2.5 Exponential families. These will be families $\{P_\theta,\ \theta \in \Theta\}$ of laws, including many of the best-known special families such as the binomial and normal laws, and for which there is a natural vector-valued sufficient statistic, whose dimension stays constant as the sample size n increases, and which has the Lehmann-Scheffé property.

Definition. A family $\mathcal{P} = \{Q_\psi :\ \psi \in \Psi\}$ of laws on a measurable space $(X, \mathcal{B})$, containing at least two different laws, is called an *exponential family* if there exist a σ-finite measure μ on $(X, \mathcal{B})$, a positive integer k, and real functions θ_j on Ψ and measurable h with $0 < h(x) < \infty$ and T_j on X for $j = 1, \dots, k$, such that for all $\psi \in \Psi$, Q_ψ is absolutely continuous with respect to μ, and for some $C(\theta(\psi)) > 0$, where $\theta(\psi) := (\theta_1(\psi), \dots, \theta_k(\psi))$,

$$(2.5.1) \qquad (dQ_\psi/d\mu)(x) = C(\theta(\psi))h(x)\exp(\textstyle\sum_{j=1}^k \theta_j(\psi)T_j(x)).$$

If we replace μ by ν where $d\nu(x) = h(x)d\mu(x)$, the factor $h(x)$ can be omitted, and ν is still a σ-finite measure. Given the θ_j, T_j, h, and μ, the number $C(\theta(\psi))$ is determined by normalization, so it is, in fact, a function of $\theta(\psi) := \{\theta_j(\psi)\}_{j=1}^k$. Thus, given T_j, h, and μ, Q_ψ is determined by the values of $\theta_j(\psi)$.

It follows from the factorization theorem, Corollary 2.1.5, that for any exponential family, the vector-valued statistic $(T_1(x), \dots, T_k(x))$ is a sufficient statistic. The structure of an exponential family is essentially preserved by taking n i.i.d. observations, as follows. Let $\{Q_\psi,\ \psi \in \Psi\}$ be any exponential family and let $X_1, \dots, X_n$ be i.i.d. (Q_ψ). Then the distribution Q_ψ^n of $(X_1, \dots, X_n)$ is an exponential family for the σ-finite measure μ^n on X^n, replacing $T_j(x)$ by $\sum_{i=1}^n T_j(X_i)$, $h(x)$ by $\Pi_{j=1}^n h(X_j)$, and $C(\theta(\psi))$ by $C(\theta(\psi))^n$. It follows that for n i.i.d. observations from the exponential family, the k-vector $\{\sum_{i=1}^n T_j(X_i)\}_{j=1}^k$ is a sufficient statistic.

Since exponentials are strictly positive, any exponential family is an equivalent family as defined in the last section. The T_j will be called *affinely dependent* if for some constants $c_0,\ c_1, \dots, c_k$, not all 0, $c_0 + c_1T_1 + \cdots + c_kT_k = 0$ almost everywhere for μ. Then $c_i \neq 0$ for some $i \geq 1$, and we can solve for T_i as a linear combination of other T_j and a constant. Then we can eliminate the T_i term and reduce k by 1, adding constants times $\theta_i(\cdot)$ to each $\theta_j(\cdot)$ for $j \neq i$. Iterating this, we can assume that $T_1, \dots, T_k$ are affinely independent, i.e. they are not affinely dependent. Likewise, we can define affine independence for the functions θ_j, where now the linear relations among the $\theta_j(\cdot)$ and a constant would hold everywhere rather than almost everywhere (at this point we are not assuming a prior given on the parameter space Ψ). We can eliminate terms until $\theta_j(\cdot)$ are also affinely independent. We will always still have $k \geq 1$ since $\mathcal{P}$ contains at least two laws.

Let Θ be the range of the function $\psi \mapsto \theta(\psi) := (\theta_1(\psi), \dots, \theta_k(\psi))$ from Ψ into $\mathbb{R}^k$. Then clearly $\theta_1(\cdot), \dots, \theta_k(\cdot)$ are affinely independent if and only if Θ is not included in any $(k-1)$-dimensional hyperplane in $\mathbb{R}^k$. Likewise, $T_1, \dots, T_k$ are affinely independent (as defined above) if and only if for $T := (T_1, \dots, T_k)$ from X into $\mathbb{R}^k$, the measure $\mu \circ T^{-1}$ is not concentrated in any $(k-1)$-dimensional hyperplane in $\mathbb{R}^k$. For each $\theta \in \Theta$, let P_θ be the law on X with $(dP_\theta/d\mu)(x) = C(\theta)h(x)e^{\theta\cdot T(x)}$ where $\theta \cdot T := \sum_{j=1}^k \theta_jT_j$. Then $Q_\psi = P_{\theta(\psi)}$ for all $\psi \in \Psi$ and $\mathcal{P} = \{P_\theta :\ \theta \in \Theta\}$.

A representation (2.5.1) of an exponential family will be called *minimal* if $T_1, \dots, T_k$ are affinely independent, as are $\theta_1(\cdot), \dots, \theta_k(\cdot)$.

A *functionoid* is an equivalence class of functions for the relation of almost sure equality for a measure. The well-known Banach spaces L^p of p-integrable functions, such as the Hilbert spaces L^2 of square-integrable functions, are actually spaces of functionoids. For an exponential family, or any other equivalent family, almost sure equality is the same for P_θ for all θ.

2.5.2 Theorem. Every exponential family $\mathcal{P} := \{Q_\psi : \psi \in \Psi\}$ has a minimal representation (2.5.1), and then k is uniquely determined.

Proof. We already saw that the $T_j(\cdot)$ can be taken to be affinely independent, as can the $\theta_j(\cdot)$, so that the representation (2.5.1) is minimal. As we also saw, the family $\mathcal{P}$ can be written as $\{P_\theta,\ \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$, where

$$(dP_\theta/d\mu)(x) \ = \ C(\theta)h(x)e^{\theta \cdot T(x)}.$$

Then the likelihood ratios are all of the form

$$R_{\theta,\phi} \ := \ R_{P_\theta/P_\phi} \ = \ C(\theta)C(\phi)^{-1}\exp\{\textstyle\sum_{j=1}^k(\theta_j - \phi_j)T_j(x)\}.$$

The logarithms of these likelihood ratios (log likelihood ratios) plus constants span a real vector space V_T of functionoids on X, included in the vector space W_T of functionoids spanned by $1, T_1, \ldots, T_k$. Then W_T is $(k+1)$-dimensional since $T_1, \ldots, T_k$ are affinely independent by minimality. Also, since $\theta_1, \ldots, \theta_k$ are affinely independent on Θ, $V_T = W_T$. Now $V := V_T$ is determined by the family $\mathcal{P}$, not depending on the choice of μ or T, so V and k are uniquely defined for the family $\mathcal{P}$. □

The number k will be called the *order* of the exponential family. From here on it will be assumed that the representation of an exponential family is minimal unless it is specifically said not to be.

