Feb. 24, 2003

CHAPTER 2. SUFFICIENCY AND ESTIMATION

2.1 Sufficient statistics. Classically, a "statistic" is a measurable function of the observations, say $f(X_1, \ldots, X_n)$. The concept of "statistic" differs from that of "random variable" in that a random variable is defined on a probability space, so that one probability measure is singled out, where for statistics we have in mind a family of possible probability measures,

R. A. Fisher (1922) called a statistic sufficient "when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter to be estimated." For example, given a sample $(X_1, \ldots, X_n)$ where the X_j are i.i.d. $N(\theta, 1)$, the sample mean $\overline{X} := (X_1 + \cdots X_n)/n$ turns out to be a sufficient statistic for the unknown parameter θ. J. Neyman (1935) gave one form of a "factorization theorem" for sufficient statistics. This section will give more general forms of the definitions and theorem due to Halmos and L. J. Savage (1949).

In general, given a measurable space $(S, \mathcal{B})$, that is a set S with a σ-algebra $\mathcal{B}$ of subsets, and another measurable space $(Y, \mathcal{F})$, a *statistic* is a measurable function T from S into Y. Often, $Y = \mathbb{R}$ or a Euclidean space $\mathbb{R}^d$ with Borel σ-algebra.

Let $T^{-1}(\mathcal{F}) := \{T^{-1}(F) : F \in \mathcal{F}\}$. Then $T^{-1}(\mathcal{F})$ is a σ-algebra and is the smallest σ-algebra on S for which T is measurable.

For any measure μ on $\mathcal{B}$, $\mathcal{L}^1(\mu) := \mathcal{L}^1(S, \mathcal{B}, \mu)$ denotes the set of all real-valued, $\mathcal{B}$-measurable, μ-integrable functions on S. For any probability measure P on $(S, \mathcal{B})$, sub-σ-algebra $\mathcal{A} \subset \mathcal{B}$, and $f \in \mathcal{L}^1(P)$, we have a *conditional expectation* $E_P(f|\mathcal{A})$, defined as a function $g \in \mathcal{L}^1(S, \mathcal{A}, P)$, so that g is measurable for $\mathcal{A}$, such that for all $A \in \mathcal{A}$, $\int_A g dP = \int_A f dP$. Such a g always exists, as a Radon-Nikodym derivative of the signed measure $A \mapsto \int_A f \, dP$ with respect to the restriction of P to $\mathcal{A}$ (RAP, Theorems 5.5.4, 10.1.1). If $f = 1_B$ for some $B \in \mathcal{B}$ let $P(B|\mathcal{A}) := E_P(1_B|\mathcal{A})$. Then almost surely $0 \le P(B|\mathcal{A}) \le 1$. Conditional expectations for P are only defined up to equality P-almost surely. Now, we will need the fact that suitable $\mathcal{A}$-measurable functions can be brought outside conditional expectations (RAP, Theorem 10.1.9), much as constants can be brought outside ordinary expectations:

2.1.1 Lemma. $E_P(fg|\mathcal{A}) = fE_P(g|\mathcal{A})$ whenever both g and fg are in $\mathcal{L}^1(S, \mathcal{B}, P)$ and f is $\mathcal{A}$-measurable.

Now, here are precise definitions of sufficiency:

Definition. Given a family $\mathcal{P}$ of probability measures on $\mathcal{B}$, a sub-σ-algebra $\mathcal{A} \subset \mathcal{B}$ is called *sufficient* for $\mathcal{P}$ if and only if for every $B \in \mathcal{B}$ there is an $\mathcal{A}$-measurable function $f_B \ge 0$ such that $f_B = P(B|\mathcal{A})$ P-almost surely for every $P \in \mathcal{P}$. A statistic T from $(S, \mathcal{B})$ to $(Y, \mathcal{F})$ is called *sufficient* if and only if $T^{-1}(\mathcal{F})$ is sufficient.

The σ-algebra $\mathcal{A}$ is *pairwise sufficient* for $\mathcal{P}$ iff for every $P, Q \in \mathcal{P}$ the likelihood ratio $R_{Q/P}$ is $\mathcal{A}$-measurable; more precisely, $R_{Q/P}$ can be chosen to be $\mathcal{A}$-measurable, and any choice of $R_{Q/P}$ is equal $(P + Q)$-almost everywhere to such an $\mathcal{A}$-measurable function.

