484

COMPARISON OF EXPERIMENTS AND INFORMATION MEASURES*

by

Prem K. Goel and Morris H. DeGroot
Purdue University and Carnegie-Mellon University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #484

February 1977

*This research was supported in part by the National Science Foundation under grant #SOC 74-02071 A02 at Carnegie-Mellon University.

Abstract

COMPARISON OF EXPERIMENTS AND INFORMATION MEASURES

by

Prem K. Goel and Morris H. DeGroot

Let $\mathcal{E}_X = \{X, S_X; P_\theta, \theta \in \Theta\}$ and $\mathcal{E}_Y = \{Y, S_Y; Q_\theta, \theta \in \Theta\}$ be two statistical experiments with the same parameter space Θ. Some implications of the sufficiency of $\mathcal{E}_X$ for $\mathcal{E}_Y$, according to Blackwell's definition, are given in terms of (i) Kullback-Leibler information contained in the experiment for discriminating between the marginal distributions of the random variable with respect to arbitrary prior distributions ξ_1 and ξ_2 on Θ, and (ii) Fisher information matrices when $\Theta \subset R^k$. For a scale parameter θ, and $k_1 > k_2 > 0$, the experiment with parameter θ^{k_1} is proved to be sufficient for the experiment with parameter θ^{k_2} for a wide class of distributions, which includes the gamma density and the normal density with known mean. Some results of Stone (1961) are generalized to the class of experiment with both location and scale parameters. A method of weakening Blackwell's definition of sufficiency is proposed in which $\mathcal{E}_X$ is more informative than $\mathcal{E}_Y$ if, for every decision problem involving θ, the expected Bayes risk from $\mathcal{E}_X$ is not greater than that from $\mathcal{E}_Y$. This concept is then applied to present a definition of marginal Bayesian sufficiency when there are nuisance parameters in the decision problem.

AMS 1970 Subject classification, Primary 62B15, Secondary 62B10
Key words and Phrases. Comparison of experiments, Kullback-Leibler information, Fisher information matrix, expected Bayes risk, marginal Bayesian sufficiency, more informative experiment, self-decomposable characteristic functions.

COMPARISON OF EXPERIMENTS AND INFORMATION MEASURES*

by

Prem K. Goel and Morris H. DeGroot
Purdue University and Carnegie-Mellon University

1. Introduction and Summary. Let $\mathcal{E}_X = \{X, S_X; P_\theta, \theta \in \Theta\}$ denote a statistical experiment in which a random variable or random vector X defined on some sample space S_X is to be observed, and the distribution P_θ of X depends on a parameter θ, whose value is unknown and lies in some parameter space Θ. Also, let $\mathcal{E}_Y = \{Y, S_Y; Q_\theta, \theta \in \Theta\}$ denote another statistical experiment with the same parameter space Θ. Blackwell's method (1951) for comparing two experiments states that the experiment $\mathcal{E}_X$ is sufficient for the experiment $\mathcal{E}_Y$ (denoted $\mathcal{E}_X \geq \mathcal{E}_Y$) if there exists a stochastic transformation of X to a random variable Z(X) such that, for each $\theta \in \Theta$, the random variables Z(X) and Y have identical distributions. It was proved by Blackwell (1953) that this method of comparison is equivalent to Bohnenblust, Shapley and Sherman's method for comparing two experiments [see Blackwell (1951)] which states that $\mathcal{E}_X$ is more informative than $\mathcal{E}_Y$ if for every decision problem involving θ and every prior distribution on Θ, the expected Bayes risk from $\mathcal{E}_X$ is not greater than that from $\mathcal{E}_Y$. LeCam (1964) generalized this notion to a concept of approximate sufficiency or ε-deficiency of $\mathcal{E}_X$ relative to $\mathcal{E}_Y$. Some other papers on this topic are DeGroot (1962) and (1966), Torgersen (1972) and (1977) and Feldman (1972). We shall now give a summary of the results presented in this paper.

---

*This research was supported in part by the National Science Foundation under grant #SOC 74-02071 A02 at Carnegie-Mellon University.

In Theorem 1 of Section 2, we prove that if $\mathcal{E}_X \geq \mathcal{E}_Y$, then for every pair of prior distributions ξ_1 and ξ_2 on Θ, the Kullback-Leibler (K-L) information contained in $\mathcal{E}_X$ for discriminating between the marginal distributions of X with respect to ξ_1 and ξ_2 is at least as the K-L information contained in $\mathcal{E}_Y$ for discriminating between the marginal distributions of Y. This theorem strengthens result by Lindley (1956) and Sakaguchi (1964). Now suppose that Θ is an open subset of the k-dimensional Euclidean space R^k and the usual regularity conditions that permit the computation of the Fisher information matrices $\underset{\sim}{i}_X(\underset{\sim}{\theta})$ and $\underset{\sim}{i}_Y(\underset{\sim}{\theta})$ are satisfied. It is proved in Theorem 2 that if $\mathcal{E}_X \geq \mathcal{E}_Y$, then $\underset{\sim}{i}_X(\underset{\sim}{\theta})-\underset{\sim}{i}_Y(\underset{\sim}{\theta})$ is non-negative definite for each $\theta \in \Theta$ (denoted $\mathcal{E}_X \geq_F \mathcal{E}_Y$). For k=1, this theorem provides an alternative proof of a result by Stone (1961). A counter example is given to show that the converse of this result does not hold.

In Section 3, it is assumed that θ is a scale parameter in the distribution of Y and that X has this same distribution except that θ is replaced by $\theta^k (k>1)$. Let W denote a random variable with distribution identical to that of X with $\theta=1$, and let $\varphi(t)$ denote the characteristic function of log W. In Lemma 2, we show that if $\varphi(t)/\varphi(t/k)$ is a characteristic function, then $\mathcal{E}_X \geq \mathcal{E}_Y$. It is noted that $\varphi(t)$ satisfies this condition if and only if $\varphi(t)$ is a self-decomposable characteristic function [see Lukacs (1970), p.161]. Let $G_n(a,b)$ denote an experiment in which a random sample of size n is taken from a gamma distribution G(a,b) with parameters $a > 0$ and $b > 0$, for which the density function is

$$g(w) = \frac{w^{a-1}}{b^a \Gamma(a)} \exp(-w/b) \text{ for } w > 0. \tag{1.1}$$

Also, let $N_n(\mu,\sigma^2)$ denote an experiment in which a random sample of size n is taken from a normal distribution with mean μ and variance σ^2. The result in Lemma 2 is used to prove that for any known numbers a>0 and b>0, $G_n(a,b\theta^k) \geq$

$G_n(a,b\theta)$ for all k>1, and that $N_n(0,\sigma^{2k}) \geq N_n(0,\sigma^2)$ for all k>1.

In Section 4, we extend some of the results obtained by Stone (1961) for a class of experiments with location parameter θ to the class of experiments with both a location parameter μ and a scale parameter σ. For μ unknown and σ known, it is easy to see and well known that $N_n(\mu,\sigma^2) \geq N_n(\mu,k\sigma^2)$ for all k>1. We prove that when both μ and σ are unknown, $N_n(\mu,\sigma^2)$ is not sufficient for $N_n(\mu,k\sigma^2)$. Next let $\mathcal{E}^*(c)$ denote an experiment in which a random variable will be observed for which the pdf has the form $(c/\sigma)f[c(x-\mu)/\sigma]$, and consider two given values $c_1 > c_2 > 0$. We prove that if f is a symmetric function on the real line and the usual regularity conditions that permit the calculation of the Fisher information matrix $\underset{\sim}{i}_c(\mu,\sigma)$ are satisfied, then $\mathcal{E}^*(c_1) \geq_F \mathcal{E}^*(c_2)$, even though $\mathcal{E}^*(c_1)$ may not be sufficient for $\mathcal{E}^*(c_2)$.

