MAJORIZATION AND APPROXIMATE MAJORIZATION FOR FAMILIES OF MEASURES. APPLICATIONS TO LOCAL COMPARISON OF EXPERIMENTS AND THE THEORY OF MAJORIZATION OF VECTORS IN R^n (SCHUR CONVEXITY).

By Erik Torgersen, University of Oslo.

The general theory of comparison (and of approximate comparison) of experiments generalizes (as shown by this author in 1969), without too many complications, to a theory of comparison which is applicable to general families of measures. After having introduced a suitable notion of differentiability for statistical experiments it was then shown that local comparison of statistical experiments may be reduced to comparisons of certain finite and quite natural sets of measures. Thus we have at our hand a possible foundation for a nonasymptotic theory for local comparison and in particular for local sufficiency. Another natural application of this general notion of comparison is to the theory of majorization of vectors in R^n (Schur convexity).

CONTENT.

MAJORIZATION AND APPROXIMATE MAJORIZATION (COMPARISON OF FAMILIES OF MEASURES).

1. INTRODUCTION

One of the main contributions of the theory of comparison of experiments was to show that natural and apparently different criteria for comparison actually was equivalent. Thus we have equivalent criteria in terms of, over all comparison of risk functions, in terms of Bayes risk for a fixed prior distribution, in terms of sub linear functionals, in terms of performance functions of decision rules and in terms of randomizations.

The underlying framework of these results is the decision theory of Abraham Wald.The first results, of Blackwell and others, was concerned with criteria for the ordering "being more informative than". Later on LeCam in his 1964 paper on sufficiency and approximate sufficiency introduced the concept of a deficiency of one experiment w.r.t. another. As experiments usually are not ordered and since these deficiencies always exists this extended considerably the applicability of the theory. Combining Blackwell's idea of comparison for k-decision problems with LeCam's deficiencies this author in 1970 considered deficiencies for k-decision problems.

The theory has found important applications to asymptotic statistical theory - but also to non asymptotic theory. Besides LeCam's works the reader might consult Millar's recent survey on asymptotic minimax theory. Expositions of the theory of statistical experiments may be found in Heyer 1973, LeCam 1974 and in Torgersen 1976. A recent contribution to non asymptotic theory is Lehmann 1983.

The idea that the underlying distributions which constitute a statistical experiment should be proper probability distributions is of course essential for the interpretations of most results from the theory of comparison of experiments. In particular the interpretation of risk as expected loss depends on this assumption. If we however look over the various arguments which are used then we see that this assumption is often not needed or may be avoided by a suitable reformulation.

Even the formal expression of the risk of a decision procedure as the value of the corresponding bilinear functional for the underlying distribution and the loss function remains well defined within a much more general set up.

We shall now see that such a generalization, or extension, to "non proper" experiments actually yields interesting results on "proper experiments". Another benefit is a unified approach to several results concerning systems of inequalities which are frequently encountered in mathematical statistics. The theory we shall formulate here may be considered as a generalization of the theory of majorization as it is described in the book by Marshall and Olkin from 1979, and in later generalizations of Dahl and of Karlin from 1983.

Let us at this point briefly indicate by examples some directions of applications:

Example 1.1. (Comparison of modified risks)

Assume that all losses are bounded by 1 and let a, b and ε be given functions of the unknown parameter θ.

When are we entitled to claim for all decision problems (of a certain type) and given experiments $\mathcal{E}$ and $\mathcal{F}$ that there to each risk function s obtainable in $\mathcal{F}$ corresponds a risk function r obtainable in $\mathcal{E}$ such that $a(\theta)r(\theta)<b(\theta)s(\theta)+\varepsilon(\theta)$ for all θ?

If $a(\theta)\underset{\theta}{\equiv}1$ and $b(\theta)\underset{\theta}{\equiv}1$ then this is just a description of the situation where $\mathcal{E}$ is ε-deficient w.r.t $\mathcal{F}$ according to LeCam's definition of deficiency.

Theoretically at least this problem, for general a, b and ε, is easily described in terms of general families of measures. We shall in the next section see that one possible answer may be obtained by a straightforward generalization of LeCam's randomization (Markov kernel) criterion.

Example 1.2. (Local (nonasymptotic) comparison of experiments).

Consider experiments having the same sub set Θ of R^m as their parameter set. Let θ^0 be an interior point of Θ. If $\mathcal{E} = (P_\theta : \theta\in\Theta)$ is an experiment then, assuming differentiability, we may consider the measure P_{θ^0} along with the partial derivatives $[\partial P_\theta/\partial\theta_i]_{\theta=\theta^0}$; $i = 1,\dots,m$. This family of measures is not an experiment since the total masses of the measures $[\partial P_\theta/\partial\theta_i]_{\theta=\theta^0}$ are all zero.

If θ is close to P_{θ^0} then the measure P_θ may be approximated by: $P_{\theta^0}+\sum_{i=1}^{m}(\theta-\theta^0)_i[\partial P_\theta/\partial\theta_i]_{\theta=\theta^0}$. Thus it is not surprising that if θ is close to θ^0 and if quantities of smaller order than the distance from θ to θ^0 are considered negligible then the local behavior of $\mathcal{E}$ around θ^0 may be completely described in terms of this family of measures. It is however interesting that, as we shall se in section 3 of this paper, the local comparison of experiments within small neighbourhoods of θ^0 may be expressed quite naturally and, in theory at least, simply in terms of such families of measures.

Example 1.3. (Comparison of measures. Dilations).

Systems of inequalities of the following type have received attention in various connections.

Let μ and ν be measures on a measurable space $(\chi,\mathcal{A})$. Consider also a convex set H of integrable functions which contains the constants and which is closed under the formation of maxima of finite sub sets. What conditions on (H,μ,ν) ensure that the inequality $\int hd\mu > \int hd\nu$ holds for every function h in H?

The ordering of experiments is a closely related case with μ and ν being conical measures of experiments.

We may also consider, as we shall, systems of inequalities of the form $\int h d\mu > \int h d\nu - \varepsilon_h$ for some given function $h \to \varepsilon_h$ on H. Thus we are led to compare the families $(h\mu : h \in H)$ and $(h\nu : h \in H)$ of measures.

We shall in section 4 see how the general theory of comparison of families of measures provides a method of attack for such problems.

Another related and most interesting discussion of such problems may be found in Karlin 1983.

Example 1.4. (Distributions with given marginals).

The fundamental papers by LeCam 1964 and by Strassen 1965 both show that there is a close relationship between the theory of comparison of experiments on the one hand and various existence problems for joint distributions with given marginals on the other hand. This relationship is also apparent from earlier works of e.g. of Blackwell 1953 and Boll 1955.

Thus the dilation criterion for ordering of experiments is, as shown by Strassen, related to the problem of whether or not a sequence of distributions might be the sequence of distributions of a martingale. Strassen considered in his paper the problem of deciding whether one distribution P is the convolution factor of another distribution Q. In his paper Boll had then already shown that this is tantamount to the problem of deciding whether the translation experiment defined by P is at least as informative as the translation experiment defined by Q. The statistical significance of Strassen's criterion in terms of minimax risk was, using the results in LeCam's paper, clarified in a later paper by Torgersen 1972c which also generalized these results to the case of ε-deficiency.

Among other results which bring out the connection between these fields we shall here only mention the various criteria for stochastic orderings of distributions on a general partially ordered set. The basic criterion in terms of probabilities of "monotone" events follows directly from Strassen's paper. It was shown later, Torgersen 1982, that this criterion is also the criterion for "being more informative" for experiments associated with sampling plans defined by these distributions. More generally it was shown that the case of ε-deficiency corresponded nicely with one distribution being up to an amount of ε stochastically larger than the other distribution.

We shall in section 5 show that this and other results may be deduced from general principles for comparison of families of measures.

Example 1.5. (Majorization).

The theory of comparison of measure families may be considered as a natural generalization of most of the various theories of majorization. In particular a funtional of experiments which is monotonically increasing (decreasing) for the ordering "being more informative" may be considered as Schur convex (concave). Thus Fisher information and statistical distance may be considered Schur convex while the Hellinger transform is Schur concave.

Let us in order to make this clearer comment on the best known particular case which is the case of majorization of vectors in R^n. An important source of information for this case is Marshall and Olkin 1979. Using the notation of this book we shall write $(x_{[1]},x_{[2]},\ldots,x_{[n]})$ for the vector in R^n obtained from the vector $(x_1,\ldots,x_n)$ by arranging the x'ses in decreasing order

Let $p = (p_1,\ldots,p_n)$ and $q = (q_1,\ldots,q_n)$ be two vectors in R^n such that $p_1+\ldots+p_n = q_1+\ldots+q_n$. Consider also p and q as $1\times n$ row matrices. Then the following conditions are known to be equivalent:

(i) $p_{[1]}+\dots+p_{[j]} \geq q_{[1]}+\dots+q_{[j]}$; $j = 1,\dots,n$.

(ii) $\sum_j \phi(p_j) \geq \sum_j \phi(q_j)$ when the function ϕ is convex on R.

(iii) $\sum_j (p_j-c)^+ \geq \sum_j (q_j-c)^+$; $c \in R$.

(iv) $pM = q$ for a doubly stochastic $n\times n$ matrix M.

Assume now that the vectors p and q are probability vectors. Define probability distributions P_0, Q_0, P_1 and Q_1 on $\{1,\dots,n\}$ by putting $P_0(j) = 1/n$, $P_1(j) = p_j$, $Q_0(j) = 1/n$ and $Q_1(j) = q_j$; $1 = 1,\dots,n$. Put $\Theta = \{0,1\}$ and consider the experiments (dichotomies) $\mathcal{E} = (P_0,P_1)$ and $\mathcal{F} = (Q_0,Q_1)$.

Using the Neyman-Pearson lemma on the problem of testing "$\theta = 0$" against "$\theta = 1$" we see that (i) tells us that $\mathcal{E}$ yields at least as large power as $\mathcal{F}$ for any level of significance. By Blackwell 1951 and 1953 this is equivalent to the condition that $\mathcal{E}$ is at least as informative as $\mathcal{F}$. Conditions (ii), (iii) and (iv) express by the same papers of Blackwell the same thing. Multiplying (ii) and (iii) by (-1) we see that condition (ii) is the over all minimum Bayes risk criterion for comparability and that (iii) is the minimum Bayes risk criterion for testing problems. Finally condition (iv) states that $P_\theta M = Q_\theta$; $\theta = 0,1$ for a Markov kernel M from the set $\{1,2,\dots,n\}$ into itself. Note in particular that the condition that the Markov matrix M is doubly stochastic expresses that the corresponding Markov kernel preserves the uniform distribution i.e. that $P_0M = Q_0$.

Replacing the distribution $P_0 = Q_0$ with more general distributions we arrive at more general notions of majorization as e.g. the notion of π-majorization considered in Marshall and Olkin's book.

Now majorization is a partial ordering on all of R^n and not just on the probability simplex. The validity of the arguments (e.g. the Neyman-Pearson lemma) used above does not, however, depend on the condition that the vectors p and q are probability vectors. We shall see in section 6 that most of the mathematics remain valid, after some straightforward modifications, for general families of measures. Thus we may, as in Dahl 1983, consider ε-deficiency as well as multivariate majorization.

Multivariate majorization was also considered by Karlin 1983.

As is amply demonstrated by Marshall and Olkin in their book on inequalities the concept of majorization in R^n is a very useful tool for establishing inequalities in statistics and elsewhere.

The point we are making is that this concept fits naturally into the extension of the framework of comparison of experiments which we shall consider here. Majorization in R^n is, however, a distinguished particular case with several features which appear difficult to deduce from this general theory of comparison of families of measures.

We shall in the forthcoming sections take a closer look at the situations described in the examples of this section. It is, however, necessary that we first became aquinted with some basic principles for comparison of families of measures. These principles are therefor the topic of the next section.

2. EQUIVALENT CRITERIA FOR COMPARISON OF FAMILIES OF MEASURES.

We shall in this section be concerned with general principles for comparing families of measures. It will be assumed that all the measures belonging to a given family are real valued and that they are defined on a common measurable space.

We have found it convenient to use the term measure family to denote a family of measures satisfying these two requirements. Thus the measures in a measure family are all bounded but we do not assume that they are non negative.

A measure family $\mathcal{E}$ with measurable space $(\chi, \mathcal{A})$ may be denoted as $\mathcal{E} = (\chi, \mathcal{A}; \mu_\theta : \theta \in \Theta)$ or just as $\mathcal{E} = (\mu_\theta : \theta \in \Theta)$ where $(\mu_\theta : \theta \in \Theta)$ is the family of finite measures which constitutes $\mathcal{E}$. The set Θ is called the parameter set of $\mathcal{E}$. This set may be any set but we shall assume, unless otherwise stated, that if two measure families $\mathcal{E}$ and $\mathcal{F}$ are compared then their parameter sets are the same.

A measure family of probability measures is called an (statistical) experiment and the measurable space of an experiment is usually called the sample space of the experiment.

Consider now in addition to the parameter set Θ a set T and a family $L = (L_\theta : \theta \in \Theta)$ of real valued functions on T. If the measure families under consideration are experiments then we may below think of T as a set of decisions and L as a loss function.

Let us agree to use the notation $\|g\|$ for the supremum norm $\sup_t |g(t)|$ of a real valued function g on T. Risking the slight possibility of a conflict of interpretations we shall use the notation $\|\mu\|$ to denote the total variation of a measure μ. Finally we shall use the term deficiency function for any non negative function on the parameter set.

Assuming that the parameter set Θ is finite we shall now list four principles for comparing measure families $\mathcal{E} = (\chi, \mathcal{A}, \mu_\theta : \theta \in \Theta)$ and $\mathcal{F} = (\mathcal{Y}, \mathcal{B}, \nu_\theta : \theta \in \Theta)$. The comparison shall be expressed in terms of a fixed deficiency function ε and a fixed set T.

We precede the statement of each of the listed principles by a "headline" indicating the interpretations of the principles when the measure families are experiments.

It should be noted that the only aspect of the set T which matters here is its cardinality which we shall denote by k.

In the statement of the third of the principles below and also at several other places we are using the convenient notation $\int h(d\mu_\theta:\theta\in\Theta)$ to denote $\int h(f_\theta:\theta\in\Theta)d\mu$ whenever the finite measure μ_θ, for each θ, has density f_θ w.r.t. the non negative measure μ and h is a non negative homogenuous measurable function on R^θ. It is then easily checked that neither the existence nor the value of this integral depends on the choice of the majorizing measure μ nor on the specifications of the densities.

If Θ is finite then the first four principles mentioned in the introduction may be formulated as follows:

(i) (Pointwise comparison of risks.)

To each family $L_\theta:\theta\in\Theta$ of real valued functions L on T and each Markov kernel σ from $\mathcal{F}$ to T corresponds a Markov kernel ρ from $\mathcal{E}$ to T so that:

$$\mu_\theta\rho L_\theta \leq \nu_\theta\sigma L_\theta+\varepsilon_\theta\|L_\theta\|;\quad \theta\in\Theta.$$

(ii) (Comparison of Bayes risks.)

To each family $L_\theta:\theta\in\Theta$ of real valued functions L on T and each Markov kernel σ from $\mathcal{F}$ to T corresponds a Markov kernel ρ from $\mathcal{E}$ to T so that:

$$\sum_\theta\mu_\theta\rho L_\theta \leq \sum_\theta\nu_\theta\sigma L_\theta+\sum_\theta\varepsilon_\theta\|L_\theta\|$$

(iii) <u>(Comparison of maximum Bayes utilities. The sub linear function criterion)</u>.

$\int\psi(d\mu_\theta:\theta\in\Theta) > \int\psi(d\nu_\theta:\theta\in\Theta) - \sum_\theta\varepsilon_\theta[\psi(-e^\theta)\vee\psi(e^\theta)]$ for each function ψ on R^Θ which is a maximum of $k = \#T$ linear functionals.

Here, as elsewhere, $e^\theta = (0,\ldots,\overset{\theta}{1},\ldots,0)$ denotes the θ-th unit vector in R^Θ.

(iv) <u>(Comparison of performance functions)</u>.

To each Markov kernel σ from $\mathcal{F}$ to T corresponds a Markov kernel ρ from $\mathcal{E}$ to T so that:

$$\|\mu_\theta\rho-\nu_\theta\sigma\| < \varepsilon_\theta;\quad \theta\in\Theta.$$

Following Blackwell and LeCam and proceeding as in Torgersen 1969 we may now state and prove the basic:

<u>Theorem 2.1.</u> <u>(Equivalent rules for comparison)</u>.

Assume that the parameter set Θ is finite. Then the criteria (i)-(iv) are all equivalent.

<u>Proof</u>: Choose non negative measures μ and ν so that μ_θ, for each θ, has a density f_θ w.r.t. μ while ν_θ, for each θ, has density g_θ w.r.t. ν.

If $L = (L_\theta:\theta\in\Theta)$ is a family of real valued functions on Θ then $\min\limits_\rho \sum\limits_\theta \mu_\theta\rho L_\theta = \int[\bigwedge\limits_t\sum\limits_\theta L_\theta(t)f_\theta]d\mu$ and

$\min\limits_{\sigma} \sum\limits_{\theta} \nu_\theta \sigma L_\theta = \int [\bigwedge\limits_{t} \sum\limits_{\theta} L_\theta(t) g_\theta] d\nu$. It follows that (ii) may, after having been multiplied with (-1), be written:

$\int [\bigvee\limits_{t} \sum\limits_{\theta} U_\theta(t) f_\theta] d\mu > \int [\bigvee\limits_{t} \sum\limits_{\theta} U_\theta(t) g_\theta] d\nu - \sum\limits_{\theta} \varepsilon_\theta \|U_\theta\|$ where $U = -L$.

Putting $\phi(x) = \bigvee\limits_{t} \sum\limits_{\theta} U_\theta(t) x_\theta$; $x \in R^\Theta$ we see that (iii) is just a reformulation of (ii).

