STAT 450

Lecture 24

Today's notes
Reading for Today's Lecture:
Goals of Today's Lecture:

- Motivate sufficiency and completeness

What can we do to find UMVUEs when the CRLB is a strict inequality?

Example: Suppose X has a Binomial(n,p) distribution. The score function is

$$U(p) = \frac{1}{p(1-p)}X - \frac{n}{1-p}$$

Thus the CRLB will be strict unless $T = cX$ for some c. If we are trying to estimate p then choosing $c = n^{-1}$ does give an unbiased estimate $\hat{p} = X/n$ and $T = X/n$ achieves the CRLB so it is UMVU.

A different tactic proceeds as follows. Suppose $T(X)$ is some unbiased function of X. Then we have

$$E_p(T(X) - X/n) \equiv 0$$

because $\hat{p} = X/n$ is also unbiased. If $h(k) = T(k) - k/n$ then

$$E_p(h(X)) = \sum_{k=0}^{n} h(k)\binom{n}{k} p^k (1-p)^{n-k} \equiv 0$$

The left hand side of the $\equiv$ sign is a polynomial function of p as is the right. Thus if the left hand side is expanded out the coefficient of each power p^k is 0. The constant term occurs only in the term $k=0$ and its coefficient is

$$h(0)\binom{n}{0} = h(0)$$

Thus $h(0) = 0$. Now $p^1 = p$ occurs only in the term $k = 1$ with coefficient $nh(1)$ so $h(1) = 0$. Since the terms with $k = 0$ or 1 are 0 the quantity p^2 occurs only in the term with $k = 2$ with coefficient

$$n(n-1)h(2)/2$$

so $h(2) = 0$. We can continue in this way to see that in fact $h(k) = 0$ for each k and so the *only* unbiased function of X is X/n.

Now a Binomial random variable is just of sum of n iid Bernoulli(p) random variables. If $Y_1, \ldots, Y_n$ are iid Bernoulli(p) then $X = \sum Y_i$ is Binomial(n, p). Could we do better by than $\hat{p} = X/n$ by trying $T(Y_1, \ldots, Y_n)$ for some other function T?

Let's consider the case $n = 2$ so that there are 4 possible values for Y_1, Y_2. If $h(Y_1, Y_2) = T(Y_1, Y_2) - [Y_1 + Y_2]/2$ then again

$$E_p(h(Y_1, Y_2)) \equiv 0$$

and we have

$$E_p(h(Y_1, Y_2)) = h(0,0)(1-p)^2 + [h(1,0) + h(0,1)]p(1-p) + h(1,1)p^2$$

This can be rewritten in the form

$$\sum_{k=0}^{n} w(k)\binom{n}{k} p^k (1-p)^{n-k}$$

where $w(0) = h(0,0)$, $w(1) = [h(1,0) + h(0,1)]/2$ and $w(2) = h(1,1)$. Just as before it follows that $w(0) = w(1) = w(2) = 0$. This argument can be used to prove that for any unbiased estimate $T(Y_1, \ldots, Y_n)$ we have that the average value of $T(y_1, \ldots, y_n)$ over vectors $y_1, \ldots, y_n$ which have exactly k 1s and $n - k$ 0s is k/n. Now let's look at the variance of T:

$$\begin{aligned}
\mathrm{Var(T)} =& E_p([T(Y_1, \ldots, Y_n) - p]^2) \\
=& E_p([T(Y_1, \ldots, Y_n) - X/n + X/n - p]^2) \\
=& E_p([T(Y_1, \ldots, Y_n) - X/n]^2) \\
& + 2E_p([T(Y_1, \ldots, Y_n) - X/n][X/n - p]) \\
& + E_p([X/n - p]^2)
\end{aligned}$$

I claim that the cross product term is 0 which will prove that the variance of T is the variance of X/n plus a non-negative quantity (which will be positive unless $T(Y_1, \ldots, Y_n) \equiv X/n$). We can compute the cross product term by writing

$$\begin{aligned}
& E_p([T(Y_1, \ldots, Y_n) - X/n][X/n - p]) \\
& \quad = \sum_{y_1, \ldots, y_n} [T(y_1, \ldots, y_n) - \sum y_i/n][\sum y_i/n - p] p^{\sum y_i}(1-p)^{n - \sum y_i}
\end{aligned}$$

