Quantitative Genomics and Genetics
BTRY 4830/6830; PBSB.5201.01

Lecture 8: MLEs, estimation topic and hypothesis testing intro

Jason Mezey
jgm45@cornell.edu
Feb. 22, 2018 (Th) 8:40-9:55
Announcements

• Reminder: Computer lab and office hours today at the normal times!
Summary of lecture 8

• Last lecture, we discussed likelihood

• Today we will introduce Maximum Likelihood Estimators (MLEs)

• We will also (briefly) consider estimation topics (e.g., bias, the concept of confidence intervals)

• Time permitting, we will also begin our discussion of hypothesis testing
Conceptual Overview

System

Question

Experiment

Sample

Inference

Prob. Models

Statistics

Assumptions
Estimators

Estimator: \( T(x) = \hat{\theta} \)

\[ \begin{align*}
X_1 &= x_1, \ldots, X_n = x_n \\
X &= x \\
\mathcal{X} &= X(\omega), \omega \in \Omega \\
\mathcal{F} &= \text{(Sigma Algebra)} \\
\Omega &= \text{(Sample Space)}
\end{align*} \]

Estimator Sampling Distribution: \( P_r(T(X)|\theta), \theta \in \Theta \)

\[ \begin{align*}
\mathcal{X} &= \text{Random Variable} \\
\mathcal{F} &= P_r(\mathcal{F}) \\
\Omega &= \text{Experiment} \\
\mathcal{X} &= P_r(\mathcal{X}) \\
\Omega &= P_r(\Omega) \\
\mathcal{F} &= P_r(\mathcal{F}) \\
\Theta &= \text{Parameter Space} \\
\mathcal{X} &= P_r(\mathcal{X}) \\
\Omega &= P_r(\Omega) \\
\mathcal{F} &= P_r(\mathcal{F}) \\
\Theta &= \text{Parameter Space}
\end{align*} \]
Review of essential concepts

- **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

- **System, Experiment, Experimental Trial, Sample Space, Sigma Algebra, Probability Measure, Random Vector, Parameterized Probability Model, Sample, Sampling Distribution, Statistic, Statistic Sampling Distribution, Estimator, Estimator Sampling distribution**
Review of estimators

- **Estimator** - a statistic defined to return a value that represents our best evidence for being the true value of a parameter.

- In such a case, our statistic is an estimator of the parameter: \( T(x) = \hat{\theta} \)

- Note that ANY statistic on a sample can in theory be an estimator.

- However, we generally define estimators (=statistics) in such a way that it returns a reasonable or “good” estimator of the true parameter value under a variety of conditions.

- How we assess how “good” an estimator depends on our criteria for assessing “good” and our underlying assumptions.
Review of maximum likelihood estimator (MLE) concept

• We will generally consider maximum likelihood estimators (MLE) in this course.

• Now, MLE’s are very confusing when initially encountered...

• However, 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.
Review Likelihood I

- **Likelihood** - (not technical def yet!) has the form of a probability function which we consider to be a function of the parameters NOT the sample.
- Note that likelihoods are NOT probability functions, i.e. they need not conform to the axioms of probability (!!)
- They have the appealing property that for an i.i.d. sample,

\[ L(\theta|\mathbf{x}) = L(\theta|x_1)L(\theta|x_2)\ldots L(\theta|x_n) \]

- They have other appealing properties, including they are sufficient statistics, the invariance principal, etc.

\[ X_1 = 0. \]
Review Likelihood 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:

![Normal likelihood: n=10, sigma=1](image)
• **Likelihood** - a function with the form of a probability function which we consider to be a function of the parameters \( \theta \) for a fixed the sample \([X = x] \)

• The form of a likelihood is therefore the sampling distribution (the probability distribution!) of the i.i.d sample but there are (at least) three major differences:
  
  • We have parameter values as input and the sample *we have observed* acts as a parameter
  
  • The likelihood function does not operate as a probability function (they can violate the axioms of probability)
  
  • For continuous cases, we can interpret the likelihood of a parameter (or combination of parameters) as the likelihood of the point
Introduction to MLE’s

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

- Recall that this statistic still takes in a sample and outputs a value that is our estimator (!!) Note that likelihoods are NOT probability functions, i.e. they need not conform to the axioms of probability (!!)