It suffices now, since the implications (iv)=>(i)=>(ii) are trivial to show that (ii)=>(iv). Assume then that (ii) is satisfied. Let σ be a Markov kernel from $\mathcal{F}$ to T and put $\Pi(L,\rho) = \sum\limits_{\theta} \mu_\theta \rho L_\theta - \sum\limits_{\theta} \nu_\theta \sigma L_\theta - \sum\limits_{\theta} \varepsilon_\theta \|L_\theta\|$ when $L = (L_\theta : \theta \in \Theta)$ is a family of real valued functions on T and ρ is a Markov kernel from $\mathcal{E}$ to T. Then Π is concave-convex in (L,ρ). Furthermore Π is continuous in ρ, for fixed L, if we topologize $\{\rho\}$ by the notion of convergence which states that a net $\{\rho^{(n)}\}$ of Markov kernels from $\mathcal{E}$ to T converges to a Markov kernel ρ from $\mathcal{E}$ to T if and only if $\int \delta \rho^{(n)}(t|\cdot) d\mu_\theta \to \int \delta \rho(t|\cdot) d\mu_\theta$ for each bounded measurable function δ on the sample space of $\mathcal{E}$. The weak compactness lemma implies that the set of Markov kernels from $\mathcal{E}$ to T is compact for this topology. Using minimax theory we see that there is a Markov kernel ρ_0 from $\mathcal{E}$ to T so that $\sup\limits_{L} \Pi(L,\rho_0) = \inf\limits_{\rho} \sup\limits_{L} \Pi(L,\rho) = \sup\limits_{L} \inf\limits_{\rho} \Pi(L,\rho)$ and the last quantity here is, by (iv), non positive. Thus $\Pi(L,\rho_0) < 0$ for each L. Choosing a particular real valued function g on T and a point $\theta^0 \in \Theta$ and then putting $L_\theta = g$ or $=0$ as $\theta = \theta^0$ or $\theta \neq \theta^0$ we see that $0 > \Pi(L,\rho_0) = (\mu_{\theta^0} \rho_0 - \nu_{\theta^0} \sigma)(g) - \varepsilon_{\theta^0} \|g\|$. Varying g and θ^0 we find that $\|\mu_\theta \rho_0 - \nu_\theta \sigma\| < \varepsilon_\theta$; $\theta \in \Theta$. □

Restricting attention in (iii) to funtions ϕ such that ϕ or $-\phi$ are projections on the coordinate spaces of R^Θ we see that the equivalent conditions (i)-(iv), implies:

$$(\S) \quad |\mu_\theta(\mathcal{X}) - \nu_\theta(\mathcal{Y})| < \varepsilon_\theta; \ \theta \in \Theta$$

which is conditions (i)-(iv) when T posesses just one element.

Assuming that the deficiency function ε satisfies (§), which it always does when $\mathcal{E}$ and $\mathcal{F}$ are experiments, we find that the "deficiency term" $\sum\varepsilon_\theta[\phi(-e^\theta)\vee\phi(e^\theta)]$ in (iii) may be replaced by the linear (in ϕ) term $\sum_\theta\varepsilon_\theta[\phi(-e^\theta)+\phi(e^\theta)]/2+\sum_\theta(\nu_\theta(\mathcal{Y})-\mu_\theta(\chi))[\phi(e^\theta)-\phi(-e^\theta)]/2$. We may then even assume that $\phi(-e^\theta)\underset{\theta}{\equiv}\phi(e^\theta)$ and thus get the simple expression $\sum_\theta\varepsilon_\theta\phi(e^\theta)$ for the deficiency term. We may also pass from a general sub linear function ϕ on R^Θ to a sub linear function which is monotonically decreasing (increasing) by replacing $\phi(x)$ with $\phi(x)-\sum_\theta x_\theta\phi(e^\theta)$ (with $\phi(x)+\sum_\theta x_\theta\phi(-e^\theta)$).

A substantial reduction is available when Θ is a two point set. In this case the fact that convex polygons may be decomposed as vector sums of triangles and line segments tells us that it suffices to consider 3 point sets T and thus functions ϕ which are maxima of at most 3 linear functionals. Actually additional assumptions will often guarantee that it suffices to consider two point sets T and thus functions ϕ which are maxima of at most 2 linear functionals. This is the case whenever $\mathcal{E}$ and $\mathcal{F}$ are experiments (i.e. dichotomies) or if $\mu_\theta,\nu_\theta>0$ and $\varepsilon_\theta=0$ for one of the two points θ in Θ. The latter case covers the usual case of majorization as well as the cases of weak majorization in Marshall and Olkin 1979. In the case of dichotomies a proof may be found in Torgersen 1970 and the other case follows by almost the same arguments. It is however not true that we in general may reduce comparison to comparison for two point sets T (testing problems) whenever Θ is a two point set.

Let us return to a general (i.e. not necessarily finite) parameter set Θ and let ε be a deficiency function. Then we shall say that the measure family $\mathcal{E}=(\mu_\theta:\theta\in\Theta)$ is ε-deficient w.r.t. the measure family $\mathcal{F}=(\nu_\theta:\theta\in\Theta)$ for k-point sets if the equivalent conditions (i)-(iv) are satisfied for the restrictions $\mathcal{E}|\Theta_0$, $\mathcal{F}|\Theta_0$ and $\varepsilon|\Theta_0$ for any non empty finite sub set Θ_0 of Θ. If $\mathcal{E}$ and $\mathcal{F}$ are experiments then the qualification "for k-point sets" may be replaced by the qualification "for k-decision problems".

If $\mathcal{E}$ is ε-deficient w.r.t. $\mathcal{F}$ for k-point sets for $k = 1,2,...$ then we shall say that $\mathcal{E}$ is ε-deficient w.r.t. $\mathcal{F}$ for ∞-point sets or just say that $\mathcal{E}$ is ε-deficient w.r.t. $\mathcal{F}$. Thus the notion of ε-deficiency for k-point sets is defined for all positive integers k and for $k = \infty$.

It is easily checked that $\mathcal{E}$ is ε'-deficient w.r.t. $\mathcal{F}$ for k'-point sets problems whenever $\mathcal{E}$ is ε-deficient w.r.t. $\mathcal{F}$ for k-point sets for $\varepsilon < \varepsilon'$ and $k > k'$.

We shall see below, without using any new ideas, that the randomization criterion is also valid for general measure families. The framework of measure theory which we are using here does not quite permit the full elegance of LeCam's formulations. Thus we might have worked with families in abstract L-spaces rather than with families of measures. No generality is however lost since any abstract L-space is, by Kakutani 1941, isomorphic to a L-space of bounded measures on some measurable space.

LeCam's randomization criterion generalized significantly earlier results in this direction by e.g. Blackwell and Bahadur (the theory of sufficiency). The formulation in terms of transitions has also to a large extent clarified the distinction between the essential statistical features behind this result on the one hand and various "technical" problems (often quite interesting) which occurs when we try to formulate this and related results within the traditional framework.

Let us before we embark on the extension of the randomization criterion introduce two concepts which we shall need in this connection. Firstly if $\mathcal{E} = (\chi, \mathcal{A}, \mu_\theta : \theta \in \Theta)$ is a measure family then we shall define the L-space of $\mathcal{E}$ as the linear space of finite measures on $\mathcal{A}$ which are dominated by finite measures of the form $\sum\{c_\theta|\mu_\theta| : \theta \in \Theta_0\}$ where the c_θ's are non negative and Θ_0 is a countable sub set of Θ. We shall also use the notation $\Gamma(\mathcal{E})$ for the Banach space of bounded measurable functions on $\mathcal{A}$ equipped with the supremum norm. It is then known that the adjoint space $\Gamma^*(\mathcal{E})$ of bounded linear functionals on $\Gamma(\mathcal{E})$ may be represented as the Banach space of bounded additive set functions on $\mathcal{A}$. The L-space $L(\mathcal{E})$ of $\mathcal{E}$ is then a closed sub space of $\Gamma^*(\mathcal{E})$.

If $\mathcal{E} = (\chi,\mathcal{A},\mu_\theta:\theta\in\Theta)$ and $\mathcal{F} = (\mathcal{Y},\mathcal{B},\nu_\theta:\theta\in\Theta)$ then any Markov kernel M from $(\chi,\mathcal{A})$ to $(\mathcal{Y},\mathcal{B})$ induces a map $\mu\to\mu M$ from $L(\mathcal{E})$ into $\Gamma(\mathcal{F})^*$, in fact into the space of finite measures on $\mathcal{B}$. This map, is linear, non negative (i.e. μM is non negative when μ is non negative) and it preserves total mass. We shall, following LeCam, call any map from $L(\mathcal{E})$ into $\Gamma(\mathcal{F})^*$ having these properties a transition. A very usefull property of the set of all transitions is that it, by Tychonoff, is compact for the pointwise topology on $L(\mathcal{E})\times\Gamma(\mathcal{F})$.

Assume now that $\mathcal{E} = (\chi,\mathcal{A},\mu_\theta:\theta\in\Theta)$ is ε-deficient w.r.t. $(\mathcal{Y},\mathcal{B},\nu_\theta:\theta\in\Theta)$. Let N denote the set of all triples $n = (\pi,\eta,F)$ where $\pi = (B_1,\ldots,B_k)$ is a measurable partitioning of $\mathcal{Y}$, $\eta = (y_1,\ldots,y_k)$ where $y_i\in B_i$, $i = 1,\ldots,k$ and F is a finite sub set of Θ. Direct the set N by defining that $n_1 = (\pi_1,\eta_1,F_1) > n_2 = (\pi_2,\eta_2,F_2)$ whenever $F_1 \supseteq F_2$ and each set in π_2 is a union of sets in π_1. Let $n = (\pi,\eta,F)\in N$ be described as above and consider $\{y_1,\ldots,y_k\}$ as a choice for the set T appearing in the statements of conditions (i)-(iv). Let σ_n denote the Markov kernel (function) from $\mathcal{F}$ to $\{y_1,\ldots,y_k\}$ which maps y into y_i when $y\in B_i$. Then, by (iv), there is a Markov kernel ρ_n from $\mathcal{E}$ to the same space such that $\|\mu_\theta\rho_n-\nu_\theta\sigma_n\|<\varepsilon_\theta$ when $\theta\in F$. Put $M_n(B|x) = \sum I_B(y_i)\rho_n(y_i|x)$ when $B\in\mathcal{B}$ ans $x\in\chi$. Then M_n is a Markov kernel from $(\chi,\mathcal{A})$ to $(\mathcal{Y},\mathcal{B})$ and thus defines a transition from $L(\mathcal{E})$ to $\Gamma(\mathcal{F})^*$. By compactness there is a transition M which is a point of accumulation for the net of transitions defined by the net $\{M_n\}$ of Markov kernels. It follows then that $\mu_\theta Mg-\nu_\theta g<\varepsilon_\theta\|g\|$ for all $\theta\in\Theta$ and for all $g\in\Gamma(\mathcal{F})$. Hence $\|\mu_\theta M-\nu_\theta\|<\varepsilon_\theta$; $\theta\in\Theta$. This proves:

Theorem 2.2. (The randomization criterion for comparison of measure families.)

The measure family $\mathcal{F} = (\mu_\theta:\theta\in\Theta)$ is ε-deficient w.r.t. the measure family $\mathcal{F} = (\nu_\theta:\theta\in\Theta)$ if and only if there is a transition M from the L-space of $\mathcal{E}$ to the L-space of bounded additive set functions on $\mathcal{B}$ such that:

$$\|\mu_\theta M-\nu_\theta\| < \varepsilon_\theta;\ \theta\in\Theta.$$

Remark 1. We may, using the same arguments as in LeCam 1964 always modify M such that it maps $L(\mathcal{E})$ into $L(\mathcal{F})$.

Remark 2. It follows that one possible answer to the problem raised in example (1.1) is:

"If and only if there is a transition M so that $\|a_\theta P_\theta M - b_\theta Q_\theta\| < \varepsilon_\theta$ for all $\theta \in \Theta$". If $a_\theta \underset{\theta}{\equiv} 1$ then this implies that $|b_\theta - 1| < \varepsilon_\theta$; $\theta \in \Theta$. If $b_\theta = 1-\varepsilon_\theta$ then the "θ-th" inequality states that $P_\theta M > (1-\varepsilon_\theta)Q_\theta$. If, on the other hand, $b_\theta = 1+\varepsilon_\theta$ then the θ-th inequality states that $P_\theta M < (1+\varepsilon_\theta)Q_\theta$.

If we want to ensure that the transition M appearing in the randomization criterion is representable as a Markov kernel then various regularity conditions may be invoked. Say that a measurable space $(\chi,\mathcal{A})$ is Euclidean if it is Borel isomorphic to a Borel sub set of $[0,1]$. Say also that a measure family $\mathcal{E} = (\mu_\theta : \theta \in \Theta)$ is coherent if all bounded linear functionals on $L(\mathcal{E})$ are representable as bounded measurable functions. If $\mathcal{E} = (\mu_\theta : \theta \in \Theta)$ is coherent and if $(\mathcal{Y},\mathcal{B})$ is Euclidean then any transition from $L(\mathcal{E})$ to the L-space of finite measures on $(\mathcal{Y},\mathcal{B})$ is representable as a Markov kernel.

If the families $\mathcal{E}$ and $\mathcal{F}$ as well as the deficiency function ε are invariant under the actions of a group then the kernel M may, see e.g. LeCam 1964 and 1974 and Torgersen 1972c, under amenability conditions be chosen invariant.

We shall not pursue the general theory of comparison of measure families further here. Most of the results are more or less straight forward generalizations of known results from the theory of comparison of experiments. Thus we may define the deficiency $\delta_k(\mathcal{E},\mathcal{F})$ of $\mathcal{E}$ w.r.t. $\mathcal{F}$ for k-point sets (k-decisions when $\mathcal{E}$

and $\mathcal{F}$ are experiments) as the smallest (it exists) constant ε such that $\mathcal{E}$ is ε-deficient w.r.t. $\mathcal{F}$ for k-point sets. Symmetrizing the deficiency we define the deficiency distance $\Delta_k(\mathcal{E},\mathcal{F})$ between the measure families *$\mathcal{E} = (\mu_\theta : \theta\in\Theta)$ and $\mathcal{F} = (\nu_\theta : \theta\in\Theta)$ for k-point sets* as the largest of the numbers $\delta_k(\mathcal{E},\mathcal{F})$ and $\delta_k(\mathcal{F},\mathcal{E})$.

If $\delta_k(\mathcal{E},\mathcal{F}) = 0$ then we shall say that *$\mathcal{E}$ majorizes $\mathcal{F}$ for k-point sets* and write this $\mathcal{E} \underset{k}{>} \mathcal{F}$. Then $\underset{k}{>}$ is an ordering for measure families and this ordering extends the notion of one experiment being at least as informative as another for k-decision problems. Similarly we shall say that $\mathcal{E}$ and $\mathcal{F}$ are equivalent for k-point sets if $\mathcal{E} \underset{k}{>} \mathcal{F}$ and $\mathcal{F} \underset{k}{>} \mathcal{E}$ i.e. if $\Delta_k(\mathcal{E},\mathcal{F}) = 0$.

The qualification "for k-point sets" as well as the subscript k may be omitted when $k = \infty$.

Although most of the known results from the theory of comparison of experiments generalizes easily there are a few surprises. Thus equivalence for two point sets (i.e. for testing problems in the case of experiments) does no longer imply full equivalence. Equivalence for 3 point sets does however imply full equivalence and this in turn is equivalent to the condition that the vector lattices generated by the measure families are isometrically isomorphic by a (and hence the) correspondence making the θ-th measures correspond to each other for each $\theta\in\Theta$. We may proceed, as in Torgersen 1972 a and b, and generalize the theory of sufficiency to the case of general measure families.

A useful characteristic of statistical experiments are, as shown by LeCam, certain functionals called conical measures. These are essentially the functionals which to a sub linear function ψ and for a given measure family $(\mu_\theta : \theta\in\Theta)$ associates the number $\int\psi(d\mu_\theta : \theta\in\Theta)$ according to the recipe given before our statements of principles (i)-(iv). Most of the basic properties of this characteristic extend without difficulties to measure families.

Let us now return to examples 1.2-1.5 in the introduction and see if this theory can contribute something in each of these situations.

3. LOCAL (FIXED SAMPLE SIZE) COMPARISON OF STATISTICAL EXPERIMENTS.

We shall assume throughout this section that the parameter set Θ is a sub set of R^m for some positive integer m. We shall be concerned with local comparison within small neighbourhoods of a given point θ^0 belonging to the interior of Θ.

An experiment $\mathcal{E} = (\chi, \mathcal{A}, P_\theta : \theta\in\Theta)$ will be called <u>differentiable (in the first mean) at θ^0</u> if the map $\theta\to P_\theta$ from Θ to the Banach space of finite measures on $\mathcal{A}$ is Frechet differentiable.

If we let $e^i \equiv (0,\ldots,\overset{(i)}{1},\ldots,0)$ denote the i-th unit vector in R^m then $\mathcal{E}$ is differentiable at θ^0 if and only if the limits (partial derivatives) $[\partial P_\theta/\partial\theta_i]_{\theta^0} = \lim_{t\to 0}(P_{\theta^0+te^i} - P_{\theta^0})/t$ exists; $i = 1,\ldots,m$ and $\|P_\theta - P_{\theta^0} - \sum_{i=1}^m [\partial P_\theta/\partial\theta_i]_{\theta^0}(\theta_i-\theta_i^0)\|/\|\theta-\theta^0\|\to 0$ as $\theta\to\theta^0$.

The theory for differentiable experiments which will be presented below is by and large self contained. We have however described some results without giving complete proofs. The missing proofs are then, if not otherwise stated, given in Torgersen 1972 a and b.

The notion of differentiability used here is weaker than the usual notion of differentiability in quadratic mean. The latter notion leads, see LeCam 1974 or Millar 1983, to basic results concerning the asymptotic behaviour of replicated experiments. If we merely assume differentiability in the first mean, then these results need not hold. We shall not be concerned with asymptotic theory in this sense here and then the chosen notion of differentiability appears to be appropriate - although there are many other possibilities for weaker as well as for stronger notions.

