Lecture Notes 6

The *likelihood function* plays an important role in statistical inference. In these notes we define the likelihood function. This function will be used later for many different tasks.

1 The Likelihood Function

Definition. Let $X^n = (X_1, \cdots, X_n)$ have joint density $p(x^n; \theta) = p(x_1, \ldots, x_n; \theta)$ where $\theta \in \Theta$. The **likelihood function** $L : \Theta \to [0, \infty)$ is defined by

$$L(\theta) \equiv L(\theta; x^n) = p(x^n; \theta)$$

where x^n is fixed and θ varies in Θ. The log-likelihood function is

$$\ell(\theta) = \log L(\theta).$$

1. The likelihood function is a function of θ.
2. The likelihood function is **not** a probability density function.
3. If the data are iid then the likelihood is

$$L(\theta) = \prod_{i=1}^{n} p(x_i; \theta) \qquad \text{iid case only.}$$

4. The likelihood is only defined up to a constant of proportionality. In other words, it is an equivalence class of functions.
5. The likelihood function is used (i) to generate estimators (the maximum likelihood estimator) and (ii) as a key ingredient in Bayesian inference.

Example 1 *Suppose that* $X = (X_1, X_2, X_3) \sim \text{Multinomial}(n, p)$ *where*

$$p = (p_1, p_2, p_3) = (\theta, \theta, 1 - 2\theta).$$

So

$$p(x; \theta) = \binom{n}{x_1\ x_2\ x_3} p_1^{x_1} p_2^{x_2} p_3^{x_3} = \theta^{x_1 + x_2} (1 - 2\theta)^{x_3}.$$

Suppose that $X = (1, 3, 2)$. *Then*

$$L(\theta) = \frac{6!}{1!\ 3!\ 2!} \theta^1 \theta^3 (1 - 2\theta)^2 \propto \theta^4 (1 - 2\theta)^2.$$

Now suppose that $X = (2, 2, 2)$. *Then*

$$L(\theta) = \frac{6!}{2!\ 2!\ 2!} \theta^2 \theta^2 (1 - 2\theta)^2 \propto \theta^4 (1 - 2\theta)^2.$$

Hence, the likelihood function is the same for these two datasets.

Example 2 *$X_1, \cdots, X_n \sim N(\mu, 1)$. Then,*

$$L(\mu) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \propto \exp\left\{-\frac{n}{2}(\overline{x} - \mu)^2\right\}.$$

Example 3 *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Then*

$$L(p) \propto p^X (1-p)^{n-X}$$

for $p \in [0, 1]$ where $X = \sum_i X_i$.

Theorem 4 *Write $x^n \sim y^n$ if $L(\theta|x^n) \propto L(\theta|y^n)$. The partition induced by $\sim$ is the minimal sufficient partition.*

Proof. Homework. ■

Example 5 *Let $X_1, \ldots, X_n \sim \text{Unif}(0, \theta)$. Then*

$$L(\theta) \propto \left(\frac{1}{\theta}\right)^n \; I(\theta \geq X_{(n)}).$$

Example 6 *Let $X_1, \ldots, X_n \sim N(\mu, \Sigma)$. Then*

$$L(\mu, \Sigma) \propto |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_i (X_i - \mu)^T \Sigma^{-1}(X_i - \mu)\right)$$

and so

$$\ell(\mu, \Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_i (X_i - \mu)^T \Sigma^{-1}(X_i - \mu).$$

Let $\text{tr}(A)$ denote the trace of a matrix A (the sum of the diagonal entries). Then,

$$\begin{aligned}
\sum_i (X_i - \mu)^T \Sigma^{-1}(X_i - \mu) &= \sum_i (X_i - \overline{X})^T \Sigma^{-1}(X_i - \overline{X}) + n(\overline{X} - \mu)^T \Sigma^{-1}(\overline{X} - \mu) \\
&= \sum_i \text{tr}[(X_i - \overline{X})^T \Sigma^{-1}(X_i - \overline{X})] + n(\overline{X} - \mu)^T \Sigma^{-1}(\overline{X} - \mu) \\
&= \sum_i \text{tr}[\Sigma^{-1}(X_i - \overline{X})(X_i - \overline{X})^T] + n(\overline{X} - \mu)^T \Sigma^{-1}(\overline{X} - \mu) \\
&= \text{tr}\left(\sum_i [\Sigma^{-1}(X_i - \overline{X})(X_i - \overline{X})^T]\right) + n(\overline{X} - \mu)^T \Sigma^{-1}(\overline{X} - \mu) \\
&= \text{tr}\left(\Sigma^{-1}\sum_i [(X_i - \overline{X})(X_i - \overline{X})^T]\right) + n(\overline{X} - \mu)^T \Sigma^{-1}(\overline{X} - \mu) \\
&= n\text{tr}(\Sigma^{-1} S) + n(\overline{X} - \mu)^T \Sigma^{-1}(\overline{X} - \mu)
\end{aligned}$$

and so

$$\ell(\mu, \Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{n}{2}\text{tr}(\Sigma^{-1}S) - \frac{n}{2}(\overline{X} - \mu)^T\Sigma^{-1}(\overline{X} - \mu).$$

2 Likelihood, Sufficiency and the Likelihood Principle

The likelihood function is a minimal sufficient statistic. That is, if we define the equivalence relation: $x^n \sim y^n$ when $L(\theta; x^n) \propto L(\theta; y^n)$ then the resulting partition is minimal sufficient.

Does this mean that the likelihood function contains all the relevant information? Some people say yes it does. This is sometimes called *the likelihood principle.* That is, the likelihood principle says that the likelihood function contains all the infomation in the data.

This is FALSE. Here is a simple example to illustrate why. Let $\mathcal{C} = \{c_1, \ldots, c_N\}$ be a finite set of constants. For simplicity, asssume that $c_j \in \{0, 1\}$ (although this is not important). Let $\theta = \frac{1}{N}\sum_{j=1}^N c_j$. Suppose we want to estimate θ. We proceed as follows. Let $S_1, \ldots, S_n \sim \text{Bernoulli}(\pi)$ where π is known. If $S_i = 1$ you get to see c_i. Otherwise, you do not. (This is an example of survey sampling.) The likelihood function is

$$\prod_i \pi^{S_i}(1-\pi)^{1-S_i}.$$

The unknown parameter does not appear in the likelihood. In fact, there are no unknown parameters in the likelihood! The likelihood function contains no information at all.

But we can estimate θ. Let

$$\widehat{\theta} = \frac{1}{N\pi}\sum_{j=1}^N c_j S_j.$$

Then $\mathbb{E}(\widehat{\theta}) = \theta$. Hoeffding's inequality implies that

$$\mathbb{P}(|\widehat{\theta} - \theta| > \epsilon) \leq 2e^{-2n\epsilon^2\pi^2}.$$

Hence, $\widehat{\theta}$ is close to θ with high probability.

Summary: the minimal sufficient statistic has all the information you need to compute the likelihood. But that does not mean that all the information is in the likelihood.