, \dots, c_m$  be real numbers. Consider the class  $\Delta$  of test functions  $\delta$  which are such that

$$\int \delta f_i d\mu = c_i ; \quad i = 1, 2, \dots, m \quad (1)$$

Suppose that  $\delta_0 \in \Delta$  is such that for some  $k_1, \dots, k_m$ ,

$$\delta_0(x) = 1 \quad \text{if} \quad f(x) > \sum_{i=1}^m k_i f_i(x) \quad (2)$$

$$\delta_0(x) = 0 \quad \text{if} \quad f(x) < \sum_{i=1}^m k_i f_i(x) \quad (3)$$

then

$$\int \delta_0 f d\mu \geq \int \delta f d\mu \quad (4)$$

for all  $\delta \in \Delta$ .

Proof: Since  $\int \delta f_i d\mu = \int \delta_0 f_i d\mu$ , we have

$$\int \delta_0 f d\mu - \int \delta f d\mu = \int (\delta_0 - \delta)(f - \sum k_i f_i) d\mu \quad (5)$$

and since, by (2) and (3), the integrand on the right hand side is  $\geq 0$  for all  $x$ , (4) follows, Q.E.D.

It is easily seen that if "=" is replaced by  $\leq$  in (1) for all  $i$  for which  $k_i \geq 0$ , then the conclusion (4) still remains true.

We assume the reader to be familiar with the use of this lemma in the case of testing simple (completely specified) hypotheses. The case  $m = 1$  is well known from elementary texts. Then  $c_1$  = level of significance,  $f_1$  = density under the null-hypothesis and  $f$  = density under the alternative.

The existence of an optimum  $\delta_0$  is not guaranteed by the lemma. It can be shown to be the case if the space is Euclidian,  $\mu$  is sigma-finite, and there are (of course) certain restrictions on  $c_1, \dots, c_m$ . This result will not be needed in the sequel.

## B. Conditional testing.

Let  $X$  be an observed random variable the distribution  $P$  of which is known to belong to a family  $\mathcal{P}$ . The problem is to construct a test  $\delta(X)$  with level  $\varepsilon$  for testing the null-hypothesis  $P \in \mathcal{P}_0$  where  $\mathcal{P}_0$  is a subfamily of  $\mathcal{P}$ . Hence we search for an optimum test  $\delta_0$  among all  $\delta$  satisfying  $\int \delta(x) dP \leq \varepsilon$  for  $P \in \mathcal{P}_0$ , and perhaps some other side conditions.

Conditional testing is said to be performed if a test  $\delta$  is constructed in the following manner. First a statistic  $Y(X)$  is chosen and the conditional distribution  $P(A|y)$  is considered for all  $P \in \mathcal{P}$ . Hence for each  $y$  an a priori family  $\mathcal{P}^y$  of distributions  $P(A|y)$  is generated by varying  $P$  in  $\mathcal{P}$  and similarly a family  $\mathcal{P}_0^y$  is generated by varying  $P \in \mathcal{P}_0$ .

Now, for each  $y$  a testfunction  $\delta(\cdot; y)$  of  $x$  is constructed such that

$$\int \delta(\cdot; y) dP(\cdot|y) \leq \varepsilon \quad \text{if } P \in \mathcal{P}_0$$

If  $Y(X) = y$ , the null hypothesis is rejected with probability  $\delta(x; y)$  when  $X = x$  is observed. But this is the same as rejecting the hypothesis with probability  $\delta(x; Y(x)) = \delta(x)$  if  $X = x$ . Writing  $\delta(x)$  in the form  $\delta(x; Y(x))$  does not, of course, restrict  $\delta(x)$ , but we have in addition

$$E[\delta(X) | Y(X) = y] \leq \varepsilon \quad \text{for } P \in \mathcal{P}_0 \quad (1)$$

So this is really the condition defining conditional testing. Of course, we then have for  $P \in \mathcal{P}_0$ ,

$$E\delta(X) = E E(\delta(X)|Y) \leq \varepsilon$$

so that the test has indeed level  $\varepsilon$ .

Sometimes it is argued that it is a priori reasonable to perform conditional testing given a very specific statistic  $Y(x)$ . Then a test procedure for testing  $P(\cdot|y) \in \mathcal{P}_0^Y$  against  $P(\cdot|y) \in \mathcal{P}^Y - \mathcal{P}_0^Y$ , is constructed. The last optimum problem is sometimes much simpler, involving just one parameter, namely the one of interest.

Example 1.  $X_1, X_2$  are independent and Poisson-distributed with  $EX_1 = \lambda_1$ ,  $EX_2 = \lambda_2$ . We shall test  $\lambda_2 \leq a\lambda_1$  against  $\lambda_2 > a\lambda_1$ .

We transform from parameters  $\lambda_1, \lambda_2$  to parameters  $\pi = \frac{\lambda_1}{\lambda_1 + \lambda_2}$ ,  $\lambda = \lambda_1 + \lambda_2$ . Thus we shall test  $\pi \geq \frac{1}{1+a} = \pi_0$  against  $\pi < \frac{1}{1+a} = \pi_0$ . We condition with respect to  $X = X_1 + X_2$ . Given  $X = n$  the distribution of  $X_1$  is binomial  $(n, \pi)$  and we have a simple problem of testing the probability  $\pi$  in a binomial distribution. In the conditional testproblem only the parameter of interest  $\pi$  is involved. The nuisance parameter  $\lambda = \lambda_1 + \lambda_2$  has been eliminated. Furthermore, in this case the variable  $X$  with respect to which we have conditioned depends only on the nuisance parameter. However, this is not a general property of the conditional test.

Example 2.  $X_1, X_2$  are independent and binomially distributed, respectively  $(n_1, \pi_1)$  and  $(n_2, \pi_2)$ . We want to test  $\pi_1 \leq \pi_2$  against  $\pi_1 > \pi_2$ . Again by conditioning with respect to  $X = X_1 + X_2$  we get a distribution which depends only on  $\pi_1/\pi_2$ , and if  $\pi_1 = \pi_2$ , this distribution is hypergeometric. However, in this case the distribution of  $X$  depends effectively on both

$\tilde{\pi}_1$  and  $\tilde{\pi}_2$ .

Example 3.  $X_1, \dots, X_n$  are independent normal  $(\xi, \sigma)$ . A test of the null-hypothesis  $\xi = K\sigma^2$  against  $\xi > K\sigma^2$  is wanted. Since  $V = \sum X_i$ ,  $Y = \sum X_i^2$  is a set of sufficient statistics, we will confine ourselves to consider families of distributions for  $(V, Y)$ .

We introduce  $\tilde{\xi} = \xi/\sigma^2$  and shall thus test  $\tilde{\xi} = K$  against  $\tilde{\xi} > K$ . In order to obtain a test problem involving just  $\tilde{\xi}$  (and not  $\sigma$ ), we consider the conditional model given  $Y = \sum X_i^2$ .

The joint distribution of  $V = \sum X_i$  and  $Z = \sum (X_i - \bar{X})^2 = Y - \frac{V^2}{n}$  is well-known. Transforming to  $(V, Y)$  we obtain for the joint density of  $(V, Y)$

$$\frac{1}{\sigma \sqrt{n}} g\left(\frac{v - n\tilde{\xi}}{\sigma \sqrt{n}}\right) \frac{1}{\sigma^2} Y_{n-1}\left(\frac{yn - v^2}{n\sigma^2}\right), \quad (2)$$

where  $g$  is the standard Gaussian density and  $Y_{n-1}$  is the chi-square density with  $n-1$  degrees of freedom. Now, we know that  $Y/\sigma^2$  is chi-square distributed with  $n$  degrees of freedom and eccentricity  $\lambda = n\tilde{\xi}^2/\sigma^2$ . Hence we can write down the density of  $Y$  and divide (2) by it. We then obtain for the conditional density of  $V$  given  $Y$ ,

$$\beta_n(v; y, \tilde{\xi}) = \frac{1}{\sqrt{n} \Gamma(\frac{n-1}{2})} \frac{1}{v \sqrt{y}} \left(1 - \frac{v^2}{ny}\right)^{\frac{n-3}{2}} Q(n\tilde{\xi}^2 y)^{-1} e^{tv}; \quad (3)$$

$$|v| \leq \sqrt{yn};$$

where

$$Q(t) = \sum_{j=0}^{\infty} \frac{t^j}{(2j)!} \frac{\Gamma(j + \frac{1}{2})}{\Gamma(\frac{n}{2} + j)} \quad (4)$$

We have a test problem involving just one parameter  $\tilde{\xi}$ ;  $\sigma$  has disappeared. Furthermore it is seen that we have a Darmais-

Koopman class of distributions relatively to the sufficient statistic  $V = \sum X_i$ .

The test problem is now easily solved by means of theorem II. A.1. We find the uniformly most powerful conditional test among all level  $\epsilon$ -tests, given by the rejection region  $V > c(Y)$ , where  $c(y)$  is determined by

$$\int_{c(y)}^{\infty} \beta_n(v; y, K) dv = \epsilon \quad (5)$$

This test could be given a different form by introducing the Student statistic

$$T = \bar{X} \sqrt{n}/S \quad (6)$$

where  $\bar{X} = V/n$  and  $S^2 = Z/(n-1)$ . Since we can write

$$T = \sqrt{\frac{n-1}{n}} \frac{V}{\sqrt{Y - \frac{V^2}{n}}} \quad (7)$$

and this is increasing with  $V$ , it is seen that the region of rejection could be written  $T > t(Y)$ , where  $t(Y)$  is given by replacing  $(T, V)$  by  $(t(Y), c(Y))$  in (7). Thus we have conditional Student testing given  $\sum X_i^2$ .

Consider now the special case with  $K = 0$ . Then we have a Student test situation of testing  $\xi = 0$  against  $\xi > 0$ . Under the null hypothesis  $\tau = K = 0$ , the density (3) reduces to

$$\beta_n(v; y, 0) = \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{1}{2}) \sqrt{n} \Gamma(\frac{n-1}{2})} \frac{1}{\sqrt{y}} \left(1 - \frac{v^2}{ny}\right)^{\frac{n-3}{2}}$$

The conditional density of  $W = V/\sqrt{Y}$  given  $Y$ , is obtained by a simple transformation from  $V$  to  $W$ , giving,

$$\mathcal{H}(w) = \frac{\Gamma(\frac{n}{2})}{\sqrt{n\pi} \Gamma(\frac{n-1}{2})} \left(1 - \frac{w^2}{n}\right)^{\frac{n-3}{2}} \quad (w \leq \sqrt{n})$$

which is independent of  $Y$ . By (7)  $T = \sqrt{\frac{n-1}{n}} \frac{W}{\sqrt{1 - \frac{W^2}{n}}}$ . Hence Student's  $T = \bar{X}\sqrt{n}/S$  and  $\Sigma X_i^2$  are independent if  $\xi = 0$ .

We now have from (5), when  $\xi = 0$   
 $\epsilon = \Pr(V > C(Y) | Y=y) = \Pr(V/Y > C(Y)/Y | Y=y) =$   
 $\Pr(V/Y > C(y)/y | Y=y) = \Pr(V/Y > C(y)/y) = \Pr(T > t(y))$ .  
Hence  $t(y) = t$  is independent of  $y$  and we have obtained the ordinary Student test. Trivially it is a conditional test given  $\Sigma X_i^2$ , since  $T$  is independent of  $\Sigma X_i^2$ . What is, however, interesting is that it is the uniformly most powerful level  $\epsilon$  test among all conditional tests given  $\Sigma X_i^2$ .

### C. Unbiased and similar tests.

As in section A we consider a situation where it is a priori known that the observation  $X$  has a probability distribution  $P$  which belongs to a class  $\mathcal{P}$  of distributions. We want to test the null hypothesis that  $P \in \mathcal{P}_0$ , where  $\mathcal{P}_0 \subset \mathcal{P}$ . An arbitrary test is exhibited as a function  $\delta(x)$  giving the conditional probability of rejecting the null hypothesis given that  $X = x$ . The power function of  $\delta$  is

$$\beta(P; \delta) = E_P \delta(X) = \int \delta(x) dP \quad (1)$$

and, as in section A, we shall require that  $\delta$  has level  $\epsilon$ , i.e.

$$\beta(P; \delta) \leq \varepsilon \quad \text{for } P \in \mathcal{P}_0 \quad (2)$$

In this section we shall also require that  $\delta$  is unbiased, i.e.

$$\beta(P; \delta) \geq \varepsilon \quad \text{for } P \in \mathcal{P} - \mathcal{P}_0 \quad (3)$$

The important implication of such a requirement is given in the following theorem.

Theorem II.C.1. Suppose that it is possible to define a topology in  $\mathcal{P}$  (i.e. a limit concept for  $P$ ) such that  $\beta(P; \delta)$  is a continuous function of  $P$  for each  $\delta$ . Let  $\mathcal{P}_0'$  be the class of boundary "points" for  $\mathcal{P}_0$  according to this topology. Then any test  $\delta$  which is unbiased with level  $\varepsilon$  must be similar on  $\mathcal{P}_0'$  with level  $\varepsilon$ , i.e.

$$\beta(P; \delta) = \varepsilon \quad \text{for } P \in \mathcal{P}_0' \quad (4)$$

Proof: Suppose that there exists a  $P \in \mathcal{P}_0'$  such that  $\beta(P; \delta) < \varepsilon$ . We can then find a  $P_1 \in \mathcal{P} - \mathcal{P}_0$  sufficiently close to  $P$  such that  $\beta(P_1; \delta) < \varepsilon$ , which violates (3). In the same manner it is seen that  $\beta(P; \delta) > \varepsilon$  for  $P \in \mathcal{P}_0'$  leads to a violation of (2). Q.E.D.

We shall now investigate the consequence of similarity on  $\mathcal{P}_0'$  when this family of distributions is of the regular Darms-Koopman exponential type, i.e.  $\mathcal{P}_0' = \{P_\tau\}_{\tau \in \omega}$ , where

$$dP_\tau = A(\tau) e^{\sum_{j=1}^s \tau_j Y_j(x)} dP_0, \quad (5)$$

and  $\omega$  contains  $\tau = 0$  as an inner point.

Theorem II.C.2. Consider a regular Darmonis-Koopman exponential family of distributions  $\{P_{\tilde{\lambda}}\}_{\tilde{\lambda} \in \omega}$  relatively to the statistic  $Y(x)$ , where  $\omega$  has  $\tilde{\lambda} = 0$  as an inner point. A necessary and sufficient condition for  $\delta$  to be similar relatively to the family, i.e.

$$E_{\tilde{\lambda}} \delta(X) = \int \delta(x) dP_{\tilde{\lambda}} = \varepsilon \quad (6)$$

for all  $\tilde{\lambda} \in \omega$ , is that it is a conditional test given  $Y$ ; i.e.

$$E_0(\delta(X) | Y) = \varepsilon \quad \text{a.e.} \quad (7)$$

or equivalently

$$E_{\tilde{\lambda}}(\delta(X) | Y) = \varepsilon \quad \text{a.e.} \quad (8)$$

for all  $\tilde{\lambda} \in \omega$ .