Since two experiments $\mathcal{E}_X$ and $\mathcal{E}_Y$ may not be comparable in Blackwell's sense, Feldman (1972) introduced a weakened definition in which $\mathcal{E}_X$ is more informative than $\mathcal{E}_Y$ for a fixed decision problem involving θ if, for every prior distribution on Θ, the expected Bayes risk from $\mathcal{E}_X$ is not greater than that from $\mathcal{E}_Y$. In Section 5, we propose an alternative method of weakening the definition in which $\mathcal{E}_X$ is more informative than $\mathcal{E}_Y$ for a fixed prior distribution on Θ if, for every decision problem involving θ, the expected Bayes risk from $\mathcal{E}_X$ is not greater than that from $\mathcal{E}_Y$. We then apply this concept to problems in which $\underset{\sim}{\theta}$ is a vector with a given prior distribution, and we are interested in decision problems involving only some of the components of $\underset{\sim}{\theta}$. We present a definition of the marginal Bayesian sufficiency of $\mathcal{E}_X$ for $\mathcal{E}_Y$ in this context. Some examples are given to illustrate the usefulness of this concept.

2. Relationships between Sufficiency and Information. Consider again two arbitrary experiments $\mathcal{E}_X$ and $\mathcal{E}_Y$ with the same parameter space Θ as defined at the beginning of Section 1. We shall assume that there exists generalized probability density functions (gpdf's) $p(x|\theta)$ and $q(y|\theta)$ for the distributions P_θ and Q_θ, with respect to some σ-finite measures μ and ν respectively. We shall now investigate the implications of the relation $\mathcal{E}_X \geq \mathcal{E}_Y$ in terms of some well known information measures. Let Ξ denote the class of all prior distributions on the parameter space Θ. Given two prior distributions $\xi_1, \xi_2 \in \Xi$, let $p_i(x)$ denote the marginal gpdf $\int_\Theta p(x|\theta)d\xi_i(\theta)$, for i=1,2, and let $I_X(\xi_1,\xi_2)$ denote the K-L information contained in $\mathcal{E}_X$ for discriminating between $p_1(x)$ and $p_2(x)$, defined by

$$I_X(\xi_1,\xi_2) = \int_{S_X} p_1(x) \log \frac{p_1(x)}{p_2(x)} d\mu(x) \tag{2.1}$$

If ξ_1 assigns probability 1 to a point $\theta=\theta_0$, we shall denote $I_X(\xi_1,\xi_2)$ by $I_X(\theta_0,\xi_2)$. The K-L information $I_Y(\xi_1,\xi_2)$ contained in $\mathcal{E}_Y$ is defined analogously.

Lindley (1956) has shown that if $\mathcal{E}_X \geq \mathcal{E}_Y$, then the Shannon information contained in $\mathcal{E}_X$ is at least as large as that contained in $\mathcal{E}_Y$. That is, if $\mathcal{E}_X \geq \mathcal{E}_Y$ then

$$\int_\Theta I_X(\theta,\xi)d\xi(\theta) \geq \int_\Theta I_Y(\theta,\xi)d\xi(\theta) \text{ for all } \xi \in \Xi. \tag{2.2}$$

If (2.2) holds for $\mathcal{E}_X$ and $\mathcal{E}_Y$, we shall denote it by $\mathcal{E}_X \succeq_L \mathcal{E}_Y$. The following stronger version of Lindley's result was proved by Sakaguchi (1964).

Lemma 1. If $\mathcal{E}_X \succeq \mathcal{E}_Y$, then

$$I_X(\theta_0,\xi) \geq I_Y(\theta_0,\xi), \text{ for all } \theta_0 \in \Theta \text{ and } \xi \in \Xi. \tag{2.3}$$

It was also stated by Sakaguchi that (2.3) does not imply $\mathcal{E}_X \succeq \mathcal{E}_Y$, although he gave no counterexample. We shall now prove the following result which is stronger than Lemma 1.

Theorem 1. Let $\mathcal{E}_X$ and $\mathcal{E}_Y$ be two statistical experiments with the same parameter space Θ. If $\mathcal{E}_X \geq \mathcal{E}_Y$, then

$$I_X(\xi_1,\xi_2) \geq I_Y(\xi_1,\xi_2) \text{ for all } \xi_1,\xi_2 \in \Xi. \tag{2.4}$$

Proof. If $\mathcal{E}_X \geq \mathcal{E}_Y$, then there exists a nonnegative function $h(y|x)$ satisfying the relations [See DeGroot (1970), p.434]

$$q(y|\theta) = \int_{S_X} h(y|x)\ p(x|\theta)\ d\mu(x) \text{ for every } \theta \in \Theta \text{ and } y \in S_y, \tag{2.5}$$

$$\int_{S_Y} h(y|x)\ d\nu(y) = 1. \tag{2.6}$$

It follows from (2.5) and a change in the order of integration that

$$\begin{aligned} q_i(y) &= \int_{\Theta} q(y|\theta)d\xi_i(\theta) \\ &= \int_{S_X} h(y|x)p_i(x)d\mu(x), \qquad i=1,2 \end{aligned} \tag{2.7}$$

Let

$$t(y) = q_1(y)\log \frac{q_1(y)}{q_2(y)}. \tag{2.8}$$

Then from (2.7) and Corollary 3.1 of Kullback (1968),

$$t(y) \leq \int_{S_X} h(y|x)p_1(x)\log \frac{p_1(x)}{p_2(x)}\ d\mu(x). \tag{2.9}$$

It now follows from (2.9), a change in the order of integration, and (2.6) that

$$\begin{aligned} I_Y(\xi_1,\xi_2) &= \int_{S_Y} t(y)d\nu(y) \\ &\leq \int_{S_X} p_1(x)\log \frac{p_1(x)}{p_2(x)}\ d\mu(x) \\ &= I_X(\xi_1,\xi_2). \ \blacksquare \end{aligned} \tag{2.10}$$

<u>Example 1</u>. Let $\mathcal{E}(\theta_1,\theta_2,\theta_3)$ denote an experiment in which a coin with unknown probability of heads θ is flipped n times and the parameter space Θ contains only three points $0 \le \theta_1 \le \theta_2 \le \theta_3 \le 1$. Blackwell (1951) remarks that the experiment $\mathcal{E}_X \equiv \mathcal{E}(0,\frac{1}{2},1)$ is not sufficient for the experiment $\mathcal{E}_Y \equiv \mathcal{E}(0,\frac{1}{2},\frac{1}{2})$ even though our intuition suggests the contrary. In other words, suppose that there are only three possible states of nature and we have a choice of either (i) observing n flips of a coin $\mathcal{E}_X$ for which the probability of heads is $0, \frac{1}{2}$, or 1 according as θ_1, θ_2 or θ_3 is correct, or (ii) observing n flips of a coin $\mathcal{E}_Y$ for which the probability of heads is 0 if θ_1 is correct, but is $\frac{1}{2}$ if either θ_2 or θ_3 is correct. It would seem at first glance that $\mathcal{E}_X$ must always be at least as useful as $\mathcal{E}_Y$, but Blackwell pointed out that this conclusion is not correct.

