# Likelihood and Maximum Likelihood Estimation

## 3.1 Motivation

- Consider a (possibly biased) coin for which $P$(Head) = $\theta$ and $P$(Tail) = $(1-\theta)$, where $\theta$ is an unknown parameter taking values in $(0,1)$ which we wish to estimate.

  One way of gaining information about $\theta$ might be to perform an experiment in which we repeatedly toss the coin and count the number of tosses until we get a Head. Assume the outcome of each toss is independent of all the other tosses, and let $X$ denote the number of the toss on which we first get a Head. Then $X \sim$ Geom$(\theta)$ and

$$p\,(x;\theta) = (1-\theta)^{x-1}\theta \qquad x = 1,2,3,\ldots; \quad \theta \in (0,1).$$

- Say we perform the experiment once and get a single observation $x = 4$ (so the first head was observed on the fourth toss). Write $L(\theta)$ [or more properly $L(\theta; x)$] for the probability of getting this particular observation $x$ as a function of the unknown parameter $\theta$. Thus in this case $L(\theta)$ is got by putting $x = 4$ in the above expression for $p\,(x;\theta)$, giving

$$L(\theta) = p\,(4;\theta) = (1-\theta)^3\theta \qquad \theta \in (0,1).$$

  We call $L(\theta)$ the **likelihood function** for the given observation. A graph of $L(\theta)$ against $\theta$ is shown below.

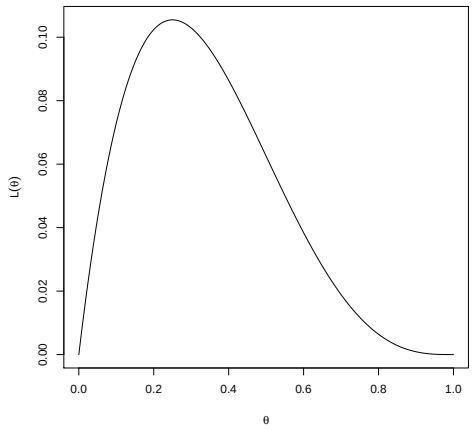


Figure 1: The likelihood function $L(\theta)$ as a function of the unknown parameter $\theta$

- The value of $\theta$ that maximises the likelihood function $L(\theta)$, i.e. that maximises the probability of getting this particular observation, is called the **maximum likelihood estimate** of $\theta$.

  We can see that the maximum here is a turning point, so here the maximising value satisfies the equation

$$\frac{dL(\theta)}{d\theta} = 0 \quad \text{where} \quad \frac{dL(\theta)}{d\theta} = (1-\theta)^3 - 3\theta(1-\theta)^2 = (1-\theta)^2(1-4\theta)$$

  and you can check that the likelihood function is maximised at $\theta = 1/4 = 0.25$.

Thus this single observation $x = 4$ has a greater likelihood of occurring when the parameter $\theta$ takes the value $\theta = 0.25$ than when $\theta$ takes any other possible value in $(0, 1)$. We call this value $0.25$ the **maximum likelihood estimate** of $\theta$ for the single observation $x = 4$ and we denote it by $\widehat{\theta}$ [or more properly $\widehat{\theta}(x)$ or even $\widehat{\theta}_{\text{mle}}(x)$] and write $\widehat{\theta} = 0.25$.

- We can easily generalise our analysis to the case of several observations. For example, say we repeat the experiment three times and get three independent observations $x_1 = 4, x_2 = 5$, and $x_3 = 1$. Then the corresponding random variables $X_1, X_2, X_3$ are independent, each with the same Geometric distribution as $X$ above, so they have joint probability mass function

$$p_{X_1,X_2,X_3}(x_1, x_2, x_3; \theta) = p(x_1; \theta)\, p(x_2; \theta)\, p(x_3; \theta)$$

where the expression for $p(x; \theta)$ is given above.

In this case the likelihood function denotes the probability of observing these three numerical values $x_1, x_2, x_3$ as a function of the unknown parameter $\theta$ and is given by

$$\begin{aligned} L(\theta) \equiv L(\theta; x_1, x_2, x_3) &= p_{X_1,X_2,X_3}(x_1, x_2, x_3; \theta) \\ &= p(x_1; \theta)\, p(x_2; \theta)\, p(x_3; \theta) \\ &= (1-\theta)^{x_1-1}\theta\,(1-\theta)^{x_2-1}\theta\,(1-\theta)^{x_3-1}\theta \\ &= (1-\theta)^{3}\theta\,(1-\theta)^{4}\theta\,(1-\theta)^{0}\theta \\ &= (1-\theta)^{7}\theta^{3}. \end{aligned}$$

A graph of this new $L(\theta)$ against $\theta$ is shown below. You can easily show that $L(\theta)$ is now maximised at the value $\theta = 3/10 = 0.3$.

