Lecture 7: Maximum likelihood estimators and estimation topics

Jason Mezey
jgm45@cornell.edu
Sept. 16, 2014 (T) 8:40-9:55
Announcements

• Homework #1 is graded (!!) and key is posted
  • In Ithaca, email Amanda to get your homework
  • In NYC, email or see Jin

• Reminder: Homework #2 is due 11:59 PM, Weds. (tomorrow!) - turn in separate answers and code (!!)

• Graded Homework #3 will be available tomorrow as well (it will be due Tues.!!)

• Amanda will hold office hours today 3-5PM in Ithaca

• Jason will hold office hours TOMORROW (=Weds. not Thurs., this week only!) in Ithaca and NYC at the same locations as always
Summary of lecture 7

- Last lecture, we discussed estimators and began our discussion of maximum likelihood estimators (MLE’s)
- Today, we will continue our discussion of MLE’s
Conceptual Overview

System

Question

Sample

Inference

Prob. Models

Statistics

Assumptions
This concept is often introduced to us as a mathematical operator that takes an input and produces an output. Before we consider the specifics of how we define a function, let's consider the intuitive definition of a function:

$$\hat{\theta} \quad \theta \in \Theta$$

$$= \quad T(X), \ Pr(T(X)|\theta)$$

$$[X_1 = x_1, \ldots, X_n = x_n], \ Pr(X_1 = x_1, \ldots, X_n = x_n|\theta)$$

Sample of size $n$  

$$X = x, \ Pr(X|\theta)$$

Sampling Distribution

$X$  

Random Variable

$\mathcal{E}(\Omega)$  

$X(\Omega)$  

$Pr(\mathcal{F})$

Experiment  

$\Omega$  

$\mathcal{F}$

Disjoint Sets

Intersection

Union

Empty Set

Disjoint Sets

5 Probability Functions

For example, we can have the function $Pr$ which we could have written $\theta u_15v$.
Review of essential concepts

- **System** - a process, an object, etc. which we would like to know something about

- **Experiment** - a manipulation or measurement of a system that produces an outcome we can observe

- **Experimental trial** - one instance of an experiment

- **Sample Space** ($\Omega$) - a set comprising all possible outcomes associated with an experiment

- **Sigma Algebra** ($\mathcal{F}$) - a collection of all events of a sample space

- **Probability function** $\text{Pr}(\mathcal{F})$ - maps sigma algebra to the reals (Axioms of probability!)

- **Random variable / vector** $X(\Omega)$ - a real valued function on the sample space

- **Parameterized probability model** - a family of probability models indexed by a constant $\theta$ (=parameter) belonging to a family of probability models
Review of essential concepts II

- **Inference** - the process of reaching a conclusion about the true probability distribution (from an assumed family of probability distributions indexed by parameters) on the basis of a sample

- **Sample** - repeated observations of a random variable $X$, generated by experimental trials (= random vector!)

- **Sampling distribution** - probability distribution on the sample random vector (usually assume i.i.d.!!)

$$Pr(X_1 = x_1) = Pr(X_2 = x_2) = \ldots = Pr(X_n = x_n)$$

$$Pr(X = x) = Pr(X_1 = x_1)Pr(X_2 = x_2)\ldots Pr(X_n = x_n)$$

- **Statistic** $T(X)$ - a function on a sample

- **Statistic sampling distribution** - probability distribution on the statistic

- **Estimator** $\hat{T}(x) = \hat{\theta}$ - a statistic defined to return a value that represents our best evidence for being the true value of a parameter
Review of maximum likelihood estimators (MLE)

• We will generally consider maximum likelihood estimators (MLE), which is one class of estimators, in this course.

• The critical point to remember is that an MLE is just an estimator (a function on a sample!!),

• i.e. it takes a sample in, and produces a number as an output that is our estimate of the true parameter value.

• These estimators also have sampling distributions just like any other statistic!