Lindley (1956) showed that $\mathcal{E}_X \succeq_L \mathcal{E}_Y$ for these experiments. However, it can be shown that $I_X(\theta_2,\xi) \ge I_Y(\theta_2,\xi)$ if and only if $3\xi_1 + \xi_2 \le 1$, where $\xi_1 = \xi(\theta_i)$. Hence (2.3) does not hold for $\theta_0 = \theta_2$ and a prior distribution ξ for which $3\xi_1 + \xi_2 > 1$.

Since $\mathcal{E}_X$ is not sufficient for $\mathcal{E}_Y$ in this example, there must be a decision problem in which the expected Bayes risk from $\mathcal{E}_Y$ is less than that from $\mathcal{E}_X$. The following simple decision problem has this property. Suppose that the hypothesis H_0: $\theta = \theta_2$ is to be tested against the alternative H_1: $\theta \ne \theta_2$ with the usual 0-1 loss function and the prior distribution ξ defined by $\xi(\theta_2) = \lambda$ and $\xi(\theta_1) = \xi(\theta_3) = (1-\lambda)/2$. If λ satisfies $2^{n-1} < \lambda/(1-\lambda) < 2^{n-1} + \frac{1}{2}$, then it can be shown that the Bayes rule for the experiment $\mathcal{E}_Y$ is to reject H_0 if the number of heads is 0 and the Bayes rule for the experiment $\mathcal{E}_X$ is to accept H_0 regardless of the outcome. Since the outcome of $\mathcal{E}_X$ is of no value to the experimenter in this decision problem, it follows that the expected Bayes risk from $\mathcal{E}_X$ is larger than that from $\mathcal{E}_Y$. ∎

Example I shows that Sakaguchi's result (Lemma 1) is stronger than that of Lindley. We believe that Theorem 1 is stronger than Lemma 1, although we do not have an example in which (2.3) is satisfied but not (2.4). Furthermore, we conjecture that the converse of Theorem 1 does not hold.

We shall now consider the implication of the relation $\mathcal{E}_X \geq \mathcal{E}_Y$ in terms of Fisher information. Suppose that Θ is an open subset of the real line R and that the Fisher information $i_X(\theta)$ in the experiment $\mathcal{E}_X$ is defined, as usual, by the relation

$$i_X(\theta) = \int_{-\infty}^{\infty} p(x|\theta) \ [\frac{\partial}{\partial\theta} \log p(x|\theta)]^2 d\mu(x). \tag{2.11}$$

The Fisher information $i_Y(\theta)$ is defined analogously. Stone (1961) proved that if (2.2) was satisfied, then under certain regularity conditions,

$$i_X(\theta) \geq i_Y(\theta) \text{ for all } \theta \in \Theta, \tag{2.12}$$

i.e., $\mathcal{E}_X \geq_F \mathcal{E}_Y$. It follows from Lindley's result that under these same conditions, if $\mathcal{E}_X \geq \mathcal{E}_Y$, then $\mathcal{E}_X \geq_F \mathcal{E}_Y$. We will now generalize this result to the case when $\underset{\sim}{\theta}$ is a parameter vector.

Let $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_k)$ and suppose that Θ is an open subset of R^k. Let $p^{(i)}(x|\underset{\sim}{\theta}) = \frac{\partial}{\partial\theta_i} p(x|\underset{\sim}{\theta}), i=1,2,\ldots,k$. Furthermore, let $\underset{\sim}{i}_X(\underset{\sim}{\theta})$ denote the kxk Fisher information matrix for the experiment X whose $(i,j)^{th}$ term is given by

$$\int_{S_X} \frac{p^{(i)}(x|\underset{\sim}{\theta})p^{(j)}(x|\underset{\sim}{\theta})}{p(x|\underset{\sim}{\theta})} d\mu(x),$$

and let $\underset{\sim}{i}_Y(\underset{\sim}{\theta})$ be defined analogously.

Suppose that $\mathcal{E}_X \geq \mathcal{E}_Y$. If $p(x|\underset{\sim}{\theta})$ and $q(y|\underset{\sim}{\theta})$ satisfy the usual regularity conditions that permit differentiation inside the integral, then it follows from (2.5) that

$$q^{(i)}(y|\underset{\sim}{\theta}) = \int_{S_X} h(y|x)p^{(i)}(x|\underset{\sim}{\theta})d\mu(x), \tag{2.13}$$

for all $\theta \in \Theta$ and $y \in S_Y$. For $b' = (b_1,\ldots,b_k)$, it follows from the definition of $i_X(\theta)$ that

$$b' \, i_X(\theta) b = \int_{S_X} \frac{[\sum_1^k b_i p^{(i)}(x|\theta)]^2}{p(x|\theta)} d\mu(x). \tag{2.14}$$

Let $T = \sum_1^k b_i p^{(i)}(x|\theta)/p(x|\theta)$. By (2.6), we can write

$$\begin{aligned} b' \, i_X(\theta) b &= \int_{S_X} T^2 p(x|\theta) [\int_{S_Y} h(y|x) d\nu(y)] d\mu(x) \\ &= \int_{S_Y} [\int_{S_X} T^2 \frac{p(x|\theta)h(y|x)}{q(y|\theta)} d\mu(x)] q(y|\theta) d\nu(y). \end{aligned} \tag{2.15}$$

Since $p(x|\theta)\, h(y|x)/q(y|\theta)$ is a gpdf [see (2.5)] and $E(T^2) \geq [E(T)]^2$, it follows from (2.15) that

$$b' \, i_X(\theta) b \geq \int_{S_Y} [\int_{S_X} \frac{\sum_1^k b_i p^{(i)}(x|\theta)}{p(x|\theta)} \frac{p(x|\theta)h(y|x)}{q(y|\theta)} d\mu(x)]^2 \, q(y|\theta) d\nu(y). \tag{2.16}$$

Hence, by (2.13)

$$\begin{aligned} b' i_X(\theta) b &\geq \int_{S_Y} \frac{[\sum_1^k b_i q^{(i)}(y|\theta)]^2}{q(y|\theta)} d\nu(y) \\ &= b' \, i_Y(\theta) b. \end{aligned} \tag{2.17}$$

Thus, if $\mathcal{E}_X \geq \mathcal{E}_Y$ and we can differentiate inside the integral sign in (2.5), then (2.17) holds for all $\theta \in \Theta$ and $b \in R^k$. This result is summarized in the following theorem.

Theorem 2. Suppose that Θ is an open subset of R^k. Suppose also that $\mathcal{E}_X \geq \mathcal{E}_Y$ and that the usual regularity conditions which permit differentiation inside the integral sign in (2.5) are satisfied. Then the matrix $i_X(\theta)-i_Y(\theta)$ is non-negative definite for every $\theta \in \Theta$.

Remark 1. It should be noted that if $i_X(\theta)-i_Y(\theta)$ is non-negative definite and $i_Y(\theta)$ is positive definite, then $|i_X(\theta)| \geq |i_Y(\theta)|$ [See Rao (1973), p.70].

In other words, the generalized Fisher information in $\mathcal{E}_X$ is at least as large as that in $\mathcal{E}_Y$.

The following counterexample shows that the converse of Theorem 2 does not necessarily hold.

