Stochastic Orders and Decision under Risk
IMS Lecture Notes – Monograph Series (1991)

STOCHASTIC ORDERS AND COMPARISON OF EXPERIMENTS

BY ERIK TORGERSEN

University of Oslo

Exploring criteria for majorization, exact and approximate, univariate and multivariate, we relate them to criteria for information orderings of statistical experiments. After providing some basic criteria for comparison of experiments, we observe their straightforward generalizations to general families of measures. Thus LeCam's randomization criterion extends to a criterion for comparing families of measures. Reversing the randomizations, we obtain dilation like kernels mapping densities, exactly or approximately, into densities.

Using this, we derive criteria for comparison of measures in terms of integrals of given functions. In particular we obtain well-known criteria for one measure being a dilation of another measure and for stochastic orderings of distributions on partially ordered sets.

Experiments having two point parameters sets, i.e. dichotomies, enjoy a variety of striking properties which are not shared by experiments in general. Dichotomies may be studied in terms of their Neyman-Pearson functions, which are functions describing the relationships between the probabilities of errors of the two kinds for most powerful tests. These functions are the inverses of the Lorenz functions of econometrics. Observing this, we readily obtain various criteria for one distribution being approximately Lorenz majorized by another.

1. Introduction. Majorization and Comparison Experiments.

The purpose of this paper is to discuss relationships between developments within the theory of comparison of statistical experiments on the one hand, and various notions of "stochastic" orders on the other. As we shall see, the theory of comparison of experiments not only throws light on standard notions of stochastic order, but also provides interesting generalizations of well-known results.

The paper provides the required results from the theory of statistical experiments. However proofs are often incomplete. The reader who wants

---

AMS 1980 Subject Classification: Primary 62B15; Secondary 62P20

Key words and phrases: Lorenz functions, Neyman-Pearson functions, majorization, Schur convex functions, comparison of experiments, the Gini index.

more background might consult e.g. Heyer (1982), LeCam (1986), Strasser (1985) or Torgersen (1985) and (1991).

In order to indicate that there is a connection, let us begin by considering the notion of majorization in $\mathbb{R}^n$.

Considering two vectors p and q in $\mathbb{R}^d$ one possible definition proclaims that p majorizes q, notation $p \succ q$, if $\sum_i p_i = \sum_i q_i$ and $p_{(1)} + \ldots + p_{(r)} \leq q_{(1)} + \ldots + q_{(r)}$; $r = 1, \ldots, d$. Here, $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(d)}$, for a vector x in $\mathbb{R}^d$, denotes the coordinates $x_1, x_2, \ldots, x_d$ arranged in increasing order.

We shall employ the notations and $a \wedge b$ and $a \vee b$ for, minimum $\{a, b\}$ and maximum $\{a, b\}$ respectively. More generally $\wedge$ and $\vee$ may be used to denote infimum and supremum.

In order to describe other criteria for majorization, we shall need the following notations for a vector (point) $x = (x_1, \ldots, x_d)$ in $\mathbb{R}^d$ and for numbers α and λ in $[0, 1]$:

$$\|x\|_1 = \sum_{i=1}^{d} |x_i|$$

A_x =the subset of $\mathbb{R}^2$ consisting of all pairs

$$((\delta_1 + \ldots + \delta_d)/d, \delta_1 x_1 + \ldots + \delta_d x_d) = \sum_{i=1}^{d} \delta_i (1/d, x_i)$$

where $\delta_1, \ldots, \delta_d$ vary independently in [0,1].

$$b_x(\lambda) = \sum_{i=1}^{d} ((1 - \lambda)/d)) \wedge (\lambda x_i) \quad \text{when} \quad x_1, \ldots, x_d \geq 0.$$

$$\beta_x(\alpha) = \sup\{v : (\alpha, v) \in A_x\}$$
$$= \sup\{v : (u, v) \in A_x \text{ for some } u \leq \alpha\}$$

so that, by Neyman-Pearson's lemma:

$$\beta_x((d - r + 1)/d) = x_{(r)} + \ldots + x_{(d)}; \quad r = 1, \ldots, d.$$

We shall also use the notation $e = (1, \ldots, 1)$. Thus e/d is the probability vector in $\mathbb{R}^d$ corresponding to the uniform distribution on $\{1, \ldots, d\}$.

A selection of majorization criteria are collected in the following theorem.

THEOREM 1.1. (Majorization criteria). *The following conditions are equivalent for vectors $p = (p_1, \ldots, p_d)$ and $q = (q_1, \ldots, q_d)$ in $\mathbb{R}^d$:*

(i) $p \succ q$

(ii) $\sum_i p_i = \sum_i q_i$ *and* $p_{(r)} + \ldots + p_{(d)} \geq q_{(r)} + \ldots + q_{(d)}$; $r = 1, \ldots, d$.

The inequality for $r = 1$ is, by the first condition, necessarily an equality.

(iii) $\sum_i (p_i - c)^+ \geq \sum_i (q_i - c)^+$; $c \in \mathbb{R}$.

(iv) $\sum_i (p_i - c)^- \geq \sum_i (q_i - c)^-$; $c \in \mathbb{R}$.

(v) $\|p - ce\|_1 \geq \|q - ce\|_1$; $c \in \mathbb{R}$.

(vi) $\sum_i g(p_i) \geq \sum_i g(q_i)$ *when* g *is convex on* $\mathbb{R}$.

(vii) $\varphi(p) \geq \varphi(q)$ *when* φ *is quasiconvex and permutation symmetric on* $\mathbb{R}^d$.

(viii) $q = Mp$ *for a* $d \times d$ *doubly stochastic matrix* M.

(ix) $q \in \langle \{\pi(p) : \pi \in \Pi\} \rangle$ *where* $\langle\ \rangle$ *denotes convex hull and* Π *is the group of coordinate permutations on* $\mathbb{R}^d$.

(x) $A_p \supseteq A_q$.

(xi) $\beta_p(\alpha) \geq \beta_q(\alpha)$; $0 \leq \alpha \leq 1$.

(xii) $\binom{e/d}{p} M = \binom{e/d}{q}$ *for a (necessarily doubly stochastic)* $d \times d$ *Markov matrix* M.

(xiii) The empirical distribution function based on the observations $p_1, \ldots,$ p_d *is a dilation of the empirical distribution function based on the observations* $q_1, \ldots, q_d$.

If p *and* q *are probability vectors then these conditions are equivalent to*

(xiv) $b_p(\lambda) \leq b_q(\lambda)$; $0 \leq \lambda \leq 1$.

Criteria (i)–(ix) are well-known and may be found in e.g. Marshall and Olkin (1979). The other criteria are not so well-known. Criteria (x)–(xii) and (xiv) are discussed in Dahl (1983).

Criterion (xiv) is only stated for probability vectors p and q. This restriction does not however amount to much. Indeed, for $c \in \mathbb{R}$ and $t > 0$, $p \succ q$ if and only if $\frac{1}{t}(p - ce) \succ \frac{1}{t}(q - ce)$. If $\sum p_i = \sum q_i$, then the last vectors are probability vectors provided $c < \min(p_{(1)}, q_{(1)})$ and $t = \sum_i (p_i - c)$.

If the vectors p and q are probability vectors in $\mathbb{R}^d$, then several criteria of the above theorem have interesting interpretations in terms of statistical decision theory.

Consider a statistical model obtained by observing a random variable X whose distribution P_θ depends on an unknown parameter θ. Assume for the moment that we know that θ is one of the numbers 0 and 1 and that X is one of the numbers $1, 2, \ldots, d$. Assume also that X is uniformly distributed when $\theta = 0$, while the distribution of X is given by the probability vector p when $\theta = 1$. In other words:

$$P_0(X = i) = \frac{1}{d}; \quad i = 1, \ldots, d$$

and

$$P_1(X = i) = p_i; \quad i = 1, \ldots, d.$$

An experiment (model) having a two point parameter set is called a *dichotomy*. If parameter values are identified with positions of coordinates, then a dichotomy may be represented as an ordered pair of probability distributions.

Let $\mathcal{E}_p$ be the statistical experiment realized by the observation X. Then $\mathcal{E}_p$ is the dichotomy (P_0, P_1). For this dichotomy we find that:

> A_p is the set of power functions in $\mathcal{E}_p$ for testing "$\theta = 0$" against "$\theta = 1$",
>
> $\beta_p(\alpha)$ is the power of the most powerful level α test in $\mathcal{E}_p$ for the same testing problem

and

> $b_p(\lambda)$ is the minimum Bayes risk in $\mathcal{E}_p$ for estimating θ with 0–1 loss and for the prior distribution assigning mass λ to the parameter value $\theta = 1$.

The interpretations of conditions (x), (xi) and (xiv) follows immediately from this.

Condition (xii) is just a rephrasing of condition (viii). Its statistical interpretation is nevertheless fundamental since it says that the experiment $\mathcal{E}_q$ may be obtained from the experiment $\mathcal{E}_p$ by a known chance mechanism. This mechanism is given by the (necessarily doubly stochastic) Markov matrix M.

Proceeding to more general decision spaces we may infer, see e.g. Torgersen (1970), that condition (vi) simply states that maximum Bayes utility in $\mathcal{E}_p$ is always at least as large as any Bayes utility obtainable in $\mathcal{E}_q$. Of course, we might replace the inequality in (vi) with the reverse inequality for concave functions g. Doing so we see that (vi) also amounts to the condition that the minimum Bayes risk in $\mathcal{E}_p$ is at most equal to any obtainable Bayes risk in $\mathcal{E}_q$.

Condition (xiii) expresses that under P_0, the distribution of the likelihood dP_1/dP_0 is a dilation of the corresponding distribution in $\mathcal{E}_q$.

By Blackwell (1953) conditions (x)–(xiv) are all equivalent ways of saying that the statistical experiment $\mathcal{E}_p$ is at least as informative as the experiment $\mathcal{E}_q$.

EXAMPLE 1.2. (Monotone likelihood and majorization). Put $s_t(d) = 1^t + 2^t + \ldots + d^t$ and, for each number t, denote as p_t the probability vector in $\mathbb{R}^d$ whose ith coordinate is $i^t/s_t(d)$. Then, $p_{t_1} \prec p_{t_2}$ if $0 \leq t_1 \leq t_2$ or if $t_2 \leq t_1 \leq 0$ (see p. 130 in Marshall and Olkin (1979)). From an informational point of view this may be inferred from the observation that the statistical model $(p_t : t \in \mathbb{R})$ has monotone likelihood in $X(i) \equiv i$.

This observation implies that the dichotomy (p_{u_1}, p_{u_2}) provides at most as much information as the dichotomy (p_{v_1}, p_{v_2}) when $v_1 \leq u_1 \leq u_2 \leq v_2$. If $0 \leq t_1 \leq t_2$ then we may choose $u_1 = v_1 = 0$, $u_2 = t_1$ and $v_2 = t_2$ and thereby obtain the desired conclusion. If $t_2 \leq t_1 \leq 0$ then we may choose $u_2 = v_2 = 0$, $u_1 = t_1$ and $v_1 = t_2$ and we arrive at the same conclusion by reparametrization. For a model $(e/d, p)$ reparametrization amounts to replacing it with the reversed model $(p, e/d)$.

Generalizing this idea, we see that $p \prec q$ for probability vectors p and q in $\mathbb{R}^d$ provided the fraction q_i/p_i is monotonically increasing in p_i as long as this fraction is defined, i.e. as long as $p_i + q_i > 0$. Indeed then the likelihood ratio of the experiment $(e/d, p, q)$ is monotonically increasing in p_i and thus the dichotomy $(e/d, p)$ is at most as informative as the dichotomy $(e/d, q)$.

The weaker concepts of weak submajorization and of weak supermajorization fit nicely into a decision theoretical framework, too. However, before discussing these ideas we consider approximate majorization.

Recall the notation $\|x\|_1 = \sum_{i=1}^d |x_i|$ for a vector $x \in \mathbb{R}^d$. The notation reflects the fact that $\|x\|_1$ is the L_1 norm of x based on the counting distribution on subsets of $\{1, \ldots, d\}$.

Considering two vectors p and q in $\mathbb{R}^d$ and a constant $\epsilon \geq 0$ we shall say that p *ϵ-majorizes* q if p majorizes a vector $\tilde{q}$ such that $\|\tilde{q} - q\|_1 \leq \epsilon$. Thus p majorizes q if and only if p 0-majorizes q. On the other hand, p ϵ-majorizes q whenever $\epsilon \geq \|p - q\|_1$.

Again there is a variety of equivalent conditions.

Before deriving the analogs of the criteria listed in Theorem 1.1, let us note some reformulations of ϵ-majorization. Observe first that p ϵ-majorizes q if and only if q admits a decomposition $q = \tilde{q} + v$, where $\tilde{q} \prec p$ and $\|v\|_1 \leq \epsilon$. It follows that the support function of the convex set consisting of vectors q which are ϵ-majorized by p is $a \to \vee_\pi \sum_{i=1}^d a_{\pi(i)} p_i + \epsilon \vee_{i=1}^d |a_i|$, where π runs through the permutation group on $\{1, \ldots, n\}$. Hence q is ϵ-majorized by p if and only if $(q, a) \leq \vee_\pi (\pi(a), p) + \epsilon \vee_i |a_i|; a \in \mathbb{R}^d$, where $\pi(a) \equiv (a_{\pi(1)}, \ldots, a_{\pi(d)})$.

Observe next that it suffices to consider vectors a such that $\vee_i |a_i| \leq 1$. Furthermore, a vector a satisfies this condition if and ony if it is of the form $2b - e$, where $0 \leq b_i \leq 1$; $i = 1, \ldots, d$. Thus q is ϵ-majorized by p if and only if $\vee_\pi (\pi(b), p) \geq (b, q) - \frac{1}{2} \sum_i (q_i - p_i) - \epsilon/2$; $0 \leq b \leq e$. Now the set of extreme points of the order interval $[0, e]$ consists precisely of the vectors b whose coordinates are 0 or 1.

It follows that q is ϵ-majorized by p if and only if $\vee_\pi (\pi(b), p) \geq (\vee_\pi (\pi(b), q) - \frac{1}{2} \sum (q_i - p_i) - \epsilon/2$ for all vectors b whose coordinates are 0 or 1. By Neyman-Pearson's lemma this amounts to the condition that $\sum_{i=r}^d p_{(i)} \geq \sum_{i=r}^d q_{(i)} -$

$\frac{1}{2}(\sum_{i=1}^{d}(q_i - p_i)) - \epsilon/2$; $r = 1, \ldots, d+1$, where we put $\sum_{i=d+1}^{d} = 0$.

