

# STOCHASTIC ORDERS AND COMPARISON OF EXPERIMENTS

Erik Torgersen  
University of Oslo

## Abstract

Exploring criteria for majorization, exact and approximate, univariate and multivariate, we relate them to criteria for information orderings of statistical experiments. After having provided some of the basic criteria for comparison of experiments we observe their straight forward generalizations to general families of measures. Thus LeCam's randomization criterion extends to a criterion for comparison of 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 of 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 throw 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. If the reader wants more background then he or

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

- (iii)  $\sum_i (p_i - c)^+ \geq \sum_i (q_i - c)^+$ ;  $c \in R$ .
- (iv)  $\sum_i (p_i - c)^- \geq \sum_i (q_i - c)^-$ ;  $c \in R$ .
- (v)  $\|p - ce\|_1 \geq \|q - ce\|_1$ ;  $c \in R$ .
- (vi)  $\sum_i g(p_i) \geq \sum_i g(q_i)$ ; when  $g$  is convex on  $R$ .
- (vii)  $\varphi(p) \geq \varphi(q)$  when  $\varphi$  is quasiconvex and permutation symmetric on  $R^d$ .
- (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  $\Pi$  is the group of coordinate permutations on  $R^d$ .
- (x)  $A_p \supseteq A_q$ .
- (xi)  $\beta_p(\alpha) \geq \beta_q(\alpha)$ ;  $0 \leq \alpha \leq 1$ .
- (xii)  $\begin{pmatrix} e/d \\ p \end{pmatrix} M = \begin{pmatrix} e/d \\ q \end{pmatrix}$  for a (necessarily doubly stochastic)  $d \times d$  Markov matrix  $M$ .
- (xiii) The empirical distribution function based on the observations  $p_1, \dots, p_d$  is a dilation of the empirical distribution function based on the observations  $q_1, \dots, q_d$ .

If  $p$  and  $q$  are probability vectors then these conditions are equivalent with

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

---

**Remark.** 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 if  $c \in R$  and  $t > 0$  then  $p \succ q$  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  $R^d$  then several of the criteria of the theorem have interesting interpretations in terms of statistical decision theory.

Consider a statistical model obtained by observing a random variable  $X$  whose distribution  $P_\theta$  depends on an unknown parameter  $\theta$ . Assume for the moment that we know that  $\theta$  is one of the numbers 0 and 1 and that  $X$  is one of the numbers  $1, 2, \dots, d$ . 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:

$t_2 \leq t_1 \leq 0$  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 here 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  $R^d$  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 experiment  $(e/d, p, q)$  has monotonically increasing likelihood ratio in  $p_i$  and thus the dichotomy  $(e/d, p)$  is at most as informative as the dichotomy  $(e/d, q)$ .

---

Two weaker concepts of majorization are those of weak sub majorization and of weak super majorization. Also these concepts fit nicely into a decision theoretal framework. Before discussing that however we shall find it convenient to consider approximate majorization.

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

Considering two vectors  $p$  and  $q$  in  $R^d$  and a constant  $\epsilon \geq 0$  we shall say that  $p$   $\epsilon$ -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  $p$   $\epsilon$ -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  $\epsilon$ -majorization. Observe first that  $p$   $\epsilon$ -majorizes  $q$  if and only if  $q$  admits a decomposition  $q = \tilde{q} + v$  where  $\tilde{q} \prec p$  while  $\|v\|_1 \leq \epsilon$ . It follows that the support function of the convex set consisting of vectors  $q$  which is  $\epsilon$ -majorized by  $p$  is  $a \rightarrow \vee_{\pi} \sum_{i=1}^d a_{\pi(i)} p_i + \epsilon \vee_{i=1}^d |a_i|$  where  $\pi$  runs through the permutation group on  $\{1, \dots, n\}$ . Hence  $q$  is  $\epsilon$ -majorized by  $p$  if and only if  $(q, a) \leq \vee_{\pi} (\pi(a), p) + \epsilon \vee_i |a_i|$ ;  $a \in R^d$  where  $\pi(a) \equiv (a_{\pi(1)}, \dots, a_{\pi(d)})$ .