Any exponential family $\mathcal{P}$ can be parameterized by a subset of $\mathbb{R}^k$, replacing $\theta_j(\psi)$ by θ_j, with $\Theta = \{\theta(\psi) : \psi \in \Psi\}$, and

$$(2.5.3) \qquad (dP_\theta/d\mu)(x) \ = \ C(\theta)h(x)\exp(\textstyle\sum_{j=1}^k \theta_j T_j(x)), \quad \theta \in \Theta \subset \mathbb{R}^k,$$

where now $Q_\psi = P_{\theta(\psi)}$ for all $\psi \in \Psi$. The parameterization in (2.5.3) is then one-to-one:

2.5.4 Theorem. If an exponential family has a minimal representation (2.5.3), then for any $\theta \neq \phi$ in Θ, $P_\theta \neq P_\phi$.

Proof. If $P_\theta = P_\phi$, then for $\theta \cdot T := \sum_j \theta_j T_j$, we have almost everywhere

$$\theta \cdot T - \log C(\theta) \ = \ \phi \cdot T - \log C(\phi),$$

or $(\theta - \phi) \cdot T = c$ for some c not depending on x. But $\theta \neq \phi$ means that the T_j are affinely dependent, contradicting minimality. □

Any subset of an exponential family is also an exponential family with the same T_j and ν, recalling that $d\nu(x) := h(x)d\mu(x)$. It can be useful to take an exponential family as large as possible. Given ν and T_j, $j = 1, \ldots, k$ the *natural parameter space* of the exponential family is the set of all $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ such that

$$(2.5.5) \qquad K(\theta) \ := \ \int \exp(\textstyle\sum_{j=1}^k \theta_j T_j(x)) d\nu(x) \ < \ \infty.$$

Clearly $K(\theta) > 0$ for all θ. For any θ in the natural parameter space, we can define $C(\theta) := 1/K(\theta)$ and get a probability measure P_θ given by (2.5.3). So we have a family of laws P_θ indexed by the natural parameter space. The family doesn't extend to values of θ outside the natural parameter space since then normalization is not possible.

2.5.6 Theorem. For any given σ-finite ν and measurable functions T_j on $(X, \mathcal{B})$, the natural parameter space is a convex set in $\mathbb{R}^k$.

Proof. First, for any real y (which can be positive or negative), $\theta \mapsto e^{y\theta}$ is a convex function of $\theta \in \mathbb{R}$ (its second derivative is positive, so its first derivative is increasing, which implies convexity). It follows that for any real $y_1, \ldots, y_k$, the function

$$\theta \ = \ (\theta_1, \ldots, \theta_k) \mapsto \exp(y_1\theta_1 + \cdots + y_k\theta_k)$$

is convex on $\mathbb{R}^k$. The inequalities defining convexity are preserved when integrated with respect to a nonnegative measure, so $K(\theta)$ is a convex function, whose values may be infinite for some θ (just those θ outside the natural parameter space). The set where a convex function $< +\infty$ is clearly a convex set. □

2.5.7 Proposition. For any exponential family, the natural parameter space is the same for any number n of i.i.d. observations.

Proof. If $K_n(\theta)$ is the integral $K(\theta)$ for n observations, then from the definitions and the Tonelli-Fubini theorem, $K_n(\theta) = K_1(\theta)^n$ for all n, so $K_n(\theta)$ is finite if and only if $K_1(\theta)$ is. □

2.5.8 Theorem. For an exponential family as in (2.5.3) let U be the interior of the natural parameter space. Then for $\xi = (\xi_1, \ldots, \xi_k)$ in U and $\eta = (\eta_1, \ldots, \eta_k) \in \mathbb{R}^k$, let $W \ := \ \{\zeta = \xi + i\eta : \ \xi \in U, \ \eta \in \mathbb{R}^k\}$ so that $\zeta_j = \xi_j + i\eta_j$ for $j = 1, \ldots, k$. Then the function $K(z)$ in (2.5.5) is, on W, an analytic (holomorphic) function of z, representable by a power series in the k coordinates $z_j - \zeta_j$ in the neighborhood of any point ζ in W. In particular K has, on W, continuous partial derivatives of all orders with respect to z, which can be obtained by differentiating under the integral sign. In other words, for any $p = (p_1, \ldots, p_k)$, where the $p(i) \ := \ p_i$ are nonnegative integers and $[p] \ := \ p_1 + \cdots p_k$, the partial derivative $D^pK \ := \ \partial^{[p]}K(z)/\partial z_1^{p(1)} \cdots \partial z_k^{p(k)}$ exists and is continuous, and equals $\int T(x)^p \exp(\sum_{j=1}^k z_j T_j(x)) d\nu(x)$, where $t^p \ := \ t_1^{p(1)} \cdots t_k^{p(k)}$. For any $\xi \in U$, $E_\xi T^p \ = \ D^pK(\xi)/K(\xi)$.

Proof. Let $\zeta = \xi + i\eta \in W$, so $\xi \in U$ and $\eta \in \mathbb{R}^k$. Take $\varepsilon > 0$ small enough so that if $|u_j - \xi_j| \le \varepsilon$ for all $j = 1, \ldots, k$ then $u \in U$, so $u + iv \in W$ for any $v \in \mathbb{R}^k$. Then for any $T = T(x) \in \mathbb{R}^k$,

$$|e^{(u+iv)\cdot T}| \ = \ e^{u\cdot T} \ = \ e^{(u-\xi)\cdot T} e^{\xi\cdot T}.$$

Thus, replacing $d\nu(x)$ by $e^{\xi\cdot T(x)}d\nu(x)$, we can assume that $\xi = 0$. Then $|u_j| \le \varepsilon$ for $j = 1, \ldots, k$.