If $d \leq 2$, as observed by Dahl (1983) when $\sum p_i = \sum q_i$, then the last condition may be reduced to 0-majorization (i.e. majorization), since it says that $p^\epsilon \succ q$, where:

$$p_1^\epsilon = p_{(1)} - \frac{1}{2}\epsilon + \frac{1}{2}\sum_i (q_i - p_i)$$

$$p_i^\epsilon = p_{(i)}; \quad i = 1, \ldots, d-1$$

and

$$p_d^{(\epsilon)} = p_{(d)} + \frac{1}{2}\epsilon + \frac{1}{2}\sum_i (q_i - p_i).$$

Listing the analogs of the criteria for majorization we obtain:

THEOREM 1.3. (Criteria for ϵ-majorization). *The following conditions are equivalent for vectors p and q in $\mathbb{R}^d$:*

(i_1) : *p ϵ-majorizes q.*

(i_2) : $(a, q) \leq \vee_\pi(\pi(a), p) + \epsilon \vee_i |a_i|$; $a \in \mathbb{R}^d$.

(ii_1) : $p_{(r)} + \ldots + p_{(d)} \geq q_{(r)} + \ldots + q_{(d)} - \epsilon/2 - \frac{1}{2}\sum_i (q_i - p_i)$, $r = 1, \ldots, d$ *and* $\sum_i (p_i - q_i) \leq \epsilon$.

(ii_2) : *(If $d \geq 2$):* $p^\epsilon \succ q$.

(iii) $\sum (p_i - c)^+ \geq \sum (q_i - c)^+ - \frac{1}{2}\epsilon - \frac{1}{2}\sum_i (q_i - p_i)$; $c \in \mathbb{R}$.

(iv) $\sum (p_i - c)^- \geq \sum (q_i - c)^- - \frac{1}{2}\epsilon - \frac{1}{2}\sum_i (p_i - q_i)$; $c \in \mathbb{R}$.

(v) $\|p - ce\|_1 \geq \|q - ce\|_1 - \epsilon$; $c \in \mathbb{R}$.

(vi_1) $\sum g(p_i) \geq \sum g(q_i) - [|g'(-\infty)| \vee |g'(\infty)|]\epsilon$, *when g is convex on $\mathbb{R}$.*

(vi_2) $\sum g(p_i) \geq \sum g(q_i) - \frac{1}{2}[g'(-\infty) + g'(\infty)] \sum_{i=1}^{d}(q_i - p_i) - \frac{1}{2}[g'(\infty) - g'(-\infty)]\epsilon$, *when g is convex on $\mathbb{R}$ and the quantities $g'(-\infty) = \lim_{x \to -\infty} g'(x)$ and $g'(\infty) = \lim_{x \to \infty} g'(x)$ are both finite.*

(vii) $g(p) \geq g(q) - \sup\{[g(q) - g(\tilde{q})] : \|\tilde{q} - q\|_1 \leq \epsilon\}$, *when g is Schur convex on $\mathbb{R}$ (i.e. g is monotonically increasing for the majorization ordering.)*

(viii) $\|q - Mp\|_1 \leq \epsilon$ *for a doubly stochastic matrix M.*

(ix) The $\| \|_1$ distance between q and the convex hull of points obtained by permuting the coordinates of p is at most ϵ.

(x) $A_p + \{0\} \times [-\frac{1}{2}\epsilon, \frac{1}{2}\epsilon] + (0, \frac{1}{2}\sum_{i=1}^{d}(q_i - p_i)) \supseteq A_q$.

(xi) $\beta_p(\alpha) \geq \beta_q(\alpha) - \frac{1}{2}\epsilon - \frac{1}{2}\sum_{i=1}^{d}(q_i - p_i)$; $0 \leq \alpha \leq 1$.

(xii) $eM = e$ and $\|pM - q\|_1 \leq \epsilon$ for a Markov (necessarily doubly stochastic) matrix M.

(xiii) *The empirical distribution function F_p based on the observations $p_1, \ldots, p_d$ is a $(F_q, \epsilon/d)$-dilation of the empirical distribution function F_q based on $q_1, \ldots, q_d$. Here a (F_q, ϵ) dilation is a Markov kernel D such that $\int |\int x D(dx|y) - y| F_q(dy) \leq \epsilon$.*

If p and q are probability vectors, then these conditions are equivalent to:

(xiv) $b_p(\lambda) \leq b_q(\lambda) + \frac{1}{2}\epsilon\lambda$; $0 \leq \lambda \leq 1$.

REMARK 1. Dahl (1983) established the equivalence of conditions (i)-(xii) and (xiv) when $\sum p_i = \sum q_i$.

REMARK 2. By the terminology of Torgersen (1985) these conditions amount to the condition that the measure pair $(e/d, p)$ is $(0, \epsilon)$ deficient w.r.t. the measure pair $(e/d, q)$. The equivalence of the above conditions follows then from the general theory of measure familes. It may however be instructive to consider first the direct proof given here.

PROOF OF THEOREM 1.3. We have observed above that conditions (i_1)–(ii_2) were all equivalent. By Theorem 1.1 these conditions are also equivalent with conditions (viii), (ix) and (xii).

If (i_1) holds, then $\|q - \tilde{q}\|_1 \leq \epsilon$ for a vector $\tilde{q}$ such that $p \succ \tilde{q}$. Then $\|q - ce\|_1 - \|p - ce\|_1 \leq \|q - ce\|_1 - \|\tilde{q} - ce\|_1 \leq \|q - \tilde{q}\|_1 \leq \epsilon$. Thus $(i_1) \Rightarrow$(v). Furthermore, by the identities $z^{\pm} \equiv \frac{1}{2}(z \pm |z|)$, conditions (iii) and (iv) are both equivalent to condition (v).

If condition (vi_1) holds and if the convex function g is such that the quantities $g'(\pm\infty) = \lim_{x \to \pm\infty} g'(x)$ are finite, then we may replace g in the inequality in (vi) with the function $x \to g(x) - \frac{1}{2}[g'(-\infty) + g'(\infty)]x$. This shows that the inequality in (vi_2) holds for g.

Applying (vi_2) to $g(x) \equiv (x - ce)^+$ we see that (iii) holds. Letting $c \to \pm\infty$ in (iii) we find that $|\sum p_i - \sum q_i| \leq \epsilon$. If so and if (vi_2) holds for g, then the quantity

$$\frac{1}{2}[g'(-\infty) + g(\infty)] \sum_{i=1}^{d} (q_i - p_i) + \frac{1}{2}[g'(\infty) - g'(-\infty)]\epsilon$$

is between $g'(\infty)\epsilon$ and $-g'(-\infty)\epsilon$. Thus (vi_2) implies that the inequality in (vi_1) holds when the quantities $g'(\pm\infty)$ are finite. If, however, one of the quantities $g'(\pm\infty)$ is infinite, then (vi) is trivial for g, unless $\epsilon = 0$. By the above observation, $(vi_2) \Rightarrow$(iii) and by Theorem 1.1, (iii) amounts to the condition that $p \succ q$, when $\epsilon = 0$. Thus, by Theorem 1.1 again, (vi_2) always implies (vi_1). This shows that conditions (vi_1) and (vi_2) are equivalent and that these conditions imply conditions (iii)–(v). On the other hand, if (iii) holds, then condition (vi_2) holds whenever g is of the form $g(x) \equiv l(x) +$

$\sum_{i=1}^{s} b_i(x - t_i)^+$, where l is linear and $b_1, \ldots, b_s \geq 0$. Any polygonal convex function g is of this form and thus (vi_2) follows by approximation. Altoghether this shows that conditions (iii)–(vi_2) are equivalent.

Note next that the support function of the planar convex set $A_x = \{1/d((\delta_1 + \ldots + \delta_d), \sum_{i=1}^{d} \delta_i x_i) : 0 \leq \delta \leq e\}$ is $(\xi, \eta) \to \sum_i (\xi/d + \eta x_i)^+$, while the support function of the segment $\{0\} \times [-\epsilon/2, \epsilon/2]$ is $(\xi, \eta) \to \frac{1}{2}|\eta|\epsilon$. Thus condition (x) may be expressed as

$$\sum_i (\eta p_i + \xi/d)^+ + \frac{1}{2}|\eta|\epsilon + \frac{1}{2}\eta \sum_i (q_i - p_i) \geq \sum_i (\eta q_i + \xi/d)^+; \xi, \eta \in \mathbb{R}.$$

We may here assume that $\eta \neq 0$ and thus that $\eta = \pm 1$. Replacing ξ/d with ξ, the last set of inequalities reduces to:

$$\sum_i (p_i \pm \xi)^+ + \frac{1}{2}\epsilon \pm \frac{1}{2}\sum (q_i - p_i) \geq \sum_i (q_i \pm \xi)^+; \ \xi \in \mathbb{R}.$$

It follows that condition (x) is equivalent to onditions (iii)–(vi_2). Condition (x) implies also that $(\alpha, \beta_q(\alpha)) = (u, v) + \theta(0, \epsilon/2) + (0, \frac{1}{2}\sum_i (q_i - p_i))$ for some point $(u, v) \in A_p$, whenever $0 \leq \alpha \leq 1$. If so, then $u = \alpha$ and $v \leq \beta_p(\alpha)$, so that $\beta_q(\alpha) \leq \beta_p(\alpha) + \epsilon/2 + \frac{1}{2}\sum_{i=1}^{d}(q_i - p_i), 0 \leq \alpha \leq 1$. Thus (x)⇒(xi) and, substituting $\alpha = \frac{d-r+1}{d}, r = 1, \ldots, d$, by Neyman-Pearson's lemma, (xi)⇒(ii_1). However, we observed above that (i_1) (which is equivalent with (ii_1)) implies condition (v).

It follows that conditions (i)–(vi_2) and (viii)–(xii) are all equivalent.

If $p \succ \tilde{q}$ and g is Schur convex, then $g(p) = g(q) - (g(q) - g(p)) \geq g(q) - (g(q) - g(\tilde{q}))$ and thus (i)⇒(vii). On the other hand, if (vii) holds for $g(x) \equiv \sum_i |x_i - c|$, then

$$\sum_i (p_i - c) \geq \sum (q_i - c) - \sup \left\{ \sum_i |q_i - c| - \sum_i |\tilde{q}_i - c| : \|q - \tilde{q}\|_1 \leq \epsilon \right\}$$
$$\geq \sum |q_i - c| - \epsilon.$$

Hence condition (vii) is equivalent to the other conditions treated so far. Furthermore, if p and q are probability vectors, then condition (xiv) is just a reformulation of conditions (iii)–(v).

It remains to consider condition (xiii).

Assume first that p ϵ-majorizes q. By (viii) $\|q - Mp\|_1 \leq \epsilon$ for a doubly stochastic matrix M. For each $y \in \{q_1, \ldots, q_d\}$, let $D(\cdot|y)$ be the probability distribution which assigns mass $\sum\{M(j|i) : p_i = x, q_j = y\}/\#\{j : q_j = y\}$

to $x \in \{p_1, \ldots, p_d\}$. Then D is a Markov kernel such that $F_p = D\, F_q$ and, putting $\tilde{q} = Mp$, we find:

$$\left|\sum_x xD(x|y)F_q(y) - yF_q(y)\right| = \frac{1}{d}\left|\sum\{\tilde{q}_j - q_j : q_j = y\}\right|$$
$$\leq \frac{1}{d}\sum\{|\tilde{q}_j - q_j| : q_j = y\}$$

so that $\int |\int xD(dx|y) - y|F_q(dy) \leq \epsilon/d$.

The proof is now completed by checking that if $F_p = DF_q$ for a $(F_q, \epsilon/d)$ dilation D, then $\|pM - q\|_1 \leq \epsilon$ for the doubly stochastic matrix M given by:

$$M(j|i) = D(p_i|q_j)/\#\{k : p_k = p_i\}; \quad i = 1, \ldots, d,\ j = 1, \ldots, d. \qquad \blacksquare$$

Consider now the particular case where p ϵ-majorizes q for $\epsilon = \sum_i (q_i - p_i)$. By Theorem 1.3, this amounts to the condition that

$$p_{(1)} + \ldots + p_{(r)} \leq q_{(1)} + \ldots + q_{(r)}; \quad r = 1, \ldots, d.$$

By the terminology in Marshall and Olkin (1979) this is precisely the condition that p weakly supermajorizes q.

Consider next the case where $\epsilon = \sum_i (p_i - q_i)$. Again by Theorem 1.3, this amounts to the condition that $p_{(r)} + \ldots + p_{(d)} \geq q_{(r)} + \ldots + q_{(d)}$; $r = 1, \ldots, d$ and this is the condition that p weakly submajorizes q.

We state this as a corollary.

COROLLARY 1.4. (Weak majorization). *Let p and q be vectors in $\mathbb{R}^d$. Then:*

(i) p weakly submajorizes q if and only if $\epsilon = \sum_i p_i - \sum_i q_i \geq 0$ and p ϵ-majorizes q.

(ii) p weakly super majorizes q if and only if $\epsilon = \sum_i q_i - \sum_i p_i \geq 0$ and p ϵ-majorizes q.

Thus Theorem 1.3 furnishes equivalent criteria for weak majorization.

By Torgersen (1985) these concepts of weak majorization extend naturally to general measure familes.

Theorem 1.3 provides several expressions for the smallest quantity ϵ such that p ϵ-majorizes q. Denoting this quantity by $\dot{\delta}(p, q)$ we obtain from criteria

(ii$_1$), (v) and (xi) the expressions:

$$\dot{\delta}(p,q) = \bigvee_{r=1}^{d+1} 2\left[\sum_{r\le i\le d}(q_{(i)} - p_{(i)}) - \sum_i (q_i - p_i)\right]$$
$$= \bigvee_c [\|q - ce\|_1 - \|p - ce\|_1]$$
$$= 2 \bigvee_{0\le\alpha\le 1}\left[\beta_q(\alpha) - \beta_p(\alpha) - \sum_i (q_i - p_i)\right]$$

Trivially, $0 \le \dot{\delta}(p,q) \le \|p - q\|_1$ and $\dot{\delta}(p,q) = 0$ if and only if p majorizes q.

Furthermore $\dot{\delta}(p',p''') \le \dot{\delta}(p',p'') + \dot{\delta}(p'',p''')$ for any three vectors p', p'' and p''' in $\mathbb{R}^d$.

Symmetrizing, we obtain the majorization pseudo metric $\dot{\Delta}$ on $\mathbb{R}^d$ defined as follows:

$$\dot{\Delta}(p,q) = \max(\dot{\delta}(p,q), \dot{\delta}(q,p))$$
$$= \bigvee_{r=1}^{d+1}\left|2\sum_{r\le i\le d}(q_{(i)} - p_{(i)}) - \sum_i (q_i - p_i)\right|$$
$$= \bigvee_c |\|p - ce\|_1 - \|q - ce\|_1|$$
$$= 2 \bigvee_{0\le\alpha\le 1}\left|\beta_p(\alpha) - \beta_q(\alpha) - \sum_i (p_i - q_i)\right|.$$

EXAMPLE 1.5. (Majorization between vectors of possibly different dimensions). Let $p = (p_1, \ldots, p_m)$ and $q = (q_1, \ldots, q_n)$ be probability vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. For $k = 1, 2, \ldots$, let $u^{(k)}$ denote the probability vector $(1/k, \ldots, 1/k)$ in $\mathbb{R}^k$.