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 only if it is of the form  $2b - e$  where  $0 \leq b_i \leq 1$ ;  $i = 1, \dots, d$ . Thus  $q$  is  $\epsilon$ -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  $\epsilon$ -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  $\sum_{i=r}^d p_{(i)} \geq \sum_{i=r}^d q_{(i)} - \frac{1}{2} (\sum_{i=1}^d (q_i - p_i)) - \epsilon/2$ ;  $r = 1, \dots, d+1$  where we put  $\sum_{i=d+1}^d = 0$ .

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

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

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

and

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

---

**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 conditions above follows then from the general theory of measure families. It may however be instructive first to consider the direct proof given here.

---

**Proof:** We have observed above that conditions (i<sub>1</sub>)–(ii<sub>2</sub>) were all equivalent. By theorem 1.1 these conditions are also equivalent with conditions (viii), (ix) and (xii).

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

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

Applying (vi<sub>2</sub>) to  $g(x) \equiv (x - ce)^+$  we see that (iii) holds. Letting  $c \rightarrow \pm\infty$  in (iii) we find that  $|\sum p_i - \sum q_i| \leq \epsilon$ . If so and if (vi<sub>2</sub>) holds for  $g$  then

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

is between  $g'(\infty)\epsilon$  and  $-g'(-\infty)\epsilon$ . Thus (vi<sub>2</sub>) implies that the inequality in (vi<sub>1</sub>) holds when the quantities  $g'(\pm\infty)$  are finite. If, however, one of the quantities  $g'(\pm\infty)$  are infinite then (vi) is trivial for  $g$  unless  $\epsilon = 0$ . By the above observation (vi<sub>2</sub>)  $\Rightarrow$  (iii) and (iii) amounts, by Theorem 1.1, to the condition that  $p \succ q$  when  $\epsilon = 0$ . Thus, by theorem 1.1 again, (vi<sub>2</sub>) implies (vi<sub>1</sub>) in any case. This shows that conditions (vi<sub>1</sub>) and (vi<sub>2</sub>) are equivalent and that these conditions imply conditions (iii)–(v). On the other hand if (iii) holds then condition (vi<sub>2</sub>) 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  $b_1, \dots, b_s \geq 0$ . Any polygonal convex function  $g$  is of this form and thus (vi<sub>2</sub>) follows by approximation. Altogether this shows that conditions (iii)–(vi<sub>2</sub>) are equivalent.

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

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

**Corollary 1.4 (Weak majorization).**

Let  $p$  and  $q$  be vectors in  $R^d$ . Then:

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

Thus theorem 1.3 furnishes equivalent criteria for weak majorization.

---

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

---

Theorem 1.3 provides several expressions for the smallest quantity  $\epsilon$  such that  $p$   $\epsilon$ -majorizes  $q$ . Denoting this quantity by  $\dot{\delta}(p, q)$  we obtain from criteria (ii<sub>1</sub>), (v) and (xi) the expressions:

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

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

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

Symmetrizing we obtain the majorization pseudo metric  $\dot{\Delta}$  on  $R^d$  which to vectors  $p$  and  $q$  assigns the distance

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

**Example 1.5 (Majorization between vectors of possibly different dimensions).**

Let  $p = (p_1, \dots, p_m)$  and  $q = (q_1, \dots, q_n)$  be probability vectors in, respectively,  $R^m$  and  $R^n$ . Let also, for  $k = 1, 2, \dots$ , the probability vector  $(1/k, \dots, 1/k)$  in  $R^k$  be denoted as  $u^{(k)}$ .

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

- (ii)  $\sum_{i=1}^d \varphi(p_{t_1}(i), \dots, p_{t_r}(i)) \geq \sum_{i=1}^d \varphi(q_{t_1}(i), \dots, q_{t_r}(i))$  whenever  $t_1, \dots, t_r \in T$  and  $\varphi$  is convex on  $R^r$ . (Actually it suffices to consider functions  $\varphi$  which are maxima of at most  $d$  linear functionals.)
- (iii) The empirical distribution function  $F_p$  based on the observations  $p.(1), \dots, p.(d)$  is a dilation of the empirical distribution function  $F_q$  based on the observations  $q.(1), \dots, q.(d)$ . (The observations are all real valued functions on  $T$ .)

---

Proceeding to  $\epsilon$ -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 families of probability vectors in  $R^d$ .

