Probability and Statistics

Fall 2000

LECT. 12: EVALUATING ESTIMATORS, SUFFICIENCY, CB 7.3.1, CB 6.1.1

We are still in the context of a random variable X with a probability density (mass) function $f_X(x;\theta^*)$, for some unknown θ^*, and with the function $f_X(\cdot)$ known. We have considered three approaches to estimating θ^* – Bayesian, Method of Moments and Maximum Likelihood. How should we think about comparing these estimators. Given an estimator $W(X)$, and a realization x, our estimate is $W(x)$. The estimation error is

$$W(x) - \theta^*.$$

We clearly want to make this error as close to zero as possible. Suppose we measure the "loss" as a result of the estimation error as quadratic in the error:

$$\text{Loss} = (W(x) - \theta^*)^2.$$

Given two estimators $W_1(X)$ and $W_2(X)$ we can then attempt to choose between them on the grounds of expected loss, in this case mean squared error. For an estimator $W(X)$ this equals

$$E[(W(X) - \theta^*)^2.$$

Is this a useful criterion for choosing estimators? Consider a random variable X with an exponential distribution with mean μ. We have a single observation. Recall

$$E[X] = \mu,$$

$$E[X^2] = 2\mu^2.$$

So consider the four estimators

$$W_1(X) = X,$$

$$W_2(X) = 2 \cdot X,$$

and, based on the fact that $E[X^2] = 2\mu^2$,

$$W_3(X) = \sqrt{X^2/2} = X/\sqrt{2},$$

and finally,

$$W_4(X) = 3.$$

Which of these estimators is best in terms of mean squared error? In general, for linear estimators,

$$E\left[(a \cdot X + b - \mu)^2\right] = (E[a \cdot X + b - \mu])^2 + V(a \cdot X + b)$$

$$= ((a-1) \cdot \mu + b)^2 + a^2 \cdot \mu^2 = b^2 + b \cdot (a-1) \cdot \mu + ((a-1)^2 + 1) \cdot \mu^2.$$

Calculating this for all four estimators we get

$$E[(W_1 - \mu)^2] = \mu^2,$$

$$E[(W_2 - \mu)^2] = 5 \cdot \mu^2,$$

$$E[(W_3 - \mu)^2] = (4 - 2 \cdot \sqrt{2}) \cdot \mu^2,$$

and

$$E[(W_4 - \mu)^2] = (3 - \mu)^2.$$

We can rule out the second and third estimator (the latter only barely), but we cannot determine whether W_4 or W_1 is better; the answer depends on the value of μ which is exactly what we are trying to estimate. The conclusion is the mean squared error is not a

generally useful criterion for evaluating estimators because there are many estimators that cannot be ranked according to this criterion.

We therefore add a condition to the estimators, implying that the set of estimators we look at is limited enough for the criterion to be useful. If you look at the example you see the reason that we cannot rule out W_3, although that clearly is a silly estimator, is that it does extremely well in some cases, namely when μ is close to 3. It does so, however, at the expense of doing really badly for other values of μ. We therefore limit the search to estimators that do well in some sense everywhere. Specifically we restrict ourselves to unbiased estimators, that is estimators with

$$E[W(X)] = \theta^*,$$

for all possible values of θ^*. This clearly rules out W_3 in the above example.

This does not, however, come without a price. It rules out some estimators that are perfectly sensible. Suppose X has a normal distribution with mean μ and variance σ^2, with both μ and σ^2 unknown. Consider the two estimators for σ^2:

$$W_1 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2,$$

and the maximum likelihood estimator

$$W_2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2.$$

The first estimator W_1 is unbiased while $W_2 = W_1(N-1)/N$ is not. To see this, consider

$$E\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right] = E\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x} + \bar{x} - \mu)^2\right]$$

$$= E\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right] + E\left[\frac{1}{N}\sum_{i=1}^{N}(\bar{x} - \mu)^2\right] + E\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})\cdot(\bar{x} - \mu)\right].$$

The expectation is equal to zero because the sum is equal to zero:

$$\sum_{i=1}^{N}(x_i - \bar{x})\cdot(\bar{x} - \mu) = (\bar{x} - \mu)\cdot\sum_{i=1}^{N}(x_i - \bar{x}) = 0.$$

Hence

$$\sigma^2 = E\left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2\right] = E\left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right] + \sigma^2/N,$$

which implies

$$E\left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right] = \sigma^2 \cdot \frac{N-1}{N}.$$

However, the mean squared error of W_2 is lower than the mean squared error of W_1. In other words, by restricting ourselves to unbiased estimators we are limiting ourselves in a serious way, and not just getting rid of useless estimators.

Now let us consider the following problem. We have a random variable X from a distribution $f_X(x;\theta^*)$. We are looking for the estimator that minimizes mean squared error among all unbiased estimators, or what is called the minimum variance unbiased estimator. This is a "big" problem, and we will put some structure on this search. The first step is to consider sufficient statistics.

