Notes on Decision Theory

by

Herman Rubin
Department of Statistics
Purdue University
West Lafayette, IN 47907

Technical Report # 89–30

Department of Statistics
Purdue University

August 1989

△

NOTES ON DECISION THEORY

by

Herman Rubin

Department of Statistics

Purdue University

West Lafayette, IN 47907

ABSTRACT

1. The General Decision Problem

Our problem has the following elements:

A parameter space Ω;

An action space A and a σ-field $\mathcal{A}$ on A;

A "sample space" Z on which all possible sample values are defined, a σ-field $\mathcal{B}$ on Z, and a probability measure ρ_ω on $\mathcal{B}$ for each $\omega \in Omega$.

A loss function; $L : \Omega \times A \times Z \to R$ (extended real line).

A decision procedure is a measurable function: $d : Z \to A$.

A mixed strategy is a measure ν on $Z \times \mathcal{A}$ such that $\nu(\cdot|z)$ is a probability measure on $\mathcal{A}$ and $\nu(E|\cdot)$ is a measurable function from Z to R for each $E \in \mathcal{A}$.

The risk function is the function ρ defined by

$$\rho(\omega, \nu) = \int \int L(\omega, a, z) d\nu(a|z) d\rho(z|\omega).$$

The above symbols will have these meanings throughout except where otherwise stated.

We first give an extension of a well-known game theory result.

Theorem. Let G be a game, X, Y the sets of strategies for players I and II respectively, and M the payoff function. Then if

(i) X and Y are convex and X is compact;

(ii) M is concave in X, convex in Y, and upper semi-continuous in X for each $y \in Y$, i.e.,

$M(x,y) \leq \inf_{x \in U \text{ open}} \sup_{z \in U} M(z,y)$, for each y;

(iii) $\gamma_G > -\infty$ where γ_G is $\sup_x \inf_y M(x,y)$;

then the game has a value and player I has a good strategy, i.e., $\max_n \inf_y M(x,y) = \inf_y \max_x M(x,y)$.

Proof Let

$$\mathcal{F} = \{F : F \subseteq Y, F \text{ is finite}, F \neq \phi\}$$

$$\mathcal{G} = \{G : G \subseteq X, G \text{ is finite}, G \neq \phi\}.$$

Consider the finite game (G, F, M^*), where M^* is the mixed extension of M. This game has a value, and both players have good strategies.

Hence, there are procedures ξ_{GF} and η_{GF} such that

$$\int_G M(x,y) d\xi_{GF}(x) \geq v_{GF} \geq \int_F M(x,y) d\eta_{GF}(y)$$

for all $y \in F, x \in G$, where v_{GF} is the value of the game (Gm, F, M^*).

Now $\mathcal{G}$ and $\mathcal{F}$ are directed by inclusion (The least upper bound of A and B is $A \cup B$) and, for any $F, v_{G_1 F} \leq v_{G_2 F}$ whenever $G_1 \leq G_2$.

Fix F and consider the net $\{(\int x d\xi_{GF}(x), \eta_{GF})\}$ where $\int x d\xi_{GF}(x) = \sum_{x_i \in G} (\xi_{GF})_i x_i \in X$ since X is convex. Now η_{GF} is a vector in the set $\{\alpha : \alpha_i \geq 0 \text{ for all } i \text{ and } \Sigma \alpha_i = 1\}$ in the euclidean space with dimension $\overline{\overline{F}}$, cardinality of F.

Hence the net has each coordinate in a compact space, so there is a convergent subnet $\{(\int x d\xi_{G_2 F}(x), \eta_{G_\alpha F})\}$ such that

$$\int x d\xi_{G_\alpha F}(x) \longrightarrow x_F^* \in X$$

$$\eta_{G_\alpha F} \longrightarrow \eta_F^*$$

Now M is concave and so

$$M(\int x d\xi_{GF}(x), y) = M(\sum_{x_i \in G} (\xi_{GF})_i x_i, y)$$

$$\geq \sum_{x_i \in G} (\xi_{GF})_i M(x_i, y) = \int_G M(x, y) d\xi_{GF}(x) \geq v_{GF} \text{ for all } y \in F.$$

Also $\overline{\lim} M(\int x d\xi_{G_\alpha F}(x), y) \leq M(x_F^*, y)$ since M is upper semi-continuous.

Hence for all $y \in F$,

$$M(x_F^*, y) \geq \overline{\lim} M(\int x d\xi_{G_\alpha F}(x), y)$$

$$\geq \overline{\lim} v_{G_\alpha F}$$

$$= \sup_G v_{GF},$$

since for any G there is an α such that $G \leq G_\alpha$ and so $v_{GF} \leq v_{G_\alpha F}$.

Since the net indexed by G_α is a subnet of the net indexed by G, and since $x_F^* \in X$, we have $x_F^* \in G_\alpha$ for α sufficiently large.

So if α is sufficiently large

$$\int M(x_F^*, y) d\eta_{G_\alpha F}(y) \leq v_{G_\alpha F} \leq \sup_G v_{GF}.$$

But $\int M(x_F^*, y) d\eta_{G_\alpha F}(y) = \Sigma M(x_F^*, y) \eta_{G_\alpha F}\{y\}$ converges to $\Sigma M(x_F^*, y) \eta_F^*\{y\}$ since $\eta_{G_\alpha F} \to \eta_F^*$.

Hence $\int M(x_F^*, y) d\eta_F^*(y) \leq \sup_G v_{GF}$. So if player II is restricted to F, the game has a value and both players have good strategies.

Now let $y_F = \int y d\eta_F^*(y) \in Y$ since Y is convex and then $M(x_F^*, y_F) \leq \sup_G v_{GF} = v_F$, since M is convex in Y. So we have $M(x_F^*, y) \geq v_F$ for all $y \in F$ and $M(x, y_F) \leq v_F$ for all $x \in X$. Note that this last inequality implies $v_F \geq \gamma_G$.

The net $\{x_F^*, F \in \mathcal{F}\}$ must have a cluster point since X is compact, so there is a subnet $\{x_{F_\beta}^*\}$ such that $x_{F_\beta}^* \to x'$, say.

By the upper semi-continuity of M, for any $y \in Y, M(x', y) \geq \overline{\lim} M(x_{F_\beta}^*, y)$. But if $y \in F_\beta, M(x_{F_\beta}^*, y) \geq v_{F_\beta}$. Since for β sufficiently large $y \in F_\beta, M(x', y) \geq \lim v_{F_\beta} = \inf v_{F_\beta}$. However, $\inf_y M(x', y) \leq \inf_F M(x', y_F) \leq \inf_F v_F$. Consequently, the game has a value and x' is a good strategy for player I.

learly, if both X and Y are compact, and M is lower semi-continuous in y for each fixed x, then both players have good strategies.

This result is due to Maurice Sion, 'On general minimax theorems' *Pac. J. of Math.*, 8, (1958), 171–176.

Principles of Choice

This section is a brief digression to give an axiomatic foundation for principles which we shall use later.

Suppose A is a convex class of actions – $a, b, \in A \to \alpha a + (1-\alpha)b \in A$ for all $\alpha \in [0, 1]$.

Axiom 1: There is a class $\mathcal{A}$ of non-empty subsets of A, including all subsets with three elements or less, and a function C mapping each element E of $\mathcal{A}$ into a non-empty subset of its convex hull, $H(E)$.

Axiom 2: If $F \leq H(E)$ and $H(F) \cap C(E) \neq \phi$ then $C(F) = H(F) \cap C(E)$.

Axiom 3: $C(\alpha a + (1-\alpha)E) = \alpha a + (1-\alpha)C(E)$.

The function C is a choice function, i.e., a method for choosing the best acts from a

set E of possible acts. Axiom 2 says, roughly, that if we reduce the set of allowable acts, we will choose the same acts as before which are available whenever there are any. Axiom 3 states, roughly, that the treatment of a problem should be independent of the probability that the problem will be faced.

These axioms are sufficient to prove $C(E)$ convex:

(i) Suppose $E = \{a, b\}$ and $c = \alpha a + (1 - \alpha)b \in C(E), 0 < \alpha < 1$. Let $F = \{a, c\} = \alpha a + (1 - \alpha)E$. Then by Axiom 2, $c \in C(F)$; hence $c \in \alpha a + (1 - \alpha)C(E)$ by Axiom 3, so that $b \in C(E)$. Similarly $a \in C(E)$. So if $C(E)$ contains any point of $[a, b]$ other than an endpoint, it contains both endpoints.

For any $d = \beta a + (1 - \beta)b, 0 < \beta < 1$ let $G = \{a, d\} = \beta a + (1 - \beta)E$ so $\{a, d\} \subseteq C(G)$ by Axiom 3.

Hence, by Axiom 2, $\{a, d\} \subseteq C(E) \cap H(F)$ so $d \in C(E)$ i.e., if $E = \{a, b\}$ then $C(E) = \{a\}$ or $\{b\}$ or $\{d : d = \alpha a + (1 - \alpha)b. 0 \leq \alpha \leq 1\}$.