We can do the sum by summing over those $y_1, \ldots, y_n$ whose sum is an integer x and then summing over x. We get

$$\begin{aligned}
&E_p([T(Y_1,\ldots,Y_n) - X/n][X/n-p]) \\
&= \sum_{x=0}^{n} \sum_{\sum y_i = x} [T(y_1,\ldots,y_n) - \sum y_i/n][\sum y_i/n - p] p^{\sum y_i}(1-p)^{n-\sum y_i} \\
&\qquad = \sum_{x=0}^{n} \left[\sum_{\sum y_i = x} [T(y_1,\ldots,y_n) - x/n] \right] [x/n - p] p^x (1-p)^{n-x}
\end{aligned}$$

We have already shown that the sum in [] is 0!

This long, algebraically involved, method of proving that $\hat{p} = X/n$ is the UMVUE of p is one special case of a general tactic.

To get more insight I begin by rewriting

$$\begin{aligned}
&E_p(T(Y_1,\ldots,Y_n)) \\
&\qquad \sum_{x=0}^{n} \sum_{\sum y_i = x} T(y_1,\ldots,y_n) P(Y_1 = y_1,\ldots,Y_n = y_n) \\
&= \sum_{x=0}^{n} \sum_{\sum y_i = x} T(y_1,\ldots,y_n) P(Y_1 = y_1,\ldots,Y_n = y_n | X = x) P(X = x) \\
&\qquad = \sum_{x=0}^{n} \frac{\sum_{\sum y_i = x} T(y_1,\ldots,y_n)}{\binom{n}{x}} \binom{n}{x} p^x (1-p)^{n-x}
\end{aligned}$$

Notice that the large fraction in this formula is the average value of T over values of y when $\sum y_i$ is held fixed at x. Notice that the weights in this average do not depend on p. Notice that this average is actually

$$\begin{aligned}
&E(T(Y_1,\ldots,Y_n|X=x)) \\
&\qquad \sum_{y_1,\ldots,y_n} T(y_1,\ldots,y_n) P(Y_1 = y_1,\ldots,Y_n = y_n | X = x)
\end{aligned}$$

Notice that the conditional probabilities do not depend on p. In a sequence of Binomial trials if I tell you that 5 of 17 were heads and the rest tails the actual trial numbers of the 5 Heads are chosen at random from the 17 possibilities; all of the 17 choose 5 possibilities have the same chance and this chance does not depend on p.

Notice that in this problem with data $Y_1, \ldots, Y_n$ the log likelihood is

$$\ell(p) = \sum Y_i \log(p) - (n - \sum Y_i) \log(1-p)$$

and

$$U(p) = \frac{1}{p(1-p)} X - \frac{n}{1-p}$$

as before. Again we see that the CRLB will be a strict inequality except for multiples of X. Since the only unbiased multiple of X is $\hat{p} = X/n$ we see again that $\hat{p}$ is UMVUE for p.

Sufficiency

In the binomial situation the conditional distribution of the data $Y_1, \ldots, Y_n$ given X is the same for all values of θ; we say this conditional distribution is **free** of θ.

Definition: A statistic $T(X)$ is sufficient for the model $\{P_\theta; \theta \in \Theta\}$ if the conditional distribution of the data X given $T = t$ is free of θ.

Intuition: Why do the data tell us about θ? Because different values of θ give different distributions to X. If two different values of θ correspond to the same joint density or cdf for X then we cannot, even in principle, distinguish these two values of θ by examining X. We extend this notion to the following. If two values of θ give the same conditional distribution of X given T then observing T in addition to X does not improve our ability to distinguish the two values.

Mathematically Precise version of this intuition: If $T(X)$ is a sufficient statistic then we can do the following. If $S(X)$ is any estimate or confidence interval or whatever for a given problem but we only know the value of T then:

- Generate an observation X^* (via some sort of Monte Carlo program) from the conditional distribution of X given T.