Let us return to differentiability (in the first mean) as defined above. If $\mathcal{E} = (P_\theta : \theta\in\Theta)$ is differentiable at θ^0 then the family consisting of the (m+1) measures $(P_{\theta^0}, [\partial P_\theta/\partial\theta_i]_{\theta=\theta^0}; i = 1,\ldots,m)$ will be called the <u>first order characterization of $\mathcal{E}$ at θ^0</u>. We might instead have used the measure family $(\mu_h : h\in R^m)$ where $\mu_h = P_{\theta^0} + \sum_{i=1}^m h_i[\partial P_\theta/\partial\theta_i]_{\theta=\theta_0}$; $h\in R^m$. If we insist that the local approximation should be an experiment then we might replace μ_h by $|\mu_h|/\|\mu_h\|$ here.

We shall find it convenient to write $\dot{P}_{\theta^0,i}$ for the partial derivative of P_θ w.r.t. θ_i at θ^0 i.e. $\dot{P}_{\theta^0,i} = [\partial P_\theta/\partial\theta_i]_{\theta^0}$. The first order characterization of $\mathcal{E}$ at θ^0 will be denoted as $\dot{\mathcal{E}}_{\theta^0}$. Thus $\dot{\mathcal{E}}_{\theta^0} = (P_{\theta^0},\dot{P}_{\theta^0,1},\ldots,\dot{P}_{\theta^0,m})$ if $\mathcal{E} = (P_\theta:\theta\in\Theta)$ is differentiable at $\theta = \theta^0$.

Let $e_i = (0,\ldots,\overset{(i)}{1},\ldots,0)$ denote the i-th unit vector of R^m. If $\mathcal{E} = (P_\theta:\theta\in\Theta)$ is differentiable at θ^0 then the experiments $(P_{\theta^0+te^i}:|t|<\varepsilon)$ are all well defined and differentiable in θ^0 provided ε is sufficiently small. Furthermore P_θ is approximable by $\mu_\theta = P_{\theta^0}+\sum_{i=1}^{m}(P_{\theta^0+(\theta_i-\theta_i^0)e^i}-P_{\theta^0})$ in the sense that $\|P_\theta-\mu_\theta\|/\|\theta-\theta^0\|\to 0$ as $\theta\to\theta^0$.

Conversely these two conditions guarantee that $\mathcal{E}$ is differentiable at $\theta = \theta^0$. The usual arguments from calculus imply that $\mathcal{E}$ is differentiable at θ^0 provided the partial derivatives exists and are continuous within some neighbourhood of θ^0.

What kind of a measure family is the first order approximation of a differentiable experiment? Note first that if $\dot{\mathcal{E}}_{\theta^0} = (P_{\theta^0},\dot{P}_{\theta^0,1},\ldots,\dot{P}_{\theta^0,m})$ is the first order approximation of $\mathcal{E}$ at θ^0 then $P_{\theta^0}>>\dot{\mathcal{E}}_{\theta^0}$ and each measure $\dot{P}_{\theta^0,i}$; $i = 1,\ldots,m$, has total mass zero. All general properties of first order approximations may be deduced from these two properties. In fact any measure family $(\sigma,\sigma_1,\ldots,\sigma_m)$ such that the measures $\sigma_1,\ldots,\sigma_m$ all have total mass zero and such that σ is a probability measure dominating $\sigma_1,\ldots,\sigma_m$ is of the form $\dot{\mathcal{E}}_{\theta^0}$ where $\mathcal{E} = (P_\theta:\theta\in\Theta)$ and P_θ = constant$\cdot|\sigma+\sum_i(\theta_i-\theta_i^0)\sigma_i|$ when $\theta\in\Theta$.

It is easily checked that if $\mathcal{E}$ is differentiable at $\theta = \theta^0$ and if $\mathcal{E}$ is at least as informative as $\mathcal{F}$ then $\mathcal{F}$ is also differentiable at $\theta = \theta^0$. In fact if $\mathcal{E} = (P_\theta:\theta\in\Theta)$ is 0-deficient w.r.t. $\mathcal{F} = (Q_\theta:\theta\in\Theta)$ then $Q_\theta \equiv P_\theta M$ for a transition M and then $\dot{Q}_{\theta^0,i} = \dot{P}_{\theta^0,i}M$; $i = 1,\ldots,m$. Thus $\dot{\mathcal{F}}_{\theta^0}$ is obtained from $\dot{\mathcal{E}}_{\theta^0}$ by the transition M.

Likewise finite products of experiments which are differentiable at θ^0 are themselves differentiable at θ^0. More precisely if $\mathcal{E} = (P_\theta:\theta\in\Theta)$ and $\mathcal{F} = (Q_\theta:\theta\in\Theta)$ are differentiable at θ^0 and if $\mathcal{G} = \mathcal{E}\times\mathcal{F}$ then the first order approximation $\mathcal{G}_{\theta^0}$ of $\mathcal{G}$ is the measure family:

$$(P_{\theta^0}\times Q_{\theta^0},\ P_{\theta^0}\times\dot{Q}_{\theta^0,i} + \dot{P}_{\theta^0,i}\times Q_{\theta^0};\ i = 1,\dots,m).$$

The local properties of the differentiable experiment $\mathcal{E} = (\chi,\mathcal{A},P_\theta:\theta\in\Theta)$ at θ^0 is, as we shall make clearer later on, determined by the distribution $F(\cdot|\theta^0,\mathcal{E})$ of the random vector $(d\dot{P}_{\theta^0,i}/dP_{\theta^0};\ i = 1,\dots,m)$ under P_{θ^0}.

Clearly $\int xF(dx|\theta^0,\mathcal{E}) = 0$ and any probability distribution on R^m having expectation zero is of the form $F(\cdot|\theta^0,\mathcal{E})$ for some differentiable experiment $\mathcal{E}$.

We shall here only point out that if $\mathcal{E}$ and $\mathcal{F}$ are both differentiable at θ^0 and if $\mathcal{G} = \mathcal{E}\times\mathcal{F}$ then $F(\cdot|\theta^0,\mathcal{G}) = F(\cdot|\theta^0,\mathcal{E})*F(\cdot|\theta^0,\mathcal{F})$ where $*$ denotes convolution.

We shall later on see how $F(\cdot|\theta^0,\mathcal{E})$ and $F(\cdot|\theta^0,\mathcal{F})$ must be related in order to ensure that $\mathcal{E}$ is locally at least as informative as $\mathcal{F}$ at θ^0.

Consider now two experiments $\mathcal{E} = (P_\theta:\theta\in\Theta)$ and $\mathcal{F} = (Q_\theta:\theta\in\Theta)$ which both are differentiable at $\theta = \theta^0$. Equip R^m with the L-norm $\| \ \|:x\to\sum_i|x_i|$. Let for each $\varepsilon>0$, the restrictions of $\mathcal{E}$ and $\mathcal{F}$ to the ε-ball $N(\theta^0,\varepsilon) = \{\theta:\|\theta-\theta^0\|<\varepsilon\}$ be denoted by, respectively $\mathcal{E}_\varepsilon$ and $\mathcal{F}_\varepsilon$.

The problem of local comparison of $\mathcal{E}$ and $\mathcal{F}$ within small neighbourhoods of θ^0 may now be discussed in terms of the behaviour of the deficiency of $\delta(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)$ for small values of ε. If $\mathcal{M}_i$ denotes the totally non informative experiment then continuity implies that the deficiencies $\delta(\mathcal{M}_i,\mathcal{F}_\varepsilon)$ and $\delta(\mathcal{M}_i,\mathcal{E}_\varepsilon)$ both $\to0$ as $\varepsilon\to0$. Thus $\delta_k(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)\to0$ as $\varepsilon\to0$ for each $k = 1,2,\dots,\infty$.

Let us next determine the rate of the convergence of $\delta_k(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)$ as $\varepsilon\to 0$. If $\mathcal{E} = (P_\theta : \theta\in\Theta)$ is differentiable in θ^0 then we may expand P_θ as $P_\theta = P_{\theta^0}+\sum(\theta-\theta^0)_i\dot{P}_{\theta^0,i}+\tau(\mathcal{E},\theta^0,\theta)$ where the measure $\tau(\mathcal{E},\theta^0,\theta)$ is defined by this expansion. The differentiability assumption implies then that $\|\tau(\mathcal{E},\theta^0,\theta)\|/\|\theta-\theta^0\|\to 0$ as $\theta\to\theta^0$.

We shall in the following find it convenient to utilize the symbol o in the usual way i.e. o denotes any real valued function on $]0,\infty[$ such that $o(t)/t\to 0$ as $t\to 0$.

If $\mathcal{E} = (P_\theta : \theta\in\Theta)$ and $\mathcal{F} = (Q_\theta : \theta\in\Theta)$ are both differentiable in θ^0 then the smallest (it exists) constant η such that $\dot{\mathcal{E}}_{\theta^0}$ is $(0,\eta,\ldots,\eta)$ deficient w.r.t. $\dot{\mathcal{F}}_{\theta^0}$ for k-point sets will be called the local deficiency at θ^0 of $\mathcal{E}$ w.r.t. $\mathcal{F}$ for k-decision problems. The local deficiency at θ^0 of $\mathcal{E}$ w.r.t. $\mathcal{F}$ for k-decision problems will be denoted as $\dot{\delta}_{k,\theta^0}(\mathcal{E},\mathcal{F})$. Here, as elsewhere, we may omit the qualification "for k-decision problems" and the subscript k if $k = \infty$.

The local deficiency determines the rate of convergence of $\delta_k(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)/\varepsilon$ as $\varepsilon\to 0$ by:

Theorem 3.1. (Asymptotic behaviour of deficiencies within small neighbourhoods).

With the notations introduced above for differentiable experiments $\mathcal{E} = (P_\theta : \theta\in\Theta)$ and $\mathcal{F} = (Q_\theta : \theta\in\Theta)$ we have

$$\delta_k(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon) \leq \varepsilon\dot{\delta}_{k,\theta^0}(\mathcal{E},\mathcal{F}) + \bar{o}(\varepsilon)$$

where $\bar{o}(\varepsilon) = \sup\{\|\tau(\mathcal{E},\theta^0,\theta)\|+\|\tau(\mathcal{F},\theta^0,\theta)\| : \|\theta-\theta^0\|\leq\varepsilon\}$. Furthermore

$$\delta_k(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)/\varepsilon \to \dot{\delta}_{k,\theta^0}(\mathcal{E},\mathcal{F}) \quad\text{as}\quad \varepsilon\to 0.$$

Remark 1. The local deficiency may by the randomization criterion be expressed as:

$$\dot{\delta}_{\theta^0}(\mathcal{E},\mathcal{F}) = \min\{\max_i \|\dot{P}_{\theta^0,i}M-\dot{Q}_{\theta^0,i}\|:P_{\theta^0}M = Q_{\theta^0}\}$$

where M varies within the set of transitions (randomizations) from $L(\mathcal{E})$ to the L-space $\Gamma(\mathcal{E})^*$ of bounded additive set functions on the sample space of $\mathcal{F}$.

We may, see remark 1 after theorem 2.2, limit our attention to transitions from $L_1(P_{\theta^0})$ to $L_1(Q_{\theta^0})$.

Remark 2. The proof implies that the statements of the theorem remain true, if, for each $\varepsilon>0$, $\mathcal{E}_\varepsilon$ and $\mathcal{F}_\varepsilon$ are replaced by the restrictions of the experiments $\mathcal{E}$ and $\mathcal{F}$ to the sub set of $N(\theta^0,\varepsilon)$ consisting of the 2m points (vertices) $\theta^0\pm(0,\ldots,\overset{(i)}{\varepsilon},\ldots,0)$; $i = 1,\ldots,m$.

Proof: Consider first the case $k = \infty$. If $P_{\theta^0}M = Q_{\theta^0}$ for a transition M and if $\|\theta-\theta^0\|<\varepsilon$ then:

$$\begin{aligned}
&\|P_\theta M-Q_\theta\| = \|(P_\theta-P_{\theta^0})M-(Q_\theta-Q_{\theta^0})\| \\
&= \|\sum_i(\theta-\theta^0)_i(\dot{P}_{\theta^0,i}M-\dot{Q}_{\theta^0,i}) + \tau(\mathcal{E},\theta^0,\theta)M-\tau(\mathcal{F},\theta^0,\theta)\| \\
&< \sum_i|(\theta-\theta^0)_i|\max_i\|\dot{P}_{\theta^0,i}M-\dot{Q}_{\theta^0,i}\| + \|\tau(\mathcal{E},\theta^0,\theta)\| + \|\tau(\mathcal{F},\theta^0,\theta)\| \\
&< \varepsilon\max_i\|\dot{P}_{\theta^0,i}M-\dot{Q}_{\theta^0,i}\| + o(\varepsilon)
\end{aligned}$$

so that $\delta(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)<\varepsilon\dot{\delta}_{\theta^0}(\mathcal{E},\mathcal{F})+o(\varepsilon)$. It follows that $\limsup\delta(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)/\varepsilon<\dot{\delta}_{\theta^0}(\mathcal{E},\mathcal{F})$ as $\varepsilon\to0$.

Consider on the other and any number $c>\liminf_{\varepsilon\to0}\delta(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)/\varepsilon$. Then $\delta(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)<c\varepsilon$ for all ε belonging to a some sequence $\varepsilon_1,\varepsilon_2,\ldots$ which decreases to zero.- Assume that ε belongs to this sub sequence. The randomization criterion (theorem

2.2) yields then a transition M_ε from $L(\mathcal{E})$ to $\Gamma(\mathcal{F})^*$ so that $\|P_\theta M_\varepsilon - Q_\theta\| < c\varepsilon$ when $\theta \in N(\theta^0, \varepsilon)$. We may, by compactness, assume that M_ε converges pointwise on $L(\mathcal{E}) \times \Gamma(\mathcal{F})$ to a transition M from $L(\mathcal{E})$ to $\Gamma(\mathcal{F})^*$.

Expanding around θ^0 we find that

$$\|P_{\theta^0}M_\varepsilon - Q_{\theta^0} + \sum_i(\theta_i - \theta_i^0)(\dot{P}_{\theta^0,i}M_\varepsilon - \dot{Q}_{\theta^0,i}) + \tau(\mathcal{E},\theta^0,\theta)M_\varepsilon - \tau(\mathcal{F},\theta^0,\theta)\| < c\varepsilon$$

when $\|\theta - \theta^0\| \leq \varepsilon$. In particular $\|P_{\theta^0}M_\varepsilon - Q_{\theta^0}\| < c\varepsilon$. $\varepsilon \to 0$ yields then $P_{\theta^0}M = Q_{\theta^0}$.

Putting $\theta' = \theta^0 + (0,\ldots,\overset{(i)}{\varepsilon},\ldots,0)$ and $\theta'' = \theta^0 - (0,\ldots,\overset{(i)}{\varepsilon},\ldots,0)$ we find that

$$\|\varepsilon(\dot{P}_{\theta^0,i}M_\varepsilon - \dot{Q}_{\theta^0,i}) + P_{\theta^0}M_\varepsilon - Q_{\theta^0} + \tau(\mathcal{E},\theta^0,\theta')M_\varepsilon - \tau(\mathcal{F},\theta^0,\theta')\| < c\varepsilon$$

and

$$\|\varepsilon(\dot{P}_{\theta^0,i}M_\varepsilon - \dot{Q}_{\theta^0,i}) - (P_{\theta^0}M_\varepsilon - Q_{\theta^0}) - (\tau(\mathcal{E},\theta^0,\theta'')M_\varepsilon - \tau(\mathcal{F},\theta^0,\theta''))\| < c\varepsilon.$$

Hence $2\varepsilon\|\dot{P}_{\theta^0,i}M_\varepsilon - Q_{\theta^0,i}\| + o(\varepsilon) < 2c\varepsilon$. Dividing through by ε and letting $\varepsilon \to 0$ we find that $\|\dot{P}_{\theta^0,i}M - \dot{Q}_{\theta^0,i}\| < c$. Thus $c > \dot{\delta}_{\theta^0}(\mathcal{E},\mathcal{F})$. It follows that $\liminf \delta(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)/\varepsilon > \dot{\delta}_{\theta^0}(\mathcal{E},\mathcal{F})$ as $\varepsilon \downarrow 0$ so that $\delta(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)/\varepsilon \to \dot{\delta}_{\theta^0}(\mathcal{E},\mathcal{F})$ as $\varepsilon \to 0$. This completes the proof when $k = \infty$.

The case of a finite k may be reduced to the case $k = \infty$ as follows: Assume that $k < \infty$ and let $(\mathcal{Y},\mathcal{B})$ be the sample space of $\mathcal{F}$. Let $\mathcal{B}_0$ be a sub algebra of $\mathcal{B}$ containing at most 2^k sets. The "$k = \infty$" part of the theorem implies that $\delta(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon|\mathcal{B}_0) < \varepsilon\dot{\delta}_{\theta^0}(\mathcal{E},\mathcal{F}|\mathcal{B}_0) + o(\varepsilon)$ where $o(\varepsilon)$ is defined as in the theorem and, consequently, does not depend on $\mathcal{B}_0$. Taking the supremum for all such algebras $\mathcal{B}_0$ we find by corollary 6 in Torgersen 1970 that $\delta_k(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon) < \varepsilon \sup \dot{\delta}_{\theta^0}(\mathcal{E},\mathcal{F}|\mathcal{B}_0) + o(\varepsilon)$. The sumpremum on the right hand side is, by the essentially the same argument as was used in the proof of this corollary precisely the "dotted" deficiency $\dot{\delta}_{k,\theta^0}(\mathcal{E},\mathcal{F})$. Hence $\delta_k(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon) < \varepsilon\dot{\delta}_{k,\theta^0}(\mathcal{E},\mathcal{F}) + o(\varepsilon)$. If follows that $\limsup_{\varepsilon \to 0} \delta_k(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)/\varepsilon < \dot{\delta}_{k,\theta^0}(\mathcal{E},\mathcal{F})$.