Proof: By theorem I.A.2 we know that  $Y$  is sufficient such that  $E_{\tilde{\lambda}}(\delta(X) | Y)$  is independent of  $y$ . Thus (7) and (8) are equivalent. Furthermore

$$E_{\tilde{\lambda}} \delta(X) = E_{\tilde{\lambda}} E_{\tilde{\lambda}} [\delta(X) | Y] = E_{\tilde{\lambda}} E_0 [\delta(X) | Y]$$

Hence it is seen that a necessary and sufficient condition for  $\delta$  to be similar with level  $\varepsilon$ , is that

$$E_{\tilde{\lambda}} \{ E_0 [\delta(X) | Y] - \varepsilon \} = 0 \quad (9)$$

However, (7) obviously implies (9). Vice versa, if (9) is true, then it follows from the completeness property of  $\{P_{\tilde{\lambda}}^{Y^{-1}}\}_{\tilde{\lambda} \in \omega}$  (see theorem I.B.3) that (7) must be true, since  $E_0 [\delta(X) | Y] - \varepsilon$  is a function of  $Y$ . Q.E.D.

Example. The result of a trial can be classified according to two factors A and B with levels  $A_1, \dots, A_r$  and  $B_1, \dots, B_s$  respectively. The results of  $n$  such trials could then be given in a table as follows,

A \ B	B <sub>1</sub>	B <sub>2</sub>	...	B <sub>s</sub>	
A <sub>1</sub>	X <sub>11</sub>	X <sub>12</sub>	...	X <sub>1s</sub>	X <sub>1.</sub>
A <sub>2</sub>	X <sub>21</sub>	X <sub>22</sub>	...	X <sub>2s</sub>	X <sub>2.</sub>
⋮	⋮	⋮		⋮	⋮
A <sub>r</sub>	X <sub>r1</sub>	X <sub>r2</sub>	...	X <sub>rs</sub>	X <sub>r.</sub>
	X <sub>.1</sub>	X <sub>.2</sub>	...	X <sub>.s</sub>	n

Thus  $X_{ij}$  is the number of times the levels  $A_i$  and  $B_j$  appear together in a trial. We define

$$X_{i.} = \sum_j X_{ij}, \quad X_{.j} = \sum_i X_{ij} \quad (10)$$

and have

$$\sum_{i,j} X_{ij} = \sum_i X_{i.} = \sum_j X_{.j} = n \quad (11)$$

We want to test if A and B are "independent" factors. We shall consider two different meanings of this concept, according as we use model I and model II below.

Model I (Independence testing). We have one multinomial trial sequence of  $n$  independent trials. In each trial we have

$$\Pr(A_i \cap B_j) = p_{ij} \quad (12)$$

and hence

$$\Pr(A_i) = \sum_j p_{ij} = p_{i.}, \quad \Pr(B_j) = \sum_i p_{ij} = p_{.j} \quad (13)$$

Thus we have

$$\Pr \left[ \bigcap_{i,j} (X_{ij} = x_{ij}) \right] = \frac{n!}{x_{11}! \dots x_{rs}!} p_{11}^{x_{11}} \dots p_{rs}^{x_{rs}} \quad (14)$$

(where  $\{x_{ij}\}$  is a set of natural numbers with  $\sum_{i,j} x_{ij} = n$ ). We want to test if A and B are stochastically independent, i.e. if

$$p_{ij} = p_{i.} \cdot p_{.j} \quad (15)$$

Obviously the assumption of theorem 1 is fulfilled in this case and thus unbiasedness leads to similarity under the null-hypothesis (since  $\mathcal{G}_0 = \mathcal{G}_0'$ ). In order to see if the assumption of theorem 2 is fulfilled, we write (14) when (15) is true as

$$dP = (rs)^n e^{\sum_{i=1}^r x_{i.} \log p_{i.} + \sum_{j=1}^s x_{.j} \log p_{.j}} dP_0$$

where P and  $P_0$  are the joint probability distributions of all  $X_{ij}$  respectively when  $p_{ij} = p_{i.} \cdot p_{.j}$  and  $p_{ij} = \frac{1}{rs}$ . Now, this does not immediately show that we have a regular Darmois-Koopman class of distributions, since  $\sum p_{i.} = 1$ , i.e. there is a functional relationship between the parameters appearing in the exponent. However, by means of (11) we get

$$dP = (rs p_{r.} p_{.s})^n e^{\sum_{i=1}^{r-1} x_{i.} \log \frac{p_{i.}}{p_{r.}} + \sum_{j=1}^{s-1} x_{.j} \log \frac{p_{.j}}{p_{.s}}} dP_0 \quad (16)$$

which is obviously of the form (5) with the assumption of theorem 2 fulfilled. We have  $Y = (X_{1.}, X_{2.}, \dots, X_{r-1.}, X_{.1}, X_{.2}, \dots, X_{.s-1})$  in this case. Thus we are led to conditional testing given all marginals in the table above.

Model II (Homogeneity testing). We have  $r$  multinomial trial sequences, called  $A_1, \dots, A_r$  respectively. In sequence  $A_i$  there are  $n_i = X_i$  trials and in each trial the probabilities of  $B_1, \dots, B_s$  are  $q_{i1}, \dots, q_{is}$ , respectively;  $\sum_j q_{ij} = 1$ . All  $n$  trials are independent. We find

$$\Pr\left[\bigcap_{i,j} (X_{ij} = x_{ij})\right] = \prod_{i=1}^r \frac{n_i!}{x_{i1}! \dots x_{is}!} q_{i1}^{x_{i1}} \dots q_{is}^{x_{is}} \quad (17)$$

when  $\{x_{ij}\}$  is a set of natural numbers with  $\sum_j x_{ij} = n_i$  for each  $i$ ). The null hypothesis is to the effect that the  $r$  trial sequences are identical; i.e.

$$q_{1j} = \dots = q_{rj} \quad \text{for all } j \quad (18)$$

From (17) and (18) we get, with  $q_{ij} = q_j$ ,

$$dP = (sq_s)^n e^{\sum_{j=1}^{s-1} x_{.j} \log \frac{q_j}{q_s}} dP_0 \quad (19)$$

where  $P$  and  $P_0$  are the joint probability distributions of all  $X_{ij}$  respectively when  $q_{ij} = q_j$  and  $q_{ij} = \frac{1}{s}$ . It is seen that the assumptions of theorems 1 and 2 are true with  $Y = (X_{.1}, \dots, X_{.s-1})$ . Thus we are again led to conditional testing given the marginals.

We shall find the conditional distributions encountered above under models I and II respectively. Under the null hypothesis, we have for the joint distribution of the marginals, in the case of model II,

$$\Pr\left[\bigcap_j (X_{.j} = x_{.j})\right] = \frac{n!}{x_{.1}! \dots x_{.s}!} q_1^{x_{.1}} \dots q_s^{x_{.s}} \quad (20)$$

Hence, dividing (17) (with  $q_{ij} = q_j$ ) by (20) we get for the conditional elementary probability function

$$h(x) = \frac{\prod_{i=1}^r \frac{n_i!}{x_{i1}! \dots x_{is}!}}{\frac{n!}{x_{.1}! \dots x_{.s}!}}$$

(when  $\{x_{ij}\}$  is a set of natural numbers with  $\sum_i x_{ij} = x_{.j}$  for all  $j$ ,  $\sum_j x_{ij} = n_i$  for all  $i$ ), which is the multivariable hypergeometric distribution.

In the case of model I, we first find the conditional distribution given the marginals  $(X_1, \dots, X_r) = (n_1, \dots, n_r)$  by dividing (14) by

$$\frac{n!}{x_{1.}! \dots x_{r.}!} p_1^{x_{1.}} \dots p_r^{x_{r.}}$$

We then obtain (17) with  $q_{ij} = q_j = p_j$ , i.e. model II is obtained by conditioning on  $X_1, \dots, X_r$ . Conditioning again with respect to  $X_{.1}, \dots, X_{.s}$ , we obtain, as shown above, the hypergeometric distribution (21). Thus both under model I and model II, the unbiasedness of the test  $\delta$  implies that

$$\sum_x \delta(x) h(x) = \epsilon \quad (22)$$

(for each possible set of marginals). The adjustment to level  $\epsilon$  should be made by means of the hypergeometric distribution.

#### D. Unbiased one-sided tests. Neyman-Pearsons fundamental theorem.

We shall consider how to find unbiased tests which are power optimal relatively to a certain alternative. The situation and notations are the same as those described in connection with theorem II. C.2.

Theorem II. D.1. We make the same assumptions as in theorem II.B.2. Let  $P \in \mathcal{P} - \mathcal{P}_0$  be a specific alternative and assume that  $P$  is absolutely continuous with respect to the distributions under  $\mathcal{P}_0'$ , such that  $dP = f(x)dP_0$ . Let  $\delta_0(x)$  be such that

$$\begin{aligned}\delta_0(x) &= 1 & \text{if } f(x) > c(Y(x)) \\ \delta_0(x) &= 0 & \text{if } f(x) < c(Y(x))\end{aligned}\tag{1}$$

$$E_0[\delta_0(X)|Y] = \epsilon\tag{2}$$

Then  $\delta_0$  is the most powerful test relatively to the alternative  $P$  among all similar tests with level  $\epsilon$ . If it is unbiased then it is also most powerful among all unbiased tests.

Proof: The last statement follows from the first and the fact that unbiasedness implies similarity (theorem II.C.1). Since similarity implies conditional testing (theorem II.C.2) we know that if  $\delta$  is similar, then

$$E_0[\delta(X)|Y] = \epsilon\tag{3}$$

Consider now the power of  $\delta$  under  $P$ .

$$E_P \delta(X) = E_0 \delta(X)f(X) = E_0 E_0[\delta(X)f(X)|Y]\tag{4}$$

It follows that we can maximize

$$E_0[\delta(X)f(X)|Y]\tag{5}$$

with respect to  $\delta$  under the side condition (3). By Neyman-Pearson's lemma we are then led to (1) as a sufficient condition for optimality. More precisely it follows (as in the proof of Neyman-Pearson's lemma) from (2), (3) and (4) that

$$E_P \delta_0(X) - E_P \delta(X) = E_0 \left[ (\delta_0(X) - \delta(X)) (f(X) - c(Y(X))) \right] \quad (6)$$

From (1) it now follows that the "integrand" on the right hand side is  $\geq 0$ . Hence  $E_P \delta_0(X) \geq E_P \delta(X)$ , Q.E.D.

Example 1. The components of  $X = (X_1, \dots, X_n, Z)$  are independent. All  $X_j$  are normal  $(0, \sigma^2)$  and  $Z/\sigma^2$  is eccentric chi-square distributed with  $m$  degrees of freedom and eccentricity  $\lambda$ .  $\lambda$  and  $\sigma$  are the unknown parameters. We want to test the nullhypothesis  $\lambda = 0$ .

A priori the joint density of  $X_1, \dots, X_n, Z$  can be written

$$(2\pi)^{-\frac{n}{2}} \sigma^{-n-2} e^{-\frac{\lambda}{2}} e^{-\frac{1}{2\sigma^2} \sum x_i^2} \gamma_m(z/\sigma^2) q_m(\lambda z/\sigma^2) \quad (7)$$

where  $\gamma_m$  is the central chi-square density with  $m$  degrees of freedom and

$$q_m(t) = \frac{\Gamma(\frac{m}{2})}{\Gamma(\frac{m}{2} + j)} \sum_{j=0}^{\infty} \frac{(\frac{t}{4})^j}{j! \Gamma(\frac{m}{2} + j)}$$

We let  $P_0$  be the probability distribution when  $\sigma = 1$ ,  $\lambda = 0$ . We then get for the distribution  $P$  of  $X$  under an arbitrary alternative

$$dP = e^{-\frac{\lambda}{2}} \sigma^{-m-n} e^{(\frac{1}{2} - \frac{1}{2\sigma^2})(\sum x_i^2 + z)} q_m(\lambda z/\sigma^2) dP_0 \quad (8)$$

Thus, if  $\lambda = 0$  we have a Darms-Koopman family of distributions with  $Y = \sum X_i^2 + Z$ . It is easily found in the general case that the powerfunction of a test must be a continuous function of  $\lambda$ . Let us now maximize the power relatively to  $(\lambda, \sigma)$  with  $\lambda > 0$ . From  $f(x) > c(Y(x))$  with  $f = dP/dP_0$  given by (8)

we get

$$\begin{aligned} q_m(\lambda Z/\sigma^2) &> e^{\frac{\lambda}{2}\sigma^{m+n}} e^{(\frac{1}{2\sigma^2} - \frac{1}{2})(\sum X_i^2 + Z)} c(\sum X_i^2 + Z) = \\ &= c_1(\sum X_i^2 + Z) \end{aligned}$$

Since  $q_m$  is increasing, this inequality could be written

$$Z > K(\sum X_i^2 + Z) \quad (9)$$

We choose  $\delta_0 = 1$  if and only if (9) is true, and  $\delta_0 = 0$  otherwise. The function  $K(y)$  must be adjusted such that the conditional level is  $\mathcal{E}$ , i.e. such that (2) is true. We then need the conditional distribution of  $Z$  given  $Y = \sum X_i^2 + Z$  when  $\lambda = 0, \sigma = 1$ . From the fact that  $Z$  and  $U = \sum X_j^2$  have joint density  $\gamma_m(z) \cdot \gamma_n(u)$ , we find the conditional density of  $Z$  given  $Y$  equal to

$$\frac{1}{B(\frac{m}{2}, \frac{n}{2})} \frac{1}{y} \left(\frac{z}{y}\right)^{\frac{m}{2}-1} \left(1 - \frac{z}{y}\right)^{\frac{n}{2}-1} \quad (0 \leq y \leq z) \quad (10)$$

Thus it is seen that the conditional density of  $W = Z/Y$  given  $Y$  is,

$$\beta_{m,n}(w) = \frac{1}{B(\frac{m}{2}, \frac{n}{2})} w^{\frac{m}{2}-1} (1-w)^{\frac{n}{2}-1} \quad (0 \leq w \leq 1) \quad (11)$$

i.e.  $W = Z/Y$  and  $Y$  are independent. From (2) we now get

$$\begin{aligned} \mathcal{E} = E_0[\delta_0(X)|Y=y] &= P_0(Z > K(Y)|Y=y) = P_0\left(\frac{Z}{Y} > \frac{K(Y)}{y} | Y=y\right) = \\ &= P_0(W > \frac{K(y)}{y}) \end{aligned} \quad (12)$$

Thus  $K(y)/y = c$  where  $\int_0^1 \beta_{m,n}(w) dw = \mathcal{E}$ . We reject the

hypothesis if  $Z > c(\sum X_i^2 + Z)$ , i.e.

$$F = \frac{Z}{\sum X_i^2} \frac{n}{m} > \frac{n}{m} \frac{c}{1-c} = f_{m,n} \quad (13)$$

where  $f_{m,n}$  is the  $1-\alpha$  fractile of the Fisher distribution with  $m$  and  $n$  degrees of freedom. The test is independent of the alternative from which we started. Hence  $E_P \delta_0(X) \geq E_P \delta(X)$  for any  $P \notin \mathcal{S}_0$  and for any  $\delta(X)$  which is unbiased with level  $\alpha$ . Letting in particular  $\delta(x) \equiv \alpha$ , we get  $E_P \delta_0(X) \geq \alpha$ . Thus  $\delta_0$  is unbiased and it is the uniformly most powerful among all unbiased tests.

