STAT 450

Lecture 21

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

- Describe the method of moments.
- Introduce the use of Mean Squared Error to compare estimators.
- Introduce unbiased estimation.

Today's notes

Other methods of estimation: Method of Moments

Basic strategy: set sample moments equal to population moments and solve for the parameters.

Definition: The r^{th} sample moment (about the origin) is

$$\frac{1}{n}\sum_{i=1}^{n} X_i^r$$

The r^{th} population moment is

$$\mathrm{E}(X^r)$$

Central moments are

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r$$

and

$$\mathrm{E}\left[(X-\mu)^r\right] .$$

If we have p parameters we can estimate the parameters $\theta_1, \ldots, \theta_p$ by solving the system of p equations:

$$\mu_1 = \bar{X}$$

$$\mu_2' = \overline{X^2}$$

and so on to

$$\mu_p' = \overline{X^p}$$

You need to remember that the population moments μ_k' will be formulas involving the parameters.

Gamma Example

The Gamma(α, β) density is

$$f(x; \alpha, \beta) = \frac{1}{\beta\Gamma(\alpha)} \left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left[-\frac{x}{\beta}\right] 1(x > 0)$$

and has

$$\mu_1 = \alpha\beta$$

and

$$\mu_2' = \alpha(\alpha+1)\beta^2$$

This gives the equations

$$\begin{aligned} \alpha\beta &= \overline{X} \\ \alpha(\alpha+1)\beta^2 &= \overline{X^2} \end{aligned}$$

Subtract the square of the first from the second to get

$$\alpha\beta^2 = \overline{X^2} - (\overline{X})^2$$

Divide this by the first equation to find the method of moments estimate of β is

$$\tilde{\beta} = \frac{\overline{X^2} - (\overline{X})^2}{\overline{X}}$$

Then from the first equation get

$$\tilde{\alpha} = \overline{X}/\tilde{\beta} = \frac{(\overline{X})^2}{\overline{X^2} - (\overline{X})^2}$$

The equations are much easier to solve than the likelihood equations which involve the function

$$\psi(\alpha) = \frac{d}{d\alpha} \log(\Gamma(\alpha))$$

called the digamma function. The score function in this problem has components

$$U_\beta = \frac{\sum X_i}{\beta^2} - n\alpha/\beta$$

and

$$U_\alpha = -n\psi(\alpha) + \sum \log(X_i) - n\log(\beta)$$

You can solve for β in terms of α to leave you trying to find a root of the equation

$$-n\psi(\alpha) + \sum \log(X_i) - n\log(\sum X_i/(n\alpha)) = 0$$

To use Newton Raphson on this you begin with the preliminary estimate $\hat\alpha_1 = \tilde\alpha$ and then compute iteratively

$$\hat\alpha_{k+1} = \frac{\overline{\log(X)} - \psi(\hat\alpha_k) - \log(\overline{X}/\hat\alpha_k)}{1/\alpha - \psi'(\hat\alpha_k)}$$

until the sequence converges. Computation of ψ', the trigamma function, requires special software. Web sites like *netlib* and *statlib* are good sources for this sort of thing.

Optimality theory for point estimates

Why bother doing the Newton Raphson steps? Why not just use the method of moments estimates? The answer is that the method of moments estimates are not usually as close to the right answer as the mles.

Rough principle: A good estimate $\hat\theta$ of θ is usually close to θ_0 if θ_0 is the true value of θ. Closer estimates, more often, are better estimates.

This principle must be quantified if we are to "prove" that the mle is a good estimate. In the Neyman Pearson spirit we measure average closeness.

Definition: The **Mean Squared Error** (MSE) of an estimator $\hat\theta$ is the **function**

$$MSE(\theta) = E_\theta[(\hat\theta - \theta)^2]$$

Standard identity:

$$MSE = \text{Var}_\theta(\hat\theta) + Bias^2_{\hat\theta}(\theta)$$

where the bias is defined as

$$Bias_{\hat\theta}(\theta) = E_\theta(\hat\theta) - \theta\,.$$

Primitive example: I take a coin from my pocket and toss it 6 times. I get $HTHTTT$. The MLE of the probability of heads is

$$\hat{p} = X/n$$

where X is the number of heads. In this case I get $\hat{p} = \frac{1}{3}$.

An alternative estimate is $\tilde{p} = \frac{1}{2}$. That is, $\tilde{p}$ ignores the data and guesses the coin is fair. The MSEs of these two estimators are

$$MSE_{\text{MLE}} = \frac{p(1-p)}{6}$$

and

$$MSE_{0.5} = (p - 0.5)^2$$

If p is between 0.311 and 0.689 then the second MSE is smaller than the first. For this reason I would recommend use of $\tilde{p}$ for sample sizes this small.

Now suppose I did the same experiment with a thumbtack. The tack can land point up (U) or tipped over (O). If I get $UOUOOO$ how should I estimate p the probability of U? The mathematics is identical to the above but it seems clear that there is less reason to think $\tilde{p}$ is better than $\hat{p}$ since there is less reason to believe $0.311 \le p \le 0.689$ than with a coin.

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 we already know that the score has mean 0 we see that

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

Now remember that correlations are between -1 and 1 or

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

Squaring gives the inequality

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

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 some non-random constants A and B (which might 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 $\mathrm{Var}_\theta(T(X)) \equiv 1/I(\theta)$ then $T(X)$ is the UMVUE. For instance in the $N(\mu, 1)$ example the Fisher information is n and $\mathrm{Var}(\overline{X}) = 1/n$ 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 function.
- Densities of the exponential form given above are somehow special. (The form is called an exponential family.)

- For most problems the inequality will be strict. It is strict unless The score is an affine function of a statistic T and T (possibility divided by some constant which doesn't depend on θ) 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

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

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

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

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

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

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

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

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.