Consider also a family  $\epsilon = (\epsilon_t : t \in T)$  of non negative 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

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

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

- (iii) The empirical distribution function  $F_p$  based on the observations  $p.(1), \dots, p.(d)$  is a  $(F_q, \epsilon/d)$  dilation of the empirical distribution function  $F_q$  based on the observations  $q.(1), \dots, q.(d)$ . Here a Markov kernel  $D$  is called a  $(F_q, \epsilon)$  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 considered 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 non sequential 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

sup norm.

As sample spaces also 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  $\sup_t |f(t)|$  for a real valued function  $f$  on  $T$ . Considering the finite decision spaces  $T_k = \{1, \dots, k\}$ ;  $k=1, 2, \dots$ , it is tacitly assumed that all sub sets are measurable.

We shall admit as a possible loss function  $L$  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  $\mathcal{E} = (\chi, \mathcal{A}; P_\theta : \theta \in \Theta)$  is just a Markov kernel from the sample space  $(\chi, \mathcal{A})$  to the decision space  $(T, \mathcal{S})$ .

Decision rules, being Markov kernels, transport distributions forwards and functions backwards. Thus if  $\rho$  is a decision rule from  $\mathcal{E}$  to the decision space  $(T, \mathcal{S})$  and  $\mu$  is a finite measure on the sample space  $(\chi, \mathcal{A})$  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  $\rho$  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 we in addition to the decision rule  $\rho$  are given both a finite measure  $\mu$  on  $(\chi, \mathcal{A})$  and a bounded measurable function  $g$  on  $(T, \mathcal{S})$ . It is a fundamental fact that then 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 without ambiguity may write this quantity as  $\mu\rho g$ .

As a function of the pair  $(\mu, g)$  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  $\rho$  up to equivalence.

Considering the map  $\mu \rightarrow \mu\rho$  as a map from the  $L$ -space of finite measure on  $\mathcal{A}$  to the  $L$ -space of finite measures on  $\mathcal{S}$  we observe that it is linear, non negative (images of non negative elements are non negative) and preserves total masses. A map from one  $L$ -space to another having these properties is called a *transition*. Thus the decision rule  $\rho$  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  $\rho$  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 regularity conditions.

If the set-functions  $P_\theta \rho : \theta \in \Theta$  are all  $\sigma$ -additive, if  $(T, \mathcal{S})$  is Euclidean and if  $\mathcal{E}$  is dominated then the generalized decision rule  $\rho$  is definable in terms of a decision rule  $\rho$  from  $\mathcal{E}$  to  $(T, \mathcal{S})$ . If  $\rho$  is a generalized decision rule and  $g$  is a bounded measurable function on  $(T, \mathcal{S})$  then  $\rho g$  is the image of  $g$  by the conjugate map  $\rho^*$ , which also will be denoted as  $\rho$ . 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  $\mu\rho g$ .

**Theorem 3.1 (Deficiency for  $k$ -decision problems).**

Consider the set  $T_k = \{1, 2, \dots, k\}$  as a decision space.

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

Then the following conditions are equivalent for experiments  $\mathcal{E} = (\chi, \mathcal{A}; P_\theta : \theta \in \Theta)$  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)  $\sigma$  from  $\mathcal{F}$  to  $T$  corresponds a generalized decision rule (transition)  $\rho$  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  $\Theta_0$  of  $\Theta$  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)  $\sigma$  from  $\mathcal{F}$  to  $T$  corresponds a decision rule (Markov kernel)  $\rho$  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 sub linear 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  $\Theta_0$  of  $\Theta$  and for each function  $\psi$  on  $R^{\Theta_0}$  which is a maximum of  $k$  linear functionals.

Here  $e^\theta = (0, \dots, \overset{\theta}{1}, \dots, 0)$  denotes the  $\theta$ -th unit vector in  $R^{\Theta_0}$ .

