4 Maximum Likelihood Estimation

In previous sections, mathematical descriptions of random variation were used to model the procedure of taking a random sample from a population, and to develop methods for making inferences about populations from samples. Those inferences involved finding confidence intervals for (or testing hypotheses about) parameters of the model used to describe the population. This section will look at a method for estimating the parameters of a model using data. The aim is to find the most likely value for a parameter, given some data which relate to it.

As an example, some unknown proportion of the population supports the Conservative party. This proportion can be viewed as a parameter p_c, which we would like to estimate. The obvious way to estimate it is to take a random sample from the population and to determine what proportion $\hat{p}_c$ vote Conservative. This proportion can be used as an estimate of the population parameter p_c. In this example, the model for the population is extremely simple ('A proportion p_c of the population vote Conservative') and it is obvious how to obtain a suitable parameter estimate. In other cases it is not so obvious how to estimate parameters, and a general method is needed. Sticking with the political example, we might believe that the probability of being a Conservative voter depends on income in a linear way, so that $P(\text{Votes Conservative}) = \alpha + \beta \times \text{income}$. Again, we would take a random sample of voters and collect data on income and voting intentions for each. How could the parameters α and β be estimated?

A general method for using data to estimate model parameters is the method of maximum likelihood. It is a simple idea, most easily grasped by example. Consider rolling a loaded die, which has probability p of coming up with a 6. You roll it 10 times and get 4 sixes. The probability of this happening is

$$P(\text{4 sixes in 10 rolls}) = \binom{10}{4} p^4 (1-p)^6.$$

What is the value of p which will make this probability as high as possible?

$$\frac{\mathrm{d}}{\mathrm{d}p}(cp^4(1-p)^6) = 4cp^3(1-p)^6 - 6cp^4(1-p)^5 = 0$$

$$\Rightarrow 4(1-\hat{p}) = 6\hat{p} \Rightarrow 4 = 10\hat{p} \Rightarrow \hat{p} = \frac{4}{10}$$

This estimate is known as the *maximum likelihood estimate* of p. The idea is that the most likely value for the parameter is the one which makes the data appear most probable. Maximum likelihood estimation is the process of finding the parameters which make a set of data look as probable as possible.

In greater generality, maximum likelihood estimation works as follows. Consider a set of observations $x_1, \ldots, x_n$ which are modelled as observations of independent discrete random variables with probability function $f(x; \boldsymbol{\theta})$ which depends on some (vector of) parameters $\boldsymbol{\theta}$. According to the model, the probability of obtaining the observed data is proportional to the product of the p.f.'s for each observation, i.e.

$$L(\boldsymbol{\theta}; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}).$$

We seek the parameters of the model that make the data look most probable, so we seek to maximise $L(\boldsymbol{\theta}; x_1, \ldots, x_n)$ w.r.t. $\boldsymbol{\theta}$. When $L(\boldsymbol{\theta}; x_1, \ldots, x_n)$ is considered as a function of the parameters in this way, it is known as the *likelihood* of the parameters (rather than the probability of the data). Note that the logarithm of a function is maximised at the same set of parameters as the function itself. Very often it is easier to maximise the *log-likelihood*

$$l(\boldsymbol{\theta}; x_1, \ldots, x_n) = \log L(\boldsymbol{\theta}; x_1, \ldots, x_n),$$

rather than the likelihood $L(\boldsymbol{\theta}; x_1, \ldots, x_n)$. Note that

$$l(\boldsymbol{\theta}; x_1, \ldots, x_n) = \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta}).$$

(Recall that $\log ab = \log a + \log b$.)

Verify that you get the same estimate of p for the loaded die example.

$$l = \log[P(4 \text{ sixes in 10 rolls})] = \log \binom{10}{4} + 4 \log p + 6 \log(1-p)$$

so that

$$\frac{\mathrm{d}l}{\mathrm{d}p} = 4/p - 6/(1-p)$$

Equating this to 0 gives $4(1-p) = 6p$, from which $\hat{p} = 4/(6+4) = 0.4$.

Example: Suppose that you have 4 observations x_1, x_2, x_3, x_4 on independent Poisson distributed r.v.'s, each with p.f.

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x \geq 0.$$

First form the likelihood of λ and then the corresponding log-likelihood.

$$L = \prod_{i=1}^{4} e^{-\lambda}\lambda^{x_i}/x_i! \Rightarrow l = \log L = \sum_{i=1}^{4}[x_i \log(\lambda) - \lambda - \log(x_i!)]$$

Now maximise the likelihood w.r.t. λ.

$$\frac{\partial l}{\partial \lambda} = \sum_{i=1}^{4}\left(\frac{x_i}{\lambda} - 1\right) = 0 \Rightarrow \hat{\lambda} = \sum x_i/4$$

$$\frac{\partial^2 l}{\partial \lambda^2} = -\frac{\sum x_i}{\lambda^2} \text{ negative } \Rightarrow \hat{\lambda} \text{ is } \textbf{maximum} \text{ likelihood estimator}$$

The expression that you have just derived is known as an *estimator* if you consider it as a *function* (of $x_1, \ldots, x_n$). The value of this function which is obtained by evaluating it on observation values $x_1, \ldots, x_n$ is an *estimate*. Suppose that the observations in this case are 1, 3, 8 and 2. What is the maximum likelihood estimate?

$$\hat{\lambda} = 14/4 = 3.5$$

Note that, in general, you should check that you have obtained a *maximum* likelihood estimator and not a *minimum* likelihood estimator.