By sufficiency the dichotomy $(u^{(m)}, p)$ is at least as informative as the dichotomy $(u^{(n)}, q)$ if and only if the product dichotomy $(u^{(m)}, p) \times (u^{(n)}, u^{(n)})$ is at least as informative as the product dichotomy $(u^{(m)}, u^{(m)}) \times (u^{(n)}, q)$. In other words $(u^{(m)}, p)$ is at least as informative as $(u^{(n)}, q)$ if and only if the $m \cdot n$ dimensional probability vector $(p_i/n : i = 1, \ldots, m, j = 1, \ldots, n)$ majorizes the $m \cdot n$ dimensional probability vector $(q_j/m : i = 1, \ldots, m, j = 1, \ldots, n)$.

The underlying idea is the obvious fact that information is not altered if we supplement our observations with ancillary observations which are independent of the original observations.

Turning to multivariate majorization, consider for a fixed set T a family $\mathcal{V} = (p_t : t \in T)$ of vectors in $\mathbb{R}^d$. Extend T by adding a point θ_0 not

belonging to T and take $\Theta = T \cup \{\theta_0\}$ as a parameter set. Let μ_{θ_0} be the uniform probability distribution on $\{1, \ldots, d\}$ given by the probability vector $u^{(d)} = (1/d, \ldots, 1/d)$. If $\theta \in T$, then let μ_θ be the measure on the subsets of $\{1, \ldots, d\}$ which assigns mass $p_\theta(i)$ to the point i.

In this way we assign to the family $\mathcal{V}$ a family $\widehat{\mathcal{V}}$ of measures on $\{1, \ldots, d\}$. If the vectors p_t; $t \in T$ are all probability vectors then $\widehat{\mathcal{V}}$ is a statistical experiment. In general $\widehat{\mathcal{V}}$ is a measure family as defined in Section 3.

Considering two families $\mathcal{V} = (p_t : t \in T)$ and $\mathcal{W} = (q_t : t \in T)$ of probability vectors we may consider the impact of the requirement that the statistical model $\widehat{\mathcal{V}}$ is at least as informative as $\widehat{\mathcal{W}}$. By Blackwell (1953) this amounts to the condition that $q_t \equiv Mp_t$ for a doubly stochastic matrix M. In the same paper he proves that this is equivalent to

$$\sum_{i=1}^{d} \varphi(p_{t_1}(i), p_{t_2}(i), \ldots, p_{t_r}(i)) \geq \sum_{i=1}^{d} \varphi(q_{t_1}(i), q_{t_2}(i), \ldots, q_{t_r}(i))$$

whenever $t_1, \ldots, t_r \in T$ and φ is convex on $\mathbb{R}^r$. Actually it suffices to consider functions φ which are maxima of at most d affine functions on $\mathbb{R}^r$.

The above inequality may be written

$$\int \varphi(p_{t_1}, \ldots, p_{t_r}) dF_p \geq \int \varphi(q_{t_1}, \ldots, q_{t_r}) dF_q$$

where F_p is the empirical distribution function based on the function valued observations $p.(1), \ldots, p.(d)$ and F_q is the empirical distribution function based on the observations $q.(1), \ldots, q.(d)$. We may infer from this that the equivalence of these conditions is a very special case of the dilation criterion in Strassen (1965).

Appealing to the results of Section 3 we see that the condition that the vectors are probability vectors does not play any essential role. This yields the following result of Dahl (1983).

THEOREM 1.6. (Multivariate majorization). *The following conditions are equivalent for families $(p_t : t \in T)$ and $(q_t : t \in T)$ of vectors in $\mathbb{R}^d$:*

(i) $q_t \equiv_t Mp_t$ *for a doubly stochstic matrix* M.

(ii) $\sum_{i=1}^{d} \varphi(p_{t_1}(i), \ldots, p_{t_r}(i)) \geq \sum_{i=1}^{d} \varphi(q_{t_1}(i), \ldots, q_{t_r}(i))$ *whenever* $t_1, \ldots, t_r \in T$ *and* φ *is convex on* $\mathbb{R}^r$. *(Actually it suffices to consider functions* φ *which are maxima of at most* d *linear functionals.)*

(iii) The empirical distribution function F_p *based on the observations* $p.(1), \ldots, p.(d)$ *is a dilation of the empirical distribution function* F_q *based on*

the observations $q.(1), \ldots, q.(d)$. *(The observations are all real valued functions on* T.*)*

Proceeding to ϵ-deficiency, see Section 3, this extends as follows:

THEOREM 1.7. (Approximate multivariate majorization). *Let* $(p_t : t \in T)$ *and* $(q_t : t \in T)$ *be two familes of probability vectors in* $\mathbb{R}^d$.

Consider also a family $\epsilon = (\epsilon_t : t \in T)$ *of nonnegative numbers. Then the following conditions are equivalent:*

(i) $\|q_t - Mp_t\|_1 \leq \epsilon_t; t \in T$ *for a doubly stochastic matrix* M.

(ii) $|\sum_i p_t(i) - \sum_i q_t(i)| \leq \epsilon_t; t \in T$ *and*

$$\sum_{i=1}^{d} \psi(1, p_{t_1}(i), \ldots, p_{t_r}(i)) \geq \sum_{i=1}^{d} \psi(1, q_{t_1}(i), \ldots, q_{t_r}(i)) - \frac{1}{2} \sum_{\nu=1}^{r} \left[\sum_{i=1}^{d} (q_{t_\nu}(i) - p_{t_\nu}(i))(\psi(0, e_\nu) - \psi(0, -e_\nu)) \right] - \frac{1}{2} \sum_{\nu=1}^{r} (\psi(0, e_\nu) + \psi(0, -e_\nu))\epsilon_\nu,$$

whenever $t_1, \ldots, t_r \in T$ *and* ψ *is sublinear on* $\mathbb{R}^{r+1}$. *Here* $e_\nu = (0, \ldots 1, \ldots 0); \nu = 1, \ldots, r$ *is the* ν*th unit vector in* $\mathbb{R}^r$.

(iii) The empirical distribution function F_p *based on the observations* $p.(1), \ldots, p.(d)$ *is a* $(F_q, \epsilon/d)$ *dilation of the empirical distribution function* F_q *based on the observations* $q.(1), \ldots, q.(d)$. *Here a Markov kernel* D *is called a* (F_q, ϵ) *dilation if* $\int |\int x_t D(dx|y) - y_t| F_q(dy) \leq \epsilon_t$ *when* $t \in T$.

The analogous results for infinite populations will be consided in Section 5. Before doing so, however, we shall provide some useful tools from decision theory and in particular from the theory of statistical experiments.

2. The Framework of Decision Theory. A nonsequential statistical decision problem is defined by a statistical model (experiment) along with a loss function defined on some decision space. The problem is to select an appropriate decision rule. Adopting the point of view, as we shall, that the quality of a decision rule resides in its risk function, the problem amounts to make a choice within the set of available risk functions.

For a given decision rule the decision taken is a random variable. The distribution of this random variable as a function of the unknown parameter is the *performance function* of the given decision rule. The loss, being a function of this variable, is also a random variable and its expectation as a function of the unknown parameter is the risk function of the given decision rule.

Let us describe these objects within the standard measure theoretical set up.

A statistical experiment (model) is a family of probability measures on a common measurable space called *the sample space* of the experiment.

The indexing set is *the parameter set* of the experiment. Thus an experiment $\mathcal{E}$ having parameter set Θ and sample space $(\mathcal{X}, \mathcal{A})$ is a family $(P_\theta : \theta \in \Theta)$ of probability measures on $(\mathcal{X}, \mathcal{A})$. This experiment may be denoted as $\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_\theta : \theta \in \Theta)$ or just as $\mathcal{E} = (P_\theta : \theta \in \Theta)$.

It may be desirable to permit more general objects as statistical experiments. Thus we may omit any requirement of σ-additivity of the set functions P_θ. Even more generally we may consider experiments $\mathcal{E} = (P_\theta : \theta \in \Theta)$ where the P_θ's are nonnegative normalized elements of abstract L-spaces. Although these creatures may appear strange, they do not represent anything radically new from the statistical point of view. Indeed LeCam (1964) showed that any experiment is statistically equivalent to an experiment in the traditional form.

Mathematically it is often convenient to replace the sample space of $\mathcal{E} = (P_\theta : \theta \in \Theta)$ with the vector lattice of finite measures which are dominated by countably infinite convex combinations of the measures P_θ. Equipped with the total variation norm $\|\cdot\|$, this space becomes a Banach lattice $L(\mathcal{E})$ having the additional property that the norm is additive on the cone of nonnegative elements. A Banach lattice having the latter property is called a *L-space* and $L(\mathcal{E})$ is called the *L-space of* $\mathcal{E}$.

Equipped with the dual ordering, the conjugate space $M(\mathcal{E}) = L(\mathcal{E})^*$ of the Banach space $L(\mathcal{E})$ becomes another Banach lattice. The additivity property of the norm on $L(\mathcal{E})$ implies that the norm of a maximum of nonnegative elements of $M(\mathcal{E})$ equals the maximum of the norms of these elements. A Banach lattice having this property is called a *M-space* and $M(\mathcal{E})$ *is* called *the M-space of* $\mathcal{E}$. Each bounded real variable on $\mathcal{E}$ defines an element of $M(\mathcal{E})$ and together these elements are dense in $M(\mathcal{E})$ in its $L(\mathcal{E})$-topology. The elements of $M(\mathcal{E})$ may be regarded as generalized bounded random variables. If the conclusion of the weak compactness lemma is valid for $\mathcal{E}$ then $M(\mathcal{E})$ is the space of (equivalence classes of) bounded random variables equipped with a modified supremum norm which takes into account that certain sets are $\mathcal{E}$-negligible.

Consider e.g. the situation where $\mathcal{E} = (P_\theta : \theta \in \Theta)$ is dominated by a convex combination $\pi = \sum_{i=1}^{\infty} c_i P_{\theta_i}$. Thus $c_1, c_2 \ldots \geq 0$ and $\sum_{i=1}^{\infty} c_i = 1$. Then $L(\mathcal{E})$ is just the space of finite π-absolutely continuous measures. By the Radon-Nikodym theorem this space may be identified with the L-space $L_1(\pi)$. Since the M-space $M(\mathcal{E})$ is the conjugate of $L(\mathcal{E})$, it may be identified with $L_\infty(\pi) = L_1(\pi)^*$. It follows that if we disregard π-null sets, then $M(\mathcal{E})$ is the space of bounded random variables equipped with the π-essential sup norm.

Just like sample spaces, decision spaces come with their measurable subsets. Mathematically, a decision space $(T, \mathcal{S})$ is just a measurable space. We shall find it convenient to write $\|f\|$ for the supremum norm $\sup_t |f(t)|$ of a real valued function f on T. Considering the finite decision spaces $T_k = \{1, \ldots, k\}$; $k = 1, 2, \ldots$, it is tacitly assumed that all subsets are measurable.

We shall admit as possible loss functions L on a decision space $(T, \mathcal{S})$ any family $L = (L_\theta : \theta \in \Theta)$ of real valued measurable functions on $(T, \mathcal{S})$. In order to ensure existence of expected loss we shall here assume that the functions $L_\theta : \theta \in \Theta$ are all bounded from below.

Within this set up, a decision rule in an experiment $\mathcal{E} = (\mathcal{X}, \mathcal{A}; P_\theta : \theta \in \Theta)$ is just a Markov kernel from the sample space $(\mathcal{X}, \mathcal{A})$ to the decision space $(T, \mathcal{S})$.

Since decision rules are Markov kernels, they transport distributions forwards and functions backwards. Thus if ρ is a decision rule from $\mathcal{E}$ to the decision space $(T, \mathcal{S})$ and μ is a finite measure on the sample space $(\mathcal{X}, \mathcal{A})$ of $\mathcal{E}$, then $\mu\rho$ is the measure on $\mathcal{S}$ assigning mass $\int \rho(S|\cdot) d\mu$ to a set S in $\mathcal{S}$. It is also convenient to have the notation $\mu \times \rho$ for the unique measure on $\mathcal{A} \times \mathcal{S}$ assigning mass $\int_A \rho(S|\cdot) d\mu$ to $A \times S$ when $A \in \mathcal{A}$ and $S \in \mathcal{S}$.

The decision rule ρ transports a bounded measurable function g on $(T, \mathcal{S})$ into the bounded measurable function $\rho g = \int g(t) \rho(dt|\cdot)$ on the sample space of $\mathcal{E}$.

Assume now that, in addition to the decision rule ρ, we are given both a finite measure μ on $(\mathcal{X}, \mathcal{A})$ and a bounded measurable function g on $(T, \mathcal{S})$. It is a fundamental fact that the three integrals $\int g d\mu\rho$, $\int (\rho g) d\mu$ and $\int g d(\mu \times \rho)$ are all equal and thus that we may write this quantity as $\mu\rho g$ without ambiguity.

As a function of the pair (μ, g), where $\mu \in L(\mathcal{E})$ and g is bounded measurable on $(T, \mathcal{S})$ the quantity $\mu\rho g$ is bilinear and this functional describes ρ up to equivalence.

Considering the map $\mu \to \mu\rho$ as a map from the L-space of finite measures on $\mathcal{A}$ to the L-space of finite measures on $\mathcal{S}$ we observe that it is linear, nonnegative (images of nonnegative elements are nonnegative) and preserves

total masses. A map from one L-space to another having these properties is called a *transition*. Thus the decision rule ρ defines a transition from the L-space $L(\mathcal{E})$ of $\mathcal{E}$ into the L-space of bounded additive set functions on $\mathcal{S}$.

Just as the concept of a bounded random variable, also the concept of a decision rule is too narrow for many purposes. We shall here admit any transition ρ from $L(\mathcal{E})$ to the L-space $ba(T,\mathcal{S})$ of bounded additive set functions as a *generalized decision rule*. As the class of decision rules (of the Markov kernel type) is dense within the class of generalized decision rules for pointwise convergence on $L(\mathcal{E}) \times ba(T,\mathcal{S})$ this is not a dramatic extension. Permitting generalized decision rules, we are however able to provide smoother statements, which otherwise would require cumbersome requlority conditions.

If the set-functions $P_\theta\rho : \theta \in \Theta$ are all σ-additive, if $(T,\mathcal{S})$ is Euclidean and if $\mathcal{E}$ is dominated, then the generalized decision rule ρ is definable in terms of a decision rule ρ from $\mathcal{E}$ to $(T,\mathcal{S})$. If ρ is a generalized decision rule and g is a bounded measurable function on $(T,\mathcal{S})$, then ρg is the image of g by the conjugate map ρ^*, which will also be denoted as ρ. If in addition $\mu \in L(\mathcal{E})$, then the fundamental identity $\int g d\mu\rho = \int (\rho g) d\mu$ may be expressed as $(\mu\rho, g) = (\rho, \mu g)$ and again this number is written $\mu\rho g$.