Remark: Under an arbitrary  $(\lambda, \sigma)$ , denote the distribution of  $F$  by  $K_{m,n}(f; \lambda)$ . (It is independent of  $\sigma$ .) Thus the power of  $\delta_0$  is

$$\beta = 1 - K_{m,n}(f_{m,n}; \lambda)$$

where

$$K_{m,n}(f_{m,n}; 0) = 1 - \alpha$$

Consider now a test  $\delta_0'$  with rejection region

$$\frac{n'}{m} Z / \sum_{i=1}^{n'} X_i^2 \geq f_{m,n'}$$

where  $n' < n$ . The power  $\beta'$  of this test is less than  $\beta$ , according to what we have just proved. From this we get the following inequality for the eccentric Fisher distribution

$$K_{m,n'}(f_{m,n'}; \lambda) > K_{m,n}(f_{m,n}; \lambda) \quad (14)$$

if  $n' < n$ . This is a useful inequality when discussing designs

of experiments.

We shall now consider Darmais-Koopman alternatives and obtain the very famous result by Neyman and Pearson (1933).

Theorem II. D.2. Neyman-Pearson's fundamental theorem

Let it be a priori known that  $X$  has a distribution  $P_{\tau, \rho}$  given by

$$dP_{\tau, \rho} = A(\tau, \rho) e^{\sum_{j=1}^s \tau_j Y_j(x) + \rho V(x)} dP_0 \quad (15)$$

where  $\tau$  varies in  $\omega$ , and  $\rho$  varies in an interval  $[A, B]$  where  $A \leq 0$  and  $B > 0$ . Assume that  $\tau = 0$  is an inner point of  $\omega$ . For testing the null-hypothesis  $\rho \leq 0$ , there is a uniformly most powerful unbiased level  $\epsilon$  test  $\delta_0(X) = \psi(V(X), Y(X))$ ; where

$$\begin{aligned} \psi(v, y) &= 1 & \text{if } v > c(y) \\ \psi(v, y) &= \gamma(y) & \text{if } v = c(y) \\ \psi(v, y) &= 0 & \text{if } v < c(y) \end{aligned} \quad (16)$$

and where  $c(y)$  and  $\gamma(y)$  are given by  $E_0[\psi(V, Y)|Y] = \epsilon$ .

Hence

$$1 - F(c(y)|y) + \gamma(y)[F(c(y)|y) - F(c(y) - 0|y)] = \epsilon \quad (17)$$

where  $F(v|y) = P_0(V \leq v|Y = y)$ .

The test is also uniformly least powerful relatively to values of  $\rho < 0$  among all unbiased level  $\epsilon$  tests, i.e. the chance of false rejection is less than for any other unbiased test.

Proof: Consider first testing  $\rho = 0, \tau \in \omega$  against a particular alternative  $(\rho_1, \tau_1)$ , where  $\rho_1 > 0$  and  $\tau_1 \in \omega$ . By applying (1) we then get the rejection region

$$e^{\rho_1 V} > C(Y(X))A(\tau_1, \rho_1)^{-1} e^{-\sum_{j=1}^S \tau_{j1} Y_j(X)} = C'(Y(X)) ,$$

which is equivalent to  $V > c(Y(X))$  . Hence we obtain from theorem 1, equations (16) and (17). The existence of such a test  $\delta_0$  , i.e. of a  $c(y)$  and a  $\gamma(y)$  , is easily seen from equation (17). Since the test  $\delta_0$  is independent of  $(\rho_1, \tau_1)$  it is the uniformly most powerful test among all similar tests. Hence  $E_{\rho, \tau} \delta_0(X) \geq E_{\rho, \tau} \delta(X)$  for all tests for which  $E_{0, \tau} \delta(X) = \epsilon$ . In particular if  $\delta(X)$  is the trivial test  $\delta(X) \equiv \epsilon$  , then (since it is similar)  $E_{\rho, \tau} \delta_0(X) \geq E_{0, \tau} \epsilon = \epsilon$  . Thus  $\delta_0$  is unbiased. Since furthermore any unbiased  $\delta$  is similar, we have proved that  $\delta_0$  is the uniformly most powerful unbiased test for the hypothesis  $\rho = 0$  .

Consider now minimizing  $E_{\rho, \tau} \delta(X)$  for  $\rho < 0$  subject to similarity, i.e.  $E_{0, \tau} \delta(X) = \epsilon$  . This is the same as maximizing  $E_{\rho, \tau} (1 - \delta(X))$  with respect to  $1 - \delta$  subject to  $E_{0, \tau} (1 - \delta(X)) = 1 - \epsilon$  . By means of theorem 1, we obtain in the same manner as above that  $1 - \delta_0 = 1$  , i.e.  $\delta_0 = 0$  , if  $e^{\rho V} > C'(Y)$  , i.e.  $V < c(Y)$  , etc., where  $E_0(1 - \delta_0 | Y) = 1 - \epsilon$  , i.e.  $E_0[\delta_0 | Y] = \epsilon$  . Hence we have precisely the same  $\delta_0$  as above. This  $\delta_0$  is the uniformly least powerful for  $\rho < 0$  among all similar tests. Hence comparing again  $\delta_0$  with the trivial test  $\delta \equiv \epsilon$  , we get  $E_{\rho, \tau} \delta_0(X) \leq \epsilon$  for  $\rho \leq 0$  . Thus  $\delta_0$  really has level  $\epsilon$  relatively to the null-hypothesis  $\rho \leq 0$  .

(The measurability of  $c(y)$  and  $\gamma(y)$  as defined by (17) needs proof. We leave out this annoying detail, since it is obvious in each particular application.)

Example 2. Consider the example of II.C with model I and  $r = s = 2$  . Thus we have a multinomial trial sequence with

two factors, A and B, each with two levels, A, A\* and B, B\* respectively. The definition of the relevant probabilities and the observed result of n trials can be summarized in the following tables

	B	B*		B	B*	
A	p <sub>11</sub>	p <sub>12</sub>	A	V	M-V	M
A*	p <sub>21</sub>	p <sub>22</sub>	A*	W	N-W	N
				L	n-L	n

Thus  $L = V+W$ ,  $M+N = n$ . The joint elementary probability function of  $V$ ,  $M-V$ ,  $W$ ,  $N-W$  is

$$\frac{n!}{V!(M-V)!W!(N-W)!} p_{11}^V p_{12}^{M-V} p_{21}^W p_{22}^{N-W} \quad (18)$$

We shall test independence, i.e. the null hypothesis

$$p_{11} = (p_{11}+p_{12})(p_{11}+p_{21}) \quad (19)$$

which is equivalent to

$$p_{11} p_{22} = p_{12} p_{21} \quad (19)'$$

against the one sided alternative that there is "positive" dependence between A and B,

$$p_{11} > (p_{11}+p_{12})(p_{11}+p_{21}) \quad (20)$$

i.e. in the long

run A and B occur together more often than would be the case if they were independent. (20) is equivalent to

$$p_{11} p_{22} > p_{12} p_{21} \quad (20)'$$

Now, denote the measure corresponding to (18) by  $P$ , and the measure corresponding to  $p_{11} = p_{12} = p_{21} = p_{22} = \frac{1}{4}$  by  $P_0$ .

Then (13) can be written

$$dP = (4p_{22})^n e^{M\tilde{\gamma}_1 + L\tilde{\gamma}_2 + V\mathcal{J}} dP_0 \quad (21)$$

where

$$\tilde{\gamma}_1 = \log \frac{p_{12}}{p_{22}}, \quad \tilde{\gamma}_2 = \log \frac{p_{21}}{p_{22}}, \quad \mathcal{J} = \log \frac{p_{11} p_{22}}{p_{12} p_{21}} \quad (22)$$

Thus, we shall test  $\mathcal{J} = 0$  against  $\mathcal{J} > 0$ , and according to theorem 1 we shall condition with respect to  $(M, L)$ . We are led to rejecting when

$$V = (\text{number of } A \cap B) > c \quad (23)$$

and rejection with probability  $\Upsilon$  if  $V = c$ , where

$$\sum_{c \neq 1}^M f(v) + \Upsilon f(c) = \xi \quad (24)$$

and

$$f(v) = \frac{\binom{M}{v} \binom{n-M}{L-v}}{\binom{n}{L}} \quad (25)$$

The test is the uniformly most powerful unbiased test. This is also true about the same test in the case of homogeneity testing, where there are two binomial trial sequences with  $M$  and  $N$  trials, respectively.

Example 3.  $X_1, X_2, \dots, X_n$  are independent normal  $(\xi, \sigma)$ . We shall test

$$\xi \leq K\sigma^2 \quad (26)$$

against  $\xi > K\sigma^2$ . (See the example of section B.) We have

$$dP = Ae^{\lambda \sum x_i^2 + \rho \bar{x}_i} dP_0 \quad (27)$$

where  $P$  corresponds to a general  $(\xi, \sigma)$ ,  $P_0$  corresponds to  $(\xi, \sigma) = (K, 1)$ ,  $\lambda = \frac{1}{2} - \frac{1}{2\sigma^2}$ ,  $\rho = \frac{\xi}{\sigma^2} - K$ . Hence we shall test  $\rho \leq 0$  against  $\rho > 0$ . We are led to conditional testing given  $Y = \sum X_i^2$  and we obtain the test developed in the example in section A. Thus the conditional Student test given  $\sum X_i^2$  is the uniformly most powerful unbiased test. In particular this is the case when  $K = 0$ , in which case we get the ordinary Student test, since  $T$  and  $\sum X_i^2$  are independent when  $\xi = 0$ .

Example 4.  $X_1, \dots, X_n$  are independent normal  $(\xi, \sigma)$ . We shall test  $\sigma \leq \sigma_0$  against  $\sigma > \sigma_0$ . We find

$$dP = Ae^{\lambda \sum x_i^2 + \rho \sum x_i} dP_0 \quad (28)$$

where  $P$  is the measure corresponding to an arbitrary  $(\xi, \sigma)$  and  $P_0$  corresponds to  $(\xi, \sigma) = (0, \sigma_0)$ . Furthermore  $\rho = \frac{1}{2\sigma_0^2} - \frac{1}{2\sigma^2}$ ,  $\lambda = \xi/\sigma^2$ . By theorem 2 we shall reject if and only if

$$\sum X_i^2 > c(\sum X_i) \quad (29)$$

where  $c(y)$  should be determined such that the conditional probability of (29) given  $\sum X_i = y$  is  $\alpha$ . Subtracting  $\frac{1}{n}(\sum X_i)^2$  on both sides of (29) we obtain

$$\sum (X_i - \bar{X})^2 > c_1(\sum X_i)$$

Now, since  $\sum (X_i - \bar{X})^2$  and  $\sum X_i$  are independent we conclude that  $c_1$  does not depend on  $\sum X_i$  and that we can adjust  $c_1$

to an unconditional level  $\Sigma$ , hence  $c_1 = z\sigma_0^2$ , where  $z$  is the  $1-\Sigma$  fractile of the chi-square distribution with  $n-1$  degrees of freedom. Hence the test with rejection region

$$\sum (X_i - \bar{X})^2 > z\sigma_0^2 \quad (30)$$

is the uniformly most powerful unbiased test.

We have seen in several examples that it was possible to change a conditional test into an unconditional one by transforming the variable  $V$  into a new variable which is independent of  $Y$ . The advantage of this was obvious. We shall now state and prove a theorem which is useful when trying to find such transformations and which explains why such transformations often exist.

Theorem II.D.3 (Basu). Let  $\mathfrak{P} = \{P\}$  be a class of probability measures and let  $Y$  be a sufficient and complete statistic relatively to  $\mathfrak{P}$  (i.e.  $\mathfrak{P}Y^{-1} = \{PY^{-1}\}$  is a complete class of measures). Suppose that the distribution of  $W$  is the same for all  $P \in \mathfrak{P}$ . Then  $Y$  and  $W$  are stochastically independent for all  $P \in \mathfrak{P}$ .

Proof: Let  $f(W)$  be an arbitrary function for which  $E_P f(W)$  exists for all  $P \in \mathfrak{P}$ . We then have for all  $P$  and  $P_0$  in  $\mathfrak{P}$ .

$$c_f = E_P f(W) = E_P E_P [f(W)|Y] = E_P E_{P_0} [f(W)|Y]$$

since  $Y$  is sufficient. However,  $c_f$  is independent of  $P$  since the distribution of  $W$  is independent of  $P$ . From

$$E_P \left\{ E_{P_0} [f(W) | Y] - c_f \right\} = 0$$

it then follows that

$$E_{P_c} [f(W) | Y] = c_f$$

a.e. In particular, we then get if  $f$  is the indicator function of a set  $D$  in the  $W$ -space,

$$c_{I_D} = P_0 W^{-1}(D|Y) = \Pr(W \in D | Y)$$

But this proves the theorem, since  $c_{I_D}$  does not depend on  $Y$ . Q.E.D.

Examples. Let  $X_1, \dots, X_n$  be independent normal  $(\xi, \sigma)$ . Consider the class of distributions for  $(X_1, \dots, X_n)$  obtained by varying  $\xi$  and keeping  $\sigma$  fixed. In that case  $\bar{X} = \frac{\sum X_i}{n}$  is seen to be complete and sufficient. But  $\sum (X_j - \bar{X})^2$  has a distribution which is independent of  $\xi$ , since we can write  $\sum (X_j - \bar{X})^2 = \sum (Y_j - \bar{Y})^2$ , where  $Y_j = X_j - \xi$  is normal  $(0, \sigma)$ . Hence  $\bar{X}$ ,  $\sum (X_j - \bar{X})^2$  are independent.

Suppose now that  $\xi = 0$  and consider the class obtained by varying  $\sigma$ . In that case  $\sum X_i^2$  is complete and sufficient. On the other hand  $T = \frac{\bar{X} \sqrt{n}}{S} = \frac{\bar{X}}{\sqrt{\sum (X_i - \bar{X})^2}} \sqrt{n(n-1)} = \frac{\bar{Y} \sqrt{n(n-1)}}{\sqrt{\sum (Y_i - \bar{Y})^2}}$ , where  $Y_i = X_i/\sigma$  is normal  $(0, 1)$ . Hence the distribution of  $T$  is independent of  $\sigma$  and it follows that  $T$  and  $\sum X_i^2$  are independent if  $\xi = 0$ .