We have $e^{u\cdot T} = \Pi_{j=1}^k \exp(u_jT_j)$,

$$\exp(u_1T_1) = \textstyle\sum_{r=0}^\infty (u_1T_1)^r/r!, \quad |(u_1T_1)^r| = |u_1|^r|T_1|^r, \quad \text{and}$$
$$\textstyle\sum_{r=0}^\infty |u_1T_1|^r/r! = \exp(|u_1T_1|) \le \exp(-\varepsilon T_1) + \exp(\varepsilon T_1),$$

and likewise for any $j = 2, \ldots, k$ in place of $j = 1$. By choice of ε,

$$\textstyle\int \cdots \int \Pi_{j=1}^k \exp(\pm\varepsilon T_j) d\nu(x_1) \cdots d\nu(x_k) < \infty$$

for any choices of $\pm$, where $T_j := T_j(x_j)$ for each j, so the sum over all 2^k possible choices of $\pm$ of the integrals is finite. Thus by dominated convergence, the series

$$e^{u\cdot T} = \Pi_{j=1}^k \textstyle\sum_{r_j=0}^\infty (u_jT_j)^{r_j}/r_j!$$

converges absolutely if $|u_j| \le \varepsilon$ for all j, and can be interchanged with

$$\int\int \cdots \int \cdot d\nu(x_1) \cdots d\nu(x_k).$$

The integral yields a power series in $u_1, \ldots, u_k$. In the above, u_j can be replaced by $u_j + iv_j$ if $|u_j + iv_j| \le \varepsilon$ for each j. So we get a power series converging to $K(z)$ for $z = u + iv$. Since such a series exists in some neighborhood of each point in W, $K(\cdot)$ is holomorphic on W as stated.

To show that derivatives can be taken under the integral sign, first let $k = 1$ and $p = 1$. If $0 < t < c$ and $y > 0$ then for $\lambda := t/c$ and $x := cy$, by convexity $e^{\lambda x} \le \lambda e^x + (1-\lambda)e^0 \le \lambda e^x + 1$, so $(e^{ty} - 1)/t \le e^{cy}/c$. Likewise, for $0 < |t| < c$ and all y, $|(e^{ty} - 1)/t| \le (e^{cy} + e^{-cy})/c$. For u in U, and c small enough, $u \pm c \in U$, so the functions $\{(e^{(t+u)T(x)} - e^{uT(x)})/t : 0 < |t| < c\}$ are dominated by an integrable function. So

$$\frac{d}{d\theta}\int e^{\theta T(x)} d\nu(x)|_{\theta=u} = \int T(x)e^{uT(x)} d\nu(x).$$

Also, $|y| \le (e^{cy} + e^{-cy})/c$.

For $p > 1$, $c^{-p}(e^{cy} + e^{-cy})^p \le (2/c)^p(e^{pcy} + e^{-pcy})$. For fixed p, $u \pm pc \in U$ for c small enough, so we can again apply dominated convergence to get

$$(d^p/d\theta^p)\int e^{\theta T(x)} d\nu(x)|_{\theta=u} = \int T(x)^p e^{uT(x)} d\nu(x).$$

Now for $k > 1$, and any $p \in \mathbb{N}^k$, the 2^k (or fewer) points $(u_1 \pm cp_1, \ldots, u_k \pm cp_k)$ are all in U if c is small enough. Dominated convergence applies once more, so the derivatives can be interchanged with integrals as stated.

The final statement follows easily since $C(\theta) \equiv 1/K(\theta)$, finishing the proof. □

Suppose given an exponential family as in (2.5.3) and let $j(\theta) := \log K(\theta) = -\log C(\theta)$, so that $dP_\theta/d\nu = \exp(-j(\theta) + \theta\cdot T)$ where $\theta \cdot T := \sum_j \theta_jT_j(x)$. Since the vector $T := \{T_j\}_{j=1}^k$ gives a sufficient statistic for the family, the means and variances of its components are of interest. They have nice expressions in terms of derivatives of the function j. The gradient of j is the vector-valued function $\nabla j := (\partial j/\partial\theta_1, \ldots, \partial j/\partial\theta_k)$.

2.5.9 Corollary. For any θ in the interior of the natural parameter space of the exponential family (2.5.3), $E_\theta T = \nabla j(\theta)$ and for any $r, s = 1, \dots, k$,

$$\mathrm{cov}_\theta(T_r, T_s) \;=\; E_\theta(T_r T_s) - E_\theta T_r E_\theta T_s \;=\; \partial^2 j(\theta)/\partial\theta_r \partial\theta_s.$$

Proof. Theorem 2.5.8 gives $E_\theta(T_r T_s) \;=\; (\partial^2 K/\partial\theta_r\partial\theta_s)/K(\theta)$ and

$$E_\theta T_r \;=\; (\partial K/\partial\theta_r)/K(\theta) \;=\; \partial j(\theta)/\partial\theta_r.$$

This gives the first conclusion. Taking $\partial/\partial\theta_s$ of both sides of the last equation, the latter conclusion follows. □

Any convex set in $\mathbb{R}^k$ either has non-empty interior or is included in some lower-dimensional affine subspace (RAP, 6.2.6). So for an exponential family of order k, the natural parameter space in a minimal representation, and so in $\mathbb{R}^k$, has non-empty interior. (Here $k \geq 1$ since by definition an exponential family contains at least two laws.) So for an exponential family in a minimal representation on its natural parameter space the hypothesis of the following theorem holds:

2.5.10 Theorem. If for an exponential family (2.5.3), Θ includes a non-empty open set U, then T is a Lehmann-Scheffé statistic.

Proof. As noted above, $T = (T_1, \dots, T_k)$ is sufficient. Let $p = (p_1, \dots, p_k) \in U$. Replacing $d\nu$ by $\exp(\sum_j p_j T_j(x))d\nu$, we can assume that $p = 0$. For some $\varepsilon > 0$, U includes the cube $C_\varepsilon \;:=\; \{\theta : \; |\theta_j| \leq \varepsilon \;\text{ for }\; j = 1, \dots, k\}$.

Recall that a function measurable for $T^{-1}(\mathcal{F})$, where in this case $\mathcal{F}$ is the Borel σ-algebra in $\mathbb{R}$, is of the form $f \circ T$ for some measurable f (Theorem 2.1.3). Let $f : \mathbb{R}^k \to \mathbb{R}$ be such that $E_\theta f(T(x)) = 0$ for all $\theta \in \Theta$. Let $f = f^+ - f^-$ where $f^+ \;:=\; \max(f, 0)$ and $f^- \;:=\; -\min(f, 0)$, so that $f^+ \geq 0$ and $f^- \geq 0$. Let $m \;:=\; \nu \circ T^{-1}$ on on $\mathbb{R}^k$. Then for all $\theta \in C_\varepsilon$,

$$u(\theta) \;:=\; \int \exp(\textstyle\sum_j \theta_j t_j) f^+(t) dm(t) \;=\; v(\theta) \;:=\; \int \exp(\sum_j \theta_j t_j) f^-(t) dm(t).$$

From $u(0) \;=\; v(0)$ we see that $\int f^+ dm \;=\; \int f^- dm$. Multiplying f by a constant, we can assume $\int f^+ dm \;=\; 1$ (if it's zero, we are done). Then letting $dP^+ \;:=\; f^+ dm$ and $dP^- \;:=\; f^- dm$, P^+ and P^- are probability measures.