Example 2. Suppose that $\mathcal{E}_X$ is the experiment $N_n(\mu,\sigma^2)$ and $\mathcal{E}_Y$ is the experiment $N_n(\mu,2\sigma^2+\varepsilon)$, where μ and σ are unknown parameters and ε is a given nonnegative constant. Then $\underset{\sim}{i}_X(\mu,\sigma)-\underset{\sim}{i}_Y(\mu,\sigma)$ is a diagonal matrix with elements $n(\sigma^2+\varepsilon)/[\sigma^2(2\sigma^2+\varepsilon)]$ and $2n\varepsilon(\varepsilon+4\sigma^2)/[\sigma^2(2\sigma^2+\varepsilon)]$, which is positive definite for $\varepsilon > 0$ and nonnegative definite for $\varepsilon = 0$. However, it will be proved in Section 4 that $\mathcal{E}_X$ is not sufficient for $\mathcal{E}_Y$ for any $\varepsilon \geq 0$. ∎

3. Comparison of normal experiments with known mean and unknown variance.

In this section we shall consider experiments in which a random sample can be taken from a normal distribution for which the mean is known and the variance is unknown. Without loss of generality we shall assume that the known value of the mean is 0.

To begin we note that $N_n(0,\sigma^2) \geq N_n(0,\sigma^2 + k^2)$ where k is a given constant. To see this, suppose that the random variable X is distributed as $N(0,\sigma^2)$, the random variable Y is distributed as $N(0,\sigma^2 + k^2)$, and the random variable W is independent of X and has the distribution $N(0,k^2)$. Then X+W has the same distribution as Y for every possible value of σ^2. Hence, $N_1(0,\sigma^2) \geq N_1(0,\sigma^2+k^2)$. However, it is well known that if $\mathcal{E}_X \geq \mathcal{E}_Y$ when only one observation is taken in each experiment, then this same relation holds when a random sample of n observations is taken from each experiment [See DeGroot (1970), p.433]. It now follows that $N_n(0,\sigma^2) \geq N_n(0,\sigma^2+k^2)$.

Next, we note that for any given constant k, the experiments $N_n(0,\sigma^2)$ and $N_n(0,k^2\sigma^2)$ are equivalent in the sense that each is sufficient for the other.

We turn now to the much more difficult problem of determining whether either of the experiments $N_n(0,\sigma^2)$ and $N_n(0,\sigma^{2k})$ is sufficient for the other, where k is a given positive constant. We shall prove that for $k_1 > k_2 > 0$, $N_n(0,\sigma^{2k_1}) \geq N_n(0,\sigma^{2k_2})$. First we obtain some related results.

Lemma 2. Let W be a non-negative random variable with pdf g(w) and let $\varphi(t)$ denote the characteristic function of log W. Let $\theta > 0$ be an unknown parameter, let $k > 0$ be a given constant, and let $G_n(\theta^k)$ denote the experiment in which a random sample of n observations is taken from the distribution with pdf $(1/\theta^k)$ $g(w/\theta^k)$. For any given constant $c > 0$, define

$$\psi_c(t) = \frac{\varphi(t)}{\varphi(ct)}, \quad -\infty < t < \infty. \tag{3.1}$$

If $\psi_{k_2/k_1}(t)$ is a characteristic function, then $G_n(\theta^{k_1}) \geq G_n(\theta^{k_2})$.

Proof. We can assume that $k_2 = 1$ without loss of generality, since we could redefine the parameter to be θ^{k_2}, and then to simplify the notation we replace k_1 by k. Let X and Y denote the observations in the experiments $G_1(\theta^k)$ and $G_1(\theta)$ respectively. Since Y/θ and X/θ^k have distributions identical to that of W, it follows that $\frac{1}{k}\log X - \log\theta$ has the same distribution as $\frac{1}{k}\log W$ and that $\log Y - \log\theta$ has the same distribution as log W. Let Z be a random variable, independent of W and X, with the characteristic function $\psi_{1/k}(t)$ given by (3.1). Then the distribution of $\frac{1}{k}\log W + Z$ is the same as that of log W, and it follows that the distribution of $X^{1/k} e^Z$ is identical to that of Y for every possible value of θ. Therefore $G_1(\theta^k) \geq G_1(\theta)$, which in turn implies that $G_n(\theta^k) \geq G_n(\theta)$. ∎

It should be noted that $\psi_c(t)$, defined in (3.1), is a characteristic function for all $c \in (0,1)$ if and only if $\varphi(t)$ belongs to the class of self-decomposable characteristic functions, introduced by P. Lévy and A. Ya. Khinchine [See Lukacs (1970), §5.11]. Some interesting properties of this class [also called L-class by Gnedenko and Kolmogorov (1954)] are as follows.

(i) All self-decomposable characteristic functions are infinitely divisible.

(ii) If $\varphi(t)$ is a self-decomposable characteristic function; then $\psi_c(t)$ is infinitely divisible.

(iii) All stable characteristic functions are self-decomposable.

(iv) The necessary and sufficient conditions for $\varphi(t)$ to be self-decomposable in terms of Lévy's and Kolmogorov's canonical representations of an infinitely divisible characteristic functions are given in Theorems 1 and 2, of Chapter 6 in Gnedenko and Kolmogorov (1954).

We shall now assume that g is the density function of a gamma distribution G(a,b), defined in (1.1), with known values of a and b, and prove that the assumptions in Lemma 2 hold for this pdf.

Theorem 3. Let $G_n(a,b)$ denote the experiment in which a random sample of n observations is taken from the gamma distribution G(a,b) with pdf (1.1). Then $G_n(a,b\theta^{k_1}) \geq G_n(a,b\theta^{k_2})$, where $\theta > 0$ is an unknown parameter, and a,b,k_1 and k_2 are given positive constants with $k_1 > k_2$.

Proof. Let $\varphi(t)$ denote the characteristic function of $\log_e W$, where W is a random variable with pdf (1.1), and let $\alpha = k_2/k_1$. Then

$$\varphi(t) = \int_0^\infty \frac{1}{b^a\Gamma(a)} w^{it} w^{a-1} e^{-w/b} dw$$

$$= b^{it}\, \Gamma(a+it)/\Gamma(a). \tag{3.2}$$

For $\alpha < 1$, consider

$$\psi_\alpha(t) = \frac{\varphi(t)}{\varphi(\alpha t)} = \frac{b^{it(1-\alpha)}\Gamma(a+it)}{\Gamma(a+i\alpha t)} \quad . \tag{3.3}$$

The Wierstrass expansion of $1/\Gamma(z)$ [See Whittaker and Watson (1935), p.236] is

$$\frac{1}{\Gamma(z)} = z\, e^{\gamma z} \prod_{j=1}^{\infty}\{(1+\frac{z}{j})\exp(-z/j)\} \tag{3.4}$$

where γ is the Euler's constant. Therefore, after some algebraic manipulation, $\psi_\alpha(t)$ can be written as

$$\psi_\alpha(t) = b^{it(1-\alpha)}\exp\{-\gamma it(1-\alpha)\ (\frac{\frac{1}{2}+it\alpha}{\frac{1}{2}+it}) \prod_{j=1}^{\infty} (\frac{j+\frac{1}{2}+it\alpha}{j+\frac{1}{2}+it})\ e^{\frac{it}{j}(1-\alpha)} \tag{3.5}$$

or, equivalently, as

$$\psi_\alpha(t) = \exp\{-it(1-\alpha)(\gamma-\ln b)\}[\alpha+(1-\alpha)(1+\frac{it}{1/2})^{-1}] \prod_{j=1}^{\infty} [\alpha+(1-\alpha)(1+\frac{it}{j+\frac{1}{2}})^{-1}]e^{\frac{it}{j}(1-\alpha)} . \tag{3.6}$$