Let  $Z_1/\sigma^2$  and  $Z_2/\sigma^2$  be independent and chi-square distributed with  $n_1$  and  $n_2$  degrees of freedom. Consider the class of distributions of  $(Z_1, Z_2)$  obtained by varying  $\sigma$ . It is seen that  $Z_1 + Z_2$  is complete and sufficient, where as

$F = \frac{Z_1}{Z_2} \frac{n_2}{n_1}$  has a distribution which is independent of  $\sigma$ .

Hence Fisher's statistic and  $\frac{Z_1 + Z_2}{n_1}$ , the sum of the numerator and denominator in  $F \frac{n_1}{n_2}$ , are independent.

The independence of a regression coefficient and the residual sum square can also be proved by means of Basu's theorem.

Furthermore in the binormal case, the empirical correlation coefficient is independent of the means and empirical variances of the two variables if the theoretical correlation coefficient is 0.

The way of applying Basu's theorem to situations described in theorem 2, is as follows.

Suppose that we obtain a rejection region of the form  $V \geq c(Y)$  (non-randomized). Then one should try to find a  $W = T(V, Y)$  which, (i) for each  $Y$  is an increasing function of  $V$ , (ii) has a distribution independent of  $\tau$  when  $\varphi = 0$ . In that case the uniformly most powerful unbiased test takes the form  $W \geq c_1$  where  $c_1$  is such that  $P_0(W \geq c_1) = \epsilon$ .

#### E. The general Student hypothesis

(i) Statement of the result to be proved. We shall apply theorem II.D.2 to show that in linear normal situations the one-sided Student tests are uniformly most powerful unbiased tests.

Assume that the observations  $X_1, X_2, \dots, X_n$  are independent normal with variance  $\sigma^2$  and

$$\xi_i = EX_i = \sum_{j=1}^p a_{ij} \beta_j \quad (p \leq n) \quad (1)$$

where

$$\sum_{j=1}^p \beta_j c_{jk} = 0 \quad ; \quad k = 1, 2, \dots, r < p \quad (2)$$

and we want to test

$$\sum_{j=1}^p \beta_j c_j \leq 0 \quad \text{against} \quad \sum_{j=1}^p \beta_j c_j > 0 \quad (3)$$

We write also  $c_j = c_{jr+1}$  and assume that  $a = \{a_{ij}\}$  and  $\{c_{jl}\}_{j=1,2,\dots,p; l=1,2,\dots,r+1}$  are of full rank.  $\sigma, \beta_1, \dots, \beta_p$  are the unknown parameters.  $a_{ij}, c_{jk}$  are known.

We shall include the case when there are no relations (2) ( $r=0$ ). Note that we have included the situation where  $\sum_j a_{ij} \xi_j = 0$  ;  $i = 1, 2, \dots, p$  ; and we want to test  $\sum_i \xi_i \leq 0$  . We then only set  $a = I$  , i.e.  $EX_i = \xi_i = \beta_i$  .

Typical situations are such as the two sample Student hypothesis, the testing that a main treatments effect  $\alpha_1 < 0$ , etc.

The standard test applied in such situations could conveniently be described as follows. We find  $\hat{\beta}_1, \dots, \hat{\beta}_p$  by minimizing

$$\sum_i (X_i - \sum_{j=1}^p a_{ij} \beta_j)^2 + 2 \sum_{k=1}^r \lambda_k \sum_{j=1}^p \beta_j c_{jk} \quad (4)$$

w.r.t.  $\beta_1, \dots, \beta_p, \lambda_1, \dots, \lambda_r$  . The estimator  $S$  of  $\sigma$  is now given by

$$S^2 = \frac{1}{n-p+r} \sum_i (X_i - \sum_{j=1}^p a_{ij} \hat{\beta}_j)^2 \quad (5)$$

$\hat{\beta}_1, \dots, \hat{\beta}_p$  are known to be linear functions of  $X_1, \dots, X_n$ , hence

$$\text{var } \hat{\beta}_1 = K_j \sigma^2, \quad (6)$$

defining  $K_j$  . The Student test with level  $\epsilon$  now consists in stating  $\sum \beta_j c_j > 0$  if and only if

$$\sum_{j=1}^p \hat{\beta}_j c_j / \sqrt{c_j^2 K_j} S > t \quad (7)$$

where  $t$  is the  $1-\alpha$  fractile of the Student distribution with  $\mu - p + r$  degrees of freedom. We shall show that this test has the stated optimum property.

(ii) Reformulation of the result to be proved. Let

$(c_{1k}, \dots, c_{pk})$  ;  $k = r+2, \dots, p$  be such that  $(c_{1k}, \dots, c_{pk})$  ;  $k = 1, 2, \dots, p$  are linearly independent. Define

$$\tilde{\beta}_k = \sum_{j=1}^p \beta_j c_{jk} \quad ; \quad k = 1, 2, \dots, p \quad (8)$$

Thus  $\tilde{\beta}_k = 0$  ;  $k = 1, 2, \dots, r$  and we shall test

$$\tilde{\beta}_{r+1} \leq 0 \quad \text{against} \quad \tilde{\beta}_{r+1} > 0 \quad (9)$$

We write (1) and (8) ,  $\xi = EX = a\beta$  and  $\tilde{\beta} = c'\beta$  , respectively, and get

$$\xi = EX = a(c')^{-1} \tilde{\beta} = \sum_{j=r+1}^p g_j \tilde{\beta}_j$$

where we have introduced  $g_{r+1}, \dots, g_p$  . We now see that (7) could be written

$$\hat{\tilde{\beta}}_{r+1} / AS > t$$

where  $\hat{\tilde{\beta}}_{r+1}$  is the least square estimator of  $\tilde{\beta}_{r+1}$  and  $\text{var } \hat{\tilde{\beta}}_{r+1} = A^2 \sigma^2$  .

Changing the notations and changing the meaning of  $\beta_j$  we write the above

$$\xi = EX = g\beta = \sum_{j=1}^q g_j \beta_j \quad , \quad (10)$$

where we shall test

$$\beta_1 \leq 0 \quad \text{against} \quad \beta_1 > 0 \quad , \quad (11)$$

and where  $q = p - r$  ,  $g_1, \dots, g_q$  are  $n$ -dimensional vectors and  $X$  is an  $n$ -dimensional vector with components  $X_1, \dots, X_n$  .

The test criterion (7) may now be written

$$T = \hat{\beta}_1 / AS > t \quad (12)$$

where  $\hat{\beta}_1$  is the least square estimator of  $\beta_1$ ,

$$\text{var } \hat{\beta}_1 = A^2 \sigma^2 \quad (13)$$

and  $t$  is the  $1-\epsilon$  fractile of the Student distribution with  $n-q$  degrees of freedom. It is well known that  $A^2$  is the leading element in the matrix  $(g'g)^{-1}$ .

(iii) Reduction of the situation to canonical form.

Let now  $H$  denote the "boundary" hypothesis  $\beta_1 = 0$  and let  $Q_a$  and  $Q_H$  be respectively the minimum of  $Q = \sum (X_j - \xi_j)^2$  a priori and under  $H$ . We then know (see Erling Sverdrup: "Laws and chance variations" vol. II, p.211-215 eq. 7.23 and 7.27 ; "Lov og tilfeldighet" bind II p.191-195, lign (23) og (27)) that

$$Q_H - Q_a = \hat{\beta}_1^2 / A^2 \quad (14)$$

$$Q_a = (n-q)S^2 = \sum_{j=1}^n (X_j - \sum_{j=1}^r g_{ij} \hat{\beta}_j)^2 \quad (15)$$

$$T^2 = (n-q)(Q_H - Q_a) / Q_a \quad (16)$$

where  $\{g_{ij}\} = g$ .

Let now  $c$  be such that  $c'g'gc = I$  and introduce  $\gamma = c^{-1}\beta$ ,  $gc = b$ . Then  $\gamma$  has least square estimator  $\hat{\gamma} = c^{-1}\hat{\beta}$ . We get

$$EX = \sum g\beta = bc^{-1}\beta = b\gamma = \sum_{j=1}^q b_j \gamma_j \quad (17)$$

$$b'b = I, \text{ i.e., } b_j' b_j = \delta_{ij} \quad (18)$$

where the definition of  $b_j$  is obvious. From  $\beta = c\gamma$  it is seen that the hypothesis (11) can now be written

$$\beta_1 = \sum_{j=1}^q c_{1j} \gamma_j \leq 0 \quad (19)$$

We now define  $b_{q+1}, \dots, b_n$  such that  $B = (b_1, \dots, b_n)$  is orthogonal and we introduce  $Y$  by

$$X = BY = \sum_{i=1}^n b_i Y_i \quad (20)$$

We find from (20) and (17) that

$$EY = B'EX = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_q \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

i.e.

$$EY_i = \gamma_i ; \quad i = 1, 2, \dots, q ; \quad EY_i = 0 ; \quad i = q+1, \dots, n \quad (21)$$

Finally we transform from  $Y$  to  $Z$  orthogonally by means of

$$Z_i = \sum_{j=1}^q e_{ij} Y_j ; \quad j = 1, 2, \dots, q$$

$$Z_i = Y_i ; \quad j = q+1, \dots, n$$

where  $e_{1j} = c_{1j}$  (and of course  $\{e_{ij}\}$  is orthogonal) . We introduce  $EZ_i = \zeta_i$  and obtain the following canonical form of our situation

$$\begin{aligned} & \underline{Z_1, \dots, Z_n \text{ are independent normal, } \text{var } Z_i = \sigma^2, EZ = \zeta_i,} \\ & \underline{i = 1, 2, \dots, q ; EZ_i = 0 ; i = q+1, \dots, n . \text{ To test that}} \\ & \underline{\zeta_1 \leq 0 \text{ against } \zeta_1 > 0 .} \end{aligned}$$

(iv) Construction of the optimum test. Relatively to

$Z_1, \dots, Z_n$  , let now  $P$  denote the probability measure a priori and  $P_0$  the probability measure when  $\sigma = 1$  ,  $\zeta_1 = \dots = \zeta_q = 0$  .

We then have

$$dP = \exp \left[ \left( \frac{1}{2} - \frac{1}{2\sigma^2} \right) \sum_{j=1}^n Z_j^2 + \sum_{j=1}^q \frac{\zeta_j}{\sigma^2} Z_j \right] dP_0 \quad (22)$$

From the fundamental Neyman-Pearson theorem (II.D.2) we see that the uniformly most powerful unbiased test of  $\zeta_1 \leq 0$  consists in rejecting whenever  $Z_1 > c(Y)$ , where

$$P_0[Z_1 > c(Y) | Y] = \epsilon \quad (23)$$

and where

$$Y = \left( \sum_{j=1}^n Z_j^2, Z_2, \dots, Z_q \right)$$

This  $Y$  is in a one-to-one correspondence with

$$Y' = \left( Z_1^2 + \sum_{j=q+1}^n Z_j^2, Z_2, \dots, Z_q \right)$$

Hence  $Y$  may be replaced by  $Y'$  in (23). Now it is seen that  $(Z_2, \dots, Z_q)$  is independent of  $Z_1$  and

$$Y_0 = Z_1^2 + \sum_{j=q+1}^n Z_j^2 = Z_1^2 + U \quad (24)$$

Hence the conditioning in (23) could be made with respect to  $Y_0$  and  $c(Y)$  would depend on  $Y'$  only through  $Y_0$ . Thus  $Y$  may be replaced by  $Y_0$  in (23). We introduce  $W = Z_1/\sqrt{Y_0}$  and the test amounts to rejection whenever  $W > K(Y_0) = c(Y_0)/\sqrt{Y_0}$  where  $K(Y_0)$  is given by

$$P_0(W > K(Y_0) | Y_0) = \epsilon \quad (25)$$

Let us now consider the situation for any specification consistent with  $H$ . Then  $W$  has a distribution which is independent of all the parameters  $\zeta_2, \dots, \zeta_q, \sigma$ , where as  $(Y_0, Z_2, \dots, Z_q)$  is a sufficient and complete set of statistics.

By Basu's theorem (II.D.3) we then know that  $W$  and  $(Y_0, Z_2, \dots, Z_q)$  are independent, hence  $W$  and  $Y_0$  are independent. It follows that  $K(Y_0) = K$  independent of  $Y_0$  is determined by  $P_0(W > K) = \epsilon$ . Introducing

$$T_0 = Z_1 \sqrt{n-q} / \sqrt{U}, \quad U = \sum_{j=1}^n Z_j^2, \quad (26)$$

this is equivalent to  $P_0(T_0 > t) = \epsilon$ . Hence we shall reject when  $T_0 > t$  where  $t$  is the  $1-\epsilon$  fractile of the Student distribution with  $n-q$  degrees of freedom.

This completes the construction of the optimum test.

(v) Identification of the optimum constructed test with the standard test. The standard test was described in (i), eq. (7) and was proved to be equivalent to (12) in (iii). Thus we have to show that our optimum test is equivalent to (12), where  $T$  is also given by (16).

Now we have from the canonical form in (iii),

$$Q_a = \sum_{j=1}^n Z_j^2 = U, \quad Q_H - Q_a = Z_1^2 \quad (27)$$

Comparing with (14) and (15) we obtain

$$U = (n-q)S^2 \quad (28)$$

and  $Z_1 = \pm \hat{\beta}_1 / A$ . But then  $\zeta_1 = \pm \beta_1 / A$ . Now  $\zeta_1 \leq 0$  is equivalent to  $\beta_1 \leq 0$ , hence we must have + sign, i.e.

$$Z_1 = \hat{\beta}_1 / A \quad (29)$$

Introducing (28) and (29) in (26), we see that (26) is equivalent to (12).

Q.E.D.

F. Performance unbiased three-decision tests.

We have above only considered situations where there is a choice between two decisions, either rejecting the null hypothesis or saying nothing. We shall below consider an important type of three-decision problems. However, first we shall make some remarks about a general type of decision situations.

We shall consider a situation where it is a priori known that the observed random variable  $X$  has a distribution

$P$  which is contained in a class  $\mathcal{P}$  of distributions. The purpose of the investigation is defined by means of a decision space  $D$ , where the points  $d$  in  $D$  are the possible results of the investigation. Very often the decisions  $d$  could be identified with a statement  $P \in \mathcal{P}_d$  where  $\mathcal{P}_d \subset \mathcal{P}$ . This is the case in classical test problems where  $\mathcal{P}_d = \mathcal{P} - \mathcal{P}_0$  when  $d =$  "reject", where as  $\mathcal{P}_d = \mathcal{P}$  if  $d =$  "do not reject". Note that the different  $\mathcal{P}_d$  may be overlapping and one  $\mathcal{P}_d$  may even be a subset of another.