For complex $\theta_j \;=\; \xi_j + i\eta_j$, in the strip $S : \; |\xi_j| < \varepsilon, \; j = 1, \dots, k$, the above integrals $u(\theta)$ and $v(\theta)$ converge absolutely. By Theorem 2.5.8, they represent holomorphic (analytic) functions of θ in S. Since $u \equiv v$ for θ real (all $\eta_j = 0$), u and v have the same derivatives of all orders at 0. So $u \equiv v$ in a complex neighborhood of 0, and by analytic continuation (e.g. Bochner and Martin, 1948, p. 34), $u \;\equiv\; v$ throughout S. Taking $\xi_j = 0$ for all j, we see that P^+ and P^- on $\mathbb{R}^k$ have the same characteristic function. So $P^+ = P^-$ by the uniqueness theorem (RAP, Theorem 9.5.1). (Here we have obtained a uniqueness theorem for Laplace transforms in a neighborhood of 0.) So $f^+ = f^-$ and $f = 0$ almost everywhere for ν, proving the Lehmann-Scheffé property. □

Theorem 2.4.15 shows that the lower bound in the information inequality is attained, for densities continuously differentiable in θ, if and only if the family of laws is exponential as in (2.5.1) where the estimator T is an affine (linear) function of T_1, $T = aT_1 + b$ for some constants a, b. On the other hand by Theorem 2.5.10, in such a case T_1 is an LS sufficient statistic. By Corollary 2.5.9, it is an unbiased estimator of $j'(\theta)$ where $j(\theta) = \log K(\theta)$.

Let η be any function not equal (even almost everywhere) to an affine function, such that $E_\theta(\eta(T_1)^2) < \infty$ for all θ. Then $\eta(T_1)$ is an unbiased estimator of $y(\theta) := E_\theta h(T_1)$ which, among unbiased estimators, has smallest possible variance and minimal risk for all convex loss functions by Theorem 2.3.5, but where the information inequality lower bound is not attained for some θ. So the information inequality produces sharp results only in rather special cases. It may be more useful in the form allowing bias, Theorem 2.4.12, or in an asymptotic form, Theorem 3.8.3 below.

The existence of a k-dimensional sufficient statistic $T = (T_1, \ldots, T_k)$ for an exponential family extends to any sample size n for n i.i.d. observations, as noted previously, replacing each T_i by $\sum_{j=1}^n T_i(X_j)$. When R. A. Fisher first defined exponential families, one of the main properties he pointed out was the possibility of data reduction in this way. Moreover, he stated that if the data can be reduced, in other words if for i.i.d. $X_1, \ldots, X_n$ there is a sufficient statistic of dimension $k < n$ (even for one value of n) then the family of laws must be exponential. This is true under some regularity conditions, one of which is that the family be equivalent. For example, the family of uniform distributions on intervals $[0, \theta]$, $0 < \theta < \infty$, has a 1-dimensional sufficient statistic, the largest order statistic $X_{(n)}$, but is evidently not equivalent and (so) not exponential. Other regularity conditions of continuity and differentiability will be assumed. If there were no such conditions, the "dimension" of a sufficient statistic would not be meaningful. For example, if X and Y are any two uncountable Borel sets in complete separable metric spaces, such as $X = \mathbb{R}^k$ and $Y = \mathbb{R}^m$, then there is always a 1-1, Borel measurable function from X onto Y with measurable inverse (RAP, Sec. 13.1). Any Borel measurable function is continuous when restricted to sets having nearly full measure (Lusin's theorem, RAP, Theorem 7.5.2). Also, for any m there is a continuous function from $\mathbb{R}^m$ into $\mathbb{R}$, 1-1 almost everywhere for Lebesgue measure (Denny, 1964).

The following example may illustrate the point. Let x and y be two numbers in $[0, 1]$, each represented by its decimal expansion, $x = \sum_{n\geq 1} x_n/10^n$ where each x_n is $0, 1, \ldots$, or 9, and likewise for y. By alternating digits define a real number z with digits $z_{2n-1} = x_n$ and $z_{2n} = y_n$ for $n = 1, 2, \ldots$. This gives a correspondence between ordered pairs (x, y) of real numbers and individual real numbers z. Although it is not quite well-defined, because of ambiguities such as $0.099999999\ldots = 0.100000\ldots$, 1-1 or continuous, the correspondence illustrates a reduction of dimension (from 2 to 1) which is not a real reduction in the sense of statistical interest. The example also shows why some regularity conditions such as differentiability may be expected in proofs about data reduction implying that a family is exponential.

Let $\mathcal{P}$ be an equivalent family of probability measures. Let Q be a fixed law in the family. If T is a sufficient statistic for $\{P^n : P \in \mathcal{P}\}$, the family of laws of n i.i.d. observations $X_1, \ldots, X_n$ with laws in $\mathcal{P}$, then by Corollary 2.1.5 for each P in $\mathcal{P}$ there is a function ρ_P with

$$\Pi_{j=1}^{n} R_{P/Q}(x_j) \;=\; \rho_P(T(x_1,\ldots,x_n)) \tag{2.5.11}$$

for almost all $x_1,\ldots,x_n$. T will be called *strongly sufficient* (with respect to given choices of Q and of $R_{P/Q}$ for all x and all $P \in \mathcal{P}$) if (2.5.11) holds for *all* (and not only almost all) x.

Let $\phi_P(x) := \log R_{P/Q}(x)$ for any P in $\mathcal{P}$. We will be considering families for which the likelihood ratios $R_{P/Q}$ are continuous non-zero functions of x, so that ϕ_P is continuous, and where every neighborhood of each point in the sample space has positive measure for each law in $\mathcal{P}$, so that ϕ_P is determined everywhere by continuity and not only almost everywhere. So, strong sufficiency is a reasonable assumption.

