

Lecture Note 13: Evaluating Estimators Part 1 (CB 7.3, CB 6.1-6.2)

Suppose you do a Bayesian analysis, and you feel comfortable with the subjectivist interpretation that the distribution over parameters represents your beliefs about the likelihood of different parameter values. Moreover, you feel comfortable with the choice of prior and the form of the likelihood. Then, the posterior distribution completely captures your updated beliefs about the parameter.

However, in practice, you might have chosen a prior partly on the basis of convenience. Or you might not like the subjectivist interpretation. Or you might use another estimator (method of moments, maximum likelihood). In that case, it is useful to simply regard the estimator as a *procedure* for transformation data into point estimates of parameters, and try to evaluate these estimators from a repeated sampling perspective. That is, we will assume that the parameter is equal to some particular value, say θ^* , and study the properties of the estimator assuming the data are generated according to θ^* .

So, let us consider a situation where we observe a variable X , with density $f_X(x; \theta^*)$. (The variable X might be a scalar, corresponding to a single observation, or it might be a vector, corresponding to a random sample of size n . In the latter case, we would interpret $f_X(x; \theta^*)$ as the joint density of all the X_i .) Suppose that θ^* is scalar.

Any point estimator is simply a function of X . So let's consider an arbitrary estimator $W(x)$. The estimation error is

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

We want to capture the idea that the estimation error should be small. A convenient way is to define the loss as a result of the estimation error to be the square of the error:

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

(Other choices for the loss are possible, for example absolute error loss.) The expected loss, or risk, of the estimator is

$$\begin{aligned} R(W, \theta^*) &= E[(W(X) - \theta^*)^2] \\ &= \int (W(x) - \theta^*)^2 f_X(x; \theta^*) dx. \end{aligned}$$

This is often called mean squared error. An important fact is that the mean squared error can be decomposed into the variance plus the squared bias:

$$E[(W(X) - \theta^*)^2] = V[W(X)] + \{E[W(X) - \theta^*]\}^2.$$

Example 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,

$$\begin{aligned} E[(a \cdot X + b - \mu)^2] &= (E[a \cdot X + b - \mu])^2 + V(a \cdot X + b) \\ &= ((a - 1) \cdot \mu + b)^2 + a^2 \cdot \mu^2. \end{aligned}$$

Calculating this for all four estimators we get

$$\begin{aligned} E[(W_1 - \mu)^2] &= \mu^2, \\ E[(W_2 - \mu)^2] &= 5 \cdot \mu^2, \\ E[(W_3 - \mu)^2] &= \left(2 - \frac{2}{\sqrt{2}}\right) \cdot \mu^2 < \mu^2, \end{aligned}$$

and

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

The estimator W_3 is better than W_1 or W_2 , but we cannot say whether W_4 or W_3 is better; the answer depends on the value of μ which is exactly what we are trying to estimate. \square

From the previous example, we conclude that different estimators may do well (in terms of risk) for different values of the parameter. In order to choose the “best” estimator, we have to either (a) define some overall measure of the estimator’s performance, or (b) restrict attention to a more limited class of estimators.

An example of approach (a), is to compare estimators by their worst-case risk:

$$r(W) = \max_{\theta^* \in \Theta} R(W, \theta^*).$$

An estimator that minimizes worst-case risk is called minimax. One drawback of the min-max approach, is that it can be very conservative, giving up good performance for most reasonable values of the parameter to avoid do badly in certain extreme parts of the parameter space. In addition, minmax estimators can be hard to calculate, although for many standard models they have been worked out.¹

We will focus on approach (b), and limit the search to estimators that satisfy the additional property of unbiasedness:

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

We will look for the estimator that minimizes mean squared error subject to the restriction that it be unbiased. This is equivalent to minimizing the *variance* subject to unbiasedness, so we are looking for the minimum variance unbiased estimator, or MVUE.

