Maximum Likelihood

Rahul Telang

Maximum Likelihood

- ML is one of the most common methods used to estimate parameters. This is particularly useful when one can deploy least square approach (linear regression)
- ML is a technique to find the most likely value of parameters β in the population,
  - Data we observe.
  - Given the distribution we assume.
- Essentially, it finds the parameters of distribution such that the distribution best fit the data.
- For OLS, ML will produce the same results as a regression, but it is a general technique that can be applied to many different types of models

Likelihood Functions

- The distribution of data is described by a “probability density function” (PDF)
  - PDF outlines the relative probability, or likelihood, to observe a certain value given the parameters of the distribution
- We usually think of a function with its parameters and generate data.

  $Y = F(x;\theta)$
- For different values of x, one can generate Y.
  - Think of $F(\theta)$ as a function or a model that you have in mind with parameter θ.
- Now consider, if you see data Y and want to guess what value of θ would potentially generate this data?

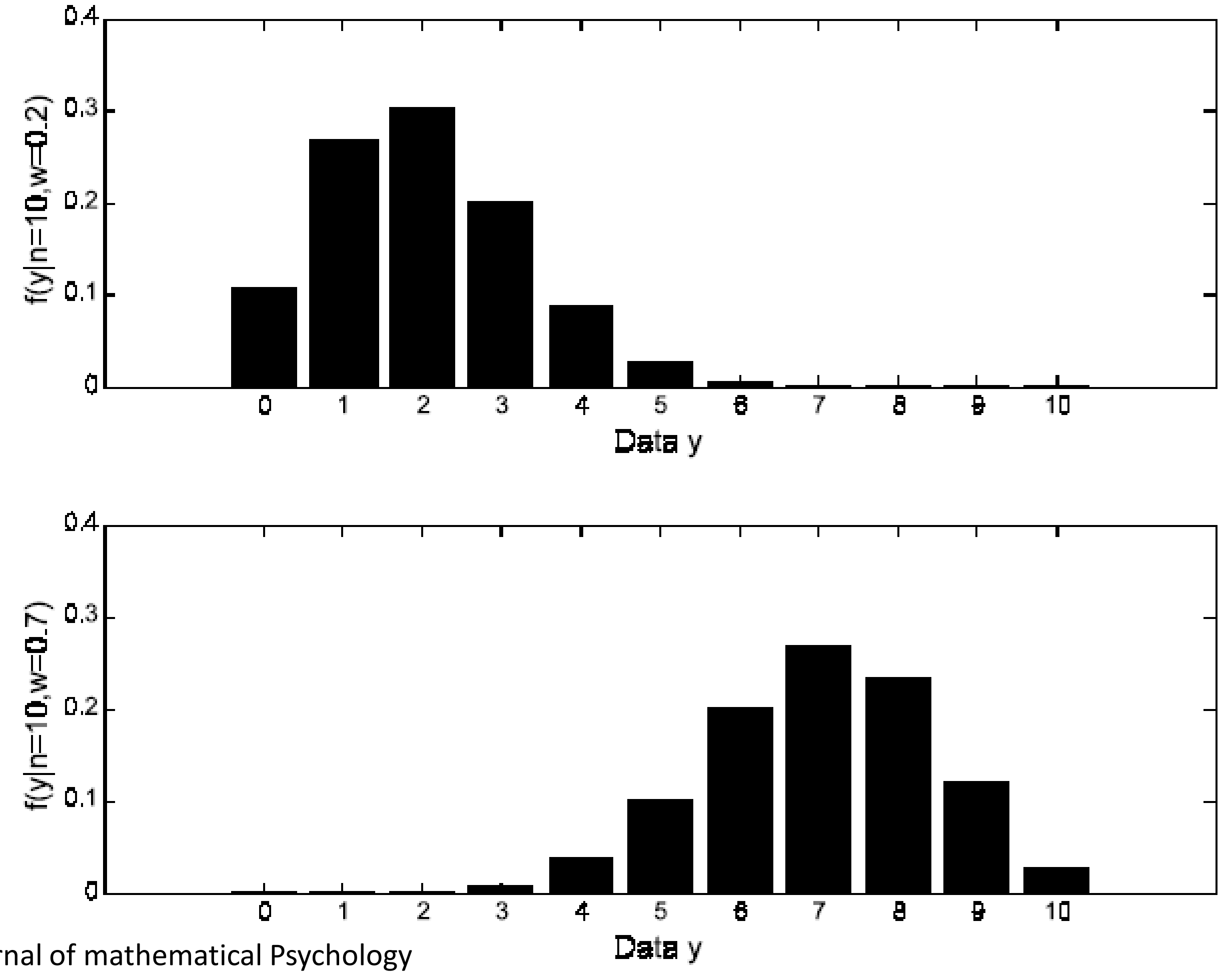
Example

- For example, suppose you conduct a Bernoulli trial (say toss a coin) and probability of success is given by a parameter w. If you conduct n trials (lets fix n=10), what is the probability of observing y number of success?
  - In short what is the distribution of y, given the parameter w?
  - It's a Binomial distribution

