18.466 Notes, Feb. 19, 2013, R. Dudley

Sufficiency

This topic corresponds to Section 1.5 of Bickel and Doksum. In this handout, starred sections contain measure-theoretic material which is not in Bickel and Doksum. These sections will not be covered on exams or problem sets. They are included so as to have mathematically more complete or correct formulations. Unstarred sections will be close to Bickel and Doksum's presentation.

1. Introduction

R. A. Fisher, a leading British statistician, in a 1922 article, 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." Fisher's very interesting but rather informal definition was formalized later, as we'll see. Statistics can be real- or vector-valued or can even have function values. For now, here is an

Example. (This is essentially Example 1.5.4 of Bickel and Doksum.) Suppose $X_1, ..., X_n$ are i.i.d. with a $N(\mu, \sigma^2)$ distribution, with μ and σ^2 both unknown. Let $X := (X_1, ..., X_n)$ and $\theta := (\mu, \sigma)$. Then the likelihood function of X and θ (the joint density, evaluated at the observed X_j) is

$$\frac{1}{(\sigma\sqrt{2\pi})^n}\prod_{j=1}^n \exp\left(-\frac{(X_j-\mu)^2}{2\sigma^2}\right) = \frac{1}{(\sigma\sqrt{2\pi})^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^n (X_j-\mu)^2\right)$$

$$= \frac{1}{(\sigma\sqrt{2\pi})^n}\exp\left(-\frac{1}{2\sigma^2}\left[n\mu^2+\sum_{j=1}^n X_j^2 - 2\mu\sum_{j=1}^n X_j\right]\right). \tag{1}$$

This likelihood function $f(X, \theta)$ depends on X only by way of the 2-dimensional vector-valued statistic $T(X) = (T_1(X), T_2(X))$ where

$$T_1(X) := \sum_{j=1}^n X_j, \quad T_2(X) := \sum_{j=1}^n X_j^2.$$

$T(X)$ is (or will be, after further definitions) an example of a sufficient statistic in this case.

It will always be true that if a likelihood function $f(X, \theta)$ can be written as $G(T(X), \theta)$ for some statistic T, then T will be sufficient.

But more generally, suppose $f(X,\theta)$ can be written as

$$f(X,\theta) = G(T(X),\theta)h(X) \tag{2}$$

where h doesn't depend on θ. Then for any two values θ_1 and θ_2 of θ, the likelihood ratio

$$\frac{f(X,\theta_1)}{f(X,\theta_2)} = \frac{G(T(X),\theta_1)}{G(T(X),\theta_2)}$$

depends on X only through $T(X)$. So, for one thing, to find the maximum likelihood estimate of θ given X (if it exists), it's enough to know $T(X)$. It turns out that the same is true for any other inference we may want to make about θ given X. Specifically, suppose we have a factorization (2) of a likelihood function $f(X,\theta)$ and we have a prior probability density $\pi(\theta)$ for θ. Then for a given X, we get the posterior density $\pi_X(\theta)$ by taking $\pi(\theta)f(X,\theta)$ and normalizing it to be a probability density with respect to θ, giving

$$\pi_X(\theta) = \frac{\pi(\theta)f(X,\theta)}{\int \pi(\phi)f(X,\phi)d\phi}. \tag{3}$$

Considering the "bivariate" probability density $\pi(\theta)f(X,\theta)$, π_X is the conditional density of θ given X. (Here X and θ may each be multidimensional.) Now, for a given X, $h(X)$ has a fixed value, which gives a constant multiple in both the numerator and denominator of (3) and so divides out. So $f(X,\psi)$ can be replaced by $G(T(X),\psi)$ for $\psi = \theta$ or ϕ in (3). Thus the conditional distribution of θ given X is the same as its conditional distribution given $T(X)$, and so in this sense also, we see how $T(X)$ is sufficient. This relates to the paragraph "Sufficiency and Bayes models" on p. 46 of Bickel and Doksum, containing Theorem 1.5.2.

