March 22, 2003

2.3 Minimal sufficiency and the Lehmann-Scheffé property. If a statistic T, for example a real-valued statistic, is sufficient for a family $\mathcal{P}$ of laws, then for any other statistic U, say with values in $\mathbb{R}^k$, the statistic (T, U) with values in $\mathbb{R}^{k+1}$ is also sufficient. In terms of σ-algebras, if the family $\mathcal{P}$ is defined on a σ-algebra $\mathcal{B}$ and a sub-σ-algebra $\mathcal{A}$ is sufficient for $\mathcal{P}$, then any other σ-algebra $\mathcal{C}$ with $\mathcal{A} \subset \mathcal{C} \subset \mathcal{B}$ is also sufficient. But since the idea of sufficiency is data reduction, one would like to have a sufficient σ-algebra as small as possible, or a sufficient statistic of dimension as small as possible.

A σ-algebra $\mathcal{A}$ will be called *minimal sufficient* for $\mathcal{P}$ if it is sufficient and for any sufficient σ-algebra $\mathcal{C}$, and each $A \in \mathcal{A}$, there is a $C \in \mathcal{C}$ such that $1_A = 1_C$ a.s. for each $P \in \mathcal{P}$. So, $\mathcal{A}$ is included in $\mathcal{C}$ up to almost sure equality of sets. Then, a statistic T with values in a measurable space $(Y, \mathcal{F})$ will be called *minimal sufficient* iff $T^{-1}(\mathcal{F})$ is a minimal sufficient σ-algebra.

Example. Let $\mathcal{P}$ be a family of symmetric laws on $\mathbb{R}$, such as the set of all normal laws $N(0, \sigma^2)$, $\sigma > 0$. Considering $n = 1$ for simplicity, the identity function x is (always) a sufficient statistic, but it is not minimal sufficient in this case, where $|x|$ is also sufficient.

For dominated families, a minimal sufficient σ-algebra always exists:

2.3.1 Theorem (Bahadur). Let $\mathcal{P}$ be a family of laws on a measurable space $(S, \mathcal{B})$, dominated by a σ-finite measure μ. Then there is always a minimal sufficient σ-algebra $\mathcal{A}$ for $\mathcal{P}$. Also, there is such a σ-algebra $\mathcal{A}$ containing all sets B in $\mathcal{B}$ for which $P(B) = 0$ for all $P \in \mathcal{P}$, and such an $\mathcal{A}$ is unique.

Proof. Take a law ν equivalent to $\mathcal{P}$ from Lemma 2.1.6(d). Choose densities $dP/d\nu$ for all $P \in \mathcal{P}$ and let $\mathcal{A}$ be the smallest σ-algebra for which all the $dP/d\nu$ are measurable. Then by Theorem 2.1.4, $\mathcal{A}$ is sufficient. Next, let $\mathcal{C}$ be any sufficient σ-algebra for $\mathcal{P}$. Let $\mathcal{A}_1$ be the collection of sets A in $\mathcal{A}$ for which there exists a $C \in \mathcal{C}$ with $1_C = 1_A$ a.s. for every $P \in \mathcal{P}$. Then $\mathcal{A}_1$ is a σ-algebra, since if $1_A = 1_C$ a.s. for all $P \in \mathcal{P}$, the same is true for the complements, with $1 - 1_A = 1 - 1_C$, and if $1_{A(j)} = 1_{C(j)}$ a.s. for all P the same is true for the union of the sequences $A(j)$ and $C(j)$. By the proof of Theorem 2.1.4, (c) implies (b), each $dP/d\nu$ must equal a $\mathcal{C}$-measurable function a.s. (ν). Thus the sets $\{dP/d\nu > t\}$ for each $P \in \mathcal{P}$ and real number t are in $\mathcal{A}_1$. Since these sets generate $\mathcal{A}$ (RAP, Theorem 4.1.6), $\mathcal{A}_1 = \mathcal{A}$ and $\mathcal{A}$ is minimal sufficient.