If, for example, $\mathcal{A}$ is finite, generated by a finite partition, and is sufficient for $\mathcal{P}$, then on each atom A of $\mathcal{A}$, f_B, being $\mathcal{A}$-measurable, must be constant, and

$$Q_A(B) \ := \ \int_A f_B dP/P(A) \ = \ P(A \cap B)/P(A) \ = \ P(B|A)$$

is the same for all $P \in \mathcal{P}$ with $P(A) > 0$. Or if $P(A) = 0$ for all $P \in \mathcal{P}$ but A is non-empty, choose any $y \in A$ and let $Q_A \ := \ \delta_y$ (point mass at y). In either case, Q_A is a probability measure concentrated on A such that for all $B \in \mathcal{B}$ and $P \in \mathcal{P}$, $P(B \cap A) = Q_A(B)P(A)$. This is a simple case of "factorization" into a product where one factor depends on B but not P and the other on P but not B. Then to estimate P, once we know x is in the atom A of $\mathcal{A}$, it can give no further information about P in $\mathcal{P}$, since each P in $\mathcal{P}$ conditioned on A reduces to the one law Q_A. Also, for any μ and ν in $\mathcal{P}$, the likelihood ratio $R_{\nu/\mu}$ is constant on the atom A.

Some facts on measurable functions will be needed. For a σ-algebra $\mathcal{S}$ and set A (not necessarily in $\mathcal{S}$) let $\mathcal{S}_A$ be the "relative" σ-algebra $\{C \cap A : \ C \in \mathcal{S}\}$. For any functions f, g such that f is defined on the range of g let $f \circ g$ denote the composition, $f \circ g(x) \ := \ f(g(x))$ for all x in the domain of g. Recall (RAP, Theorems 4.2.5, 4.2.8):

2.1.2 Theorem. Let $(Y, \mathcal{S})$ be any measurable space and A an arbitrary subset of Y. Let f be any real-valued function on A measurable for $\mathcal{S}_A$. Then there is an $\mathcal{S}$-measurable extension of f to all of Y.

2.1.3 Theorem. Given a set S, a measurable space $(Y, \mathcal{F})$, and a function T from S into Y, a real-valued function f on S is $T^{-1}(\mathcal{F})$ measurable if and only if $f = g \circ T$ for some $\mathcal{F}$-measurable function g.

Theorem 2.1.2 is immediate if A is in $\mathcal{S}$, since f can be given any constant value on the complement of A. But sets A not in $\mathcal{S}$ can arise: for example, the range of a Borel measurable function $T : \ \mathbb{R} \to \mathbb{R}$ need not be Borel (RAP, Theorem 13.2.5).

Recall, as in Section 1.3, that if a family $\mathcal{P} = \{P_\theta, \ \theta \in \Theta \ \}$ of laws is dominated by a σ-finite measure μ, meaning that every $P \in \mathcal{P}$ is absolutely continuous with respect to μ, then there exist Radon-Nikodym derivatives $dP/d\mu$, which give us likelihood functions $f(\theta, x) \ := \ (dP_\theta/d\mu)(x)$, $\theta \in \Theta$, $x \in S$, and allow us to write in integrals $dP_\theta(x) = f(\theta, x)d\mu(x)$.

Neyman (1935), Halmos and Savage (1949), and Bahadur (1954) proved the equivalence of (a) and (b) for dominated families in the following theorem.

2.1.4 Theorem (Factorization theorem). Let $(S, \mathcal{B})$ be a measurable space, $\mathcal{A}$ a sub-σ-algebra of $\mathcal{B}$, and $\mathcal{P}$ a non-empty family of probability measures on $\mathcal{B}$.
(I) If $\mathcal{P}$ is dominated by a σ-finite measure μ, the the following are equivalent:

(a) $\mathcal{A}$ is sufficient for $\mathcal{P}$;

(b) there is a $\mathcal{B}$-measurable function $h \geq 0$ such that for all $P \in \mathcal{P}$, there is an $\mathcal{A}$-measurable function f_P with $dP/d\mu \ = \ f_P h$. We can take $h \in \mathcal{L}^1(S, \mathcal{B}, \mu)$.

