

Statistical Prerequisite

A crash course on Maximum Likelihood theory

Assume X_1, \dots, X_k are independent identically distributed random variables with density functions $f_{X_i}(\cdot; \theta)$. Note that θ can be a vector. Assume furthermore that we have observed the data x_1, \dots, x_n . The likelihood function is defined by

$$L(\theta; x_1, \dots, x_k) = \prod_i f_{X_i}(x_i; \theta)$$

The likelihood is thus the probability of the parameter θ given the observed data. An estimate of θ is denoted $\hat{\theta}$.

The Maximum Likelihood Estimate (MLE) is the estimate $\hat{\theta}$ that maximises $L(\theta; x_1, \dots, x_k)$:

$$L(\hat{\theta}; x_1, \dots, x_k) \geq L(\tilde{\theta}; x_1, \dots, x_k) \quad \forall \tilde{\theta} \in \Theta$$

For the purpose of finding the MLE it is often more convenient to work with a sum rather than a product. This is achieved by considering the log-likelihood function

Example

Let X_1, \dots, X_k be independent binomial variables:

$$f_i(x_i; n_i, p) = c_i \cdot p^{x_i} \cdot (1-p)^{n_i - x_i}$$

The log-likelihood function is therefore given by

$$\begin{aligned} l(p; x_1, \dots, x_k) &= \sum \log(f_i(x_i; n_i, p)) \\ &= \sum_i c_i + \sum x_i \cdot \log(p) + (n_i - x_i) \cdot \log(1-p) \end{aligned}$$

The first term is a constant and can be ignored. By usual optimisation it is easily shown that the MLE is

$$\hat{p} = \frac{\sum_i x_i}{\sum_i n_i} = \frac{x_+}{n_+}$$

Maximum likelihood estimation owes its significance to a number of attractive properties of the MLE. The most important one under certain regularity conditions are (in non-technical terms)

- Sufficiency: The estimator contains all information in the data about the parameter
- Consistency: The estimator has the smallest variance
- Efficiency: The estimator tends to the true value as the sample-size increases
- Asymptotically normal: $\hat{\theta} - \theta \approx N(0, i(\theta)^{-1})$

Where $i(\theta) = -E\left\{\frac{\partial^2 l}{\partial \theta^2}\right\}$ is the Fisher information.

Efron and Hinkley (1978) argue that the *observed information* $I(\theta) = -\left\{\frac{\partial^2 l}{\partial \theta^2}\right\}_{\theta=\hat{\theta}}$ should be used instead, since it, in some sense, is closer to the data. This appears to be commonly adopted by many statisticians.

Example (continued):

The second derivative of the log-likelihood function is given by

$$\begin{aligned} \frac{d^2}{dp^2} l(p; x_1, \dots, x_k) &= \sum_i \frac{d^2}{dp^2} (x_i \cdot \log(p) + (n_i - x_i) \cdot \log(1-p)) \\ &= -\sum_i \frac{x_i}{p^2} + \frac{n_i - x_i}{(1-p)^2} \end{aligned}$$

Since $E_p(X_i) = n_i \cdot p$ it follows that

$$\begin{aligned} i(p) &= -E_p\left(\frac{d^2 l}{dp^2}\right) \\ &= \sum_i \frac{n_i \cdot p}{p^2} + \frac{n_i \cdot (1-p)}{(1-p)^2} \\ &= \left(\frac{1}{p} + \frac{1}{1-p}\right) \cdot \sum_i n_i \\ &= \frac{n_+}{p \cdot (1-p)} \end{aligned}$$

The variance of the estimate is therefore $i(p)^{-1} = \frac{p \cdot (1-p)}{n_+}$.

An estimate of this variance is obtained by plugging in the estimate $\hat{p} = \frac{x_+}{n_+}$. This gives

$$i(\hat{p})^{-1} = \frac{x_+ \cdot (n_+ - x_+)}{n_+^2}$$

Further reading: Lehmann (1983) gives a very thorough but also very technical presentation of the subject. McCullagh & Nelder (1989) has a small annex on the subject. This provides an excellent excerpt of the most important aspects of maximum likelihood theory.

Below is given some theorems that are essential to the derivation of models in selectivity analysis. These are primarily within the field of probability theory.

Theorem 1.

If $N \sim Po(\mu)$ and $a \in R_+$ then $a \cdot N \sim Po(a \cdot \mu)$

Theorem 2.

Assume N_1, \dots, N_k are k independent Poisson distributed variables:

$$N_i \sim Po(\mu_i)$$

and set $N_+ = N_1 + N_2 + \dots + N_k$

The conditional distribution of (N_1, \dots, N_k) given $N_+ = n_+$ is multinomial:

$$(N_1, \dots, N_k | N_+ = n_+) \sim m(\varphi_1, \dots, \varphi_k; n_+)$$

where $\varphi_i = \frac{\mu_i}{\sum_j \mu_j}$

Theorem 3.

Assume X_1, \dots, X_k are k independent random variables with $X_i \sim N(0;1)$.

Then $\sum_i X_i^2 \sim \chi(k)$