Section 6, 10/8/19

Parameter Estimations: Maximum Likelihood Estimation

Suppose that a set of random variable $X_1, \ldots, X_n$ are distributed according to a joint density function $f(\cdot|\theta)$ dependent on some set of parameter θ. Then for a given set of observations $x_1, \ldots, x_n$, we write the density at these observations as $f(x_1, \ldots, x_n|\theta)$. The likelihood of the observations given some θ is exactly $lik(\theta) = f(x_1, \ldots, x_n|\theta)$. The **maximum likelihood estimate** (**mle**) of θ is the value for which $lik(\theta)$ is maximized. Intuitively this is the value for which the data observed is most likely to have been observed.

MLEs of Multinomial Cell Probabilities

What follows is an example of the application of the MLE to data from a given distribution. Consider the multinomial distribution, a generalization of the binomial distribution where, given a set of probabilities $p_1, \ldots, p_m$ that sum to les 1, the probability p_i determines the probability of an observation being in "cell" i. The random variables $X_1, \ldots, X_m$ are the number of counts in cells $1, \ldots, m$ for a total count of n. Given a set of observed counts $x_1, \ldots, x_m$, we wish to estimate the unknown probabilities. The distribution is defined as follows:

$$f(x_1, \ldots, x_m|p_1, \ldots, p_m) = \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i}$$

We seek the p_i that maximize this quantity. It is usually easier to maximize the log likelihood instead.

$$l(p_1, \ldots, p_m) = log(n!) - \sum_{i=1}^m log(x_i!) + \sum_{i=1}^m x_i log(p_i)$$

This is subject to $\sum p_i = 1$. So we use Lagrange Multipliers. Recall that if we wish to optimize some function $f(x)$ subject to the constrain $g(x) = 0$, then we can equivalently optimize the Lagrangian function $\mathcal{L}(x, \lambda) = f(x) + \lambda g(x)$. Thus, we can maximize

$$\mathcal{L}(p_1, \ldots, p_m, \lambda) = log(n!) - \sum_{i=1}^m log(x_i!) + \sum_{i=1}^m x_i log(p_i) + \lambda\big(\sum_{i=1}^m p_i - 1\big)$$

We do so by taking partial derivatives with respect to the p_i and setting the result to 0. This yields, by symmetry, $\hat{p}_i = -\frac{x_i}{\lambda}, \quad i = 1, \ldots, m$. Since the the sum of these estimates must be 1, $1 = \frac{-n}{\lambda}$ and so $\lambda = -n$. Therefore $\hat{p}_i = \frac{x_i}{n}, \quad i = 1, \ldots, m$. This is the sample proportion in each class, an intuitive answer.

Example A: Hardy-Weinberg Equilibrium

See textbook page 274

Fisher Information

The results will be proven in class, but here we simply define terms and attempt to build an intuition.

Given some function with density $f(X|\theta_0)$ parameterized by some true θ_0 which we estimate with our mle $\hat{\theta}$. It can be shown that under *reasonable conditions* of smoothness, $\hat{\theta}$ converges to θ_0 as our sample size n goes to infinity.

Define the score function of the likelihood as

$$\frac{\partial}{\partial\theta} log\, f(X|\theta)$$

And the **Fisher Information** as

$$I(\theta) = E\big[\frac{\partial}{\partial\theta} log\, f(X|\theta)\big]^2$$

Back to what we said above, specifically $\hat{\theta}$ converges to θ_0 in that the distribution of $\hat{\theta}$ asymptotically converges to a normal distribution with mean θ_0 and variance $1/[nI(\theta_0)]$. We say that $\hat{\theta}$ is *asymptotically unbiased*.

By Theorem B on page 227, under smoothness condition on f, the distribution of $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ approaches the standard normal distribution $N(0, 1)$. A large Fisher Information quantity implies a strong confidence of our estimated parameter $\hat{\theta}$. One of the most important regularization conditions is that the support of the distribution is independent of the parameter choice θ. The support is the set of values x where $f(x) > 0$. For instance, the uniform distribution $unif(0, \theta)$ does **not** satisfy this condition.