STAT 450

Lecture 25

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

- Motivate sufficiency and completeness

Completeness and Sufficiency

Example: $Y_1, \ldots, Y_n$ iid Bernoulli(p), $\sum Y_i = X \sim$Binomial(n, p):

- If $\mathrm{E}[h(X)] = p$ for all p then $h(X) = X/n$.
- If $\mathrm{E}[h(Y_1, \ldots, Y_n)] = p$ for all p then

$$\frac{\sum_{\sum y_i = x} h(y_1, \ldots, y_n)}{\binom{n}{x}} = \frac{x}{n}$$

- Fact if $\sum y_i = x$:

$$P(Y_1 = y_1, \cdots, Y_n = y_n | X = x) = \frac{1}{\binom{n}{x}}$$

- So:

$$\sum_{\sum y_i = x} h(y_1, \ldots, y_n)$$
$$\times P(Y_1 = y_1, \cdots, Y_n = y_n | X = x) = \frac{x}{n}$$

Sufficiency

Notice: conditional distribution of $Y_1, \ldots, Y_n$ given X same for all θ; we say 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

$$\begin{bmatrix} I_{n\times n} & \mathbf{1}_n/n \\ \mathbf{1}_n^t/n & 1/n \end{bmatrix}$$

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. $\mathrm{Var}_\theta(E(T|S)) \leq \mathrm{Var}_\theta(T)$ 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

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

and

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

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

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

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.