A function f on a region in $\mathbb{R}^k$ is called C^1 if it has continuous first partial derivatives with respect to each of the k variables. It will be called BC^1 if these derivatives are also bounded. A real-valued function f on an interval $U \subset \mathbb{R}$ will be called *piecewise* BC^1 if f is continuous on U and there is a finite set $F \subset U$ such that f is BC^1 on $U \setminus F$, i.e. f is BC^1 on each of finitely many open intervals whose endpoints are in F or are endpoints of U. Now a fact can be stated:

2.5.12 Theorem. Let $\mathcal{P}$ be a family of laws defined on a connected open set U in $\mathbb{R}^r$ and having continuous densities f_P, $P \in \mathcal{P}$, with respect to Lebesgue measure λ on U, with $f_P(x) > 0$ for all $x \in U$ and P in $\mathcal{P}$ (so $\mathcal{P}$ is equivalent). Suppose that all the functions f_P are continuous on U and that for some positive integers $k < n$, there is a statistic T, continuous from U^n into $\mathbb{R}^k$, strongly sufficient for $\{P^n : P \in \mathcal{P}\}$, where $R_{P/Q} := f_P/f_Q$. Then

(a) If $k = 1$, $\mathcal{P}$ is an exponential family of order 1.

(b) If all the densities f_P are BC^1, or if $r = 1$ and they are piecewise BC^1, then $\mathcal{P}$ is exponential of order at most k.

Proof. For a given n, let $S := S_T$ be the set of all continuous real functions ϕ such that for some function ζ,

$$\phi(x_1) + \cdots + \phi(x_n) \;=\; \zeta(T(x_1,\ldots,x_n)) \tag{2.5.13}$$

for all $x_1,\ldots,x_n$ in U. (Such an equation results from taking logarithms in (2.5.11) with $\phi(x) := \phi_P(x)$.) Clearly S is a vector space of functions containing the constants. Suppose S has dimension $k+1$. Then it has a basis consisting of the constant 1 and k other functions $\phi_1,\ldots,\phi_k$, and each function ϕ_P is of the form $a_0(P) + a_1(P)\phi_1(x) + \cdots + a_k(P)\phi_k(x)$, so the family is exponential of order at most k.

Case 1: $r = 1$, so U is an open interval in $\mathbb{R}$, and $k = 1$. It will be enough to prove that the dimension of S is at most 2.

Suppose the hypothesis holds for some $n > 2$. It will be shown that it holds for $n = 2$. Let's indicate the dependence of ρ_P and T on n in (2.5.11) by writing them as $\rho_P^{(n)}$ and $T^{(n)}$ respectively. By our choice of $R_{P/Q} = f_P/f_Q$, we have $0 < R_{P/Q}(y) < \infty$ for all $y \in U$. Fix any $y_3, ..., y_n \in U$. Then (2.5.11) holds for $n = 2$ with

$$T^{(2)}(x_1,x_2) \;:=\; T^{(n)}(x_1,x_2,y_3,...,y_n) \;\text{ and }\; \rho_P^{(2)}(t) \;:=\; \rho_P^{(n)}(t)/\Pi_{j=3}^{n} R_{P/Q}(y_j)$$

for any $t \in \mathbb{R}$. Clearly $T^{(2)}$ is continuous. So the hypothesis holds for $n = 2$ and it suffices to treat that case. The following will be useful:

2.5.14 Lemma. If $k = 1$, $n = 2$, $\phi \in S$, x_0, y and z are in U, $\phi(y) = \phi(z)$ and $s := T(x_0, y) \neq t := T(x_0, z)$, then ϕ is constant in some neighborhood of x_0.

Proof. We can assume that $s < t$ and then that $y < z$, otherwise interchanging x and $-x$. Let $y_1 := \sup\{u :\ u < z \text{ and } T(x_0, u) = s\}$. Then by (2.5.13) and continuity of ϕ, $\phi(y) = \phi(y_1)$. So we can assume $y = y_1$. Likewise, we can assume $z = \inf\{u :\ y_1 < u$ and $T(x_0, u) = t\}$. Then by continuity of T and the intermediate value theorem,

$$s < T(x_0, u) < t \ \text{ for } \ y < u < z. \tag{2.5.15}$$

Since $\phi(y) = \phi(z)$, by Rolle's theorem there is some v with $y < v < z$ at which ϕ attains either its absolute maximum or absolute minimum on the closed interval $[y, z]$. By symmetry, suppose it is an absolute minimum. Let $t_0 := T(x_0, v)$. Then for ζ in (2.5.13), $\zeta(t_0) = \min\{\zeta(w) :\ s \leq w \leq t\}$ since $T(x_0, \cdot)$ takes $[y, z]$ onto $[s, t]$ by the intermediate value theorem and $\phi(u) \equiv \zeta(T(x_0, u)) - \phi(x_0)$ (continuity of ζ is not assumed or needed here).

Now $s < t_0 < t$ by (2.5.15). For x in some neighborhood V of x_0, by continuity, $T(x, y) < t_0 < T(x, z)$ and $s < T(x, v) < t$. By the intermediate value theorem again, for each x in V there is some $g(x)$ with $T(x, g(x)) = t_0$ and $y < g(x) < z$. Then for each $x \in V$, by (2.5.13) twice,

$$\zeta(t_0) \ = \ \phi(x) + \phi(g(x)) \ \geq \ \phi(x) + \phi(v) \ = \ \zeta(T(x, v)) \ \geq \ \zeta(t_0).$$

So both inequalities just above are equations. Also, $\zeta(t_0) = \zeta(T(x_0, v)) = \phi(x_0) + \phi(v)$, so $\phi(x) = \phi(x_0)$, proving Lemma 2.5.14. □

Now let g and γ be in S and suppose g is not constant, so $g(y) \neq g(z)$ for some y and z in U. Then for any $x_0 \in U$, $T(x_0, y) \neq T(x_0, z)$. For some $c \in \mathbb{R}$, $\gamma(y) - cg(y) = \gamma(z) - cg(z)$. Let $\phi := \gamma - cg$. Then $\phi \in S$ and $\phi(y) = \phi(z)$, so by Lemma 2.5.14, ϕ is constant on a neighborhood of x_0. Since x_0 was arbitrary in U and U is connected, $\phi \equiv b$ on U for some constant b, so $\gamma \equiv cg + b$ and S is at most 2-dimensional, finishing the proof for Case 1.

Case 2: $r = 1$ and $k > 1$. Here we use:

2.5.16 Lemma. Let $f_1, \ldots, f_n$ be piecewise BC^1 functions from an open interval U into $\mathbb{R}$, $x = (x_1, \ldots, x_n) \in U^n$, and $G(x) := \{\sum_{j=1}^n f_i(x_j)\}_{i=1}^n$. If $1, f_1, \ldots, f_n$ are linearly independent, then for some $y \in U^n$, $\det J(y) \neq 0$ where J is the Jacobian matrix of G, $J_{ij} := f_i'(x_j)$ for $i, j = 1, \ldots, n$.

