STAT 801: Mathematical Statistics

Unbiased Estimation

The problem above illustrates a general phenomenon. An estimator can be good for some values of θ and bad for others. When comparing $\hat\theta$ and $\tilde\theta$, two estimators of θ we will say that $\hat\theta$ is better than $\tilde\theta$ if it has *uniformly* smaller MSE:

$$MSE_{\hat\theta}(\theta) \le MSE_{\tilde\theta}(\theta)$$

for **all** θ. Normally we also require that the inequality be strict for at least one θ.

The definition raises the question of the existence of a *best* estimate – one which is better than every other estimator. There is no such estimate. Suppose $\hat\theta$ were such a best estimate. Fix a θ^* in Θ and let $\tilde p \equiv \theta^*$. Then the MSE of $\tilde p$ is 0 when $\theta = \theta^*$. Since $\hat\theta$ is better than $\tilde p$ we must have

$$MSE_{\hat\theta}(\theta^*) = 0$$

so that $\hat\theta = \theta^*$ with probability equal to 1. This makes $\hat\theta = \tilde\theta$. If there are actually two different possible values of θ this gives a contradiction; so no such $\hat\theta$ exists.

Principle of Unbiasedness: A good estimate is unbiased, that is,

$$E_\theta(\hat\theta) \equiv 0\,.$$

WARNING: In my view the Principle of Unbiasedness is a load of hog wash.

For an unbiased estimate the MSE is just the variance.

Definition: An estimator $\hat\phi$ of a parameter $\phi = \phi(\theta)$ is **Uniformly Minimum Variance Unbiased** (UMVU) if, whenever $\tilde\phi$ is an unbiased estimate of ϕ we have

$$\mathrm{Var}_\theta(\hat\phi) \le \mathrm{Var}_\theta(\tilde\phi)$$

We call $\hat\phi$ the UMVUE. ('E' is for Estimator.)

The point of having $\phi(\theta)$ is to study problems like estimating μ when you have two parameters like μ and σ for example.

Cramér Rao Inequality

If $\phi(\theta) = \theta$ we can derive some information from the identity

$$E_\theta(T) \equiv \theta$$

When we worked with the score function we derived some information from the identity

$$\int f(x,\theta)dx \equiv 1$$

by differentiation and we do the same here. If $T = T(X)$ is some function of the data X which is unbiased for θ then

$$E_\theta(T) = \int T(x) f(x,\theta)dx \equiv \theta$$

Differentiate both sides to get

$$\begin{aligned}
1 &= \frac{d}{d\theta}\int T(x)f(x,\theta)dx \\
&= \int T(x)\frac{\partial}{\partial\theta}f(x,\theta)dx \\
&= \int T(x)\frac{\partial}{\partial\theta}\log(f(x,\theta))f(x,\theta)dx \\
&= E_\theta(T(X)U(\theta))
\end{aligned}$$

where U is the score function. Since score has mean 0

$$\text{Cov}_\theta(T(X), U(\theta)) = 1$$

Remember correlations between -1 and 1 or

$$1 = |\text{Cov}_\theta(T(X), U(\theta))| \le \sqrt{\text{Var}_\theta(T)\text{Var}_\theta(U(\theta))}$$

Squaring gives the inequality

$$\text{Var}_\theta(T) \ge \frac{1}{I(\theta)}$$

which is called the Cramér Rao Lower Bound. The inequality is strict unless the correlation is 1 which would require that

$$U(\theta) = A(\theta)T(X) + B(\theta)$$

for non-random constants A and B (may depend on θ.) This would prove that

$$\ell(\theta) = A^*(\theta)T(X) + B^*(\theta) + C(X)$$

for some further constants A^* and B^* and finally

$$f(x,\theta) = h(x)e^{A*(\theta)T(x)+B^*(\theta)}$$

for $h = e^C$.