(ii) Suppose $E = \{a, b, c\}$, and $C(E)$ contains some interior point x where $x = \alpha_1 a + \alpha_2 b + \alpha_3 c, \alpha_1 + \alpha_2 + \alpha_3 = 1, \alpha_i > 0$. Then there is a

$$d = \beta b + (1 - \beta)c, 0 < \beta < 1 \text{ such that}$$

$$x = \alpha a + (1 - \alpha)d, 0 < \alpha < 1.$$

Then by Axiom 2, $x \in C\{a, d\}$ and by the argument in (i), $d \in C(E)$, and by (1) again, $[b, c] \in C(E)$. Proceeding in this way, we can obviously show $C(E) = H(E)$ unless $C(E)$ is either a corner point (a, b or c) or a boundary line of E.

We now introduction a partial ordering ("preference ordering") of elements of A.

$$\text{Define: } a \geq b \text{ if } a \in c\{a,b\}$$

$$a > b \text{ if } \{a\} = c\{a,b\}$$

$$a \sim b \text{ if } [a,b] = c\{a,b\}.$$

Note that $\sim$ is an equivalence relation and also that $a \sim b \Rightarrow \alpha c + (1-\alpha)a \sim \alpha c + (1-\alpha)b$ for all $c \in A$ and all $\alpha \in [0,1]$, because

$$\begin{aligned} C\{\alpha c + (1-\alpha)a, \alpha c + (1-\alpha)b\} &= C(\alpha c + (1-\alpha)\{a,b\}) \\ &= \alpha c + (1-\alpha)C\{a,b\} \\ &= \alpha c + (1-\alpha)(\beta a + (1-\beta)b) 0 \leq \beta \leq 1 \\ &= [(\alpha c + (1-\alpha)a), (\alpha c + (1-\alpha)b]. \end{aligned}$$

So $\sim$ is a substitution relation.

If $a \geq b$ and $b \geq c$, consideration of $C\{a,b,c\}$ shows easily that $a \geq c$.

Similarly if $a > b$ and $b > c$ then $a > c$.

If we add a further axiom:

Axiom 4: If $a > b > c$, there is an $\alpha \in (0,1)$ with $b \sim \alpha a + (1-\alpha)c$, we have enough for the derivation of the von Neumann result: there is a real valued linear function on A such that $a \geq b$ iff $f(a) \geq f(b)$.

Now let Ω be the class of possible states of nature, and assume that for each $\omega \in$ *Omega* there is a choice function C_ω on subsets of A satisfying the Axioms 1–4. Furthermore, let Axioms 1–4 hold for A and also

Axiom 5: If for all $\omega \in \Omega, a \geq_\omega b$, then $a \geq b$.

Then there is a relation $\geq_\omega$ on A, and a function $g_\omega(\cdot)$ such that $g_\omega(a) \geq g_\omega(b)$ iff $a \geq_\omega b$, for each $\omega \in \Omega$. So to each $a \in A$ corresponds a function $g_\omega(a)$ on Ω.

Then we can set up a linear functional L_1 so that we have a real valued function $f(a) = L(g_\cdot(a))$ where $L(\alpha h_1 + (1-\alpha)h_2) = \alpha L(h_1) = (1-\alpha)L(h_2)$ and we can say $a \leq b$ if $L(g_\cdot(a)) \leq L(g_\cdot(b))$. L will be the expected value of $g_\cdot(a)$ with respect to some prior distribution.

As we have just seen, under Axioms 1–5, any consistent procedure is to maximize expected utility. Unfortunately, these axioms are valid only if one has an infinitely fast, infinitely large, costless computing machine available. For example, let it be desired to estimate the number of redheads in a student body of 30,000. Suppose we are given the names, school addresses, home addresses, sex, and marital status of each, and we are allowed to sample 500 students. We certainly have some useful information; but to carry out the procedure, assuming symmetry in the utilities given the state of nature, would require us to first list the coefficients of L ("prior probabilities") for each of the $2^{30,000}$ sets of students. Since electronic computers currently have memories on the order of 2^{20}, and it is estimated that the number of particles in the universe is on the order of 2^{250}, we see that this is out of the question.

There are a number of methods proposed for selecting "best" decisions:

Unbiasedness.

This principle is applicable in the two-action case, $A = \{a_1, a_2\}$, when one of the actions, say a_1, is distinguished from the other. If $\Omega = \Omega_1 \cup \Omega_2$, a disjoint union, and we

want to take action a_1 if $\omega \in \Omega_1$ and a_2 if $\omega \in \Omega_2$, then a decision function d is said to be unbiased if $P_{\theta_2}(d(z) = a_2) \geq P_{\theta_1}(d(z) = a_2)$, for all $\theta_2 \in \Omega_2, \theta_1 \in \Omega_1$.

More generally, if for each $\theta \in \Omega$ there is a unique correct act, each act is correct for some $\theta \in \Omega_1$ and $L(\theta_2, d) = L(\theta_1, d)$ for all d if the same act is correct for both θ_1 and θ_2, then a procedure ν is unbiased if, for all $\theta, \theta' \in \Omega$,

$$\int \int L(\theta', a) d\nu(a|z) dP(z|\theta) \geq \int \int L(\theta, a) d\nu(a|z) dP(z|\theta),$$

i.e. "on the average", the procedure ν gives a smaller loss for the true θ than for any other $\theta' \in \Omega$, that is, it comes "closer", on the average, to the correct decision than to any wrong one.

The following example, due to Hodges and Lehmann, shows that this is not a good rule. Let X be distributed with density $\frac{1}{\Gamma(\alpha)} \frac{x^{\alpha-1}}{\theta^\alpha} e^{-x/\theta}$, α known, and let the loss of announcing d be $(\theta - d)^2$. Now $E(X) = \alpha\theta$ and $V(X) = \alpha\theta^2$. The estimate $d(X) = \frac{x}{\alpha+1}$ is not unbiased and has risk $\frac{\theta^2}{\alpha+1}$, while the best unbiased estimate $\frac{x}{\alpha}$ has risk θ^2/α.

Maximum Likelihood.

Broadly, this procedure chooses the act which minimizes the loss for that θ which is most likely to yield the observed result, e.g. in the discrete case, with $\Omega = \Omega_1 \cup \Omega_2, A = \{a_1, a_2\}$, we look at

(1) $\frac{\sup_{\theta \in \Omega_1} p_\theta\{x\}}{\sup_{\theta \in \Omega_2} p_\theta\{x\}}$, choosing a_1 if this is $> 1, a_2$ otherwise. Though this is intuitively a very reasonable procedure, there are examples where it gives poor results, e.g. $F = \Omega =$

$\{(n,i) : n = 1,2,\ldots N; i = 1,2\}$

$$A = \{1,2\}, P[\{(m,j)\} : (n,i)] = \alpha/N \text{ if } i = j$$

$$= (1-\alpha)\delta_{mn} \text{ if } i \neq j.$$

$L[(n,i),a] = 1 - \delta_{ia}$. So $\Omega_1 = \{(n,i) : i = 1\}, \Omega_2 = \{(n,i) : i = 2\}$.

Then $\theta \in \Omega_1\ P_\theta(\{(n,1)\}) = \alpha/N$, $\sup_{\theta\in\Omega_2} P_\theta(\{(n,1)\}) = 1-\alpha$.

Hence the likelihood ratio, (1), is $\frac{\alpha}{N(1-\alpha)}$.

But if, say, $\alpha = .9999$ and $N = 10^{20}$ then $\frac{\alpha}{N(1-\alpha)} \simeq 10^{-15}$. So if the 2nd coordinate is 1, the maximum likelihood procedure tells us to choose act a_2; similarly if the 2nd coordinate is 2, we would choose act a_1.

The probability of error in this procedure is .9999, i.e. It is much better to do the opposite.

Minimax risk. Here it is assumed that the $g_\omega(a)$ can be compared. Then choose a to minimize $\sup_\omega g_\omega(a)$. This procedure gives too much weight to catastrophes; for example, if g is

	a_1	a_2
ω_1	0	999
ω_2	1000	999,

a_2 should be preferred. This leads to the suggestion of

Minimax regret. Define $h_\omega(a) = g_\omega(a) - \inf_b g_\omega(b)$. This leads to the following paradox: there can be actions a, b, c, d so that a is the choice, but of a, b, c, the action b will be chosen.

Because of the difficulties of deciding what is the best act, it is often useful to try to reduce the number of acts available. There are two ways to doing this: we can try to

eliminate those acts which are clearly not best, and in a wide range of problems which exhibit certain symmetries we can try to reduce consideration to those acts which also exhibit these symmetries.

The first of these procedures leads us to the notions of admissibility and completeness; the second to the notion of invariance.

We will later give an example to show the lack of general desirability of invariance; this renders impossible an attempt at "objectivity".

2. Classes of Actions: Completeness and Admissibility.

Definition A procedure d is admissible iff

$$\rho(\theta, d') \leq \rho(\theta, d) \text{ for all } \theta \text{ implies } \rho(\theta, d') = \rho(\theta, d) \text{ for all } \theta.$$

(See §1 for notation). Let $\mathcal{I}$ be the class of admissible procedures. Unfortunately, $\mathcal{I}$ may be empty, so we need to define a wider class of "good" procedures.

