3 Maximum Likelihood Estimation

3.1 Motivating example

We now come to the most important idea in the course: *maximum likelihood estimation*.

Let us begin with a special case. Our data is a a Binomial random variable X with parameters 10 and p_0. The parameter p_0 is a fixed constant, unknown to us. That is,

$$f(x; p_0) = P_{p_0}(X = x) = \binom{n}{x} p_0^x (1 - p_0)^{n-x} .$$

Suppose that we observe $X = 3$. This we regard as our fixed data.

Our goal, as in all point estimation problems, is to estimate the actual parameter value p_0 based on the available data.

We consider now some thought experiments. We do not know p_0, but we can consider the scenario in which the value of p_0 is $1/2$. *Under this particular assumption*, the probability of generating the data *which we actually saw* – namely $X = 3$ – is

$$f(3; 0.5) = P_{0.5}(X = 3) = \binom{10}{3} (0.5)^3 (0.5)^7 \approx 0.117 .$$

We can calculate this probability under the assumption that $p_0 = p$ for each $p \in [0, 1]$. For a given p, this probability is

$$f(3; p) = P_p(X = 3) = \binom{10}{3} p^3 (1 - p)^7 .$$

We thus obtain a function $p \mapsto f(3; p)$. This function is called the *likelihood function*. We write $L(p; 3)$ for the value of this function at $p = 3$.

The principle of maximum likelihood says we should use as our estimate of p_0 the value p which makes $L(p; 3)$ as large as possible. This is a reasonable idea: we pick the parameter value p which makes the observed data most likely when assuming p_0 equals p.

Notice that since log is an increasing function, the value of p which maximizes $L(p; 3)$ is the same value which maximizes $\log L(p; 3)$. It is often convenient to maximizes the logarithm of the likelihood function instead of the function itself, so we give this function a name and notation: we write $\ell(p; 3)$ for the *log-likelihood* function, defined as $\ell(p; 3) \stackrel{\text{def}}{=} \log L(p; 3)$.

Here,

$$\ell(p;3) = 3\log p + 7\log(1-p) + \log(\binom{n}{k}).$$

We use calculus to maximize this function. We can first graph ℓ to see its general shape – see Figure 1.

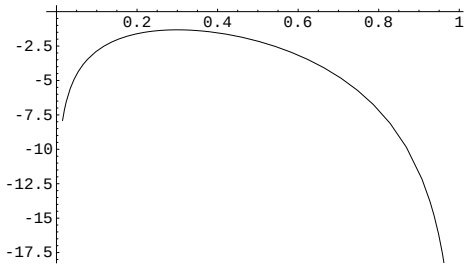

Figure 1: Graph of $\ell(p;3)$.

We note in particular the ℓ has a unique maximum at the single critical point. [A *critical point* of a function is a point in the domain where the derivative is zero.]

We compute:

$$\begin{aligned} 0 &= \frac{\partial \ell}{\partial p}(p;3) \\ 0 &= \frac{3}{p} - \frac{7}{1-p} \\ 0 &= 3(1-p) - 7p \\ p &= \frac{3}{10}. \end{aligned}$$

Thus the value of p maximizing $\ell(p;3)$ is $p = \frac{3}{10}$. We call this the *maximum likelihood estimate* of p_0, for the data $X = 3$.

It is clear that if we observed $X = k$, where $k = 0, 1 \ldots, n$, the maximum likelihood estimate of p_0 would be k/n. Thus, the estimate is determined by the value of X, and we have the estimator $\hat{p} = X/n$. This is a statistic (a function of our sample, which in this case consists only of X).

3.2 Definitions

The generic situation is that we observe a n-dimensional random vector $\boldsymbol{X}$ with probability density (or mass) function $f(\boldsymbol{x};\theta)$. It is assumed that θ is a fixed, unknown constant belonging to the set $\Theta \subset \mathbb{R}^k$.

Definition 7. For $\boldsymbol{x} \in \mathbb{R}^n$, the *likelihood function* of θ is defined as

$$L(\theta;\boldsymbol{x}) = f(\boldsymbol{x};\theta)\,.$$

$\boldsymbol{x}$ is regarded as fixed, and θ is regarded as the variable for L. The *log-likelihood function* is defined as $\ell(\theta;\boldsymbol{x}) = \log L(\theta;\boldsymbol{x})$.

Definition 8. The *maximum likelihood estimate* (or mle) is the value $\hat{\theta} = \hat{\theta}(\boldsymbol{x}) \in \Theta$ maximizing $L(\theta;\boldsymbol{x})$, provided it exists:

$$L(\hat{\theta}(\boldsymbol{x})) = \max_{\theta\in\Theta} L(\theta,\boldsymbol{x})\,.$$

4 Examples

We see that the problem of finding a maximum likelihood estimate is now reduced to the problem of optimizing the likelihood function. As in any optimization problem, one must be careful; we give some examples and pitfalls here.

Example 1 (Poisson). Let $X_1,\ldots,X_n$ be an i.i.d. collection of Poisson(μ) random variables, where $\mu > 0$. Thus the likelihood function is

$$\begin{aligned} L(\mu;\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{x_i}}{x_i!} \\ &= e^{-n\mu}\mu^{\sum_{i=1}^n x_i}\frac{1}{\prod_{i=1}^n x_i!} \\ \ell(\mu;\boldsymbol{x}) &= -n\mu + \sum_{i=1}^{n} x_i \log\mu - \log\prod_{i=1}^{n} x_i!\,. \end{aligned}$$

We note that $\ell(\mu; \boldsymbol{x})$ is a differentiable function over the domain $(0, \infty)$, and so we first find the critical points:

$$\begin{aligned} 0 &= \frac{\partial \ell}{\partial \mu}(\mu; \boldsymbol{x}) \\ &= -n + \frac{\sum_{i=1}^n x_i}{\mu} \\ \mu &= \bar{x}\,. \end{aligned}$$

[Here $\bar{x}$ denotes $n^{-1}\sum_{i=1}^n x_i$.]