In order to define a randomized statistical procedure, we introduce a sigmafield  $\mathcal{A}_D$  in  $D$ , which contains all one-point sets  $\{d\}$ . An arbitrary randomized procedure  $\psi(\bar{D}|X)$  gives the conditional probability of choosing a  $d \in \bar{D}$  given  $X$ , for all  $\bar{D} \in \mathcal{A}_D$ . (Of course,  $\psi$  is measurable as a function of  $X$  for each  $\bar{D}$ , and a measure as a function of  $\bar{D}$  for each  $X$ .) The unconditional probability of choosing a  $d$  in  $\bar{D}$  is now

$$\beta(\bar{D}; P, \psi) = \int \psi(\bar{D}|x) dP \quad (1)$$

This is the performance function. The effort of a statistician should be aimed at finding a  $\psi$  which makes  $\beta$  "nice".

When we are choosing a decision function  $\psi$ , we must consider the possible erroneous interpretations. These may be of two kinds. For any  $(P, d)$ , the decision  $d$  may, if it is erroneous, either be considered to be a false statement (error of type I) or an error by default (error of type II, "unnlatel-sessynd"). Thus error of type II implies failing to discover an interesting feature of  $P$ .

We make a false statement if we state  $d$ , when  $\mathcal{P}_d$  does not cover  $P$ . Let  $D^f(P)$  be the class of all  $d \in D$  which are false relative to  $P$ . The level of a procedure is

$\in$  if

$$P(D^f(P) | \psi) \leq \varepsilon \quad (2)$$

for all  $P$ . In order to define performance unbiasedness we would have to exclude  $d_0 = \text{"no statement"}$  (if there is such a  $d$  in  $D$ ). A procedure is performance unbiased, if

$$P(D - D^f(P) - \{d_0\} | \psi) \geq \varepsilon \quad (3)$$

for all  $P$ . This requirement gives a minimal safeguard against errors by default. One might try to exclude some other  $d$  besides  $d_0$  in (3). In particular one might consider "strictest statements" under  $P$ .  $d$  is a strictest statement under  $P$  if  $P \in \mathcal{P}_d$  and there is no  $d_1$  such that  $P \in \mathcal{P}_{d_1}$  and  $\mathcal{P}_{d_1} \subset \mathcal{P}_d$ . Under appropriate compactness properties about  $D$  this would be a statistically meaningful concept. Let  $D^S(P)$  be the class of strictest statements corresponding to  $P$ . It is clear that  $P(D^S(P) | \psi)$  should be large, but it would in most cases be asking too much to require that it should be  $\geq \varepsilon$ . (See, however, the three-decision problems treated below.)

An example of a procedure satisfying (2) is Scheffé's multiple comparison procedure.

Let us now consider a situation where we are interested in a parameter  $\mathcal{J} = \mathcal{J}(P)$ . We have a choice between the three decisions,

$$\begin{aligned} d_1 &= \text{state that } \mathcal{J} \leq 0; & d_2 &= \text{state that } \mathcal{J} \geq 0; \\ d_3 &= \text{make no inference.} \end{aligned}$$

We define in this case the statistical method  $\psi$  by  $\psi_i(X) =$  the conditional probability of choosing  $d_i$  given  $X$ ;  $i = 1, 2, 3$ . Thus

$$\sum_{i=1}^3 \psi_i(X) = 1 \quad (4)$$

The performance function of  $\psi$  is now

$$\Pr(\text{choosing } d_i) = E_P \psi_i(X) \quad (5)$$

considered as a function of  $i = 1, 2, 3$  and  $P \in \mathcal{P}$  for given  $\psi$ .

In this case it is natural to require,

Optimum requirement A.

(i). The level should be  $\varepsilon$ , i.e. the probabilities of falsely stating  $\mathcal{G} \leq 0$  and  $\mathcal{G} \geq 0$  should be at most  $\varepsilon$ .

$$\begin{aligned} E_P \psi_1(X) &\leq \varepsilon \text{ for } \mathcal{G}(P) > 0 \\ E_P \psi_2(X) &\leq \varepsilon \text{ for } \mathcal{G}(P) < 0 \end{aligned} \quad (6)$$

(ii). The probabilities of correct statements should be at least  $\varepsilon$ .

$$\begin{aligned} E_P \psi_1(X) &\geq \varepsilon \text{ for } \mathcal{G}(P) < 0 \\ E_P \psi_2(X) &\geq \varepsilon \text{ for } \mathcal{G}(P) > 0 \end{aligned} \quad (7)$$

This is the requirement of performance unbiasedness.

(iii). Among all  $\psi$  satisfying (4), (6) and (7) we want to find one  $\psi$  which maximizes  $E_P \psi_1(X)$  for  $\mathcal{G}(P) < 0$  and maximizes  $E_P \psi_2(X)$  for  $\mathcal{G}(P) > 0$ .

It is seen that if we disregard (4) we have really two separate problems of finding uniformly most powerful unbiased tests  $\psi_1$  and  $\psi_2$  respectively. If we solve these problems and (hopefully) obtain  $\psi_1 + \psi_2 \leq 1$  for all  $x$ , then we may set  $\psi_3 = 1 - \psi_1 - \psi_2$ , and we have obtained a procedure which uniformly maximizes the performance among all performance unbiased level  $\varepsilon$  procedures.

The above could be taken as an interpretation of what we are really interested in when we want to test if  $\theta$  is "significantly" different from 0.

Another well-known (classical) interpretation is the following. We want to decide if  $\theta \neq 0$  or not. Let the test  $\delta$  be defined by  $\delta(X)$  = the conditional probability of stating that  $\theta \neq 0$  given  $X$ . Then we want,

Optimum requirement B.

$$(i). E_P \delta(X) \leq \epsilon \text{ for } \theta(P) = 0$$

$$(ii). E_P \delta(X) \geq \epsilon \text{ for } \theta(P) \neq 0$$

(iii). We want to maximize  $E_P \delta(X)$  for  $\theta(P) \neq 0$  subject to (i) and (ii).

A method  $\delta$  could be compared with a method  $\psi$  since  $\psi_1 + \psi_2$  is the conditional probability of rejecting  $\theta = 0$  under  $\psi$ . Thus, any three-decision method  $\psi$  is also a test of significance with test function  $\delta = \psi_1 + \psi_2$ . On the other hand, any  $\delta$  combined with some "good" point estimator  $\hat{\theta}$  would give rise to a three-decision function  $\psi$ ; since it is "understood" that, provided  $\theta$  is significantly different from 0, then we should state that  $\theta < 0$  if  $\hat{\theta} < 0$  and that  $\theta > 0$  if  $\hat{\theta} > 0$ . Thus  $\psi_1(X) = \delta(X) I_{\hat{\theta} < 0}$ ,  $\psi_2(X) = \delta(X) I_{\hat{\theta} > 0}$ , where  $I_{\hat{\theta} < 0}$  and  $I_{\hat{\theta} > 0}$  are indicator functions for the sets  $(\hat{\theta} < 0)$  and  $(\hat{\theta} > 0)$ , respectively. (We assume here, for the sake of convenience, that  $\Pr(\hat{\theta} = 0) = 0$ ).

Under optimum requirement B we are interested in making

$$E_P \delta(X) = E_P \delta(X) I_{\hat{\theta} < 0} + E_P \delta(X) I_{\hat{\theta} > 0} \quad (8)$$

large. However, for  $\theta(P) < 0$  it is only the first term we want large, and for  $\theta(P) > 0$ , it is only the second term which

we want large. Thus, it seems, that in most circumstances the classical optimum requirement B is inadequate.

Example 1.  $X_1, \dots, X_n$  are independent normal  $(\xi, \sigma)$ . We want to decide if  $\xi < 0, \xi > 0$  or if no statement should be made. Regarding this as two separate testings, applied simultaneously, we obtain from example 3 of section II.C that we should state  $\xi < 0, \xi > 0$  or make no statement, according as

$$\bar{X} < -t_{1-\varepsilon} S/\sqrt{n}, \bar{X} > t_{1-\varepsilon} S/\sqrt{n}, |\bar{X}| \leq t_{1-\varepsilon} S/\sqrt{n} \quad (9)$$

where

$$\bar{X} = \frac{1}{n} \sum X_j \quad (n-1)S^2 = \sum (X_j - \bar{X})^2,$$

and  $t_{1-\varepsilon}$  is the  $1-\varepsilon$  fractile in the Student distribution with  $n-1$  degrees of freedom. This method has the uniformly largest performance among all performance unbiased methods with level  $\varepsilon$ . Rejecting  $\xi=0$  if  $|\bar{X}| > t_{1-\varepsilon} S/\sqrt{n}$  is the uniformly most powerful unbiased level  $2\varepsilon$  test, as we shall see later. Since  $\bar{X}$  is the "natural" estimate of  $\xi$ , we are led to the same method whether we use optimum requirement A or B above.

However, the following example shows that this is not always the case.

Example 2.  $X_1, \dots, X_n$  are independent normal  $(\xi, \sigma)$ . We want to decide if  $\sigma < \sigma_0, \sigma > \sigma_0$  or no statement should be made. We obtain as in example 1, by using the result of example 4 in section II.C, that we should state that  $\sigma < \sigma_0$  or  $\sigma > \sigma_0$  according as  $(n-1)S^2 < z_{\varepsilon/2}^2 \sigma_0^2$  or  $> z_{1-\varepsilon/2}^2 \sigma_0^2$ , where  $z_\alpha$  is the  $\alpha$ -fractile of the chi-square distribution with

$n-1$  degrees of freedom. Otherwise we should state nothing.

Hence we have again obtained a performance optimum test.

However, as we shall see below, the unbiased power optimum

test would consist in rejecting  $\sigma = \sigma_0$  if  $(n-1)S^2 < z' \sigma_0^2$

or  $> z'' \sigma_0^2$ , where  $z'$  and  $z''$  are determined by

$\int_{n-1}^{\infty} (z') + 1 - \int_{n-1}^{\infty} (z'') = 2\epsilon$ ,  $z' \gamma_{n-1}(z') = z'' \gamma_{n-1}(z'')$ . Here  $\gamma$

and  $\gamma$  are respectively the cumulative probability function

and the probability density of the chi-square distribution with

$\nu$  degrees of freedom. Now, combining this with the fact that

$S^2$  is an unbiased estimate of  $\sigma^2$ , we are led to stating

$\sigma < \sigma_0$  or  $\sigma > \sigma_0$  according as  $(n-1)S^2 < z' \sigma_0^2$  or  $> z'' \sigma_0^2$ .

This last method is based on the requirement that

$\Pr(\text{stating } \sigma < \sigma_0) + \Pr(\text{stating } \sigma > \sigma_0) \geq 2\epsilon$  and that this

expression should be maximized for  $\sigma \neq \sigma_0$ . However, as we

have pointed out above in a general context, if  $\sigma < \sigma_0$  it is

only the first term we want to maximize, and if  $\sigma > \sigma_0$  it is

only the last term we want to maximize.

#### G. Unbiased two-sided tests.

We shall consider the same family of distributions

as in section II.C, but we are now interested in testing  $\rho = 0$

against two-sided alternatives  $\rho \neq 0$ .

Theorem II.E.1. Let it be a priori known that  $X$  has a distribution  $P_{\tau, \rho}$  given by

$$dP_{\tau, \rho} = A(\tau, \rho) e^{\sum_{j=1}^s \tau_j Y_j(x) + \rho V(x)} dP_0 \quad (1)$$

where  $\tau \in \omega$  and  $\mathcal{S}$  belongs to an interval containing 0 as an inner point. Assume also that  $\tau = 0$  is an inner point of  $\omega$ . For testing the null hypothesis  $\mathcal{S} = 0$  against  $\mathcal{S} \neq 0$ , the uniformly most powerful unbiased level  $\mathcal{E}$  test is  $\delta_0(X) = \psi(V(X), Y(X))$ , where

$$\begin{aligned}\psi(v, y) &= 1 \quad \text{if } v < c_1(y) \text{ or } > c_2(y), \\ \psi(v, y) &= \gamma_i(y) \quad \text{if } v = c_i(y); i = 1, 2; \\ \psi(v, y) &= 0 \quad \text{if } c_1(y) < v < c_2(y),\end{aligned}\tag{2}$$

and where  $c_1(y), c_2(y), \gamma_1(y), \gamma_2(y)$  are determined such that

$$\begin{aligned}E_0[\delta_0(X)|Y] &= \mathcal{E}, \\ E_0[V(X)\delta_0(X)|Y] &= \mathcal{E} E_0[V(X)|Y],\end{aligned}\tag{3}$$

i.e. such that

$$\begin{aligned}F(c_1-0) + 1 - F(c_2) + \gamma_1[F(c_1) - F(c_1-0)] + \gamma_2[F(c_2) - F(c_2-0)] &= \mathcal{E}, \\ \int_{-\infty}^{c_1-0} v dF(v) + \int_{c_2+0}^{\infty} v dF(v) + \gamma_1 c_1[F(c_1) - F(c_1-0)] + \gamma_2 c_2[F(c_2) - F(c_2-0)] &= \\ \mathcal{E} \int_{-\infty}^{+\infty} v dF(v)\end{aligned}\tag{4}$$

Here  $F(v) = P_0(V \leq v | Y = y)$  and  $c_i = c_i(y), \gamma_i = \gamma_i(y)$ .