Definition A class C of procedures is complete iff, for all $d \notin C$, there is a $d' \in C$ such that

(i) $\rho(\theta, d') \leq \rho(\theta, d)$ for all θ and

(ii) $\rho(\theta_0, d') < \rho(\theta_0, d)$ for some $\theta_0 \in \Omega$.

If only (i) holds, C is "essentially complete". It is easy to see that

(a) The intersection of any finite number of complete classes is complete;

(b) The admissible class is contained in every complete class;

(c) The intersection of all complete classes is the admissible class.

We are interested in the question: under what circumstances is the admissible class complete?

Some of the conditions which imply completeness are given in Blackwell and Girshick, Theory of Games and Statistical Decisions (Wiley, 1954). In particular, Theorem 5.7.1 gives conditions under which the class of admissible procedures is complete. We give here an extension:

Theorem If every (transfinite) sequence of procedures, $\{\nu_\alpha\}$, which is decreasing in the sense that $\rho(\theta, \nu_\gamma) \geq \rho(\theta, \nu_\eta)$ (with strict inequality holding for at least one θ) whenever $\gamma < \eta$, and which is of length less than $\mathcal{N}_{\alpha+1}$, has a lower bound, and either

(a) Ω has a dense subset of cardinality $< \mathcal{N}_{\alpha+1}$ and all risk functions $\rho(\theta, \nu)$ are continuous in θ, or

(b) Every subset of Ω has a dense subset of cardinality less than $\mathcal{N}_{\alpha+1}$ and all risk functions of lower semi-continuous, then the class of admissible procedures is complete.

Proof. We note that if ρ is the risk, then $\rho^* = \frac{\rho}{1+|\rho|}$ preserves the order properties of ρ, but $|\rho^*| \leq 1$. Hence we can assume ρ bounded. In the following, we will identify any procedure ν with the function $\nu(\theta) = \rho(\nu, \theta)$; and we will write $\nu < \eta$ if $\rho(\nu, \theta) \leq o(\eta, \theta)$ for all $\theta \in \Omega$, with strict inequality holding for at least one $\theta \in \Omega$.

If the admissible class is not complete, there is an inadmissible procedure ν_0, say, such that $\nu < \nu_0$ implies ν is inadmissible. We will show that this leads to a contradiction.

Let Q be a dense subset of Ω such that the cardinality of Q, $\overline{\overline{Q}}$, is less than $\mathcal{N}_{\alpha+1}$.

Then if R is the set of rationals, $\overline{\overline{Q \times R}} = \lambda < \mathcal{N}_{\alpha+1}$. We well-order the elements (q_β, r_β) of $Q \times R$.

If the theorem is false, we can establish a (transfinite) sequence, $\nu_0, \nu_1, \ldots, \nu_\beta, \ldots$, such that

$$\nu_{\beta+1} < \nu_\beta \text{ and, in fact, if } M)q_\beta) = \inf\{\eta(q_\beta) : \eta < \nu_\beta\}$$

we can choose $\nu_{\beta+1}$ from $\{\eta : \eta < \nu_\beta\}$ so that $\nu_{\beta+1}(q_\beta) < M(q_\beta) + r_\beta$. Then $\eta(q_\beta) > \nu_{\beta+1}(q_\beta) - r_\beta$ for all $\eta < \nu_\beta$. If β is a limit ordinal, we have $\nu_\beta \leq \nu_\gamma$ for all $\gamma < \beta$.

Since the sequence β has cardinality less than $\mathcal{N}_{\alpha+1}$, the sequence has a lower bound ζ. That is

$$\zeta \leq \nu_\beta \text{ for all } \beta < \lambda.$$

Further, if $\eta < \zeta$, then $\eta < \nu_\beta$ for all β and hence, for any $q \in Q$,

$$\eta(q) \geq \nu_{\beta+1}(q) - r_\beta \text{ for every } \beta \text{ for which } q_\beta = q$$

$$\geq \zeta(q) - r_\beta \text{ for every } \beta \text{ for which } q_\beta = q.$$

Since $\inf\{r_\beta : q_\beta = q\} = 0, \eta(q) = \zeta(q)$.

Thus ζ cannot be improved on Q and, since Q is dense in Ω, ζ cannot be improved on any open set in Ω. If the risk functions are all continuous, $\eta(\omega) < \zeta(\omega)$ implies $\eta < \zeta$ on an open set, which provides the required contradiction: ζ must be an admissible procedure.

For the case when the risk functions are lower semi-continuous, we define $f^*(\theta) = \inf_{\theta \in U \text{ open}} \sup_{\varphi \in U} f(\varphi)$ for any function f. Clearly f^* is upper semi-continuous.

Remark: If $f \geq g, f$ is lower semi-continuous, and $f^* \neq g^*$, then $f - g > 0$ on some open set.

For there exists an $\varepsilon > 0$ and a θ such that $f^*(\theta) > g^*(\theta) + 3\varepsilon$, and

(i) There is an open set U such that, for all $\varphi \in U$,

$$g(\varphi) < g^*(\theta) + \varepsilon, \text{ by definition of } g^*;$$

(ii) There is a $\varphi_0 \in U$ such that

$$f(\varphi_0) > f^*(\theta) - \varepsilon, \text{ by definition of } f^*;$$

(iii) There is an open neighbourhood V of φ_0 such that, for all $\Psi \in V$,

$$f(\Psi) > f(\varphi_0) - \varepsilon \text{ since } f \text{ is lower semi-continuous.}$$

Hence if $\Psi \in U \cap V$, open, $f(\Psi) > f(\varphi_0) - \varepsilon > f^*(\theta) - 2\varepsilon > g^*(\theta) + \varepsilon > g(\Psi)$.

Now let the ζ obtained above be ζ_0, and suppose we can get a sequence of procedures $\zeta_0, \zeta_1, \ldots \zeta_\beta, \ldots, \beta < \omega_{\omega+1}$, such that $\zeta_{\beta+1} < \zeta_\beta$ for all β.

We consider the upper semi-continuous functions $\zeta^* - \zeta_\beta$ and the closed sets

$$E_\beta(r) - \{\theta : \zeta^* - \zeta_\beta \geq r\}.$$

Clearly $E_\beta(r) \subseteq E_{\beta+1}(r)$ for all β. Let $E(r) = \bigcup_\beta E_\beta(r)$. By assumption, $E(r)$ has a dense subset $Z(r)$ such that $\overline{\overline{Z(r)}} < \mathcal{N}_{\alpha+1}$. If $q \in Z(r)$, the set $\{\beta : q \in E_\beta(r)\}$ is non-empty, so has a least element, say β_q. The set $\{\beta_q : q \in Z(r)\}$ has cardinality $< \mathcal{N}_{\alpha+1}$ and so has an upper bound γ_r, with $\gamma_r < \mathcal{N}_{\alpha+1}$.

Since $E_{\beta_q}(r)$ is increasing and $\beta_q < \gamma_r, q \in Z(r)$ implies $q \in E_{\gamma_r}(r)$. Hence $E_{\gamma_r}(r)$ is a closed, dense subset of $E(r)$. So $E_{\gamma_r}(r) = E(r)$. The set $\{\gamma_r : r \text{ is rational}\}$ is countable, and therefore has an upper bound δ, since no sequence is cofinal in $\omega_{\alpha+1}$.

Then for every $r, E(r) = E_{\gamma_r}(r) \subseteq E_\delta(r)$. Now for any real number a, $E_\beta(a) = \bigcap_{\substack{r<a \\ r \text{ rational}}} E_\beta(r) \subseteq \bigcap_{\substack{r<a \\ r \text{ rational}}} E_\delta(r) = E_\delta(a)$. Hence $E(a) = E_\delta(a)$.

Thus for any $\beta < \omega_{\alpha+1}$ and any $a, \{\theta : \zeta^* - \zeta_\beta \geq a\} \subseteq \{\theta : \zeta^* - \zeta_\beta \geq a\}$. Hence $\zeta_\beta \geq \zeta_\delta$, so that $\zeta_{\delta+1} \geq \zeta_\delta$ which contradicts the construction of the sequence $\{\zeta_\beta\}$.

Hence the sequence $\{\zeta_\beta\}$ must terminate with an admissible strategy ζ_δ, and the class of admissible procedures is complete.

Invariance

A decision problem is invariant if there is a group $\mathcal{G}$ of ordered triples of transformation preserving the problems.

That is, we assume that if $g \in \mathcal{G}, g = (g_\Omega, g_Z, g_Z), g_\Omega, g_Z, g_A$ are 1–1 mappings of Ω, Z, and A, respectively, onto themselves, g_Z and g_A are measurable, and $L(g_\Omega, (\theta), g_Z(z), g_A(a)) = L(\theta, z, a)$ and $P(g_Z(B)|g_\Omega(\theta)) = P(B|\theta)$ for all $\theta \in \Omega, z \in Z, a \in A, B \in \mathcal{B}$, and $g \in \mathcal{G}$. Furthermore $(gh^{-1})_\Omega = g_\Omega - (h_\Omega)^{-1}$, etc.