By choice of ν, the collection $\mathcal{Z}$ of sets B (in $\mathcal{B}$) with $P(B) = 0$ for all $P \in \mathcal{P}$ is the same as $\{B \in \mathcal{B} : \nu(B) = 0\}$. The σ-algebra $\mathcal{Y}$ generated by $\mathcal{Z}$ and $\mathcal{A}$ is easily seen to be minimal sufficient. If we start with any other minimal sufficient σ-algebra $\mathcal{C}$ in place of $\mathcal{A}$, it follows easily from the minimal sufficiency of both $\mathcal{A}$ and $\mathcal{C}$ that the resulting $\mathcal{Y}$ will be the same. So $\mathcal{Y}$ is uniquely determined. $\square$

The σ-algebra $\mathcal{Y}$ just treated may be called "*the* minimal sufficient σ-algebra," although as a collection of sets it is actually the largest of all minimal sufficient σ-algebras.

An idea closely related to minimal sufficiency is the Lehmann-Scheffé property, as follows:

Definition. Given a collection $\mathcal{P}$ of laws on a measurable space $(S, \mathcal{B})$, a sub-σ-algebra $A \subset \mathcal{B}$ will be called a *Lehmann-Scheffé (LS)* σ-algebra for $\mathcal{P}$ iff whenever f is an $\mathcal{A}$-measurable function with $\int f dP = 0$ for all $P \in \mathcal{P}$, we have $f = 0$ a.s. for all $P \in \mathcal{P}$. A statistic will be called an LS statistic for $\mathcal{P}$ iff the smallest σ-algebra for which it is measurable is LS for $\mathcal{P}$.

Lehmann and Scheffé called σ-algebras satisfying their property *complete*. This is different from the notion of complete class of decision rules. Also, in measure theory, a σ-algebra $\mathcal{S}$ may be called complete for a measure μ if it contains all subsets of sets of μ-measure 0. The Lehmann-Scheffé property is, evidently, quite different. So, it seemed appropriate to name it here after its authors. It is equivalent to uniqueness of $\mathcal{A}$-measurable unbiased estimators:

2.3.2 Theorem. A sub-σ-algebra $\mathcal{A}$ is LS for $\mathcal{P}$ if and only if for every real-valued function g on $\mathcal{P}$ having an unbiased $\mathcal{A}$-measurable estimator, the estimator is unique up to equality a.s. for all $P \in \mathcal{P}$.

Proof. The constant function 0 always trivially has an unbiased estimator by the statistic which is identically 0 (and so measurable for any $\mathcal{A}$). Uniqueness of this estimator up to equality a.s. for all $P \in \mathcal{P}$ yields the definition of the LS property. Conversely if $\mathcal{A}$ is LS for $\mathcal{P}$, suppose T and U are both $\mathcal{A}$-measurable and both unbiased estimators of a function g on $\mathcal{P}$. Then $T - U$ has integral 0 for all $P \in \mathcal{P}$, so $T - U = 0$ a.s. and $T = U$ a.s. for all $P \in \mathcal{P}$. □

Some σ-algebras are LS just because they are small. For example, the trivial σ-algebra $\{\emptyset, S\}$ is always LS. For any measurable set A, the σ-algebra $\{\emptyset, A, A^c, S\}$ is LS for $\mathcal{P}$ unless $P(A)$ is the same for all P in $\mathcal{P}$. So LS σ-algebras will be interesting only when they are large enough. One useful measure of being large enough is sufficiency. If a function g on $\mathcal{P}$ has an unbiased estimator U and $\mathcal{A}$ is a sufficient σ-algebra, then $T = E_P(U|\mathcal{A})$, which doesn't depend on $P \in \mathcal{P}$ by Theorem 2.1.8, is an unbiased, $\mathcal{A}$-measurable estimator as in Corollary 2.2.3 and Theorem 2.3.2.