- Use $S(X^*)$ instead of $S(X)$. Then $S(X^*)$ has the same performance characteristics as $S(X)$ because the distribution of X^* is the same as that of X.

You can carry out the first step **only** if the statistic T is sufficient; otherwise you need to know the true value of θ to generate X^*.
Example 1: IF $Y_1, \ldots, Y_n$ are iid Bernoulli(p) then given $\sum Y_i = y$ the indexes of the y successes have the same chance of being any one of the $\binom{n}{y}$ possible subsets of $\{1, \ldots, n\}$. This chance does not depend on p so $T(Y_1, \ldots, Y_n) = \sum Y_i$ is a sufficient statistic.
example 2: If $X_1, \ldots, X_n$ are iid $N(\mu, 1)$ then the joint distribution of $X_1, \ldots, X_n, \overline{X}$ is multivariate normal with mean vector whose entries are all μ and variance covariance matrix which can be partitioned as

$$\left[\begin{array}{cc} I_{n\times n} & \mathbf{1}_n/n \\ \mathbf{1}_n^t/n & 1/n \end{array}\right]$$

where $\mathbf{1}_n$ is a column vector of n 1s and $I_{n\times n}$ is an $n \times n$ identity matrix.

You can now compute the conditional means and variances of X_i given $\overline{X}$ and use the fact that the conditional law is multivariate normal to prove that the conditional distribution of the data given $\overline{X} = x$ is multivariate normal with mean vector all of whose entries are x and variance-covariance matrix given by $I_{n\times n} - \mathbf{1}_n\mathbf{1}_n^t/n$. Since this does not depend on μ we find that $\overline{X}$ is sufficient.
WARNING: Whether or not a statistic is sufficient depends on the density function and on Θ.

Rao Blackwell Theorem

Theorem: Suppose that $S(X)$ is a sufficient statistic for some model $\{P_\theta, \theta \in \Theta\}$. If T is an estimate of some parameter $\phi(\theta)$ then:

1. $E(T|S)$ is a statistic.
2. $E(T|S)$ has the same bias as T; if T is unbiased so is $E(T|S)$.
3. $\text{Var}_\theta(E(T|S)) \le \text{Var}_\theta(T)$ and the inequality is strict unless T is a function of S.
4. The MSE of $E(T|S)$ is no more than that of T.

Proof: It will be useful to review conditional distributions a bit more carefully at this point. The abstract definition of conditional expectation is this:

Definition: $E(Y|X)$ is any function of X such that

$$E\left[R(X)E(Y|X)\right] = E\left[R(X)Y\right]$$

for any function $R(X)$.
Definition: $E(Y|X=x)$ is a function $g(x)$ such that

$$g(X) = E(Y|X)$$

Fact: If X, Y has joint density $f_{X,Y}(x,y)$ and conditional density $f(y|x)$ then

$$g(x) = \int y f(y|x) dy$$

satisfies these definitions.
Proof:

$$\begin{aligned}
E(R(X)g(X)) &= \int R(x)g(x)f_X(x)dx \\
&= \int\int R(x)yf(y|x)f_X(x)dydx \\
&= \int\int R(x)yf_{X,Y}(x,y)dydx \\
&= E(R(X)Y)
\end{aligned}$$

You should simply think of $E(Y|X)$ as being what you get when you average Y holding X fixed. It behaves like an ordinary expected value but where functions of X only are like constants.

Proof of the Rao Blackwell Theorem

Step 1: The definition of sufficiency is that the conditional distribution of X given S does not depend on θ. This means that $E(T(X)|S)$ does not depend on θ.
Step 2: This step hinges on the following identity (called Adam's law by Jerzy Neyman – he used to say it comes before all the others)

$$E[E(Y|X)] = E(Y)$$

which is just the definition of $E(Y|X)$ with $R(X) \equiv 1$.