Summary of Implications

- You can recognize a UMVUE sometimes. If $\text{Var}_\theta(T(X)) \equiv 1/I(\theta)$ then $T(X)$ is the UMVUE. In the $N(\mu, 1)$ example the Fisher information is n and $\text{Var}(\overline{X}) = 1/n$ so that $\overline{X}$ is the UMVUE of μ.
- In an asymptotic sense the MLE is nearly optimal: it is nearly unbiased and (approximate) variance nearly $1/I(\theta)$.
- Good estimates are highly correlated with the score.
- Densities of exponential form (called *exponential family*) given above are somehow special.
- Usually inequality is strict — strict unless score is affine function of a statistic T and T (or T/c for constant c) is unbiased for θ.

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

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

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.

Different tactic: 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

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

LHS of $\equiv$ sign is 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.

A Binomial random variable is a sum of n iid Bernoulli(p) rvs. If $Y_1, \ldots, Y_n$ 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?

Try $n = 2$. 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

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

and we have

$$\begin{aligned} E_p(h(Y_1, Y_2)) &= h(0,0)(1-p)^2 \\ &\quad +[h(1,0) + h(0,1)]p(1-p) \\ &\quad +h(1,1)p^2 . \end{aligned}$$

This can be rewritten in the form

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

where

$$\begin{aligned} w(0) &= h(0,0) \\ 2w(1) &= h(1,0) + h(0,1) \\ w(2) &= h(1,1) . \end{aligned}$$

So, as before $w(0) = w(1) = w(2) = 0$. 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:

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

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

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

Sum over those $y_1, \ldots, y_n$ whose sum is an integer x; then sum over x:

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

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 rewrite

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

Notice large fraction in formula is 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

$$
E\{T(Y_1,\ldots,Y_n)|X = x\} = \sum_{y_1,\ldots,y_n} T(y_1,\ldots,y_n) \times P(Y_1 = y_1,\ldots,Y_n = y_n | X = x)
$$

Notice 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: with data $Y_1,\ldots,Y_n$ 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 CRLB is strict except for multiples of X. Since only unbiased multiple of X is $\hat{p} = X/n$ UMVUE of p is $\hat{p}$.

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.

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

Intuition: Data tell us about θ **if** different values of θ give different distributions to X. If two different values of θ correspond to same density or cdf for X we cannot distinguish these two values of θ by examining X. Extension of this notion: if two values of θ give same conditional distribution of X given T then observing T in addition to X doesn't improve our ability to distinguish the two values.

Mathematically Precise version of this intuition: Suppose $T(X)$ is sufficient statistic and $S(X)$ is any estimate or confidence interval or ... If you only know 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: $Y_1, \ldots, Y_n$ iid Bernoulli(p). 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 $\{1, \ldots, n\}$. Chance does not depend on p so $T(Y_1, \ldots, Y_n) = \sum Y_i$ is sufficient statistic.
Example 2: $X_1, \ldots, X_n$ iid $N(\mu, 1)$. Joint distribution of $X_1, \ldots, X_n, \overline{X}$ is MVN. All entries of mean vector are μ. Variance covariance matrix 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 column vector of n 1s and $I_{n\times n}$ is $n \times n$ identity matrix.

Compute conditional means and variances of X_i given $\overline{X}$; use fact that conditional law is MVN. Conclude conditional law of data given $\overline{X} = x$ is MVN. Mean vector has all entries x. Variance-covariance matrix is $I_{n\times n} - \mathbf{1}_n\mathbf{1}_n^t/n$. No dependence on μ so $\overline{X}$ is sufficient.
WARNING: Whether or not statistic is sufficient depends on density function and on Θ.
Theorem: [Rao-Blackwell] Suppose $S(X)$ is a sufficient statistic for model $\{P_\theta, \theta \in \Theta\}$. If T is an estimate of $\phi(\theta)$ 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. $\mathrm{Var}_\theta(E(T|S)) \le \mathrm{Var}_\theta(T)$ and the inequality is strict unless T is a function of S.
4. MSE of $E(T|S)$ is no more than MSE of T.