From here on, the LS property will be considered for sufficient σ-algebras. These must be minimal sufficient:

2.3.3 Theorem. For any collection $\mathcal{P}$ of laws on a measurable space $(S, \mathcal{B})$, any LS, sufficient σ-algebra $\mathcal{C}$ is minimal sufficient.

Proof. If not, there is a sufficient σ-algebra $\mathcal{A}$ and a set $C \in \mathcal{C}$ such that there is no set A in $\mathcal{A}$ for which $1_C = 1_A$ a.s. for all $P \in \mathcal{P}$. Let $f := E_P(1_C|\mathcal{A})$ for all $P \in \mathcal{P}$ by Theorem 2.1.8. For some $P \in \mathcal{P}$, f is not equal to 1_C a.s. (P), otherwise letting $A = \{f = 1\}$ would give a contradiction. We have $\int (1_C - f) f dP = 0$, as can be seen by taking the conditional expectation of the integrand with respect to $\mathcal{A}$ and bringing f outside the conditional expectation by Lemma 2.1.1 as in the proof of Theorem 2.1.4; or, see RAP, Theorem 10.2.9 (conditional expectation is an orthogonal projection in L^2). It follows from this orthogonality that

$$P(C) \ = \ \int 1_C^2 dP \ = \ \int (1_C - f)^2 dP + \int f^2 dP > \int f^2 dP.$$

Let $g := E_P(f|\mathcal{C})$. This actually doesn't depend on P since $\mathcal{C}$ is also sufficient. Then $1_C - g$ is a $\mathcal{C}$-measurable function whose integral over all of S is 0 for every $P \in \mathcal{P}$ since S is in both $\mathcal{A}$ and $\mathcal{C}$. It is not possible that $1_C = g$ a.s., since $\int g^2 dP \leq \int f^2 dP < P(C)$. This contradicts the fact that $\mathcal{C}$ is LS. □

From Theorems 2.3.1 and 2.3.3 we see that for a dominated family $\mathcal{P}$ there is an LS sufficient σ-algebra if and only if the minimal sufficient σ-algebra is LS. It may not be:

2.3.4 Proposition. There exists a density f on $\mathbb{R}$ and n such that if P_θ has density $x \mapsto f(x-\theta)$ for $x \in \mathbb{R}$ and $\theta \in \mathbb{R}$, the family of laws P_θ^n, $\theta \in \mathbb{R}$, has the n-tuple of order statistics $(X_{(1)}, \ldots, X_{(n)})$ as a minimal sufficient statistic which is not LS.

Proof. For any densities $f(x, \theta)$, the density for n i.i.d. variables is

$$\Pi_{j=1}^n f(X_j, \theta) = \Pi_{j=1}^n f(X_{(j)}, \theta),$$

so the order statistics are sufficient. Take the Cauchy density $f(x) := 1/(\pi(1+x^2))$ as $f(x)$ for all $x \in \mathbb{R}$. For any n, we can take a measure ν equivalent to $\{P_\theta^n : \theta \in \mathbb{R}\}$ as P_0^n. Then the density of P_θ^n with respect to ν is