On the other hand $\liminf_{\varepsilon\to 0} \delta_k(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)/\varepsilon > \liminf_{\varepsilon\to 0} \delta(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon|\mathcal{B}_0)/\varepsilon = \dot{\delta}_{\theta^0}(\mathcal{E},\mathcal{F}|\mathcal{B}_0)$ for any sub algebra $\mathcal{B}_0$ of $\mathcal{B}$ containing at most 2^k sets.

Taking the sumpremum for all such algebras $\mathcal{B}_0$ we see that $\liminf_{\varepsilon\to 0} \delta_k(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon)/\varepsilon > \dot{\delta}_{k,\theta^0}(\mathcal{E},\mathcal{F})$. □

Define the <u>local deficiency distance at θ^0 between $\mathcal{E}$ and $\mathcal{F}$</u> for k-decision problems as the largest of the numbers $\dot{\delta}_{k,\theta^0}(\mathcal{E},\mathcal{F})$ and $\dot{\delta}_{k,\theta^0}(\mathcal{E},\mathcal{F})$. This number will be denoted ad $\dot{\Delta}_{k,\theta^0}(\mathcal{E},\mathcal{F})$. Here the subscript k and the qualification "for k-decision problems" may be omitted when $k = \infty$. It is easily checked that $\dot{\Delta}_{k,\theta^0}$ is a pseudodistance for experiments. The theorem yields readily:

<u>Corollary 3.2. (Asymptotic behaviour of deficiency distances within small neighbourhoods)</u>.

$\Delta_k(\mathcal{E}_\varepsilon,\mathcal{F}_\varepsilon) < \varepsilon\dot{\Delta}_{k,\theta^0}(\mathcal{E},\mathcal{F}) + o(\varepsilon)$ where $o(\varepsilon)$ is defined as in the theorem. Furthermore "=" holds for some function $o(\varepsilon)$.

Related to the local deficiencies are the local orderings and the local equivalences of experiments. Thus we shall say that <u>$\mathcal{E}$ is locally at least as informative as $\mathcal{F}$ for k-decision problems at $\theta = \theta^0$</u> if $\dot{\delta}_{k,\theta^0}(\mathcal{E},\mathcal{F}) = 0$. This define, for $k = 1,2,\ldots,\infty$, a partial ordering $\underset{k,\theta^0}{>}$ of differentiable experiments. If $\underset{k,\theta^0}{>}$ and $\mathcal{F} \underset{k,\theta^0}{>} \mathcal{E}$ i.e. if $\dot{\Delta}_{k,\theta^0}(\mathcal{E},\mathcal{F}) = 0$ then we shall say that <u>$\mathcal{E}$ and are locally equivalent for k-decision problems at $\theta = \theta^0$</u>.

This defines for each $k = 1,2,\ldots,\infty$ a local equivalence $\underset{k,\theta^0}{\sim}$ for differentiable experiments. It turns out, however, that the non trivial equivalences $\underset{k,\theta^0}{\sim}$; $k = 2,3,\ldots,\infty$ are all the same. More generally it may be shown that the pseudodistances $\dot{\Delta}_{k,\theta^0}$; $k = 2,3,\ldots,\infty$ all define the same notion of convergence.

What is the statistical significance of local deficiencies and related notions? Some insight may be gained from the following characterization in terms of performance functions:

Proposition 3.3. (Local comparison of performance functions).

Let $\mathcal{E} = (P_\theta : \theta \in \Theta)$ and $\mathcal{F} = (Q_\theta : \theta \in \Theta)$ both be differentiable at $\theta = \theta^0$.

Let $(T, \mathcal{S})$ be a decision space and consider a decision rule σ in $\mathcal{F}$. Then there is a decision rule ρ in $\mathcal{E}$ so that

$$\limsup_{\theta \to \theta^0} \| P_\theta \rho - Q_\theta \sigma \| / \| \theta - \theta^0 \| < \dot{\delta}_{\theta^0}(\mathcal{E}, \mathcal{F}).$$

Furthermore $(T, \mathcal{S})$ and σ may be chosen so that

$$\limsup_{\theta \to \theta^0} \| P_\theta \rho - Q_\theta \sigma \| / \| \theta - \theta^0 \| > \dot{\delta}_{\theta^0}(\mathcal{E}, \mathcal{F})$$

for all decision rules ρ in $\mathcal{E}$.

Remark. If $(T, \mathcal{S})$ is a k-decision space then the first inequality may be sharpened by replacing $\dot{\delta}_{\theta^0}(\mathcal{E}, \mathcal{F})$ by the usually smaller number $\dot{\delta}_{k,\theta^0}(\mathcal{E}, \mathcal{F})$. Furthermore if $(T, \mathcal{S})$ is a k-decision space then σ in $\mathcal{F}$ may be chosen so that

$$\limsup_{\theta \to \theta^0} \| P_\theta \rho - Q_\theta \sigma \| / \| \theta - \theta^0 \| > \dot{\delta}_{k,\theta^0}(\mathcal{E}, \mathcal{F})$$

for all decision rules ρ in $\mathcal{E}$.

This may be seen by applying the proposition to the restrictions of $\mathcal{F}$ to algebras of events containing at most 2^k events.

Proof: Note first that for any transition M:

$\limsup_{\theta\to\theta_0}\|P_\theta M-Q_\theta\|/\|\theta-\theta^0\| = \max_i\|\dot{P}_{\theta^0,i}M-\dot{Q}_{\theta^0,i}\|$ or $=\infty$ as $P_{\theta^0}M = Q_{\theta^0}$ or $P_{\theta^0}M\neq Q_{\theta^0}$. The first statement of the proposition follows now by putting $\rho = M\sigma$ where $P_{\theta^0}M = Q_{\theta^0}$ and $\max_i\|\dot{P}_{\theta^0,i}M-\dot{Q}_{\theta^0,i}\| = \dot{\delta}_{\theta^0}(\mathcal{E},\mathcal{F})$. The second statement follows by observing that we may let $(T,\mathcal{S})$ be the sample space of $\mathcal{F}$ and then choose σ as the identity map. □

If ϕ is a sub linear function on $R\times R^m$ and if $\mathcal{E} = (P_\theta:\theta\in\Theta)$ is differentiable at $\theta = \theta^0$ then $\int\phi(dP_{\theta^0},d\dot{P}_{\theta^0,1},\ldots,d\dot{P}_{\theta^0,m}) = \int\phi(1,x_1,\ldots,x_m)F(dx|\theta^0,\mathcal{E})$. It follows readily that $\mathcal{E} = (P_\theta:\theta\in\Theta)$ is locally at least as informative as $\mathcal{F}$ at $\theta = \theta^0$ if and only if $\int\phi dF(\cdot|\theta^0,\mathcal{E})\geq\int\phi dF(\cdot|\theta^0,\mathcal{F})$ for all convex functions ϕ on R^m. This in turn is, as we shall see in the next section, equivalent to the condition that $F(\cdot|\theta^0,\mathcal{E}) = DF(\cdot|\theta^0,\mathcal{F})$ for a Markov kernel (randomization) D from R^m to R^m such that $\int yD(dx|y) = y$ for all points y in R^m. A Markov kernel having the latter property is called a dilation.

The Fisher information matrix $I(\theta^0,\mathcal{E})$ is the covariance matrix of $F(\cdot|\theta^0,\mathcal{E})$ - provided of course that $F(\cdot|\theta^0,\mathcal{E})$ posesses finite second order moments. It follows that if $\mathcal{E}$ is locally at least as informative as $\mathcal{F}$ at θ^0 and if the Fisher information matrix of $\mathcal{E}$ at θ^0 exists, then the Fisher information matrix of $\mathcal{F}$ at θ^0 also exists and then the difference matrix $I(\theta^0,\mathcal{E})-I(\theta^0,\mathcal{F})$ is non negative definite. This proves the local, and hence the "global", monotonicity of the Fisher information matrix.

Example 3.4. (Local orderings of linear normal models).

Let $\mathcal{E}_A$ for each $n_A\times p$ matrix A' denote the linear normal experiment $(N(A'\beta,I_A):\beta\in R^p)$ where I_A denotes the $n_A\times n_A$ unit matrix. The Fisher information matrix of $\mathcal{E}_A$ is the $p\times p$ matrix AA'. If B is another matrix with p rows and if $\mathcal{E}_A$ is locally at least as informative as $\mathcal{E}_B$ then, by the remarks above, $AA'\geq BB'$. The ordering "$\geq$" for matrices which is used in this example is the ordering which declares that $M\geq N$ if $M-N$ is non negative definite.

If $AA' \geq BB'$ then $\mathcal{E}_A \sim \mathcal{E}_B \times \mathcal{E}_M$ where M is the non negative definite squareroot of $AA'-BB'$. It follows that the local orderings as well as the global orderings of linear normal models with known variances coincides with the usual ordering of Fisher information matrices.

When we turn to the case of unknown variances then the matters are a bit more involved. Let $\mathcal{F}_A$ denote the experiment $(N(A'\beta, \sigma^2 I_A): \beta \in R^p, \sigma > 0)$ where A' and I_A are as above. The $(p+1) \times (p+1)$ Fisher information matrix of $\mathcal{F}_A$ w.r.t. the unknown parameters $\beta_1, \ldots, \beta_p$ and σ is:

$$I(\beta, \sigma, \mathcal{F}_A) = \begin{pmatrix} AA'/\sigma^2, & 0 \\ 0, & 2n_A/\sigma^2 \end{pmatrix}$$

By Hansen and Torgersen 1974 the experiment $\mathcal{F}_A$ is at least as informative as $\mathcal{F}_B$ if and only if $AA' \geq BB'$ and $n_A \geq n_B + \text{rank}(AA'-BB')$. In fact, Torgersen 1984, this is equivalent to the condition that $\mathcal{F}_A \sim \mathcal{F}_B \times \mathcal{F}_C$ (i.e. $AA' = BB'+CC'$ and $n_A = n_B + n_C$) for some $n_C \times p$ matrix C'.

The above mentioned result of Hansen and Torgersen was extended by Lehmann 1983 to the case of multivariate regression. In that case $\mathcal{F}_A$ is, for each $n_A \times p$ matrix A', realized by observing a random $n_A \times q$ matrix X such that $EX = A'\beta$ for a unknown $p \times q$ matrix β while the rows of X are independent and multinormally distributed with the same unknown non singular covariance matrix σ^2. (Actually Lehmann assumes that A is in a "reduced" form where rank $A = p$.) If we compare the distribution of the minimal sufficient statistics we see again that $\mathcal{F}_A \sim \mathcal{F}_B \times \mathcal{F}_C$ when $AA' = BB'+CC'$ and $n_A = n_B + n_C$. On the other hand if we consider e.g. the first column of β and restrict β and σ^2 such that all the other entries of β are 0 while σ^2 is a diagonal matrix such that all the diagonal elements are known except the first one then we are back in the univariate case with $q = 1$. It follows that the above mentioned criteria extends directly to the multivariate case.

Returning to the univariate case we see that the Fisher information matrix of $\mathcal{F}_A$ majorizes the Fisher information matrix of $\mathcal{F}_B$ if and only if $AA' > BB'$ and $n_A > n_B$. This amounts to the condition that the two restrictions we may obtain from $\mathcal{F}_A$ by assuming that exactly one of the quantities β and σ are known are at least as informative as the corresponding restrictions of $\mathcal{F}_B$. It follows that the local orderings, the global ordering as well as the ordering by the Fisher information matrices are all stronger than the ordering of Fisher information matrices for known σ. Assuming that $AA' > BB'$ we may distinguish the three cases "$n_A < n_B$", "$n_A > n_B$ and $n_A < n_B + \text{rank}(AA'-BB')$" and "$n_A > n_B + \text{rank}(AA'-BB')$".

In the first case $\mathcal{F}_A$ and $\mathcal{F}_B$ are not comparable for any of the mentioned orderings. In the last case $\mathcal{F}_B$ is as we noted above, a factor of $\mathcal{F}_A$ so that $\mathcal{F}_A$ majorizes $\mathcal{F}_B$ for the global ordering and thus also for the weaker orderings. In the second case we know that $\mathcal{F}_A$ does not majorize $\mathcal{F}_B$ globally while it does majorize $\mathcal{F}_B$ for the ordering of Fisher information matrices.

It remains therefor only to consider the problem of local ordering when $n_B + \text{rank}(AA'-BB') > n_A > n_B$.

As a particular case consider univariate normally distributed variables X and Y such that X is distributed as $N(\xi, \sigma^2)$ while Y is distributed as $N(\xi, \sigma^2/2)$. Assume that ξ and $\sigma > 0$ are both unknown. Replacing X with $\sqrt{2}X$ we see that we are in this situation with $n_A = n_B = 1$, $AA' = 2$ and $BB' = 1$. We shall see below, in spite of the fact that the experiments defined by X and Y are not comparable, that X is locally more informative than Y at any point (ξ, σ).

If fact we shall now show that the local ordering of these experiments coincides with the usual ordering of the Fisher information matrices. Thus let us assume that $AA' > BB'$ and that $n_A > n_B$. Using the dilation criterion mentioned above we shall prove our claim by showing that $\dot{\delta}_{\beta^0, \sigma^0}(\mathcal{F}_A, \mathcal{F}_B) = 0$ for all $\beta^0 \in R^p$ and all $\sigma^0 > 0$.

The consequence that the local orderings do not depend on where localization takes place follows, as we shall see, from the arguments below. This fact follows however also by general considerations on invariance under groups acting transitively and "smoothely" on the parameter set.

Differentiating the log likelihoods we find that $F(\bullet|\sigma,\beta,\mathcal{F}_A) = \mathcal{L}((AX)'/\sigma,(\sum_1^{n_A} X_i^2-n_A)/\sigma)$ where X is distributed as $N(0,I_A)$. Similarly $F(\bullet|\sigma,\beta,\mathcal{F}_B) = \mathcal{L}((BY)'/\sigma,(\sum_1^{n_B} Y_i^2-n_B)/\sigma)$ where Y is distributed as $N(0,I_B)$. Thus we must show that $E\phi((AX)'/\sigma,(\sum_1^{n_A} X_i^2-n_A)/\sigma) >$ the same expression for $\mathcal{F}_B$ when ϕ is convex on R^{p+1} and, say, a maximum of a finite number of linear functions. Replacing ϕ with $\phi(\bullet/\sigma)$ we see that we may without loss of generality assume that $\beta = 0$ (in R^p) and $\sigma = 1$. Proceeding as in Hansen and Torgersen 1974 we consider first the case where AA' is the $p\times p$ identity matrix while $BB' = \Lambda$ is a $p\times p$ diagonal matrix. Let λ_i denote the (i,i)th element of Λ. We shall, for convenience of formulation, assume that the diagonal elements of Λ are ordered in decreasing order. If $s = \text{rank } B = \text{rank } \Lambda$ then $\lambda_i>0$ or $=0$ as $i<s$ or $i>s$.

As the rows of A are orthonormal we may add n_A-p rows to A so that the extended $n_A\times n_A$ matrix $\tilde{A}$ is orthonormal. Putting $\tilde{X} = \tilde{A}X$ we see that $AX = (\tilde{X}_1,\ldots,\tilde{X}_p)'$ and that $\sum X_i^2 = \sum \tilde{X}_i^2$. Thus, since $\tilde{X}$ is also distributed as $N(0,I_A)$, we find that $E\phi((AX)',\sum X_i^2-n_A) = E\phi(X_1,\ldots,X_p,X_1^2+\ldots+X_{n_A}^2-n_A)$.

Likewise the rows of B are orthogonal. The first s rows became orthornormal after having been divided by, respectively, $\sqrt{\lambda_1},\sqrt{\lambda_2},\ldots,\sqrt{\lambda_s}$. The $p-s$ remaining rows are all the $1\times n_B$ zero matrix. Extend the described orthonormal system of row matrices to a $n_B\times n_B$ orthornormal matrix U and put $\tilde{Y} = UY$. Then $\tilde{Y}$ is distributed as Y and $BY = (\sqrt{\lambda_1}\tilde{Y}_1,\ldots,\sqrt{\lambda_s}\tilde{Y}_s,0,\ldots,0)'$ while $\sum Y_i^2 = \sum \tilde{Y}_i^2$. Hence $E\phi((BY)',\sum Y_i^2-n_B) =$ $E\phi(\sqrt{\lambda_1}Y_1,\ldots,\sqrt{\lambda_s}Y_s,0,\ldots,0,\sum Y_i^2-n_B)$. Our task is therefor to show that $E\phi(X_1,\dots,X_p,\sum_1^{n_A} X_i^2-n_A) > E\phi(\sqrt{\lambda_1}X_1,\dots,\sqrt{\lambda_p}X_p,\sum_1^{n_B} X_i^2-n_B)$ when $X_1,X_2,\dots$ are independent $N(0,1)$ variables. Assume first that $n_B>p$. Then, by Jensen's inequality, $E\phi(X_1,\dots,X_p,\sum_1^{n_A} X_i^2-n_A) = EE\phi(\dots)|X_1,\dots,X_{n_B}) > E\phi(X_1,\dots,X_p,\sum_1^{n_B} X_i^2-n_B)$. If, on the other hand, $n_B<p$ then, since $n_B>s$, the same argument implies that $E\phi(\sqrt{\lambda_1}X_1,\dots,\sqrt{\lambda_p}X_p,\sum_1^{n_B} X_i^2-n_B) < E\phi(\sqrt{\lambda_1}X_1,\dots,\sqrt{\lambda_p}X_p,\sum_1^{p} X_i^2-p)$ and that $E\phi(X_1,\dots,X_p,\sum_1^{p} X_i^2-p) < E\phi(X_1,\dots,X_p,\sum_1^{n_A} X_i^2-n_A)$. Putting $t = \max(p,n_B)$ we see that we in both cases will be through if we can show that $E\phi(X_1,\dots,X_p,\sum_1^{t} X_i^2-t) > E\phi(\sqrt{\lambda_1}X_1,\dots,\sqrt{\lambda_p}X_p,\sum_1^{t} X_i^2-t)$. Let the variables $\xi_1 = \pm 1, \xi_2 = \pm 1,\dots,\xi_p = \pm 1$ be independent and independent of $(X_1,\dots,X_t)$. Assume that $Pr(\xi_i = 1) = (\sqrt{\lambda_i}+1)/2$. (This is feasible since $\lambda_i \in [0,1]$.) Then $E\xi_i = \sqrt{\lambda_i}$. By symmetry and Jensen's inequality we obtain:

$$E\phi(X_1,\dots,X_p,\sum_1^{t} X_i^2) = E\phi(\xi_1X_1,\dots,\xi_pX_p,\sum_1^{t} X_i^2) > \phi(\sqrt{\lambda_1}X_1,\dots,\sqrt{\lambda_p}X_p,\sum_1^{t} X_i^2).$$

This establishes the desired inequality when AA' is the $p\times p$ unit matrix and BB' is a diagonal matrix. If, more generally, rank $A = p$ then there is a $p\times p$ matrix F so that $FAA'F'$ is the $p\times p$ unit matrix while $FBB'F'$ is a diagonal matrix (with the diagonal elements being in decreasing order). Then $AX = F^{-1}FAX$ and $BY = F^{-1}FBY$ and we are back to the previous case with ϕ replaced by $\tilde{\phi}$ where $\tilde{\phi}(x_1,\dots,x_p,z) \equiv \phi((F^{-1}x)',z)$.