(c) $\mathcal{A}$ is pairwise sufficient for $\mathcal{P}$.

(II) In general, if $\mathcal{P}$ is not necessarily dominated, (a) always implies (c).

Theorem 2.1.4 has the following consequence, via Theorem 2.1.3:

2.1.5 Corollary. Let $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ be a non-empty family of laws on the sample space $(S, \mathcal{B})$, dominated by a σ-finite measure μ, and let T be a statistic on $(S, \mathcal{B})$. Then the following are equivalent:

(a′) T is sufficient for $\mathcal{P}$;

(b′) The likelihood functions $f(\theta, x)$ can be written as $G(\theta, T(x))h(x)$, where for each θ, $G(\theta, \cdot)$ is a measurable function from Y to $\mathbb{R}$, and h is a measurable function on $(S, \mathcal{B})$;

(c′) For any $P, Q \in \mathcal{P}$, the likelihood ratio $R_{Q/P}$ can be written as a measurable function of $T(x)$.

Thus, when the parameter space contains just two points, giving laws P and Q, the likelihood ratio $R_{Q/P}$ is a sufficient statistic, since $R_{P/Q} = 1/R_{Q/P}$. Before beginning the proofs, here is an example. Consider the family of gamma distributions with scale parameter equal to 1, namely, let μ be Lebesgue measure on $(0, \infty)$ and let $f(\theta, x) := x^{\theta-1}e^{-x}/\Gamma(\theta)$ for $x > 0$ and 0 for $x \leq 0$, where $0 < \theta < \infty$. Then the likelihood function for n i.i.d. observations $X_1, ..., X_n$ is

$$\Pi_{j=1}^n f(\theta, X_j) \ = \ (X_1 X_2 \cdots X_n)^{\theta-1} \exp\left(\textstyle\sum_{j=1}^n X_j\right) / \Gamma(\theta)^n.$$

It follows from Corollary 2.1.5 that $X_1 X_2 \cdots X_n$ is a sufficient statistic for the family. The function $h(x) := \exp(X_1 + \cdots + X_n)$ in this case doesn't depend on θ, thus it divides out in forming likelihood ratios.

A first step in the proof of Theorem 2.1.4, that (b) implies (c) in (I), is easy: for any $P, Q \in \mathcal{P}$, (b) implies that we can take $R_{Q/P}(x) = (f_Q(x)h(x))/(f_P(x)h(x)) = f_Q(x)/f_P(x)$, which is $\mathcal{A}$-measurable. We use the usual conventions for likelihood ratios, $y/0 := +\infty$ if $y > 0$ and 0 if $y = 0$. From (b), $P(\{x :\ h(x) = 0\}) = 0$ for all $P \in \mathcal{P}$, so $h(x)/h(x)$ does not become indeterminate except on negligible sets.

To continue the proof of Theorem 2.1.4 we will need some facts about dominated families. Two measures are called *equivalent* if each is absolutely continuous with respect to the other. This notion extends to families of laws as follows.

Definition. Two families $\mathcal{P}$ and $\mathcal{Q}$ of probability measures on a σ-algebra $\mathcal{B}$ are called *equivalent* if and only if for all $B \in \mathcal{B}$, $(P(B) = 0$ for all $P \in \mathcal{P})$ is equivalent to $(Q(B) = 0$ for all $Q \in \mathcal{Q})$. If $\mathcal{Q}$ consists of just one law Q we will say $\mathcal{P}$ is equivalent to Q.

2.1.6 Lemma. Assume that a non-empty family $\mathcal{P}$ of probability measures is dominated by a σ-finite measure μ. Then:

(a) $\mathcal{P}$ is dominated by a probability measure μ'.

(b) $\mathcal{P}$ is equivalent to some probability measure μ_0.

(c) There is a countable subfamily $\{Q_k\}_{k\geq 1}$ of $\mathcal{P}$ equivalent to $\mathcal{P}$.

(d) $\mathcal{P}$ is equivalent to a probability measure ν given by $\nu = \sum_{k=1}^\infty Q_k/2^k$ for some sequence $\{Q_k\}_{k\geq 1} \subset \mathcal{P}$.