A decision procedure ν is said to be invariant if, for all $g \in \mathcal{G}, E \in \mathcal{A}, z \in Z, \nu(g_A^{-1}(E)|g_Z^{-1}(z)) = \nu(E|z)$.

Again, it seems intuitively reasonable that invariant problems should have reasonable invariant solutions. The following example shows that this is not always the case.

Stein's Example.

Let $\Omega = \{(\Sigma, \alpha) : \alpha = 10^{10}$ or 10^{-10} and Σ is a positive definite 2×2 matrix$\}$. $Z = \{(z_1, z_2) : z_1, z_2 \in E_2$, and z_1, z_2 are linearly independent$\}$. $P_{(\Sigma,\alpha)}$ is given by $z_1 \sim N(0, \Sigma), z_2 \sim N(0, \alpha\Sigma)$ and z_1, z_2 are independent.

$$A = \{10^{10}, 10^{-10}\}; L((\Sigma, \alpha),\ z, a) = \begin{matrix} 0 & \text{if } \alpha = a \\ 1 & \text{if } \alpha \neq a \end{matrix}$$

$G = \{$all 2×2 non-singular matrices$\}$ with, for $g \in G, g_\Omega(\Sigma, \alpha) = g\Sigma g', \alpha); g_A(a) = a; g_Z(z_1, z_2) = (gz_1, gz_2)$.

The only invariant procedure is to guess, since under G any result (z_1, z_2) is equivalent to any other result (z_3, z_4). In particular, (z_1, z_2) is equivalent to (z_2, z_1). hence the probability of error is $\frac{1}{2}$.

However the non-invariant procedure that consists of looking at the first coordinates of z_1 and z_2, z_{11} and z_{21}, and chooses "$\alpha = 10^{10}$" if $|z_{11}| < |z_{21}|$ has an error probability of $\frac{2}{\pi}$ artan 10^{-10}, since $\frac{z_{11}}{z_{21}}$ has the Cauchy distribution with density $\frac{\alpha}{\pi(\alpha^2 + x^2)}$.

There have been a number of investigations of conditions under which invariant procedures give "good" results. The most important of these is the generalized Hunt-Stein Theorem, which we discuss only briefly, referring the reader to Wesler, A.M.S., Vol. 30 (1959), pp. 1–20, for details.

We let C be a Borel field of subsets of $\mathcal{G}$, and assume that $(g, z) \to g_Z(z)$ is $C \times \mathcal{B} - \mathcal{B}$ measurable and $(g_1, g_2) \to g_1 g_2$ is $C^2 - C$ measurable. We say that a measure μ on $(\mathcal{G}, C)$ is right invariant (left invariant) if $\mu(Cg) = \mu(C)(\mu(gC) = \mu(C))$ for every $g \in \mathcal{G}, C \in C$; and that a net $\{\mu_\alpha\}$ of measures on $(\mathcal{G}, C)$ is asymptotically right invariant if

$\lim(\mu_\alpha(Cg)) - \mu(C) = 0$ for every $g \in \mathcal{G}, C \in C$.

For example, if $\mathcal{G}$ is the additive group of reals, C the Borel sets, and μ the Lebesgue measure on the reals, then μ is both right and left invariant. Although no invariant probability measure exists for this group, the sequence of measures $\{\mu_n\}$ given by $\mu_n(C) = \mu(C \cap [-n, n])/2n$, is asymptotically invariant (in fact uniformly so).

(i) The testing problem. Let $\Omega = \Omega_0 \cup \Omega_1$, a disjoint union, and suppose we wish to test $H_0 : \omega \in \Omega_0$ against $H_1 : \omega \in \Omega_1$. We assume that there is a σ-finite measure ν such that $P_\omega << \nu$ for all $\omega \in \Omega$, and $\frac{dP_\omega}{d\nu} = p(z/\omega)$.

The group $\mathcal{G}$ keeps the testing problem invariant iff $\omega \in \Omega_i$ implies $g_\Omega(\omega) \in \Omega_i, i = 0, 1$, for every $g \in \mathcal{G}$. A randomized test is a $\mathcal{B}$-measurable real-valued function $\Phi : Z \to [0, 1]$ where $\Phi(z)$ is the probability of rejecting H_0 when z is observed.

The test Φ is invariant if $\Phi(g_Z(z)) = \Phi(z)$ except on a set of ν measure zero, for all $g \in \mathcal{G}$. If the exceptional null set depends on g, Φ is almost invariant.

Then the Hunt-Stein Theorem is:

If, for the above testing problem, there is an asymptotically right invariant sequence $\{\mu_n\}$ of probability measures on $(\mathcal{G}, C)$, then there exists an invariant test Φ_0 among all tests Φ for which $\int \Phi(z)p(z|\omega)d\nu(z) \leq \alpha$ for all $\omega \in \Omega_0$, which maximizes $\inf_{\omega \in \Omega} \int \Phi(z)$ $p(z|\omega)d\nu(z)$; i.e., there is an invariant minimax test at any level α.

(ii) For the "estimation" case, we observe firstly that if $P_\omega << \nu'$ for all ω, then there is a ν' such that $P_\omega << \nu'$ for all and $P_\omega(B) = 0$ for all ω implies $\nu'(B) = 0$. (Reference given by Wesler).

We write $\Omega = \bigcup_{s \in S} \Omega_s$ where the Ω_s are sets of the form $\{\omega : \omega = g(s)$ for some $g \in \mathcal{G}, s \in \Omega\}$, and $S \subseteq \Omega$ serves as an index set. (It is obvious that $s \in \Omega_S$, and $\Omega_S \cap \Omega_t \neq \phi$ implies $\Omega_s = \Omega_t$). We define a strategy Φ to be at least as good as a strategy Ψ in the modified minimax sense if $\sup_{\omega \in \Omega_s} \rho(\omega, \Phi) \leq \sup_{\omega \in \Omega_s} \rho(\omega, \Psi)$ for all $s \in S$. A class C of strategies is called essentially complete in the modified minimax sense if, for every $\Psi \notin C$, there is a $\Phi \in C$ which is at least as good as Ψ in the modified minimax sense.

Then we have:

The Generalized Hunt-Stein Theorem.

If $(Z, \mathcal{B}, \Omega, P, A, \mathcal{A}, L)$ is a statistical problem invariant under a group $\mathcal{G}$, and

(i) There exists a measure ν for which the Radon-Nikodym Theorem holds with respect to which all P_ω are absolutely continuous,

(ii) A is separable metric, $calA$ is the Borel field generated by the compact subsets of A, and L is such that, for each $\omega \in \Omega, L(\omega, a)$ is non-negative and lower semi-continuous in a and, for every real number τ, the set $\{a : L(\omega, a) \leq \tau\}$ is compact, and

(iii) C is a Borel fields of subsets of $\mathcal{G}$ and there is an asymptotically right invariant net of probability measures, $\{\mu_\alpha\}$, on $(\mathcal{G}, C)$. Then the class Φ^* of almost invariant decision procedures is essentially complete in the modified minimax sense. If in addition,

(iv) $\mathcal{G}$ is a locally compact, σ-compact, topological group with C generated by the compact subsets of $\mathcal{G}$, then the class Φ^{**} of invariant procedures is essentially complete in this sense.

The important condition in this theorem, the condition which determines which groups satisfy the requirements for the Hunt-Stein Theorem, is that requiring the existence of an asymptotically right invariant net of probability measures on $\mathcal{G}$.

Peisakoff has shown that it is essential that the group be orgodic, and consequently the following groups are among those which satisfy the conditions for the theorem:

Groups which have ergodic normal sub-groups, ergodic factor groups, and sufficient continuity;

Groups which are the direct limits of ergodic groups, i.e. $\mathcal{G} = \bigcup_{\alpha \in A} \mathcal{G}_\alpha$ where $\mathcal{G}_\alpha$ is ergodic and A is directed; e.g. the group of those permutations of the positive integers which permute only a finite number of integers;

Abelian Groups;

Solvable Groups;

Compact Groups;

Connected Lie groups which are compact module their maximal solvable normal sub-group. All other connected Lie groups can be shown not to be ergodic. Stein's example, given previously, is a case of this.

There still remains the question of admissibility of variant procedures. If measurability is not imposed, Blackwell has shown that admissibility need not occur for the translation problem for squared error loss for distributions concentrated on a finite number of points. The best result for admissibility in the univariate translation parameter problem with

$L(\omega, a) = |\omega - a|^k$ is that $E(|X|^{k+1}) < \infty$ (Brown), in which case if the best invariant estimator is unique it is admissible almost everywhere. Farrell has shown that uniqueness is needed. If the P_ω are absolutely continuous on all orbits with respect to Lebesque measure, the "almost everywhere" can be deleted, but Fox and Rubin have given an example to show that continuity is not enough for $L(\omega, a) = |\omega - a|$. For testing, Lehmann and Stein have shown that if the optimal invariant test is unique, $E(|X|) < \infty$ is sufficient for admissibility, and for a very powerful definition of admissibility of confidence intervals, Joshi has shown that a first moment guarantees admissibility. Perng has shown that the additional moment required cannot be replaced by $1 - \varepsilon$ moments for any ε.