Finally if rank $A = r<p$ then we may choose a basis $v_{\cdot 1},v_{\cdot 2},\dots,v_{\cdot r}$ for the column space of A. Any vector in the column space of B belongs to this space since $AA'>BB'$. If $a_{\cdot j}$ is the j-th column in A and $b_{\cdot j}$ is the j-th column in B then we may write $a_{\cdot j} = \sum_i s_{ij}v_{\cdot i}$ and $b_{\cdot j} = \sum t_{ij}v_{\cdot i}$. Putting $S = \{s_{ij}\}$ and $T = \{t_{ij}\}$ we obtain $A = VS$ and $B = VT$ where V is the $p\times r$ matrix $(v_{\cdot 1},\dots,v_{\cdot r})$. Then S and T have, respectively, dimensions $r\times n_A$ and $r\times n_B$. Furthermore rank $S = r$ and

$SS'>TT'$. If $\phi_V(y_1,..,y_r,z)\equiv\phi(Vy,z)$ then $\phi((AX)',\sum X_i^2-n_A) = \phi_V((SX)',\sum X_i^2-n_A)$ and $\phi((BY)',\sum Y_i^2-n_B) = \phi_V((TY)',\sum Y_i^2-n_B)$. Thus we may apply the previous arguments to $\mathcal{F}_S$ and $\mathcal{F}_T$.

It would be interesting to know if there are general and manageable expressions for the local deficiencies between linear normal models. The reader is referred to LeCam 1975 and Swensen 1980 for information on the "global" deficiencies in this case.

The statistical significance of local deficiencies is particularily transparent in the one dimensional (i.e. $m = 1$) case. In this case, as explained in section 2, the deficiencies $\dot{\delta}_{k,\theta^0}(\mathcal{E},\mathcal{F})$; $k = 2,3,...$ are all the same. They may then be expressed in terms of powers of most powerful tests or in terms of slopes of power-functions of locally most powerful tests.

These results may be expressed as follows:

Let for each (ordered) pair (P_1,P_2) of probability measures on the same measurable space the power of the most powerful level α-test for testing "P_1" against "P_2" be denoted by $\beta(\alpha|P_1,P_2)$. Then:

$$\dot{\delta}_{\theta^0}(\mathcal{E},\mathcal{F}) = 2\lim_{\varepsilon\to 0}\sup_{\alpha}[\beta(\alpha|Q_{\theta^0},Q_{\theta^0+\varepsilon})-\beta(\alpha|P_{\theta^0},P_{\theta^0+\varepsilon})]/\varepsilon.$$

In terms of locally most powerful tests - or rather slope maximizing tests - we have:

$$\dot{\delta}_{\theta^0}(\mathcal{E},\mathcal{F}) = \sup_{\alpha}[\tau(\alpha|\mathcal{F})-\tau(\alpha|\mathcal{E})]^+$$

where the quantity $\tau(\alpha|\mathcal{E})$ for a differentiable experiment $\mathcal{E} = (P_\theta:\theta\in\Theta)$ and a number $\alpha\in[0,1]$ is the maximal slope at θ^0 for powerfunctions of level α tests for testing "$\theta = \theta^0$" against "$\theta>\theta^0$". The function $\tau(\cdot|\mathcal{E})$ is, and may be any, continuous concave function on $[0,1]$ vanishing at $\alpha = 0$ and $\alpha = 1$.

Actually $\dot{\mathcal{E}}_{\theta^0}\sim(\lambda,\tau)$ where λ is the uniform distribution on $[0,1]$ and τ is the measure on $[0,1]$ posessing $\tau(\cdot|\mathcal{E})$ as distribution function.

In particular $\mathcal{E}$ is locally at least as informative as $\mathcal{F}$ at $\theta = \theta^0$ if and only if $\tau(\alpha|\mathcal{F})<\tau(\alpha|\mathcal{E})$ for all numbers $\alpha\in[0,1]$. Thus the function $\tau(\cdot|\mathcal{E})$ characterizes $\mathcal{E}$ up to local equivalence at θ^0. The local deficiency distance $\dot{\Delta}_{\theta^0}$ becomes then just the sup norm distance for such functions.

If G is a probability distribution on the real line then the corresponding translation family is differentiable at θ^0, and hence for all θ, if and only if G has an absolutely continuous density g such that $\int_{-\infty}^{\infty}|g'(x)|dx<\infty$. Any experiment $\mathcal{E} = (P_\theta:\theta\in\Theta)$ which is differentiable at $\theta = \theta^0$ and which is not locally equivalent to a totally non informative experiment (i.e. $\dot{P}_{\theta^0}\neq 0$) is locally equivalent to a differentiable translation experiment $G(\cdot,-\theta):\theta\in\Theta)$. In fact we may require that G is strongly unimodal and then G is determined up to a translation.

It follows that from the point of view of local comparison which we have described here that all differentiable exeriments which are not locally non informative are equivalent to an essentially unique strongly unimodal translation experiment. This construction is as follows: Let $\mathcal{E} = (P_\theta:\theta\in\Theta)$ be differentiable at θ^0 and assume that $\dot{P}_{\theta^0}\neq 0$. Then the set of solutions of the differential equation $G' = \tau(1-G|\mathcal{E})$ is the set of translates of a strongly unimodal distribution function G. If $\mathcal{G} = (G(\cdot-\theta):\theta\in\Theta)$ is the translation experiment determined by G then it follows that $\tau(\cdot|\mathcal{E}) = \tau(\cdot|\mathcal{G})$ so that $\dot{\Delta}_{\theta^0}(\mathcal{E},\mathcal{G}) = 0$.

If two strongly unimodal translation experiments are locally equivalent then they are "globally" equivalent as well. In general, however, translation experiments may be locally equivalent without being "globally" equivalent. On the other hand, whatever the dimension m, $\mathcal{E}$ is locally at least as informative as $\mathcal{E}$ at θ^0 whenever $\mathcal{E}$ is "globally" at least as informative as $\mathcal{F}$. In fact any transition which carries $\mathcal{E}$ onto $\mathcal{F}$ also carries $\dot{\mathcal{E}}_{\theta^0}$ onto $\dot{\mathcal{F}}_{\theta^0}$.

It should be kept in mind that the theory for local comparison which have been outlined here is only conserned with first order approximations. There are certainly cases of interest where the first order approximations are quite inadequate. Thus any experiment $\mathcal{E} = (P_\theta : \theta \in \Theta)$ such that $[\partial P_\theta / \partial \theta_i]_{\theta^0} = 0$; $i = 1, \ldots, m$ is locally equivalent (for first order approximations) to a totally non informative experiment. If, however, we want to test "$\theta = \theta^0$" against close alternatives "$\theta = \theta'$" then such experiments may behave quite differently. Using the theory in section 2 we may then proceed and consider higher order approximations.

Along with the notion of local information considered here goes a notion of local sufficiency. Thus if $\mathcal{E} = (\chi, \mathcal{A}, P_\theta : \theta \in \Theta)$ is differentiable at $\theta = \theta^0$ and if $\mathcal{F}$ is the restriction $\mathcal{E}|\mathcal{B}$ to some sub σ-algebra $\mathcal{B}$ of $\mathcal{A}$ then $\dot{\Delta}_{\theta^0}(\mathcal{E}, \mathcal{E}|\mathcal{B}) = 0$ if and only if $d\dot{P}_{\theta^0,i} / dP_{\theta^0}$ may be specified $\mathcal{B}$ measurable for all $i = 1, \ldots, m$. Thus these Radon-Nikodym derivatives describe the minimal locally sufficient sub σ-algebra of $\mathcal{A}$.

4. COMPARISON OF MEASURES. DILATIONS.

Orderings of experiments and of measures are often expressed in terms of dilations. The usual notion of a dilation as a kernel which dilates is, however, too narrow and should for our purposes be replaced by some notion of (almost) density preserving kernels. The typical situation is as follows:

Assume that we are given two measure families $\mathcal{E} = (\chi,\mathcal{A},\mu_\theta:\theta\in\Theta)$ and $\mathcal{F} = (\mathcal{Y},\mathcal{B},\nu_\theta:\theta\in\Theta)$ such that the measurable spaces $(\chi,\mathcal{A})$ and $(\mathcal{Y},\mathcal{B})$ are both Euclidean.

Assume also that we are given a point $\theta^0\in\Theta$ such that the measures μ_{θ^0} and ν_{θ^0} are non negative and dominates, respectively, $\mathcal{E}$ and $\mathcal{F}$. Deficiency for deficiency functions vanishing at θ^0 may then be characterized by:

<u>Theorem 4.1. (Almost density preserving kernels)</u>.

Assume that the requirements described above are satisfied and that $\varepsilon_{\theta^0} = 0$. Let, for each $\theta\in\Theta$, f_θ and g_θ be densities of, respectively, μ_θ and ν_θ w.r.t., respectively μ_{θ^0} and ν_{θ^0}.

Then $\mathcal{E}$ is ε-deficient w.r.t. $\mathcal{F}$ if and only if $\mu_{\theta^0} = D\nu_{\theta^0}$ for a Markov kernel D from $(\mathcal{Y},\mathcal{B})$ to $(\chi,\mathcal{A})$ such that:

$$\int\left|\int f_\theta(x)D(dx|y)-g_\theta(y)\right|\nu_{\theta^0}(dy) \leq \varepsilon_\theta;\ \theta\in\Theta.$$

<u>Proof</u>: It follows from theorem 2.2 that $\mathcal{E}$ is ε-deficient w.r.t. $\mathcal{F}$ if and only if there is a Markov kernel M from $(\chi,\mathcal{A})$ to $(\mathcal{Y},\mathcal{B})$ so that $\|\mu_\theta M-\nu_\theta\|\leq\varepsilon_\theta;\ \theta\in\Theta$. If M has this property then, since $\varepsilon_{\theta^0} = 0$, $\mu_{\theta^0}M = \nu_{\theta^0}$. It follows that there is a joint distribution on $(\chi\times\mathcal{Y}, \mathcal{A}\times\mathcal{B})$ with marginals μ_{θ^0} and ν_{θ^0} so that the conditional distribution of the second coordinate given $x\in\chi$ is $M(\cdot|x)$. Let D be a regular distribution of the first coordinate (in χ) given the second coordinate (in $\mathcal{Y}$). Then $\mu_{\theta^0}\times M = D\times\nu_{\theta^0}$. It is easily checked

that $y \to \int f_\theta(x)D(dx|y)$ is a density of $\mu_\theta M$ w.r.t. $\nu_{\theta 0}$ so that $\varepsilon_\theta > \|\mu_\theta M - \nu_\theta\| = \int|\int f_\theta(x)D(dx|y) - g_\theta(y)|\nu_{\theta 0}(dy)$.

Assume conversely that a randomization D mapping $\nu_{\theta 0}$ onto $\mu_{\theta 0}$ and having this property exists. Let M be a regular conditional distribution of the second coordinate (in $\mathcal{Y}$) given the first coordinate (in χ) for the distribution $D \times \nu_{\theta_0}$ on $\mathcal{A} \times \mathcal{B}$. Thus $D \times \nu_{\theta 0} = \mu_{\theta 0} \times M$. We find again that $y \to \int f_\theta(x)D(dx|y)$ is a density of $\mu_\theta M$ w.r.t. $\nu_{\theta 0}$ and thus $\|\mu_\theta M - \nu_\theta\| = \int|\int f_\theta(x)D(dx|y) - g_\theta(y)|\nu_{\theta 0}(dy) < \varepsilon_\theta$; $\theta \in \Theta$. □

The assumption that the kernels should map $\nu_{\theta 0}$ into $\mu_{\theta 0}$ (or $\mu_{\theta 0}$ into $\nu_{\theta 0}$) is a generalized form of the condition that a Markov matrix should be doubly Markov (stochastic). We obtain the latter condition if we impose the condition that χ and $\mathcal{Y}$ are the same finite sets and the measures $\mu_{\theta 0}$ and $\nu_{\theta 0}$ both are the uniform distribution on this set.

Let us express these facts in terms of matrices in the slightly more general situation where $\mu_{\theta 0}$ is a non negative distribution μ on $\{1,\ldots,m\}$ and $\nu_{\theta 0}$ is a non negative distribution ν on $\{1,\ldots,n\}$.

Let us also, for simplicity, assume that the measure families are finite. Thus we assume that we are given a finite family $\mathcal{E} = (\mu,\mu_1,\ldots,\mu_r)$ of measures on $\{1,\ldots,m\}$ and a finite family $\mathcal{F} = (\nu,\nu_1,\ldots,\nu_r)$ of measures on $\{1,\ldots,n\}$.

Identify $\mathcal{E}$ with (μ,A) where μ is the row matrix $(\mu(1),\ldots,\mu(m))$ and A is the $r \times m$ matrix whose (k,i)th element is the number $\mu_k(i)$; $k = 1,\ldots,r$, $i = 1,\ldots,m$.

Identify $\mathcal{F}$ with (ν,B) where ν is the row matrix $(\nu(1),\ldots,\nu(n))$ and B is the $r \times n$ matrix whose (k,j)th element is $\nu_k(j)$; $k = 1,\ldots,r$, $j = 1,\ldots,n$.

Thus the entries of the row matrices μ and ν are non negative and we shall assume that they are positive. Otherwise there is no restrictions on the (real valued) entries of the matrices μ, A, ν and B.

The theorem yields the following comparison rules for matrices.

Corollary 4.2. (Informational inequalities for matrices).

Let $\varepsilon_1,\dots,\varepsilon_r$ be non negative numbers and let A and B be matrices with real valued entries and with, respectively, dimensions $r\times m$ and $r\times n$.

Let also μ and ν be row matrices with positive entries and of dimensions m and n respectively. Put $\tilde{A}_{k,i} = A_{k,i}\mu_i$ and $\tilde{B}_{k,j} = B_{k,j}\nu_j$; $k = 1,\dots,r$; $i = 1,\dots,m$; $j = 1,\dots,n$.

Then the following conditions are equivalent:

(i) $\mu M = \nu$ for a $m\times n$ Markov matrix M such that $\sum_j |(\tilde{A}M)_{k,j}-\tilde{B}_{k,j}| \leq \varepsilon_k$; $k = 1,\dots,r$.

(ii) $\mu = \nu D$ for a $n\times m$ Markov matrix D such that $\sum_j |(AD)_{k,j}-B_{k,j}|\nu_j \leq \varepsilon_k$; $k = 1,\dots,r$.

(iii) $\sum_i \phi(\mu_i, A_{1,i}, A_{2,i},\dots,A_{r,i}) \geq \sum_j \phi(\nu_j, B_{1,j}, B_{2,j},\dots,B_{r,j}) - \sum_{k=1}^{r} \varepsilon_k \lfloor \phi(0,\overset{(1)}{0},\dots,\overset{(k)}{-1},\dots,\overset{(r)}{0}) \vee \phi(0,\overset{(1)}{0},\dots,\overset{(k)}{1},\dots,\overset{(r)}{0})$ for each sub linear function ϕ on $R\times R^r$.

Remark. We leave it to the reader to simplify these statements when μ and ν are uniform. If $\varepsilon_1 = \varepsilon_2 = \dots = \varepsilon_k = 0$ then (iii) may be phrased:

(iii') $\sum_i \phi(A_{1i}/\mu_i,\dots,A_{ri}/\mu_i)\mu_i \geq \sum_j \phi(B_{1j}/\nu_j,\dots,B_{rj}/\nu_j)\nu_j$ for each convex function ϕ on R^r.

Consider so a non negative measure T on R^m and a Markov kernel D from R^m to R^m. If $\varepsilon_1,\dots,\varepsilon_m$ are non negative numbers and if $\varepsilon = (\varepsilon_1,\dots,\varepsilon_m)$ then we shall say that D is a (T,ε) dilation if $\int|(\int x_i D(dx|y)-y_i|T(dy)<\varepsilon_i$; $i = 1,\dots,m$. A Markov kernel D from R^m to R^m such that $\int x_i D(dx|y)\underset{y}{\equiv}y_i$; $i = 1,\dots,m$ is called a dilation. Clearly a dilation is a (0,T) dilation and conversely any (0,T) dilation is T equivalent to a dilation.

Here is a more general corollary of the theorem which includes some of the standard results on dilations in R^m as special cases.

Corollary 4.3. (ε-dilations in R^m).

Let S and T be non negative measures on R^m such that the projections are integrable. Let the measure S_i (the measure T_i) for each $i = 1,\dots,m$, be the measure having the projection on the i-th coordinate space as a density w.r.t. the measure S (the measure T). Then the following conditions are equivalent for non negative numbers $\varepsilon_1,\dots,\varepsilon_k$.