Whether ρ is a decision rule (of the Markov kernel variety) or a generalized decision rule, it defines a family $\mathcal{E}\rho = (P_\theta\rho : \theta \in \Theta)$ of probability set functions. This family is *the performance function of the decision rule* ρ. If ρ is an ordinary decision rule, then the performance function is an experiment and $P_\theta\rho$ is the distribution of the random decision when ρ is used and θ prevails. If a loss function $L = (L_\theta : \theta \in \Theta)$ on $(T,\mathcal{S})$ is given and if all the functions L_θ are bounded then *the risk function of the generalized decision rule* ρ is the function $\theta \to P_\theta\rho L_\theta$. In the general case *the risk function of* ρ may be defined as the function $\theta \to \lim_{N\to\infty} P_\theta\rho \min(L_\theta, N)$. If $P_\theta\rho$ is σ-additive or if L_θ is bounded, the latter quantity is the integral $\int L_\theta dP_\theta\rho$.

Many important functionals of experiments are expressible as integrals of homogeneous functions on the likelihood space. Those among them which appear in this paper may be described as follows:

Consider a family $\mu_\theta : \theta \in \Theta$ of finite (possibly signed) measures on a common measurable space $(\mathcal{X},\mathcal{A})$. Let h be a homogeneous measurable function on the product space $\mathbb{R}^\Theta$.

Then there is a countable subset Θ_0 of Θ such that $h(x)$ depend on x only via the restriction $x|\Theta_0$. Let σ be a nonnegative measure on $\mathcal{A}$ such that μ_θ has a density f_θ w.r.t. σ, when $\theta \in \Theta_0$. The crucial point to be noted is that neither the existence nor the value of the integral $\int h(d\mu_\theta/d\sigma : \theta \in \Theta) d\sigma$ depend on how σ otherwise is chosen. We may therefore supress σ in the notation of the

integral and, without ambiguity, we can write it as $\int h(d\mu_\theta : \theta \in \Theta)$, provided it exists.

As a particular case consider probability measures P and Q on a common measurable space. Then $\int |dP - dQ|$ is the statistical distance between P and Q while $\int \sqrt{dPdQ}$ is the affinity between P and Q

3. Comparisons of Experiments and Measure Families. Considering statistical experiments $\mathcal{E} = (P_\theta : \theta \in \Theta)$ and $\mathcal{F} = (Q_\theta : \theta \in \Theta)$ there are many possibilities for appealing definitions of $\mathcal{E}$ being at least as informative as $\mathcal{F}$. With most of these definitions, it is typically the case that two experiments are not comparable. One may however always ask for numbers quantifying how much we lose by relying on the experiment $\mathcal{E}$ rather than on the experiment $\mathcal{F}$. The scope of the early comparison theory was therfore vastly extended when LeCam (1964) provided a theory of approximate comparison. This theory is expressed in terms of a notion of ϵ-deficiency of one experiment w.r.t. another. Here $\epsilon = (\epsilon_\theta : \theta \in \Theta)$ is a family of nonnegative numbers.

Again there is a great variety of appealing, and apparently different, ways of expressing that one experiment is ϵ-deficient w.r.t. another. Thus one might e.g. base oneself on pointwise comparison of risk functions, on comparison of Bayes (or weighted) risks or on the natural extension of the operational definition of sufficiency. Perhaps the most important aspect of LeCam's theory is that his notion of ϵ-deficiency is very natural from any of these points of view.

Considering first a fixed finite decision space, this may be expressed as follows:

THEOREM 3.1. (Deficiency for k-decision problems). *Consider the set* $T_k = \{1, 2, \ldots, k\}$ *as a decision space.*

Let $\epsilon = (\epsilon_\theta : \theta \in \Theta)$ *be a nonnegative real valued function on the parameter set* Θ.

Then the following conditions are equivalent for experiments $\mathcal{E} = (\mathcal{X}, \mathcal{A} : P_\theta : \theta \in \Theta)$ *and* $\mathcal{F} = (\mathcal{Y}, \mathcal{B} : Q_\theta : \theta \in \Theta)$:

(i) *Pointwise comparison of risks:*
To each loss function L *(family* $L_\theta : \theta \in \Theta$ *of real valued functions on* T_k*) and each decision rule (Markov kernel)* σ *from* $\mathcal{F}$ *to* T *there corresponds a generalized decision rule (transition)* ρ *from* $\mathcal{E}$ *to* T *so that:*

$$P_\theta \rho L_\theta \leq Q_\theta \sigma L_\theta + \epsilon_\theta \|L_\theta\|; \quad \theta \in \Theta.$$

(ii) *Comparison of Bayes risks:*
To each finite subset Θ_0 *of* Θ *and to each loss function* L *(family* $L_\theta : \theta \in \Theta$ *of real valued functions on* T_k*) and each decision rule (Markov kernel)* σ *from* $\mathcal{F}$ *to* T *there corresponds a decision rule (Markov kernel)* ρ *from* $\mathcal{E}$ *to* T *so that:*

$$\sum_{\Theta_0} P_\theta \rho L_\theta \leq \sum_{\Theta_0} Q_\theta \sigma L_\theta + \sum_{\Theta_0} \epsilon_\theta \|L_\theta\|$$

(iii) *Comparison of maximum Bayes utilities. The sublinear function criterion:*
$\int \psi(dP_\theta : \theta \in \Theta_0) \geq \int \psi(dQ_\theta : \theta \in \Theta_0) - \sum_{\Theta_0} \epsilon_\theta[\psi(-e^\theta) \vee \psi(e^\theta)]$ *for each finite subset* Θ_0 *of* Θ *and for each function* ψ *on* $\mathbb{R}^{\Theta_0}$ *which is a maximum of* k *linear functionals.*
Here $e^\theta = (0, \ldots, \overset{\theta}{1}, \ldots, 0)$ *denotes the* θ*th unit vector in* $\mathbb{R}^{\Theta_0}$.

(iv) *Comparison of performance functions:*
To each decision rule (Markov kernel) σ *in* $\mathcal{F}$ *corresponds a generalized decision rule* ρ *in* $\mathcal{E}$ *so that:*

$$\|P_\theta \rho - Q_\theta \sigma\| \leq \epsilon_\theta; \quad \theta \in \Theta.$$

PROOF. The implications (iv)⇒(i)⇒(ii) are all more or less immediate. Replacing the loss function L with the utility function $U = -L$ the inequality of (ii) may be written:

$$\sum_{\Theta_0} P_\theta \rho U_\theta \geq \sum_{\Theta_0} Q_\theta \sigma U_\theta - \sum_{\Theta_0} \epsilon_\theta \|U_\theta\|.$$

Maximizing first w.r.t. ρ and then w.r.t. σ it may be seen that (iii) is essentially a reformulation of (ii). The implication (ii)⇒(iv) follows by standard minimax theory (see e.g. Torgersen (1970)). ∎

The theorem is stated in order to make the generalization to general mass distributions more or less obvious. Knowing however that the distributions P_θ and Q_θ have the same total masses the deficiency term $\sum_{\Theta_0} \epsilon_\theta[\psi(-e_\theta) \vee \psi(e_\theta)]$ in (iii) may be replaced with the linear (in ψ) term $\frac{1}{2}\sum_{\Theta_0} \epsilon_\theta[\psi(-e_\theta) + \psi(e_\theta)]$. Actually we may in this case restrict attention to functions ψ such that $\psi(-e_\theta) \equiv_\theta \psi(e_\theta)$ and then both expressions reduce to $\sum_{\Theta_0} \epsilon_\theta \psi(e_\theta)$. This is also so for general measures provided we supplement (iii) with the condition that $|P_\theta(\mathcal{X}) - Q_\theta(\mathcal{Y})| \leq \epsilon_\theta : \theta \in \Theta$. This amounts to require that the inequality in (iii) holds for linear functions ψ.

If the equivalent conditions of the theorem are fulfilled, then we shall say that $\mathcal{E}$ *is* ϵ*-deficient w.r.t.* $\mathcal{F}$ *for* k*-decision problems.* If $\mathcal{E}$ is ϵ-deficient w.r.t. $\mathcal{F}$ for 2-decision problems, then we shall also say that $\mathcal{E}$ *is* ϵ*-deficient w.r.t.* $\mathcal{F}$ *for testing problems.* If $\mathcal{E}$ is ϵ-deficient w.r.t. $\mathcal{F}$ for k-decision problems for all $k = 1, 2, 3, \ldots$, then we shall just say that $\mathcal{E}$ *is* ϵ*-deficient w.r.t.* $\mathcal{F}$.

The smallest constant $\epsilon \geq 0$ (it exists) such that $\mathcal{E}$ is ϵ-deficient w.r.t. $\mathcal{F}$ (for k-decision problems) is *the deficiency of* $\mathcal{E}$ *w.r.t.* $\mathcal{F}$ *(for* k*-decision problems).* This number is denoted as $\delta_{(k)}(\mathcal{E}, \mathcal{F})$. The *deficiency distance between* $\mathcal{E}$ *and* $\mathcal{F}$ *(for* k*-decision problems)* is the quantity $\Delta_{(k)}(\mathcal{E}, \mathcal{F}) = \max\{\delta_{(k)}(\mathcal{E}, \mathcal{F}), \delta_{(k)}(\mathcal{F}, \mathcal{E})\}$.

The deficiency distances $\Delta_1 = 0, \Delta_2, \Delta_3, \ldots$ and Δ are all pseudometrics. The nontrivial ones, $\Delta_2, \Delta_3, \ldots$ and Δ, all determine the same notion of equivalence. In other words the requirements $\Delta_2(\mathcal{E}, \mathcal{F}) = 0, \Delta_3(\mathcal{E}, \mathcal{F}) = 0, \ldots$ and $\Delta(\mathcal{E}, \mathcal{F}) = 0$ are the same for experiments $\mathcal{E}$ and $\mathcal{F}$. Experiments $\mathcal{E}$ and $\mathcal{F}$ such that $\Delta(\mathcal{E}, \mathcal{F}) = 0$ are called equivalent and we may express this by writing $\mathcal{E} \sim \mathcal{F}$.

The collection of experiments which are equivalent with a given experiment $\mathcal{E}$ is called *the type of* $\mathcal{E}$. Although types are not well defined sets it may be argued that there is a well defined set containing all types of experiments having the same parameter set Θ.

If $\delta_{(k)}(\mathcal{E}, \mathcal{F}) = 0$ then we shall say that $\mathcal{E}$ *is at least as informative as* $\mathcal{F}$ *(for* k*-decision problems)* and write this $\mathcal{E} \geq_{(k)} \mathcal{F}$. The relations $\geq_1, \geq_2, \geq_3, \ldots$ and $\geq$ are all partial orderings. Here is LeCam's fundamental randomization criterion for ϵ-deficiency:

THEOREM 3.2. (The randomization criterion for ϵ-deficiency). *The experiment* $\mathcal{E} = (P_\theta : \theta \in \Theta)$ *is* ϵ*-deficient w.r.t. the experiment* $\mathcal{F} = (Q_\theta : \theta \in \Theta)$ *if and only if* $\|P_\theta M - Q_\theta\| \leq \epsilon_\theta$; $\theta \in \Theta$ *for a transition* M *from* $L(\mathcal{E})$ *to* $L(\mathcal{F})$.

REMARK. If, say, $\mathcal{E}$ admits the conclusion of the weak compactness lemma and if the sample space of $\mathcal{F}$ is Euclidean, then M may be chosen as a Markov kernel.

PROOF OF THEOREM 3.2. The "if" is immediate. Using the previous theorem the "only if" may be argued by approximation from the "finite" case, as in Torgersen (1985). ∎

This is not the place to dwell on the many statistical interesting implications of the above theorems. We shall instead proceed by observing that the theorems remain valid if we permit the distributions P_θ and Q_θ to be finite (possibly signed) measures.

It is then however not clear what meaning should be attached to statistical concepts as sample space, decision space, loss function, decision rule, performance function and risk. Thus conditions (i)–(iv) of Theorem 3.1 should be read without the headings but with the parenthetical insertions.

The concept of a statistical experiment $\mathcal{E} = (P_\theta : \theta \in \Theta)$ has to be replaced by the concept of a *measure family* consisting of a family $(P_\theta : \theta \in \Theta)$ of finite measures on a common measurable space. Doing that, the other comments and definitions remain valid except that the equivalence induced by the deficiency distance Δ_2 is no longer the same as the equivalences induced by the deficiencies $\Delta_3, \Delta_4, \ldots$ and Δ. The latter are however still the same. Of course it may not make much sense to interpret the orderings $\geq_1, \geq_2, \ldots$ and $\geq$ as information orderings.

We summarize these observations with the following theorem.

THEOREM 3.3. (Comparison of measure families). *Theorems 3.1–3.2 remain valid for general measure families* $\mathcal{E} = (P_\theta : \theta \in \Theta)$ *and* $\mathcal{F} = (Q_\theta : \theta \in \Theta)$ *provided they are read as explained above.*

The sublinear function criterion, condition (iii) of Theorem 3.1, may be linearized by adding the requirement that $\epsilon_\theta \geq |P_\theta(\mathcal{X}) - Q_\theta(\mathcal{Y})|$; $\theta \in \Theta$ and then replacing the deficiency term $\sum_\theta \epsilon_\theta[\psi(-e_\theta) \vee \psi(e_\theta)]$ by

$$\frac{1}{2}\sum_\theta \epsilon_\theta[\psi(-e_\theta) + \psi(e_\theta)] + \frac{1}{2}\sum_\theta [Q_\theta(\mathcal{Y}) - P_\theta(\mathcal{X})][\psi(e_\theta) - \psi(-e_\theta)].$$

As remarked before, we may then even restrict attention to sublinear functions ψ such that $\psi(-e_\theta) \equiv_\theta \psi(e_\theta)$ and then the deficiency term in both cases reduces to $\sum_\theta \epsilon_\theta \psi(e_\theta)$.

Let us conclude this section with some remarks on functionals of experiments having a common finite parameter set Θ.

We observed at the end of the previous section how we might construct an integral $\overline{h}(\mathcal{E}) = \int h(dP_\theta : \theta \in \Theta)$ for an experiment $\mathcal{E} = (P_\theta : \theta \in \Theta)$ and for a homogeneous measurable function on $[0, \infty[^\Theta$. If h is nonnegative or if h is bounded on bounded sets then $\overline{h}(\mathcal{E})$ is defined this way for all experiments $\mathcal{E}$.

In both cases the functional $\mathcal{E} \to \overline{h}(\mathcal{E})$ behave as an affine function for the operation of mixing experiments according to known mixing distributions.