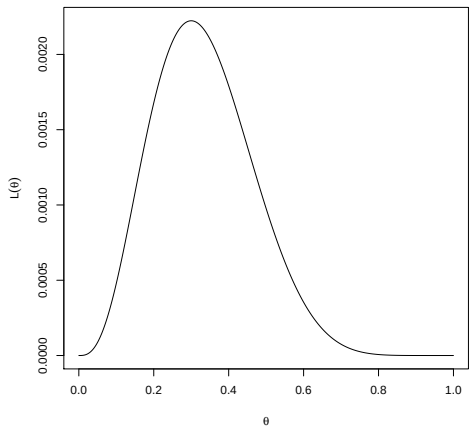


Figure 2: The likelihood function $L(\theta; x_1, x_2, x_3)$ as a function of the unknown parameter $\theta$

Thus, as a function of the unknown parameter $\theta$, the likelihood of these three numerical observations $x_1 = 4, x_2 = 5, x_3 = 1$ occurring is maximised by taking $\theta = 0.3$. Again, we say that 0.3 is the maximum likelihood estimate of $\theta$ and we write $\widehat{\theta}(x_1, x_2, x_3) = 0.3$ – or, since the context is clear, $\widehat{\theta} = 0.3$.

## 3.2 Definition – Likelihood function

- **General case:** Assume the data $x_1, x_2, \ldots, x_n$ are the observed numerical values of random variables $X_1$, $X_2$, …, $X_n$, whose joint distribution depends on one or more unknown parameters $\theta$. The **likelihood function** $L(\theta) \equiv L(\theta; x_1, x_2, \ldots, x_n)$ is just the joint probability mass function (discrete case) or joint probability density function (continuous case) regarded as a function of the unknown parameter $\theta$ for these fixed numerical values of $x_1, x_2, \ldots, x_n$.

- **Usual random sample case:** If $X_1, X_2, \ldots, X_n$, is a random sample of size $n$ from a distribution with probability mass function $p(x;\theta)$ (or probability density function $f(x;\theta)$) then the $X_i$ are i.i.d.r.v.s and their joint distribution factorises into the product of the individual marginal distributions. Thus for a random sample

$$L(\theta) \equiv L(\theta; x_1, x_2, \ldots, x_n) = \begin{cases} p(x_1;\theta)\ p(x_2;\theta) \cdots p(x_n;\theta) & \text{discrete case} \\ \\ f(x_1;\theta)\ f(x_2;\theta) \cdots f(x_n;\theta) & \text{continuous case} \end{cases}$$

- $L(\theta)$ is a function of $\theta$ for fixed data $x_1, x_2, \ldots, x_n$.

- We interpret $L(\theta)$ as a combined measure of of how well the value $\theta$ explains the set of observations, and hence of the 'plausibility' of $\theta$. e.g. if $\{x_i\}$ are collectively unlikely as observations from $f_X(x;\theta)$ then $L(\theta)$ is small, and vice-versa.

## 3.7 Maximum likelihood estimate of $\tau(\theta)$

- The actual quantity of interest may be a function $\tau(\theta)$ of $\theta$, not $\theta$ itself.

  In this case the mle of $\tau(\theta)$ is simply

$$\widehat{\tau(\theta)} = \tau(\widehat{\theta})$$

  (just the 'plug-in' estimate)

- This is the *invariance* property of maximum likelihood estimation – we can get the mle for $\tau$ without going to the trouble of re-parameterizing the problem in terms of $\tau$ and then solving $\partial\ell(\tau)/\partial\tau = 0$ for $\widehat{\tau}$ (which would give the same answer, at least when the function $\tau(\theta)$ is 1-1 (injective)).

  To see this last point, note that by the chain rule for differentiation applied to $\ell(\tau(\theta))$, $\partial\ell(\tau(\theta))/\partial\theta = \partial\ell(\tau)/\partial\tau \times \tau'(\theta)$, so that provided $\tau'(\theta) \neq 0$, $\partial\ell(\tau(\theta))/\partial\theta = 0$ if and only if $\partial\ell(\tau)/\partial\tau = 0$.