Stein (1960) has shown that for the multivariate normal distribution with squared error loss, admissibility occurs in 1 and 2 dimensions, but in no more. This result has been generalized by many others.

Sufficiency and Informativeness

For purposes of economy similar to those which suggest the use of concepts of invariance, admissibility and completeness, it is desirable to keep the sample space, Z, as "small" as possible; that is, it is desirable to be able to ignore those parts of the data that are uninformative about the true parameter θ, or are otherwise unhelpful in reducing the risk of a decision. Consequently we introduce the following definitions:

Informativeness. An experiment, z, is a quadruple, $\{Z, \mathcal{B}, \Omega, P_z\}$ where Z is a space of outcomes, $\mathcal{B}$ a σ-field of subsets of Z, Ω the parameter space, and P_z a function on $\mathcal{B} \times \Omega$ such that, for each $\omega \in \Omega, P_z(0|\omega)$ is a probability measure on $\mathcal{B}$. The experiment

$z = \{Z, \mathcal{B}, \Omega, P_z\}$ is more informative than $\mathcal{W} = \{W, \mathcal{D}, \Omega, P_{\mathcal{W}}\}$("$\mathcal{W} \subset z$") if for every action space and every loss function, any risk function attainable from a decision procedure on $\mathcal{W}$ is also attainable from a decision procedure on z.

Sufficiency. To introduce this concept we need to discuss briefly the idea of conditional expectation. Let $(\mathcal{H}, \mathcal{A}, \mu)$ be a measure space. Let $\mathcal{A}_0$ be a sub-σ-field of $\mathcal{A}_1$ and let f be a non-negative, integrable function. Then, since $\int_A f d\mu$ exists for all $A \in \mathcal{A}$ and hence for all $A_0 \in \mathcal{A}_0$, under suitable restrictions the Radon-Nikodym theorem asserts that there is an almost-everywhere defined a function f_0, integrable $(\mathcal{A}_0, \mu)$ (i.e. $\mathcal{A}_0$ measurable and μ integrable), such that $\int f d\mu = \int_{A_0} f d\mu$ for all $A_0 \in \mathcal{A}_0$. We call f_0 the conditional expectation of f with respect to $\mathcal{A}_0$ and μ and write $f_0 = E(f|\mathcal{A}_0)$. If f is integrable, but not non-negative, we define $E(f|\mathcal{A}_0) = Ef^+|\mathcal{A}_0) - E(f^-|\mathcal{A}_0)$.

The restrictions on $(\mu, \mathcal{A}_0)$ are satisfied automatically if μ is a finite measure. Since μ is usually a probability measure, there is usually no difficulty. They can be summarized as follows: For every set of positive measure in $\mathcal{A}_0$, there is a subset positive finite measure in $\mathcal{A}_0$, and any family of sets in $\mathcal{A}_0$ has a μ-least upper bound in $\mathcal{A}_0$, i.e., if $\mathcal{H} \subset \mathcal{A}_0$ there is an $H \in \mathcal{A}_0$ such that

(a) if $A \in \mathcal{H}$, then $\mu(A \sim H) = 0$,

(b) if $K \in \mathcal{A}_0$ and $A \in \mathcal{H}$ imply $\mu(A \sim K) = 0$, then $\mu(K \sim H) = 0$.

A sub-σ-field $\mathcal{A}_0$ of $\mathcal{A}$ is defined to be sufficient for a family $\mathcal{P} = \{P_\omega : \omega \in \Omega\}$ of probability measures if $E_\omega(f|\mathcal{A}_0)$ can be defined to be independent of ω for every integrable function f on $(\mathcal{H}, \mathcal{A})$. (It is easy to see that only bounded functions need be considered).

A statistic T (i.e. an $\mathcal{A}$-measurable function $T : (\mathcal{H}, \mathcal{A}) \to (Y, \mathcal{I})$) is sufficient if the subfield it induces on $\mathcal{H}$ (i.e. the subfield $\mathcal{B}\{T^{-1}(D)_s : D \in \mathcal{D}\}$) is sufficient.

We note that it is possible that $\mathcal{A}_0$ and $\mathcal{A}_1$ are sub-σ-fields of $\mathcal{A}$, with $\mathcal{A}_0 \subset \mathcal{A}_1$, and $\mathcal{A}_0$ is sufficient while $\mathcal{A}_1$ is not. For if X is $R \times R$, where R is the real line, $\mathcal{P} = \{$all measures $p \times p$ where p is a probability measure on $R\}$.

$\mathcal{A}$ = Borel sets, T_0 = order statistics, $\mathcal{A}_0$ is the sub-σ-field induced by T_0, $T_1[(x_1, x_2)] = \begin{matrix}(x_1,x_2) \text{ if } (x_1,x_2)\in B \\ T_0(x_1,x_2) \text{ otherwise}\end{matrix}$ where B is not a Borel set and $(x_1, x_2) \in B \Rightarrow x_1 > x_2$, $\mathcal{A}_1$ is the sub-σ-field induced by T_1, then T_0 and $\mathcal{A}_0$ are sufficient but T_1 and $\mathcal{A}_1$ are not. This example is given by D.L. Burkholder, AMS 31, 1960, p. 232. However if $\mathcal{P}$ is domianted by a σ-finite measure, $\mathcal{A}_0 \subseteq \mathcal{A}_1$ and $\mathcal{A}_0$ is sufficient imply $\mathcal{A}_1$ is sufficient (Bahadur, AMS, 25, p. 440).

A slightly weaker concept than sufficiency is that of pairwise sufficiency: A sub-σ-field is pairwise sufficient for a family $\mathcal{P}$ if it is sufficient for each pair, $\{P_\theta, P_\omega\}$, of elements of $\mathcal{P}$.

Halmos and Savage (AMS, 20, 2, 1949, p. 236) give an example of a statistic which is pairwise sufficient but not sufficient. It is of interest, especially since pairwise sufficiency may be much easier to establish than sufficiency, to know what further restrictions, together with pairwise sufficiency, will ensure sufficiency.

To discuss this we need the notion of dominated sets of measures.

Definition

(1) If μ and ν are measures on a σ-field $\mathcal{S}$, μ is dominated by ν("$\mu << \nu$") if $E \in \mathcal{S}$ and $\nu(E) = 0$ implies $\mu(E) = 0$. If also $\nu << \mu$, ν is equivalent to μ.

(2) If $\mathcal{M}$ and N are sets of measures, $\mathcal{M}$ is dominated by $\mathcal{N}$("$\mathcal{M} << \mathcal{N}$") if $E \in \mathcal{S}$ and $\nu(E) = 0$ for all $\nu \in \mathcal{N}$ implies $\mu(E) = 0$ for all $\mu \ni \mathcal{N}$. If also $\mathcal{N} << \mathcal{M}, \mathcal{N}$ is equivalent to $\mathcal{M}$.

$\mathcal{M}$ is domianted by (equivalent to) ν if $\mathcal{M}$ is dominated by (equivalent to) $\{\nu\}$.

Halmos and Savage have proved the following two results:

(1) If $\mathcal{P}$ is a set of measures on $\mathcal{B}$, dominated by a σ-finite measure ν, then a sub-σ-field C of $\mathcal{B}$ is sufficient for $\mathcal{P}$ if and only if there is a σ-finite measure λ on $\mathcal{B}$ such that $\mathcal{P} << \lambda, \lambda << \mathcal{P}$ and $\frac{dP\omega}{d\lambda}$ (Radon-Nikodym derivative) is measurable with respect to C for each ω.

(2) If $\mathcal{P}$ is dominated by a σ-finite measure on $\mathcal{B}$, then T is sufficient if and only if T is pairwise sufficient.

We note that a sufficient condition for the sufficienty of C for $\mathcal{P}$, is that there exist a measure λ with respect to which the Radon-Nikodym theorem holds such that $\frac{d\mu}{d\lambda}$ exists and is C-measurable for every $\mu \in \mathcal{P}$. For if this is the case, and g is any integrable function, then if $g_\mu = E_\mu(g|C)$ and $g_\lambda = E_\lambda(g|C)$ where λ is the dominating measure, $F \in C$ iplies $\int_F g d_\mu = \int_F g_\mu d_\mu = \int_F g_\mu \frac{d\mu}{d\lambda} d\lambda$. But $\int_F g d_\mu = \int_F g \frac{d\mu}{d\lambda} d\lambda = \int_F g_\lambda \frac{d\mu}{d\lambda} d\lambda$. Since $\frac{d\mu}{d\lambda}$ is C-measurable. This holds for all g and all $F \in C$, so $g_\mu \frac{d\mu}{d\lambda} = g_\lambda \frac{d\mu}{d\lambda}[\lambda]$; since g_μ is arbitrary where $\frac{d\mu}{d\lambda} = 0$ we can take $g_\mu = g_\lambda$ everywhere; so g_μ is independent of μ, and hence C is sufficient.