$$f(y/w, n=10) = \frac{n!}{y!(10-y)!} w^y (1-w)^{n-y}$$

Plot of pdf

- We can plot pdf for different values of “w”.



Src:Myung, 2003, journal of mathematical Psychology

Another way..

- Another way to think about this is
  - We are observing **f(y/w).** If we change w, f(y/w) will change too. What is the value of “w” that makes it mostly likely to generate f(y/w) that we are observing.
  - For example, in the previous slides we saw two distributions depending on two different value of w.
- Put another way, we usually can find a distribution $f(y/\theta)$ if we know θ.
  - But when we are “estimating” parameters, we are interested in inferring θ with the data we have in hand.
  - Or, we want to know $f(\theta/y)$.
- If the data is distributed in the first plot, low value of w is the most likely outcome; for the second plot it will be high value of w.

Likelihood Function

- The likelihood principle works on the same idea.
  - It looks at the data and asks, what “parameter” value can generate this data. Or, what is the probability of observing this data given the parameter?
- How to write this down formally?
- We are interested in knowing g(w/y) when we are given f(y/w). Using Bayes rule, one can show that

$$\mathrm{g}(w/y) \propto f(y/w)$$

- This is also called likelihood function such that
  - $L(w/y) \propto \mathrm{g}(w/y) \propto f(y/w)$
- Likelihood function is proportional to the density function. Thus
  - For discrete distribution, likelihood is simply probability p(x/θ) = observing data.
- The best estimator for “w” can be estimated by maximizing $L(.)$ for a sample of n observations.

Put it to use now

- Our data has “n” observation or n data points for a given problem.
- Since we have n observations, we can write the distribution f(y/w) as follows
  - $f(y_1, y_2, y_{3\ldots\ldots}y_n /w)$
  - Notice this is a joint distribution
- The principle of maximum likelihood says that the “most appropriate” w can be recovered by maximizing this function.
- This function is also called the “likelihood” function. So.
  - $L(w/y) = f(y_1, y_2 \ldots . y_n/\mathrm{w})$
- Since data comes from “n” <u>independent</u> samples, we can multiple each data points. So

$$L(w/\, y_1, y_2, \ldots..y_n) = f(y_1, y_2, \ldots..y_n \,/\, w) = \prod_n f(y_n \,/\, w)$$

Maximum Likelihood (ML)

- ML then requires that

$$\max_w L(w / y_1, y_2, ...y_n) = \max_w \prod_n f(y_n / w)$$

- Taking logs would simplify this enormously (multiplication would turn to addition) and hence

$$\max_w LogL = \max_w \sum_n Log\left(f(y_n / w) \right)$$

- Thus, we usually refer to this as maximum Log likelihood.
- Since we are maximizing a function, it must be that

$$\frac{\partial LL}{\partial w} = 0; \qquad \frac{\partial^2 LL}{\partial^2 w} < 0$$

- Solving the first should readily provide us the value of “w” that maximizes the function. The second ensures that we find the maxima and not minima.

Key steps for applying ML

- What is the likelihood of observing the data one wants to model?
- That probability is given by the pdf – f(y). The likelihood is simply this pdf. So, likelihood of observing data is – f(y)
- Since we have n independent sample – the likelihood of observing the sample is simply - $\prod_{i=1}^{n} f_i(y)$
- One maximizes this expression to recover the estimates.

Maximum Likelihood (ML)

- In practice finding a maxima of a function cannot be solved analytically, and different algorithms have been developed and extensively used in practice.
  - There is a whole stream of work in optimization which is devoted to developing robust algorithm.
  - Many of these functions are available in most packages.
- Data analytics almost always requires that predictions are close to the actual data.
  - The error is minimized.
- Majority of data estimation are solved using likelihood technique.
  - Most ML methods use some maximization algorithm to find optimal solution.