Proof. Part (d) implies the other parts, which are steps in its proof.

(a): If $\mu(S) < \infty$ let $\mu'(B) := \mu(B)/\mu(S)$ for all $B \in \mathcal{B}$. Otherwise, $S = \bigcup_{i=1}^\infty A_i$ where $0 < \mu(A_i) < \infty$ for each i and $A_i \cap A_j = \emptyset$ for $i \neq j$. For each $B \in \mathcal{B}$ let $\mu'(B) := \sum_{i=1}^\infty \mu(B \cap A_i)/(2^i \mu(A_i))$. Then in either case μ' is a probability measure equivalent to μ, and both dominate $\mathcal{P}$.

(b): Take $\mu = \mu'$ from (a). Let $\gamma := \sup\{\mu(C) : P(C) = 0 \text{ for all } P \in \mathcal{P}\}$. Take $C_i \in \mathcal{B}$ with $P(C_i) = 0$ for all $P \in \mathcal{P}$ and $\mu(C_i) \uparrow \gamma$. Let $C := \bigcup_i C_i$. Then $\mu(C) = \gamma$ and $P(C) = 0$ for all $P \in \mathcal{P}$. We have $\mu(C^c) > 0$ since μ dominates $\mathcal{P}$. Let $\mu_0(B) := \mu(B \cap C^c)/\mu(C^c)$, a probability measure. Then clearly μ_0 dominates $\mathcal{P}$. On the other hand if $P(A) = 0$ for all $P \in \mathcal{P}$ then $\mu_0(A) = 0$, otherwise we could replace C by $C \cup A$ and violate the definition of γ. So μ_0 is equivalent to $\mathcal{P}$.

(c): Take $\mu = \mu_0$ from (b). For any $\{P_i\}_{i\geq 1} \subset \mathcal{P}$ let $\gamma(\{P_i\}_{i\geq 1}) := \sup\{\mu(C) : P_i(C) = 0 \text{ for all } i\}$. Let $\alpha := \inf\{\gamma(\{P_i\}_{i\geq 1}) : \{P_i\}_{i\geq 1} \subset \mathcal{P}\}$. For $j = 1, 2, ...,$ take $\{P_{ji}\}_{i\geq 1}$ such that $\gamma(\{P_{ji}\}_{i\geq 1}) < \alpha + (1/j)$. Let $\{Q_k\}_{k\geq 1} = \{P_{ji}\}_{j\geq 1, i\geq 1}$. Then $\gamma(\{Q_k\}_{k\geq 1}) = \alpha$. By the proof of (b), for some $C \in \mathcal{B}$, $\mu(C) = \alpha$ and $Q_k(C) = 0$ for all k. It will be shown that $\alpha = 0$. If not, $\alpha > 0$. Then by (b), for some $P \in \mathcal{P}$, $P(C) > 0$. Letting $Q_0 := P$ we have $\gamma(\{Q_k\}_{k\geq 0}) \leq \mu(\{x \in C : (dQ_0/d\mu)(x) = 0\}) < \alpha$, a contradiction. So $\alpha = 0$ and $\mathcal{P}$ is equivalent to $\{Q_k\}_{k\geq 1}$.

It is straightforward that (c) implies (d), so the Lemma is proved. □

Now to continue the proof of Theorem 2.1.4, we have already proved (b) implies (c) in (I). To prove (c) implies (a) and (b), take $\nu = \sum_{k=1}^{\infty} Q_k/2^k$ from Lemma 2.1.6(d). For any $P \in \mathcal{P}$, we have $R_{\nu/P} = \sum_{k=1}^{\infty} R_{Q_k/P}/2^k$, which is $\mathcal{A}$-measurable. Thus, so is $dP/d\nu = R_{P/\nu} = 1/R_{\nu/P}$, with the usual conventions $1/\infty := 0$, $1/0 := +\infty$. Thus we can write

$$\frac{dP}{d\mu} = \frac{dP}{d\nu}\frac{d\nu}{d\mu},$$

which gives (b) since $dP/d\nu$ is $\mathcal{A}$-measurable, with $h := d\nu/d\mu \in \mathcal{L}^1(\mu)$.