From this we deduce that

$$E_\theta[E(T|S)] = E_\theta(T)$$

so that $E(T|S)$ and T have the same bias. If T is unbiased then

$$E_\theta[E(T|S)] = E_\theta(T) = \phi(\theta)$$

so that $E(T|S)$ is unbiased for ϕ.
Step 3: This relies on the following very useful decomposition. (In regression courses we say that the total sum of squares is the sum of the regression sum of squares plus the residual sum of squares.)

$$\mathrm{Var(Y)} = \mathrm{Var}(E(Y|X)) + E[\mathrm{Var}(Y|X)]$$

The conditional variance means

$$\mathrm{Var}(Y|X) = E[(Y - E(Y|X))^2|X]$$

This identity is just a matter of squaring out the right hand side

$$\mathrm{Var}(E(Y|X)) = E[(E(Y|X) - E[E(Y|X)])^2] = E[(E(Y|X) - E(Y))^2]$$

and

$$E[\mathrm{Var}(Y|X)] = E[(Y - E(Y|X)^2]$$

Adding these together gives

$$E[Y^2 - 2YE[Y|X] + 2(E[Y|X])^2 - 2E(Y)E[Y|X] + E^2(Y)]$$

The middle term actually simplifies. First, remember that $E(Y|X)$ is a function of X so can be treated as a constant when holding X fixed. This means

$$E[Y|X]E[Y|X] = E[YE(Y|X)|X]$$

and taking expectations gives

$$E[(E[Y|X])^2] = E[E[YE(Y|X)|X]] = E[YE(Y|X)]$$

This makes the middle term above cancel with the second term. Moreover the fourth term simplifies

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

so that

$$\mathrm{Var}(E(Y|X)) + E[\mathrm{Var}(Y|X)] = E[Y^2] - E^2(Y)$$

We apply this to the Rao Blackwell theorem to get

$$\text{Var}_\theta(T) = \text{Var}_\theta(E(T|S)) + E[(T - E(T|S))^2]$$

The second term is non negative so that the variance of $E(T|S)$ must be no more than that of T and will be strictly less unless $T = E(T|S)$. This would mean that T is already a function of S. Adding the squares of he biases of T (or of $E(T|S)$ which is the same) gives the inequality for mean squared error.

Examples:

In the binomial problem $Y_1(1 - Y_2)$ is an unbiased estimate of $p(1 - p)$. We improve this by computing

$$E(Y_1(1 - Y_2)|X)$$

We do this in two steps. First compute

$$E(Y_1(1 - Y_2)|X = x)$$

Notice that the random variable $Y_1(1 - Y_2)$ is either 1 or 0 so its expected value is just the probability it is equal to 1:

$$\begin{aligned}
E(Y_1(1 - Y_2)|X = x) &= P(Y_1(1 - Y_2) = 1|X = x) \\
&= P(Y_1 = 1, Y_2 = 0|Y_1 + Y_2 + \cdots + Y_n = x) \\
&= \frac{P(Y_1 = 1, Y_2 = 0, Y_1 + Y_2 + \cdots + Y_n = x)}{P(Y_1 + Y_2 + \cdots + Y_n = x)} \\
&= \frac{P(Y_1 = 1, Y_2 = 0, Y_3 + Y_4 + \cdots + Y_n = x - 1)}{\binom{n}{x} p^x (1 - p)^{n-x}} \\
&= \frac{p(1 - p)\binom{n-2}{x-1} p^{x-1} (1 - p)^{(n-1)-(x-1)}}{\binom{n}{x} p^x (1 - p)^{n-x}} \\
&= \frac{\binom{n-2}{x-1}}{\binom{n}{x}} \\
&= \frac{x(n - x)}{n(n - 1)}
\end{aligned}$$

This is simply $n\hat{p}(1-\hat{p})/(n-1)$ (which can be bigger than $1/4$ which is the maximum value of $p(1-p)$).

Example: If $X_1, \ldots, X_n$ are iid $N(\mu, 1)$ then $\bar{X}$ is sufficient and X_1 is an unbiased estimate of μ. Now

$$\begin{aligned} E(X_1|\bar{X}) &= E[X_1 - \bar{X} + \bar{X}|\bar{X}] \\ &= E[X_1 - \bar{X}|\bar{X}] + \bar{X} \\ &= \bar{X} \end{aligned}$$

which is the UMVUE.