Proof. The proof will be by induction on n. The result clearly holds for $n = 1$. Suppose it holds for $n - 1$ but fails for n, for some $f_1, \ldots, f_n$. Expanding the determinant by the minors of the last column gives

$$0 \ = \ a_1(x_1, \ldots, x_{n-1}) f_1'(x_n) + \cdots + a_n(x_1, \ldots, x_{n-1}) f_n'(x_n) \tag{2.5.17}$$

for all $x \in U^n$ with x_n not in the finite union of finite sets where f_i' don't exist. Here $a_n(x_1, \ldots, x_{n-1})$ is the determinant for the $n - 1$ case and $f_1, \ldots, f_{n-1}$, so for some

$z \in U^{n-1}$, $a_n(z_1, \ldots, z_{n-1}) \neq 0$. Let $x_i = z_i$ for $i = 1, \ldots, n-1$ and integrate (2.5.17) with respect to x_n over an interval, say from (a constant) b to (a variable) x, so

$$0 \equiv \textstyle\sum_{j=1}^n a_j(z_1, \ldots, z_{n-1})(f_j(x) - f_j(b)).$$

Since $a_n(z_1, \ldots, z_{n-1}) \neq 0$, this contradicts the linear independence of $1, f_1, \ldots, f_n$, proving Lemma 2.5.16. □

Now to prove Case 2, it is enough to show that the dimension of the space S_1 of C^1 functions in S is at most $k+1$. If not, let $1, f_1, \ldots, f_n$ be linearly independent functions in S_1 with $n = k+1$. The hypotheses hold for $n = k+1$ since they hold for some $n > k$, as in Case 1. By Lemma 2.5.16 there is a point $x \in U^n$ with $\det J(x) \neq 0$. Then by the inverse function theorem, e.g. Rudin (1976, Theorem 9.24), G is 1-1 on some neighborhood of x. But also, $G = H(T)$ where $T = (t_1, \ldots, t_k)$ and $H(T) = (\psi_1(T), \ldots, \psi_n(T))$ for some functions ψ_i, so T must be 1-1 on the neighborhood. But this is impossible since T is continuous and reduces the dimension (see Appendix B), finishing the proof in Case 2.

Case 3: $r > 1$. Consider part (b). To show that the dimension of S_1 is at most $k+1$, since it contains the constants, is equivalent to showing that for any $\{f_1, \ldots, f_{k+1}\} \subset S_1$, the range of the function $x \mapsto (f_1(x), \ldots, f_{k+1}(x))$ is included in some k-hyperplane (i.e. k-dimensional hyperplane) in $\mathbb{R}^{k+1}$, in other words $1, f_1, ..., f_{k+1}$ are linearly dependent.

Otherwise, there exist such f_i and points $x_1, \ldots, x_{k+2}$ in U such that the points $\{f_j(x_i)\}_{j=1}^{k+1}$ for $i = 1, \ldots, k+2$ are not all in any k-hyperplane in $\mathbb{R}^{k+1}$. To see this, we can recursively select $x_1, x_2, ..., x_{k+2}$ such that $x_2 \neq x_1$ and for $j \geq 3$, x_j is not in the unique $(j-2)$-dimensional hyperplane containing $x_1, ..., x_{j-1}$. For each j, since $f_j \in S$, take ψ_j such that $f_j(y_1) + \cdots + f_j(y_n) \equiv \psi_j(T(y_1, \ldots, y_n))$. Let γ be a BC^1 curve, i.e. a BC^1 function from an open interval $I_1 := (a, b)$ into U, whose range contains all the points $x_1, \ldots, x_{k+2}$. Such a γ exists since U is open and connected. In more detail, for each $u \in U$ the open ball $B(u, r) \subset U$ for some $r > 0$ where $B(u, r) := \{y : |y - u| < r\}$. Let $W(u)$ be the set of all v such that for some $n < \infty$ there exist $u_0 = u, u_1, ..., u_n = v$ and $r_i > 0$, $i = 0, 1, ..., n$ such that $B(u_i, r_i) \subset U$ for each i and $B(u_i, r_i) \cap B(u_{i-1}, r_{i-1}) \neq \emptyset$ for each $i = 1, ..., n$. It is easily seen that $W(u)$ is an open set included in U and that two sets $W(u)$ and $W(u')$, if they intersect, are identical. Thus by connectedness, $W(u) = U$ for each $u \in U$. Now if $v \in W(u)$ it is easy to construct a BC^1 curve within U joining u to v.

Let

$$V(t_1, \ldots, t_{k+1}) := T(\gamma(t_1), \ldots, \gamma(t_{k+1}))$$

for any t_i in I_1. Then V is continuous. Let $g_i(t) := f_i(\gamma(t))$ for any t in I_1. Then

$$g_i(t_1) + \cdots + g_i(t_{k+1}) = \psi_i(V(t_1, \ldots, t_{k+1}))$$

for any $t_1, \ldots, t_{k+1} \in I_1$ and $i = 1, \ldots, k+1$. Since f_i and γ are BC^1 functions, so are g_i, and they belong to the space S_1 for $r = 1$, I_1 in place of U, and V in place of T. But the range of $t \mapsto (g_1(t), \ldots, g_{k+1}(t))$ for $t \in I_1$ is not included in any k-hyperplane, contradicting Case 2. If $k = 1$, then f_i need not be BC^1 and Case 1 is applied instead, finishing the proof of Theorem 2.5.12. □

Example. This will show why the connectedness of U is, or the continuity hypotheses are, needed in Theorem 2.5.12. Let $U := (0,1) \cup (2,3)$ (which is not connected). Let the dominating measure ν be the sum of Lebesgue measures on the two intervals. For $0 < \lambda < 1, 0 < \theta < \infty$ let

$$\begin{aligned} \phi_{\theta,\lambda}(x) &:= \lambda\theta e^{\theta x}/(e^\theta - 1), \quad 0 < x < 1; \\ &:= (1-\lambda)\theta e^{\theta(x-2)}/(e^\theta - 1), \quad 2 < x < 3. \end{aligned}$$

It is straightforward to check that this is a probability density for each θ and λ. Let x_1, x_2 be i.i.d. with this density. Then the likelihood function is

$$u(\theta, \lambda, x_1, x_2)\theta^2 \exp(\theta(x_1 + x_2))/(e^\theta - 1)^2$$

where $u(\theta, \lambda, x_1, x_2) := \lambda^2$ for $0 < x_1 + x_2 < 2$, $2\lambda(1-\lambda)e^{-2\theta}$ for $2 < x_1 + x_2 < 4$, and $(1-\lambda)^2 e^{-4\theta}$ for $4 < x_1 + x_2 < 6$. It follows by Corollary 2.1.5, not only because of the factor $\exp(\theta(x_1 + x_2))$ but because the ranges for different formulas for $u(\cdot, \cdot, x_1, x_2)$ also are functions of $x_1 + x_2$, that $x_1 + x_2$ is a $k = 1$-dimensional sufficient statistic for the family with $n = 2$. Let $\gamma(\theta, \lambda) := \log[\theta/(e^\theta - 1)]$. Then one can check that

$$\log \phi_{\theta,\lambda}(x) = \gamma(\theta, \lambda) + \log \lambda + \theta x + [\log((1-\lambda)/\lambda) - 2\theta]1_{2<x<3}.$$

Since the functions x and $1_{2<x<3}$ are affinely independent, as are the functions θ and $\log((1-\lambda)/\lambda)$, we see that the family is exponential of order 2, not 1. Thus the conclusion of Theorem 2.5.12(a) does not hold in this case. The connectedness of the interval U is used in the proof more than once, by way of the intermediate value theorem.