We are now in a position to prove an extension of the "sufficient" assertion of (2):

Theorem Let C be a pairwise sufficient sub-σ-field for a family $\mathcal{P}$ of measures on a

space $(X, \mathcal{A})$. Then if every family of C-measurable sets has a common least upper bound with respect to all measures in $\mathcal{P}, C$ is sufficient for $\mathcal{P}$.

[Note: A least upper bound, with respect to P, of a family of $\mathcal{A}$ C-measurable sets is a C-measurable set E such that $P(F \sim E) = 0$ for every $F \in \mathcal{A}$ and that if $P(E \sim H) > 0$ then H does not have this property.

If $\mathcal{P}$ is equivalent to a finite measure λ, and $\mathcal{U}$ is any collection of C-measurable sets, let $S = \sup\{\lambda(V) : V$ is a finite union of elements of $\mathcal{U}\}$. For each n let U_n be a finite union of elements of $\mathcal{U}$ such that $\lambda(U_n) > S - \frac{1}{n}$. Let $U = \bigcup_{n=1}^{\infty} U_n$. Then $U \in C, \lambda(U) = S, V \in \mathcal{U}$ implies $\lambda(V - U) = 0$ since otherwise there is an n such that $\lambda(\bigcup_{i=1}^{n} U_i \cup V) > S$ contradicting the definition of S, and $\lambda(U \sim H) > 0$ iplies $\lambda(H \cap U) < \lambda(U)$ which means there is an n such that $\lambda(H \cap U_n) < \lambda(U_n)$ so that $\lambda(U_n \sim H) > 0$ so H cannot almost contain all members of $\mathcal{U}$. Hence U is a least upper bound of $\mathcal{U}$ with respect to λ, and hence is a least upper bound with respect to all measures in $\mathcal{P}$. Halmos and Savage show that if $\mathcal{P}$ is dominated by a σ-finite measure μ, it is equivalent to a finite measure λ. Hence the conditions of the Halmos-Savage theorem imply the conditions stated above.]

Fix the non-negative bounded measurable function f. Let $\mathcal{M} = \{(g, P) : g = E_p(f|C), P \in \mathcal{P}$ and $g \geq 0\}$. For each $P \in \mathcal{P}$, there is a set $A_P = g.l.b. \{A : (g, P) \in \mathcal{M}$ and $A = \{X : g(x) \neq 0\}\}$. Note that A_P is not unique. Let $\mathcal{F} = \{g :$ for some $P \in \mathcal{P}(g, P) \in \mathcal{M}$ and for every $(h, P) \in \mathcal{M} g \leq h$ a.e. Q for all $Q \in \mathcal{M}\}$. If $(g, P) \in \mathcal{M}$, then $g\chi_{A_P} \in \mathcal{F}$. Now let B_r be a least upper bound of $\{C :$ for some $g \in \mathcal{F}, C = (X : g(x) > r)\}$, for r rational. We may assume $B_r \subseteq B_s$ if $r > s$. Define $f^*_{(x)} = \sup\{r : \chi \in B_r\}$.

Observe that if $g, h \in \mathcal{F}$ and $(g, P), (h, Q) \in \mathcal{M}$ then

(i) on $A_P \cap A_Q g = h$ a.e. R for all $R \in \mathcal{P}$,

(ii) on $A_P \sim A_Q h = 0$ a.e. R for all $R \in \mathcal{P}$,

(iii) on $A_Q \sim A_P g = 0$ a.e. R for all $R \in \mathcal{P}$,

(iv) on $\sim A_P \sim A_Q g = h = 0$ a.e. R for all $R \in \mathcal{P}$, since C is pairwise sufficient.

Consequently if $(g, P) \in \mathcal{M}, g \in \mathcal{F}$, and for some $h \in \mathcal{F}, C = \{x : h(x) > r\}$, then $C \cap A_P \sim \{x : g(x) > r\}$ has R-measure 0 for all $R \in \mathcal{P}$. Thus $B_r \cap A_P \sim \{x : g(x) > r\}$ has R-measure 0 for all $R \in \mathcal{P}$. Since the reverse relation holds for all r and P, if $(g, P) \in \mathcal{M}$ and $g \in \mathcal{F}, f^* = g$ except on a set of P-measure 0. Thus f^* is a common conditional expectation of f given C for all $P \in \mathcal{P}$. Therefore $\mathcal{P}$ is sufficient.

We can obtain the analog of the other Halmos-Savage theorem also.

Theorem: Let C be a pairwise sufficient sub-σ-field for a family $\mathcal{P}$ of probability measures on a space $(X, \mathcal{A})$. Then if every family of C-measurable sets has a common least upper bound with respect to all measures in $\mathcal{P}$, and $\mathcal{P}$ is equivalent to the well-ordered family $\mathcal{Q}$, there is a measure μ with respect to which the Radon-Nikodym theorem holds for C-measurable functions such that for all $P \in \mathcal{P}$, $\frac{dP}{d\mu}$ is C-measurable.

Proof Well-Order $\mathcal{Q}$ into $Q_1, Q_2, \ldots, Q_\alpha, \ldots$. Let $\{E_\alpha\}$ be a net of C-measurable sets such that

(a) $Q_\alpha(\underset{\gamma \leq \alpha}{\ell.u.b} E_\gamma) = 1$

(b) $Q(E_\beta \cap E_\alpha) = 0$ for all whenever $\beta < \alpha$. (This is certainly possible, since we need

only choose E_α to be disjoint from $\ell u.b._{\gamma<\alpha} E_\gamma$).

(c) $Q_\gamma(E_\alpha)$ is minimal for $Q_\gamma \in \mathcal{Q}$ i.e. If $\mathcal{U}$ is collection of sets satifying (a) and (b), then E_α is a greatest lower bound of $\mathcal{U}$ with respect to $\mathcal{P}$.

We define $\mu(E) = \sum_\alpha Q_\alpha(E \cap E)\alpha)$ for any set $E \in \mathcal{A}$. ("$\sum_\alpha Q_\alpha(E \cap E_\alpha)$" means the supreme of the finite sums).

Then μ is countably additive since if $\{F_i\}$ are disjoint, $\mu(\bigcup_i F_i) = \sum_\alpha Q_\alpha(\bigcup_i F_i \cap E_\alpha) = \sum_\alpha \sum_i Q_\alpha(F_i \cap E_\alpha)$ while $\sum_i \mu(F_i) = \sum_i \sum_\alpha Q_\alpha(F_i \cap E_\alpha)$ and the order of summation can be reversed since all terms are non-negative.

For any $E \in \mathcal{A}, \mu(E) = 0$ implies $Q_\alpha(E) = 0$ since $Q_\alpha(E \cap E_\alpha) \leq \mu(E)$, if $\gamma < \alpha, Q_\gamma(E \cap E_\gamma) = 0$ implies $Q_\alpha(E \cap E_\gamma) = 0$ by the minimality of E_γ, and by (a) and as $Q_\alpha(E) = \sum_{\gamma \in \alpha} Q_\alpha(E \cap E_\gamma)$. So $Q << \mu$ for all $Q \in \mathcal{Q}$.

Since $\mathcal{P}$ is equivalent of $\mathcal{Q}$, $P << \mu$ for all $P \in \mathcal{P}$. Condition (b) shows that $Q << Q_\alpha$ on E_α for all $Q \in \mathcal{Q}$, hence for all $P \in \mathcal{P}$. Since Q_α is a finite measure on E_α, by the Halmos-Savage theorem $\frac{dP}{dQ_\alpha}$ can be taken to be C-measurable on E_α; i.e., there is a C-measurable function $\lambda_{P,\alpha}$ vanishing off E_α such that for all $F \in \mathcal{A}, F \subseteq E_\alpha, P(F) = \int_F \lambda_{P,\alpha} dQ_\alpha$. But by (a) and (b), only countably many $P(E_\alpha)$ are positive; hence the sum of those $\lambda_{P,\alpha}$ is a Radon-Nikodym derivative of P with respect to μ. That the Radon-Nikodym theorem holds for (μ, C) is a consequence of every family of C-measurable sets having a common $\mathcal{Q}$-least upper bound, which is also a μ-least upper bound.

(It may seem that the axiom of choice was used in the proof; however, the E_γ are unique up to sets of P-measure 0 for all $P \in \mathcal{P}$, and hence all that is required is the

countable number of choices needed to prove the Radon-Nikodym theorem.)

Investigations of the relationship between the concepts of sufficiency and informativeness have been carried out by Blackwell and C.H. Boll.

Formally, $z = (Z, \mathcal{B}, \Omega, P_Z)$ is more informative than $\mathcal{W} = (W, \mathcal{I}, \Omega, P_W)$ if, for every action space $(A, \mathcal{A})$ and every loss function L,

$$R(A, \mathcal{A}, L|z) \supseteq R(A, \mathcal{A}, L|\mathcal{W})$$

where

$$R(A, \mathcal{A}, L|z) = \{\rho_d : d \in D\}$$

where $D = \{d : d \text{ is a } \mathcal{B} - \mathcal{A} \text{ measurable function from } Z \text{ to } A\}$ and ρ_d is the function

$$\rho_d(\omega) = \int L(\omega, d(Z)) dP_\omega(Z).$$

We assume throughout that the measures $\{P_\omega : \omega \in \Omega\}$ on Z are dominated by a measure μ such that $\frac{dP_\omega}{d\mu}$ exists for all $\omega \in \Omega$.