The first factor in (3.6) is the characteristic function of a degenerate random variable T_0 with probability one at the point $[-(1-\alpha)(\gamma-\ln b)]$, and the factor $e^{it(1-\alpha)/j}$ is the characteristic function of a degenerate random variable T_j with probability one at the point $(1-\alpha)/j, j=1,2,\ldots$. Furthermore, $[\alpha+(1-\alpha)(1+\frac{it}{j+\frac{1}{2}})^{-1}]$ is the characteristic function of a random variable Z_j, $j=0,1,2,\ldots$, which takes the value 0 with probability α and, with probability $(1-\alpha)$, has the pdf

$$f_j(z) = \begin{cases} (j+\frac{1}{2})\exp[(j+\frac{1}{2})z] & \text{for } z \le 0 \\ 0 & \text{for } z > 0 \end{cases} \tag{3.7}$$

Let $\{T_i,\ i=0,1,2,\ldots\}$ and $\{Z_i, i=0,1,2,\ldots\}$ be independent sequences of independent random variables with the distributions defined above. Define $S_j = \sum_{\ell=0}^{j}(T_\ell+Z_\ell)$, and let $\psi_j(t)$ denote the characteristic function of S_j. It follows from (3.6) and the above discussion that

$$\psi_\alpha(t) = \lim_{j\to\infty} \psi_j(t). \tag{3.8}$$

It is obvious that $\psi_\alpha(0) = 0$. Furthermore, since the gamma function $\Gamma(z)$ is analytic for complex arguments except at the points $z=0,-1,-2,\ldots$, where it has simple poles, $\psi_\alpha(t)$ is continuous at $t=0$. Hence by the Continuity Theorem [Theorem 3.6.1, Lukacs (1970)], $\psi_\alpha(t)$ is a characteristic function. It now follows from Lemma 2 that $G_n(a,b\theta^{k_1}) \geq G_n(a,b\theta^{k_2})$ for all $k_1 > k_2 > 0$.

Remark. (i) An alternative way to prove that $\sum_{i=0}^{j} (Z_i + T_i)$ converges in distribution to a random variable Z, is to use Theorem 3.7.3, Lukacs (1970). Since $\sum_{i=0}^{\infty} \mathrm{Var}(Z_i + T_i) = \alpha(1-\alpha) \sum_{i=0}^{\infty} 1/(i+\frac{1}{2})^2\} < \infty$, $\psi_j(t)$ converges to a characteristic function $\psi_\alpha(t)$ as $j \to \infty$.

(ii) Another proof of fact that $\psi_\alpha(t)$ is a characteristic function could be given by proving that $\varphi(t)$ is self-decomposable by using either Theorem 1 or Theorem 2 in Chapter 6 of Gnedenko and Kolmogorov (1954). However, we prefer the proof given above, because it gives the specific form of the random variable Z in the proof of Lemma 2.

(iii) Using Theorem 3.7.6 of Lukacs (1970), it can be shown that the distribution function of the random variable Z is continuous.

We shall now prove the main result of this section.

Theorem 4. Let k_1 and k_2 be given constants satisfying $k_1 > k_2 > 0$. Then $N_n(0,\sigma^{2k_1}) \geq N_n(0,\sigma^{2k_2})$.

Proof. As explained earlier, we can assume without loss of generality that $k_2=1$ and $k_1=k$. Let X and Y denote the observations in the experiments $N_1(0,\sigma^{2k})$ and $N_1(0,\sigma^2)$, respectively. Since X^2/σ^{2k} and Y^2/σ^2 have the same χ^2 distribution, it follows from Theorem 3 that the experiments in which X^2 is observed is sufficient for that in which Y^2 is observed. Furthermore, since X^2 is a sufficient statistic for the experiment $N_1(0,\sigma^{2k})$ and Y^2 is a sufficient statistic for the experiment $N_1(0,\sigma^2)$, it follows that $N_1(0,\sigma^{2k}) \geq N_1(0,\sigma^2)$. Hence, $N_n(0,\sigma^{2k}) \geq N_n(0,\sigma^2)$.

If X and Y have the distributions specified in the proof of Theorem 4, we now know how to generate a random variable equivalent to an observation on Y from an observation on X.

Let the random variable Z be as defined in the proof of Theorem 3, independently of X, with $a=b=\frac{1}{2}$ and $\alpha = k_2/k_1$, and let Y' be defined as follows

$$Y' = \begin{cases} |X|^{\alpha} e^{Z/2} \text{ with probability } \frac{1}{2}, \\ -|X|^{\alpha} e^{Z/2} \text{ with probability } \frac{1}{2}. \end{cases}$$

Then Y' has the same distribution as Y for every possible value of σ^2.

4. Comparison of experiments with location and scale parameters. Stone (1961) considers the class of experiments $\{\mathcal{E}(c); c > 0\}$ where $\mathcal{E}(c)$ is the experiment in which an observation is taken from the pdf $cf[c(x-\theta)]$, for a fixed pdf f and $\Theta = R$. For given values of c_1 and c_2, he obtains conditions under which $\mathcal{E}(c_1) \succeq \mathcal{E}(c_2)$, $\mathcal{E}(c_1) \succeq_L \mathcal{E}(c_2)$, or $\mathcal{E}(c_1) \succeq_F \mathcal{E}(c_2)$. Let $\varphi(t)$ denote the characteristic function of the pdf f. Stone shows that if $f(\cdot)$ is bounded and $c_1 > c_2 > 0$, then a sufficient condition for $\mathcal{E}(c_1) \succeq \mathcal{E}(c_2)$ is that

$$\psi(t) = \frac{\varphi(t/c_2)}{\varphi(t/c_1)} \tag{4.1}$$

be a characteristic function. However, it can be seen from our proof of Lemma 2, that the boundedness of $f(\cdot)$ is not needed in this result. Furthermore, it follows that if $\varphi(t)$ is a self decomposable characteristic function, then $\mathcal{E}(c_1) \succeq \mathcal{E}(c_2)$ for all $c_1 > c_2 > 0$.

Stone also established that if $f(\cdot)$ is bounded and the family of pdf's $\{f(u-\theta);\ \theta \in R\}$ is boundedly complete, then a necessary condition that $\mathcal{E}(c_1) \succeq \mathcal{E}(c_2)$ whenever $c_1 > c_2 > 0$ is that $\psi(t)$ be a characteristic function. In addition, if all the cumulants of $f(\cdot)$ exist, Stone proves that $\psi(t)$ is a characteristic function, only if (i) $f(\cdot)$ is a normal density or (ii) the even-order cumulants of $f(\cdot)$ are positive. After proving this result, he states that "it is possible that condition (ii) is inconsistent with $\mathcal{E}(c_1) \geq \mathcal{E}(c_2)$ whenever $c_1 > c_2$, in which event, yet another characterization of the normal distribution would be provided". However, for $f(u) = \exp(-u), u > 0$, and $c_1 > c_2 > 0$, it can be shown that $\psi(t)$ is a characteristic function and therefore $\mathcal{E}(c_1) \geq \mathcal{E}(c_2)$. Furthermore, all the cumulants of $f(\cdot)$ exist, all the even order cumulants of $f(\cdot)$ are positive, $f(\cdot)$ is bounded, and the family of distributions $\{f(u-\theta);\ \theta \in R\}$ is boundedly complete. Hence, this result does not provide yet another characterization of the normal distribution, as suggested by Stone.