• The structure of this particular estimator / statistic is complicated but just keep this big picture in mind.
Likelihood I

- To introduce MLE’s we first need the concept of *likelihood*
- Recall that a probability distribution (of a r.v. or for our purposes now, a statistic) has fixed constants in the formula called *parameters*
- The function is therefore takes different inputs of the statistic, where different sample inputs produce different outputs
- For example, for a normally distributed random variable

\[
Pr(X = x | \mu, \sigma^2) = f_X(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

\[
Pr(\mu, \sigma^2 | X = x) = L(\mu, \sigma^2 | X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

- **Likelihood** - a probability function which we consider to be a function of the parameters for a fixed sample
- Likelihoods have the structure of probability functions but they are NOT probability functions, e.g. they are functions of parameters and they are used for estimation
- Intuitively, a probability function represents the frequency at which a specific realization of \(T(X)\) will occur, while a likelihood is our supposition that the true value of the parameter (that determines the probability of the values of \(X\) and \(T(X)!\)) is a specific value
Normal model example I

- As an example, for our heights experiment / identity random variable, the (marginal) probability of a single observation in our sample is $x_i$ is:

$$Pr(X_i = x_i | \mu, \sigma^2) = f_{X_i}(x_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- The joint probability distribution of the entire sample of $n$ observations is a multivariate (n-variate) normal distribution

- Note that for an i.i.d. sample, we may use the property of independence

$$Pr(X = x) = Pr(X_1 = x_1)Pr(X_2 = x_2)...Pr(X_n = x_n)$$

to write pdf of this entire sample as follow:

$$P(X = x | \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- The likelihood is therefore:

$$L(\mu, \sigma^2 | X = x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$
Normal model example II

- Let’s consider a sample of size $n=10$ generated under a standard normal, i.e.
  $$X_i \sim N(\mu = 0, \sigma^2 = 1)$$

- So what does the likelihood for this sample “look” like? It is actually a 3-D plot where the $x$ and $y$ axes are $\mu$ and $\sigma^2$ and the z-axis is the likelihood:
  $$L(\mu, \sigma^2|X = x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- Since this makes it tough to see what is going on, let’s set just look at the marginal likelihood for $\sigma^2 = 1$ when using the sample above:

[Image of a 2-D plot showing normal likelihood for $\mu$ and $X$ with $n=10, \sigma=1$.]
Introduction to MLE’s

- A maximum likelihood estimator (MLE) has the following definition:
  \[ MLE(\hat{\theta}) = \hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta|x) \]

- Again, recall that this statistic still takes in a sample and outputs a value that is our estimator (!!)

- Sometimes these estimators have nice forms (equations) that we can write out and sometimes they do not

- For example the maximum likelihood estimator of our single coin example:
  \[ MLE(\hat{p}) = \frac{1}{n} \sum_{i=1}^{n} x_i \]

- And for our heights example:
  \[ MLE(\hat{\mu}) = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \]
  \[ MLE(\hat{\sigma}^2) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]
Getting to the MLE

- To use a likelihood function to extract the MLE, we have to find the maximum of the likelihood function $L(\theta|x)$ for our observed sample.
- To do this, we take the derivative of the likelihood function and set it equal to zero (why?)
- Note that in practice, before we take the derivative and set the function equal to zero, we often transform the likelihood by the natural log ($ln$) to produce the log-likelihood:

$$l(\theta|x) = ln[L(\theta|x)]$$

- We do this because the likelihood and the log-likelihood have the same maximum and because it is often easier to work with the log-likelihood.
- Also note that the domain of the natural log function is limited to $[0, \infty)$ but likelihoods are never negative (consider the structure of probability!)
MLE under a normal model

- Recall that the likelihood for a sample of size $n$ generated under a normal model has the following likelihood:

$$L(\mu, \sigma^2|X = x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - \mu)^2/2\sigma^2}$$

- By remembering the properties of $\ln$, we can derive the log-likelihood for this model:

$$l(\mu, \sigma^2|X = x)) = -n\ln(\sigma) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i}^{n} (x_i - \mu)^2$$

1. $\ln\frac{1}{a} = -\ln(a)$
2. $\ln(a^2) = 2\ln(a)$
3. $\ln(ab) = \ln(a) + \ln(b)$
4. $\ln(e^a) = a$
5. $e^a e^b = e^{a+b}$

- To obtain the maximum of this function with respect to $\mu$ we can then take the partial (!!) derivative with respect to $\mu$ and set this equal to zero, then solve (this is the MLE!):

$$\frac{\partial l(\theta|X = x)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i}^{n} (x_i - \mu) = 0$$

$$\text{MLE}(\hat{\mu}) = \frac{1}{n} \sum_{i}^{n} x_i$$
MLE under a normal model II