(iv) *Comparison of performance functions:*

To each decision rule (Markov kernel)  $\sigma$  in  $\mathcal{F}$  corresponds a generalized decision rule  $\rho$  in  $\mathcal{E}$  so that:

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

**Proof:** The implications (iv)  $\Rightarrow$  (i)  $\Rightarrow$  (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.  $\rho$  and then w.r.t.  $\sigma$  it may be seen that (iii) is essentially a reformulation of (ii). The implication (ii)  $\Rightarrow$  (iv) follows, see e.g. Torgersen (1970), by standard minimax theory.  $\square$

The theorem is stated in order to make the generalization to general mass distributions more or less obvious. Knowing however that the distributions  $P_\theta$  and  $Q_\theta$  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  $\psi$ ) term  $\frac{1}{2} \sum_{\Theta_0} \epsilon_\theta [\psi(-e_\theta) + \psi(e_\theta)]$ . Actually we may in this case restrict

common measurable space. Doing that the other comments and definitions remain valid except that the equivalence induced by the deficiency distance  $\Delta_2$  is no longer the same as the equivalences induced by the deficiencies  $\Delta_3, \Delta_4, \dots$  and  $\Delta$ . The latter are however still the same. Of course it may not make much sense of interpreting the orderings  $\geq_1, \geq_2, \dots$  and  $\geq$  as information orderings.

We summarize these observations as:

**Theorem 3.3 (Comparison of measure families).**

Theorems 3.1–2 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.

---

**Remark.** The sub linear 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 sub linear functions  $\psi$  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 by some remarks on functionals of experiments having a common finite parameter set  $\Theta$ .

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

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

If, in addition,  $h$  is continuous then  $\bar{h}$  is continuous for the topology of the deficiency distance  $\Delta$ . By Torgersen (1990), any affine continuous functional of experiments is of the form  $\mathcal{E} \rightarrow \bar{h}(\mathcal{E})$  for a continuous homogeneous function  $h$  on  $[0, \infty[^\ominus$ . If, furthermore,  $h$  is sub linear on  $R^\ominus$  then this functional is, by theorem 3.2, monotonically increasing. Conversely any continuous affine monotonically increasing functional is of this form for a sub linear function  $h$  on  $R^\ominus$ .

**Example 3.4 (Multivariate Gini index).**

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

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.  $R^\ominus$  valued) measure  $\mathcal{E} = (\mu_\theta : \theta \in \Theta)$ . This set defines  $\mathcal{E}$  up to  $\Delta_2$  equivalence and for experiments  $\Delta_2$  equivalence is the same as full equivalence i.e.  $\Delta$ -equivalence. It follows

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

$$(*) \quad \sum_{1 \leq j_1 < \dots < j_n \leq r} |\det(a_{i,j_1}, \dots, a_{i,j_n})| = \text{Volume}(\langle 0, a_{\cdot,1} \rangle + \dots + \langle 0, a_{\cdot,r} \rangle)$$

where  $\langle \rangle$  denotes convex hull.

If the vectors  $a_{\cdot,1}, \dots, a_{\cdot,r}$  are linearly dependent then both sides of (\*) are zero.

If  $r = n$  and  $a_{\cdot,1}, \dots, a_{\cdot,r}$  are linearly independent then (\*) may, by (i), 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  $= s \geq 0$  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  $\mathcal{E} = (\chi, \mathcal{A}; \mu_\theta : \theta \in \Theta)$  is  $\epsilon$ -efficient 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  $\mu_{\theta_0}$  and  $\nu_{\theta_0}$  are non negative and dominates, respectively,  $\mathcal{E}$  and  $\mathcal{F}$ . Assume also that  $\epsilon_{\theta_0} = 0$ . Then  $\mathcal{E}$  is  $\epsilon$ -efficient w.r.t.  $\mathcal{F}$  if and only if  $\nu_{\theta_0} = \mu_{\theta_0} M$  for a transition  $M$  such that  $\|\mu_\theta M - \nu_\theta\| \leq \epsilon_\theta; \theta \in \Theta$ .

Let us, in order to escape difficult technical problems, assume that the underlying measurable spaces  $(\chi, \mathcal{A})$  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  $\theta$ , within an error of at most  $\epsilon_\theta$  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  $\epsilon$ -efficiency

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

(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  $P_1, \dots, P_k$  and  $Q_1, \dots, Q_k$  are not necessarily probability distributions. They may be non negative, 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  $\mu$  is a measure and  $h$  is a measurable function then  $h\mu$  denotes the measure (if it exists) having density  $h$  w.r.t.  $\mu$ .

