On sufficiency and related issues.

Moulinath Banerjee

University of Michigan

November 16, 2001

In what follows we shall assume the following set-up: $X_1, X_2, \ldots X_n$ are i.i.d. with common distribution P_θ where $\theta \in \Theta$, the parameter space. Let the common density or frequency function of the X_i's be denoted by $f(x, \theta)$. The vector of observations $(X_1, X_2, \ldots, X_n)$ is denoted by X. The joint density of the observations is then given by,

$$f(X, \theta) = \Pi_{i=1}^n f(X_i, \theta)\,.$$

A statistic $T(X)$ is said to be **sufficient** for θ if the conditional distribution of X given $T(X)$ is independent of θ (independence as used here should not be confused with stochastic independence of random variables). Thus, once you know $T(X)$ you have captured all information about θ that your original data has to offer - in the presence of $T(X)$, your original data is completely non-informative about θ. Note that, by transforming from X to a function $T(X)$ we will generally lose some information, but not any information that is relevant to making inference on θ. Thus sufficiency is essentially the art of effective data-condensation (without compromising information about the parameter of interest). Clearly $T(X) = X$ is trivially sufficient, but quite useless from a practical standpoint.

We have discussed some examples of sufficient statistics in class, where sufficiency was deduced from first principles - that is, by actually computing the conditional distribution of X, the data, given a statistic $T(X)$. Some of these examples will be illustrated later; the point that I want you to note is that the "bare-hands"' or "first-principle" technique is in general a time consuming and involved exercise, involving how to figure out conditional distributions. In order to fruitfully and efficiently employ the sufficiency criterion one clearly needs a more straightforward and " cut and dried"

method of arriving at a sufficient statistic. The factorization theorem that we discussed in class does precisely that.

The factorization theorem says that $T(X)$ is a sufficient statistic if and only if the joint density of the observations $f(X,\theta) = g(\theta, T(X))\, h(X)$, for known functions g and h. Thus, the joint density/likelihood function splits up into two parts: the part $h(X)$ has nothing to do with the parameter and does not contribute to inference about θ. The part $g(\theta, T(X))$ is the relevant part, bearing information about the parameter and depends on the data only through $T(X)$. The factorization theorem shows that the MLE $\hat{\theta}$ is a function of any sufficient statistic $T(X)$, because to maximize $f(X,\theta)$ as a function of θ it just suffices to maximize $g(\theta, T(X))$ as a function of θ and clearly if $T(X) = T(X^{'})$ for two data vectors X and $X^{'}$, the maximizer of $g(\theta, T(X))$ must coincide with the maximizer of $g(\theta, T(X^{'}))$. Another conclusion that we easily draw from the factorization theorem is that a 1-1 function of a sufficient statistic is a sufficient statistic. If $T(X)$ is sufficient and $T^{'}(X)$ is a 1-1 function of $T(X)$, then $T^{'}(X) = \psi(T(X))$ for an invertible function ψ; thus $T(X) = \psi^{-1}(T^{'}(X))$. Since $T(X)$ is sufficient,

$$f(X,\theta) = g(\theta, T(X))\, h(X) = g(\theta, \psi^{-1}(T^{'}(X)))\, h(X) = \tilde{g}(\theta, T^{'}(X))\, h(X)\,,$$

where $\tilde{g}(\theta, T^{'}(X)) = g(\theta, \psi^{-1}(T^{'}(X)))$. This shows that $T^{'}(X)$ is sufficient.

Thus, we draw two important conclusions.

- (i) The MLE is a function of every sufficient statistic $T(X)$.
- (ii) A 1-1 function of a sufficient statistic is sufficient

We next discuss the concept of a minimal sufficient statistic. A **minimal sufficient statistic** $T_M(X)$ is a sufficient statistic that can be written as a function of every sufficient statistic. Thus $T_M(X) = \phi(T(X))$ for every sufficient statistic $T(X)$ and ϕ is a function depending on T. Note that a minimal sufficient statistic provides maximal condensation of the data without compromising information on θ - in other words, no further condensation of the data is possible without compromising some information on θ. Once again, it is easy to see that there is no unique minimal sufficient statistic - any 1-1 transformation of a minimal sufficient statistic is minimal sufficient. The "data-condensation" idea can be explained nicely in terms of the partitioning of a sample space into "orbits" of a statistic $T(X)$; we will not pursue that angle here and now. Now, note that if the MLE is a 1-1

function of a sufficient statistic, it is itself sufficient. But, by observation (i), the MLE is also a function of every sufficient statistic $T(X)$. Therefore, the MLE if sufficient, is actually minimal sufficient.

Let's list down our two major observations from this discussion.

- (iii) A 1-1 function of a minimal sufficient statistic is itself minimal sufficient.
- (iv) If the MLE is a 1-1 function of a sufficient statistic, then it is also minimal sufficient.