We note that the requirement

$$R(A, \mathcal{A}, L|z) \supseteq R(A, \mathcal{A}, L|\mathcal{W})$$

for all action spaces A is necessary in the definition of "more informative", in the sense that there do exist examples in which the above inclusion relation holds for one action space but not for another. We exhibit a simple one here, due to Blackwell.

Let $\Omega = \{1, 2, \ldots, n\}, Z = \{x_1, \ldots, x_n, x_{n+1}\}, W = \{y_1, \ldots, y_n\}, A_1 = \{1, \ldots, n\}$,

$L(\omega, a) = 1 - \delta_{\omega a}, P_Z(x_i|\omega =$ q if $\omega = i$ and let

r if $i = n + 1$

0 otherwise

$p_W(y_i|\omega) = \begin{matrix} \frac{1}{n-1} & \text{if } \omega \neq i \\ 0 & \text{if } \omega = i \end{matrix}$. If the action space is A_1, the procedure $d(y_i) = i$ on W has risk $\rho_d(\omega) = \Sigma L(\omega, d(y_i))p(y_i|\omega) = 0$ since $L(\omega, i) = 0$ unless $p(y_i|\omega) = 0$. But if d^* is any procedure on $Z, P(i|x_{n+1}, d^*) > 0$ for some i, and then $\rho_{d^*}(i) \geq p(i|x_{n+1}, d^*)p(x_{n+1}|\omega = i) > 0$.

However if there are at most $n - 1$ actions, let d be any pure decision procedure on W (i.e., for each y_i there is an $a \in A_2$ such that $p(a|y_i, d) = 1$. We let this a be $d(y_i)$). Then there is some $a^* \in A_2$ such that $d(y_i) = d(y_j) = a^*$ for some $y_i \neq y_j$, since there are ny's and only $(n - 1)a$'s. If $r \leq \frac{1}{n-1}$ we can define d^* by $d^*(x_{n+1}) = a^*$ and extend the definition to other observations.

Definition. We say that an experiment $z = (Z, \mathcal{B}, \Omega, P)$ is sufficient for any experiment $\mathcal{W} = (W, \mathcal{D}, \Omega, Q)$ if there exists a linear transformation $K : M(W) \to M(Z)$ (where $M(X)$ is the set of bounded measurable functions on X) such that, for each $f \in M(W)$,

$$\int f dQ_\omega = \int K(f) dP_\omega \text{ for all } \omega \in \Omega,$$

and $a \leq f(w) \leq b$ for all w implies $a \leq K9f)(z) \leq b$ for all Z.

We remark that this definition does not correspond exactly to the usual definition of sufficiency because there may exist no transformation $T : W \to Z$ (a trivial example would be $W = \{0\}$). That is, a sufficient statistic (in the usual sense) is only "as good as" a

point in the sample space, but an observation from Z may actually be more informative than one from W.

However if there is a measure preserving transformation, $T : W \to Z$, then under mild further conditions the two definitions are closely connected.

If T is a sufficient statistic, then defining $K(f)(z) = E(f|T = z)$ for all $f \in M(W)$ shows that z is a sufficient experiment for $\mathcal{W}$.

The converse is rather more involved. Suppose z is a sufficient experiment for $\mathcal{W}$. If $A = \{a_1, a_2\}$, and d is any procedure on

$$\begin{aligned} W, \rho(\omega, d) &= L(a_1, \omega) \int p(a_1|w, d) dQ_\omega(w) \\ &\quad + L(a_2, \omega) \int f(a_2|w, d) dQ_\omega(w) \\ &= L(a_1, \omega) \int g_1(z) dp_\omega(z) \\ &\quad + L(a_2, \omega) \int g_2(z) dP_\omega(z) \end{aligned}$$

where $g_i(z) = K(P(a_i|o, d))(z), i = 1, 2$, so that

$$g_1(z) + g_2(z) = K(P(a_1)) + K(p(a_2)) = K[p(a_1) + p(a_2)] = K(1) = 1.$$

So the strategy d^* on Z such that $p[a_i|z, d^*] = g_i(z), i = 1, 2$, has the same risk for all ω as does d. Also if d^* is any procedure on z, then $d(w) = d^*(T(w))$ satisfies

$$\begin{aligned} \rho(\omega, d) &= L(a_1, \omega) \int p(a_1|w, d) dQ_\omega(w) + L(a_2, \omega) \int p(a_2, |w, d) dQ_\omega(w) \\ &= L(a_1, \omega) \int p(a_1|T(w), d^*) dQ_\omega(w) + L(a_2, \omega) \int p(a_2|T(w), d^*) dQ_\omega(w) \\ &= L(a_1, \omega) \int p(a_1|z, d^*) dp_\omega(z) + L(a_2, \omega) \int p(a_2|z, d^*) dp_\omega(z) = \rho(\omega, d^*) \end{aligned}$$

since T is a measure-preserving transformation. Hence z and $\mathcal{W}$ are equally informative for all 2-action problems and in fact $(Z, \mathcal{B}, \Omega, P), (W, \mathcal{D}, \Omega, Q)$ and (W, C, Ω, Q) are thus equally

informative for all 2-action problems, where $C = T^{-1}(\mathcal{B}) \subseteq \mathcal{D}$. Now let $\{\omega_1, \omega_2\} \subseteq \Omega$, and let

$F_{\mathcal{D}}(t) = (p_{\omega_1}+p_{\omega_2})\{w : \frac{dp_{\omega_1}}{d(p_{\omega_1}+p_{\omega_2})_{\mathcal{D}}}(w) \leq t\}$ where $p_{\omega_1}, p_{\omega_2}$ are taken as measures on $(W, \mathcal{D})$

$F_C(t) = (p_{\omega_1}+p_{\omega_2})\{w : \frac{dp_{\omega_1}}{d(p_{\omega_1}+p_{\omega_2})_C}(w) \leq t\}$ where $p_{\omega_1}, p_{\omega_2}$ are taken as measures on (W, C).

It is a triviality that if z and $\mathcal{W}$ are equally informative when the state space is Ω, they are equally informative when the state space is $\Omega_1 \subseteq \Omega$. Consequently the Bayes risks against any prior on $\{\omega_1, \omega_2\}$ must be equal. Let $\frac{dp_{\omega_1}}{d(p_{\omega_1}+p_{\omega_2})_{\mathcal{D}}}(w) = f(x), \xi$ be the prior probability that ω_1 is the true state, $L(a_1, \omega_1) = L(a_2, \omega_2) = 0, L(a_1, \omega_2) = L(a_2, \omega_1) = 1$, and suppose we use a procedure d on $(W, \mathcal{D})$ which causes us to take act a_1 with probability $Q(w)$. Then the risk of d is

$$\begin{aligned} R(d) &= (1-\xi)\int Q(w)dp_{\omega_2}(w) + \xi\int -Q(w)dp_{\omega_1}(w) \\ &= \int (1-\xi)Q(w)(1-f(w)) + \xi(1-Q(w))f(w)d(p_{\omega_1}+p_{\omega_2})w. \end{aligned}$$

For fixed w, the minimum of the integrand occurs if Q is the function:

$$Q(w) = \begin{array}{ll} 1 & \text{if } \xi f > (1-\xi)(1-f); \text{ i.e. } f > 1-\xi \\ 0 & \text{if } f < 1-\xi \end{array}$$

so that the Bayes risk is

$$\begin{aligned} R^*_{\mathcal{D}} &= \int_{f(w)>1-\xi} (1-\xi)\ (1-f(w))d(p_{\omega_1}+p_{\omega_2})(w) + \int_{f(w)<1-\xi} \xi f(w)\ \ d(p_{\omega_1}+p_{\omega_2})(w) \\ &= \int_{1-\xi}^{1+} (1-\xi)(1-t)dF_{\mathcal{D}}(t) + \int_{0-}^{1-\xi} \xi t dF_{\mathcal{D}}(t) \\ &= \int_{0-}^{1+} (1-\xi)(1-t)dF_{\mathcal{D}}(t) + \int_{0-}^{1-\xi} \xi + t - 1 dF_{\mathcal{D}}(t) \end{aligned}$$

But $F(0-) = 0$, and $\xi+t-1$ is continuous and zero at $t = 1-\xi$, so integration by parts gives $\int_{0-}^{1-\xi}(\xi+t-1)dF_{\mathcal{D}}(t) = -\int_0^{1-\xi} F_{\mathcal{D}}(t)dt$. Also $(1-\xi)\int_0^1 dF_{\mathcal{D}}(t) = 2(1-\xi)$ and $\int_0^1 t dF_{\mathcal{D}}(t) =$