$$\Pi_{j=1}^n (1 + X_{(j)}^2)/(1 + (X_{(j)} - \theta)^2),$$

the reciprocal of a polynomial $J(\theta)$ of degree $2n$ in θ. By the proof of Theorem 2.3.1, a minimal sufficient σ-algebra is the smallest σ-algebra making these functions measurable in $X_1, \ldots, X_n$ for each θ. The coefficients of the 0th through $2n$th powers of θ are determined by the values of J at any $2n+1$ values of θ, say $\theta = 0, 1, \ldots, 2n$, and are linear and thus Borel measurable functions of these values, and so are $\mathcal{A}$-measurable. The roots of J are the complex numbers $X_{(j)} \pm i$ for $j = 1, \ldots, n$. The order statistics $X_{(j)}$ are Borel measurable functions of the coefficients, and so are also $\mathcal{A}$-measurable, as follows: first, $X_{(1)}$ is measurable because it is the limit of the sequence t_k where t_k is the infimum of rational numbers r such that $|J(r+i)| < 1/k$. Recall that for any non-constant polynomial J, $|J(z)| \to \infty$ as $|z| \to \infty$, so $|J|$ only has arbitrarily small values near roots of J. Once we have found $X_{(1)}$, we can divide the polynomial J by $1 + (X_{(1)} - \theta)^2$ to get a new polynomial of degree $2n-2$ in θ whose coefficients are $\mathcal{A}$-measurable, and iterate to get that all the order statistics $X_{(1)}, \ldots, X_{(n)}$ are $\mathcal{A}$-measurable, so the order statistics are indeed minimal sufficient.

Now for $n \geq 4$, $E(X_{(n-1)} - X_{(2)})$ is finite by Problem 5(c) and is a constant c_n not depending on θ, so $X_{(n-1)} - X_{(2)} - c_n$ is a non-zero $\mathcal{A}$-measurable function with expectation 0 for all θ, so $\mathcal{A}$ is not LS. □

2.3.5 Theorem. If $\mathcal{A}$ is an LS, sufficient σ-algebra for $\mathcal{P}$ and g is a real-valued function on $\mathcal{P}$ for which an unbiased estimator exists, then there is an $\mathcal{A}$-measurable unbiased estimator T for g, unique up to almost sure equality for all $P \in \mathcal{P}$, and which attains the minimum possible risk for unbiased estimators simultaneously for all convex loss functions and all $P \in \mathcal{P}$.

Proof. Let U be any unbiased estimator of g and $T = E(U|\mathcal{A})$, which doesn't depend on $P \in \mathcal{P}$ by Theorem 2.1.8. Then T is an $\mathcal{A}$-measurable estimator and is unbiased for g

since $E_P E_P(U|\mathcal{A}) = E_P U$ for each P from the definition of conditional expectation. T is unique with these properties up to almost sure equality by Theorem 2.3.2. Here T could have been obtained as $E(V|\mathcal{A})$ for any other unbiased estimator V, and so by Corollary 2.2.3, T attains the minimum possible risk for unbiased estimators for any $P \in \mathcal{P}$ and any convex loss function. □

In the example in Prop. 2.3.4 for $n = 6$, $\overline{X}_{2,5} := (X_{(2)} + X_{(5)})/2$ and $\overline{X}_{3,4} := (X_{(3)} + X_{(4)})/2$ are two different unbiased estimators of θ, both measurable for the minimal sufficient σ-algebra, which is not LS. An unbiased estimator measurable for a minimal sufficient σ-algebra doesn't necessarily attain minimum risk for such estimators for any θ. In fact, $\overline{X}_{2,5}$ has infinite risk for squared-error loss, while $\overline{X}_{3,4}$ has finite risk, by Problem 5(d). In such cases there is no LS sufficient σ-algebra $\mathcal{A}$ since by Theorem 2.3.3, $\mathcal{A}$ would be minimal and by Theorem 2.3.1, for a family dominated by a σ-finite measure, the minimal sufficient σ-algebra is essentially unique.

Theorem 2.2.2 (the Rao-Blackwell theorem) shows that estimators become no worse for convex loss functions by conditioning (taking conditional expectation) with respect to a sufficient σ-algebra, but it is not clear when conditioning makes estimators strictly better. For that we need the notion of strict convexity. A function f from a convex set C into $\mathbb{R}$ is called *strictly convex* if for any $x \neq y$ in C and $0 < u < 1$ we have

$$f(ux + (1-u)y) \;<\; uf(x) + (1-u)f(y).$$

Thus $f(x) := x^2$ is strictly convex but $f(x) := |x|$ is not, specifically when x and y have the same sign.