Of course, connectedness is only meaningful in connection with continuity of some functions. We could take $U := (0,2)$ to be connected in the example while $\phi_{\theta,\lambda}$ and T are discontinuous by replacing $(2,3)$ by $[1,2)$ and letting $x_1(x) = x$ for $0 < x < 1$ and $x_1(x) = x + 1$ for $1 \leq x < 2$, while taking x_2 as an i.i.d. copy of x_1.

Or, replacing the union of two intervals by a union of as many intervals as we like, we can get the exponential family to be of arbitrarily high order for $n = 2$. Similarly, by spreading the intervals farther apart, for example taking $(0,1) \cup (n, n+1) \cup (n^2+n, n^2+n+1), ...$, we can get a 1-dimensional sufficient statistic for any number n of i.i.d. observations, again if U is not connected or the densities and T are not continuous.

Note that for $r > 1$, the dimension of the full data vector $(X_1, \ldots, X_n)$ is nr, but that the assumption in Theorem 2.5.12 is $k < n$ (and not $k < nr$). Suppose we consider a family of distributions $\mathbb{R}^r$ having densities with respect to some measure (not Lebesgue measure) which are functions of the first coordinate x_1. Then x_1 is a sufficient statistic. For n i.i.d. variables there is an n-dimensional sufficient statistic and $n < nr$, but the family need not be exponential. So the assumption $k < n$ in Theorem 2.5.12 is sharp.

Example. For a normal distribution $N(m, \sigma^2)$ on $\mathbb{R}$, we can write the density as

$$(2\pi\sigma^2)^{-1/2} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{mx}{\sigma^2} - \frac{m^2}{2\sigma^2}\right).$$

These laws form an exponential family with $T_1(x) = x^2$, $Q_1(m, \sigma^2) = -1/(2\sigma^2)$, $T_2(x) = x$, and $Q_2(m, \sigma^2) = m/\sigma^2$. For $N(m, \sigma^2)^n$ on $\mathbb{R}^n$, we get $T_1(x) = \sum_{j=1}^n x_j^2$, $T_2(x) = \sum_{j=1}^n x_j$.

2.5.18 Proposition. Let $X_1, \dots, X_n$ be i.i.d. with law $N(m, \sigma^2)$ and $n \geq 2$. Then $s^2 := (n-1)^{-1} \sum_{j=1}^n (X_j - \overline{X})^2$ has, among unbiased estimators of σ^2, smallest risk for squared-error loss, for all (m, σ^2). The risk of s^2 is $2\sigma^4/(n-1)$.

Proof. We know that s^2 is an unbiased estimator of σ^2. Let $\mathcal{S}$ be the smallest σ-algebra for which T_1 and T_2 are measurable in this case. Then $\mathcal{S}$ is sufficient. It is Lehmann-Scheffé by Theorem 2.5.10. Since $s^2 = (n-1)^{-1}(T_1 - T_2^2/n)$, s^2 is $\mathcal{S}$-measurable. Then it follows from Theorem 2.3.5 that s^2 has minimum risk for squared-error loss (which is convex).

To find the variance of s^2, first note that if X has distribution $N(0, \sigma^2)$, then $EX^4 = 3\sigma^4$, by integration by parts or the moment generating function. Also, $Es^2 = \sigma^2$ and we can assume $m = 0$. Make an orthogonal change of coordinates from $(X_1, \dots, X_n)$ to $(Y_1, \dots, Y_n)$ where the Y_1 axis is in the direction of $(1, 1, \dots, 1)$, so that $Y_1 = n^{1/2}\overline{X}$. Then the Y_j are i.i.d. $N(0, \sigma^2)$ and $s^2 = (n-1)^{-1} \sum_{j=2}^n Y_j^2$. So

$$E((s^2)^2) \;=\; (n-1)^{-2}\sigma^4[3(n-1) + (n-1)(n-2)] \;=\; (n+1)\sigma^4/(n-1),$$

and $\mathrm{var}(s^2) = 2\sigma^4/(n-1)$. □

Recall that for an unbiased estimator, the risk for squared-error loss is the same as the variance. Proposition 2.5.18 completes example (3) following Theorem 2.4.10 and shows that the information inequality lower bound (2.4.3) cannot be attained in this case.

Unbiased estimators which are LS sufficient statistics are thereby optimal among unbiased estimators, but may fail in other ways. In Sec. 2.2, just after the definition of unbiased estimator, an example was given of an inadmissible unbiased estimator for $e^{-\lambda}$ where λ is the parameter of a Poisson random variable X which is observed conditional on $X \geq 1$. Now, note that Poisson distributions conditional on $X \geq 1$ form an exponential family of order 1. So the pathology of unbiased estimators occurs even in this case where we have an LS sufficient one-dimensional statistic. We had also noted that a constant estimator may be admissible though biased.

From what has been said so far it might seem that an estimator which is both unbiased and admissible, especially for an exponential family of order 1, might be a good estimator. But it may not be: consider the binomial distributions for $n = 2$ (for the number X of successes in 2 independent trials) and probability p of success, where $0 < p < 1$. Suppose the problem is to estimate p^2. It's easily seen that the unique unbiased estimator U measurable for the minimal sufficient σ-algebra is $U(0) = U(1) = 0$ (here $U(1) = 0$ is surprising and bad) and $U(2) = 1$. Now, it will be shown that U is admissible for a wide class of loss functions. If the loss is a continuous function f of the difference $T - p^2$ for an estimate T, with $f \geq 0$ and $f(x) = 0$ if and only if $x = 0$, the risk for a given p is

$$r(p, T) \;=\; (1-p)^2 f(T(0) - p^2) + 2p(1-p) f(T(1) - p^2) + p^2 f(T(2) - p^2),$$

so that $r(p, U) \;=\; (1-p)^2 f(-p^2) + 2p(1-p) f(-p^2) + p^2 f(1 - p^2)$. If for some $s > 1/2$, $f(x) = O(|x|^s)$ as $|x| \to 0$, which holds if f is convex for $s = 1$, and where we can assume

$s \le 1$, then $r(p, U) = O(p^{2s})$ as $p \downarrow 0$. If T is an estimator with $r(p, T) \le r(p, U)$ for all p, then taking $p \downarrow 0$, and noting that $2s > 1$, it is easily seen that $T(0) = 0$ and then that $T(1) = 0$. Then letting $p \uparrow 1$ shows that $T(2) = 1$. So $T = U$ and U is admissible.