- How about the $\sigma^2$? Use the same approach:

$$l(\mu, \sigma^2 | X = x) = -n \ln(\sigma) - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i} (x_i - \mu)^2$$

$$\frac{\partial l(\theta | X = x)}{\partial \sigma^2} = 0$$

$$MLE(\hat{\sigma}^2) = \frac{1}{n} \sum_{i} (x_i - \bar{x})^2$$

- This equation will give us the maximum of the log-likelihood with respect to this parameter

- Will this produce the true value of $\sigma^2$ (??)
A discrete example I

- As an example, for our coin flip / number of tails random variable
- The probability distribution of one sample is:
  \[ Pr(x_i|p) = p^{x_i} (1 - p)^{1-x_i} \]
- The joint probability distribution of an i.i.d sample of size n is is an n-variate Bernoulli
  \[ Pr(x|p) = \prod_{i=1}^{n} p^{x_i} (1 - p)^{1-x_i} \]
- A TRICK (!!): it turns out that we can get the same MLE of p for this model by considering \( x = \text{total number of tails in the entire sample} \):
  \[ Pr(x|p) = \binom{n}{x} p^x (1 - p)^{n-x} \]
- Such that we can consider the following likelihood:
  \[ L(p|X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \]
A discrete example II

- To find the MLE, we will use the same approach by taking the log-likelihood:

\[ L(p | X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \]

\[ l(p | X = x) = \ln \left( \binom{n}{x} \right) + x \ln(p) + (n - x) \ln(1 - p) \]

- Taking the first derivative set to zero, then solve (again \( x = \text{number tails!} \))

\[ \frac{\partial l(p | X = x)}{\partial p} = \frac{x}{p} - \frac{n - x}{1 - p} \]

\[ \text{MLE}(\hat{p}) = \frac{x}{n} \]

- Question: in general, how do we know this is a maximum?

- We can check by looking at the second derivative and making sure that it is always negative (why?):

\[ \frac{\partial^2 l(p | X = x)}{\partial p^2} = -\frac{x}{p^2} + \frac{x - n}{(1 - p)^2} \]
Last general comments (for now) on maximum likelihood estimators (MLE)

• In general, *maximum likelihood estimators* (MLE) are at the core of most standard “parametric” estimation and hypothesis testing (stay tuned!) that you will do in basic statistical analysis.

• Both likelihood and MLE’s have many useful theoretical and practical properties (i.e. no surprise they play a central role) although we will not have time to discuss them in detail in this course (e.g. likelihood has strong connections to the concept of sufficiency, likelihood principal, etc., MLE have nice properties as estimators, ways of obtaining the MLE, etc.

• Again, for this course, the critical point to keep in mind is that when you calculate an MLE, you are just calculating a statistic (estimator!)
Brief Introduction: Properties of estimators I

• Remember (!!) for all the complexity in thinking about, deriving, etc. MLE’s these are still just estimators (!!), i.e. they are statistics that take a sample as input and output a value that we consider an estimate of our parameter.

• MLE in general have nice properties (and we will largely use them in this class!), but there are many other estimators that we could use.

• This is because there is no “perfect” estimator and each estimator that we can define has different properties, some of which are desirable, some are less desirable.

• In general, we do try to use estimators that have “good” properties based on well defined criteria.

• In this class, we will briefly consider two: *unbiasedness* and *consistency*.
Properties of estimators II

- We measure the bias of an estimator as follows (where an unbiased estimator has a bias of zero):
  \[ \text{Bias}(\hat{\theta}) = E\hat{\theta} - \theta \]

- We consider an estimator to be consistent if it has the following property
  \[ \lim_{n \to \infty} \Pr(|\hat{\theta} - \theta| < \epsilon) = 1 \]

- Note that one can have an estimator that is consistent but not unbiased (and vice versa!)

- As an example of the former, the following MLE is biased but consistent
  \[ \text{MLE}(\hat{\sigma}^2) = \frac{1}{n} \sum_{i} (x_i - \bar{x})^2 \]

- An unbiased estimator of this parameter is the following:
  \[ \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i} (x_i - \bar{x})^2 \]
That’s it for today

• Next lecture, we will continue our discussion of hypothesis testing

• This will complete our introduction / review of basic concepts in probability and statistics!