If, in addition, h is continuous, then $\overline{h}$ is continuous in the topology of the deficiency distrance Δ. Indeed, by Torgersen (1991), any affine continuous functional of experiments is of the form $\mathcal{E} \to \overline{h}(\mathcal{E})$ for a continous homogeneous function h on $[0, \infty[^\Theta$. If, furthermore, h is sublinear on $\mathbb{R}^\Theta$, then by

Theorem 3.2, this functional is monotonically increasing. Conversely any continuous affine monotonically increasing functional is of this form for a sublinear function h on $\mathbb{R}^{\Theta}$.

EXAMPLE 3.4. (Mulivariate Gini index). Consider measure families $\mathcal{E} = (\mu_\theta : \theta \in \Theta)$ having a common finite parameter set Θ.

An interesting set valued functional of measure families is the functional which to a given measure family assigns the convex hull $\tau(\mathcal{E})$ of the range of the vector valued (i.e. $\mathbb{R}^{\Theta}$ valued) measure $\mathcal{E} = (\mu_\theta : \theta \in \Theta)$. This set defines $\mathcal{E}$ up to Δ_2 equivalence and, for experiments, Δ_2 equivalence is the same as full equivalence i.e. Δ-equivalence. It follows that any functional of experiments which respects equivalence, i.e. is defined for types, is a functional of these sets.

For a given experiment $\mathcal{E}$, the set $\tau(\mathcal{E})$ consists precisely of all available power functions of tests in $\mathcal{E}$. It follows readily that τ is monotonically increasing for the information ordering and also that it behaves affinely under mixing. If Θ contains at most two points then τ actually determines the deficiency ordering i.e. then $\mathcal{E} \geq \mathcal{F}$ if and only if $\tau(\mathcal{E}) \supseteq \tau(\mathcal{F})$. If $\#\Theta \geq 3$, then, however, the set inequality $\tau(\mathcal{E}) \supseteq \tau(\mathcal{F})$ does not imply that $\mathcal{E}$ and $\mathcal{F}$ are comparable.

A particularily interesting real valued functional is the functional G which to a measure family $\mathcal{E}$ assigns the volume $G(\mathcal{E})$ of $\tau(\mathcal{E})$. The functional $\mathcal{E} \to G(\mathcal{E}) = \text{Volume } \tau(\mathcal{E})$ may appropriately be considered as a multivariate (at least if $\#\Theta \geq 3$) Gini index.

This index provides a monotonically information increasing functional of experiments. Furthermore, by the Brunn-Minkowski theorem, its nth root is concave for mixing.

For a given measure family $\mathcal{E} = (\mu_1, \ldots, \mu_n)$ the multivariate Gini index may be expressed as:

$$G(\mathcal{E}) = \text{Volume } \tau(\mathcal{E}) = \frac{1}{n!} \left\| \sum_{\pi} (\operatorname{sgn} \pi)(\mu_{\pi(1)} \times \ldots \times \mu_{\pi(n)}) \right\|,$$

where $\| \cdot \|$ denotes total variation and where π runs through the set of all permutations of $\{1, \ldots, n\}$.

The structure of this formula may become more apparent if we write it

as:

$$G(\mathcal{E}) = \frac{1}{n!}\left\| \det \begin{pmatrix} \mu_1, \ldots, \mu_n \\ \ldots\ldots\ldots \\ \mu_1, \ldots, \mu_n \end{pmatrix} \right\|.$$

If $\mu_1, \mu_2, \ldots, \mu_n$ are dominated by the nonnegative σ-finite measure σ and if $f_i = d\mu_i/d\sigma; i = 1, \ldots, n$ then $G(\mathcal{E})$ may be computed as the integral:

$$G(\mathcal{E}) = \int \frac{1}{n!}\left| \det \begin{pmatrix} f_1(x_1), \ldots, f_n(x_1) \\ \ldots\ldots\ldots\ldots\ldots \\ f_1(x_n), \ldots, f_n(x_n) \end{pmatrix} \right| \sigma^n(d(x_1, \ldots, x_n)).$$

The formula:

$$\text{Volume } \tau(\mathcal{E}) = \frac{1}{n!}\left\| \sum_{\pi} (\text{sgn } \pi)\mu_{\pi(1)} \times \ldots \times \mu_{\pi(n)} \right\|$$

may be established by verifying that:

(i) If $T = (t_{i,k}; i, k = 1, \ldots, n)$ is a nonsingular $n \times n$ matrix and if we replace $\mu_1, \ldots, \mu_n$ with $\nu_1 = \sum_k t_{1k}\mu_k, \ldots, \nu_n = \sum_k t_{nk}\mu_k$, respectively, then both sides of the equality are multiplied by $|\det T|$.

(ii) If $\mathcal{X} = \{1, \cdots, r\}$ and $\mu_i(j) = a_{ij}$; $i = 1, \ldots, n, j = 1, \ldots, r$ then the identity reduces to:

(*) $\sum_{1 \le j_1 < \ldots < j_n \le r} |\det(a_{\cdot,j_1}, \ldots, a_{\cdot,j_n})| = \text{Volume}(\langle 0, a_{\cdot,1}\rangle + \ldots + \langle 0, a_{\cdot,r}\rangle)$ where $\langle\ \rangle$ denotes convex hull.

If the vectors $a_{\cdot,1}, \ldots, a_{\cdot,n}$ are linearily dependent then both sides of (*) are zero.

If $r = n$ and $a_{\cdot,1}, \ldots, a_{\cdot\ldots,n}$ are linearily independent then, by (i), (*) may be reduced to the statement that the volume of a cube is the product of the lengths of its sides.

The validity of (*) follows now by induction on r. (Using (i) we may assume that $a_{i,r} = 0$ or $= s \ge 0$ as $i < n$ or $i = n$.)

(iii) Both sides of the desired equality are continuous for weak convergence of standard measures. (These measures are defined in the next section.) It suffices therefore, since the set of finitely supported standard measures is dense, to consider the finite case.

4. Comparison in Terms of Densities. Dilations. By the randomization criterion, Theorem 3.3, the measure family $\mathcal{E} = (\mathcal{X}, \mathcal{A}; \mu_\theta : \theta \in \Theta)$

is ϵ-deficient w.r.t. the measure family $\mathcal{F} = (\mathcal{Y}, \mathcal{B}; \nu_\theta : \theta \in \Theta)$ if and only if $\|\mu_\theta M - \nu_\theta\| \leq \epsilon_\theta; \theta \in \Theta$ for some transition M from $L(\mathcal{E})$ to $L(\mathcal{F})$.

Before applying this, note that most of the measure families encountered in Section 1 admitted a particular parameter value $\theta = \theta_0$ such that the distributions for this parameter value was uniform. Furthermore the concepts of approximate majorization required that approximation should be exact when this parameter value prevailed. Within the context of Section 1 this amounted to the condition that certain Markov matrices were doubly stochastic.

Generalizing this, let us assume that there is a distinguished parameter value $\theta = \theta_0$ such that the measures μ_{θ_0} and ν_{θ_0} are nonnegative and dominate, respectively, $\mathcal{E}$ and $\mathcal{F}$. Assume also that $\epsilon_{\theta_0} = 0$. Then $\mathcal{E}$ is ϵ-deficient w.r.t. $\mathcal{F}$ if and only if $\nu_{\theta_0} = \mu_{\theta_0} M$ for a transition M such that $\|\mu_\theta M - \nu_\theta\| \leq \epsilon_\theta; \theta \in \Theta$.

In order to escape difficult technical problems, let us assume that the underlying measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ are both Euclidean. Then the transition M may be represented by a Markov kernel which, by abuse of notation, also will be denoted by M. The joint distribution $\mu_{\theta_0} \times M$ on $\mathcal{A} \times \mathcal{B}$ may be factorized as $\mu_{\theta_0} \times M = D \times \nu_{\theta_0}$ for a Markov kernel D from $\mathcal{F}$ to $\mathcal{E}$. This implies in particular that $\mu_{\theta_0} = D\nu_{\theta_0}$.

The basic property of the kernel D is that it, for each θ, within an error of at most ϵ_θ maps the density $f_\theta = d\mu_\theta / d\mu_{\theta_0}$ in $\mathcal{E}$ into the corresponding density $g_\theta = d\nu_\theta / d\nu_{\theta_0}$ in $\mathcal{F}$. Indeed $d\mu_\theta M / d\nu_{\theta_0} = \int f_\theta(x) D(dx|\cdot)$ and thus $\|\mu_\theta M - \nu_\theta\| = \int |\int f_\theta(x) D(dx|y) - g_\theta(y)| \nu_{\theta_0}(dy)$. Hence, by ϵ-deficiency

$$\int \left| \int f_\theta(x) D(dx|y) - g_\theta(y) \right| \nu_{\theta_0}(dy) \leq \epsilon_\theta : \theta \in \Theta .$$

Regarding the likelihood functions $f.(x)$ and $g.(y)$ as markings of positions of the points x and y, we may interpret the last set of inequalities as saying that the Markov kernel D behaves ϵ-approximately as a dilation.

If, conversely, the Markov kernel D has these properties then the last equality holds for any Markov kernel M from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$ such that $D \times \nu_{\theta_0} = \mu_{\theta_0} \times M$. This proves the following theorem.

THEOREM 4.1. (Deficiencies, ϵ-dilating kernels and densities). *Assume that the nonnegative finite measures μ_{θ_0} and ν_{θ_0} dominate the measure familiy $\mathcal{E} = (\mathcal{X}, \mathcal{A}; \mu_\theta : \theta \in \Theta)$ and the measure family $\mathcal{F} = (\mathcal{Y}, \mathcal{B}; \nu_\theta : \theta \in \Theta)$, respectively.*

Assume also that the underlying measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ are both Euclidean. Let $\epsilon_\theta : \theta \in \Theta$ be non negative numbers such that $\epsilon_{\theta_0} = 0$. Put, for each θ, $f_\theta = d\mu_\theta / d\mu_{\theta_0}$ and $g_\theta = d\nu_\theta / d\nu_{\theta_0}$.

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 such that:

$$\int \left| \int f_\theta(x) D(dx|y) - g_\theta(y) | \nu_{\theta_0}(dy) \leq \epsilon_\theta; \theta \in \Theta.$$

The theorem furnishes alternative proof of several well-known existence theorems. Perhaps the most famous among them is the dilation criterion for probability distributions on $\mathbb{R}^k$ of Blackwell (1953) and LeCam (1964). In Strassen (1965) this is generalized to probability distributions on convex compact metrizable subsets of of locally convex topological vector spaces.

If P and Q are probability vectors on the same finite dimensional linear space V and if both their expectation vectors exist in V then a slight generalization of this criterion states that $P = DQ$ for a dilation D from V to V if and only if $\int \varphi dP \geq \int \varphi dQ$, when φ is real valued and convex on V.

Here a dilation from V to V is a Markov kernel D such that, for each $y \in V$, the expectation vector of $M(\cdot|y)$ is y. In order to see how this result fits into the framework used here we shall incorporate it into the following theorem.

THEOREM 4.2. (Dilations of distributions on finite dimensional spaces). *Let P and Q be probability distributions on the same finite dimensional linear space V.*

Let also $v_1', \ldots, v_k'$ be a basis for the algebraic dual V' of V. For $i = 1, \ldots, k$ denote by P_i and Q_i the distribution on V which have density v_i' w.r.t. P and Q, respectively.

Let Φ be the set of $P+Q$ integrable convex functions and for each $\varphi \in \Phi$, let the quantity $\int \varphi dP - \int \varphi dQ$ be denoted as ϵ_φ.

Then the following conditions are equivalent:

(i) $(P, P_1, \ldots, P_k) \geq (Q, Q_1, \ldots, Q_k)$.

(ii) $\int \varphi dP \geq \int \varphi dQ$; $\varphi \in \Phi$.

(iii) *P is a dilation of Q.*

(iv) *The measure family $(\varphi P : \varphi \in \Phi)$ is $(\epsilon_\varphi : \varphi \in \Phi)$ deficient w.r.t. the measure family $(\varphi Q : \varphi \in \Phi)$.*

REMARK 1. The distributions $P_1, \cdots, P_k$ and $Q_1, \cdots, Q_k$ are not necessarily probability distributions. They may be nonnegative, non positive or neither. In any case "$\geq$" in (i) is in the sense of the left hand side being 0-deficient w.r.t the right hand side.

REMARK 2. If μ is a measure and h is a measurable function, then $h\mu$ denotes the measure (if it exists) having density h w.r.t. μ.

PROOF. By Theorem 3.3 condition (i) amounts to the condition that $\int \psi(1, v_1', \ldots, v_k')dP \geq \int \psi(1, v_1', \ldots, v_k')dQ$, when ψ is sublinear on $\mathbb{R}^{k+1}$. Putting $\varphi(x) = \psi(1, v_1'(x), \ldots, v_k'(x))$ when $x \in V$, this inequality may also be written $\int \varphi dP \geq \int \varphi dQ$. As φ is convex, it is clear that (ii)$\Rightarrow$(i). The converse implication is a consequence of the fact that a convex function φ on V which is a maximum of a finite set of affine functionals is of the form $x \to \psi(1, v_1'(x), \ldots, v_k'(x))$ for a sublinear function ψ on $\mathbb{R}^{k+1}$.

On the other hand condition (i) is, by Theorem 4.1, equivalent to the condition that $P = DQ$ for a Markov kernel D such that $\int v_i'(x)D(dx|y) = v_i'(y)$; $i = 1, \ldots, k$ when $y \in V$. As any linear functional on V is a linear combination of $v_1', \ldots, v_k'$, the last requirement on D expresses that D is a dilation. Thus also conditions (i) and (iii) are equivalent. Furthermore the very statement of condition (iv) implies that the quantities $\epsilon_\varphi : \varphi \in \Phi$ are nonnegative i.e. that (ii) holds.