2.3.6 Theorem. Assume given a decision problem where the action space A is a convex Borel subset of some $\mathbb{R}^k$ and the loss function $W(P, \cdot)$ is strictly convex on A and Borel measurable, for each $P \in \mathcal{P}$. Let $\mathcal{A}$ be a sufficient σ-algebra for $\mathcal{P}$ and U a decision rule such that $\int \|U\| dP < \infty$ for all $P \in \mathcal{P}$ and such that for some $P \in \mathcal{P}$, U is not equal P-almost surely to an $\mathcal{A}$-measurable function. Then for such a P, $T := E_P(U|\mathcal{A})$, if it has finite risk $r(P, T)$, has strictly smaller risk for P and W than U has.

Proof. Recall that $E_Q(U|\mathcal{A})$ doesn't depend on $Q \in \mathcal{P}$ by Theorem 2.1.8. We will first have:

2.3.7 Theorem. Let C be a convex Borel set in some $\mathbb{R}^k$. Let $(\Omega, \mathcal{A}, P)$ be a probability space and X a random variable on Ω with values in C, $E|X| < \infty$ and such that X is not a constant a.s.. Let f be a strictly convex, Borel measurable real-valued function on C, such that $E|f(X)| < \infty$. Then $EX \in C$,

(a) $Ef(X) > f(EX)$, and

(b) If $\mathcal{C}$ is any sub-σ-algebra of $\mathcal{A}$, then $E(X|\mathcal{C}) \in C$ a.s., and if X is not equal almost everywhere to a $\mathcal{C}$-measurable function, then $E(f(X)|\mathcal{C}) > f(E(X|\mathcal{C}))$ with probability > 0.

Proof. From Jensen's inequality (RAP, Theorem 10.2.6) and its proof, $EX \in C$ and there is a constant c and a linear function g such that $f(x) \geq c - g(x)$ for all $x \in C$ and $f(EX) = c - g(EX)$. If $y \in C$, $y \neq EX$ and $f(y) = c - g(y)$, then $x := (y + EX)/2 \in C$

and $f(x) < (f(EX) + f(y))/2 = c - g(x)$, a contradiction. So $f(y) > c - g(y)$. Thus if X is not constant a.s., $Ef(X) > c - Eg(X) = f(EX)$. So (a), a strict form of Jensen's inequality, holds.

Then, the conditional Jensen inequality (RAP, Theorem 10.2.7) and its proof apply since a Borel set in a Polish space (complete separable metric space, here $\mathbb{R}^k$) is Borel-isomorphic to a Polish space (RAP, Theorem 13.1.1), so regular conditional probabilities for X given $\mathcal{C}$ exist (RAP, Theorem 10.2.2). If X is not equal a.s. to a $\mathcal{C}$-measurable function, specifically to $E(X|\mathcal{C})$, then with positive probability, the regular conditional probabilities are not concentrated at single points. Since the conditional expectation can be defined by integrating with respect to regular conditional probabilities (RAP, Theorem 10.2.5), it follows from part (a) that $E(f(X)|\mathcal{C}) > f(E(X|\mathcal{C}))$ with positive probability. □

Now, the conclusion of Theorem 2.3.6 clearly holds if $r(P, U) = +\infty$, or if $r(P, U) < \infty$, it follows directly from Theorem 2.3.7. □

2.3.8 Corollary. Under the conditions of Theorem 2.3.6, if there exists a decision rule U with $\int \|U\| dQ < \infty$ and $r(Q, U) < \infty$ for all $Q \in \mathcal{P}$, then the $\mathcal{A}$-measurable decision rules, and those equal to them a.s. for all P, form a complete class.