Now to prove (a), for each $B \in \mathcal{B}$, let $f_B := \nu(B\mathcal{A}) := E_\nu(1_B|\mathcal{A})$. For any $P \in \mathcal{P}$, since $dP/d\nu$ is $\mathcal{A}$-measurable we have by Lemma 2.1.1 that

$$E_\nu\left(\frac{dP}{d\nu}1_B|\mathcal{A}\right) = \frac{dP}{d\nu}f_B.$$

Using this and the definition of conditional expectation twice, we have for any $A \in \mathcal{A}$,

$$\int_A f_B dP = \int_A f_B \frac{dP}{d\nu}d\nu = \int_A E_\nu\left(\frac{dP}{d\nu}1_B|\mathcal{A}\right)d\nu = \int_A 1_B\frac{dP}{d\nu}d\nu = \int_A 1_B dP.$$

Thus f_B satisfies the definition of $P(B|\mathcal{A})$, proving (a).

(a) implies (b): Again take ν from Lemma 2.1.6(d). For any $B \in \mathcal{B}$ take f_B from (a). Since $0 \leq f_B \leq 1$ P-almost surely for all P in $\mathcal{P}$, also $0 \leq f_B \leq 1$ almost surely for ν. We have $Q_k(B \cap A) = \int_A f_B dQ_k$ for all k and all $A \in \mathcal{A}$, so $\nu(B \cap A) = \int_A f_B d\nu$ so $f_B = E_\nu(1_B|\mathcal{A})$.

To show that for each $P \in \mathcal{P}$, $dP/d\nu$ is $\mathcal{A}$-measurable, note that for each B in $\mathcal{B}$, by the definition of likelihood ratio and since $S \in \mathcal{A}$),

$$\int_B \frac{dP}{d\nu}d\nu = P(B) = \int f_B dP = \int f_B\frac{dP}{d\nu}d\nu = \int E_\nu\left(f_B\frac{dP}{d\nu}|\mathcal{A}\right)d\nu,$$

which by Lemma 2.1.1 twice equals

$$\int E_\nu(1_B|\mathcal{A})E_\nu\left(\frac{dP}{d\nu}|\mathcal{A}\right)d\nu \;=\; \int E_\nu\left(1_B E_\nu\left(\frac{dP}{d\nu}|\mathcal{A}\right)|\mathcal{A}\right)d\nu \;=\; \int 1_B E_\nu\left(\frac{dP}{d\nu}|\mathcal{A}\right)d\nu.$$

Since $dP/d\nu$ and $E_\nu(dP/d\nu|\mathcal{A})$ have the same integral over all sets in $\mathcal{B}$, they must be equal P-a.s., so that $dP/d\nu$ is equal P-a.s. to an $\mathcal{A}$-measurable function. So (a) implies (b), again with $h = d\nu/d\mu \in \mathcal{L}^1(\mu)$ since ν is a probability measure, and part (I) of Theorem 2.1.4 is proved.

For (II), to show (a) implies (c) in general, if $\mathcal{P}$ is not necessarily dominated, take any $P, Q \in \mathcal{P}$. Then $\{P, Q\}$ is dominated, e.g. by $(P+Q)/2$, and $\mathcal{A}$ is sufficient for it, so by (a) implies (c) in the dominated case, we get that (c) holds. Theorem 2.1.4 is proved. □

The next fact gives an example showing that pairwise sufficiency does not imply sufficiency in general.

2.1.7 Proposition. There exists a family of laws $\mathcal{P}$ on a sample space $(S, \mathcal{B})$ and a sub-σ-algebra $\mathcal{A} \subset \mathcal{B}$ which is pairwise sufficient for $\mathcal{P}$ but not sufficient.

Proof. Let $S = [0, 1]$ with Borel σ-algebra $\mathcal{B}$. Let $\mathcal{A}$ be the sub-σ-algebra of countable sets and their complements. Let $\mathcal{P}$ be the family of all purely atomic laws on S. Thus $P \in \mathcal{P}$ if and only if $P(M) = 1$ for some countable set M. Let $P, Q \in \mathcal{P}$ and take a countable $A \subset S$ such that $P(A) = Q(A) = 1$. Defining $R_{Q/P}(x) := 0$ for $x \notin A$, it is clear that $R_{Q/P}$ is $\mathcal{A}$-measurable, so $\mathcal{A}$ is pairwise sufficient for $\mathcal{P}$.