$\int_W \frac{dp_{\omega_1}}{d(p_{\omega_1}+p_{\omega_2})}(w)d(p_{\omega_1}+p_{\omega_2}) = \int_W^0 dp_{\omega_1}(w) = 1$. Hence $R^*_{\mathcal{D}}(1-\xi) = \int_0^{1-\xi} F_{\mathcal{D}}(t)dt$ for procedures on $(W,\mathcal{D})$. Obviously for procedures on (W,C) we have similarly

$$R^*_C = (1-\xi) - \int_0^{1-\xi} F_C(t)dt.$$

Since $(W,\mathcal{D})$ and (W,C) are equally informative, we must have

$$\int_0^{1-\xi} F_{\mathcal{D}}(t)dt = \int_0^{1-\xi} F_C(t)dt \text{ for all } \xi \text{ in } [0,1]$$

so that $F_{\mathcal{D}}(t) = F_C(t)$ a.e. So if $dR = \frac{1}{2}d(p_{\omega_1}+p_{\omega_2})$, f is as before, and $g = \frac{dp_{\omega_1}}{d(p_{\omega_1}+p_{\omega_2})_C}$ we have

$$R(f \leq a) = R(g \leq a).$$

But $g = E_R(f|C)$. Hence $f = g$ a.e. $[R]$, so f is C-measurable and hence C is pairwise sufficient for any $\{\omega_1,\omega_2\} \subseteq \Omega$. Since we assume the measures $\{p_\omega : \omega \in \Omega\}$ on Z are domianted by a measure with respect to which the Radon-Nikodym theorem holds, C is then sufficient for Ω. Hence the statistic $T(w) \in Z$ is sufficient in the ordinary sense for Ω.

For a further comparison of the two definitions of 'sufficient', suppose that $\mathcal{Y} = (Z \times W, \mathcal{B} \times \mathcal{D}, \Omega, R)$ is a 'combined' experiment with $R_\omega(B \times W) = P_\omega(B)$ and $R_\omega(Z \times D) = Q_\omega(D)$ for all $\omega \in \Omega, B \in \mathcal{B}, D \in \mathcal{D}$. Then what is the relationship between (1)"$\mathcal{Z} = (Z,\mathcal{B},\Omega,P)$ is sufficient for $\mathcal{W} = (W,\mathcal{D},\Omega,Q)$" and (2) "the statistic T given by $T(z,w) = z$ is sufficient on $Z \times W$ for Ω"? We have been able to show that (2) $\Rightarrow$ (1) but not that (1) $\Rightarrow$ (2). If, in the definition of 'sufficient experiment', we insist that K be countably additive, we can show easily that (1) $\Rightarrow$ (2), but not that (2) $\Rightarrow$ (1) unless the conditional probabilities obtained from T are true probabilities (i.e. countably additive), which will be

true if there is a locally compact, separable metric topology $\mathcal{T}$ on W, and $\mathcal{D}$ is the Borel field generated by $\mathcal{T}$.

We now return to the relationship between sufficiency and informativeness in the comparison of experiments problem. We showed above that if $\mathcal{Z}$ is sufficient for $\mathcal{W}$, then $\mathcal{Z}$ is more informative than $\mathcal{W}$ for all 2-action problems; obviously this can be extended to all finite action problems. If, further, the transformation K is countably additive (so that, when applied to indicator functions, it gives conditional probabilities which are true probability measures) then $\mathcal{Z}$ is more informative than $\mathcal{W}$ for arbitrary action spaces. Again, an extension of the argument for finite action spaces can be used to show that $\mathcal{Z}$ is more informative than $\mathcal{W}$ if the action space is separable metric.

We now give a theorem about the converse:

If $\mathcal{Z} = (Z, \mathcal{B}, \Omega, P)$ is more informative than $\mathcal{W} = (W, \mathcal{D}, \Omega, Q)$ and the measures $\{p_\omega : \omega \in \Omega\}$ are dominated by a measure μ with respect to which the Radon-Nikodym theorem holds, then $\mathcal{Z}$ is sufficient for $\mathcal{W}$.

Proof. Let

$$M(W) = \{\text{all bounded measurable functions on } W\}$$

$$L_1(\mu) = \{\text{all } L_1(\mu) \text{ functions on } Z\}$$

$$\mathcal{H} = \{H : H : M(W) \times L_1(\mu) \to R \text{ and } H \text{ is bilinear, and}$$

$$|H(f,g)| \leq ||f||\ ||g|| \text{ for all } f \in M_1(W), g \in L_1(\mu)\}$$

where $||f|| = \sup |f|$, $||g|| = \int_Z |g|\ d\mu$. Then $\mathcal{H} \subseteq \underset{M(W)\times L_1(\mu)}{X} [-||f||\ ||g||,\ ||f||\ ||g||]$, a compact space; and if $\{H_\alpha\}$ is a net of elements of $\mathcal{H}_1$ and $F \in X[-||f||\ ||g||,\ ||f||\ ||g||]$ but F is not bilinear, then F is not an accumulation point of $\mathcal{H}_1$ since if $f_1 \in M(W), g_i \in$

$L_1(\mu), i = 1, \ldots, n$, and $\Sigma\alpha_i f_i g_i = 0$ but $\Sigma\alpha_i F(f_i, g_i) \neq 0$, then $\{G : \Sigma\alpha_i G(f_i, g_i) > 0\}$ is an open set containing F but not containing any element of $\{H_\alpha\}$. Hence $\mathcal{H}$ is a closed subset of a compact space so $\mathcal{H}$ is compact.

Let $\mathcal{A} = \{A : A \subseteq \Omega$ and A is finite$\}$ and for each $A \in \mathcal{A}$ let $\mathcal{H}_A = \{H : H \in \mathcal{H}$, and $H(f, p_\omega) = \int f dQ_\omega$ for each $\omega \in A\} f \in M(W)$ where $p_\omega = \frac{dp_\omega}{d\mu}$.

Blackwell (AMS, 24, 2; June 1953, pp 265–72, Thm. 8) has shown that if $\mathcal{Z}$ is more informative than $\mathcal{W}$, then $\mathcal{H}_A \neq \phi$ for every $A \in \mathcal{A}$. Also, if $\{H_\alpha\}$ is a net of elements of $\mathcal{H}_A$, and $F \in \mathcal{H}$ such that for some $f \in M(W), \omega \in A, F(f, p_\omega) \neq \int f dQ_\omega$, then $\{G : G(f, p_\omega) \neq \int f dQ_\omega\}$ is an open set containing F but containing no element of $\mathcal{H}_A$. hence $\mathcal{H}_A$ is closed.

If $\{A_k\} k = 1, \ldots, n$ is a sequence of elements of $\mathcal{A}$, then $\bigcup_{k=1}^{r} A_k \in \mathcal{A}$ and hence

$$\bigcap_{k=1}^{n} \mathcal{H}_{A_k} = \bigcup_{k=1}^{\mathcal{H}_n} A_k \neq \phi.$$

Hence the collection of closed sets, $\{\mathcal{H}_A : A \in \mathcal{A}\}$ has the finite intersection property so that

$$\mathcal{H}_\Omega = \{h : H \in \mathcal{H} \text{ and } H(f, p_\omega) = \int f dQ_\omega \text{ for all } \omega \in \Omega, f \in M(W)\}$$

$$= \bigcap_{A \in \mathcal{A}} \mathcal{H}_A \neq \phi \text{ since } \mathcal{H} \text{ is compact.}$$

Now choose any $H \in \mathcal{H}_\omega$, and let, for each $f \in M(W)$ and each $A \in \mathcal{B}$,

$$\nu_f(A) = H(f, \chi_A).$$

Then

$$|\nu_f(A)| \leq ||f||\ \mu(A)$$

so

$$\nu_f << \mu$$

and hence $\frac{d\nu_f}{d\mu} = K(f)$ exists. Then for each $\omega \in \Omega$, and $f \geq 0$

$$\int K(f)(x)p_\omega(x)d_\mu(x) = \int p_\omega(x)d\nu_f(x) \text{ by definition of } K(f),$$
$$= \lim \int \Sigma a_i \chi_{A_i}(x)d\nu_f(x)$$

where $\Sigma a_i \chi_{A_i}(x) \uparrow p_\omega(x)$

$$= \lim \Sigma a_i H(f, \chi_{A_i})$$

$$= \lim H(f, \Sigma a_i \chi_{A_i}) \text{ since } H \text{ is bilinear}$$

$$= H(f, p_\omega) \text{ since bounded linear functionals are continuous}$$

$$= \int f dQ_\omega \text{ since } H \in \mathcal{H}_\Omega.$$

Clearly $|K(f)| \leq ||f||$ a.e. $[\mu]$, so we can take $|K(f)| \leq ||f||$ everywhere. Then $K : M(W) \to M(Z)$ has the properties: (i) $|K(f)| \leq ||f||$ everywhere; (ii) $K(\sum_1^n a_i f_i) = \sum_1^n a_k K(f_i)$, since H is bilinear; (iii) $\int K(f)dp_\omega = \int f dQ_\omega$ for all $\omega \in \Omega, f \in M(W)$. Hence $\mathcal{Z}$ is sufficient for $\mathcal{W}$.