Proof: It can, in fact, be proved that we can always determine  $c_i(y), \gamma_i(y)$  from (4). Thus, there is always a uniformly most powerful test  $\delta_0$ . We shall content ourselves with proving that if a  $\delta_0$  can be found satisfying (4), then it is the uniformly most powerful unbiased level  $\mathcal{E}$  test. We first prove that any  $\delta$  which is unbiased with level  $\mathcal{E}$  must satisfy (3). However,

$$E_c[\delta(X)|Y] = \varepsilon \quad (5)$$

follows immediately from theorems II.B.1 and 2. As to the second equation (3), let us consider the power  $\beta = E_{\rho, \tau} \delta(X)$  of  $\delta$ . It is an analytic function of  $\rho$  (see theorem I.B.2). Hence it has continuous derivatives. Since  $\beta$  shall attain the minimum value  $\varepsilon$  for  $\rho = 0$ , we have  $\frac{\partial \beta}{\partial \rho} \Big|_{\rho=0} = 0$ . Furthermore, the derivative can be found by derivation under the integral sign in the integral expression for  $\beta$ . We get

$$\frac{\partial \beta}{\partial \rho} = \frac{\partial A}{\partial \rho} \int \delta(x) e^{\sum \tau_j Y_j(x) + \rho V(x)} dP_0 + A \int \delta(x) V(x) e^{\sum \tau_j Y_j(x) + \rho V(x)} dP_0$$

Hence for any  $\delta$

$$\frac{\partial \beta}{\partial \rho} \Big|_{\rho=0} = A_1(\tau) E_0(\delta(X) e^{\sum \tau_j Y_j(X)}) + A(\tau) E_0(\delta(X) V(X) e^{\sum \tau_j Y_j(X)}), \quad (6)$$

where  $A(\tau) = A(\tau, 0)$ ,  $A_1(\tau) = \frac{\partial A}{\partial \rho} \Big|_{\rho=0}$ . In particular for  $\delta(x) \equiv 1$ , since  $E_0 e^{\sum \tau_j Y_j(X)} = A(\tau)^{-1}$ , we get

$$A_1(\tau) = -A(\tau)^2 E_0(V(X) e^{\sum \tau_j Y_j(X)}), \quad (7)$$

Introducing this expression for  $A_1$  in (6), we find

$$\begin{aligned} \frac{\partial \beta}{\partial \rho} \Big|_{\rho=0} &= \\ &= A(\tau) \left[ E_0(\delta(X) V(X) e^{\sum \tau_j Y_j(X)}) - A(\tau) E_0(\delta(X) e^{\sum \tau_j Y_j(X)}) E_0(V(X) e^{\sum \tau_j Y_j(X)}) \right] \end{aligned} \quad (8)$$

When  $\delta$  is unbiased with level  $\varepsilon$  we can substitute

$\frac{\partial \beta}{\partial \rho} \Big|_{\rho=0} = 0$  and  $A(\tau) E_0(\delta(X) e^{\sum \tau_j Y_j(X)}) = \varepsilon$  in (8). We get

$$0 = E_0 \left[ (\delta(X) V(X) - \varepsilon V(X)) e^{\sum \tau_j Y_j(X)} \right]$$

Hence, multiplying by  $A(\tilde{\gamma})$

$$0 = E_{\tilde{\gamma}} [\delta(X)V(X) - E V(X)] = E_{\tilde{\gamma}} E_0 [(\delta(X)V(X) - E V(X)) | Y]$$

Since this is true for all  $\tilde{\gamma} \in \omega$  and the class  $\{P_{0,\tilde{\gamma}}\}_{\tilde{\gamma} \in \omega}$  is complete, we get

$$E_0 [V(X)\delta(X) | Y] = E E_0 [V(X) | Y], \quad (9)$$

which is the second equation (3) for an arbitrary unbiased  $\delta$ .

We shall now use the side conditions (5) and (9) instead of the original ones that  $\delta$  should be unbiased with level  $E$ . Maximizing the power under a particular alternative  $(\beta_1, \tilde{\gamma}_1), \beta_1 \neq 0$ ,

$$E_{\beta_1, \tilde{\gamma}_1} \delta(X) = A E_0 \left[ e^{\sum \tilde{\gamma}_{j1} Y_j(X)} E_0 (e^{\beta_1 V(X)} \delta(X) | Y) \right],$$

leads to maximizing

$$E_0 [e^{\beta_1 V(X)} \delta(X) | Y]$$

subject to (5) and (9). By the analogue to Neyman-Pearson's lemma, we then get

$$\begin{aligned} \delta_0 &= 1 & \text{if } e^{\beta_1 V} > a(Y)V + b(Y) \\ \delta_0 &= 0 & \text{if } e^{\beta_1 V} < a(Y)V + b(Y) \end{aligned} \quad (10)$$

where  $a(Y)$  and  $b(Y)$  should be determined such that (3) is fulfilled. More explicitly and precisely we get from (3), (5) and (9)

$$E_{\beta_1, \tilde{\gamma}_1} \delta_0(X) - E_{\beta_1, \tilde{\gamma}_1} \delta(X) = A E_0 e^{\sum \tilde{\gamma}_{j1} Y_j} (e^{\beta_1 V - a(Y)V - b(Y)})(\delta_0 - \delta) \quad (11)$$

It is seen from (10) that the integrand in (11) is non-negative,

hence  $E_{\mathcal{F}_1, \mathcal{Z}_1} \delta_0(X) \geq E_{\mathcal{F}_1, \mathcal{Z}_1} \delta(X)$ . Now, (10) is equivalent to (2). Assuming that  $c_i(y)$  and  $Y_i(y)$  could be determined such that (3) is true, we have a test  $\delta_0$  which is independent of  $(\mathcal{F}_1, \mathcal{Z}_1)$ . Thus  $\delta_0$  is uniformly most powerful among all tests satisfying (5) and (9), i.e.

$$E_{\mathcal{F}_1, \mathcal{Z}_1} \delta_0(X) \geq E_{\mathcal{F}_1, \mathcal{Z}_1} \delta(X) \quad (12)$$

if  $\delta$  satisfies (5) and (9). Substituting in particular  $\delta(X) \equiv \mathcal{E}$ , we conclude that  $E_{\mathcal{F}_1, \mathcal{Z}_1} \delta_0(X) \geq \mathcal{E}$ . Thus  $\delta_0$  is unbiased. Since the class of all unbiased tests is a subset of the class of all tests satisfying (5) and (9), the assertion in the theorem follows, Q.E.D.

Example 1. We consider again example 2 in section II.C where we wanted to test the dependence of two factors A and B, which each attained two levels. Using the same notations as in II.C we shall test the null-hypothesis  $p_{11} = (p_{11}+p_{12})(p_{11}+p_{21})$  against  $p_{11} \neq (p_{11}+p_{12})(p_{11}+p_{21})$ . By II.C.eq.(21) and (22) we can apply the theorem above, and we obtain that we shall reject independence if  $V =$  (the frequency of  $A \cap B$ ) is  $< c_1$  or  $> c_2$ . We shall reject with probability  $\gamma_i$  if  $V = c_i$ ;  $i = 1, 2$ . Otherwise we shall not reject that A and B are independent. Here  $c_i$  and  $\gamma_i$  depend on the marginals M and L, and are determined by (4), where

$$F(v) = \sum_{j=0}^v \binom{M}{j} \binom{n-M}{L-Y} / \binom{n}{L} \quad (13)$$

Example 2.  $(X_1, \dots, X_n)$  are independent normal  $(\xi, \sigma)$ . We shall test  $\xi = 0$  against  $\xi \neq 0$ . Using the theorem just proved and II.C.eq.(27) with  $K = 0$  we shall reject if and

only if  $\sum X_j \geq c_2(\sum X_j^2)$  or  $\leq c_1(\sum X_j^2)$ , where  $c_1(\ )$  and  $c_2(\ )$  are determined by (3), i.e.

$$\begin{aligned} E_0[\psi(V, Y) | Y] &= \xi \\ E_0[V\psi(V, Y) | Y] &= \xi E_0(V | Y) \end{aligned} \quad (14)$$

with  $V = \sum X_j$  and  $Y = \sum X_j^2$  as usual. Now, by Basu's theorem (or II. B)  $W = \frac{V}{\sqrt{Y}}$  and  $Y$  are independent if  $\xi = 0$ . Hence, introducing  $\psi(W, Y) = \phi(W)$  we get from the first equation (14)

$$\xi = E_0(\phi(W, Y) | Y = y) = E_0(\phi(W, y) | Y = y) = E_0\phi(W, y) \quad (15)$$

Dividing the last equation (14) by  $\sqrt{Y}$ , we get similarly

$$\xi E_0 W = E_0(\phi(W, y) W) \quad (16)$$

From these two equations it is seen that  $\phi(W, y) = \phi(W)$ , independently of  $y$ .  $\phi = \psi$  is either 0 or 1 according as  $W$  is inside or outside a certain interval. Hence (15) and (16) may be written

$$\begin{aligned} \int_{k_1}^{k_2} \mathcal{L}(w) dw &= 1 - \xi, \\ \int_{k_1}^{k_2} w \mathcal{L}(w) dw &= (1 - \xi) E_0 W, \end{aligned}$$

where  $\mathcal{L}$  is the density of  $W$ , given in II.A. It is seen that  $\mathcal{L}(w) = \mathcal{L}(-w)$  and  $E_0 W = 0$ . Hence, the last equation is satisfied when  $k_1 = -k_2 = -k$ , and the first equation reduces to  $\int_{-k}^k \mathcal{L}(w) dw = \frac{1}{2} \xi$ . Introducing

$$T = \frac{\bar{X}\sqrt{n}}{S} = \sqrt{\frac{n-1}{n}} \frac{W}{\sqrt{1 - \frac{W^2}{n}}}$$

which is an increasing function of  $W$ , it is seen that we have obtained the Student equal tailed test. This test is thus the uniformly most powerful unbiased test.

Example 3.  $X_1, \dots, X_n$  are independent normal  $(\xi, \sigma)$ . We want to test  $\sigma = \sigma_0$  against  $\sigma \neq \sigma_0$ . By means of II. D. eq. (28) and the theorem just proved, we obtain that the test function  $\psi$  should be 0 or 1 according as  $\sum X_i^2$  is inside or outside a certain interval depending on  $\sum X_i$ . This interval is determined by

$$E_0[\psi(\sum X_i^2, \sum X_i) | \sum X_i] = \varepsilon$$

$$E_0[\psi(\sum X_i^2, \sum X_i) \sum X_i^2 | \sum X_i] = \varepsilon E_0(\sum X_i^2 | \sum X_i)$$

We introduce  $Z = \sum (X_i - \bar{X})^2 = \sum X_i^2 - \frac{1}{n}(\sum X_i)^2$ , multiply the first equation by  $\frac{1}{n}(\sum X_i)^2$  and subtract from the last. We then get

$$E_0[\phi(Z, \sum X_i) | \sum X_i] = \varepsilon$$

$$E_0[\phi(Z, \sum X_i) Z | \sum X_i] = \varepsilon E_0(Z | \sum X_i)$$

where we have introduced  $\phi$  instead of  $\psi$ . Since  $Z$  is independent of  $\sum X_i$ , we can leave out the conditioning and write  $\phi(Z, \sum X_i) = \phi(Z)$ , where  $\phi$  is 0 or 1 according as  $Z/\sigma_0^2 \in (z', z'')$  or  $Z/\sigma_0^2 \notin (z', z'')$ , and  $z', z''$  are determined by

$$\int_{n-1}^{\infty} (z') + 1 - \int_{n-1}^{\infty} (z'') = \varepsilon \quad (15)$$

$$\int_0^{z'} z \gamma_{n-1}(z) dz + \int_{z''}^{\infty} z \gamma_{n-1}(z) dz = \varepsilon EZ / \sigma_0^2 = (n-1)\varepsilon \quad (16)$$

Here  $\int_{\gamma}$  and  $\gamma$ , are respectively the cumulative chi-square function and the chi-square density with  $\nu$  degrees of freedom. It is easily seen that (16) may be written

$$\int_{n+1}^{\infty} (z') + 1 - \int_{n+1}^{\infty} (z'') = \varepsilon$$

Applying

$$\int_{\nu+2}^{\infty} (z) = -\frac{2}{\nu} z \gamma_{\nu}(z) + \int_{\nu}^{\infty} (z)$$

(obtained by integration by parts) and (15), we get

$$z' \gamma_{n-1}(z') = z'' \gamma_{n-1}(z'') \quad (17)$$

(15) and (17) determine  $z'$  and  $z''$ . Thus  $\sigma = \sigma_0$  is rejected if either  $Z < z' \sigma_0^2$  or  $> z'' \sigma_0^2$ , and this method is the uniformly most powerful unbiased test.

Let us now study the situation where we are interested in testing several of the parameters in the exponent simultaneously, not only one as in theorem II.E.1.

We then write the probability measure of  $X$

$$dP_{\tilde{\gamma}, \mathcal{P}} = A(\tilde{\gamma}, \mathcal{P}) e^{\sum_{j=1}^s \tilde{\gamma}_j \gamma_j(x) + \sum_{j=1}^r \mathcal{P}_j \gamma_j(x)} dP_C \quad (18)$$

where  $\tilde{\gamma} \in \omega$ , and each  $\mathcal{P}_j$  varies in an interval having 0 as inner point. We shall test  $\int_1 = \dots = \int_r = 0$ . As in theorem II.E.1 we obtain that a necessary condition for unbiasedness is that

$$E_0(\delta(X)/Y) = \xi$$

(19)

$$E_0(V_j(X)\delta(X)/Y) = \xi E(V_j(X)/Y); j = 1, 2, \dots, r$$

Maximizing the power for a particular alternative  $(\xi^0, \eta^0)$  leads to maximizing

$$E_0(e^{\sum g_j^0 V_j(X)} | Y)$$

If we use (19) as side conditions, we are led to a test

$\delta_0(X) = \psi(V(X), Y(X))$ , where

$$\psi(v, y) = 1 \quad \text{if} \quad e^{\sum g_j^0 v_j} > \sum g_i(y) v_i + g(y)$$

(20)

$$\psi(v, y) = 0 \quad \text{if} \quad e^{\sum g_j^0 v_j} < \sum g_i(y) v_i + g(y),$$

where the  $g_i$  and  $g$  should be determined such that (19) is satisfied. If such a determination is possible we have obtained a test independent of  $\xi^0$ . It is thus the most powerful test for any alternative  $(\xi, \eta^0)$ ;  $\xi \in \Omega$ ; among all tests satisfying (19). However, if  $r > 1$ , we get in general no uniformly most powerful unbiased test.

To solve this dilemma Neyman and Pearson in 1938 replaced the requirement of high power everywhere, with the requirement of large curvature upwards of the powersurface  $\beta(\beta_1, \dots, \beta_r)$  for  $\beta^0 = 0$ . We shall not deal with this method here.

# H. Testing of non-regular Darmois-Koopman classes of distribution.

Suppose that under the null-hypothesis (or the boundary points for the null-hypothesis) the class of distribution is  $\{P_{\gamma}\}_{\gamma \in \omega}$  where  $P_{\gamma}$  is given by II.B.eq. (1) but where the parameter set  $\omega$  contains no inner points. This is the situation if there are "functional" (non-linear) relations between the parameters  $\gamma_1, \dots, \gamma_s$ . It may be possible to write  $\gamma_i = \gamma_i(\theta)$ ;  $i = 1, 2, \dots, s$ ; where  $\theta = (\theta_1, \dots, \theta_r)$  varies in an open set ( $r < s$ ).

In that case, even if unbiasedness implies similarity, similarity may not imply conditional testing (see theorem II.C.2).

There are important situations where this is the case.

Example 1. (Behrens-Fisher's problem).  $X_1, X_2, \dots, X_m, Y_1, \dots, Y_n$  are independent. Each  $X_i$  is normal  $(\xi, \sigma_1^2)$  and each  $Y_j$  is normal  $(\eta, \sigma_2^2)$ . All parameters  $\xi, \eta, \sigma_1^2, \sigma_2^2$  are unknown. A test for  $\xi = \eta$  is wanted. Then under the null-hypothesis, the probability measure  $P$  of  $X_1, \dots, Y_n$  is

$$dP = A e^{(\frac{1}{2} - \frac{1}{2\sigma_1^2}) \sum x_i^2 + (\frac{1}{2} - \frac{1}{2\sigma_2^2}) \sum y_i^2 + \frac{\xi}{\sigma_1^2} \sum x_i + \frac{\xi}{\sigma_2^2} \sum y_i} dP_0$$

where  $P_0$  is the measure corresponding to  $\xi = \eta = 0$ ,  $\sigma_1^2 = \sigma_2^2 = 1$ . Hence it is seen that the four parameters  $\gamma_1, \dots, \gamma_4$  depend on the three parameters  $\sigma_1^2, \sigma_2^2, \xi$  in a non-linear manner. Thus conditioning with respect to  $Y = (\sum X_i^2, \sum Y_i^2, \sum X_i, \sum Y_i)$  is not justified and would indeed be absurd, since  $Y$  is sufficient also under the a priori

assumptions. Hence the conditional distribution would be independent of all parameters and would render no information about  $\xi - \eta$ .

