Maximum Likelihood Estimators

We examine a series of (independent, identical) experiments (for example, experiments to measure a certain unknown quantity V). Let M_1, M_2, ... denote the actual results obtained which are assumed to be numerical for simplicity.

We want to examine the probability p of obtaining the result that we experimentally observed.

For example, suppose that the different experiments are independent. Then the probability p is the product of p_i where the latter is the probability of obtaining the measurement M_i in the i-th experiment.

If the result of each experiment is a discrete random variable, then we can calculate the probability as $p_i = P(X_i = M_i)$.

On the other hand, the random variable (representing the result of the experiment) may take a continuum of values with a distribution function F_i. In this case, we look at the least count e_i of our experimental apparatus for the i-th experiment, and write

$$p_i = P(X_i \in (M_i - e_i, M_i + e_i]) = F(M_i + e_i) - F_i(M_i - e_i)$$

Usually, we can *only* write p formally, since it may not be possible to know the exact probability mass functions or probability distribution functions. However, we can very often write these functions in terms of parameters that are unknown. For example, the unknown quantity V we are try to measure could be thought of as such a parameter. In that case, we can think of p as varying with V; in other words, we can think of p as a *function* of V.

More generally, let t_1, t_2, ... be parameters that can be used to determine the probabilities p_i (and perhaps the probability mass functions or distribution functions). We can then think of p as a function $L(M_1, \dots; t_1, \dots)$ of the *results* M_i *and the parameters* t_i. This function is called the *likelihood* function.

We would like to *estimate* these parameters, based on the experiments performed. Such estimators $T_i(m_1, \dots)$ are functions of the results; so that our estimates are $t_i = T_i(M_1, \dots)$.

The method of Maximum Likelihood Estimation is to choose these estimators (functions) in such a way that $L(M_1, \dots; t_1, \dots)$ takes its maximum value when we put $t_i = T_i(M_1, \dots)$.

This method was formally proposed and analysed by Ronald A. Fisher (it had already been used earlier by Gauss and others).

It is easier to understand this theoretical framework trough some typical examples.

Coin Choice

Suppose a box contains a number of coins each having its own probability of getting a head. For example suppose we have coins C_1, C_2, C_3 and C_4 with probability of Head being 1/2, 2/3, 1/4 and 1.

We pick a coin from the box, but we don't know which one it is! We try to figure out which one by flipping it a large number N of times. Assuming that the distinct flips are independent, the probability of a given sequence of Heads and Tails is given by $p^r(1-p)^{N-r}$; where the value of p depends on which coin was chosen. For definite-ness, let us assume that there are $r = 40$ Heads out of $N = 100$ coin tosses.

We can then make a table as follows:

p	$1/2$	$2/3$	$1/4$	1
L	$1/2^{100}$	$2^{40}/3^{100}$	$3^{60}/4^{100}$	0

Here L is the likelihood that the given value of p is the "real" one. The ratio of the first value by the second value is a power of $3^5/2^7 = 243/128$ which is greater than 1. Thus, the first value is greater than the second. Similarly, the ratio of the first value by the third is a power of $2^5/3^3 = 32/27$ which is greater than 1. Hence the first value is greater than the third as well. It is clear that all the other values are greater than 0.

We conclude that the maximum likelihood is that the value of p is 1/2. In other words, we picked the first coin.

Coin Flips

Now lets take a single coin for which we are trying to find out its bias. This is like the previous experiment, but with infinitely many coins, one for each value of p such that $0 < p < 1$.

As before we assume that distinct coin flips are independent, we put the probability of Head as p and the probability of tail as $1 - p$.

As in the first experiment we record the exact sequence of Heads and Tails obtained.

We start with the results of 10000 tosses of a slightly biased coin. Let us see whether we can detect the bias.

```
> dat <- rbinom(10000,1,0.49)+1 # to make it 1,2 rather than 0,1
> coin <- as.factor(c('T','H')[dat])
```

The probability of the given sequence of heads and tails is given by $p^r(1-p)^s$ where r is the number of heads and s is the number of tails ($r+s=10000$). In our case, we get

```
> tapply(coin,coin,length)
   T    H
5145 4855
```

So the likelihood function is $L(p) = p^{5145}(1-p)^{4855}$ (or more generally, $L(r,s;p) = p^r(1-p)^s$. The question we have to resolve is: Given r and s, what is the value of p for which this takes the maximum value? Using some elementary calculus, this amounts to

$$p^r(1-p)^s\left(\frac{r}{p} - \frac{s}{1-p}\right) = 0$$

Equivalently $r(1-p) - sp = 0$ or $p = r/(r+s)$. In other words, the *estimate* that is *most likely* to fit the results obtained is $p = 0.4855$.