Thus the two "good" properties of being admissible (for a great many loss functions) and unbiased, even in combination, and for an exponential family of order 1, still allow the absurd inference that the probability of success is 0 when 1 success is observed in 2 trials. Moreover, there do not seem to be theorems providing in adequate generality that there are admissible and/or unbiased estimators which behave well. So in the search for really good estimators we will need to consider other properties, such as those in the next chapter.

PROBLEMS

In problems 1-4, show that the given family $\mathcal{P}$ of laws is exponential. Specifically, find a σ-finite measure μ such that the densities of laws in $\mathcal{P}$ with respect to μ are of the form $C(\theta)h(x)\exp(\sum_j \theta_j T_j(x))$ as in (2.5.1). Find the order of the family and a minimal representation. Give T_j, h, and $C(\theta)$ explicitly. Then find the natural parameter space (largest possible set of $\{\theta_j\}$), and indicate what functions the θ_j are of the usual parameters.

1. Let F be a finite set with k elements and $\mathcal{P}$ the family of all laws P on F which are not 0 at any point.
2. Let $\mathcal{P}$ be the family of binomial distributions $B(n, p)$, $0 < p < 1$, for any fixed n.
3. (a) $\mathcal{P}$ is the family of geometric distributions $P(k) = (1-p)^{k-1}p$ for $k = 1, 2, \ldots$, with $0 < p < 1$.

   (b) $\mathcal{P}$ is the family of Poisson distributions $P_\lambda(k) := e^{-\lambda}\lambda^k/k!$ for $k = 0, 1, \ldots$, with $0 < \lambda < \infty$.
4. Let F be the so-called "extreme value" distribution function $F(x) := \exp(-e^{-x})$ for $-\infty < x < \infty$. Let its density be $f(x)$ and consider the location family of all laws with densities $f(x - \theta)$, $\theta \in \mathbb{R}$.

- - - - - - - - - - - - - - - - -

5. Show that the family of all normal laws $N(\mu, \sigma^2)$ on $\mathbb{R}$ is exponential of order 2 and has a 1-dimensional continuous strongly sufficient statistic (for $n = 1$). Show that the family of distributions of i.i.d. normal $(X_1, \ldots, X_n)$ has a 2-dimensional continuous strongly sufficient statistic for each $n \ge 2$.
6. Let Q_ψ be distributions on $\mathbb{R}^2$ such that X has distribution $N(0, \tau^2)$ and, given X, Y has distribution $N(a + bX, \sigma^2)$, where $\psi = (\sigma^2, \tau^2, a, b)$. Find a minimal representation with functions $\theta_i(\psi)$. Describe the family as a subset of the family given by the natural parameter space.
7. Same question with $a + bX$ replaced by $a + bX + cX^2$.
8. Find the natural parameter space for the exponential family on $\mathbb{R}$ with $k = 1$, $T_1(x) \equiv x$, and

   (a) $d\mu(x) = e^{-|x|}dx$ (b) $d\mu(x) = e^{-|x|}dx/(1 + x^2)$.

NOTES

Fisher (1934) began the theory of exponential families. An important book on the topic is Barndorff-Nielsen (1978), who on p. 136 reviews the history of the subject, noting that characteristically Fisher was "mathematically somewhat imprecise." Brown (1986) is a more recent monograph.

Let's say there is "data reduction" for a family of laws if for n i.i.d. observations there is a sufficient statistic of dimension less than n. From the beginning, one of the main properties of interest for exponential families was to be equivalent families allowing data reduction. Other early references are Darmois (1935), Koopman (1936) and Pitman (1936). Their names have sometimes been used for exponential families, as in "Koopman-Darmois," "Darmois-Koopman," "Koopman-Pitman-Darmois" or "Fisher-Darmois-Koopman-Pitman" families.

The current form of Theorem 2.5.12, that under some continuity conditions the possibility of data reduction implies that a family is exponential, is due for $r = 1$ to Brown (1964) with earlier work by Dynkin (1951), and for $r > 1$ to Barndorff-Nielsen and Pedersen (1968), which was the source of the proof given for Theorem 2.5.12. Interestingly, the precise statement (let alone the proof) is not given, although it is cited, in the book of Barndorff-Nielsen (1978), and Brown (1986) doesn't cite Brown (1964) or Barndorff-Nielsen and Pedersen (1968).

REFERENCES

Barndorff-Nielsen, Ole (1978). *Information and Exponential Families.* Wiley, New York.

Barndorff-Nielsen, O., and K. Pedersen (1968). Sufficient data reduction and exponential families. *Math. Scand.* **22**, 197-202.

Bochner, Salomon, and William Ted Martin (1948). *Several Complex Variables.* Princeton University Press.

Brown, Lawrence D. (1964). Sufficient statistics in the case of independent random variables. *Ann. Math. Statist.* **35**, 1456-1474.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families. IMS Lecture Notes-Monograph Series* **9**, 283 pp.

Darmois, Georges (1935). Sur les lois de probabilité à estimation exhaustive. *Comptes Rendus Acad. Sci. Paris* **260**, 1265-1266.

Denny, J. L. (1964). A continuous real-valued function on E^n almost everywhere 1-1. *Fund. Math.* **55**, 95-99.

Dynkin, Evgenii Borisovich (1951). Necessary and sufficient statistics for a family of probability distributions. *Selected Transl. Math. Statist. and Probab.* **1** (1961), 23-41.

Fisher, Ronald Aylmer (1934). Two new properties of mathematical likelihood. *Proc. Royal Soc.* (London) Ser. A **144**, 285-307.

Koopman, L. H. (1936). On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.* **39**, 399-409.

Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. *Proc. Camb. Phil. Soc.* **32**, 567-579.

Rudin, Walter (1976). *Principles of Mathematical Analysis*, 3d ed. McGraw-Hill, New York.