Proof: Review conditional distributions: abstract definition of conditional expectation is:
Defn: $E(Y|X)$ is any function of X such that

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

for any function $R(X)$. $E(Y|X = x)$ 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:

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

Think of $E(Y|X)$ as average Y holding X fixed. Behaves like ordinary expected value but functions of X only are like constants:

$$E(\sum A_i(X)Y_i|X) = \sum A_i(X)E(Y_i|X)$$

Example: $Y_1, \ldots, Y_n$ iid Bernoulli(p). Then $X = \sum Y_i$ is Binomial(n, p). Summary of conclusions:

- Log likelihood function of X only not of $Y_1, \ldots, Y_n$.
- Only function of X which is unbiased estimate of p is $\hat{p} = X/n$.
- If $T(Y_1, \ldots, Y_n)$ is unbiased for p then average value of $T(y_1, \ldots, y_n)$ over $y_1, \ldots, y_n$ for which $\sum y_i = x$ is x/n.
- Distribution of T given $\sum Y_i = x$ does not depend on p.
- If $T(Y_1, \ldots, Y_n)$ is unbiased for p then

$$\text{Var}(T) = \text{Var}(\hat{p}) + E[(T - \hat{p})^2]$$

- $\hat{p}$ is the UMVUE of p.

This proof that $\hat{p} = X/n$ is UMVUE of p is special case of general tactic.

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.)

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

The conditional variance means

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

Square out right hand side:

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

and

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

Adding these together gives

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

Simplify remembering $E(Y|X)$ is function of X — constant when holding X fixed. So

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

taking expectations gives

$$\begin{aligned} E[(E[Y|X])^2] &= E[E[YE(Y|X)|X]] \\ &= E[YE(Y|X)] \end{aligned}$$

So 3rd term above cancels with 2nd term.

Fourth term simplifies

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

so that

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

Apply to Rao Blackwell theorem to get

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

Second term ≥ 0 so variance of $E(T|S)$ is no more than that of T; will be strictly less unless $T = E(T|S)$. This would mean that T is already a function of S. Adding the squares of the biases of T (or of $E(T|S)$) gives the inequality for MSE.

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:

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

This is simply $n\hat{p}(1-\hat{p})/(n-1)$ (can be bigger than 1/4, 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.

Finding Sufficient statistics

Binomial(n, p): log likelihood $\ell(\theta)$ (part depending on θ) is function of X alone, not of $Y_1, \ldots, Y_n$ as well.
Normal example: $\ell(\mu)$ is, ignoring terms not containing μ,

$$\ell(\mu) = \mu \sum X_i - n\mu^2/2 = n\mu\bar{X} - n\mu^2/2\,.$$

Examples of the **Factorization Criterion**:

Theorem: If the model for data X has density $f(x, \theta)$ then the statistic $S(X)$ is sufficient if and only if the density can be factored as

$$f(x, \theta) = g(S(x), \theta)h(x)$$

Proof: Find statistic $T(X)$ such that X is a one to one function of the pair S, T. Apply change of variables to the joint density of S and T. If the density factors then

$$f_{S,T}(s, t) = g(s, \theta)h(x(s, t))J(s, t)$$

where J is the jacobian, so conditional density of T given $S = s$ does not depend on θ. Thus the conditional distribution of (S, T) given S does not depend on θ and finally the conditional distribution of X given S does not depend on θ.

Conversely if S is sufficient then the $f_{T|S}$ has no θ in it so joint density of S, T is

$$f_S(s, \theta)f_{T|S}(t|s)$$

Apply change of variables formula to get

$$f_X(x) = f_S(S(x), \theta)f_{T|S}(t(x)|S(x))J(x)$$

where J is the Jacobian. This factors.

Example: If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ then the joint density is

$$(2\pi)^{-n/2}\sigma^{-n} \times$$

$$\exp\{-\sum X_i^2/(2\sigma^2) + \mu \sum X_i/\sigma^2 - n\mu^2/(2\sigma^2)\}$$

which is evidently a function of

$$\sum X_i^2, \sum X_i$$

This pair is a sufficient statistic. You can write this pair as a bijective function of $\bar{X}, \sum(X_i - \bar{X})^2$ so that this pair is also sufficient.