Note that Theorem 2.3.6 and Corollary 2.3.8 are not limited to unbiased estimators or to estimators at all; they hold for general decision rules. Also, $E(U|\mathcal{A})$ improves on U not only for an individual loss function but simultaneously for all convex loss functions. By taking a minimal sufficient σ-algebra $\mathcal{A}$, as is possible for dominated families by Theorem 2.3.1, under the hypotheses of Corollary 2.3.8, we get a complete class of decision rules for strictly convex loss functions. In other words, decision rules which are not $\mathcal{A}$-measurable (up to almost sure equality for all $P \in \mathcal{P}$) are inadmissible. By Theorem 2.3.3 an LS, sufficient σ-algebra will be minimal sufficient, so that if a statistic T is LS and sufficient and is an unbiased estimator of a function $g(\theta)$, it is optimal among unbiased estimators, while among general estimators, we can limit the choice to functions of T.

PROBLEMS

1. If $\mathcal{P} = \{P_1, \dots, P_k\}$ is a finite set of laws on $(X, \mathcal{B})$ and all P_j are absolutely continuous with respect to P_1, show that $x \mapsto (dP_2/dP_1, \dots, dP_k/dP_1)(x)$ is a minimal sufficient statistic for $\mathcal{P}$.
2. Let $\mathcal{U}$ be the set of all uniform distributions on intervals $[a, b] \subset \mathbb{R}$ for $a < b$. In Sec. 2.1, Problem 4 was to show that for n i.i.d. observations from a distribution in $\mathcal{U}$, the smallest and largest order statistics $(X_{(1)}, X_{(n)})$ form a sufficient statistic (with values in $\mathbb{R}^2$). Show that this statistic is minimal sufficient.
3. For a fixed $h > 0$, let $\mathcal{U}_h$ be the set of all uniform distributions on intervals $[\theta, \theta + h]$, $\theta \in \mathbb{R}$.
   (a) Show that $(X_{(1)}, X_{(n)})$ is also minimal sufficient for this family.
   (b) Show that this minimal sufficient statistic is not LS: give two different unbiased estimators for θ which are both functions of $(X_{(1)}, X_{(n)})$. *Hint*: If $X_1, ..., X_n$ are i.i.d. $U[0, 1]$, then $P(X_{(n)} \le x) = x^n$ for $0 \le x \le 1$. Thus $X_{(n)}$ has density nx^{n-1} for $0 \le x \le 1$ and $EX_{(n)} = n/(n + 1)$. Likewise, $EX_{(1)} = 1/(n + 1)$.

4. Show that for the family $N(0, \sigma^2)$, $\sigma > 0$, on $\mathbb{R}$, for $n = 1$ (the example before Theorem 2.3.10), $|x|$ is a minimal sufficient statistic and for n i.i.d. variables $X_1, \ldots, X_n$, a minimal sufficient statistic is $X_1^2 + \cdots + X_n^2$.
5. (a) If X has the standard Cauchy density $1/[\pi(1 + x^2)]$ for $-\infty < x < \infty$, show that $P(X > x) \sim 1/(\pi x)$ as $x \to +\infty$, where $f \sim g$ means $f/g \to 1$.

   (b). If $X_1, ..., X_n$ are i.i.d. standard Cauchy, show that for each $k = 1, ..., n$, $P(X_{(k)} \leq -x) = P(X_{(n+1-k)} \geq x) \sim \binom{n}{k}/(\pi x)^k$ as $x \to +\infty$.

   (c) In part (b), show that $E|X_{(k)}| < \infty$ if and only if $1 < k < n$. *Hint:* for a random variable $Y \geq 0$ with density g, $EY = \int_0^\infty yg(y)dy = \int_0^\infty g(y) \int_0^y dt\, dy = \int_0^\infty \int_t^\infty g(y)dy\, dt = \int_0^\infty P(Y \geq t)dt$.

   (d) Similarly, show that $E(X_{(k)}^2) < \infty$ if and only if $3 \leq k \leq n - 2$.

NOTES

The Lehmann-Scheffé property, which they called completeness, is due to Lehmann and Scheffé (1950, 1955,1956).

REFERENCES

Lehmann, Erich L., and Scheffé, Henry (1950, 1955, 1956). Completeness, similar regions and unbiased estimation. *Sankhyā* **10**, 305-340; **15**, 219-236; Correction **17**, 50.