(i) SM = T for a Markov kernel M from R^m to R^m such that $\|S_iM-T_i\|<\varepsilon_i$; $i = 1,\dots,m$.

(ii) S = DT for a (T,ε) dilation D from R^m to R^m.

(iii) $\int\phi(1,x)S(dx)>\int\phi(1,x)T(dx)-\sum_{i=1}^{m}\varepsilon_i[\phi(\overset{(0)}{0},\overset{(1)}{0},\dots,\overset{(i)}{-1},\dots,\overset{(m)}{0})\vee\phi(\overset{(0)}{0},\overset{(1)}{0},\dots,\overset{(i)}{1},\dots,\overset{(m)}{0})]$ for each sub linear function ϕ on $R\times R^m$.

Remark: If $\varepsilon_1 =\dots= \varepsilon_k = 0$ then (ii) may be written:

(ii') S = DT for a dilation D.
In this case (iii) may be written:
(iii') $\int\phi(x)S(dx)>\int\phi(x)T(dx)$ for all convex functions ϕ on R^m.

Proof: Apply the theorem and the sub linear function criterion ((iii) in section 2) to the measure families $(S,S_1,\dots,S_m)$ and $(T,T_1,\dots,T_m)$. □

Let us phrase this corollary in terms of standard measures of measure families.

If $\mathcal{E} = (\chi, \mathcal{A}, \mu_\theta : \theta \in \Theta)$ is a measure family with finite parameter set Θ then its <u>standard measure</u> is the measure induced from $\mu = \sum_\theta |\mu_\theta|$ by the map $x \to ([d\mu_\theta/d\mu]_x;\ \theta \in \Theta)$ from χ to R^Θ. Thus the corollary may be phrased:

<u>Corollary 4.4. (Ordering in terms of standard measures)</u>.

Let S and T denote, respectively, the standard measures of the measure families $\mathcal{E} = (\mu_1, \ldots, \mu_m)$ and $\mathcal{F} = (\nu_1, \ldots, \nu_m)$.

Let, for each $i = 1, \ldots, m$, the measure S_i (the measure T_i) be the measure on R^m having the projection on the i-th coordinate space as a density w.r.t. the measure S (the measure T).

Then the measure family $(\sum_i |\mu_i|,\ \mu_1, \ldots, \mu_m)$ is $(0, \varepsilon_1, \ldots, \varepsilon_m)$ deficient w.r.t. $(\sum_i |\nu_i|,\ \nu_1, \ldots, \nu_m)$ if and only if the (equivalent) conditions (i), (ii) and (iii) of the previous corollary is satisfied.

In particular if the measures $\mu_1, \ldots, \mu_m$ and $\nu_1, \ldots, \nu_m$ are all non negative then $\mathcal{E}$ is 0-deficient w.r.t. $\mathcal{F}$ and only if $S = DT$ for a dilation D.

Another application is to differentiable experiments. The result below provides in particular the criterion for "being locally at least as informative as" which we commented on and utilized in the previous section:

<u>Corollary 4.5. (Local deficiencies)</u>.

Assume that the parameter set Θ is a sub set of R^m and that the point θ^0 is an interior point of Θ.

Let the experiments $\mathcal{E} = (P_\theta : \theta\in\Theta)$ and $\mathcal{F} = (Q_\theta ; \theta\in\Theta)$ both be differentiable in θ^0 and let $\varepsilon_1,\ldots,\varepsilon_m$ be non negative numbers.

Put, using the notations of section 3, $S = F(\cdot|\theta^0,\mathcal{E})$ and $T = F(\cdot|,\theta^0,\mathcal{F})$.

Then each of (the equivalent) conditions (i)-(iii) of corollary 4.3 are equivalent to the condition that $\dot{\mathcal{E}}_{\theta^0} = (P_{\theta^0},\dot{P}_{\theta^0,1},\ldots,\dot{P}_{\theta^0,m})$ is $(0,\varepsilon_1,\ldots,\varepsilon_m)$ deficient w.r.t. $\dot{\mathcal{F}}_{\theta^0} = (Q_{\theta^0},\dot{Q}_{\theta^0,1},\ldots,\dot{Q}_{\theta^0,m})$.

In particular $\mathcal{E}$ is locally at least as informative as $\mathcal{F}$ at θ^0 if and only if $\int\phi(x)S(dx) > \int\phi(x)T(dx)$ for each convex function ϕ on R^m.

The dilation results for standard measures may be viewed as a particular case of dilation criterions for comparison of measures. A fairly general situation is described in:

Theorem 4.6. (Orderings of measures).

Let μ and ν be non negative finite measures on a measurable space $(\chi,\mathcal{A})$.

Consider also a convex set H of $\mu+\nu$ integrable functions containing 0 and having the property that $h_1 \vee h_2 \in H$ when $h_1,h_2\in H$.

Then $\int h d\mu > \int h d\nu$ for all $h\in H$ if and only if there is a transition M from $L_1(\mu)$ to $L_1(\nu)$ so that:

$$(h\mu)M > h\nu;\ h\in H.$$

If, in addition H contains the constants and if $(\chi,\mathcal{A})$ is Euclidean then this condition is also equivalent to the condition that $\mu = D\nu$ for a Markov kernel D such that $\int h(x)D(dx|y) > h(y)$ for ν almost all y for all $h\in H$.

Remark 1. If μ is a measure and if $\int h d\mu$ exists then $h\mu$ denotes the measure which to each measurable set A assigns the mass $\int_A h d\mu$.

Remark 2. If H is countable or if H is "sufficiently separable" then the exceptional sets in the last statement may be chosen as the empty set.

Proof: Consider the measure families $\mathcal{E} = (h\mu : h \in H)$ and $\mathcal{F} = (h\nu : h \in H)$. Let $h_1, \ldots, h_r \in H$ and let ϕ be a sub linear function on R^r which is both monotonically increasing and a maximum of a finite set of linear functionals on R^r. Thus $\phi(x) \underset{x}{\equiv} \bigvee_{t=1}^{n} \sum_{i=1}^{r} a_{ti} x_i$ for non negative constants a_{ti}. If $a_1, \ldots, a_r$ are non negative constants then, since $0 \in H$ and H is convex, $\frac{1}{N} \sum_{i=1}^{r} a_i h_i \in H$ when $N > \sum_{i=1}^{r} a_i$. It follows that $\frac{1}{N}\phi(h_1, \ldots, h_r) \in H$ when N is sufficiently large. Thus $\int \phi(h_1, \ldots, h_r) d\mu > \int \phi(h_1, \ldots, h_r) d\nu$. Consider so the maximum ϕ of a arbitrary finite family of linear functionals on R^r. The map $x \to \phi(x) + \sum_{i=1}^{r} x_i \phi(0, \ldots, \overset{(i)}{-1}, \ldots, 0)$ is then monotonically increasing on R^r. Thus $\int \phi(h_1, \ldots, h_r) d\mu > \int \phi(h_1, \ldots, h_r) d\nu - \sum_{i=1}^{r} \varepsilon(h_i) \phi(0, \ldots, \overset{(i)}{-1}, \ldots, 0)$ where $\varepsilon(h) = \int h d\mu - \int h d\nu$; $h \in H$. It follows that $\mathcal{E}$ is ε-deficient w.r.t. $\mathcal{F}$. Hence, by the randomization criterion, there is a transition M from $L_1(\mu)$ to $L_1(\nu)$ so that $\|(h\mu)M - (h\nu)\| < \varepsilon(h)$; $h \in H$. Now $\|(h\mu)M - (h\nu)\| = [(h\mu)M - h\nu](1) + 2\|[(h\mu)M - (h\nu)]^-\| = \varepsilon(h) + 2\|[(h\mu)M - (h\nu)]^-\|$. Thus $\|[(h\mu)M - (h\nu)]^-\| = 0$ when $h \in H$ so that $(h\mu)M > h\nu$; $h \in H$. Conversely these inequalities imply that: $\int h d\mu = (h\mu)(1) = ((h\mu)M)(1) > (h\nu)(1) = \int h d\nu$ when $h \in H$.

Assume so that this condition is satisfied, that H contains the constants and that the measurable space $(\chi,\mathcal{A})$ is Euclidean. Then $(\pm\mu)M>\pm\nu$ so that $\mu M = \nu$. Let D be a regular conditional distribution of the first coordinate given the second for the distribution $\mu\times M$. Thus $\mu\times M = \nu\times D$. If $A\in\mathcal{A}$ and $h\in H$ then $\int_A[\int h(x)D(dx|y)]\nu(dy) = \int h(x)I_A(y)(D\times\nu)(d(x,y)) = \int h(x)I_A(y)(\mu\times M)(d(x,y)) = \int M(A|x)h(x)\mu(dx) = ((h\mu)M)(A)>(h\nu)(A) = \int_A hd\nu$. Hence $\mu = D\nu$ and $\int h(x)D(dx|\cdot)>h$ a.e. ν. This in turn implies, by essentially the same computations, that $(h\mu)M>h\nu$ where the Markov kernel M satisfies $\mu\times M = D\times\nu$. □

Consider the particular case where H consists of all functions on R^m which are maxima of a finite set of (monotonically increasing) linear functionals. If $\int hd\mu>\int hd\nu$ for all $h\in H$ then, by the theorem and remark 2, $\mu = D\nu$ for a Markov kernel D such that $\int xD(dx|y)>y$ for all $y\in R^m$. Conversely the excistence of such a D implies, by Jensens's inequality that $\int hd\mu>\int hd\nu$, $h\in H$. If, more generally, $S_1,S_2,\ldots$ is a sequence of probability distributions on R^m such that the sequence $\int hdS_1, \int hdS_2,\ldots$ is monotonically increasing for each $h\in H$ then $S_{n+1} = D_nS_n$ where $\int xD_{n+1}(dx|y)>y$; $y\in R^m$. Let the joint distribution of $(X_1,X_2,\ldots)$ be determined by the requirements that $\mathcal{L}(X_1) = S_1$ and that $D_n(\cdot|X_n)$, for each n, is a conditional distribution of X_{n+1} given X_n. Then the Markov process $X_1,X_2,\ldots$ is a (sub) martingale such that X_n, for each n, has distribution S_n. This and Jensen's inequality prove the following result of Strassen 1965:

Corollary 4.7. ((Sub) Martingales with prescribed marginal distributions).

Assume that the probability distributions $S_1,S_2,\ldots$ on R^m all possess finite expectations. Then there is a (sub) martingale $X_1,X_2,\ldots$ with, respectively, marginal distributions $S_1,S_2,\ldots$ if and only if $\int hdS_1<\int hdS_2<\ldots$ whenever the function h is convex (and monotonically increasing) on R^m.

Another interesting situation occurs when we assume that the functions in H are monotonically increasing w.r.t. some partial ordering. In the latter case we obtain, see e.g. Karlin 1983, the joint distribution characterization of stochastic orderings of distributions on sets with prescribed partial orderings. As more general situations will be treated in the next section we shall not write this out here.

If we in theorem 4.6 just require that $\int h d\mu \geq \int h d\nu$ for the functions h in H which are non negative i.e. for the functions in $H_+ = \{h : h \in H,\ h \geq 0\}$ then the first part of the theorem applies with H replaced by H_+. The argument used in the proof of the last part of the theorem is, however, no longer valid since it required that $\int(-1)d\mu \geq \int(-1)d\nu$.

There is for this situation a usefull result of Fischer and Holbrook 1980. As the general underlying idea is quite simple we shall here expose it as:

Proposition 4.8. (Existence of minorizing functionals).

Let γ be a real valued convex functional on a convex sub set H of a linear space.

Let $\mathcal{P}$ be a sub convex class of real valued concave functionals on H.

Assume that $\mathcal{P}$ is compact for the weakest topology which makes P(h) lower semicontinuous in $P \in \mathcal{P}$ for each $h \in H$.

Then there is a P in $\mathcal{P}$ so that $\gamma \geq P$ on H if and only if $\gamma(h) \geq \inf\{P(h) : P \in \mathcal{P}\}$ for all $h \in H$.

Remarks. We may, for intuition, think about H as a convex set of real valued functions and of each $P \in \mathcal{P}$ as an expectation.

$\mathcal{P}$ is called sub convex if there to each pair $(P_0,P_1)\in\mathcal{P}\times\mathcal{P}$ and to each $t\in]0,1[$ corresponds a $P_t\in\mathcal{P}$ so that $P_t\leq(1-t)P_0+tP_1$. The compactness condition amounts to require that there to each net $\{P_\alpha\}$ in $\mathcal{P}$ corresponds a $P\in\mathcal{P}$ so that $\limsup_\alpha P_\alpha(h)\geq P(h)$; $h\in H$.

Proof: The "only if" is trivial and the "if" follows by applying minimax theory to the concave-convex payoff function $(P,h)\to\gamma(h)-P(h)$. □

Applying this proposition to the situation of corollary 4.3 with $\varepsilon_1=\ldots=\varepsilon_m=0$ we obtain:

Corollary 4.9. (Sub Markov kernels and non negative convex functions).

Let S and T be non negative measures on R^m such that the projections are integrable. Let the measure S_i (the measure T_i) for each $i=1,\ldots,m$ be the measure having the projection on the i-th coordinate space as a density w.r.t. the measure S (the measure T). Then the following conditions are equivalent:

(i) $SM=T$ for a sub Markov kernel M from R^m to R^m so that $S_iM=T_i$; $i=1,\ldots,m$.

(ii) $S\geq DT$ for a dilation D from R^m to R^m.

(iii) $\int\phi dS\geq\int\phi dT$ for each non negative convex function ϕ on R^m.

(iv) There is a non negative measure U on R^m such that $\int\phi dS\geq\int\phi d(T+U)$ for all convex functions ϕ on R^m. If so then the measure U may always be chosen as a one point mass distribution.

Remark. If S and T are required to possess compact supports then the equivalence of (iii) and (iv) follows from Fischer and Holbrook 1980.

<u>Proof</u>: Suppose (i) is satisfied. Then $\|S\| > \|SM\| = \|T\|$. If $\|S\| = \|T\|$ then M may be chosen as a proper Markov kernel and then we know from the remark after corollary 4.3 that $\int\phi dS > \int\phi dT$ for all convex functions ϕ on R^m. If $\|S\| > \|T\|$ then we may put $\xi_i = [\|S_i\| - \|T_i\|]/[\|S\| - \|T\|]$; $i = 1,\ldots,m$. Put $\xi = (\xi_1,\ldots,\xi_m)$ and define a Markov kernel $\tilde{M}$ by putting $\tilde{M}(B|x) = M(B|x) + [1-M(R^m|x)]I_B(\xi)$. Then $S\tilde{M} = T+U$ where U is the one point distribution which assigns mass $\|S\| - \|T\|$ to ξ. It is then a matter of checking that $S_i\tilde{M} = (T+U)_i = T_i + U_i$ where, for each measure κ, κ_i denotes the measure having $x \to x_i$ as density w.r.t. κ. Thus the remark after corollary 4.3 applies again and we get $\int\phi dS > \int\phi d(T+U) = \int\phi dT + \int\phi dU > \int\phi dT$ when ϕ is non negative and convex on R^m. Hence (i)=>(iv) and clearly (iv)=>(iii). Assume so that (ii) is satisfied. If ϕ is non negative and convex then $\int\phi dS > \int\phi d(DT) > \int\phi dT$ where the last ">" follows from the remark after corollary 4.3. Thus (ii)=>(iii).

Assume next that (iii) is satisfied. Put $\rho(x) = \vee_i |x_i|$ when $x = (x_1,\ldots,x_m) \in R^m$. Then $\|S\| = \int 1 dS > \int 1 dT = \|T\|$. If $\|S\| = \|T\|$ and if ϕ is convex and bounded from below by the constant b then $\int\phi dS = \int(\phi - b)dS + b\|S\| > \int(\phi-b)dT + b\|T\| = \int\phi dT$. This implies readily (iv) with $U = 0$. Put $\Gamma = [\|S\| - \|T\|]^{-1}(S-T)$ when $\|S\| > \|T\|$. Then $\Gamma(\phi) > 0$ for all non negative convex functions ϕ on R^m and $\Gamma(1) = 1$. Let H_N, $N = 1,2,\ldots,$ denote the set of convex functions h such that $h < \rho$ while $h(x) > 0$ when $\rho(x) > N$. Take $\mathcal{P}_N$ as the set of probability distributions on $\{x: \rho(x) < N\}$. If $h \in H_N$ then $\Gamma(h) > \inf_x h(x) = \inf\{P(h): P \in \mathcal{P}_N\}$. It follows, since $\mathcal{P}_N$ is tight, that there is a $P_N \in \mathcal{P}_N$ so that $\Gamma(h) > P_N(h)$; $h \in H_N$. In particular $\Gamma(\rho) > P_N(\rho)$. It follows that $(P_1, P_2, \ldots,)$ is tight. Let the probability distribution P be a weak limit point for this sequence. If h is convex and if $h \in H_{N_0}$ then $h \in H_N$ when $N > N_0$. Hence $\Gamma(h) > \limsup_N P_N(h) > P(h)$. In particular $\Gamma(h \vee (\rho - N_0)) > P(h \vee (\rho - N_0))$ when h is convex and $h < \rho$. $N_0 \to \infty$ yields $\Gamma(h) > P(h)$. It follows that $\Gamma(\phi) > P(\phi)$ when ϕ is convex and $\phi < k\rho$ for some positive constant k. Jensen's inequality tells us that $P(\phi) > \phi(\int x P(dx))$ when ϕ is convex. It follows that we may assume that P is a one point distribution. Alltogether this show that (iv) holds with U being the one point distribution which assigns mass $\|S\|-\|T\|$ to the point $\int xP(dx)$. Thus (iii)=>(iv). If (iv) holds then, by the remark after Corollary 4.3, there is a Markov kernel M so that $SM = T+U$ and $S_iM = T_i+U_i$; $i = 1,\dots,m$. Furthermore $T = (T+U)\hat{M}$ where $\hat{M}$ is the sub Markov kernel $(x,B)\to I_B(x)g(x)$ where g is a version of $dT/d(T+U)$ such that $0<g<1$. Then $SM\hat{M} = T$ and $S_iM\hat{M} = T_i$ for the sub-Markov kernel $M\hat{M}$. Hence (iv)=>(i). Still assuming (iv) and applying the same remark we find that $S = D(T+U)$ for a dilation D. Thus $S>DT$ so that (iv)=>(ii). The proof is now completed by combining the established implications. □

Returning to matrices again we obtain:

Corollary 4.10. (The matrix case).