Assume finally that conditions (i)–(iii) are satisfied. Let $\varphi_1, \ldots, \varphi_s \in \Phi$ and let ψ be a maximum of a finite set of nonnegative linear functionals on $\mathbb{R}^s$. Then $\psi(\varphi_1, \ldots, \varphi_s) \in \Phi$ so that

$$\int \psi(d(\varphi P), \ldots, d(\varphi_s P)) = \int \psi(\varphi_1, \ldots, \varphi_s)dP \geq$$
$$\int \psi(\varphi_1, \ldots, \varphi_s)dQ = \int \psi(d(\varphi_1 Q), \ldots, d(\varphi_s Q)).$$

Consider any maximum ψ of a finite set of linear functionals on $\mathbb{R}^s$. Putting $\tilde{\psi}(z) = \psi(z) + \sum_{i=1}^s \psi(0, \ldots, \overset{(i)}{-} 1, \ldots, 0)z_i$ when $z = (z_1, \ldots, z_s) \in \mathbb{R}^k$, we see that $\tilde{\psi}$ satisfies the above requirements. Furthermore $\int \tilde{\psi}(d(\varphi_1 P), \ldots, d(\varphi_s P)) =$ $\int \psi(\varphi_1, \ldots, \varphi_s)dP + \sum_{i=1}^s \psi(0, \ldots, \overset{(i)}{-} 1, \ldots, 0) \int \varphi_i dP$, where P throughout may be replaced by Q. Thus

$$\int \psi(d(\varphi_1 P), \ldots, d(\varphi_s P)) \geq \int \psi(d(\varphi_1 Q), \ldots, d(\varphi_s Q))$$
$$- \sum_{i=1}^s \psi(0, \ldots, \overset{i}{-} 1, \ldots, 0)\epsilon_{\varphi_i}.$$

Condition (iv) follows now by Theorem 3.3. ∎

A measure family $\mathcal{E} = (\mu_\theta : \theta \in \Theta)$ is $\epsilon = (\epsilon_\theta : \theta \in \Theta)$ deficient w.r.t. a measure family $\mathcal{F} = (\nu_\theta : \theta \in \Theta)$ if and only if $(\mu_\theta : \theta \in \Theta_0)$ is $(\epsilon_\theta : \theta \in \Theta_0)$ deficient w.r.t. $(\nu_\theta : \theta \in \Theta_0)$ for every finite subset Θ_0 of Θ. Thus

comparison problems may be reduced to the case of measure families having finite parameter sets. Very useful tools in the finite case are the concepts of *standard measure families* and of *standard measures.*

For a measure family $\mathcal{E} = (\mathcal{X}, \mathcal{A}; \mu_\theta; \theta \in \Theta)$ having a finite parameter set Θ these concepts are defined as follows.

Let f_θ, for each $\theta \in \Theta$, be a version of the density of μ_θ w.r.t. $\mu = \sum_\theta |\mu_\theta|$. Then $\sum_\theta |f_\theta| = 1$ a.e.μ. For each point $x \in \mathcal{X}$, let $()$ denote the vector $(f_\theta(x) : \theta \in \Theta)$ in $\mathbb{R}^\Theta$. Thus $f = (f_\theta : \theta \in \Theta)$ is a measurable map from $(\mathcal{X}, \mathcal{A})$ to $\mathbb{R}^\Theta$.

With these notations *the standard measure family of* $\mathcal{E}$ is the measure family $(\mu_\theta f^{-1} : \theta \in \Theta)$ on $\mathbb{R}^\Theta$. The measure $S_\mathcal{E} = (\sum_\theta |\mu_\theta|) f^{-1} = \sum_\theta |\mu_\theta f^{-1}|$ is *the standard measure* of $\mathcal{E}$.

By sufficiency, any measure family is equivalent to its standard measure family $(S_\theta : \theta \in \Theta)$. This measure family is uniquely determined by the standard measure S since the projection onto the θth coordinate space is the density of S_θ w.r.t. S. For statistical experiments Theorem 4.2 yields the following dilation criterion of Blackwell (1953) and LeCam (1964):

COROLLARY 4.3. (The dilation criterion for "being at least as informative as"). *Let* $\mathcal{E} = (P_\theta : \theta \in \Theta)$ *and* $\mathcal{F} = (Q_\theta : \theta \in \Theta)$ *be experiments having standard measures* S *and* T *on* $\mathbb{R}^\Theta$, *respectively.*

Then $\mathcal{E} \geq \mathcal{F}$ *if and only if* S *is a dilation of* T.

Here a dilation D *on* $\mathbb{R}^\Theta$ *is a Markov kernel* D *from* $\mathbb{R}^\Theta$ *to* $\mathbb{R}^\Theta$ *such that* $\int xD(dx|y) = y$ *for all points* $y \in \mathbb{R}^\Theta$.

Another application of Theorem 4.2 is to the theory of local comparison of experiments. In order to indicate this, consider experiments whose common parameter set Θ is a subset of $\mathbb{R}^k$. If θ_0 is an interior point of Θ and if the map $\theta \to P_\theta$ from Θ to $L(\mathcal{E})$ is Fréchet differentiable at $\theta = \theta_0$ then we shall simply say that $\mathcal{E}$ *is differentiable at* $\theta = \theta_0$. If this is the case, then the partial derivatives $\dot{P}_{\theta_0,i} = [\partial P_\theta / \partial \theta_i]_{\theta=\theta_0}$; $i = 1, \ldots, k$ are well defined. It turns out then that the local behaviour of $\mathcal{E}$ around $\theta = \theta_0$ is to a first approximation completely described by the measure family $(P_{\theta_0}, \dot{P}_{\theta_0,1}, \ldots, \dot{P}_{\theta_0,k})$. The characteristic property of this measure family is that the "first" member P_{θ_0} is a probability distribution while the other distributions have total mass zero and are dominated by P_{θ_0}.

If $F(\cdot|\theta_0, \mathcal{E})$ is the joint distribution under $\theta = \theta_0$ of the differentiated likelihood ratios $d\dot{P}_{\theta_0,i}/dP_{\theta_0}$; $i = 1, \ldots, k$, then $\int \psi(dP_{\theta_0}, d\dot{P}_{\theta_0,1}, \ldots, d\dot{P}_{\theta_0,k}) = \int \psi(1, x)\, F(dx|\theta_0, \mathcal{E})$ for each sublinear function ψ on $\mathbb{R}^{k+1}$. Thus first order local behaviour is determined by the distribution function $F(\cdot|\theta_0, \mathcal{E})$.

If $\mathcal{F} = (Q_\theta : \theta \in \Theta)$ is also differentiable at $\theta = \theta_0$ then we may compare the local behavior of risk functions around $\theta = \theta_0$. Then locally, around $\theta = \theta_0$, $\mathcal{E}$ is at least as informative as $\mathcal{F}$ if and only if the measure family $(P_{\theta_0}, \dot{P}_{\theta_0,1}, \ldots, \dot{P}_{\theta_0,k})$ is 0-deficient w.r.t. the measure family $(Q_{\theta_0}, Q_{\theta_0,1}, \ldots, \dot{Q}_{\theta_0,k})$ (Torgersen (1985) and (1991)).

By Theorem 4.2 this amounts to the condition that $F(\cdot|\theta_0, \mathcal{E})$ is a dilation of $F(\cdot|\theta_0, \mathcal{F})$.

Condition (iv) of Theorem 4.2 permits interesting generalizations and variations. An immediate generalization is obtained by replacing the probability distributions P and Q by nonnegative finite measures μ and ν on a measurable space $(\mathcal{X}, \mathcal{A})$ and by replacing the class Φ of convex functions by a convex set H of $\mu + \nu$ integrable functions. Looking over the proof of the theorem we see that we needed some additional structure of Φ. We shall here assume that H shares with Φ the properties that it contains the null function and that $h_1 \vee h_2 \in H$ whenever $h_1 \in H$ and $h_2 \in H$. Under these conditions, we may derive a characterization in terms of transitions of the situation where $\int h d\mu \geq \int h d\nu$ when $h \in H$. Indeed, if this is the case, and if $\epsilon_h = \int h d\mu - \int h d\nu$ when $h \in H$, then the measure family $(h\mu : h \in H)$ is $(\epsilon_h : h \in H)$ deficient w.r.t. $(h\nu : h \in H)$.

In order to see this consider functions $h_1, \ldots, h_s$ in H along with a sublinear function ψ on $\mathbb{R}^s$ which is a maximum of a finite set of nonnegative linear functionals. If $z \to \sum_{1=1}^s a_i z_i$ is one of these functionals then $a_1, \ldots, a_s \geq 0$ and thus, by convexity, $\frac{1}{N}\sum_{1=1}^s a_i h_i = (1 - \frac{1}{N}\sum_{i=1}^s a_i)0 + \sum_{i=1}^s (a_i/N)h_i \in H$ when N is suficiently large. It follows that also $\frac{1}{N}\psi(h_1, \ldots, h_s) \in H$, when N is sufficiently large and thus $\int \psi(h_1, \ldots, h_s)d\mu = N\int[\psi(h_1, \ldots, h_s)/N]d\mu \geq N\int[\psi(h_1, \ldots, h_s)/N]d\nu = \int \psi(h_1, \ldots, h_s)d\nu$.

As in the proof of Theorem 4.2, we derive from this the asserted statement on deficiencies.

By the randomization criterion this amounts to the conditions that $\|(h\mu)M - h\nu\| \leq \int h d\mu - \int h d\nu$; $h \in H$ for a transition M from $L_1(\mu)$ to $L_1(\nu)$. Now the total variation $\|\sigma\|$ of any finite measure σ may be expressed as $\|\sigma\| = \|\sigma^+\| + \|\sigma^-\| = \|\sigma^+\| - \|\sigma^-\| + 2\|\sigma^-\| = 2\|\sigma^-\| + \int 1 d\sigma$. Applying this to the measures $(h\mu)M - h\nu$ and utilizing the equality $\int 1 d[(h\mu)M - h\nu] = \int 1 d[(h\mu)M] - \int 1 d(h\nu) = \int 1 d(h\mu) - \int 1 d(h\nu) = \int h d\mu - \int h d\nu$, we find that the last inequality may be written $[(h\mu)M - (h\nu)]^- = 0$ i.e. that $(h\mu)M \geq h\nu$; $h \in H$. The last "$\geq$" then simply indicates that the measures $(h\mu)M - h\nu$, $h \in H$ are all nonnegative.

Assume now that $1 \in H$ and that the measures μ and ν have mass i.e. that $\|\mu\| = \|\nu\|$. Inserting $h = 1$ above, we find that $\mu M \geq \nu$ and hence, since $\|\mu M\| = \|\mu\| = \|\nu\|$, $\mu M = \nu$. In the Euclidean case this

yields the factorization $\mu \times M = D \times \nu$ for a Markov kernel D. Now the density of $(h\mu)M$ w.r.t. $\nu = \mu M$ may be specified as $\int h(x)D(dx|\cdot)$ and thus the above requirement in terms of densities is expressed by the inequalities: $\int h(x)D(dx|y) \geq h(y)$ for ν almost all y, whenever $h \in H$.

We summarize these considerations in the following theorem.

THEOREM 4.4. (Transition criteria for the ordering of measures by integrals of given functions). *Let H be a convex family of real valued measurable functions on a measurable space $(\mathcal{X}, \mathcal{A})$. Assume that $0 \in H$ and that $h_1 \vee h_2 \in H$ when $h_1, h_2 \in H$.*

Let μ and ν be nonnegative finite measures on $\mathcal{A}$ such that each function $h \in H$ is $\mu+\nu$ integrable. Put $\epsilon_h = \int h d\mu - \int h d\nu$; $h \in H$. Then the following conditions are equivalent:

(i) $\epsilon_h \geq 0$; $h \in H$.

(ii) $\epsilon_h \geq 0$, $h \in H$ *and the measure family* $(h\mu : h \in H)$ *is* $(\epsilon_h : h \in H)$ *deficient w.r.t. the measure family* $(h\nu : h \in H)$.

(iii) There is a transition M from $L_1(\mu)$ to $L_1(\nu)$ such that all measures $(h\mu)M - h\nu$; $h \in H$ are nonnegative.

If $(\mathcal{X}, \mathcal{A})$ is Euclidean, if $\|\mu\| = \|\nu\|$ and if $1 \in H$ then these conditions are equivalent to:

(iv) $\mu = D\nu$ for a Markov kernel D such that $\int h(x)D(dx|y) \geq h(y)$ for ν almost all y, whenever $h \in H$.

REMARK 1. If H is sufficiently separable then the exceptional set in (iv) may be chosen independently of H and then D may be modified in such a way that $\int h(x)D(dx|y) \geq h(y)$ for all $y \in \mathcal{X}$ and for all $h \in H$.

REMARK 2. Assuming that H is a set of continuous functions on a compact metric space and that the constant functions are in H the theorem may be modified in order to provide criteria for the condition that $\int h d\mu \geq \int h d\nu$ for all nonnegative functions h in H (Torgersen (1985)). Besides Karlin (1983) this result was inspired by Fisher and Holbrook (1980).

EXAMPLE 4.5. (Stochastic ordering of distributions). Assume that P and Q are probability measures on a partially ordered measurable space $(\mathcal{X}, \mathcal{A})$. By partial integration $P(A) \geq Q(A)$ for every monotonically increasing measurable set if and only if $\int h dP \geq \int h dQ$ for each function h belonging to the class H of monotonically increasing functions h which are $P + Q$ integrable. By the theorem this is the case if and only if $(hP)M \geq hQ$ for a transition M. As $1 \in H$ it is clear that $PM = Q$.

If $(\mathcal{X}, \mathcal{A})$ is Euclidean this shows that $P = DQ$ for a Markov kernel D such that $\int h(x)D(dx|\cdot) \geq h$ a.s. Q whenever h is measurable bounded and monotonically increasing. In particular $D(A|\cdot) \geq I_A$ a.e. Q when A is monotonically increasing and measurable. Assume now also that there is a countable class $\mathcal{A}_0$ of increasing measurable sets such that for any point $a \in \mathcal{X}$ the set $\{x : x \geq a\}$ is the intersection of a decreasing sequence of sets from $\mathcal{A}_0$. Then we may modify D so that $D(A|\cdot) \geq I_A$ everywhere when $A \in \mathcal{A}_0$ and then $D(\{x : x \geq a\}|y) = 1$ whenever $y \in \mathcal{X}$, $a \in \mathcal{X}$ and $y \geq a$. Putting $y = a$ this yields $D(\{x : x \geq y\}|y) = 1$; $y \in \mathcal{X}$.

It follows that there are random variables $X \geq Y$ having, respectively, distributions P and Q if and only if $P(A) \geq Q(A)$ for each increasing measurable set A.

Using the theory of comparison of statistical models we have thus obtained this important particular case of a very general existence result of Strassen (1965). We may however proceed the other way around and derive the latter result from the theory of comparison of measure families (see, e.g. Torgersen (1985)).

5. Dichotomies. Lorenz Functions and Neyman-Pearson Functions. Experiments having two point parameter sets, i.e. dichotomies, enjoy a variety of striking properties which are not shared by experiments in general.

Thus comparison of dichotomies may be expressed solely in terms of testing problems and the information ordering is in this case a lattice ordering. The crucial property of dichotomies is that they all have monotone likelihood in some statistics. Indeed, by Lehmann (1988) and Torgersen (1989), many properties of dichotomies extend, properly formulated, to such experiments.

We shall here present some of the basic properties of dichotomies. A discussion of the more general case of measure pairs, i.e. $\mathbb{R}^2$-valued measures, will appear in Torgersen (1991).

The basic assumption in this section is thus that the parameter set is a two point set and we shall proceed assuming that this set actually is $\Theta = \{0, 1\}$. Thus a dichotomy $\mathcal{D}$ is an ordered pair $\mathcal{D} = (P_0, P_1)$ of probability distribution on a common measurable space. Convenient tools are then:

(i) The relationship between level of significance and maximum power for testing, say, "$\theta = 0$" against "$\theta = 1$".

(ii) The relationship between prior distribution and minimum Bayes risk for testing "$\theta = 0$" against "$\theta = 1$" with 0-1 loss.

(iii) Variations of standard measures and Blackwell measures.

