# STAT 450

## Lecture 17

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

- Start large sample theory for the likelihood function

**Today's notes**

**Large Sample Theory**

Our goal is to study the behaviour of $\hat\theta$ and to develop approximate distribution theory for $n^{1/2}(\hat\theta-\theta)$. Here is a summary of the conclusions of the theory:

1. The log likelihood is probably bigger at $\theta_0$, the true value of $\theta$ than at any single $\theta_1 \neq \theta_0$.
2. The log-likelihood is approximately quadratic in a small region around $\theta_0$;
3. This will prove that there is probably a local maximum of $\ell$ near $\theta_0$ and that this local maximum is unique.
4. The value of the score function at $\theta_0$ is approximately normal.
5. The second derivative of the log likelihood is nearly constant for $\theta$ fairly near $\theta_0$;
6. The likelihood equations can be approximated by Taylor expansion.
7. The error in the MLE $\hat\theta$ is approximately a multiple of the score function at $\theta_0$.
8. The MLE is approximately normal.
9. The approximate standard deviation of the MLE can be estimated.

We now study the approximate behaviour of $\hat\theta$ by studying the function $U$. Notice first that $U$ is a sum of independent random variables.

**Theorem**: If $Y_1, Y_2, \ldots$ are iid with mean $\mu$ then

$$\frac{\sum Y_i}{n} \to \mu$$

This is called the law of large numbers. The strong law says

$$P(\lim \frac{\sum Y_i}{n} = \mu) = 1$$

and the weak law that

$$\lim P(|\frac{\sum Y_i}{n} - \mu| > \epsilon) = 0$$

For iid $Y_i$ the stronger conclusion holds but for our heuristics we will ignore the differences between these notions.

Now suppose that $\theta_0$ is the true value of $\theta$. Then

$$U(\theta)/n \to \mu(\theta)$$

where

$$\begin{aligned}\mu(\theta) =& E_{\theta_0}\left[\frac{\partial \log f}{\partial \theta}(X_i,\theta)\right] \\ =& \int \frac{\partial \log f}{\partial \theta}(x,\theta) f(x,\theta_0) dx\end{aligned}$$

Consider as an example the case of $N(\mu,1)$ data where

$$U(\mu)/n = \sum (X_i - \mu)/n = \bar{X} - \mu$$

If the true mean is $\mu_0$ then $\bar{X} \to \mu_0$ and

$$U(\mu)/n \to \mu_0 - \mu$$

If we think of a $\mu < \mu_0$ we see that the derivative of $\ell(\mu)$ is likely to be positive so that $\ell$ increases as we increase $\mu$. For $\mu$ more than $\mu_0$ the derivative is probably negative and so $\ell$ tends to be decreasing for $\mu > 0$. It follows that $\ell$ is likely to be maximized close to $\mu_0$.

Now we repeat these ideas for a more general case. We study the random variable $\log[f(X_i,\theta)/f(X_i,\theta_0)]$. You know the inequality

$$E(X)^2 \le E(X^2)$$

(because the difference is $Var(X) \ge 0$. This inequality has the following generalization, called Jensen's inequality. If $g$ is a convex function (non-negative second derivative, roughly) then

$$g(E(x)) \le E(g(X))$$

The inequality above has $g(x) = x^2$. We use $g(x) = -\log(x)$ which is convex because $g''(x) = x^{-2} > 0$. We get

$$-\log(E_{\theta_0}[f(X_i,\theta)/f(X_i,\theta_0)] \le E_{\theta_0}[-\log\{f(X_i,\theta)/f(X_i,\theta_0)\}]$$

But

$$\begin{aligned}E_{\theta_0}[f(X_i,\theta)/f(X_i,\theta_0)] &= \int \frac{f(x,\theta)}{f(x,\theta_0)} f(x,\theta_0) dx \\ &= \int f(x,\theta) dx \\ &= 1\end{aligned}$$