A natural extension of the above results is to consider the class of experiments $\{\mathcal{E}^*(c);\ c > 0\}$ such that $\mathcal{E}^*(c)$ is the experiment in which an observation is taken from the pdf $(c/\sigma)\, f[c(x-\mu)/\sigma]$, where f is a given pdf and the parameter space is $\Theta = \{(\mu,\sigma):\ \mu \in R,\ \sigma > \theta\}$. One may ask whether $\mathcal{E}^*(c_1) \geq \mathcal{E}^*(c_2)$ for $c_1 > c_2 > 0$. In particular, one may ask whether $N_n(\mu,\sigma^2)$ is sufficient for $N_n(\mu,\sigma^2/c^2)$, where $c < 1$ is a known constant. We will prove that the answer is negative, even though $N_n(\mu,\sigma^2) \geq_F N_n(\mu,\sigma^2/c^2)$ as mentioned in Example 2.

For any two given joint prior distributions ξ_1 and ξ_2, let $I_c(\xi_1,\xi_2)$ denote the K-L information contained in the experiment $N_n(\mu,\sigma^2/c^2)$ as defined by (2.1).

<u>Theorem 5</u>. For $c_1 > c_2 > 0$, the experiment $N_n(\mu,\sigma^2/c_1^2)$ is not sufficient for $N_n(\mu,\sigma^2/c_2^2)$.

<u>Proof</u>. Without loss of generality we will assume that $c_1 = 1$ and $c_2 = c < 1$. For $i=1,2$, let ξ_i denote the joint prior distributions of (μ,σ), such that μ has a normal distribution with mean 0 and variance τ^2 and σ takes the value σ_i with probability 1, where $\sigma_1 \neq \sigma_2$ are given positive constants. Since $\bar{X}$ and

$S^2 = \sum_1^n (X_i-\bar{X})^2$ are sufficient statistics for the experiment $N_n(\mu,\sigma^2/c^2)$, the marginal distribution p_i of $\bar{X}$ and S^2, for the experiment $\mathcal{E}^*(c)$, with respect to the prior distribution ξ_i is given by

$$p_i(\bar{X},S^2) = \left(\frac{\sigma^2}{2\sigma_i^2}\right) \frac{\left(\frac{c^2S^2}{2\sigma_i^2}\right)^{\frac{n-1}{2}-1}}{\Gamma\left(\frac{n-1}{2}\right)} \exp\left[-\frac{c^2S^2}{2\sigma_i^2}\right] \frac{1}{\sqrt{2\pi\tau_i(c)}} \exp\left[-\tfrac{1}{2}\frac{\bar{X}^2}{\tau_i(c)}\right], \quad (4.2)$$

where $\tau_i(c) = \tau^2 + \frac{1}{n}\left(\frac{\sigma_i}{c}\right)^2$.

It follows from (2.1) and (4.1) that $I_c(\xi_1,\xi_2)$ is given by

$$I_c(\xi_1,\xi_2) = \left(\frac{n-1}{2}\right)\left\{\log\frac{\sigma_1^2}{\sigma_2^2} - \left(1-\frac{\sigma_1^2}{\sigma_2^2}\right)\right\} + \tfrac{1}{2}\left[\log\frac{\tau_2(c)}{\tau_1(c)} - 1 + \frac{\tau_1(c)}{\tau_2(c)}\right]. \quad (4.3)$$

On taking the partial derivative of $I_c(\xi_1,\xi_2)$ with respect to c, we find that

$$\frac{\partial}{\partial c} I_c(\xi_1,\xi_2) < 0.$$

Therefore, for any $c < 1$,

$$I_1(\xi_1,\xi_2) < I_c(\xi_1,\xi_2). \quad (4.4)$$

Hence, it follows from Theorem 1, that $N_n(\mu,\sigma^2)$ is not sufficient for $N_n(\mu,\sigma^2/c^2)$ for any c such that $0 < c < 1$. ∎

Furthermore, as mentioned after Example 2 in Section 2, it can be shown that the experiment $N_n(\mu,\sigma^2)$ is not sufficient for the experiment $N_n(\mu,2\sigma^2+\varepsilon)$, where ε is a known positive constant. This result can be established through a proof similar to that of Theorem 5 provided only that the prior distributions ξ_1 and ξ_2 are chosen so that $\tau^2 > \varepsilon$.

It should be noted that for the degenerate prior distributions ξ_1 and ξ_2, where ξ_i assigns probability 1 to the point (μ_i,σ_0^2) and $\mu_1 \neq \mu_2$, $I_1(\xi_1,\xi_2) > I_c(\xi_1,\xi_2)$. It follows from this fact and the expression (4.4) that neither of the two experiments $N_n(\mu,\sigma^2/c_1^2)$ and $N_n(\mu,\sigma^2/c_2^2)$ will be sufficient for the other for $c_1 \neq c_2$, even though $N_n(\mu,\sigma^2/c_1^2) \succeq_F N_n(\mu,\sigma^2/c_2^2)$ for $c_1 > c_2 > 0$.

Stone (1961) also proved that in the location parameter case, $\mathcal{E}(c_1) \succeq_F \mathcal{E}(c_2)$ for $c_1 > c_2 > 0$, whenever the Fisher information exists. We shall now extend this result to the family of experiments $\mathcal{E}^*(c)$ with both location and scale parameters, defined at the beginning of this section.

<u>Theorem 6</u>. Suppose that the pdf f(x) is symmetric around x=0 and let $\underset{\sim}{i}_c(\mu,\sigma)$ denote the Fisher information matrix for the experiment $\mathcal{E}^*(c)$. If $\underset{\sim}{i}_c(\mu,\sigma)$ exists for $c=c_1$ and $c=c_2$, with $c_1 > c_2 > 0$, then the matrix $\underset{\sim}{i}_{c_1}(\mu,\sigma)-\underset{\sim}{i}_{c_2}(\mu,\sigma)$ is nonnegative definite for all values of μ and σ.

<u>Proof.</u> It can be shown that if $f(\cdot)$ is symmetric, then the matrix $\underset{\sim}{i}_{c_1}(\mu,\sigma)-\underset{\sim}{i}_{c_2}(\mu,\sigma)$ is a diagonal matrix with diagonal elements $\frac{(c_1^2-c_2^2)}{\sigma^2}\int_R f(u)[\frac{d}{du}\log f(u)]^2$ and 0. Hence, it is nonnegative definite. ■

5. <u>Marginally Sufficient Experiments.</u> In general, the relation $\mathcal{E}_X \succeq \mathcal{E}_Y$ is equivalent to the requirement that $\mathcal{E}_X$ is at least as preferred as $\mathcal{E}_Y$ for <u>every decision problem</u> involving the parameter θ and <u>every prior distribution</u> on Θ. Therefore, it is a very restrictive relation and induces only a partial ordering on the class $E(\Theta)$ of all possible experiments with parameter space Θ.

Feldman (1972) studied certain properties of orderings of $E(\Theta)$ induced by the weakened requirement that in a fixed decision problem, the expected Bayes risk from $\mathcal{E}_X$ be not greater than that from $\mathcal{E}_Y$ for every prior distribution $\xi \in \Xi$. Following DeGroot (1962), he identified the decision problem with an uncertainty function $U(\xi)$ defined on Ξ and considered the experiment $\mathcal{E}_X$ to be at least as informative as the experiment $\mathcal{E}_Y$ with respect to U if $U(\xi|X) \leq U(\xi|Y)$ for all $\xi \in \Xi$, where $U(\xi|X)$ is the expected posterior uncertainty if X is observed and the prior distribution is ξ and $U(\xi|Y)$ is the corresponding value for the observation Y.

