Lecture 9: MLEs, estimation topic and hypothesis testing intro

Jason Mezey
jgm45@cornell.edu
Feb. 18, 2020 (T) 8:40-9:55
Announcements

• Homework Announcements:
  • We do accept late homeworks… (glitch with CMS = homework #2 late submissions not accept) - this will be fixed for next homework submission (!!)
  • Homework #3 available Thurs. or Fri. this week - and will be due 11:59PM March 3 (!!)
  • Class cancelled Thurs. (Feb. 20) and no class Tues., Feb. 25 (=Cornell Winter Break) - please make a note of it (!!)
  • Office hours cancelled Mon. 24 but I will work on scheduling extra for that week + regular office hours will resume March 2 as per normal
  • THERE WILL BE COMPUTER LAB on Thurs. / Fri. - Feb. 20/21 as per normal (!!)
Summary of lecture 9

• Last lecture, we discussed likelihood introduced Maximum Likelihood Estimators (MLEs)

• Today, we will (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
- Sample
- Assumptions
- Inference
- Prob. Models
- Statistics

Diagram:

1. System -> Question
2. Question -> Sample
3. Sample -> Assumptions
4. Assumptions -> Inference
5. Inference -> Prob. Models
Estimators

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

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

$\{X_1 = x_1, \ldots, X_n = x_n\}$

$X = x$

$Pr(X)$

Random Variable

$\mathcal{X}$

$X(\omega), \omega \in \Omega$

$Pr(\mathcal{F})$

Experiment

$\Omega$

$(\text{Sample Space})$

$\mathcal{F}$

$(\text{Sigma Algebra})$
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: \( \hat{T}(x) = \hat{\theta} \)

- 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.
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:

- **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 $[X = x]$

- 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
Review: 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:

\[ MLE(\hat{\mu}) = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad MLE(\hat{\sigma}^2) = \frac{1}{n} \sum_{i} (x_i - \bar{x})^2 \]
Review: MLE under a normal model I

- 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. \( \ln \frac{1}{a} = -\ln(a) \)
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!):

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

\[
MLE(\hat{\mu}) = \frac{1}{n} \sum_{i}^{n} x_i
\]
Review: 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:
  \[ L(p|X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \]
  \[ l(p|X = x) = \ln\binom{n}{x} + x\ln(p) + (n - x)\ln(1 - p) \]

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

- 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

• 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

• 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 meant 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?)
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**
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.
Estimators

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

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

$[X_1 = x_1, ..., X_n = x_n]$

$Pr([X_1 = x_1, ..., X_n = x_n])$

$X = x$

$Pr(X)$

Random Variable

$\mathcal{X}$

$X(\omega), \omega \in \Omega$

$Pr(\mathcal{F})$

Experiment

$\Omega$

(Sample Space)

$\mathcal{F}$

(Sigma Algebra)
Hypothesis Tests

Hypothesis: \( T(x), H_0 : \theta = c \)

\[
[X_1 = x_1, \ldots, X_n = x_n]
\]

\( X = x \)

\( Pr(X) \)

\( Pr(F) \)

\( \Omega \) (Sample Space)

\( \mathcal{F} \) (Sigma Algebra)

\( X(\omega), \omega \in \Omega \)

Random Variable

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

\( Pr(T(X)|\theta), \theta \in \Theta \)

Statistical Sampling Distribution:

\( X \)

Experiment

\( \mathcal{X} \)
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:

\[
H_0 : \mu = 0
\]
Hypothesis testing II

• Our goal in hypothesis testing is to use a sample to reach a conclusion about the null hypothesis.

• To do this, just as in estimation, we will make use of a statistic (a function on the sample), where recall we know the sampling distribution (the probability distribution) of this statistic.

• More specifically, we will consider the probability distribution of this statistic, assuming that the null hypothesis is true:

\[ \Pr(T(X = x | \theta = c)) \]

• Note that this means we have a probability distribution of the statistic given the null hypothesis!!

• We will use this distribution to construct a \textit{p-value}.
Hypothesis testing III

- As example, consider our height experiment (reals as sample space) / identity random variable \( X \) / normal probability model \( \theta = [\mu, \sigma^2] \) / sample \( n=1 \) (of one height measurement) / identity statistic \( T(x) = x \) (takes the height measured height)

- Let's assume that \( \sigma^2 = 1 \) and say we are interested in testing the following null hypothesis \( H_0 : \mu = 5.5 \) such that we have the following probability distribution of the statistic under the null hypothesis:
p-value l

- We quantify our intuition as to whether we would have observed the value of our statistics given the null is true with a *p-value*.

- **p-value** - the probability of obtaining a value of a statistic $T(x)$, or more extreme, conditional on $H_0$ being true.

- Formally, we can express this as follows:

  $$pval = Pr(|T(x)| \geq t | H_0 : \theta = c)$$

- Note that a p-value is a function on a statistic (!) that takes the value of a statistic as input and produces a p-value as output in the range $[0, 1]$:

  $$pval(T(x)) : T(x) \rightarrow [0, 1]$$
p-value II

- As an intuitive example, let’s consider a continuous sample space experiment / identify r.v. / normal family / \( n=1 \) sample / identity statistic, i.e. \( T(x) = x \)

- Assume we know \( \sigma^2 = 1 \) (is this realistic?), let’s say we are interested in testing the null hypothesis \( H_0: \mu = 0 \) and let’s say that we assume that if we are wrong the value of \( \mu \) will be greater than zero (why?)
p-value III

- Same example: let's consider a continuous sample space experiment / identify r.v. / normal family / n=1 sample / identity statistic, i.e. $T(X) = X$ / assume we know $\sigma^2 = 1$ / we test the null hypothesis $H_0 : \mu = 0$ and let's assume that if we are wrong the value of $\mu$ could be in either direction (again, why?)
p-value IV

- More technically a p-value is determined not just by the probability of the statistic given the null hypothesis is true, but also whether we are considering a “one-sided” or “two-sided” test.

- For a one-sided test (towards positive values), the p-value is:

\[
pval(T(x)) = \int_{\max(T(X))}^{\infty} Pr(T(x)\mid \theta = c) \, dT(x)
\]

- For a two-sided test, the p-value is:

\[
pval(T(x)) = \int_{-\infty}^{-|T(x) - \text{median}(T(X))|} Pr(T(x)\mid \theta = c) \, dT(x) + \int_{|T(x) - \text{median}(T(X))|}^{\infty} Pr(T(x)\mid \theta = c) \, dT(x)
\]

\[
pval(T(x)) = \sum_{\min(T(X))}^{-|T(x) - \text{median}(T(X))|} Pr(T(x)\mid \theta = c) + \sum_{|T(x) - \text{median}(T(X))|}^{\max(T(X))} Pr(T(x)\mid \theta = c)
\]
That’s it for today

- Week after next: last hypothesis testing and start of genetic analysis!