---

**Proof:** By theorem 3.3 condition (i) amounts to the condition that  $\int \psi(1, v'_1, \dots, v'_k) dP \geq \int \psi(1, v'_1, \dots, v'_k) dQ$  when  $\psi$  is sub linear on  $R^{k+1}$ . Putting  $\varphi(x) = \psi(1, v'_1(x), \dots, v'_k(x))$  when  $x \in V$  this inequality may also be written  $\int \varphi dP \geq \int \varphi dQ$ . As  $\varphi$  is convex it is clear that (ii) $\Rightarrow$ (i). The converse implication is a consequence of the fact that a convex function  $\varphi$  on  $V$  which is a maximum of a finite set of affine functionals is of the form  $x \rightarrow \psi(1, v'_1(x), \dots, v'_k(x))$  for a sub linear function  $\psi$  on  $R^{k+1}$ .

On the other hand condition (i) is, by theorem 4.1, equivalent with the condition that  $P = DQ$  for a Markov kernel  $D$  such that  $\int v'_i(x) D(dx|y) = v'_i(y)$ ;  $i = 1, \dots, k$  when  $y \in V$ . As any linear functional on  $V$  is a linear combination of  $v'_1, \dots, v'_k$  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 non negative i.e. that (ii) holds.

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

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

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

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

Condition (iv) follows now by theorem 3.3. □

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 non negative finite measures  $\mu$  and  $\nu$  on a measurable space  $(\chi, \mathcal{A})$  and by replacing the class  $\Phi$  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  $\Phi$ . We shall here assume that  $H$  shares with  $\Phi$  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 so 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  $h_1, \dots, h_s$  in  $H$  along with a sub linear function  $\psi$  on  $R^s$  which is a maximum of a finite set of non negative linear functionals. If  $z \rightarrow \sum_{i=1}^s a_i z_i$  is one of these functionals then  $a_1, \dots, a_s \geq 0$  and thus, by convexity,  $\frac{1}{N} \sum_{i=1}^s a_i h_i = (1 - \frac{1}{N} \sum_{i=1}^s a_i) 0 + \sum_{i=1}^s (\frac{a_i}{N}) h_i \in H$  when  $N$  is sufficiently large. It follows that also  $\frac{1}{N} \psi(h_1, \dots, h_s) \in H$  when  $N$  is sufficiently large and thus  $\int \psi(h_1, \dots, h_s) d\mu = N \int [\psi(h_1, \dots, h_s)/N] d\mu \geq N \int [\psi(h_1, \dots, h_s)/N] d\nu = \int \psi(h_1, \dots, h_s) d\nu$ .

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  $\sigma$  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 that  $\int 1 d[(h\mu)M - h\nu] = \int 1 d[(h\mu)M] - \int 1 d(h\nu) = \int 1 d(h\mu) - \int 1 d(h\nu) = \int h d\mu - \int h d\nu$  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 " $\geq$ " indicates then simply that the measures  $(h\mu)M - h\nu$ ,  $h \in H$  are all non negative.

Assume now that  $1 \in H$  and that the measures  $\mu$  and  $\nu$  both have the same total 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  $\nu$  almost all  $y$  whenever  $h \in H$ .

We summarize these considerations as:

**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  $(\chi, \mathcal{A})$ . Assume that  $0 \in H$  and that  $h_1 \vee h_2 \in H$  when  $h_1, h_2 \in H$ .

Let  $\mu$  and  $\nu$  be non negative 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)  $\epsilon_h \geq 0$ ;  $h \in H$ .

## 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.  $R^2$ -valued measures, will appear in Torgersen (1990).

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 recognized, 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  $[0, 1]$  to 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  $\beta$  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  $\alpha$  power defines a N-P function  $\beta$  of  $\alpha$  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  $\alpha$  power for test functions belonging to a given convex class of test functions containing the constants in  $[0, 1]$ .

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, respectively, N-P functions  $\beta_1$  and  $\beta_2$  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  $\beta$  and  $\gamma$  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

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

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 round 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  $\beta_1$  and  $\beta_2$  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  $\beta_i$  is the N-P function of  $\mathcal{D}_i$  then the pointwise infimum  $\inf_i \beta_i$  is also a N-P function. Any dichotomy  $\underline{\mathcal{D}}$  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, Torgersen (1989), very naturally represented as families of N-P functions. 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, respectively N-P functions  $\beta_1$  and  $\beta_2$  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 most as informative as the product dichotomy  $\mathcal{D}_1 \times \mathcal{D}_2$ . Indeed if  $\gamma$  is the N-P function of  $\mathcal{D}_1 \times \mathcal{D}_2$