## 3.8 Example – Exp$(\theta)$

- Again let $x_1, x_2, \ldots, x_n$ be observed values of a simple random sample from the Exp$(\theta)$ distributuion with $\theta$ unknown.

  We found in §3.6 that $\widehat{\theta}_{\text{mle}} = 1/\overline{x}$.

  – Suppose we are interested in the population variance $\text{Var}(X;\theta) = 1/\theta^2$, i.e. in $\tau(\theta) = 1/\theta^2$. Then the mle of the population variance is

$$\widehat{\tau(\theta)} = \tau(\widehat{\theta}) = 1/\widehat{\theta}^2 = \overline{x}^2$$

    This is not the same as the sample variance!

  – Suppose instead that we are interested in the proportion of the population taking values $\geq 1$, that is, in $\tau(\theta) = P\{X \geq 1;\theta\} = e^{-\theta}$ for the Exp$(\theta)$ case.

    Then the mle of this proportion is

$$\widehat{\tau(\theta)} = \tau(\widehat{\theta}) = e^{-\widehat{\theta}} = \exp(-1/\overline{x})$$

    This is not the same as the sample proportion of values $\geq 1$!

## 3.11 An example with not-identically-distributed observations

- One of the strengths of the maximum likelihood approach is that it still provides answers when the observations cannot be treated as i.i.d. The likelihood is just the joint probability density or mass function of the data, regarded as a function of the parameters, and this makes sense, and can be maximised with respect to the parameters, whatever the model. Here we just consider one example, where the observations are still independent, but with different distributions.

- *Poisson data with unequal means*. The Poisson distribution is used to model counts of events that can be assumed to occur completely at random – for example, counts of photons arriving at a detector in different intervals of time. Suppose that the *rate* of arrival is $\lambda$ per unit time, and that $X_i$ is the number of arrivals in a time interval of length $t_i$. The it is natural to assume that $X_i \sim$ Poisson$(\lambda t_i)$, for $i = 1, 2, \ldots, n$. If the different intervals are not overlapping, then the counts $X_i$ should be independent. If the $t_i$ are all equal (all $t_i = t$, say), then we have a simple random sample from Poisson$(\lambda t)$.

  But it is useful and instructive to look also at the case of *unequal* $t_i$. The joint probability mass function of $X_1, X_2, \ldots, X_n$ is

$$\begin{aligned} P\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} &= P\{X_1 = x_1\} \times P\{X_2 = x_2\} \times \cdots \times P\{X_n = x_n\} \\ &= \frac{e^{-\lambda t_1}(\lambda t_1)^{x_1}}{x_1!} \times \frac{e^{-\lambda t_2}(\lambda t_2)^{x_2}}{x_2!} \times \cdots \times \frac{e^{-\lambda t_n}(\lambda t_n)^{x_n}}{x_n!} \\ &= e^{-\lambda(t_1+t_2+\cdots+t_n)} \lambda^{x_1+x_2+\cdots+x_n} \frac{t_1^{x_1} t_2^{x_2} \cdots t_n^{x_n}}{x_1! x_2! \cdots x_n!} \end{aligned}$$

  So the log-likelihood is just the logarithm of this, considered as a function of $\lambda$:

$$\ell(\lambda) = -\lambda(t_1 + t_2 + \cdots + t_n) + (\log \lambda)(x_1 + x_2 + \cdots + x_n) + \text{terms not containing } \lambda.$$

  Then

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = -(t_1 + t_2 + \cdots + t_n) + (1/\lambda)(x_1 + x_2 + \cdots + x_n)$$

  So

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = 0 \quad \text{if and only if} \quad \lambda = \widehat{\lambda} = \frac{(x_1 + x_2 + \cdots + x_n)}{(t_1 + t_2 + \cdots + t_n)}$$

  so this is the mle. (The turning value we have found is obviously a maximum, since we can see that $\partial \ell(\lambda)/\partial \lambda$ is decreasing in $\lambda$). Note that the mle $\widehat{\lambda}$ is the total count of photons divided by the total time of observation. Note that, when the $t_i$ are unequal, this is *not* the same as the average of the individual estimates $(x_i/t_i)$.