To continue with the previous example of normal likelihoods (1), now suppose σ^2 is fixed and known, say $\sigma = 1$ for simplicity, so μ is the only unknown parameter. The likelihood function becomes, for $T_1(X)$ and $T_2(X)$ defined after (1)

$$\frac{1}{(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}\left[n\mu^2 + T_2(X) - 2\mu T_1(X)\right]\right) = G(T_1(X),\mu)h(X)$$

where it's arbitrary which factor $(2\pi)^{-n/2}$ belongs to, $G(T_1(X),\mu)$ is proportional to $\exp(\mu T_1(X) - n\mu^2/2)$ and $h(X)$ is proportional to $\exp(-T_2(X)/2)$. So, for known σ, $T_1(X)$ is a sufficient statistic for μ, as is $\overline{X} = T_1(X)/n$, the sample mean.

When σ is unknown, $\overline{X}$ is still the maximum likelihood estimate of μ, but for further inference about μ, such as confidence intervals in the frequentist view or its posterior distribution in the Bayesian view, we need not only $T_1(X)$ but also $T_2(X)$.

2. *Measure-theoretic background

References to "Dudley (2002)" are to my book *Real Analysis and Probability*, Cambridge University Press edition. Facts from there can probably also be found in other texts.

Let S be any set. A collection $\mathcal{B}$ of subsets of S is called a *σ-algebra* if it satisfies the following conditions:

(a) The empty set $\emptyset$ and S are in $\mathcal{B}$;
(b) The complement $A^c := S \setminus A$ is the set of all x in S not in A. If A is in $\mathcal{B}$, so is A^c.
(c) For any sequence $A_1, A_2, \ldots$ of sets in $\mathcal{B}$, the union $\bigcup_{n=1}^{\infty} A_n$, in other words the set of all x such that $x \in A_n$ for some n, is also in $\mathcal{B}$.

It follows easily from the definition that any intersection of σ-algebras of subsets of S is a σ-algebra of subsets of S. The collection 2^S of all subsets of S is a σ-algebra. Thus for any collection $\mathcal{A}$ of subsets of S, there is a smallest σ-algebra including $\mathcal{A}$, called the σ-algebra *generated* by $\mathcal{A}$, namely, the intersection of all σ-algebras including $\mathcal{A}$, one of which is 2^S. If S is the real line $\mathbb{R}$ then an important σ-algebra of subsets of S is the *Borel* σ-algebra generated by the collection of all open intervals (a, b) for $a < b$ in $\mathbb{R}$.

If S is a set and $\mathcal{B}$ is a σ-algebra of subsets of S then $(S, \mathcal{B})$ is called a *measurable space*.

If $(S, \mathcal{B})$ is a measurable space then a function μ on $\mathcal{B}$ is called a *measure* if:

(d) $0 \leq \mu(A) \leq +\infty$ for all $A \in \mathcal{B}$;
(e) $\mu(\emptyset) = 0$;
(f) For any sequence $A_1, A_2, \ldots,$ of sets in $\mathcal{B}$ which are disjoint, in other words $A_i \cap A_j = \emptyset$ whenever $i \neq j$, we have $\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$.
Then $(S, \mathcal{B}, \mu)$ is called a *measure space*.

A measure μ is called *finite* if $\mu(S) < +\infty$. A *finite signed measure* is a function μ from $\mathcal{B}$ into the real numbers satisfying (e) and (f) above. It can be shown (by the Hahn–Jordan decomposition; Dudley, 2002, Theorem 5.6.1) that equivalently $\mu \equiv \mu^+ - \mu^-$ where μ^+ and μ^- are finite measures.

A main example of a measure space is given by $S = \mathbb{R}$, with $\mathcal{B}$ as the Borel σ-algebra, and μ as *Lebesgue measure* λ, which equals the length for intervals.

A measure space $(S, \mathcal{B}, \mu)$ is called a *probability space*, and μ is called a *probability measure*, if and only if $\mu(S) = 1$. Then μ is often written as P, or as Q if two probability measures are considered, or sometimes as Pr.

Several of the above notions are actually needed to give the usual axiomatization of probability as in beginning probability courses.

If $(S, \mathcal{B})$ is any measurable space, then a real-valued function f on S is called *measurable* if for any Borel set $A \subset \mathbb{R}$, $f^{-1}(A) := \{x : f(x) \in A\} \in \mathcal{B}$. It turns out to be equivalent that for any $t \in \mathbb{R}$, $f^{-1}((t, \infty)) := \{x : f(x) > t\} \in \mathcal{B}$. A *random variable* is a measurable function X on a probability space. $(\Omega, \mathcal{B}, P)$. The *expectation* or *mean* of X is defined by $EX = \int X\, dP$ if X is either nonnegative or integrable.