An alternative possibility for comparing experiments is to consider a fixed prior distribution ξ and study the ordering on $E(\Theta)$ induced by the requirement that the expected Bayes risk from $\mathcal{E}_X$ be not greater than that from $\mathcal{E}_Y$ for every decision problem involving θ. In this case, we will say that $\mathcal{E}_X$ is at least as informative as $\mathcal{E}_Y$ with respect to the prior distribution ξ.

It is clear that if $\mathcal{E}_X$ is at least as informative as $\mathcal{E}_Y$ with respect to an uncertainty function U, then every experimenter interested in the decision problem corresponding to U will prefer $\mathcal{E}_X$ to $\mathcal{E}_Y$. On the other hand, if $\mathcal{E}_X$ is at least as informative as $\mathcal{E}_Y$ with respect to a prior distribution ξ, then an experimenter with prior distribution ξ on Θ will prefer $\mathcal{E}_X$ to $\mathcal{E}_Y$ regardless of his decision problem. We shall now give an example to illustrate this concept.

Example 3. Let $c_1 > c_2 > 0$ be given constants and for i=1,2, let X_i denote a random variable with the normal distribution $N[\mu, \frac{\sigma^2}{c_2^2}]$. Suppose that the prior distribution of (μ,σ) is concentrated on just two points such that $Pr[(\mu,\sigma) = (0,1)] = \xi$ and $Pr[(\mu,\sigma) = (\mu_0,\sigma_0)] = 1-\xi$ where $0 < \xi < 1$, σ_0 and μ_0 are known and arbitrary. It follows from Bradt and Karlin (1956) that with respect to this two point prior distribution, $N_1(\mu,\sigma^2/c_1^2)$ is sufficient for $N_1(\mu,\sigma^2/c_2^2)$, and therefore $N_n(\mu,\sigma^2/c_1^2) \succeq N_n(\mu,\sigma^2/c_2^2)$ with respect to this two point prior.

Hence $N_n(\mu,\sigma^2/c_1^2)$ is more informative than $N_n(\mu,\sigma^2/c_2^2)$ with respect to this respect to this prior distribution. ∎

This concept of relative informativeness with respect to a prior distribution ξ is especially useful when the parameter $\underset{\sim}{\theta}$ is vector valued, $\underset{\sim}{\theta} = (\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_2)$, and the experimenter is interested only in $\underset{\sim}{\theta}_1$, i.e., $\underset{\sim}{\theta}_2$ is a nuisance parameter. For example, in the experiment $\mathcal{E}^*(c)$ defined in Section 4, corresponding to the pdf $\frac{c}{\sigma} f[\frac{c}{\sigma}(x-\mu)]$, the decision problems of interest may involve only μ or only σ. A detailed discussion on the elimination of nuisance parameters in the framework of classical statistical inference is given by Basu (1977). For a given prior distribution, $\xi(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_2) = \xi_1(\underset{\sim}{\theta}_1)\xi_2(\underset{\sim}{\theta}_2|\underset{\sim}{\theta}_1)$, a Bayesian statistician interested only in $\underset{\sim}{\theta}_1$ will eliminate $\underset{\sim}{\theta}_2$ from the analysis and use the prior pdf $\xi_1(\underset{\sim}{\theta}_1)$ together with the conditional pdf

$$g(x|\underset{\sim}{\theta}_1) = \int_{\Theta_2} p(x|\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_2)d\xi_2(\underset{\sim}{\theta}_2|\underset{\sim}{\theta}_1). \tag{5.1}$$

Consider a particular decision problem with $\Theta = \{\underset{\sim}{\theta} = (\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_2)|\underset{\sim}{\theta}_1 \in \Theta_1, \underset{\sim}{\theta}_2 \in \Theta_2\}$ and a given class D of all possible decisions d, and let $\ell(\underset{\sim}{\theta},d)$ denote the loss incurred from any decision $d \in D$ when $\underset{\sim}{\theta} \in \Theta$ is true. We shall say that the decision problem involves only $\underset{\sim}{\theta}_1$ if, for every pair $(\underset{\sim}{\theta}_1,d)$,

$$\ell[(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_2),d] = \ell[(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_2^*),d] \text{ for all } \underset{\sim}{\theta}_2^* \in \Theta_2, \tag{5.2}$$

i.e., ℓ depends only on the value of $\underset{\sim}{\theta}_1$ and the value of d, and not on the value of $\underset{\sim}{\theta}_2$. For such decision problems, we now present a natural and useful concept of <u>marginal Bayesian sufficiency</u> with respect to a given prior distribution $\xi(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_2)$.

<u>Definition</u>. The experiment $\mathcal{E}_X$ is marginally sufficient for $\mathcal{E}_Y$, denoted by $\mathcal{E}_X \succeq \mathcal{E}_Y(\underset{\sim}{\theta}_1)$, with respect to the prior distribution $\xi(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_2)$ if the expected Bayes risk from $\mathcal{E}_X$ is not greater than that from $\mathcal{E}_Y$ for every decision problem involving only $\underset{\sim}{\theta}_1$, when the prior distribution is $\xi(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_2)$.

If there exists a randomizing function h satisfying (2.6) such that (2.5) holds for all $\underset{\sim}{\theta}_1 \in \Theta_1$, with $p(x|\theta)$ replaced by $g(x|\underset{\sim}{\theta}_1)$ as given in (5.1), and $q(y|\theta)$ replaced by the corresponding conditional pdf of y given $\underset{\sim}{\theta}_1$, then it will be true that $\mathcal{E}_X \succeq \mathcal{E}_Y(\underset{\sim}{\theta}_1)$ with respect to any prior distribution ξ^* that yields the same conditional distribution $\xi_2(\underset{\sim}{\theta}_2|\underset{\sim}{\theta}_1)$ as ξ. In this case, we shall say that $\mathcal{E}_X \succeq \mathcal{E}_Y(\underset{\sim}{\theta}_1)$ with respect to the conditional prior distribution $\xi_2(\underset{\sim}{\theta}_2|\underset{\sim}{\theta}_1)$.

We shall now give some examples of marginal Bayesian sufficiency.

<u>Example 4</u>. For a given pdf f, let $\mathcal{E}^*(c)$ denote the experiment defined in Section 4, and let $\varphi(t) = \int_R e^{itu} f(u)du$. For any joint prior distribution of μ and σ, let $\xi_2(\sigma)$ denote the marginal prior distribution of σ and let $\varphi_1(t) = \int_0^\infty \varphi(t\sigma)d\xi_2(\sigma)$. It follows from (5.1) that if μ and σ are independent under their joint prior distribution, then $g(x|\mu)$ is of the form $cg^*[c(x-\mu)]$. Therefore, if $\varphi_1(t)$ is a self decomposable characteristic function, then it follows from the result of Stone, presented at the beginning of Section 4, that $\mathcal{E}^*(c_1) \succeq \mathcal{E}^*(c_2)(\mu)$ with respect to the conditional prior distribution $\xi_2(\sigma)$ for $c_1 > c_2 > 0$.