Let μ and ν be row matrices of, respectively, dimensions $1\times m$ and $1\times n$. We shall assume that the entries $\mu_1,\dots,\mu_m$ of μ and $\nu_1,\dots,\nu_n$ of ν are all positive.

Let A and B be matrices with real entries and with dimensions $r\times m$ and $r\times n$ respectively. Then the following statements are equivalent.

(i) $\mu M = \nu$ for a $m\times n$ sub Markov matrix M such that $\sum_i A_{ki}\mu_i M_{i,j} = B_{k,j}\nu_j$; $k = 1,\dots,r$, $j = 1,\dots,n$.

(ii) $\mu>\nu D$ for a $n\times m$ Markov matrix D such that $AD = B$.

(iii) $\sum_i \phi(A_{1,i}/\mu_i,\dots,A_{r,i}/\mu_i)\mu_i > \sum_j \phi(B_{1,j}/\nu_j,\dots,B_{r,j}/\nu_j)\nu_j$ for each non negative convex function ϕ on R^r.

Proof: Apply the previous corollary to S and T where S assigns mass μ_i to each point $(A_{1,i},\dots,A_{r,i})$ while T assigns mass ν_j to each point $(B_{1,j},\dots,B_{r,j})$. □

Consider again the situation treated in theorem 4.6 under the additional assumptions that the functions in H are bounded and that the constant functions are in H.

If we weaken the requirement that $\mu(h)>\nu(h)$ for all $h\in H$ to the requirement that $\mu(h)>\nu(h)$ for <u>all non negative functions</u> $h\in H$ then there is a non negative and additive set function κ on $\mathcal{A}$ so that $\mu(h)>\nu(h)+\kappa(h)$ for all $h\in H$. This may be seen as follows: The assumption that $1\in H$ implies that $\|\mu\|>\|\nu\|$. If $\|\mu\| = \|\nu\|$ and c is a lower bound for h then $\frac{1}{2}(h-c)\in H_+$ so that $\frac{1}{2}\mu(h-c)>\frac{1}{2}\nu(h-c)$ yielding $\mu(h)>\nu(h)$. Thus we may put $\kappa = 0$ in this case. If $\|\mu\|>\|\nu\|$ then we may apply proposition 4.7 to the functional $\Gamma = [\|\mu\|-\|\nu\|]^{-1}(\mu-\nu)$ yielding a finitely additive probability set function P so that $\Gamma(h)>P(h)$ for all $h\in H$. We may then put $\kappa = [\|\mu\|-\|\nu\|]P$.

If we now were able to show that κ might be chosen countably additive then proposition 4.6 would become applicable with ν being replaced by $\nu+\kappa$. A particular case where this works is, by the Riesz representation theorem, when $\mathcal{A}$ is the class of Borel sub sets of a compact Hausdorff space χ and the functions in H are continuous. This yield finally:

<u>Corollary 4.11. (Comparison for integrals of non negative functions)</u>.

Let μ and ν be non negative finite measures on the Borel class $\mathcal{A}$ of a compact metric space χ.

Let H be a convex set of continuous functions on χ such that $h_1\vee h_2\in H$ when $h_1\in H$ and $h_2\in H$. We shall also assume that H contains the constant functions. Then the following conditions are equivalent

(i) $\int h d\mu>\int h d\nu$ when $0<h\in H$.

(ii) There is a non negative measure κ on $\mathcal{A}$ so that $\int h d\mu>\int h d\nu+\int h d\kappa$ for all $h\in H$.

(iii) There is a sub Markov kernel M so that $(h\mu)M>h\nu$; $h\in H$.

(iv) $\mu>D\nu$ for a Markov kernel D so that $\int h(x)D(dx|y)>h(y)$ for ν almost all y and for all $h\in H$.

Remark. If additional assumptions on linear and topological structures are satisfied then the equivalence of (i) and (ii) follows from Fischer and Holbrook 1980.

5. APPLICATION TO A PROBLEM ON PROBABILITY DISTRIBUTIONS WITH GIVEN MARGINALS.

We shall here apply the theory in section 2 to derive a result in Strassen's 1965 paper on probability distribution with given marginals.

The result is related to the problem of determining the set of possible probabilities $Pr(X,Y)\in S)$ for a specified set S and for prescribed marginal distributions P and Q for X and Y. By convexity this set is clearly an interval and it is also fairly clear that it is closed when we permit joint distributions which are not necessarily countably additive. Thus it suffices, at least under this permission, to find the end points of this interval. As the problem of minimizing $Pr((X,Y)\in S)$ is equivalent to the problem of maximizing $Pr(X,Y)\notin S)$ it should then also suffice to determine the right end point of this interval.

We shall here, as is actually done in Strassen's paper, consider the problem of maximizing $Pr((X,Y)\in S)$ when S is closed and the sample spaces of X and Y are both complete separable metric spaces. The right end point of the above mentioned interval may then be described as follows:

Let B be any event in the sample space of Y. Then $Pr((X,Y)\in S)$ may be decomposed as:

$$Pr((X,Y)\in S\&Y\notin B)+Pr((X,Y)\in S\&Y\in B).$$

Hence:

$$Pr((X,Y)\in S)\leq Pr(Y\notin B)+Pr((X,y)\in S \quad \text{for some} \quad y\in B))$$

where the number on the right hand side is determined by P and Q.

Let us for each sub set B of the sample space $\mathcal{Y}$ of Y agree to use the notation $B^{(S)}$ for the set of points x in the sample space χ of X such that $(x,y)\in S$ for some y in B. The upper bound derived above may then be written:

$$\inf_B [Q(B^c) + P(B^{(S)})]$$

and it follows from Strassen's paper (under the above mentioned regularity conditions) that this upper bound is achieved.

Usually it is not required to consider all sets B. If e.g. X and Y take their values in the same partially ordered set $(\chi, <)$ and if $S = \{(x,y): x > y\}$ then it suffices to consider monotonically increasing sets B (B is called <u>monotonically increasing (decreasing)</u> if its indicator function is monotonically increasing (decreasing)). Thus if χ is a sub set of the real line with the usual ordering then we obtain the greatest lower bound of all numbers $1-G(x)+F(x)$; $x \in \chi$ where F and G are, respectively, the distribution functions of X and Y.

In general there is a joint distribution such that $\Pr((X,Y) \in S) > 1-\varepsilon$ if and only if $Q(B) < P(B^{(S)})+\varepsilon$ for each (measurable) sub set B of $\mathcal{Y}$.

If ρ is a real valued function on $\chi \times \mathcal{Y}$ and if $S = \{(x,y): \rho(x,y) < \varepsilon\}$ then this amounts to the requirement that $Q(B) < P(\{x: \rho(x,y) < \varepsilon$ for some $y \in B\})+\varepsilon$ when B is a (measurable) sub set of $\mathcal{Y}$. This yield in particular, as was shown in Strassen 1965, the joint distribution characterization of the Prohorov, distance for a given distance ρ on χ.

There is a rephrasing of this result which is obtained by relaxing the condition that the distribution of Y should be exactly Q while at the same time strengthening the requirement on $\Pr((X,Y) \in S)$ by requiring that $\Pr((X,Y) \in S) = 1$. Using this rephrasing we shall now see how this type of results may be deduced from the basic principles for comparing measure families.

The essential idea used in the proof may be grasped by restricting the attention to finite sets χ and $\mathcal{Y}$. The reader might then see that the arguments using measure families may be replaced by simple arguments using well known results on support functions. The point here however is to provide another example on the relationship between the principles for comparison of measure families on the one hand and excistence problems for probability distributions with given marginals on the other hand.

Except for the rephrasing, condition (iii) below, the following result is a particular case of a theorem in Strassen's paper:

Proposition 5.1. (Probabilities for a specified set assigned by joint distributions possessing prescribed marginals).

Let $(\chi,\mathcal{A},P)$ and $(\mathcal{Y},\mathcal{B},Q)$ be probability spaces and let S be a sub set of $\chi\times\mathcal{Y}$. Assume that χ and $\mathcal{Y}$ are compact metric spaces with, respectively, Borel classes $\mathcal{A}$ and $\mathcal{B}$. Assume also that S is a closed sub set of $\chi\times\mathcal{Y}$. Then the following two conditions are equivalent for a number $\varepsilon\in[0,1]$

(i) $Q(B)\leq P(B^{(S)})+\varepsilon$ for all Borel sets B.

(ii) There is a joint distribution R on $\mathcal{A}\times\mathcal{B}$ with marginals P and Q such that $R(S)\geq 1-\varepsilon$.

If, furthermore, the projection of S into χ is onto then these conditions are equivalent with:

(iii) There is a joint distribution R on $\mathcal{A}\times\mathcal{B}$ with marginals P and μ such that $R(S) = 1$ and $\|\mu-Q\|\leq 2\varepsilon$.

Remark 1. The set $B^{(S)}$ is the projection on χ of the measurable set $(\chi\times B)\cap S$ and is consequently analytic and therefor completion measurable.

Remark 2. It is apparent from this proof that several of the assumptions may be weakened. Thus we shall only make use of (i) for closed sets B.

If S is defined by a partial ordering then it suffices to consider monotonically increasing sets B in (i) and then $B^{(S)} = B$. In this case these conditions are equivalent to the condition that $Q(h)<P(h)+\varepsilon\|h\|$ when the function h is bounded, monotonically increasing and measurable.

Proof: We know allready that (ii)=>(i). Assume now that (iii) holds. Then there is, by the joint distribution characterization of the statistical distance (see e.g. Torgersen 1970), a joint distribution U on $\mathcal{B}\times\mathcal{B}$ with marginals ν and Q such that $U(\{(y,y):y\in\mathcal{Y}\})>1-\varepsilon$. Define a joint distribution V for variables X, Z and Y such that (X,Z) has distribution R while the conditional distribution of Y given (X,Z) is given by a regular conditional distribution of Y given Z in the situation where U is the joint distribution of Y and Z. Then (ii) holds if R is replaced by the joint distribution of (X,Y). Thus in any case (and we did not use compactness) (iii)=>(ii)=>(i). Assume now that (i) holds and that the projection of S into χ is onto. Put for any bounded measurable function h on $\mathcal{Y}$ and $x\in\chi$ $h(x) = \sup\{h(y):y\in S_x\}$. If t is a real number then clearly $[h>t]^{(S)} = [\hat{h}>t]$. Thus, by (i):

$$Q(h) = \int_0^{\|h\|} Q(h>t)dt< \int_0^{\|h\|} [P(\hat{h}>t)+\varepsilon]dt = P(\hat{h})+\varepsilon\|h\|$$

when h is non negative. It follows from the expression for $[\hat{h}>t]$ above and remark 1 that $\hat{h}$ is completion measurable. Using that $\hat{c}(x)\equiv c$ for a constant function c we find, since $\widehat{h-c} = \hat{h}-c$, that $Q(h)<P(\hat{h})+2\varepsilon\|h\|$ for any bounded measurable function h. Here we used the assumption that the projection of S into χ was onto. Let h vary through the set $C(\mathcal{Y})$ of continuous functions on $\mathcal{Y}$. Then the linear functional Q

is majorized by the sum of the two sub linear functionals $h\to P(\hat{h})$ and $h\to 2\varepsilon\|h\|$. It follows that Q may be decomposed as $Q = \mu+\nu$ where μ and ν are linear functionals (measures) such that $\mu(h)<P(\hat{h})$; $h\in C(\mathcal{Y})$ while $\nu(h)<2\varepsilon\|h\|$. If $h<0$ then $\hat{h}<0$ and thus $\mu(h)<0$. If $h = c$ is a constant then $\mu(c)<P(\hat{c}) = P(c) = c$. It follows that μ is a probability distribution and that ν is a measure whose total variation is at most 2ε.

Consider the measure families $\mathcal{E} = (\hat{h}P:h\in C(\mathcal{Y}))$ and $\mathcal{F} = (h\mu:h\in C(\mathcal{Y}))$. Put $\eta_h = P(\hat{h})-\mu(h)$ when $h\in C(\mathcal{Y})$.

Let $h_1,\ldots,h_r\in C(\mathcal{Y})$ and let $h = \phi(h_1,\ldots,h_r)$ where ϕ is a sub linear monotonically increasing function on R^r. Put $h = \phi(h_1,..,h_r)$. Then $\int\phi(h_1,\ldots,h_r)d\mu = \int hd\mu<\int\hat{h}dP<\int\phi(\hat{h}_1,\ldots,\hat{h}_r)dP$. If ϕ is any sub linear function on R^r then $x\to\phi(x)+\sum_i x_i\phi(0,\ldots,\overset{(i)}{-1},\ldots,0)$ is monotonically increasing. It follows then from the sub linear function criterion (iii) of section 2 that $\mathcal{E}$ is η deficient w.r.t. $\mathcal{F}$. Thus there is a Markov kernel M from $(\chi,\mathcal{A})$ to $(\mathcal{Y},\mathcal{B})$ so that $\|(\hat{h}P)M-h\mu\|<\eta_h$; $h\in C(\mathcal{Y})$. Thus $P(\hat{h})-\mu(h)+2\|[(\hat{h}P)M-h\mu]^-\|<\eta_h$. $h = 1$ yields then $PM>\mu$ so that $PM = \mu$. Let D be a regular conditional distribution of "x" given "y", for $P\times M$. Then $P\times M = D\times\mu$. If $h\in C(\mathcal{Y})$ and $B\in\mathcal{B}$ then $\int_B[\int\hat{h}(x)D(dx|y)]\mu(dy) = \int\hat{h}(x)I_B(y)(D\times M)(d(x,y)) = \int\hat{h}(x)I_B(y)(P\times M)(d(x,y)) = \int M(B|x)\hat{h}(x)(dx) = ((\hat{h}P)M)(B)>(h\mu)(B) = \int_B h(y)\mu(dy)$. Thus $\int\hat{h}(x)D(dx|y)>h(y)$ for μ almost all y when $h\in C(\mathcal{Y})$. We may now, since $C(\mathcal{Y})$ is separable, modify D so that $\int\hat{h}(x)D(dx|y)>h(y)$ for all $y\in\mathcal{Y}$ and all $h\in C(\mathcal{Y})$. This inequality extends, since $\hat{h}_n\uparrow\hat{h}$ when $h_n\uparrow h$, to all lower semicontinuous functions h on $\mathcal{Y}$. Let $y_0\in\mathcal{Y}$ and let, for $n = 1,2,\ldots$, the open ball around y_0 with radius $1/n$ be denoted as B_n. Applying the inequality to the indicator function h_n of B_n we find that $D(B_n^{(S)}|y_0) = \int\hat{h}_n(x)D(dx|y)>h_n(y_0) = 1$. Furthermore, since $B_1^{(S)}\cap B_2^{(S)}\cap\ldots = \{x:(x,y_0)\in S\}$, we find that $D(\{x:(x,y_0)\in S\}|y_0) = 1$ for all points $y_0\in\mathcal{Y}$. Thus $(P\times M)(S) = (D\times\mu)(S) = \int D(\{x:(x,y)\in S\}|y)\mu(dy) = 1$.

It follows that condition (iii) is satisfied with $R = P \times M$.

Altogether this show that conditions (i), (ii) and (iii) are equivalent when the projection of S into χ is onto.

It remains to show that (i)=>(ii) in the general case. Assume then that (i) holds. Let the metric d which metrisizes $\mathcal{Y}$ be bounded by 1. Add a point $\S$ to $\mathcal{Y}$ and extend d by putting $d(\S,\S) = 0$ while $d(\S,y) = d(y,\S) = 1$ for all points $y \in \mathcal{Y}$. Put $\mathcal{Y}_1 = \mathcal{Y} \cup \{\S\}$ and $S_1 = S \cup \{(x,\S) : x \in \chi\}$. Then (i) holds if $\mathcal{Y}$ and S are replaced by, respectively, $\mathcal{Y}_1$ and S_1 and if Q is replaced by its extension $\tilde{Q}$ to $\mathcal{Y}_1$. Thus, by what we have proved, there is a joint distribution $\tilde{R}$ on $\chi \times \mathcal{Y}_1$ with marginals P and $\tilde{Q}$ such that $\tilde{R}(S_1) > 1-\varepsilon$. Then, since $\tilde{Q}(\mathcal{Y}) = 1$, $\tilde{R}$ is supported by $\chi \times \mathcal{Y}$ and thus (ii) holds with R being the restriction of $\tilde{R}$ to $\chi \times \mathcal{Y}$. □

Using the fact that a complete separable metric space is homeomorphic to a G_δ sub set of compact metric space we obtain the theorem of Strassen mentioned above. This theorem states that (i) and (ii) are equivalent for probability spaces $(\chi, \mathcal{A}, P)$ and $(\mathcal{Y}, \mathcal{B}, Q)$ and sub sets S of $\chi \times \mathcal{Y}$ such that χ and $\mathcal{Y}$ are complete separable metric spaces and S is closed.