Example: If $Y_1, \ldots, Y_n$ are iid Bernoulli(p) then

$$\begin{aligned} f(y_1, \ldots, y_p; p) &= \prod p^{y_i}(1-p)^{1-y_i} \\ &= p^{\sum y_i}(1-p)^{n-\sum y_i} \end{aligned}$$

Define $g(x, p) = p^x(1-p)^{n-x}$ and $h \equiv 1$ to see that $X = \sum Y_i$ is sufficient by the factorization criterion.

Minimal Sufficiency

In any model $S(X) \equiv X$ is sufficient. (Apply the factorization criterion.) In any iid model the vector $X_{(1)}, \ldots, X_{(n)}$ of order statistics is sufficient. (Apply the factorization criterion.) In $N(\mu, 1)$ model we have 3 sufficient statistics:

1. $S_1 = (X_1, \ldots, X_n)$.
2. $S_2 = (X_{(1)}, \ldots, X_{(n)})$.
3. $S_3 = \bar{X}$.

Notice that I can calculate S_3 from the values of S_1 or S_2 but not vice versa and that I can calculate S_2 from S_1 but not vice versa. It turns out that $\bar{X}$ is a **minimal** sufficient statistic meaning that it is a function of any other sufficient statistic. (You can't collapse the data set any more without losing information about μ.)

Recognize minimal sufficient statistics from ℓ:

Fact: If you fix some particular θ^* then the log likelihood ratio function

$$\ell(\theta) - \ell(\theta^*)$$

is minimal sufficient. WARNING: the function is the statistic.

Subtraction of $\ell(\theta^*)$ gets rid of irrelevant constants in ℓ. In $N(\mu, 1)$ example:

$$\ell(\mu) = -n\log(2\pi)/2 - \sum X_i^2/2 + \mu \sum X_i - n\mu^2/2$$

depends on $\sum X_i^2$, not needed for sufficient statistic. Take $\mu^* = 0$ and get

$$\ell(\mu) - \ell(\mu^*) = \mu \sum X_i - n\mu^2/2$$

This function of μ is minimal sufficient. Notice: from $\sum X_i$ can compute this minimal sufficient statistic and vice versa. Thus $\sum X_i$ is also minimal sufficient.

Completeness

In Binomial(n, p) example only one function of X is unbiased. Rao Blackwell shows UMVUE, if it exists, will be a function of any sufficient statistic. Can there be more than one such function? Yes in general but no for some models like the binomial.

Definition: A statistic T is complete for a model $P_\theta; \theta \in \Theta$ if

$$E_\theta(h(T)) = 0$$

for all θ implies $h(T) = 0$.

We have already seen that X is complete in the Binomial(n, p) model. In the $N(\mu, 1)$ model suppose

$$E_\mu(h(\bar{X})) \equiv 0$$

Since $\bar{X}$ has a $N(\mu, 1/n)$ distribution we find that

$$E(h(\bar{X})) = \frac{\sqrt{n}e^{-n\mu^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(x)e^{-nx^2/2}e^{n\mu x}dx$$

so that

$$\int_{-\infty}^{\infty} h(x)e^{-nx^2/2}e^{n\mu x}dx \equiv 0$$

This is the so called Laplace transform of the function $h(x)e^{-nx^2/2}$. It is a theorem that a Laplace transform is 0 if and only if the function is 0 (because you can invert the transform). Hence $h \equiv 0$.

How to Prove Completeness

There is only one general tactic. Suppose X has density

$$f(x, \theta) = h(x)\exp\{\sum_1^p a_i(\theta)S_i(x) + c(\theta)\}$$

If the range of the function $(a_1(\theta), \ldots, a_p(\theta))$ as θ varies over Θ contains a (hyper-) rectangle in R^p then the statistic

$$(S_1(X), \ldots, S_p(X))$$

is complete and sufficient.