- Sometimes these estimators have nice forms (equations) that we can write out

- For example the maximum likelihood estimator when considering a sample for our single coin example / number of tails is:
  \[ MLE(\hat{p}) = \frac{1}{n} \sum_{i=1}^{n} x_i \]

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

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

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

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

- 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:

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

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

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

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

- taking the first derivative set to zero, then solve (again \(x=\)number tails!)
  \[ \frac{\partial l(p|X = x)}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p} \]
  \[ MLE(\hat{p}) = \frac{x}{n} \]

- 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):

\[ \text{Bias}(\hat{\theta}) = \mathbb{E}\hat{\theta} - \theta \]

• 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

\[ \text{MLE}(\sigma^2) = \frac{1}{n} \sum_{i}^{n} (x_i - \bar{x})^2 \]

• An unbiased estimator of this parameter is the following:

\[ \hat{\sigma}^2 = \frac{1}{n - 1} \sum_{i}^{n} (x_i - \bar{x})^2 \]
Confidence intervals I

- For the estimation framework we have considered thus far, our goal was to define an estimator that provides a “reasonable guess given the sample” of the true value of the parameter.

- This is called “point” estimation since the true parameter has a single value (i.e. it is a point).

- We could also estimate an interval, where our goal is to say something about the chances that the true parameter (the point) would fall in the interval.

- **confidence interval** (CI) - an estimate of an interval defined such that if it were estimated individually for an infinite number of samples, a specific percentage of the estimated intervals would contain the true parameter value.

- Don’t worry if this concept seems confusing (it is!) let’s first consider an example and then discuss some basics.
Confidence intervals II

- As an example, assume the standard normal r.v. $X \sim N(0, 1)$ correctly describes our sampling distribution if we were to produce 50 independent samples, each of size $n=10$ and we were to estimate a CI for each one, we would expect to get the following:
Confidence intervals III

- A CI is therefore calculated from a sample (and reflects uncertainty!)

- A CI is an estimate of an interval, as opposed to an estimate of a parameter, which is a point estimate (more technically, the CI is an estimate of the endpoints of the interval)

- This estimated interval of a CI (generally) includes the estimate of the parameter in the “middle”

- In general, a CI provides a measure of “confidence” in the sense that the smaller the interval, the more “confidence” we have in our estimate (if this seems circular, it is mean to be!)

- In general, we can make the CI smaller with a larger sample size \( n \) and by decreasing the probability that the interval contains the true parameter value, i.e. a 95% CI is smaller than a 99% CI

- NOTE THAT A 95% CI estimated from one sample does not contain the true parameter value with a probability of 0.95 (!!!) - the definition of a CI says if we performed an infinite number of samples, and calculated the CI for each, then 95% of these intervals would contain the true parameter value (strange?)
Estimation and Hypothesis Testing

- Thus far we have been considering a “type” of inference, estimation, where we are interested in determining the actual value of a parameter.
- We could ask another question, and consider whether the parameter is NOT a particular value.
- This is another “type” of inference called hypothesis testing.
- We will use hypothesis testing extensively in this course.
Hypothesis testing I

• To build this framework, we need to start with a definition of hypothesis

• **Hypothesis** - an assumption about a parameter

• More specifically, we are going to start our discussion with a *null hypothesis*, which states that a parameter takes a specific value, i.e. a constant

\[ H_0 : \theta = c \]

• For example, for our height experiment / identity random variable, we have \( Pr(X|\theta) \sim N(\mu, \sigma^2) \) and we could consider the following null hypothesis:

\[ H_0 : \mu = 0 \]
That’s it for today

• Next week: hypothesis testing!