6. MAJORIZATION.

The concepts of majorization defined in Marshall and Olkin 1979 generalizes as follows:

Let $\mathcal{E} = (\chi,\mathcal{A},\mu_\theta:\theta\in\Theta)$ and $\mathcal{F} = (\mathcal{Y},\mathcal{B},\nu_\theta:\theta\in\Theta)$ be two measure families. Say that $\underline{\mathcal{E}\text{ weakly supermajorizes }\mathcal{F}}$ if $\mathcal{E}$ is $\nu.(\mathcal{Y})-\mu.(\chi)$ w.r.t. $\mathcal{F}$. Say that $\underline{\mathcal{E}\text{ weakly sub majorizes }\mathcal{F}}$ if $\mathcal{E}$ is $\mu.(\chi)-\nu.(\mathcal{Y})$ deficient w.r.t. $\mathcal{F}$. Say finally that $\underline{\mathcal{E}\text{ majorizes }\mathcal{F}}$ is $\mathcal{E}$ is 0-deficient w.r.t. $\mathcal{F}$.

It is then easily checked that the two kind of weak majorization as well as majorization are partial orderings for measure families.

The basic properties of weak supermajorization are collected in:

Proposition 6.1. (Weak supermajorization).

The following properties are equivalent for measure families $\mathcal{E} = (\chi,\mathcal{A},\mu_\theta:\theta\in\Theta)$ and $\mathcal{F} = (\mathcal{Y},\mathcal{B},\nu_\theta:\theta\in\Theta)$:

(i) $\mathcal{E}$ weakly supermajorizes $\mathcal{F}$.

(ii) $\mu_\theta M\leq\nu_\theta$ for all θ for some transition M from $L(\mathcal{E})$ to $L(\mathcal{F})$.

(iii) If $\theta_1,\ldots,\theta_r\in\Theta$ and ϕ is a sub linear and monotonically decreasing function on R^r then $\int\phi(d\mu_{\theta_1},\ldots,d\mu_{\theta_r})\geq\int\phi(d\nu_{\theta_1},\ldots,d\nu_{\theta_r})$.

Remark 1. If $\mu_\theta M\leq\nu_\theta$ for a transition M and if $\mu_\theta(\chi) = \nu_\theta(\mathcal{Y})$ then actually $\mu_\theta M = \nu_\theta$. It follows that these conditions imply that the inequality in (iii) holds for a sub linear function ϕ provided ϕ is monotonically decreasing in x_i when $\mu_{\theta_i}(\chi)\neq\nu_{\theta_i}(\mathcal{Y})$.

Remark 2. If $\theta_0 \in \Theta$ is such that $\mu_{\theta_0}, \nu_{\theta_0} > 0$, $\|\mu_{\theta_0}\| = \|\nu_{\theta_0}\|$, $\mu_{\theta_0} >> \mathcal{E}$, $\nu_{\theta_0} >> \mathcal{F}$ and the sample spaces are Euclidean then we may, as we saw in section 4, express weak super majorization in terms of dilations. In that case (iii) may be replaced by

(iii') $\int \phi(d\mu_{\theta_1}/d\mu_{\theta_0}, \dots, d\mu_{\theta_r}/d\mu_{\theta_0}) d\mu_{\theta_0}$
$> \int \phi(d\nu_{\theta_1}/d\nu_{\theta_0}, \dots, d\nu_{\theta_r}/d\nu_{\theta_0}) d\nu_{\theta_0}$ when ϕ is monotonically decreasing and convex on R^r.

Remark 3. If $\Theta = \{\theta_0, \theta_1\}$, $\mu_{\theta_0} > 0$, $\nu_{\theta_0} > 0$ and $\|\mu_{\theta_0}\| = \|\nu_{\theta_0}\|$ then it suffices, as explained in section 2, to consider comparison for 2-points sets (testing problems) and we find that (iii) is equivalent to:

(iii") $\mu_{\theta_1}(\chi) < \nu_{\theta_1}(\mathcal{Y})$ and $\int [c\,d\mu_{\theta_0} - d\mu_{\theta_1}]^+ > \int [c\,d\nu_{\theta_0} - d\nu_{\theta_1}]^+$; $c \in R$.

Considering the support function:

$$(c_0, c_1) \rightarrow \begin{cases} \int [c_0 d\mu_{\theta_0} + c_1 d\mu_{\theta_1}]^+ & \text{when } c_1 < 0 \\ \infty & \text{when } c_1 > 0 \end{cases}$$

we see that (iii) now says:

(iii''') $(\mu_{\theta_1}(\chi) < \nu_{\theta_1}(\mathcal{Y})$ and $\min\{\mu_{\theta_1}(\delta) : \mu_{\theta_0}(\delta) = \alpha\} < \min\{\nu_{\theta_1}(\phi) : \nu_{\theta_0}(\phi) = \alpha\}$; $\alpha > 0$, where δ (ϕ) runs through the set of testfunctions in $\mathcal{E}$ $(\mathcal{F})$.

The minima in (iii") may be obtained in the same was as we obtain maximum power by the Neyman-Pearson lemma.

Putting $\alpha = \mu_{\theta_0}(\chi)$ we see that the condition which appears first in (iii''') and in (iii") is superfluous when $\mu_{\theta_0} >> \mu_{\theta_1}$.

Proof: Assume that $\mathcal{E}$ is $\nu.(\mathcal{Y})-\mu(\chi)$ deficient w.r.t. $\mathcal{F}$. Then there is a transition M so that $\|\mu_\theta M-\nu_\theta\|<\nu_\theta(\mathcal{Y})-\mu_\theta(\chi)$ for all $\theta\in\Theta$. This may, by the identity $\|\lambda\| = \int 1d\lambda+2\|\lambda^-\|$, be written $\nu_\theta>\mu_\theta M:\theta\in\Theta$. This in turn implies for any monotonically decreasing sub linear function ϕ on R^r that $\int\phi(d\mu_{\theta_1},\ldots,d\mu_{\theta_r})>\int\phi(d(\mu_{\theta_1}M),\ldots,d(\mu_{\theta_r}M))>\int\phi(d\nu_{\theta_1},\ldots,d\nu_{\theta_r})$. Altogether we have shown that (i)=>(ii)=>(iii). If (iii) holds and ϕ is any sub linear function on R^r then, putting $e^i = (0,\ldots,\overset{(i)}{1},\ldots,0)$, we see that $x\to\phi(x)-\int x_i\phi(e^i)$ is monotonically decreasing so that

$$\int\phi(d\mu_{\theta_1},\ldots,d\mu_{\theta_r})>\int\phi(d\nu_{\theta_1},\ldots,d\nu_{\theta_2})-\sum_{i=1}^{r}\phi(e^i)(\nu_{\theta_i}(\mathcal{Y})-\mu_{\theta_i}(\chi)).$$

Hence (i) holds. □

We omit the proof of the quite similar case of weak sub majorization:

Proposition 6.2. (Weak sub majorization).

The following conditions are equivalent for measure families $\mathcal{E} = (\chi,\mathcal{A},\mu_\theta:\theta\in\Theta)$ and $\mathcal{F} = (\mathcal{Y},\mathcal{B},\nu_\theta:\theta\in\Theta)$:

(i) $\mathcal{E}$ weakly sub majorizes $\mathcal{F}$.

(ii) $\mu_\theta M>\nu_\theta$ for all θ for some transition M from L( ) to $L(\mathcal{F})$.

(iii) $\int\phi(d\mu_{\theta_1},\ldots,d\mu_{\theta_r})>\int\phi(d\nu_{\theta_1},\ldots,d\nu_{\theta_r})$ when $\theta_1,\ldots,\theta_r\in\Theta$ and ϕ a is sub linear and monotonically increasing function on R^r.

Remarks. We may here make similar remarks as those we made after the previous proposition. Thus remarks 1 and 2 applies provided we substitute "increasing" for "decreasing". If $\Theta = \{\theta_0,\theta_1\}$, $\mu_{\theta_0},\nu_{\theta_0}>0$, $\|\mu_{\theta_0}\| = \|\nu_{\theta_0}\|$ then (iii) may be replaced by:

"$\mu_{\theta_1}(\chi)>\nu_{\theta_1}(\mathcal{Y})$ and $\int[cd\mu_{\theta_0}+d\mu_{\theta_1}]^+>\int[cd\nu_{\theta_0}+d\nu_{\theta_1}]^+$; $c\in R$"

or by

"$\mu_{\theta_1}(\chi)>\nu_{\theta_1}(\mathcal{Y})$ and $\max\{\mu_{\theta_1}(\delta):\mu_{\theta_0}(\delta)=\alpha\}>\max\{\nu_{\theta_1}(\phi):\nu_{\theta_0}(\phi)=\alpha\}$; $\alpha>0$"

where $\delta(\phi)$ runs through the set of testfunctions in $\mathcal{E}$ $(\mathcal{F})$.

The condition $\mu_{\theta_1}(\chi)>\nu_{\theta_1}(\mathcal{Y})$ is superfluous when $\mu_{\theta_0}>>\mu_{\theta_1}$.

Combining the two kinds of weak majorization we obtain:

Theorem 6.3. (Majorization criteria for measure families).

The following properties are equivalent for measure families $\mathcal{E}=(\mu_\theta:\theta\in\Theta)$ and $\mathcal{F}=(\nu_\theta:\theta\in\Theta)$

(i) $\mathcal{E}$ majorizes $\mathcal{F}$.
(ii) $\mathcal{E}$ weakly supermajorizes $\mathcal{F}$ and weakly submajorizes $\mathcal{F}$.
(iii) $\mu_\theta M\underset{\theta}{\equiv}\nu_\theta$ for some transition M from $L(\mathcal{E})$ to $L(\mathcal{F})$.
(iv) If $\theta_1,\ldots,\theta_r\in\Theta$ and ψ is sublinear on R^r then

$$\int\psi(d\mu_{\theta_1},\ldots,d\mu_{\theta_r}) > \int\psi(d\nu_{\theta_1},\ldots,d\nu_{\theta_r}).$$

Remark 1. If $\theta_0\in\Theta$ is such that $\mu_{\theta_0}>0$, $\nu_{\theta_0}>0$, $\|\mu_{\theta_0}\| = \|\nu_{\theta_0}\|$, $\mu_{\theta_0}>>\mathcal{E}$ and $\nu_{\theta_0}>>\mathcal{F}$ then we may, provided the sample spaces are Euclidean, express majorization in terms of dilations. In that case (iv) may be expressed as:

(iv') $\int\phi(d\mu_{\theta_1}/d\mu_{\theta_0},\ldots,d\mu_{\theta_r}/d\mu_{\theta_0})/d\mu_{\theta_0}$
$>\int\phi(d\nu_{\theta_1}/d\nu_{\theta_0},\ldots,d\nu_{\theta_r}/d\nu_{\theta_0})d\nu_{\theta_0}$ when ϕ is convex on R^r.

If in addition μ_{θ_0} and ν_{θ_0} are probability measures then this is a particular case of corollary 4.6.

Remark 2. If $\Theta = \{\theta_0, \theta_1\}$, $\mu_{\theta_0} \geq 0$, $\nu_{\theta_0} \geq 0$, $\|\mu_{\theta_0}\| = \|\nu_{\theta_0}\|$ and $c_1 \neq 0$ is a fixed number, then (iv) is equivalent to

(iv") $\mu_{\theta_1}(\chi) \equiv \nu_{\theta_1}(\mathcal{Y})$ and $\int [c_0 d\mu_{\theta_0} + c_1 d\mu_{\theta_1}]^+ \geq \int [c_0 d\nu_{\theta_0} + d_1 d\nu_{\theta_1}]^+$ when $c_0 \in R$.

The function $(c_0, c_1) \to \int [c_0 d\mu_{\theta_0} + c_1 d\mu_{\theta_1}]^+$ is the support function of the closed convex range of $(\mu_{\theta_0}, \mu_{\theta_1})$. Thus conditions (i)-(iv) are in this case also equivalent to the condition that

$\mu_{\theta_1}(\chi) = \nu_{\theta_1}(\mathcal{Y})$ and that the convex hull of the range of $(\mu_{\theta_0}, \mu_{\theta_1})$ contains the range of $(\nu_{\theta_0}, \nu_{\theta_1})$. Furthermore the latter condition implies the first when $\mu_{\theta_0} \gg \mu_{\theta_1}$.

If $\mu_{\theta_1}(\chi) = \nu_{\theta_1}(\mathcal{Y})$ then $(\mu_{\theta_0}, \mu_{\theta_1})$ majorizes $(\nu_{\theta_0}, \nu_{\theta_1})$ if and only if $(\mu_{\theta_0}, \mu_{\theta_1})$ super (sub) majorizes $(\nu_{\theta_0}, \nu_{\theta_1})$.

If $\Theta = \{\theta_0, \theta_1\}$, $\chi = \mathcal{Y} = \{1, \ldots, n\}$ and $\mu_{\theta_0}(i) = \nu_{\theta_0}(i) = 1$; $i = 1, \ldots, n$ then we obtain the equivalence of conditions (i)-(iv) in example 1.5.

7. REFERENCES.

Blackwell, D. 1951. Comparison of experiments. Proc. Second Berkeley Sympos. Math. Statist. Probab. 93-102.

Blackwell, D. 1953. Equivalent comparisons of experiments. Ann. Math. Statist. **24**, 265-272.

Boll, C. 1955. Comparison of experiments in the infinite case. Ph.D. thesis. Standford University.

Dahl, G. 1983. Pseudo experiments and majorization. Thesis. Univ. of Oslo. Statistical research report, 1984.

Fischer, P. and Holbrook J.A.R. 1980. Balayage defined by the non negative convex functions. Proc. Amer. Math. Soc. 79, 445.

Hansen, O.H. and Torgersen, E.N. 1974. Comparison of linear normal experiments. Ann. Statist. 2, 367-373.

Heyer, H. 1973. Mathematische Theorie statistischer Experimente. Springer Verlag.

Kakutani, S. 1941. Concrete representation of abstract (L)-spaces and the mean ergodic theorem. Ann. of Math. (2) 42, 523-537.

Karlin, S. 1983. Comparison of measures, multivariate majorization, and applications to statistics. In: Karlin, Amemiya, Goodman. Studies in econometric time series and multivariate statistics, Academic Press.

LeCam, L. 1964. Sufficiency and approximate sufficiency. Ann. Math. Statist. **35**, 1419-1455.

LeCam, L. 1974. Notes on asymptotic metods in statistical decision theory. Centre de Recherches Math. Univ. de Montréal.

LeCam, L. 1975. Distances between experiments. In: J.N. Srivastava, Ed., A survey of Statistical Design and Linear Models. North-Holland, Amsterdam, pp. 383-396.

Lehmann, E. L. 1983. Comparison of experiments for some multivariate normal situations. In: Karlin, Amemiya, Goodman. Studies in econometric time series and multivariate statistics, Academic Press.

Marshall, A. W. and Olkin, I. 1979. Inequalities: Theory of majorization and its applications. Academic Press.

Millar, P. W. 1983. The minimax principle in asymptotic statistical theory. Lecture notes in Mathematics No 976, 76-265. Springer-Verlag.

Strassen, V. 1965. The existence of probability measures with given margnals. Ann. Math. Statistics. **36**, 423-439.

Swensen, A.R. 1980. Deficiencies in linear normal experiments. Ann. Statist. 8, 1142-1155.

Torgersen, E. 1969. On ε-comparison of experiments. Mimeographed notes. Univ. of California, Berkeley.

Torgersen, E. 1970. Comparison of experiments when the parameter space is finite. Z. Wahrscheinlichkeitstheorie Verw. Geb. **16**, 219-249.

Torgersen, E. 1972a. Local comparison of experiments when the parameter set is one dimensional. Statist. res. report no. 4 1972, Univ. of Oslo.

Torgersen, E. 1972b. Local comparison of experiments. Statist. res. report no. 5, Univ. of Oslo.

Torgersen, E. 1972c. Comparison of translation experiments. Ann. Math. Stat. **43**, 1383-1399.

Torgersen, E. 1976. Comparison of statistical experiments. Scand. Journ. of Statistics. **3**, 186-208.

Torgersen, E. 1982. Comparison of some statistical experiments associated with sampling plans. Probability and Mathematical Statistics **3**, 1-17.

Torgersen, E. 1984. Orderings of linear models. Journal of statistical planning and inference 9, 1-17.

Wald, 1950. Statistical decision functions. New York, Wiley.

CORRECTIONS TO MAJORIZATION AND APPROXIMATE MAJORIZATION FOR FAMILIES OF MEASURES. STATISTICAL RESEARCH REPORT NO 8, 1984

Page	Line	Old version	Corrected version
1.3	9↓	close to P_{θ^0}	close to θ^0
1.6	8↓	$1 = 1,\dots,n$	$j = 1,\dots,n$
2.3	2↑	functions on Θ	functions on T
2.7	5↑	$\mathcal{F} = (\mu_\theta : \theta\in\Theta)$	$\mathcal{E} = (\mu_\theta : \theta\in\Theta)$
3.2	10↓	differentiable in θ^0	differentiable in $t = 0$
3.7	8↓	$\dot{\delta}_{k,\theta^0}(\mathcal{E},\mathcal{F})$	$\dot{\delta}_{k,\theta^0}(\mathcal{F},\mathcal{E})$
3.7	8↑	If $\underset{k,\theta^0}{>}$	If $\mathcal{E} \underset{k,\theta^0}{>} \mathcal{F}$
3.7	6↑–7↑	$\mathcal{E}$ and are locally	$\mathcal{E}$ and $\mathcal{F}$ are locally
3.15	4↑	as $\mathcal{E}$ at θ^0	as $\mathcal{F}$ at θ^0
5.4	12↓	marginals ν and Q	marginals μ and Q
6.1	5↓–6↓	$\nu.(\mathcal{Y})-\mu.(\chi)$ w.r.t.	$\nu.(\mathcal{Y})-\mu.(\chi)$ deficient w.r.t.
6.3	11↓	$\int\phi(d\nu_{\theta_1},\dots,d\nu_{\theta_2})$	$\int\phi(d\nu_{\theta_1},\dots,d\nu_{\theta_r})$
6.3	11↑	M from L( ) to	M from L($\mathcal{E}$) to
6.5	3↓	$\int[c_0 d\nu_{\theta_0}+d_1 d\nu_{\theta_1}]^+$	$\int[c_0 d\nu_{\theta_0}+c_1 d\nu_{\theta_1}]^+$