A more complicated example: Suppose that you have a series of measurements $y_1, \ldots, y_n$ of radioactive emission counts from samples of caesium of masses $x_1, \ldots, x_n$, respectively. You wish to model the counts as Poisson random variables, where each Y_i has mean αx_i. Obtain the maximum likelihood estimator of α (the radioactivity per unit mass).

$$L = \prod_{i=1}^{n} \frac{e^{-ax_i}(ax_i)^{y_i}}{y_i!} \quad l = \sum_{i=1}^{n}(-ax_i + y_i log(ax_i) - log(y_i!))$$

$$\frac{\partial l}{\partial a} = \sum_{i=1}^{n}(-x_i + y_i/a) = 0 \Rightarrow \hat{a} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$

4.1 Likelihood for continuous distributions

The examples met so far have dealt with the simple case of a single parameter estimated using discrete data. Maximum likelihood estimation works just as well for continuous random variables. The only difference is that we form the likelihood from the product of the p.d.f's of the random variables used to model the data. If $x_1, \ldots, x_n$ are observations of independent continuous r.v's with p.d.f's $f(x_i;\boldsymbol{\theta})$ which depend on some parameters $\boldsymbol{\theta}$ then the likelihood function is just

$$L(\boldsymbol{\theta}; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}).$$

This is still a likelihood, but can no longer be interpreted as a probability of getting the observed data, given $\boldsymbol{\theta}$; instead, it is a probability density (probability per unit interval [or per unit area or per unit volume]) of getting the observed data. This makes no difference to the actual calculations. We still maximise the likelihood w.r.t. the parameters, and it is still easier to use the log-likelihood in most cases.

Consider an example involving two parameters. Suppose that we have some observations $x_1, \ldots, x_n$, which we wish to model as observations of i.i.d. r.v's from a normal distribution with mean μ and variance σ^2, where the two parameters are unknown and so have to be estimated from data. First the likelihood function is formed by multiplying together the p.d.f's for the n r.v's, evaluated at the observed values to give

$$\begin{aligned} L(\mu, \sigma^2; x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}. \end{aligned}$$

Again, it is convenient to work with the log-likelihood

$$l(\mu, \sigma^2; x_1, \ldots, x_n) = \log L(\mu, \sigma^2; x_1, \ldots, x_n) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

To maximise this, find its partial derivatives w.r.t. σ^2 and μ, and set these equal to zero. In this particular case, it is easier to start with μ.

$$\frac{\partial l}{\partial \mu} = \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i - \mu) = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\mu$$

$$\Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Now find the partial derivative of the log-likelihood w.r.t. σ^2 (or, if you prefer, w.r.t. σ), and obtain the m.l.e. by setting this to zero and substituting for μ.

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Note that the maximum likelihood estimator $\hat{\sigma}^2$ of the variance σ^2 is **not** the sample variance s^2 (which is the usual estimator of σ^2).

Note: In general, maximum likelihood estimates are biased (although often the bias is fairly small). However, m.l.e's do have the advantage of being consistent.

4.2 Invariance of m.l.e's

The invariance property of maximum likelihood estimators:

If $\hat{\theta}$ denotes the maximum likelihood estimator of θ and g is any function of θ, then the maximum likelihood estimator of $g(\theta)$ is $g(\hat{\theta})$.

Example: Suppose that $x_1, \ldots, x_k$ are observations on independent binomial r.v's, each with n trials and unknown probability p. The likelihood of p is

$$L(p; x_1, \ldots, x_k) = \prod_{i=1}^{k} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i}.$$

Find the maximum likelihood estimator of p.

$$l = \sum_{i=1}^{k} \left[\log\left(\frac{n!}{(n-x_i)!x_i!} \right) + x_i \log(p) + (n-x_i)\log(1-p) \right]$$

$$\frac{\partial l}{\partial p} = \sum_{i=1}^{k} \frac{x_i}{p} - \sum_{i=1}^{k} \frac{n-x_i}{1-p} = 0 \Rightarrow \frac{1}{\hat{p}} - 1 = \frac{nk}{\sum_{i=1}^{k} x_i} - 1$$

$$\Rightarrow \hat{p} = \bar{x}/n$$

Using the invariance property, deduce the maximum likelihood estimators of the mean and variance of the bin(n, p) distribution.

mean $= \mu = np$ variance $= \sigma^2 = np(1-p)$

$$\hat{\mu} = \frac{\sum_{i=1}^{k} x_i}{k} = \bar{x} \quad \hat{\sigma^2} = \bar{x}(1-\bar{x}/n) \quad \text{by invariance}$$