In particular, let f(u) be the standard normal pdf and let either (i) the prior density of σ^2 be a gamma distribution of the form $G(\alpha,\beta)$, or (ii) the prior density of $(1/\sigma^2)$ be a gamma distribution of the form $G(\frac{1}{2},\beta)$. By carrying out the analysis indicated in this example, it can be shown that $N_1(\mu,\sigma^2/c_1) \succeq N_1(\mu,\sigma^2/c_2)(\mu)$ with respect to both of these conditional prior distributions of σ^2 given μ. ∎

Example 5. Let $c_1 > c_2 > 0$ be given constants and, for i=1,2, let X_i denote a random variable with the normal distribution $N[\mu,\frac{\sigma^2}{c_i}]$. Suppose that our interest lies in decision problems involving only σ^2. If μ and σ are independent under their joint prior distribution, and if the marginal distribution of μ is a normal distribution $N(m,\tau^2)$, then it follows that, given σ^2, X_i is distributed as $N[m, \frac{\sigma^2}{c_i^2} + \tau^2]$. Let W be distributed as $N[(1-\frac{c_2}{c_1})m, (1-\frac{c_2^2}{c_1^2})\tau^2]$ independently of X_2. Then it can be verified that $(\frac{c_2}{c_1})X_2 + W$ has the same distribution as X_1 for every possible value of σ^2. Hence, $N_1(\mu, \frac{\sigma^2}{c_2^2}) \succeq N_1(\mu,\frac{\sigma^2}{c_1^2})(\sigma^2)$ with respect to this conditional prior distribution of μ. In fact, using the joint distribution of $\overline{X}$ and S^2 from a random sample of n observations given σ^2 [See (4.1)], it can be shown that $N_n(\mu,\frac{\sigma^2}{c_2^2}) \succeq N_n(\mu,\frac{\sigma^2}{c_1^2})(\sigma^2)$ with respect to this conditional prior distribution of μ.

However, if the conditional prior distribution of μ given σ^2 is $N(m,\frac{\sigma^2}{\tau^2})$, then it follows that X_i is distributed as $N[m,\sigma^2(\frac{1}{c_i^2} + \frac{1}{\tau^2})]$, given σ^2. Therefore, the experiments $N_1(\mu,\frac{\sigma^2}{c_1^2})$ and $N_1(\mu,\frac{\sigma^2}{c_2^2})$ are sufficient for each other with respect to this conditional prior distribution of μ. Again, it can be shown, in fact, that the experiments $N_n(\mu,\frac{\sigma^2}{c_1^2})$ and $N_n(\mu,\frac{\sigma^2}{c_2^2})$ are sufficient for each other with respect to this conditional prior distribution of μ. ∎

Example 6. For given constants $c_1 > c_2 > 0$, consider again the normal experiments $N_n(\mu,\frac{\sigma^2}{c_2^2})$, i=1,2, and suppose that the joint prior distribution of μ and σ^2 is a conjugate normal-gamma distribution such that the conditional distribution of μ given σ^2 is $N[m,\frac{\sigma^2}{\tau^2}]$ and the distribution of $(1/\sigma^2)$ is $G(\alpha,\beta)$. It follows from Example 4, that for decision problems involving only σ^2, the experiments $N_n(\mu,\frac{\sigma^2}{c_1^2})$ and $N_n(\mu,\frac{\sigma^2}{c_2^2})$ are marginally equivalent with respect to this joint prior distribution. However, for decision problems involving only μ, it is not known whether one of these experiments is marginally sufficient for the other with respect to this conjugate joint prior distribution. Also, the question of whether one of these experiments is more informative than the other with respect to this joint prior distribution, when one is interested in all decision problems, is open. We can prove, however, that for estimating any of the functions μ, μ/σ, μ/σ^2, $\mu\sigma$ and $\mu\sigma^2$ with squared-error loss, the experiment $N_n(\mu, \frac{\sigma^2}{c_1^2})$ has a smaller expected Bayes risk than the experiment $N_n(\mu, \frac{\sigma^2}{c_2^2})$ for this conjugate prior distribution. ∎

Example 7. If a statistic $T(\underset{\sim}{Y})$ is partially sufficient for the parameter $\underset{\sim}{\theta}_1$ according to Fraser's definition (1956), then it can be proved that the experiment $\mathcal{E}_T$, in which only T is observed, satisfies $\mathcal{E}_T \geq \mathcal{E}_{\underset{\sim}{Y}} (\underset{\sim}{\theta}_1)$ with respect to any prior distribution $\xi(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_2)$ under which $\underset{\sim}{\theta}_1$ and $\underset{\sim}{\theta}_2$ are independent. For example, if $Y_1,\ldots,Y_n$ are independent and identically distributed with a gamma distribution $G(\alpha,\beta)$, then $T = \sum_1^n Y_i$ is partially sufficient for β in Fraser's sense and therefore, $\mathcal{E}_T \geq \mathcal{E}_{\underset{\sim}{Y}}(\beta)$ with respect to any prior distribution for which α and β are independent. ∎

References

[1] Basu, D. (1977). On the elimination of nuisance parameters, Part I. To appear in Jour. Amer. Stat. Assoc.

[2] Blackwell, D. (1951). Comparison of experiments, Proc. of the 2nd Berk. Symp. on Mathematical Statistics and Probability, Univ. of California Press. 93-102.

[3] Blackwell, D. (1953). Equivalent comparison of experiments. Ann. Math. Statist. **24**, 265-272.

[4] Bradt, R. N. and Karlin, S. (1956). On the design and comparison of certain dichotomons experiments. Ann. Math. Statist. **27**, 390-409.

[5] DeGroot, M. H. (1962). Uncertainty, information and sequential experiments. Ann. Math. Statist. **33**, 404-419.

[6] DeGroot, M. H. (1966). Optimal allocation of observations. Ann. Inst. Statist. Math. **18**, 13-28.

[7] DeGroot, M. H. (1970). *Optimal Statistical Decisions*, McGraw-Hill, N.Y.

[8] Feldman, D. (1972). Some properties of Bayesian ordering of experiments. Ann. Math. Statist. **43**, 1428-1440.

[9] Fraser, D. A. S. (1956). Sufficient statistics with nuisance parameters. Ann. Math. Statist. **27**, 838-842.

[10] Gnedenko, B. V. and Kolmogorov, A. N. (1954). *Limit Theorems for Sums of Independent Random Variables*, Addison-Wesley, Inc., Reading, Mass.

[11] Kullback, S. (1968). *Information Theory and Statistics*, Dover Publications Inc., N.Y.

[12] LeCam, L. (1964). Sufficiency and approximate sufficiency. Ann. Math. Statist. **35**, 1419-1455.

[13] Lindley, D. V. (1956). On a measure of the information provided by an experiment. Ann. Math. Statist. **27**, 986-1005.

[14] Lukacs, E. (1970). *Characteristic Functions*, 2nd ed. Hafner, N.Y.

[15] Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed., Wiley, N.Y.

[16] Sakaguchi, M. (1964). Information theory and decision making. Unpublished lecture notes, Statistics Department, The George Washington Univ., Washington, D.C.

[17] Stone, M. (1961). Non-equivalent comparisons of experiments and their use for experiments involving location parameters. Ann. Math. Statist. **32**, 326-332.

[18] Torgersen, E. N. (1972). Comparison of translation experiments. Ann. Math. Statist. **43**, 1383-1399.

[19] Torgersen, E. N. (1977). Prediction sufficiency when the loss function does not depend on the unknown parameter. Ann. Statist. **5**, 155-163.

[20] Whittaker, E. T. and Watson, G. N. (1935). A Course on Modern Analysis, 4th ed., Cambridge University Press, Cambridge.