(iv) The Hellinger transform.

The relationship (i) is given by functions which, in one form or another, appear to play important roles at the most diverse occasions, not all of them in statistics. Although not widely recoqnized, even among statisticians, their genesis may be regarded as rooted in the Neyman-Pearson lemma. We shall here say that a function is a *Neyman-Pearson function (N-P function)* if it is a continuous concave function from the unit interval [0,1] into itself which leaves 1 fixed. Of course concavity ensures continuity on the open interval $]0,1[$ and if, in addition, it is assumed that 1 is a fixed point then it is automatically continuous on]0,1]. Thus a function β from the unit interval to itself is a N-P function if and only if it is concave, $\beta(0+) = \beta(0)$ and $\beta(1) = 1$.

In statistics N-P functions arise in testing theory in many situations which are not directly related to the Neyman-Pearson lemma. Thus e.g. the maximin level α power defines a N-P function β of α provided we ensure that $\beta(0+) = \beta(0)$. [If the weak compactness lemma holds then this is automatic. In general we may just define $\beta(0)$ as $\beta(0+)$.]

More generally we may consider maximin level α power for test functions belonging to a given convex class of test functions containing the constants in [0,1].

In particular if $\mathcal{D} = (P_0, P_1)$ is a dichotomy, then *the N-P function* of $\mathcal{D}$ is the function $\beta(\cdot|\mathcal{D})$ which to each $\alpha \in [0,1]$ assigns the power $\beta(\alpha|\mathcal{D})$ of the most powerful level α test for testing "$\theta = 0$" against "$\theta = 1$". When convenient, this function may also be denoted as $\beta(\cdot|P_0, P_1)$.

The N-P functions and their close relatives appear in abundance in statistics and in econometry. For instance, in econonometrics spread is frequently described in terms of Lorenz functions (see e.g. Arnold (1987)). The relationship between Lorenz functions and N-P functions may be described as follows:

Let F be any distribution on $[0,\infty[$ having finite positive expectation $\mu_F = \int xF(dx) = \int_0^1 F^{-1}(p)dp$. The Lorenz function of F is the function L_F on [0,1] defined by:

$$L_F(p) \equiv_p \int_0^p F^{-1}(t)dt/\mu_F.$$

Put $F_0 = F$ and let F_1 have density $x \to x/\mu_F$ w.r.t. F_0. Let K be the distribution of dF_1/dF_0 under F_0. Then $K^{-1} = F^{-1}/\mu_F$ and $L_F(p) \equiv_p 1 - \beta(1-p|\mathcal{D}_F)$ where $\mathcal{D}_F$ in the dichotomy (F_0, F_1).

It is easily inferred that a function L is a Lorenz function if and only if it is of the form $L(\alpha) \equiv 1 - \beta(1-\alpha)$ for a N-P function β such that $\beta(0) = 0$. If L has this form where β is the N-P function of the dichotomy (P_0, P_1), then L is the inverse function of the N-P function of the reversed dichotomy (P_1, P_0).

It follows that a function is a Lorenz function if and only if it is a continuous convex function from [0,1] onto [0,1] having the origin as a fixed point.

Considering two probability distributions F and G on $[0,\infty[$ with finite positive expectations, following Arnold (1987), we may say that G *Lorenz majorizes* F *if* $L_F \geq L_G$. By the results described here, this amounts to the condition that $\mathcal{D}_G \geq \mathcal{D}_F$ where $\mathcal{D}_G$ is defined in terms of G as $\mathcal{D}_F$ above was defined in terms of F.

Another notion related to the N-P functions is the total time on test (TTT) transform in reliability theory. These are, see Klefsjø (1984), the functions of the form $\alpha \to 1 - \beta(1-\alpha) + (1-\alpha)\beta'(1-\alpha)$ for a N-P function β.

EXAMPLE 5.1. (Double dichotomies and triangular N-P functions). If β is a N-P function, then $\alpha \leq \beta(\alpha) \leq 1$ for all $\alpha \in [0,1]$. The lefthand side corresponds to the N-P function of a totally noninformative dichotomy (P,P) while the right hand side corresponds to a totally informative dichotomy (P_0,P_1) with P_0 and P_1 being mutually singular.

An interesting family of N-P functions (which include the above mentioned) are the triangular ones. These are the N-P functions of the double dichotomies. Thus the N-P function of the double dichotomy $((1-p,p),(1-q,q))$ with $p \leq q$ is the upper boundary of the triangle $((0,0),(p,q),(1,1))$.

The functions $\alpha \to \alpha^p; 0 \leq p \leq 1$ are N-P functions. If $p = 0$ or $p = 1$ then we have just seen that they arise from dichotomies. The reader might try to figure out which famous model (related to income distributions) provides these functions.

Any N-P function is the N-P function of a dichotomy and, as we shall explain soon, any dichotomy is defined up to equivalence by its N-P function. Accepting this for the moment, we realize that operations on dichotomies and on N-P functions are the same thing.

Thus if $\mathcal{D}_1$ and $\mathcal{D}_2$ have N-P functions β_1 and β_2, respectively, then the mixture $(1-p)\mathcal{D}_1 + p\mathcal{D}_2$ and the product $\mathcal{D}_1 \times \mathcal{D}_2$ have, respectively, N-P functions β and γ given by:

$$\beta(\alpha) \equiv \sup\{(1-p)\beta_1(\alpha_1) + p\beta_2(\alpha_2) : (1-p)\alpha_1 + p\alpha_2 = \alpha\}$$

and

$$\gamma(\alpha) \equiv \sup\{\int_0^1 \beta_1(\alpha(x))\beta_2(dx) : \int_0^1 \alpha(x)dx = \alpha\}.$$

It is not immediate from these formulas that products are distributive w.r.t. mixtures. This is however clear from the fact that the Hellinger transform,

which is defined for dichotomies later in this section, is multiplicative for products and affine under mixtures.

Proceeding the other way around, we find that the class of N-P functions is closed for several standard operations on numerical functions. Thus convex combinations of N-P functions are themselves N-P functions. It follows that if $\mathcal{D}_1$ and $\mathcal{D}_2$ are dichotomies having N-P functions β_1 and β_2 and if p is a number in [0,1] then there is, up to equivalence, a unique dichotomy $\mathcal{D}$ having $(1-p)\beta_1 + p\beta_2$ as its N-P function. This dichotomy is at most as informative as $(1-p)\mathcal{D}_1 + p\mathcal{D}_2$, and generally it is less informative than this mixture.

By Torgersen (1970) any dichotomy has an essentially unique decomposition as a mixture of a totally ordered family of double dichotomies.

Other interesting operations are the lattice operations derived from the information ordering and the operation of functional composition of N-P functions.

Consider a family $(\mathcal{D}_i : i \in I)$ of dichotomies. If β_i is the N-P function of $\mathcal{D}_i$ then the pointwise infimum $\inf_i \beta_i$ is also a N-P function. Any dicholomy $\underline{\mathcal{D}}$ having this function as its N-P function possesses necessarily the following properties: Firstly $\underline{\mathcal{D}} \leq \mathcal{D}_i$ for all $i \in I$. Secondly: If $\mathcal{D}$ is any dichotomy such that $\mathcal{D} \leq \mathcal{D}_i$ for all $i \in I$ then $\mathcal{D} \leq \underline{\mathcal{D}}$. Thus $\underline{\mathcal{D}}$ is a greatest lower bound (infimum) of the family $(\mathcal{D}_i : i \in I)$.

It follows that the collection of dichotomies is order complete for the informational ordering. Note, however, that the sup operation expressed for N-P functions is not the pointwise supremum. It corresponds of course to the supremum operation on N-P functions for the informational ordering.

Monotone likelihood experiments are very naturally represented as families of N-P functions (Torgersen (1989)). These families are characterized by being closed for the "natural" functional compositions. In general if $\mathcal{D}_1$ and $\mathcal{D}_2$ are dichotomies having N-P functions β_1 and β_2, respectively, then the composed function $\beta_1(\beta_2) = \beta_1 \circ \beta_2$ is also a N-P function. If $\mathcal{D}$ is a dichotomy having $\beta_1(\beta_2)$ as its N-P function then $\mathcal{D}$ is at least as informative as the product dichotomy $\mathcal{D}_1 \times \mathcal{D}_2$. Indeed if γ is the N-P function of $\mathcal{D}_1 \times \mathcal{D}_2$ then, for any $\alpha \in [0,1], \gamma(\alpha) = \sup\{\int \beta_1(\alpha(x))\beta_2(dx) : \int_0^1 \alpha(x)dx = \alpha\} \leq$ (by Jensen's inequality) $\sup\{\beta_1(\int \alpha(x)\beta_2(dx)) : \int_0^1 \alpha(x)dx = \alpha\} = \beta_1(\beta_2(\alpha))$.

As mentioned above, any N-P function arises from a dichotomy. In fact a N-P function β is also a cumulative distribution function of a probability distribution on [0,1] which is absolutely continuous on]0,1]. In fact it may be checked that β is the N-P function of the pair $(R(0,1), \beta)$ where $R(0,1)$ denotes the rectangular distribution on (0,1).

The N-P function of a dichotomy $\mathcal{D} = (P_0, P_1)$ is usually found by first finding a real valued sufficient statistic X, e.g. $X = dP_1/d(P_0 + P_1)$, such that $F_1 = \mathcal{L}(X|P_1)$ has a monotonically increasing density w.r.t. $F_0 = \mathcal{L}(X|P_0)$. By the Neyman Pearson lemma $\beta(\alpha|P_0, P_1) = 1 - F_1(F_0^{-1}(1-\alpha))$ for any $\alpha \in]0,1[$ such that $F_0(F_0^{-1}(1-\alpha)) = 1-\alpha$. In general this formula holds for any $\alpha \in]0,1[$, provided we permit a random mass in $F_0^{-1}(1-\alpha)$ distributed uniformly on $[0, m(\alpha)]$, where $m(\alpha)$ is the F_0 mass in $F_0^{-1}(1-\alpha)$.

All dichotomies having the same N-P function β are statistically equivalent with the dichotomy $(R(0,1), \beta)$. Using the terminology of LeCam (1986) we may express this by saying that $\beta(|\mathcal{D})$ defines the type of the dichotomy $\mathcal{D}$. In fact if $\hat{\alpha}$ is the observed significance level for testing "$\theta = 0$" against "$\theta = 1$" in $\mathcal{D} = (P_1, P_0)$ then $\mathcal{L}(\hat{\alpha}|P_0) = R(0,1)$ and $\mathcal{L}(\hat{\alpha}|P_1) = \beta$.

The dual of a N-P function β is the function b on [0,1] given by:

$$b(\lambda) \equiv_\lambda \min_\alpha [(1-\lambda)\alpha + \lambda(1-\beta(\alpha))]$$

The function β may be recovered from b by:

$$\beta(\alpha) \equiv_\alpha \inf_{\lambda > 0} \frac{1}{\lambda}[(1-\lambda)\alpha + \lambda - b(\lambda)]$$

The function b is clearly also concave and nonnegative and $b(\lambda) \leq \min(\lambda, 1-\lambda) \leq \frac{1}{2}$. If $\beta = \beta(\cdot|P_0, P_1)$ for a dichotomy $\mathcal{D} = (P_0, P_1)$ then $b(\lambda) = b(\lambda|P_0, P_1) = \|(1-\lambda)P_0 \wedge \lambda P_1\|$ is the minimum Bayes risk in the above testing problem with $0-1$ loss and prior distribution $(1-\lambda, \lambda)$.

The dichotomy $\mathcal{D} = (P_0, P_1)$ is, up to equivalence, completely characterized by the distribution $K(\cdot|P_0, P_1)$ of dP_1/dP_0 under P_0. The probability measure K may be probability distribution K on $[0, \infty[$ such that $\int xK(dx) \leq 1$. In fact if this condition is satisfied then $K = K(\cdot|K_0, K_1)$ where $K_0 = K$ and K_1 is the unique probability measure on $[0, \infty]$ having density $x \to x$ w.r.t. K. Thus K, assigns mass $1 - \int xK(dx)$ at ∞.

Furthermore, if $K = K(\cdot|P_0, P_1)$ for a dichotomy (P_0, P_1) then $K_i; i = 0, 1$ is the distribution under P_i of the P_1 maximal version of dP_1/dP_0. Thus, since dP_1/dP_0 is sufficient, the dichotomies (P_0, P_1) and (K_0, K_1) are equivalent. It may be checked that:

$$\beta(\alpha|P_0, P_1) \equiv_\alpha 1 - \int_\alpha^1 K^{-1}(1-p|P_0, P_1)dp$$

and

$$b(\lambda|P_0, P_1) \equiv_\lambda \int [(1-\lambda) \wedge \lambda x] K(dx|P_0, P_1)$$

and that

$$K = \mathcal{L}(\beta'(U)) \quad \text{when} \quad \mathcal{L}(U) = R(0,1).$$

The observed significance level $\hat{\alpha}$ for a dichotomy (P_0, P_1) for testing "$\theta = 0$" against "$\theta = 1$" may be expressed in terms of K by:

$$\hat{\alpha} = K(]dP_1/dP_0, \infty[) + UK(\{dP_1/dP_0\})$$

where U is independent of dP_1/dP_0 and uniformly distributed on [0,1].

Putting $\delta_\alpha = 1$ or $= 0$ as $\hat{\alpha} \leq \alpha$ or $\hat{\alpha} > \alpha$, we obtain a right continuous (in α) monotonically increasing family of test functions $\delta_\alpha : \alpha \in [0,1]$ such that

$$E_0\delta_\alpha \equiv \alpha \quad \text{while} \quad E_1\delta_\alpha \equiv \beta(\alpha|P_0, P_1).$$

Of course we do not need the random variable U. Conditioning on the sufficent statistic dP_1/dP_0 we may ensure that δ_α is the unique most powerful level α test which is functionally dependent on dP_1/dP_0. If so, then there are constants c_α and γ_α such that

$$\delta_\alpha = \begin{cases} 1 & \text{for } dP_1/dP_0 > c_\alpha \\ \gamma_\alpha & \text{for } dP_1/dP_0 = c_\alpha \\ 0 & \text{for } dP_1/dP_0 < c_\alpha. \end{cases}$$

More generally if γ is any N-P function such that $\gamma(\alpha) \leq \beta(\alpha|P_0, P_1)$ for all $\alpha \in [0,1]$, then there is a right continuous monotonically increasing family of test functions $\varphi_\alpha : \alpha \in [0,1]$ in $\mathcal{D} = (P_0, P_1)$ such that

$$E_0\varphi_\alpha \equiv \alpha \quad \text{while} \quad E_1\varphi_\alpha \equiv \gamma(\alpha).$$

If e.g. γ is given as the upper boundary of the convex hull of points $(0,b)$, (p_1,q_1), (p_2,q_2) and (1,1) where $0 \leq p_1 \leq p_2 \leq 1$ and $\gamma(0) = b, \gamma(p_i) = q_i; i = 1,2$, then we may construct the family $\varphi_\alpha : \alpha \in [0,1]$ in the following steps:

(i) Let $\delta_\alpha : \alpha \in [0,1]$ be given as above.