We can reassemble the inequality and this calculation to get

$$E_{\theta_0}[\log\{f(X_i,\theta)/f(X_i,\theta_0)\}] \le 0$$

It is possible to prove that the inequality is strict unless the $\theta$ and $\theta_0$ densities are actually the same. Let $\mu(\theta) < 0$ be this expected value. Then for each $\theta$ we find

$$\begin{aligned} n^{-1}[\ell(\theta) - \ell(\theta_0)] &= n^{-1}\sum \log[f(X_i,\theta)/f(X_i,\theta_0)] \\ &\to \mu(\theta) \end{aligned}$$

This proves that the likelihood is probably higher at $\theta_0$ than at any other single $\theta$. This idea can often be stretched to prove that the mle is **consistent**.

**Definition** A sequence $\hat\theta_n$ of estimators of $\theta$ is consistent if $\hat\theta_n$ converges weakly (or strongly) to $\theta$.

**Proto theorem**: In regular problems the mle $\hat\theta$ is consistent.

Now let us study the shape of the log likelihood near the true value of $\hat\theta$ under the assumption that $\hat\theta$ is a root of the likelihood equations close to $\theta_0$. We use Taylor expansion to write, for a 1 dimensional parameter $\theta$,

$$\begin{aligned} U(\hat\theta) &= 0 \\ &= U(\theta_0) + U'(\theta_0)(\hat\theta-\theta_0) + U''(\tilde\theta)(\hat\theta-\theta_0)^2/2 \end{aligned}$$

for some $\tilde\theta$ between $\theta_0$ and $\hat\theta$. (This form of the remainder in Taylor's theorem is not valid for multivariate $\theta$.) The derivatives of $U$ are each sums of $n$ terms and so should be both proportional to $n$ in size. The second derivative is multiplied by the square of the small number $\hat\theta-\theta_0$ so should be negligible compared to the first derivative term. If we ignore the second derivative term we get

$$-U'(\theta_0)(\hat\theta-\theta_0) \approx U(\theta_0)$$

Now let's look at the terms $U$ and $U'$.

In the normal case

$$U(\theta_0) = \sum(X_i - \mu_0)$$

has a normal distribution with mean 0 and variance $n$ (SD $\sqrt{n}$). The derivative is simply

$$U'(\mu) = -n$$

and the next derivative $U''$ is 0. We will analyze the general case by noticing that both $U$ and $U'$ are sums of iid random variables. Let

$$U_i = \frac{\partial \log f}{\partial\theta}(X_i,\theta_0)$$

and

$$V_i = -\frac{\partial^2 \log f}{\partial\theta^2}(X_i,\theta)$$

In general, $U(\theta_0) = \sum U_i$ has mean 0 and approximately a normal distribution. Here is how we check that:

$$
\begin{aligned}
E_{\theta_0}(U(\theta_0)) &= nE_{\theta_0}(U_1) \\
&= n\int \frac{\partial \log(f(x,\theta))}{\partial\theta}(x,\theta_0) f(x,\theta_0)dx \\
&= n\int \frac{\partial f/\partial\theta(x,\theta_0)}{f(x,\theta_0)}\theta f(x,\theta_0)dx \\
&= n\int \frac{\partial f}{\partial\theta}(x,\theta_0)dx \\
&= n\frac{\partial}{\partial\theta}\int f(x,\theta)dx\Big|_{\theta=\theta_0} \\
&= n\frac{\partial}{\partial\theta}1 \\
&= 0
\end{aligned}
$$

Notice that I have interchanged the order of differentiation and integration at one point. This step is usually justified by applying the dominated convergence theorem to the definition of the derivative. The same tactic can be applied by differentiating the identity which we just proved

$$\int \frac{\partial \log f}{\partial\theta}(x,\theta)f(x,\theta)dx = 0$$

Taking the derivative of both sides with respect to $\theta$ and pulling the derivative under the integral sign again gives

$$\int \frac{\partial}{\partial\theta}\left[\frac{\partial \log f}{\partial\theta}(x,\theta)f(x,\theta)\right]dx = 0$$