Note that this result depends only on the *number* of heads and the total number of coin flips since the Likelihood also depends only on those counts.

Secondly note that the estimator has the simple description as the frequency of occurrence of heads. This "classical" estimator is therefore also the maximum likelihood estimator in this case.

Poisson Density

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables following the Poisson distribution $P(X_i = k) = e^{-c}c^k/k!$. Suppose that the result of the experiment is that $X_i = k_i$. The likelihood of this result is

$$L(c) = \prod_{i=1}^{n} \frac{e^{-c}c^{k_i}}{k_i!} = \frac{e^{-cn}c^{\sum_{i=1}^n k_i}}{\prod_{i=1}^n k_i!}$$

Since $\log(L)$ is a *monotonic* increasing function of L we can calculate the maximum after taking the logarithm.

$$\log L(c) = -cn + \log(c)\sum_{i=1}^{n} k_i - \sum_{i=1}^{n} \log(k_i!)$$

Taking the derivative

$$\frac{d\log L(c)}{dc} = -n + (1/c)\sum_{i=1}^{n} k_i$$

Putting this to 0 (in order to find the extremum) we get

$$c = \frac{\sum_{i=1}^n k_i}{n}$$

Thus, the maximum likelihood estimate for the parameter c is given by the mean.

Waiting Time

Cars go past a certain checkpost with a frequency c which is unknown. A guard records the waiting times t_1, t_2,... for various cars. We want to use Maximum Likelihood Estimation to write down an estimator for c given the values t_1, t_2, ...

The waiting time distribution is given by the Poisson density $f_X(t) = ce^{-ct}$ when the frequency is c. Let us assume that the least count of the guard's stop watch is e.

The probability of the guard's i-th observation is

$$\int_{t_i-e}^{t_i+e} ce^{-ct}dt = e^{-c(t_i-e)} - e^{-c(t_i+e)}$$

We thus want to maximise the function

$$L(c) = \prod_{i=1}^{n} \left(e^{-ct_i}\left(e^{ce} - e^{-ce}\right)\right) = e^{-c\sum_i t_i}(2^n)\prod_{i=1}^{n}\sinh(ce)$$

(Here $\sinh(x) = (e^x - e^{-x})/2$ is called the hyperbolic sine function.)

By calculus, we are looking for c such that

$$0 = \frac{dL}{dc} = L(c)\left(-\sum_{i=1}^{n} t_i + e\sum_{i=1}^{n}\coth(ce)\right)$$

(Here $\coth(x) = (e^x + e^{-x})/(e^x - e^{-x})$ is the hyperbolic cotangent function.)

We can assume that e is very small, so that $(ce)\coth(ce)$ is approximately 1, then the approximate solution of this equation is given by the slightly different equation

$$0 = -\sum_{i=1}^{n} t_i + n/c \text{ or } c = \frac{n}{\sum_i t_i}$$

In other words, the estimator for the frequency c which is the ratio of the number of cars seen by the total amount of time is *close* to the maximum likelihood estimator for c.

This justifies the fact that the parameter c in the Poisson density is called the frequency!

Continuous Probability Density

The above technique can be applied more generally to the case of a continuous probability density. In that case, we have

$$P(|X - x| < e) = \int_{x-e}^{x+e} f_X(s)ds \simeq f_X(x)e$$

for small values of e. Thus, instead of finding the maximum of the likelihood function (with parameter c)

$$L(c) = \prod_{i=1}^{n} P(|X_i - x_i| < e|c)$$

we can find the maximum of the likelihood density

$$l(c) = \prod_{i=1}^{n} f(x_i|c)$$

This will give an estimator which is *close* to the maximum likelihood estimator for small values of e. Moreover, as in the case above, it is often easier to calculate than the maximum likelihood estimator.

Normal Distribution

We can apply the above to a sequence of experiments which result in a sequence X_1, X_2, ... of random variables each of which is distributed according to the normal distribution with mean m and standard deviation s. The numbers m and s are the parameters which we wish to estimate based on the results of the experiments.