Example 2. (Testing the "probit"). Let  $X_1, \dots, X_n$  be independent normal  $(\xi, \sigma)$ . We want to test  $\xi = K\sigma$ . This kind of test problem arises when we want to test if  $\Pr(X_1 \leq 0) = p$ ; i.e.  $G(-\frac{\xi}{\sigma}) = p$ ; i.e.  $\xi = -G^{-1}(p)\sigma$ . It is easily seen that the same kind of difficulties arise as in example 1.

The important problem in these situations concerns the exploration and construction of the class of similar tests. To find the "structure" of the tests in the regular case is easy, it amounts to stating that  $E_0(\xi|Y) = \xi$ . In the non-regular case, some results have been obtained, which give interesting descriptions of the class of similar tests. However, until now they have not proved very useful and we shall not deal with them here.

It should be pointed out that the tests which are commonly used in the Behrens-Fisher situation are not similar, see e.g. the test described in E. Sverdrup: Laws and Chance Variations, vol.II, p.166. Then they are not unbiased, which means that relatively to some alternatives they will have very low power. They have proved useful anyhow.

### III. Power optimum tests in non-parametric situations.

We have observations  $X_1, \dots, X_n$ . Under the null hypothesis they are assumed to be independent and identically distributed with an unknown probability density. Hence we may write for the probability measure of  $X = (X_1, \dots, X_n)$  under the null hypothesis.

$$dP_f = f(x_1) \dots f(x_n) dx_1 \dots dx_n \quad (1)$$

Let  $\omega_0$  be the class of all  $f$  defined by I.C. eq. (4) and let  $\omega \supset \omega_0$ .

We shall assume that under the hypothesis  $f$  could be any member of  $\omega$ .

As in I.C we introduce the order statistic  $Y(x) = (Y_1(x), \dots, Y_n(x))$  for a sample point  $x = (x_1, \dots, x_n)$ . We now have for the conditional probability of  $X$  given  $Y$ ,

$$\Pr(X = x | Y = y) = \begin{cases} \frac{1}{n!} & \text{if } (x_1, \dots, x_n) \text{ is a permutation} \\ & \text{of } (y_1, \dots, y_n) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Thus the conditional probability is independent of  $f$ , hence  $Y$  is sufficient for  $f$ . We have proved in I.C that  $Y$  is also complete for  $f$ ; i.e. the class  $\{P_f Y^{-1}\}_{f \in \omega}$  is complete.

Suppose now that a test  $\delta$  is required to be similar over  $\omega$  with level  $\epsilon$ . As in II.C.2 it follows that a necessary and sufficient condition for this to be the case is that  $\delta$  is a conditional test given  $Y$ ; i.e.

$$E_f(\delta(X) | Y = y) = \epsilon \quad \text{a.e.; } f \in \omega. \quad (3)$$

Let  $R(y)$  be the class of all permutations  $(x_1, \dots, x_n)$  of  $y_1, \dots, y_n$ . Then it is seen from (2) that (3) may be written

$$\sum_{x \in R(y)} \delta(x) = n! \mathcal{E} \quad (4)$$

Let us now consider the case where  $\omega$  is the class of all densities which are continuous almost everywhere. Then obviously  $\omega \supset \omega_0$  and the results above are valid.

The class of alternatives to the null hypothesis is given by

$$dP_{f,\mathcal{G}} = f(x_1 - t_1 \mathcal{G}) \dots f(x_n - t_n \mathcal{G}) dx_1 \dots dx_n \quad (5)$$

where  $f$  and  $\mathcal{G}$  are unknown.  $f$  is known to belong to  $\omega$  and  $\mathcal{G}$  is any real number.  $t_1, \dots, t_n$  are a priori specified numbers, not all equal. Let  $\mu$  be the median for the density  $f$ . Then it is seen that

$$\text{median}(X_i) = \mu + \mathcal{G} t_i \quad (6)$$

Thus the alternative to random sampling is that there is "median regression" between  $X_i$  and  $t_i$ . The null hypothesis now takes the form  $\mathcal{G} = 0$ .

A special case is the "two sample situation", where a priori  $X_1, \dots, X_m$  have a common distribution and  $X_{m+1}, \dots, X_n$  also have a common distribution. This case is obtained by setting  $t_i = 0$ ;  $i = 1, 2, \dots, m$ ;  $t_i = 1$ ;  $i = m+1, \dots, n$ .

For convenience we shall sometimes write

$$f(x_1 - t_1 \mathcal{G}) \dots f(x_n - t_n \mathcal{G}) = p(x - t\mathcal{G}; f)$$

In general, the power of a test  $\mathcal{D}$  can be written

$$\begin{aligned} \beta(\mathcal{G}) &= \int \dots \int \mathcal{D}(x_1, \dots, x_n) f(x_1 - t_1 \mathcal{G}) \dots f(x_n - t_n \mathcal{G}) dx_1 \dots dx_n = \\ &= \int \mathcal{D}(x) p(x - t\mathcal{G}; f) dx \end{aligned} \quad (7)$$

$f$  is continuous almost everywhere, hence

$$\lim_{\eta \rightarrow 0} p(x-\eta; f) = p(x; f) \quad \text{a.e.} \quad (3)$$

It then follows from Scheffé's theorem, that

$p(x-\eta; f) \rightarrow p(x; f)$  in the mean, i.e.

$$\int |p(x-\eta; f) - p(x; f)| dx \rightarrow 0$$

Hence we have

$$|\beta(\eta) - \beta(0)| \leq \int |p(x-\eta; f) - p(x; f)| dx \rightarrow 0,$$

so  $\beta(\eta)$  is continuous for  $\eta = 0$  (and everywhere). It follows that all unbiased tests must be similar for  $\eta = 0$ .

We have proved above that similarity implies conditional testing. Assuming that we want a level  $\alpha$  unbiased test, we can limit ourselves to conditional tests. Since the conditional distributions have the form (2), we are led to "combinatorial" tests, i.e. the tests of the type which has been generally recognized as "good" in non-parametric situations.

We shall now maximize the power. We then need the conditional distribution of  $X$  given  $Y$  under (5).

Suppose that  $x_1, \dots, x_n$  are all different and that  $y_1, \dots, y_n$  are  $x_1, \dots, x_n$  arranged in an increasing sequence. We make  $\Delta_1, \dots, \Delta_n$  so small that  $\bigcap_{i=1}^n (x_i \leq X_i \leq x_i + \Delta_i)$  contains no two points the coordinates of which are permutations of each other. We then get as all  $\Delta_i \rightarrow 0$ ,

$$\begin{aligned} \Pr\left[\bigcap_i (x_i \leq X_i \leq x_i + \Delta_i) | Y\right] &= \\ &= \lim \Pr\left[\bigcap_i (x_i \leq X_i \leq x_i + \Delta_i) | \bigcap_i (y_i \leq Y_i \leq y_i + dy_i)\right] = \\ &= \lim \frac{\Pr(\bigcap_i (x_i \leq X_i \leq x_i + dy_{j_i}))}{\Pr(\bigcap_i (y_i \leq Y_i \leq y_i + dy_i))}, \end{aligned}$$

since for sufficiently small  $dy_{j_i}$ ,  $(x_i \leq X_i \leq x_i + dy_{j_i})$  is a subset of  $(x_i \leq X_i \leq x_i + \Delta_i)$ . Here  $(y_{j_1}, \dots, y_{j_n}) = (x_1, \dots, x_n)$ . Hence we get

$$\Pr(X = x | Y = y) = p(x - t_\beta; f) / \sum_{x' \in R(y)} p(x' - t_\beta; f) \quad (9)$$

if  $x \in R(y)$ . Otherwise  $\Pr(X = x | Y = y) = 0$ . If  $\beta = 0$  it is seen that (9) reduces to (2). (9) could also easily have been derived from the more sophisticated definition of conditional probability.

We are interested in finding a  $\delta$  which maximizes the power  $E_{f, \beta} \delta(X)$  for  $\beta \neq 0$  subject to  $\delta$  being unbiased with level  $\epsilon$ . As we have done several times before, we shall maximize the power subject to the weaker condition (4). Now since

$$E_{f, \beta} \delta(X) = E_{f, \beta} E_{f, \beta} [\delta(X) | Y]$$

we shall maximize

$$E[\delta(X) | Y = y] = \sum_{x \in R(y)} \delta(x) p(x - t_\beta; f) / \sum_{x \in R(y)} p(x - t_\beta; f) \quad (10)$$

subject to (4).

According to Neyman-Pearson's lemma we shall then reject the hypothesis if

$$p(X-t_g;f) / \sum_{x' \in R(Y)} p(x'-t_g;f) > c(Y), \quad (11)$$

reject with probability  $\Upsilon(Y)$  if we have "=" in (11), and not reject if we have "<" in (11). Hence we shall use a  $\delta_0$  defined by

$$\begin{aligned} \delta_0(X) &= 1 && \text{if } p(X-t_g;f) > d(Y), \\ \delta_0(X) &= \Upsilon(Y) && \text{if } p(X-t_g;f) = d(Y), \\ \delta_0(X) &= 0 && \text{if } p(X-t_g;f) < d(Y), \end{aligned} \quad (12)$$

where  $d(Y)$  and  $\Upsilon(Y)$  are determined such that (4) is satisfied with  $\delta = \delta_0$ .

For given  $Y = y$  consider the  $n!$  quantities  $p(x-t_g;f)$  obtained by letting  $x$  run through all points in  $R(y)$ . Let us order them in a decreasing sequence,  $p^{(1)}, p^{(2)}, \dots, p^{(n!)}$ , and let the corresponding permutations of  $Y$  be

$$x^{(1)}, x^{(2)}, \dots, x^{(n!)} \quad (13)$$

For given  $f$ ,  $g$  and  $Y$ , the sequence (13) can in principle always be found, even if the numerical work will be prohibitive if  $n$  is large.

Let us now determine  $k$  and  $\Upsilon$  by

$$k + \Upsilon = n! \varepsilon \quad 0 \leq \Upsilon < 1 \quad (14)$$

i.e.  $k$  is the integer part and  $\Upsilon$  the decimals in  $n!\varepsilon$ .

We now see from (12) and (4), that according to  $\delta_0$ , we shall reject the hypothesis if  $Y(X) = y$  and  $X$  is one of the points  $x^{(1)}, \dots, x^{(k)}$ . We shall reject the hypothesis with probability

$\gamma$  if  $Y(X) = y$  and  $X = x^{(k+1)}$ . Otherwise when  $Y(X) = y$  we shall not reject the hypothesis.

This method is the most powerful test relatively to the particular alternative  $(f, \rho)$  among all similar tests with level  $\varepsilon$ .

Unfortunately, however, the sequence (13) depends on  $f$  and  $\rho$ . Hence  $\delta_0$  depends on  $f$  and  $\rho$ . We have therefore not obtained a uniformly most powerful test. In many cases it would not even be unbiased. It seems as if the methods applied in the case of the regular Darmois-Koopman exponential class of distributions cannot be applied in the non-parametric situations.

#### IV. Estimation in connection with Darmois-Koopman classes.

##### A. Some mathematical results.

Let  $X_i$  be a real random variable with distribution  $P$  belonging to a Darmois-Koopman class  $\mathcal{D}$ . We now assume that each  $\tau_j$  is a function of a parameter  $\theta = (\theta_1, \dots, \theta_r)$  ( $r \leq s$ ):

$$dP_\theta(x_i) = A(\tau(\theta)) e^{\sum_{j=1}^s \tau_j(\theta) Y_j(x_i)} dP_c(x_i) \quad (1)$$

Here  $\theta$  is known to belong to a set in the  $r$ -dimensional space containing an open subset  $(H)$ . Without restricting generality, we have arranged that  $\theta = 0 \in (H)$ , and that  $\tau_1(\theta), \dots, \tau_s(\theta)$  are linearly independent. However, at present the space  $\Omega$  obtained by varying  $\tau$  may contain no open subsets, i.e.  $\mathcal{D}$  may not be regular. By introducing  $\tau_0 = \log A(\tau)$ , (1) can be written

$$dP_\theta = e^{\tau_0(\theta) + \sum_{j=1}^s \tau_j(\theta) Y_j(x)} dP_c \quad (2)$$

We will now assume that all derivatives

$$\tau_{jm} = \frac{\partial \tau_j}{\partial \theta_m}, \quad \tau_{jmn} = \frac{\partial^2 \tau_j}{\partial \theta_m \partial \theta_n}; \quad j = 0, 1, \dots, s; \quad m, n = 1, \dots, r$$

exist and are continuous functions of  $\theta$ . Denoting  $\frac{dP_\theta}{dP_c}$  by  $L$  we introduce the notations

$$V_m = \frac{\partial \log L}{\partial \theta_m} = \tau_{0m} + \sum_{j=1}^s \tau_{jm} Y_j \quad (3)$$

$$W_{m,n} = \frac{\partial^2 \log L}{\partial \theta_m \partial \theta_n} = \tau_{omn} + \sum_{j=1}^s \tau_{jmn} Y_j \quad (4)$$

In matrix notations equation (3) may be written

$$V = D\tau^0 + (D\tau)' Y \quad (5)$$

where the definitions of the matrices of derivatives and of  $Y$  follow by comparison with (3). We now assume that the region of convergence  $\Omega$  of  $\int e^{\sum \tau_j Y_j(x)} dP_0$  contains all points  $(\tau_1(\theta), \dots, \tau_s(\theta))$  as inner points. Then  $EV_m$ ,  $EW_{mn}$  and

$$\eta_j = EY_j$$

exist, and we may differentiate  $\int L dP_0 = 1$  by taking differentiation under the integral sign (according to theorem I.C.2).