(ii) Put $\varphi_0 = [b/\beta(0|P_0, P_1)]\delta_0$.

(iii) Let α_1 be the smallest number $\alpha_1 \geq 0$ such that the graph of $\beta(\cdot|P_0, P_1)$ intersects the line through $(0,b)$ and (p_1,q_1) in the point $(\alpha_1, \beta(\alpha_1|P_0, P_1))$. Define φ_α so that $\varphi_\alpha = (1-\theta)\varphi_0 + \theta\delta_{\alpha_1}$ when $\alpha = (1-\theta)0 + \theta\alpha_1$ is in $[0, p_1]$.

(iv) Let α_2 be the smallest number $\alpha_2 \geq \alpha_1$ such that the line through (p_1,q_1) and (p_2,q_2) intersects the graph of $\beta(\cdot|P_0, P_1)$ in $(\alpha_2, \beta(\alpha_2|P_0, P_1))$. Define φ_α so that $\varphi_\alpha = (1-\theta)\varphi_{p_1} + \theta\delta_{\alpha_2}$ when $\alpha = (1-\theta)p_1 + \theta\alpha_2$ is in $[p_1, p_2]$.

(v) Put $\varphi_\alpha = (1-\theta)\varphi_{p_2} + \theta \cdot 1$ for $\alpha = (1-\theta)p_2 + \theta \cdot 1$ in $[p_2, 1]$.

It may be checked that $\delta_0 \geq \varphi_0, \delta_{\alpha_1} \geq \varphi_{p_1}, \delta_{\alpha_2} \geq \varphi_{p_2}$ and that $\varphi_0 \leq \varphi_{p_1} \leq \varphi_{p_2} \leq 1$, so that $\varphi_\alpha : \alpha \in [0,1]$ satisfy our requirements.

Proceeding by induction, we obtain for any polygonal $\gamma \leq \beta(\cdot|P_0, P_1)$ a representation $\varphi_\alpha : \alpha \in [0,1]$. By compactness this extends to any N-P function $\gamma \leq \beta(\cdot|P_0, P_1)$.

This procedure is closely related to the procedure whereby we may pass from a vector p to a vector q which is majorized by p by a finite number of "decreasing" steps, each modifying only two coordinates.

Suppose now that $\gamma(\cdot|Q_0, Q_1) \leq \beta(\cdot|P_0, P_1)$ for a dichotomy (Q_0, Q_1). Then $\gamma(\alpha|Q_0, Q_1) \equiv E_1\varphi_\alpha$ where $\alpha \equiv E_0\varphi_\alpha$ for an increasing right continuous family $\varphi_\alpha : 0 \leq \alpha \leq 1$ of test functions in $\mathcal{D} = (P_0, P_1)$. Let $M(\cdot|x)$, for each x in the sample space of $\mathcal{D}$ be the measure on [0,1] having distribution function $\alpha \to \varphi_\alpha(x)$. Letting $R(0,1)$ denote the uniform distrubution on (0,1) we find for any Borel set $B \subseteq [0,1]$ that:

$$R(0,1)(B) = \int M(B|x)P_0(dx)$$

while:

$$\gamma(B|P_0, P_1) = \int_B \gamma(dx|P_0, P_1) = \int M(B|x)P_1(dx)$$

Finally, the conditional distributions (which may not be regularizable), given the sufficient statistics dQ_1/dQ_0 in (Q_0, Q_1), provide a transition from $(R(0,1), \gamma)$ to (Q_0, Q_1). Combining these constructions we obtain a transition T such that $P_iT = Q_i$, $i = 0, 1$.

This establishes the transition criterion for comparison of dichotomies. It follows that comparison of dichotomies may be completely decided by testing problems (see Blackwell (1953)). The general transition criterion for ϵ-comparison of statistical experiments is established in LeCam (1964).

Some other basic comparison rules for dichotomies are collected in the following theorem.

THEOREM 5.2. (Approximate comparison of dichotomies). *Let* $\mathcal{D} = (P_0, P_1)$ *and* $\widetilde{\mathcal{D}} = (\widetilde{P}_0, \widetilde{P}_1)$ *be dichotomies. Then the following conditions are equivalent for nonnegative numbers* ϵ_0 *and* ϵ_1*:*

(i) $\beta(\alpha + \epsilon_0/2|\mathcal{D}) + \epsilon_1/2 \geq \beta(\alpha|\widetilde{\mathcal{D}})$; $0 \leq \alpha \leq 1 - \frac{1}{2}\epsilon_0$.

(ii) $b(\lambda|\mathcal{D}) \leq b(\lambda, \widetilde{\mathcal{D}}) + (1-\lambda)\epsilon_0/2 + \lambda\epsilon_1/2$; $0 \leq \lambda \leq 1$.

(iii) $\int \gamma(dP_1/dP_0)dP_0 \leq \int \gamma(dQ_1/dQ_0)dQ_0 + \frac{1}{2}\epsilon_0[\gamma(\infty) - \gamma(0)] + \frac{1}{2}\epsilon_1\gamma'(0)$, *when* γ *is a concave function on* $[0, \infty[$ *such that* $\gamma(x)/x \to 0$ *as* $x \to \infty$.

(iv) $\|P_iT - Q_i\| \leq \epsilon_i$; $i = 0, 1$ *for a transition* T *from* $\mathcal{D}$ *to* $\widetilde{\mathcal{D}}$.

If $\epsilon_0 = \epsilon_1 = 0$ this reduces to:

THEOREM 5.3. (Comparability of dichotomies). *With notations as in the previous theorem for dichotomies* $\mathcal{D}$ *and* $\widetilde{\mathcal{D}}$, *the following conditions are equivalent:*

(i) $\beta(\cdot|\mathcal{D} \geq \beta(\cdot|\widetilde{\mathcal{D}})$.

(ii) $b(\cdot|\mathcal{D}) \leq b(\cdot|\widetilde{\mathcal{D}})$.

(iii) Assuming $P_0 >> P_1$ $\int \varphi(dP_1/dP_0)dP_0 \geq \int \varphi(dQ_1/dQ_0)dQ_0$ *when* φ *is convex on* $[0, \infty[$.

(iv) $P_iT = Q_i$; $i = 0, 1$ *for a transition* T.

(v) Assuming $P_0 >> P_1$ $\mathcal{L}(dP_1/dP_0|P_0) = D\mathcal{L}(dQ_1/dQ_0|Q_0)$ *for a dilation* D *on* $[0, \infty[$.

These conditions imply all

(vi) $\int [dP_1/dP_0]^t dP_0 \leq \int [dQ_1/dQ_0]^t dQ_0$; $0 \leq t \leq 1$.

REMARKS. The equivalent conditions (i)–(v) all express that $\mathcal{D}$ is at least as informative as $\widetilde{\mathcal{D}}$. A dilation on $[0, \infty[$ is a Markov kernel D from $[0, \infty[$ to $[0, \infty[$ such that $\int xD(dx|y) = y$; $y \geq 0$.

The integral $\int (dP_1/dP_0)^t dP_0$ for a dichotomy $\mathcal{D} = (P_0, P_1)$ is, as a function of $t \in [0, 1]$, the *Hellinger transform* of $\mathcal{D}$. It defines $\mathcal{D}$ up to equivalence. However, the ordering described by (vi) does not imply that $\mathcal{D}$ is at least as informative as $\widetilde{\mathcal{D}}$ (see Torgersen (1970)). Within the theory of statistical experiments the Hellinger transforms have a similar role as characteristic functions have in probability theory.

In terms of the N-P function β of $\mathcal{D}$ the Hellinger transform may be expressed as:

$$t \to \int_0^1 [\beta'(\alpha)]^t d\alpha = \int_0^1 [K^{-1}(\alpha)]^t d\alpha,$$

where

$$K = \mathcal{L}(dP_1/dP_0|P_0).$$

Turning to econometric applications we obtain the following well-known characterizations of the Lorenz ordering:

COROLLARY 5.4. (The Lorenz ordering). *Let* X *and* Y *be nonnegative random variables having finite positive expectations.*

Let F *be the distribution of* X *and let* G *be the distribution of* Y.

Let F_1 be the distribution having density $x \to x/EX$ w.r.t. F and let G_1 be the distribution having density $y \to y/EY$ w.r.t. G.

Then the following conditions are equivalent:

(i) F *Lorenz majorizes* G.

(ii) $E(X - cEX)^{\pm}/EX \geq E(Y - cEY)^{\pm}/EY$; $c \in \mathbb{R}$ *where the* $\pm$ *signs in exponential position may either both be replaced by* $+$ *or both be replaced with* $-$ *or both be deleted provided* $(X - c)$ *and* $(Y - c)$ *are replaced by, respectively,* $|X - c|$ *and* $|Y - c|$.

(iii) $E\varphi(X/EX) \geq E\varphi(Y/EY)$ *when* φ *is convex on* $[0, \infty[$.

(iv) $G = FM$ *and* $G_1 = F_1 M$ *for a Markov kernel* M.

(v) There are random variables $\widetilde{X}$ *and* $\widetilde{Y}$ *having distributions* F *and* G, *respectively, such that:*

$$E((\widetilde{X}/E\widetilde{X}|\widetilde{Y}) \geq \widetilde{Y}/E\widetilde{Y}.$$

Consider now nonnull observation vectors $p = (p_1, \ldots, p_d)$ and $q = (q_1, \ldots, q_d)$ having nonnegative coordinates. Letting F_p and F_q be the empirical distribution functions based on, respectively, p and q we find that F_p Lorenz majorizes F_q if and only if $p/\sum p_i$ majorizes $q/\sum q_i$.

Proceeding to approximate majorization, we say that a distribution function F having Lorenz function L_F, ϵ-Lorenz majorizes the distribution function G, having Lorenz function L_G, if $L_F \leq L_G + \frac{1}{2}\epsilon$. Theorem 5.2 yields then:

COROLLARY 5.5. (Approximate Lorenz ordering). *With the notations of the previous corollary the following conditions are equivalent:*

(i) F ϵ*-Lorenz majorizes* G.

(ii) $E(|X - cEX|/EX) \geq E(|Y - cEY|/EY) - \epsilon$; $c \in \mathbb{R}$.

(iii) $E\varphi(X/EX) \geq E\varphi(Y/EY) - \frac{1}{2}\epsilon[\varphi'(\infty) - \varphi'(0)]$ *where* φ *is convex on* $[0, \infty[$ *and the quantities* $\varphi'(0) = \lim_{x \to 0}[\varphi(x) - \varphi(0)]/x$ *and* $\varphi'(\infty) = \lim_{x \to \infty} \varphi'(x)$ *are both finite.*

(iv) $\|G_1 - F_1 M\| \leq \epsilon$ *for a Markov kernel* M *such that* $G = FM$.

(v) There are random variables $\widetilde{X}$ *and* $\widetilde{Y}$ *having, respectively, distributions* F *and* G *such that:*

$$E|E((\widetilde{X}/E\widetilde{X})|\widetilde{Y}) - (\widetilde{Y}/E\widetilde{Y})| \leq \epsilon.$$

Consider again nonnull observation vectors $p = (p_1, \ldots, p_d)$ and $q = (q_1, \ldots, q_d)$ having nonnegative coordinates. Then the empirical distribution

function F_p based on p ϵ/d-Lorenz majorizes the empirical distribution function F_q based on q if and only if the probability vector $p/\sum p_i$ ϵ-majorizes the probability vector $q/\sum q_i$.

Schur convex functions of probability vectors in $\mathbb{R}^d$ are information increasing functionals of those β-functions which are linear on intervals $[\frac{i-1}{d}, \frac{1}{d}]$; $i = 1, \ldots, d$. In general a functional of types of dichotomies is monotonically increasing if and only if it may be represented as a monotonically increasing functional of N-P functions, i.e. if and only if it may be represented as a monotonically decreasing functional of Lorenz functions.

EXAMPLE 5.6. (The Gini index). If $\mathcal{D} = (P_0, P_1)$ is a dichotomy then the area

$$G = \|(P_0 \times P_1) - (P_1 \times P_0)\|/2 = 1 - \|(P_0 \times P_1) \wedge (P_1 \times P_0)\|$$

of the convex hull of the range of the vector valued measure (P_0, P_1) depends on $\mathcal{D}$ only via its type. If $\mathcal{D}$ has Neyman-Pearson function β, then it is equivalent to $(\mathbb{R}[0,1], \beta)$ and thus $\|(P_0 \times P_1) \wedge (P_1 \times P_0)\| = \int\int [\beta'(\alpha_2) \wedge \beta'(\alpha_1)] d\alpha_1 d\alpha_2 = 2\int\int_{\alpha_1 > \alpha_2} \beta'(\alpha_2) d\alpha_1 d\alpha_2 = 2\int_0^1 (1 - \beta(\alpha_1)) d\alpha_1$. Hence

$$G = 2\int_0^1 \beta(\alpha) d\alpha - 1 = 1 - 2\int_0^1 L(\alpha) d\alpha,$$

where $L(\alpha) \equiv_\alpha 1 - \beta(1 - \alpha)$ is a Lorenz function provided $P_0 >> P_1$.

REFERENCES

BLACKWELL, D. (1953). Equivalent comparisons of experiments. *Ann. Math. Statist.* **24**, 265–272.

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

FISHER, P. and HOLBROOK, J. A. R. (1980). Balayage defined by the non negative convex functions. *Proc. Amer. Math. Soc.* **79**, 445.

HEYER, H. (1982). *Theory of Statistical Experiments.* Springer-Verlag, Berlin.

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, New York.

KLEFSJÖ, B. (1984). Reliability interpretations of some concepts from economics. *Naval Res. Logist. Quart.* **31**, 301–308.

LECAM, L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Statist.* **35**, 1419–1455.

LeCam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, Berlin.

Lehmann, E. (1988). Comparing location experiments. *Ann. Statist.* **16**, 521–533.

Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and its Applications.* Academic Press, New York.

Strassen, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.* **36**, 423–439.

Strasser, H. (1985). *Mathematical Theory of Statistics.* Walter de Gruyter.

Torgersen, E. (1970). Comparison of experiments when the parameter space is finite. *Z. Wahrsch. Verw. Gebiete* **16**, 219–249.

Torgersen, E. (1976). Comparison of statistical experiments. *Scand. J. Statist.* **3**, 186–208.

Torgersen, E. (1985). 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). Lecture Notes in Statistics No. 35. Springer-Verlag.

Torgersen, E. (1989). Monotone likelihood, power function diagrams and selection. Statistical Research Report. Univ. of Oslo.

Torgersen, E. (1991). *Comparison of Statistical Experiments.* Cambridge Univ. Press, Cambridge.

Mathematical Institute
University of Oslo
P.O. Box 1053 Blindern
N-0316 Oslo 3, Norway