# Maximum Likelihood Estimation (MLE)*

Ewa Paszek

This work is produced by OpenStax-CNX and licensed under the
Creative Commons Attribution License 2.0†

**Abstract**

This course is a short series of lectures on Introductory Statistics. Topics covered are listed in the Table of Contents. The notes were prepared by Ewa Paszek and Marek Kimmel. The development of this course has been supported by NSF 0203396 grant.

## 1 MAXIMUM LIKELIHOOD ESTIMATION (MLE)

### 1.1 Likelihood function

From a statistical standpoint, the data vector $x = (x_1, x_2, ..., x_n)$ as the outcome of an experiment is a random sample from an unknown population. **The goal of data analysis is to identify the population that is most likely to have generated the sample.** In statistics, each population is identified by a corresponding probability distribution. Associated with each probability distribution is a unique value of the model's parameter. As the parameter changes in value, different probability distributions are generated. Formally, a model is defined as the family of probability distributions indexed by the model's parameters.

Let denote the **probability distribution function** (PDF) by $f(x|\theta)$ that specifies the probability of observing data **y** given the parameter **w**. The parameter vector $\theta = (\theta_1, \theta_2, ..., \theta_k)$ is a vector defined on a multi-dimensional parameter space. If individual observations, $x_i$'s are statistically independent of one another, then according to the theory of probability, the PDF for the data $x = (x_1, x_2, ..., x_n)$ can be expressed as a multiplication of PDFs for individual observations,

$$f(x, \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta),$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

**To illustrate the idea of a PDF**, consider the simplest case with one observation and one parameter, that is, $n = k = 1$. Suppose that the data **x** represents the number of successes in a sequence of 10 independent binary trials (e.g., coin tossing experiment) and that the probability of a success on any one trial, represented by the parameter, $\theta$ is 0.2. The PDF in this case is then given by

$$f(x|\theta = 0.2) = \frac{10!}{x!(10-x)!}(0.2)^x(0.8)^{10-x}, (x = 0.1, ..., 10),$$

*Version 1.3: Oct 23, 2007 3:55 pm -0500

†http://creativecommons.org/licenses/by/2.0/

which is known as the binomial probability distribution. The shape of this PDF is shown in the top panel of Figure 1 (Figure 1). If the parameter value is changed to say $\mathbf{w} = 0.7$, a new PDF is obtained as

$$f\left(x|\theta = 0.7\right) = \frac{10!}{x!\left(10-x\right)!}(0.7)^{x}(0.3)^{10-x}, (x = 0.1, ..., 10)\,;$$

whose shape is shown in the bottom panel of Figure 1 (Figure 1). **The following is the general expression of the binomial PDF for arbitrary values of $\theta$ and n:**

$$f\left(x|\theta\right) = \frac{n!}{\theta!\left(n-x\right)!}\theta^{x}(1-\theta)^{n-x}, 0 \leq \theta \leq 1, x = 0.1, ..., n;$$

which as a function of $\mathbf{y}$ specifies the probability of data $\mathbf{y}$ for a given value of the parameter $\theta$ . The collection of all such PDFs generated by varying parameter across its range (0 - 1 in this case) defines a model.

---

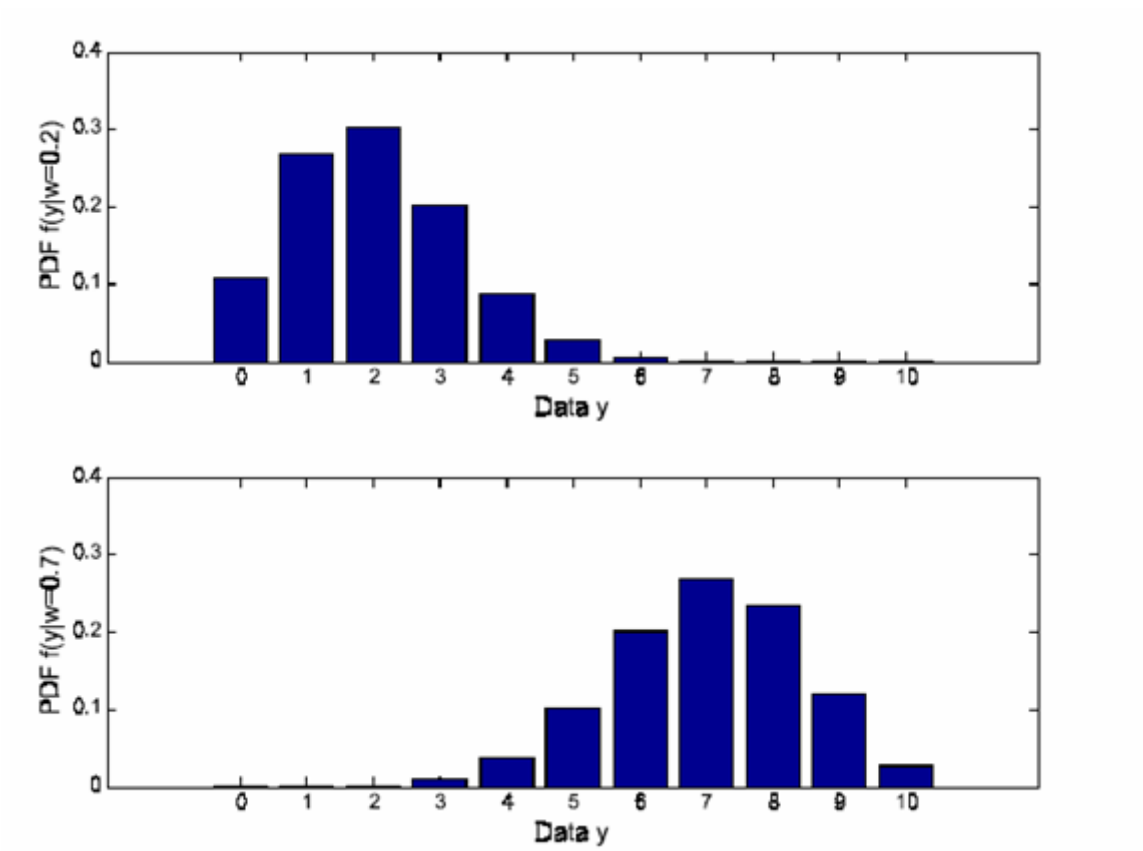


**Figure 1:** Binomial probability distributions of sample size $\mathbf{n} = 10$ and probability parameter $\theta = 0.2$ (top) and $\theta = 0.7$ (bottom).