Do the derivative and get

$$
\begin{aligned}
-\int \frac{\partial^2 \log(f)}{\partial\theta^2}f(x,\theta)dx &= \int \frac{\partial \log f}{\partial\theta}(x,\theta)\frac{\partial f}{\partial\theta}(x,\theta)dx \\
&= \int \left[\frac{\partial \log f}{\partial\theta}(x,\theta)\right]^2 f(x,\theta)dx
\end{aligned}
$$

**Definition**: The **Fisher Information** is

$$I(\theta) = -E_\theta(U'(\theta)) = nE_{\theta_0}(V_1)$$

We refer to $\mathcal{I}(\theta_0) = E_{\theta_0}(V_1)$ as the information in 1 observation.

The idea is that $I$ is a measure of how curved the log likelihood tends to be at the true value of $\theta$. Big curvature means precise estimates. Our identity above is

$$I(\theta) = Var_\theta(U(\theta)) = n\mathcal{I}(\theta)$$

Now we return to our Taylor expansion approximation

$$-U'(\theta_0)(\hat\theta - \theta_0) \approx U(\theta_0)$$

and study the two appearances of $U$.

We have shown that $U = \sum U_i$ is a sum of iid mean 0 random variables. The central limit theorem thus proves that

$$n^{-1/2}U(\theta) \Rightarrow N(0, \sigma^2)$$

where $\sigma^2 = Var(U_i) = E(V_i) = \mathcal{I}(\theta)$.

Next observe that

$$-U'(\theta) = \sum V_i$$

where again

$$V_i = -\frac{\partial U_i}{\partial \theta}$$

The law of large numbers can be applied to show

$$-U'(\theta_0)/n \to E_{\theta_0}[V_1] = \mathcal{I}(\theta_0)$$

Now manipulate our Taylor expansion as follows

$$n^{1/2}(\hat\theta - \theta_0) \approx \left[\frac{\sum V_i}{n}\right]^{-1} \frac{\sum U_i}{\sqrt{n}}$$

Apply Slutsky's Theorem to conclude that the right hand side of this converges in distribution to $N(0, \sigma^2/\mathcal{I}(\theta)^2)$ which simplifies, because of the identities, to $N(0, 1/\mathcal{I}(\theta))$.

**Summary**

In regular families:

- Under strong regularity conditions Jensen's inequality can be used to demonstrate that $\hat\theta$ which maximizes $\ell$ globally is consistent and that this $\hat\theta$ is a root of the likelihood equations.

- It is generally easier to study $\ell$ only close to $\theta_0$. For instance define $A$ to be the event that $\ell$ is concave on the set of $\theta$ such that $|\theta - \theta_0| < \delta$ and the likelihood equations have a unique root in that set. Under weaker conditions than the previous case we can prove that there is a $\delta > 0$ such that

  $$P(A) \to 1$$

  In that case we can prove that the root $\hat\theta$ of the likelihood equations mentioned in the definition of $A$ is consistent.

- Sometimes we can only get an even weaker conclusion. Define $B$ to be the event that $\ell(\theta)$ is concave for $n^{1/2}|\theta - \theta_0| < L$ and there is a unique root of $\ell$ over this range. Again this root is consistent but there might be other consistent roots of the likelihood equations.

- Under any of these scenarios there is a consistent root of the likelihood equations which is definitely the closest to the true value $\theta_0$. This root $\hat{\theta}$ has the property

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, 1/\mathcal{I}(\theta)) \,.$$

We usually simply say that the mle is consistent and asymptotically normal with an asymptotic variance which is the inverse of the Fisher information. This assertion is actually valid for vector valued $\theta$ where now $I$ is a matrix with $ij$th entry

$$I_ij = -E\left(\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}\right)$$