Let $B := [0, 1/2]$. For any $P \in \mathcal{P}$ and the countable set $A := \{x : P(\{x\}) > 0\}$, we must have $E_P(1_B|\mathcal{A})(x) = 1_B(x)$ for $x \in A$. If f_B has the property in the definition of sufficiency then $f_B(x) = 1_B(x)$ for all x, since for all x there is some $P \in \mathcal{P}$ with $P(\{x\}) > 0$. But 1_B is not $\mathcal{A}$-measurable, so $\mathcal{A}$ is not sufficient for $\mathcal{P}$. □

Note that in the example in the last proof, $\mathcal{P}$ is dominated by a measure c, namely counting measure on $[0, 1]$, and derivatives dP/dc exist and are $\mathcal{A}$-measurable, but c is not σ-finite.

Sufficiency, defined in terms of conditional probabilities of measurable sets, can be extended to suitable conditional expectations:

2.1.8 Theorem. Let $\mathcal{A}$ be sufficient for a family $\mathcal{P}$ of laws on a measurable space $(S, \mathcal{B})$. Then for any measurable real-valued function f on $(S, \mathcal{B})$ which is integrable for each $P \in \mathcal{P}$, there is an $\mathcal{A}$-measurable function g such that $g = E_P(f|\mathcal{A})$ a.s. for all $P \in \mathcal{P}$.

Proof. When f is the indicator function of a set in $\mathcal{B}$, the assertion is the definition of sufficiency. It then follows for any simple function, which is a finite linear combination of such indicators. If f is nonnegative, there is a sequence of nonnegative simple functions increasing up to f and the conclusion follows (RAP, Proposition 4.1.5 and Theorem 10.1.7). Then any f satisfying the hypothesis can be written as $f = f^+ - f^-$ for f^+ and f^- nonnegative and the result follows. □

Rather generally, for a sufficient σ-algebra $\mathcal{A}$, the $\mathcal{A}$-measurable decision rules are an essentially complete class. The idea is that for any decision rule d, since conditional probabilities $P(\cdot|\mathcal{A})(x)$ don't depend on $P \in \mathcal{P}$, given x, we choose y at random in S according to

this conditional distribution, then take decision $d(y)$, thus giving an $\mathcal{A}$-measurable randomized decision rule $e(x)$ which has the same risk function as d. For conditional probabilities allowing such a choice to be made we will need some regularity conditions such as the following.

A measurable space $(S, \mathcal{B})$ will be called *standard* if there is a 1-1 function, measurable with measurable inverse, from S onto a Borel set in a complete separable metric space or, equivalently, onto a complete separable metric space (RAP, Sec. 13.1). Most sample spaces considered in statistics are standard, where the Borel set is an open or closed subset of a Euclidean space $\mathbb{R}^k$. On a class $\mathcal{P}$ of laws on $(S, \mathcal{B})$, recall that as defined just before (1.2.3), $\mathcal{S}_\mathcal{B}$ is the smallest σ-algebra making the functions $P \mapsto P(B)$ measurable for all $B \in \mathcal{B}$.

2.1.9 Theorem. Let $\mathcal{P}$ be a class of probability measures on a standard measurable space $(S, \mathcal{B})$, dominated by a σ-finite measure μ, and $\mathcal{A}$ a sub-σ-algebra sufficient for $\mathcal{P}$. Then for any action space $(A, \mathcal{E})$, any measurable loss function L on $\mathcal{P} \times A$, and any (possibly randomized) decision rule d, there exists a decision rule e, measurable on $(S, \mathcal{A})$, with risk $r(P, e) = r(P, d)$ for all $P \in \mathcal{P}$. So, the $\mathcal{A}$-measurable decision rules form an essentially complete class.