If S is a countable set such as the set $\mathbb{N}$ of nonnegative integers, then the usual σ-algebra on S will be the collection 2^S of all its subsets. A measure μ on such a set will be called *discrete*. Then μ of any set is given by a sum, $\mu(A) = \sum_{x \in A} \mu(\{x\})$, where $\{x\}$ is the set whose only member is x. Lebesgue measure λ, on the other hand, is not given by such sums, since $\lambda(\{x\}) = 0$ for all x. On S, we have the *counting measure* c where $c(A)$ is the number of elements of A if A is finite and $c(A) = \infty$ if A is infinite. Any probability measure P on the countable set S has a density f with respect to counting measure, which in this case is called a *probability mass function*, or by Bickel and Doksum a *frequency function*. Thus $P(A) = \sum_{x \in A} f(x)$ for any set $A \subset S$.

2.1. ***The Radon–Nikodym theorem.** This theorem, one of the main facts in measure theory, is as follows:

Theorem 1 (Radon–Nikodym)**.** *Let $(S, \mathcal{B})$ be a measurable space. Let μ be a finite signed measure and ν a σ-finite measure on $(S, \mathcal{B})$. Let μ be absolutely continuous with respect to ν, meaning that for each $A \in \mathcal{B}$ with $\nu(A) = 0$, also $\mu(A) = 0$. Then there exists an integrable function $d\mu/d\nu := f$ for ν such that for each $A \in \mathcal{B}$,*

$$\mu(A) = \int_A f\, d\nu := \int f \cdot 1_A d\nu,$$

where $1_A(x) = 1$ for $x \in A$ and 0 otherwise. Here f is unique up to equality except on a set of ν-measure 0.

The theorem is given and proved in Dudley (2002, Theorem 5.5.4 and Corollary 5.6.2). The function $d\mu/d\nu$ is called the *Radon–Nikodym derivative* of μ with respect to ν.

2.2. ***Likelihood ratios.** Given two probability measures P and Q on the same $(S, \mathcal{B})$, how can one define the likelihood ratio $R_{Q/P}$ of

Q with respect to P? If both P and Q have densities, say f and g respectively, with respect to some measure ν, then one can define $R_{Q/P}(x) = (g/f)(x)$, or as $+\infty$ if $f(x) = 0 < g(x)$, or as 0 if $g(x) = f(x) = 0$. In general, P or Q could have continuous or discrete parts. In some applications, there are probabilities on $\mathbb{R}$ having point masses at 0 but having densities for $x > 0$, and so, not belonging to any "regular model" as defined by Bickel and Doksum on p. 9. But, any P and Q are always absolutely continuous with respect to $P+Q$, so that we have:

Definition. Let P and Q be any two probability measures on a measurable space $(S, \mathcal{B})$. Let $h := dP/d(P+Q)$. The *likelihood ratio* $R_{Q/P}(x)$ of Q to P at x is defined as $(1-h(x))/h(x)$, or $+\infty$ if $h(x) = 0$.

The likelihood ratio, like h, is defined up to equality $(P+Q)$-almost everywhere. We have $dQ/d(P+Q) = 1-h$.

2.3. ***Conditional expectations and probabilities.** Let $(S, \mathcal{B}, P)$ be a probability space. Let X be a random variable (measurable real-valued function) defined on S with finite expectation EX, so that $E|X| < +\infty$. Let $\mathcal{A}$ be a sub-σ-algebra of $\mathcal{B}$. Then a *conditional expectation of X with respect to $\mathcal{A}$, written $E(X|\mathcal{A})$, is a real-valued function Y on S, measurable with respect to $\mathcal{A}$, such that $\int_A Y\,dP = \int_A X\,dP$ for all $A \in \mathcal{A}$.*

A conditional expectation always exists and is unique up to equality almost surely for P. One can see this as follows. Let $\mu(B) := \int_B X\,dP$ for all $B \in \mathcal{B}$. Then μ is a finite signed measure, absolutely continuous with respect to P. Consider the restrictions $\mu_{\mathcal{A}}$ and $P_{\mathcal{A}}$ of μ and P respectively to $\mathcal{A}$, which are a finite signed measure and a probability measure on $\mathcal{A}$. By the Radon–Nikodym theorem, $Y := d\mu_{\mathcal{A}}/dP_{\mathcal{A}}$ exists and has the properties required of $E(X|\mathcal{A})$. Let Z be another random variable with these properties. Consider the sets $\{s \in S: \ Y(s) > Z(s)\}$ and $\{s \in S: \ Y(s) < Z(s)\}$. From the definitions, we see that $P = 0$ for each of these sets, and so $P(Y = Z) = 1$.