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

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 in this way a right continuous (in  $\alpha$ ) 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 w.r.t. the sufficient statistic  $dP_1/dP_0$  we may ensure that  $\delta_\alpha$  is the unique most powerful level  $\alpha$  test which is functionally dependent on  $dP_1/dP_0$ . If so then there are constants  $c_\alpha$  and  $\gamma_\alpha$  such that  $\delta_\alpha = 1, = \gamma_\alpha$  or  $= 0$  as

$$dP_1/dP_0 > c_\alpha, = c_\alpha \quad \text{or} \quad < c_\alpha.$$

More generally if  $\gamma$  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.  $\gamma$  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  $\alpha_1$  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  $(\alpha_1, \beta(\alpha_1|P_0, P_1))$ . Put so  $\varphi_\alpha = (1-\theta)\varphi_0 + \theta\delta_{\alpha_1}$  for  $\alpha = (1-\theta)0 + \theta\alpha_1$  in  $[0, p_1]$ .
- (iv) Let  $\alpha_2$  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  $(\alpha_2, \beta(\alpha_2|P_0, P_1))$ . Put so  $\varphi_\alpha = (1-\theta)\varphi_{p_1} + \theta\delta_{\alpha_2}$  for  $\alpha = (1-\theta)p_1 + \theta\alpha_2$  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  $[p_2, 1]$ .

- (i)  $\beta(\cdot|\mathcal{D}) \geq \beta(\cdot|\tilde{\mathcal{D}})$ .
- (ii)  $b(\cdot|\mathcal{D}) \leq b(\cdot|\tilde{\mathcal{D}})$ .
- (iii)  $\int \varphi(dP_1/dP_0)dP_0 \geq \int \varphi(dQ_1/dQ_0)dQ_0$  when  $\varphi$  is convex on  $[0, \infty[$ .
- (iv)  $P_i T = Q_i$ ;  $i=0, 1$  for a transition  $T$ .
- (v)  $\mathcal{L}(dP_1/dP_0|P_0) = D\mathcal{L}(dQ_1/dQ_0|Q_0)$  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) express all that  $\mathcal{D}$  is at least as informative as  $\tilde{\mathcal{D}}$ . 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. The ordering described by (vi) does not however, see Torgersen (1970), imply that  $\mathcal{D}$  is at least as informative as  $\tilde{\mathcal{D}}$ . 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  $\beta$  of  $\mathcal{D}$  the Hellinger transform may be expressed as:

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

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 non negative 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  $x \rightarrow x/EX$  w.r.t.  $F$  and let  $G_1$  be the distribution having density  $y \rightarrow y/EY$  w.r.t.  $G$ .

Then the following conditions are equivalent:

- (i)  $F$  Lorenz majorizes  $G$ .
- (ii)  $E(X - cEX)^\pm / EX \geq E(Y - cEY)^\pm / EY$ ;  $c \in R$  where the  $\pm$  signs in exponential positions 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  $\varphi$  is convex on  $[0, \infty[$ .

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

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  $\beta$  then it is equivalent with  $(R[0, 1], \beta)$  and thus  $\|(P_0 \times P_1) \wedge (P_1 \times P_0)\| = \iint [\beta'(\alpha_2) \wedge \beta'(\alpha_1)] d\alpha_1 d\alpha_2 = 2 \iint_{\alpha_1 > \alpha_2} \beta'(\alpha_2) d\alpha_1 d\alpha_2 = 2 \int_0^1 (1 - \beta(\alpha_1)) d\alpha_1$ . 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  $P_0 \gg P_1$ .

---

## 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.*
- Karlin, S. 1983. Comparison of measures, multivariate majorization, and applications to statistics. In: *Karlin, Amemija, Goodman. Studies in econometric time series and multivariate statistics, Academic Press.*
- LeCam, L. 1964. Sufficiency and approximate sufficiency. *Ann. Math. Statist.* 35, 1419-1455.
- LeCam, L. 1986. Asymptotic methods in statistical decision theory. *Springer-Verlag.*
- 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.*
- Strassen, V. 1965. The existence of probability measures with given marginals. *Ann. Math. Statistics*, 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. Wahrscheinlichkeitstheorie Verw. Geb.* 16, 219-249.
- Torgersen, E. 1976. Comparison of statistical experiments. *Scand. Journ. of Statistics* 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.*