We then get from  $\int \frac{\partial L}{\partial \theta_m} dP_0 = 0$  that

$$EV_m = \tau_{om} + \sum_{j=1}^s \tau_{jmn} \eta_j = 0 \quad (6)$$

and from  $W_{m,n} = -V_m V_n + \frac{1}{L} \frac{\partial^2 L}{\partial \theta_m \partial \theta_n}$  and  $\int \frac{\partial^2 L}{\partial \theta_m \partial \theta_n} dP_0 = 0$  that

$$\begin{aligned} \lambda_{mn} &= E(V_m V_n) = -EW_{mn} \\ &= -\tau_{omn} - \sum_{j=1}^s \tau_{jmn} \eta_j \end{aligned} \quad (7)$$

Let  $\sigma$  be the covariance matrix for  $Y$  :

$$\sigma_{ij} = E(Y_i - \eta_i)(Y_j - \eta_j) \quad (8)$$

By  $\sigma = E(Y - EY)(Y - EY)'$ , (5), (6) and (7) we get

$$\lambda = EVV' = (D\tau)' \sigma (D\tau) \quad (9)$$

$$W_{m,n} = \frac{\partial^2 \log L}{\partial \theta_m \partial \theta_n} = \tau_{omn} + \sum_{j=1}^s \tau_{jmn} Y_j \quad (4)$$

In matrix notations equation (3) may be written

$$V = D\tau^0 + (D\tau)' Y \quad (5)$$

where the definitions of the matrices of derivatives and of  $Y$  follow by comparison with (3). We now assume that the region of convergence  $\Omega$  of  $\int_{\Omega} e^{\sum \tau_j Y_j(x)} dP_0$  contains all points  $(\tau_1(\theta), \dots, \tau_s(\theta))$  as inner points. Then  $EV_m$ ,  $EW_{mn}$  and

$$\eta_j = EY_j$$

exist, and we may differentiate  $\int L dP_0 = 1$  by taking differentiation under the integral sign (according to theorem I.C.2).

We then get from  $\int \frac{\partial L}{\partial \theta_m} dP_0 = 0$  that

$$EV_m = \tau_{om} + \sum_{j=1}^s \tau_{jmn} \eta_j = 0 \quad (6)$$

and from  $W_{m,n} = -V_m V_n + \frac{1}{L} \frac{\partial^2 L}{\partial \theta_m \partial \theta_n}$  and  $\int \frac{\partial^2 L}{\partial \theta_m \partial \theta_n} dP_0 = 0$  that

$$\begin{aligned} \lambda_{mn} &= E(V_m V_n) = -EW_{mn} \\ &= -\tau_{omn} - \sum_{j=1}^s \tau_{jmn} \eta_j \end{aligned} \quad (7)$$

Let  $\sigma$  be the covariance matrix for  $Y$  :

$$\sigma_{ij} = E(Y_i - \eta_i)(Y_j - \eta_j) \quad (8)$$

By  $\sigma = E(Y - EY)(Y - EY)'$ , (5), (6) and (7) we get

$$\lambda = EVV' = (D\tau)' \sigma (D\tau) \quad (9)$$

By theorem I.B.2 we may differentiate the equation

$$\eta_j = EY_j = \int Y_j(x) L dP_0$$

by taking differentiation under the sign of integration. This gives us

$$\begin{aligned} \eta_{jm} &= \frac{\partial \eta_j}{\partial \theta_m} = \int Y_j(x) \frac{\partial L}{\partial \theta_m} dP_0 = \int Y_j(x) \frac{\partial \log L}{\partial \theta_m} L dP_0 \\ &= E(Y_j V_m), \end{aligned} \quad (10)$$

or by (3) and (6)

$$\begin{aligned} \eta_{jm} &= E(Y_j(\tau_{0m} + \sum_{i=1}^S \tau_{im} Y_i)) = E(Y_j \sum_{i=1}^S \tau_{im} (Y_i - \eta_i)) \\ &= \sum_{i=1}^S \tau_{im} E(Y_j - \eta_j)(Y_i - \eta_i) = \sum_{i=1}^S \sigma_{ji} \tau_{im}, \end{aligned}$$

which, after introduction of the matrix  $D\eta = \{\eta_{jm}\}$ , for the system of derivatives of  $\eta$  may be written

$$D\eta = \sigma(D\tau) \quad (11)$$

#### B. The information matrix.

It is well known from the asymptotic properties of the maximum likelihood estimates, and from the Frechet inequality, that the matrix

$$\lambda = - \left\{ E \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right\}_{ij} \equiv - \{EW_{ij}\} \quad (12)$$

is of special interest.  $\lambda$  is sometimes called the information

matrix. From (9) we get

$$\lambda_{mn} = \sum_{i=1}^s \sum_{j=1}^s \sigma_{ij} \tau_{im} \tau_{jn} \quad (13)$$

We now assume that  $\mathcal{S}$  is regular, so we may take derivatives with respect to  $\tau_1, \dots, \tau_s$ . This gives

$$\frac{\partial \log L}{\partial \tau_i} = \frac{\partial \tau_0}{\partial \tau_i} + Y_i ; \quad i = 1, \dots, s \quad (14)$$

( $\tau_0$  is considered as a function of  $\tau_1, \dots, \tau_s$ ), and as  $E \frac{\partial \log L}{\partial \tau_i} = 0$ ,

$$\eta_i = EY_i = - \frac{\partial \tau_0}{\partial \tau_i} \quad (15)$$

Furthermore, as an analogue to (7), we have

$$\sigma_{ij} = E \left( \frac{\partial \log L}{\partial \tau_i} \frac{\partial \log L}{\partial \tau_j} \right) = -E \frac{\partial^2 \log L}{\partial \tau_i \partial \tau_j} \quad (16)$$

But from (14) we find  $\frac{\partial^2 \log L}{\partial \tau_i \partial \tau_j} = \frac{\partial^2 \tau_0}{\partial \tau_i \partial \tau_j}$ , which by (13) and (16) gives us

$$\lambda_{mn} = - \sum_{i,j} \frac{\partial^2 \tau_0}{\partial \tau_i \partial \tau_j} \tau_{im} \tau_{jn} \quad (17)$$

Thus, no integration is needed in order to find the Frechet lower bound of variance of unbiased estimators or the asymptotic covariance matrix of maximum likelihood estimators.

C. Fisher-consistent estimates.

We return to the general theory, with  $\mathcal{S}$  not necessarily being regular.

Assume now that we have observed  $X = (X_1, \dots, X_N)'$  where  $X_1, \dots, X_N$  are independent each with probability measure  $P$ . The probability measure of  $X$  is now given by

$$\prod_{i=1}^N dP_{\theta}(x_i) = e^{N\tau_0(\theta) + N \sum_{j=1}^s \tau_j(\theta) \bar{Y}_j} \prod_{i=1}^N dP_0(x_i) \quad (18)$$

where

$$\bar{Y}_j = \frac{1}{N} \sum_{i=1}^N Y_j(x_i); \quad j = 1, \dots, s$$

An estimate  $\theta_1^*$  of  $\theta_1$  is said to be Fisher-consistent if  
(i)  $\theta_1^* = f(\bar{Y})$  depends on  $X$  only through  $\bar{Y}$ , and  $f$  is independent of  $N$ .

$$(ii) \quad f(\eta_1, \dots, \eta_s) = \theta_1 \quad (19)$$

(iii)  $f$  has continuous derivatives  $f_j = \frac{\partial f}{\partial \eta_j}$ .

Now,  $\text{plim}_{N \rightarrow \infty} \bar{Y}_j = \eta_j$ , and by (iii)  $f$  is continuous. From Slutsky's theorem it then follows that

$$\text{plim}_{N \rightarrow \infty} \theta_1^* = \text{plim}_{N \rightarrow \infty} f(\bar{Y}_1, \dots, \bar{Y}_s) = \theta_1 \quad (20)$$

Thus, every Fisher-consistent estimate is consistent.

We introduce the notation

$$\eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_s \end{pmatrix}$$

By means of the mean value theorem, we get

$$\theta_1^* = f(\bar{Y}) = \theta_1 + \sum_{j=1}^S (\bar{Y}_j - \eta_j) f_j(\eta + R(\bar{Y} - \eta)) \quad (21)$$

where  $R$  is a stochastic vector with components  $|R_i| \leq 1$ , from which it is seen that  $\sqrt{N}(\theta_1^* - \theta_1)$  has the same limit distribution as  $\sum_j (\bar{Y}_j - \eta_j) \sqrt{N} f_j(\eta)$ , and that consequently any Fisher-consistent estimate  $\theta_1^*$  is asymptotically normal with mean  $\theta_1$  and

$$\text{as.var } \theta_1^* = \frac{1}{N} \sum_{i,j} f_i f_j \sigma_{ij} = \frac{1}{N} (Df)' \sigma (Df) \quad (22)$$

where we have introduced the notation  $Df$  for the set of derivatives of  $f$ .

We shall find a lower bound for (22) under (i), (ii) and (iii). By differentiating (19) with respect to  $\theta_m$  we get

$$\sum_{j=1}^S f_j \eta_{jm} = \begin{cases} 1, & m = 1 \\ 0, & m > 1 \end{cases} \quad (23)$$

which may also be written

$$(D\eta)'(Df) = \epsilon = \begin{Bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{Bmatrix} \quad (24)$$

Let us first minimize (22) with respect to  $Df$  subject to (23). Since (i), (ii), (iii) is at least as restrictive as (23), this will certainly give us a lower bound for (22). We use the Lagrange multiplier rule, starting with minimizing

$$\sum_{i,j} f_i f_j \sigma_{ij} - 2 \sum_{m=1}^r \lambda_m \left( \sum_{i=1}^S f_i \eta_{im} - \delta_{1m} \right), \quad (25)$$

where  $\lambda_1, \dots, \lambda_r$  are the "multipliers" and  $\delta_{1m} = 1$  if  $m = 1$ , otherwise 0. The minimizing  $f_1, \dots, f_S$  and the values of  $\lambda_1, \dots, \lambda_r$  are found as solutions of the equations

$$\sum_{j=1}^s f_j \sigma_{ij} - \sum_{m=1}^r \kappa_m \eta_{im} = 0 ; \quad i = 1, \dots, s \quad (26)$$

and (23). Now (26) can also be written as

$$\sigma(Df) = (D\eta) \kappa$$

hence  $Df = \sigma^{-1}(D\eta) \kappa$  which inserted in (24) gives

$$(D\eta)' \sigma^{-1}(D\eta) \kappa = \epsilon \quad \text{or} \quad \kappa = [(D\eta)' \sigma^{-1}(D\eta)]^{-1} \epsilon \quad \text{and finally}$$

$$Df = \sigma^{-1}(D\eta) \kappa = \sigma^{-1}(D\eta) [(D\eta)' \sigma^{-1} D\eta]^{-1} \epsilon \quad (27)$$

Substituting this in  $(Df)' \sigma(Df)$  and applying (11) and (9), we get

$$\min(Df)' \sigma(Df) = \epsilon' [(D\eta)' \sigma^{-1}(D\eta)]^{-1} \epsilon = \epsilon' \lambda^{-1} \epsilon$$

Hence any Fisher-consistent estimate has asymptotic variance

$$\geq \frac{1}{N} \epsilon' \lambda^{-1} \epsilon = \frac{1}{N} (\lambda^{-1})_{11} \quad (28)$$

[where  $(\lambda^{-1})_{11}$  denotes the leading element in the matrix  $\lambda^{-1}$ ].

#### D. The maximum likelihood estimator.

Returning to the model defined by (18), the maximum likelihood estimate  $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_r)$  of  $\theta$  is defined as the solution of the equations

$$\tau_{om}(\hat{\theta}) + \sum_{j=1}^s \tau_{jm}(\hat{\theta}) \bar{Y}_j = 0 ; \quad m = 1, \dots, r \quad (29)$$

(without regard to whether  $\hat{\theta}$  actually maximizes (18)). We assume that the equations (29) have a unique solution  $\hat{\theta}$ . It is seen from (29) that the estimate  $\hat{\theta}_i$  for  $\theta_i$  satisfies condition (i), from (6) that it satisfies (ii), and from the assumptions about the derivatives  $\tau_{jmn}$  that (iii) is satisfied. Thus  $\hat{\theta}_i$  is Fisher-consistent, and especially

$$\text{plim}_{N \rightarrow \infty} \hat{\theta}_i = \theta_i$$

Now the asymptotic multinormality of  $\hat{\theta}$  is easily seen in the following manner.

By the mean value theorem we have

$$\tau_{jm}(\hat{\theta}) = \tau_{jm}(\theta) + \sum_{n=1}^r (\hat{\theta}_n - \theta_n) \tau_{jmn}(\theta + S(\hat{\theta} - \theta))$$

with  $|S_i| \leq 1$ . Introducing this, we see that (29) can be written

$$\sqrt{N} \bar{V}_m + \sum_{n=1}^r \sqrt{N} (\hat{\theta}_n - \theta_n) \sum_{j=0}^s \bar{Y}_j \tau_{jmn}(\theta + S(\hat{\theta} - \theta)) = 0 \quad (30)$$

where for convenience we use the notation  $\bar{Y}_0 = 1$ .

However, since  $\tau_{jmn}$  is continuous and  $\text{plim}_{N \rightarrow \infty} \hat{\theta}_n = \theta_n$ , we have  $\text{plim}_{N \rightarrow \infty} \tau_{jmn}(\theta + S(\hat{\theta} - \theta)) = \tau_{jmn}(\theta) = \tau_{jmn}$ . Furthermore  $\text{plim} \bar{Y}_j = \eta_j$ . By the central limit theorem, the limit cumulative distribution function of  $\sqrt{N} \bar{V} = \sqrt{N}(\bar{V}_1, \dots, \bar{V}_r)'$  is the cumulative multinormal distribution with mean 0 and covariance matrix  $\lambda$ . Hence by (30), the limit cumulative distribution function of  $\sqrt{N}(\hat{\theta} - \theta)$  is that of  $T = (T_1, \dots, T_r)'$ , where  $T$  is given by

$$U_m + \sum_{n=1}^r T_n \sum_{j=0}^s \eta_j \tau_{jmn} = 0 \quad (31)$$

and where  $U = (U_1, \dots, U_r)'$  is multinormal  $(0, \lambda)$ , and  $\eta_0 = 1$ .

Combining (31) and (7) we get

$$U = \lambda T \quad (32)$$

Hence  $T = \lambda^{-1}U$ ,  $ET = 0$ , and the covariance matrix of  $T$  is  $\lambda^{-1}\lambda(\lambda^{-1})' = \lambda^{-1}$ . Thus  $T$  is multinormal  $(0, \lambda^{-1})$ . We conclude that:

The limit cumulative distribution function of  $\sqrt{N}(\hat{\theta} - \theta)$  as  $N \rightarrow \infty$  is the cumulative multinormal with expectation 0 and covariance matrix  $\lambda^{-1}$ ; i.e.  $\hat{\theta}$  is asymptotically multinormal  $(\theta, (1/N)\lambda^{-1})$ , and in particular  $\hat{\theta}_1$  is asymptotically normal with mean  $\theta_1$  and variance  $(\lambda^{-1})_{11}/N$ .

$$\text{as. var } \hat{\theta}_1 = \frac{1}{N}(\lambda^{-1})_{11} \quad (33)$$

Comparing this with (28) we may conclude that the maximum likelihood estimate of  $\theta_1$  has asymptotic variance less than or equal to that of any other Fisher-consistent estimate for  $\theta_1$ .