For a set $B \in \mathcal{B}$ the conditional probability $P(B|\mathcal{A})$ is defined as $E(1_B|\mathcal{A})$.

For any measurable space $(S, \mathcal{B})$ and measurable function T from S into some other measurable space, for example $\mathbb{R}$ with Borel σ-algebra $\mathcal{F}$, the smallest σ-algebra with respect to which T is measurable is

$$\mathcal{A} := T^{-1}(\mathcal{F}) := \{T^{-1}(C): \ C \in \mathcal{F}\}. \tag{4}$$

3. Sufficiency more generally

Given observations $X = (X_1, ..., X_n)$, a *statistic* is a [measurable] function X, say $T(X)$. Sometimes the definition of statistic needs to be extended to allow further randomization, such as resampling from a given sample in the bootstrap method, but in this handout, it will be simply such a function $T(X)$.

Suppose we have a measurable space $(S, \mathcal{B})$ and a collection $\mathcal{P}$ of probability measures on $(S, \mathcal{B})$, which may be a parametrized family $\{P_\theta,\ \theta \in \Theta\}$. Let $X \in S$ be observed, where X is often a vector $(X_1, ..., X_n)$. A statistic $T(X)$ is called *sufficient* for $\mathcal{P}$ "if the conditional distribution of X given $T(X) = t$ does not involve θ" (Bickel and Doksum, p. 42, 3d paragraph). Another way to express this is to say that for any measurable set $B \in \mathcal{B}$ there is a measurable function f_B such that for all $P \in \mathcal{P}$, the conditional probability $P(X \in B | T(X) = t) = f_B(t)$, where conditional probability given T is defined as conditional probability given the σ-algebra defined in (4).

A family $\mathcal{P}$ of probability measures on a measurable space $(S, \mathcal{B})$ is said to be *dominated* by a σ-finite measure μ if every $P \in \mathcal{P}$ is absolutely continuous with respect to μ. Then by the Radon–Nikodym theorem (Theorem 1) there is a function $f_P \geq 0$ with $\int f_P d\mu = 1$ which is the density of P with respect to μ, or the Radon–Nikodym derivative $dP/d\mu$. In regular models as defined by Bickel and Doksum, μ is either Lebesgue (volume) measure λ^d on some d-dimensional Euclidean space $\mathbb{R}^d$, with f_P the density in the usual sense, or μ is counting measure on some countable set B such as the set $\mathbb{N}$ of nonnegative integers, where $\mu(A)$ is the cardinality of A for $A \subset B$ finite, or $+\infty$ for A infinite, and f_P is the probability mass function or frequency function.

The following factorization theorem is Bickel and Doksum's Theorem 1.5.1 in case μ is one of the usual choices as just mentioned and $\mathcal{P}$ is a parametric family.

Theorem 2 (Factorization theorem). *Let $\mathcal{P}$ be a family of probability measures on a measurable space $(S, \mathcal{B})$, dominated by a σ-finite measure μ. Then a statistic $T(X)$ is sufficient for $\mathcal{P}$ if and only if there exists a measurable function $h \geq 0$ on S such that for each $P \in \mathcal{P}$ there is a measurable function g_P such that $f_P(X) = g_P(T(X))h(X)$. If $\mathcal{P}$ is a parametric family $\{P_\theta\}_{\theta\in\Theta}$then there is a function G such that (2) holds, i.e. $f(X, \theta) := (dP_\theta/d\mu)(X) = G(T(X), \theta)h(X)$.*

Bickel and Doksum give a proof in case we have a parametric family in the discrete case, i.e. X takes values in a countable set S. They refer for more general cases to E. L. Lehmann's classic book *Testing*

Statistical Hypotheses. I also gave a proof in my OpenCourseWare notes for 18.466 (2003) but, in this course this semester, we are not concerned with the proof.

A main use of the theorem is that it may be difficult to verify the definition of sufficiency, but relatively easy to find a factorization. So, to show a statistic is sufficient, one just needs to find a factorization (2).