Proof. The conditional probability $P(B|\mathcal{A})(x)$ for $P \in \mathcal{P}$, $B \in \mathcal{B}$ and $x \in X$ is initially defined for each fixed B up to P-almost sure equality with respect to x. It will be called a *regular conditional probability* if for each $B \in \mathcal{B}$, $P(B|\mathcal{A})(\cdot)$ is a specific $\mathcal{A}$-measurable function such that for each x, $P(\cdot|\mathcal{A})(x)$ is a countably additive probability measure on $\mathcal{B}$. Regular conditional probabilities exist in this case (RAP, Theorem 10.2.2, with $Y =$ the identity and $\mathcal{A}$, $\mathcal{C}$ there $= \mathcal{B}$, $\mathcal{A}$ here respectively). Take ν from Lemma 2.1.6(d) and a regular conditional probability $\nu(\cdot|\mathcal{A})(\cdot)$. By sufficiency of $\mathcal{A}$ and the proof of Theorem 2.1.4, (c) $\Rightarrow$ (a), for any $P \in \mathcal{P}$, $\nu(\cdot|\mathcal{A})(\cdot)$ is also a regular conditional probability for P, so we can write it as $P(\cdot|\mathcal{A})(\cdot)$. Given a randomized decision rule d, with values in the set $D_\mathcal{E}$ of probability measures on $(A, \mathcal{E})$, let e be such a decision rule where for each $F \in \mathcal{E}$ and $x \in S$,

$$e(x)(F) := \int d(y)(F) P(dy|\mathcal{A})(x),$$

and for a measure m, $m(dy) := dm(y)$. In other words the function $y \mapsto d(y)(F)$, which is assumed $\mathcal{B}$-measurable, is integrated with respect to the probability measure $P(\cdot|\mathcal{A})(x)$. So the integral is well-defined. For fixed F, $e(x)(F)$ is an $\mathcal{A}$-measurable function of x, in fact equal to $E(d(\cdot)(F)|\mathcal{A})$ (RAP, Theorem 10.2.5). For each x, $e(x)(\cdot)$ is a countably additive probability measure on $\mathcal{E}$ since $d(y)(\cdot)$ is for each y and $P(\cdot|\mathcal{A})(x)$ is a probability measure. So $e(\cdot)$ is a well-defined $\mathcal{A}$-measurable randomized decision rule, and for each $P \in \mathcal{P}$ and $x \in S$,

$$\int L(P, a) e(x)(da) = \int_{y \in S} \int_{a \in A} L(P, a) d(y)(da) P(dy|\mathcal{A})(x),$$

so

$$r(P,e) \;=\; \int_{x\in S}\int_{y\in S}\int_{a\in A} L(P,a)d(y)(da)P(dy|\mathcal{A})(x)dP(x).$$

Now for any $f \in \mathcal{L}^1(P)$, by RAP, Theorem 10.2.5 again,

$$\int_{x\in S}\int_{y\in S} f(y)P(dy|\mathcal{A})(x)dP(x) \;=\; \int E_P(f|\mathcal{A})(x)dP(x) \;=\; \int f(x)dP(x).$$

Omitting the middle term, the equation holds for any measurable $f \geq 0$, taking integrable $f_n := \min(f,n)$ with $0 \leq f_n \uparrow f$. Applying this to $f(u) := \int_{a\in A} L(P,a)d(u)(da)$, we get $r(P,e) = r(P,d)$. Since P was any member of $\mathcal{P}$, the proof is complete. □

Examples. In each of the following $S = \mathbb{R}^n$, $\mathcal{B}$ = the Borel σ-algebra, and $\mathcal{P}$ is a set $\mathcal{P}(n,\mathcal{M})$ of distributions μ^n, in other words laws for which the coordinates $x_1,\ldots,x_n$ are i.i.d. with distribution μ, and where $\mu \in \mathcal{M}$, a set of laws on $\mathbb{R}$. In each case T is a sufficient statistic.

(i) (Binomial family). Let $\mathcal{M} = \{(1-p)\delta_0 + p\delta_1 : \; 0 \leq p \leq 1\}$ and $T(x_1,\ldots,x_n) = x_1 + \cdots + x_n$.

(ii) (Poisson family). Let $\mathcal{M} = \{P_\lambda : \; 0 \leq \lambda < \infty\}$ where $P_\lambda(k) = e^{-\lambda}\lambda^k/k!$ for $k = 1,2,\ldots,$ and $P_\lambda(B) = 0$ whenever B contains no nonnegative integers. Again let $T(x_1,\ldots,x_n) = x_1 + \cdots + x_n$.