---

### 1.2 Maximum Likelihood Estimation

Once data have been collected and the likelihood function of a model given the data is determined, one is in a position to make statistical inferences about the population, that is, the probability distribution that underlies the data. Given that different parameter values index different probability distributions (Figure 1 (Figure 1)), we are interested in finding the parameter value that corresponds to the desired PDF.

The principle of **maximum likelihood estimation (MLE)**, originally developed by R. A. Fisher in the 1920s, states that the desired probability distribution be the one that makes the observed data most likely, which is obtained by seeking the value of the parameter vector that maximizes the likelihood function (Section 1: MAXIMUM LIKELIHOOD ESTIMATION (MLE))$L(\theta)$. The resulting parameter, which is sought by searching the multidimensional parameter space, is called **the MLE estimate**, denoted by

$$\theta MLE = (\theta_1 MLE, ..., \theta_k MLE).$$

#### 1.2.1

Let **p** equal the probability of success in a sequence of Bernoulli trials or the proportion of the large population with a certain characteristic. The method of moments estimate for **p** is relative frequency of success (having that characteristic). It will be shown below that the maximum likelihood estimate for **p** is also the relative frequency of success.

Suppose that **X** is $b(1,p)$ so that the p.d.f. of **X** is

$$f(x;p) = p^x (1-p)^{1-x}, x = 0, 1, 0 \leq p \leq 1.$$

Sometimes is written

$$p \in \Omega = [p : 0 \leq p \leq 1],$$

where $\Omega$ is used to represent parameter space, that is, the space of all possible values of the parameter. A random sample $X_1, X_2, ..., X_n$ is taken, and the problem is to find an estimator $\mathrm{u}(\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_\mathrm{n})$ such that $\mathrm{u}(x_1, x_2, ..., x_\mathrm{n})$ is a good point estimate of **p**, where $x_1, x_2, ..., x_\mathrm{n}$ are the observed values of the random sample. Now the probability that $X_1, X_2, ..., X_n$ takes the particular values is

$$P(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p \sum x_i (1-p)^{n-\sum x_i},$$

which is the joint p.d.f. of $X_1, X_2, ..., X_n$ evaluated at the observed values. One reasonable way to proceed towards finding a good estimate of **p** is to regard this probability (or joint p.d.f.) as a function of **p** and find the value of **p** that maximizes it. That is, find the **p** value most likely to have produced these sample values. The joint p.d.f., when regarded as a function of **p**, is frequently called **the likelihood function**. Thus here the likelihood function is:

$$L(p) = L(p; x_1, x_2, ..., x_n) = f(x_1;p) f(x_2;p) \cdots f(x_n;p) = p^{\sum x_i} (1-p)^{n-\sum x_i}, 0 \leq p \leq 1.$$

To find the value of **p** that maximizes $L(p)$ first take its derivative for $0 < p < 1$ :

$$\frac{dL(p)}{dp} = \left(\sum x_i\right) p^{n-\sum x_i} (1-p)^{n-\sum x_i} - \left(n - \sum x_i\right) p^{\sum x_i} (1-p)^{n-\sum x_i - 1}.$$

Setting this first derivative equal to zero gives

$$p^{\sum x_i} (1-p)^{n-\sum x_i} \left[\frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p}\right] = 0.$$

Since $0 < p < 1$, this equals zero when

$$\frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0.$$

Or, equivalently,

$$p = \frac{\sum x_i}{n} = \overline{x}.$$

The corresponding statistics, namely $\sum X_i/n = \overline{X}$, is called **the maximum likelihood estimator** and is denoted by $\hat{p}$ ,that is,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

When finding a maximum likelihood estimator, it is often easier to find the value of parameter that minimizes the natural logarithm of the likelihood function rather than the value of the parameter that minimizes the likelihood function itself. Because the natural logarithm function is an increasing function, the solution will be the same. To see this, the example which was considered above gives for $0 < p < 1$,

$$\ln L(p) = \left(\sum_{i=1}^{n} x_i\right) \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p).$$

To find the maximum, set the first derivative equal to zero to obtain

$$\frac{d[\ln L(p)]}{dp} = \left(\sum_{i=1}^{n} x_i\right)\left(\frac{1}{p}\right) + \left(n - \sum_{i=1}^{n} x_i\right)\left(\frac{-1}{1-p}\right) = 0,$$

which is the same as previous equation. Thus the solution is $p = \overline{x}$ and the maximum likelihood estimator for **p** is $\hat{p} = \overline{X}$.

Motivated by the preceding illustration, the formal definition of maximum likelihood estimators is presented. This definition is used in both the discrete and continuous cases. In many practical cases, these estimators (and estimates) are unique. For many applications there is just one unknown parameter. In this case the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

NOTE: Maximum Likelihood Estimation - Examples [1]

[1] "Maximum Likelihood Estimation - Examples": Section MAXIMUM LIKELIHOOD ESTIMATION - EXAMPLES <http://cnx.org/content/m13500/latest/#sec_1>