M4056 Lecture Notes. **Monday, September 20, 2010**

Sufficient Statistics

The most basic task of inferential statistics is to acquire knowledge about a population from knowledge of a sample. We conceptualize this process by means of a model with two layers.

1) The first layer is the assumption that the population itself is described by a probability distribution of some sort. At the present time, we are concerned with cases in which the population distribution is assumed to belong to a *parametrized family*, and we use θ to refer to the parameter (which may be a scalar or a vector). Here are some examples:
   a) The population consists of units that are identical except with respect to a single property, which may have one of several values. (Think M&Ms and colors.) If the population is large, its state is described by the proportion of units with each of the possible values. In this case, θ is a vector that records the proportion of each kind. A special case occurs when there are exactly two possible values, which for convenience we might take to be 0 and 1.
   b) The population consists of units that are identical except with respect to a quantitative property that varies from unit to unit, and the proportion of the population in which that quantity does not exceed a given value is described by a cumulative distribution function that belongs to a parametrized family, e.g., normal(μ, σ^2) or gamma(α, β), etc. We will use θ to refer generically to such parameters.

2) The second layer is the assumption that sampling is modeled by an n-tuple of random variables, $\vec{X} = (X_1, \ldots, X_n)$, that are independent of one another, with each distributed in the same way the population is. The inferential task is to bet on the parameters of the population distribution using the properties of a sample as a source of information.

There is always a lot of information in a sample that is of no use. For example, the variables X_i are interchangeable. If they were re-labelled, the sample would have the same significance. If we have a specific task in mind—for example, if we wish to extract information about the mean of a normal population—then there may be many features of the sample that we can ignore because they care no useable information. Our goal is to formalize this.

Definition. Suppose $\vec{X} = (X_1, \ldots, X_n)$ is a sample from a population with *pdf* equal to $f_X(x|\theta)$ (so the sample distribution is $f_{\vec{X}} = f_X \times \cdots \times f_X$). Let $T(\vec{X})$ be a statistic. (T is simply a function defined on the set of all possible samples.) We say that $T(\vec{X})$ is a *sufficient statistic for* θ if the conditional distribution of $\vec{X}$ given $T(\vec{X}) = t$ is the same for all values of θ, i.e., if $f_{\vec{X}}\big((\vec{x} \mid T(\vec{X}) = t) \mid \theta\big)$ does not change when θ changes.

Example. Suppose $X_1, \ldots, X_n$ are iid Bernoulli with $P(X_i = 1) = \theta$. Let $T = T(\vec{X}) = \sum_{i=1}^n X_i$. Let $\vec{x} \in \{0,1\}^n$, and let $t = T(\vec{x}) = \sum_{i=1}^n x_i$. Then

$$P(\vec{X} = \vec{x} \mid T = t) = \frac{P(\vec{X} = \vec{x} \;\&\; T = t)}{P(T = t)}.$$

Now if $\vec{X} = \vec{x}$, then the condition $T = t$ is redundant, so

$$P(\vec{X} = \vec{x} \;\&\; T = t) = P(\vec{X} = \vec{x}) = \theta^t (1-\theta)^{n-t}.$$

On the other hand, $P(T=t) = \binom{n}{t}\theta^t(1-\theta)^{n-t}$. Thus

$$P(\vec{X} = \vec{x} \mid T = t) = 1 \Big/ \binom{n}{t}.$$

Comment. The meaning of the definition is simply that $T(\vec{X})$ tells us all there is to know about θ; we gain no additional information about θ if we know more about the sample. In the example just given, it is intuitively clear that knowing more about the sample than how many 1s were obtained could not possibly give us more information about θ, since the X_i play a symmetrical role. In general, the definition says that if $T(\vec{x}) = T(\vec{x}')$, for two different values of the sample $\vec{X}$, then whatever conclusions we draw from observing $\vec{X} = \vec{x}$ must be the same as those we draw from observing $\vec{X} = \vec{x}'$.

Let $P_\theta(A \mid B)$ be the conditional probability of A given B, as a function of the parameter θ. In the discrete case, the definition of sufficiency means that $P_\theta(\vec{X} = \vec{x} \mid T(\vec{X}) = T(\vec{x}))$ is constant as a function of θ. But

$$\begin{aligned} P_\theta(\vec{X} = \vec{x} \mid T(\vec{X}) = T(\vec{x})) &= \frac{P_\theta(\vec{X} = \vec{x} \;\&\; T(\vec{X}) = T(\vec{x}))}{P_\theta(T(\vec{X}) = T(\vec{x}))} \\ &= \frac{P_\theta(\vec{X} = \vec{x})}{P_\theta(T(\vec{X}) = T(\vec{x}))} \qquad (*) \\ &= \frac{p_{\vec{X}}(\vec{x}|\theta)}{p_T(T(\vec{x})|\theta)}. \end{aligned}$$

In other words, T is sufficient for θ if and only if the ratio of probability mass functions $\frac{p_{\vec{X}}(\vec{x}|\theta)}{p_T(T(\vec{x})|\theta)}$ is constant when regarded as a function of θ.

This generalizes to the continuous case (see the comment in the 3rd paragraph from the bottom on page 272) and gives the following (cf. Theorem 6.2.2, page 274):

Theorem. *T is sufficient for θ if and only if*

$$\frac{f_{\vec{X}}(\vec{x}|\theta)}{f_T(T(\vec{x})|\theta)}$$

is constant when regarded as a function of θ.

Example. Suppose $X_1, \ldots, X_n$ are iid $n(\mu, \sigma^2)$. We will show that $T(\vec{X}) = \overline{X}$ is a sufficient statistic for μ. Note that

$$\begin{aligned} f_{\vec{X}}(\vec{x}|\mu) &= (2\pi\sigma^2)^{-n/2} \exp\Big(-\sum_{i=1}^n (x_i - \mu)^2/(2\sigma^2)\Big) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-n}{2\sigma^2}(\overline{x} - \mu)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \overline{x})^2\right) \qquad \text{(using (5.2.4.b).} \end{aligned}$$

On the other hand, the sample mean is $n(\mu, \sigma^2/n)$:

$$f_{\overline{X}}(\overline{x}|\mu) = (2\pi\sigma^2/n)^{-1/2} \exp\left(\frac{-n}{2\sigma^2}(\overline{x} - \mu)^2\right)$$

When we form the ratio $\frac{f_{\vec{X}}(\vec{x}|\mu)}{f_{\overline{X}}(\overline{x}|\mu)}$, the terms in the exponent containing μ cancel, and the ratio is independent of μ. This shows that the sample mean is a sufficient statistic for μ.