Example
Suppose (X,Y) has a joint pdf

$$f_{YX}(y,x;\theta) = \theta\exp(-x\theta)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right).$$

In other words, X and Y are independent, Y is standard normal and X has an exponential distribution with arrival rate θ. How is Y useful for estimating θ? It clearly is not. Hence any estimator that is "optimal", should depend only on X and not depend on Y. This concept of all the information about the parameter being contained in part of the random variables is call sufficiency. □

Definition 1 *A statistic $T(X)$ is a sufficient statistic for a parameter θ if the distribution of X given T does not depend on θ.*

First consider the example above. In that case $T(X,Y) = X$ is a sufficient statistic. The distribution of (X,Y) given X does not depend on θ. This is clear: the conditional

distribution of Y given X is standard normal, and given X, X is a degenerate random variable.

Intuitively, a sufficient statistic captures all the information about θ that is available in the sample. The distribution of X given T does not change depending on the value of θ, so there is no point in using those values.

Example
Suppose $X_1, \ldots, X_N$ have Binomial distributions with parameters 1 and p. The joint pmf of $X_1, \ldots, X_N$ is

$$f_{X_1,\ldots,X_N}(x_1, \ldots, x_N; p) = p^{\sum X_i} \cdot (1-p)^{N-\sum X_i}.$$

A guess for a sufficient statistic is the number of successes, $T = \sum X_i$. The pmf of T is binomial with parameters N and p:

$$f_T(t; p) = \left(\begin{array}{c} N \\ t \end{array} \right) \cdot p^t \cdot (1-p)^{N-t}.$$

The conditional pdf of $X_1, \ldots, X_N$ given $T = t$ is

$$f_{X_1,\ldots,X_N|T}(x_1, \ldots, x_N | T = t) = 1 \Big/ \left(\begin{array}{c} N \\ t \end{array} \right),$$

for $t = \sum x_i$, and zero elsewhere. This distribution does not depend on p, and so the order of the successes and failures is not important, only the total number. □

There are two issues left. One is the question how to find sufficient statistics, and second how to actually use the result that you only have to consider estimators that are a function of the sufficient statistics.

First consider the problem of finding a sufficient statistic. The most important tool is the following factorization theorem:

Result 1 (FACTORIZATION THEOREM)
Let $f_X(x; \theta)$ denote the pdf of a random variable X. A statistic $T = t(X)$ is a sufficient

statistic for θ if and only if there are functions $g(t)$ and $h(x)$ such that the pdf can be written as

$$f_X(x;\theta) = g(t(x);\theta) \cdot h(x),$$

for all values of θ.

Proof of the Factorization Theorem (for the discrete case only).

First we proof that the factorization implies that T is indeed sufficient. Consider the marginal density of T:

$$f_T(t;\theta) = \sum_x f_{T,X}(t,x;\theta).$$

Conditional on $X = x$, T has a degenerate distribution, as it is a function of X: $pr(T = t) = 1$ for $t = t(x)$ and zero elsewhere. Hence,

$$f_{T,X}(t,x;\theta) = f_X(x;\theta) = g(t;\theta) \cdot h(x),$$

for $t = t(x)$, and zero elsewhere. Then

$$f_T(t;\theta) = g(t;\theta) \sum_x h(x).$$

The conditional density of X given $T = t$ is then

$$f_{X|T}(x|T = t) = \frac{f_{X,T}(x,t;\theta)}{f_T(t;\theta)} = \frac{g(t;\theta) \cdot h(x)}{g(t;\theta) \sum_z h(z)} = \frac{h(x)}{\sum_z h(z)},$$

for x such that $T(x) = t$, and zero elswhere.

For the second part of the result, suppose that the conditional distribution of X given $T = t$ does not depend on θ. Then the marginal density of T is

$$f_T(t;\theta) = f_{T,X}(t,x;\theta) \Big/ f_{X|T}(x|T = t).$$

Hence we can choose $h(x) = f_X(x)$ and $g(t;\theta) = f_T(t;\theta)$ and the factorization is trivial.

Example

Suppose $X_1, \ldots, X_N$ are iid exponential with arrival rate λ. The joint pdf is

$$f_{X_1,\ldots,X_N}(x_1, \ldots, x_N; \lambda) = \prod_{i=1}^{N} \lambda \exp(-x_i\lambda) = \lambda^N \exp\Big(-\sum_{i=1}^{N} x_i\lambda\Big).$$

Hence $T = \sum X_i$ is a sufficient statistic. □

Example

Suppose $X_1, \ldots, X_N$ are iid cauchy with parameter θ. The joint pdf is

$$f_{X_1,\ldots,X_N}(x_1, \ldots, x_N; \lambda) = \prod_{i=1}^{N} \frac{1}{\pi} \frac{1}{1 + (x_i - \theta)^2}.$$

This cannot be simplified and there is no one–dimensional sufficient statistic. □

Sufficient statistics are <u>not unique</u>. Any one-to-one function of a sufficient statistic is a sufficient statistic. Also, the full set of X's is always a sufficient statistic, but we are more interested in sufficient statistics of dimension lower than the sample space. Ideally we would like to find <u>minimal sufficient statistics</u>, sufficient statistics that can be written as a function of any other sufficient statistic. For example in the exponential example, $T_1 = (X_1, X_2, \ldots, X_N)$ and $T_2 = \sum X_i$ are both sufficient statistics, but T_2 can be written as a function of T_1, but not the other way around. If for <u>any</u> sufficient statistic T we can write $\tilde{T}$ as a function of T, then $\tilde{T}$ is minimal sufficient.