(iii) (Normal laws with variable location and fixed variance). Let $\mathcal{M} = \{N(\theta,1) : \; \theta \in \mathbb{R}\}$, with the same T.

(iv) (Fixed location and variable scale). Let $\mathcal{M} = \{N(0,\sigma^2) : \; 0 < \sigma^2 < \infty\}$, and $T(x_1,\ldots,x_n) = \sum_{j=1}^n x_j^2$.

(v) (Order statistics). Let $\mathcal{M}$ be the set of all Borel probability measures on $\mathbb{R}$ and let $T(x_1,\ldots,x_n) := (x_{(1)},\ldots,x_{(n)})$ where $x_{(1)},\ldots,x_{(n)}$ are $x_1,\ldots,x_n$ arranged into non-decreasing order, $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

PROBLEMS

1. Show that the given statistics are, in fact, sufficient for the binomial and Poisson families.
2. Likewise for the normal location and normal scale families.
3. Show that the n-tuple of order statistics is pairwise sufficient for the family of all laws P^n on $\mathbb{R}^n$ where P runs over the class of all laws on $\mathbb{R}$. (It is actually sufficient, but that is harder to prove, see problems 5 and 6.)
4. For $-\infty < a < b < \infty$ let $U[a,b]$ be the uniform distribution on the interval $[a,b]$, with density $1/(b-a)$ on $[a,b]$ and 0 elsewhere. Let $\mathcal{U}$ be the family of all such laws. Show that for the set of all laws P^n on $\mathbb{R}^n$ for $P \in \mathcal{U}$, $(x_{(1)}, x_{(n)})$, the minimum and maximum of the observations, provide a sufficient statistic (with values in $\mathbb{R}^2$).
5. Let $(S,\mathcal{B})$ be a sample space. Let $\mathcal{P}_0$ be the family of all laws on $(S,\mathcal{B})$. Let S^n be the Cartesian product of n copies of S with product σ-algebra $\mathcal{B}^n$. Let $\mathcal{P} := \{P^n : \; P \in \mathcal{P}_0\}$ be the set of all laws on S^n for which the coordinates $X_1, ..., X_n$ are i.i.d. Let $\mathcal{S}_n$ be the

σ-algebra of all sets in $\mathcal{B}^n$ invariant under all permutations of coordinates. Show that $\mathcal{S}_n$ is sufficient for $\mathcal{P}$. *Hint*: For any π in the set Π_n of all permutations of $\{1, ..., n\}$, and $x := (x_1, ..., x_n)$, let $f_\pi(x) := (x_{\pi(1)}, ..., x_{\pi(n)})$. Show that for any $B \in \mathcal{B}^n$ and $P \in \mathcal{P}$, $P(B|\mathcal{S}_n)(x) = \frac{1}{n!} \sum_{\pi \in \Pi_n} 1_B(f_\pi(x))$.

6. If $S = \mathbb{R}$ in the previous problem, with Borel σ-algebra, show that the vector $T := (x_{(1)}, ..., x_{(n)})$ of order statistics is a sufficient statistic. *Hint:* show that $\mathcal{S}_n$ is the smallest σ-algebra for which T is measurable.

NOTES

Fisher (1922) invented the idea of sufficiency and Neyman (1935) gave a first form of factorization theorem. Halmos and Savage (1949) proved the factorization theorem (2.1.4) in case $h \in \mathcal{L}^1(S, \mathcal{B}, \mu)$. Bahadur (1954) removed the restriction on h.

Regarding measurability of the ranges of measurable functions, as in Theorem 2.1.3, Darst (1971) showed that even an infinitely differentiable (C^∞) function can take a Borel set onto a non-Borel set.

REFERENCES

Bahadur, R. R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist.* **25**, 423-462.

Darst, R. B. (1971). C^∞ functions need not be bimeasurable. *Proc. Amer. Math. Soc.* **27**, 128-132.

Fisher, Ronald Aylmer (1922). IV. On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London* Ser. A **222**, 309-368.

Halmos, Paul R., and Leonard Jimmie Savage (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Statist.* **20**, 225-241.

Neyman, Jerzy (1935). Su un teorema concernente le cosidette statistiche sufficienti. *Giorn. Ist. Ital. Attuari* **6**, 320-334.