M4056 Lecture Notes. **Monday, September 27, 2010**

General orientation. We have decided to model reality using a probability distribution from a parametrized family $f_X(x|\theta)$, but we do not know the the value of θ that best describes reality. We wish to use the information in a random sample to select it.

The **likelihood function** is defined as follows:

$$L(\theta|\vec{x}) := f_{\vec{X}}(\vec{x}|\theta).$$

The likelihood function is the *pdf* (or *pmf*) of the sample, viewed not a function of the sample value, but as a function of the parameter.

Example 1. Suppose X is Bernoulli(p), where p (the probability of success) is unknown. Consider the general sample $\vec{X}$ of n elements, and let $\vec{x}$ be a specific observation. Then

$$L(p|\vec{x}) = f_{\vec{X}}(\vec{x}|p) = p^t(1-p)^{n-t}, \quad \text{where } t := \textstyle\sum_{i=1}^n x_i.$$

We see that $L(p) = L(p|\vec{x})$ is defined and non-negative on $[0,1]$, with $L(0) = 0 = L(1)$. We have

$$\frac{dL}{dp} = tp^{t-1}(1-p)^{n-t} - p^t(n-t)(1-p)^{n-t-1} = p^{t-1}(1-p)^{n-t-1}(t-pn),$$

so L has a critical point at $p = t/n$, and this is a maximum. Note that integration by parts ($u = p^t$, $dv = (1-p)^{n-t}$) and induction gives:

$$\int_0^1 p^t(1-p)^{n-t}dp = \frac{1}{(n+1)\binom{n}{t}}.$$

Since this is not 1 (unless $n = 0$), $L(\theta, \vec{x})$ is not a *pdf* as a function of θ.

Example 2. Suppose X is geometric(p), where p (the probability of success) is unknown. By definition, $P(X = x)$ is the probability that the first success is obtained on the x^{th} attempt. Since $f_X(x|p)$ is the probability of $x-1$ failures followed by a success, we have:

$$f_X(x|p) = (1-p)^{x-1}p\,, \quad x = 1, 2, 3, \ldots$$

Suppose $\vec{x}$ is a specific instance of an m-element sample. Then

$$L(p|\vec{x}) = \prod_{i=1}^m (1-p)^{x_i-1}p = p^m(1-p)^{w-m}, \quad \text{where } w := \textstyle\sum_{i=1}^m x_i.$$

The computation in the last example (with $t = m$ and $n = w$) shows that the maximum is at $p = m/w$.

Example 3. We extend Example 2, by allowing the recorded data to be either of the form $X = x$ or $X > x$. Note that $P(X > x) = (1-p)^x$. For definiteness, assume the first m observations are of the first kind, and observations $(m+1)$ through $m+k$ are of the second kind.

$$\begin{aligned} L(p|\vec{X}) &= \prod_{i=1}^{m}(1-p)^{x_i-1}p \cdot \prod_{i=m+1}^{m+k}(1-p)^{x_i} \\ &= p^m(1-p)^{w-m}, \quad \text{where } w := \textstyle\sum_{i=1}^{m+k} x_i. \end{aligned}$$

Once again, the maximum is at $p = m/w$.

Maximum Likelihood Estimators. One of the most widely used strategies to choose the value of the parameter is to let it have the value that makes the observed data most likely. This is denoted by the acronym "MLE".

Example 1, continued. Suppose that you have a coin that you know to be biased, and you wish to estimate the probability p of it landing on heads. You flip the coin 1000 times and observe 413 heads. The computation above shows that the likelihood of this result is maximized when $p = 0.413$.

Example 4. The main idea may be seen even in the simple example we gave at the beginning of the semester. Assume that red and green beads are available, and the Dealer places 19 beads of one color in 19 boxes, and 1 of the other color in a 20th box. You choose a box at random and examine the contents, and you then are allowed to place a bet about the dominant color. To see how this fits in the framework we have established, let X be the color of the bead you observe. The *pmf* of X depends on the unknown parameter $\theta =$ the dominant color, and it can be written as follows:

$$f_X(x|\theta) = \begin{cases} 19/20, & \text{if } x = \theta \\ 1/20, & \text{if } x \neq \theta \end{cases}.$$

Viewed as a function of θ, this is the likelihood, and clearly the likelihood is maximized when θ is the color observed.

Example 3, continued. We apply the formula for the likelihood function for the geometric to the data listed in the handout. For smokers, $m = 93$, $k = 7$, and $w = (29)(1) + (16)(2) + \cdots + (3)(12) + (7)(12) = 415$, so the MLE is $93/415 \approx 0.2241$. For nonsmokers, $m = 474$, $k = 12$ and $w = 1429$, and the MLE is $474/1429 \approx 0.3317$.

The intelligent statistician knows that the maximum likelihood estimate is not necessarily the value that will be optimal for achieving a concrete goal. For example, the state of Louisiana has recently passed legislation that mandates that public school teachers will be evaluated based on the test data of their students. In order to do this, it seems that state plans to create a statistical model of student performance. If a teacher's students perform substantially above the model's predictions, the teacher will receive a reward; if they perform substantially below, the teacher is will receive something else. Choosing to set the model parameters at the MLE may or may not produce the best policy from an economic, social or moral perspective.