To do this we write

$$l(m, s) = \frac{1}{s^n} \prod_{i=1} \exp\left(\left(-\frac{x_i - m}{s}\right)^2 /2\right)$$

Note that since the $\sqrt{2\pi}$ factor in the denominator is a constant, it can be ignored while finding the maximum of $l(m, s)$

Secondly, since $\log(y)$ is an *increasing* function of y, we can replace the problem of finding the maximum of $l(m, s)$ by finding the maximum of

$$k(m, s) = \log(l(m, s)) = -n \log(s) - \sum_{i=1}^{n} \frac{(x_i - m)^2}{2s^2}$$

Taking derivative with respect to m and putting it as 0, we find

$$0 = \frac{\partial k}{\partial m} = -\frac{1}{s^2} \sum_{i=1}^{n} (x_i - m)$$

From this we deduce the estimator for m given by the formula

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

which is called the *sample mean*.

Similarly, we take the derivative with respect to s and it as 0 to get

$$0 = \frac{\partial k}{\partial s} = -\frac{n}{s} + \frac{1}{s^3}\sum_{i=1}^{n}(x_i - m)^2$$

Then we see that we obtain an estimator for s^2 as

$$\frac{\sum_{i=1}^{n}(x_i - m)^2}{n}$$

which is similar to the usual formula for variance. (However, see the discussion below!)

A Mean Estimator

As seen above, there are a number of distributions for which there is just one parameter that determines the distribution. Moreover, the expectation for such a distribution determines and is uniquely determined by the value of the parameter.

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables that follow a distribution f of the above kind and let θ denote the parameter. Further, let $E(X_i) = \mu(\theta)$ be the expectation and assume that this is a 1-1 function of θ. In this this case, we can use μ as the parameter. Moreover, it will turn out (as we have seen above in many cases!) that the maximum likelihood estimate for μ is the *mean* $(\sum_i X_i)/n$.

The following explicit examples will clarify this matter.

- Bernoulli trials: Let X_i be a sequence of independent coin flips with a coin that shows Head (or 1) with a probability of θ. In that case we have $E(X_i) = \theta = \mu$. Further, we have seen that the maximum likelihood estimator for θ is the *frequency* of occurrence of Head which is the same as $(\sum_i X_i)/n$.
- Poisson Distribution: Let X_i be a sequence of independent random variables following the Poisson distribution $P(X_i = k) = e^{-\theta}(\theta)^k/k!$. In this case too $E(X_i) = \theta = \mu$. One has seen above that $(\sum_i X_i)/n$ is the mle for θ.
- Waiting time: Let X_i be an independent sequence of waiting times with frequency of occurrence as θ. These random variables have the probability density $p_{X_i}(t) = \theta e^{-t\theta}$. In this case $E(X_i) = 1/\theta = \mu$. We have seen above that the maximum likelihood estimate for θ is given by $n/(\sum_i X_i)$; so the maximum likelihood estimate for μ is again given by $(\sum_i X_i)/n$.

As seen above the same holds for the normal distribution. In fact, one can use the Central Limit Theorem to justify the statement that the maximum likelihood estimate for the expectation *is* the average once the number of observations is large enough.

Biased and Unbiased Estimators

As seen above an estimator is a function $f(X_1, \ldots, X_n)$ of the observational results X_i which are themselves random variables. Thus, it makes sense to talk about the mathematical expectation of this function. When the expectation $E(f(X_1, \ldots, X_n))$ turns out to be the same as the value of the parameter that we are estimating, then we say that the estimator is *unbiased*. Otherwise, we say that the estimator is *biased*.

One can show that the sample mean $(\sum_i X_i)/n$ is an unbiased estimator.

On the other hand the expression $(\sum_i (X_i - m)^2)/n$ in which m is taken to be the sample mean, is no longer an unbiased estimator. One way to understand this is to note that the variables $X_i - m$ are no longer independent since $nm = \sum_i X_i$. It turns out that $(\sum_i (X_i - m)^2)/(n-1)$ *is* an unbiased estimator for the variance s^2. Hence, it is this latter expression which is called the *sample variance*. For large values of n there is hardly any difference between this formula and the earlier one, so we need not fuss about it too much!