¹For a good introduction to minmax estimation theory, see J. Berger, *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

(Note that restricting attention to unbiased estimators is not always a good idea. It might rule out estimators that are nearly unbiased, and otherwise do better than any unbiased estimator. But we'll continue with the unbiasedness restriction and see where it gets us.)

In some cases, we can simplify the analysis by using the concept of *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 .

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, \dots, X_N have Binomial distributions with parameters 1 and p . The joint pmf of X_1, \dots, X_N is

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

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 :

$$f_T(t; p) = \binom{N}{t} \cdot p^t \cdot (1 - p)^{N-t}.$$

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

$$f_{X_1, \dots, X_N|T}(x_1, \dots, x_N|T = t) = 1 / \binom{N}{t},$$

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. \square

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 prove 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 : $P(T = t) = 1$ 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 elsewhere.

For the second part of the result, suppose that the conditional distribution of X given $T = t$ does not depend on θ . We want to show that $f_X(x; \theta)$ can be written in the factorization form for some $g(\cdot)$ and $h(\cdot)$. We can write

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

Note that $f_{X|T}(x|t) = P(X = x|T(x) = t(x))$ is a function of x . So we can set $h(x) = f_{X|T}(x|t)$ and $g(t; \theta) = f_T(t; \theta)$. \square

Example

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

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

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

Example

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

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

This cannot be simplified and there is no one-dimensional sufficient statistic. \square

Sufficient statistics are not unique. 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 minimal sufficient statistics, sufficient statistics that can be written as a function of any other sufficient statistic. For example in the exponential example, $T_1 = (X_1, X_2, \dots, X_N)$ 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 any sufficient statistic T we can write \tilde{T} as a function of T , then \tilde{T} is minimal sufficient.

Suppose we have a sufficient statistic $T(X)$. How can we use this? The Rao-Blackwell Theorem deals with this question.

Result 2 (RAO-BLACKWELL THEOREM)

Let $W = W(X)$ be any unbiased estimator for θ , and let $T = T(X)$ be a sufficient statistic for θ . Then

$$\tilde{W} = E[W|T],$$

is an unbiased estimator for θ with a variance less than or equal to that of W .

Proof:

First consider the expectation of \tilde{W} . The law of iterated expectations says in general that

$$E[Y] = E[E[Y|X]],$$

because

$$\begin{aligned} E[Y] &= \int \int y f_{XY}(x, y) dy dx = \int \int y f_{Y|X}(y|x) f_X(x) dy dx \\ &= \int E[Y|X] f_X(x) dx = E[E[Y|X]]. \end{aligned}$$

Therefore we have

$$\theta = E[W] = E[E[W|T]] = E[\tilde{W}],$$

showing that \tilde{W} is unbiased. In addition,

$$V(W) = V(E[W|T]) + E[V(W|T)] = V(\tilde{W}) + E[V(W|T)] \geq V(\tilde{W}).$$

We use here the fact that

$$Y = E[Y|X] + Y - E[Y|X],$$

so

$$\begin{aligned} V(Y) &= V(E[Y|X]) + V(Y - E[Y|X]) + 2 \cdot COV(E[Y|X], Y - E[Y|X]) \\ &= V(E[Y|X]) + E[V(Y|X)], \end{aligned}$$

where we use the fact that

$$V(Y - E[Y|X]) = E[(Y - E[Y|X])^2] = E[E[(Y - E[Y|X])^2]] = E[V(Y|X)].$$

□

So one way of using sufficient statistics is to first look for any unbiased estimator and then take its conditional expectation given the sufficient statistic. This will never make you worse off, and can actually improve things. In practice this is not an easy thing to do. The conditional distribution of the unbiased estimator given the sufficient statistic is often difficult to calculate, and calculating its expectation is often even more difficult. The main value of the result is in formalizing the notion that we can do as well with estimators that are functions of the sufficient statistic as with the general set of all estimators.

The reason we can only use this trick with sufficient statistics is that in general $E[W|S]$ depends on the unknown θ , unless S is a sufficient statistic.