Thus there is a single critical point at $\mu = \bar{x}$. Taking the second derivative gives

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\mu^{-2}\sum_{i=1}^n x_i < 0\,.$$

Thus there is a local maximum at $\mu = \bar{x}$. We then note that as $\mu \to 0$ or $\mu \to \infty$, the log-likelihood $\ell(\mu; \boldsymbol{x})$ approaches $-\infty$. Thus $\mu = \bar{x}$ is a global maximum, and the maximum likelihood estimate of μ is $\hat{\mu} = \bar{x}$.

The maximum likelihood estimator in this example is then $\hat{\mu}(\boldsymbol{X}) = \bar{X}$. Since μ is the expectation of each X_i, we have already seen that $\bar{X}$ is a reasonable estimator of μ: by the Weak Law of Large numbers, $\bar{X} \xrightarrow{\Pr} \mu$ as $n \to \infty$. We have just seen that according to the maximum likelihood principle, $\bar{X}$ is the preferred estimator of μ.

Example 2 (Multinomial). Suppose that we have n independent experiments, each of which must result in one of r mutually exclusive outcomes. On each trial, the probability of the ith outcome is p_i, for $i = 1, \ldots, r$. For example, we drop n balls into r boxes. Let N_i be the number of these experiments which result in the ith outcome, where $i = 1, \ldots, r$.

The joint mass function for $(N_1, \ldots, N_r)$ is given by

$$f(\boldsymbol{n}; \boldsymbol{p}) = \binom{n}{n_1 n_2 \ldots n_r} \prod_{i=1}^r p_i^{n_i}\,.$$

Notice the parameter space is the simplex

$$\Theta = \{\boldsymbol{p} \,:\, p_i \geq 0\,, \sum_{i=1}^r p_i = 1\} \subset \mathbb{R}^r\,.$$

Notice that Θ is a curved surface inside $\mathbb{R}^r$. We want to maximize the likelihood over this surface, not over all of $\mathbb{R}^r$. Thus we might use the method of Lagrange multipliers.

We recall the following from multi-variable calculus:

Theorem 6 (Constrained optimization). *Let $f : A \subset \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function, where A is an open set. Let $g : A \subset \mathbb{R}^n \to \mathbb{R}$ also be a continuously differentiable function. For $c \in \mathbb{R}$, let $S = g^{-1}(c) = \{x : g(x) = c\}$. If, among all points in S, the function f has an extreme point at p_0, then $\nabla f(p_0) = \lambda \nabla g(p_0)$, where λ is a real number.*

We use this theorem by finding all solutions to $\nabla f(x) = \lambda \nabla g(x)$; the extremes of f must be among these solutions.

In our example, we want to find the maximum of $\ell(\boldsymbol{p})$ among all $\boldsymbol{p} \in h^{-1}(1)$, where $h(\boldsymbol{p}) = \sum_{i=1}^n p_i$.

We first write

$$\ell(\boldsymbol{p}; \boldsymbol{n}) = \sum_{i=1}^r n_i \log p_i + \log \binom{n}{n_1 \dots n_r} .$$

We calculate

$$\begin{aligned} \nabla \ell(\boldsymbol{p}; \boldsymbol{n}) &= \left(n_1 p_1^{-1}, \dots, n_r p_r^{-1}\right) \\ \nabla g(\boldsymbol{p}) &= (1, \dots, 1) \ . \end{aligned}$$

Solving $\nabla \ell(\boldsymbol{p}; \boldsymbol{x}) = \lambda \nabla g(\boldsymbol{p})$ yields $p_i = \lambda^{-1} n_i$. Since $\sum_i p_i = 1$, and $\sum_i n_i = n$, we have $\lambda = n$ and the solution is $p_i = n_i/n$, for $i = 1, \dots, r$.

We conclude that the mle of $\boldsymbol{p}$ is

$$\hat{\boldsymbol{p}}(\boldsymbol{N}) = \left(\frac{N_1}{n}, \frac{N_2}{n}, \dots, \frac{N_r}{n}\right) .$$

An alternative way to maximize ℓ in this problem is to write $p_r = 1 - \sum_{i=1}^{r-1} p_i$, so that there are now $r - 1$ free parameters $(p_1, \dots, p_{r-1}) \in (0, 1)^{r-1}$, reducing the optimization problem to one with domain equal to an open subset of Euclidean space.

Example 3 (Uniform). Here is an example to keep in mind. Let $X_1, \dots, X_n$ be i.i.d. Uniform on the interval $[0, \theta]$, where $\theta > 0$. That is, X_i has pdf

$$f(x_i) = \theta^{-1} \mathbf{1}\left\{0 \le x_i \le \theta\right\} .$$

The indicator appearing in the density is important. In examples where the support of the density (the interval where it is positive) depends on the parameter, one must be careful to always indicate the support when writing down the density.

So

$$L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} \theta^{-1} \mathbf{1}\left\{0 \leq x_i \leq \theta\right\} = \theta^{-n} \mathbf{1}\left\{0 \leq x_{(1)} \leq x_{(n)} \leq \theta\right\} .$$

Sketching L will show that it is maximized at $\theta = x_{(n)}$, so the maximum likelihood estimator is $\hat{\theta}(\boldsymbol{X}) = X_{(n)}$. Notice that since the likelihood function has a discontinuity, the maximum is not attained at a critical point.