Here is another example, similar to Bickel and Doksum's Example 1.5.3. Consider the family of distributions $P_\theta = U[0,\theta]$ where $0 < \theta < \infty$. Given observed $X_1, ..., X_n$, putting them in order we get the order statistics $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. The likelihood function $f(X,\theta) = \theta^{-n} \prod_{j=1}^n 1_{\{0 \le X_j \le \theta\}}$ can be written, assuming that all $X_j \ge 0$, as $\theta^{-n} 1_{\{X_{(n)} \le \theta\}}$, showing that $X_{(n)}$ is a sufficient statistic for the family.

For any real-valued observations $X = (X_1, ..., X_n)$, we can form the vector of order statistics $S(X) = (X_{(1)}, \ldots, X_{(n)})$. (Here "$S$" connotes "sorted:" the R command "sort" gives $\text{sort}(X_1, ..., X_n) = (X_{(1)}, ..., X_{(n)})$.) For any family $\{P^n : P \in \mathcal{Q}\}$ of distributions on $\mathbb{R}^n$ for any family $\mathcal{Q}$ of probability distributions on $\mathbb{R}$, so that the coordinates are i.i.d. P for some $P \in \mathcal{Q}$, $S(X)$ is a sufficient statistic. We can see this easily if the distributions $P \in \mathcal{Q}$ all have densities f_P with respect to some σ-finite measure μ, because the likelihood function

$$\prod_{j=1}^n f_P(X_j) \equiv \prod_{j=1}^n f_P(X_{(j)}).$$

Here is the fact in general. One might say it is a nonparametric fact.

Theorem 3. *Let $\mathcal{Q}$ be the set of all probability measures on [the Borel sets of] $\mathbb{R}$. On $\mathbb{R}^n$ let $\mathcal{P} := \{P^n : P \in \mathcal{Q}\}$. Then $S(X)$ is a sufficient statistic for P.*

Proof. Let S_n be the set of all $n!$ permutations of $\{1, ..., n\}$. Given $X = (X_1, ..., X_n) \in \mathbb{R}^n$ and $\pi \in S_n$ let $X_{(\pi)} := (X_{(\pi(1))}, ..., X_{(\pi(n))})$. First suppose the X_j are all distinct. Then given $S(X)$, X has $n!$ values with probability $1/n!$ each. Thus for any [measurable (Borel)] set $B \subset \mathbb{R}^n$ and $P \in \mathcal{Q}$,

$$P^n(X \in B | S(X)) = \frac{1}{n!} \sum_{\pi \in S_n} 1_B(X_{(\pi)})$$

which does not depend on P, as desired.

If X_j are not all distinct, but some values occur with multiplicity > 1, then $X_{(\pi)}$ all occur with multiplicities, but the above displayed equation remains true. □

4. Minimal sufficiency

If $T = T(X)$ is a sufficient statistic for a family $\mathcal{P}$ of probability distributions, it is called *minimal sufficient* if for any other sufficient statistic $U = U(X)$, T is a function of U. Bickel and Doksum have a treatment of minimal sufficiency starting with the last paragraph of p. 46 and continuing through the first part of p. 48.

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, and x is not a function of $|x|$.

Example. Let's show that for the family of uniform $U[0, \theta]$ distributions for $\theta > 0$, the statistic $X_{(n)}$, seen to be sufficient above, is actually minimal sufficient (Problem 11, p. 86 of Bickel and Doksum). Let $T(X)$ be another sufficient statistic. By the factorization theorem, we have for all $\theta > 0$

$$f(X, \theta) = \theta^{-n} 1_{X_{(n)} \leq \theta} = G(T(X), \theta) h(X)$$

for some functions G and h. Considering large values of θ, we see that $h(X) > 0$ for any X_j, $j = 1, ..., n$, all > 0 (just take $\theta > X_{(n)}$). Thus $f(X, \theta) > 0$ if and only if $G(T(X), \theta) > 0$. Then

$$X_{(n)} = \inf\{\theta : \ f(X, \theta) > 0\} = \inf\{\theta : \ G(T(X), \theta) > 0\},$$

where the infima can be restricted to rational values of $\theta > 0$ to assure measurability. The last expression is a function of $T(X)$, so $X_{(n)}$ is indeed minimal sufficient.