You prove the sufficiency by the factorization criterion and the completeness using the properties of Laplace transforms and the fact that the joint density of $S_1, \ldots, S_p$

$$g(s_1, \ldots, s_p; \theta) = h^*(s) \exp\{\sum a_k(\theta) s_k + c^*(\theta)\}$$

Example: $N(\mu, \sigma^2)$ model density has form

$$\frac{\exp\left\{\left(-\frac{1}{2\sigma^2}\right) x^2 + \left(\frac{\mu}{\sigma^2}\right) x - \frac{\mu^2}{2\sigma^2} - \log \sigma\right\}}{\sqrt{2\pi}}$$

which is an exponential family with

$$h(x) = \frac{1}{\sqrt{2\pi}}$$

$$a_1(\theta) = -\frac{1}{2\sigma^2}$$

$$S_1(x) = x^2$$

$$a_2(\theta) = \frac{\mu}{\sigma^2}$$

$$S_2(x) = x$$

and

$$c(\theta) = -\frac{\mu^2}{2\sigma^2} - \log \sigma \, .$$

It follows that

$$(\sum X_i^2, \sum X_i)$$

is a complete sufficient statistic.

Remark: The statistic $(s^2, \bar{X})$ is a one to one function of $(\sum X_i^2, \sum X_i)$ so it must be complete and sufficient, too. Any function of the latter statistic can be rewritten as a function of the former and vice versa.

FACT: A complete sufficient statistic is also minimal sufficient.

The Lehmann-Scheffé Theorem

Theorem: If S is a complete sufficient statistic for some model and $h(S)$ is an unbiased estimate of some parameter $\phi(\theta)$ then $h(S)$ is the UMVUE of $\phi(\theta)$.

Proof: Suppose T is another unbiased estimate of ϕ. According to Rao-Blackwell, T is improved by $E(T|S)$ so if $h(S)$ is not UMVUE then there must exist another function $h^*(S)$ which is unbiased and whose variance is smaller than that of $h(S)$ for some value of θ. But

$$E_\theta(h^*(S) - h(S)) \equiv 0$$

so, in fact $h^*(S) = h(S)$.

Example: In the $N(\mu, \sigma^2)$ example the random variable $(n-1)s^2/\sigma^2$ has a χ^2_{n-1} distribution. It follows that

$$E\left[\frac{\sqrt{n-1}s}{\sigma}\right] = \frac{\int_0^\infty x^{1/2} \left(\frac{x}{2}\right)^{(n-1)/2-1} e^{-x/2} dx}{2\Gamma((n-1)/2)}$$

Make the substitution $y = x/2$ and get

$$E(s) = \frac{\sigma}{\sqrt{n-1}} \frac{\sqrt{2}}{\Gamma((n-1)/2)} \int_0^\infty y^{n/2-1} e^{-y} dy$$

Hence

$$E(s) = \sigma \frac{\sqrt{2}\Gamma(n/2)}{\sqrt{n-1}\Gamma((n-1)/2)}$$

The UMVUE of σ is then

$$s\frac{\sqrt{n-1}\Gamma((n-1)/2)}{\sqrt{2}\Gamma(n/2)}$$

by the Lehmann-Scheffé theorem.

Criticism of Unbiasedness

- UMVUE can be **inadmissible for squared error loss** meaning there is a (biased, of course) estimate whose MSE is smaller for every parameter value. An example is the UMVUE of $\phi = p(1-p)$ which is $\hat\phi = n\hat p(1-\hat p)/(n-1)$. The MSE of

  $$\tilde\phi = \min(\hat\phi, 1/4)$$

  is smaller than that of $\hat\phi$.

- Unbiased estimation may be impossible.

  Binomial(n,p) log odds is $\phi = \log(p/(1-p))$. Since the expectation of any function of the data is a polynomial function of p and since ϕ is **not** a polynomial function of p there is no unbiased estimate of ϕ

- The UMVUE of σ is not the square root of the UMVUE of σ^2. This method of estimation does not have the parameterization equivariance that maximum likelihood does.

- Unbiasedness is irrelevant (unless you average together many estimators). The property is an average over possible values of the estimate in which positive errors are allowed to cancel negative errors. An exception to this criticism is that if you plan to average a number of estimators to get a single estimator then it is a problem if all the estimators have the same bias. In assignment 5 you have the one way layout example in which the mle of the residual variance averages together many biased estimates and so is very badly biased. That assignment shows that the solution is not really to insist on unbiasedness but to consider an alternative to averaging for putting the individual estimates together.