Lecture 6: Introduction to estimators and maximum likelihood estimators (MLE)

Jason Mezey
jgm45@cornell.edu
Feb. 18, 2016 (Th) 8:40-9:55
• Homework #3 will be available TOMORROW (Fri., Feb. 19) and will be DUE in one week (11:59PM on Fri., Feb. 26)

• Homework #1 available in computer lab today

• Reminder: computer lab today (!!), office hours, etc.
Summary of lecture 6

• Last lecture, we discussed parameterized probability models, introduced the general concept of inference, defined samples and statistics

• Today, we will continue our introduction to the concept of estimators and introduce maximum likelihood estimators (mle)
**Review**

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

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

- **Sigma Algebra or Sigma Field** ($\mathcal{F}$) - a collection of events (subsets) of the sample space of interest

- **Probability Measure (=Function)** - maps a Sigma Algebra of a sample to a subset of the reals

- **Random Vector (Variable)** - (measurable) function on a sample space

- **Probability (or Cumulative) Distribution Function** - functions that describes the probability distribution of a random variable / vector

- **Sample** - results of one or more experimental trials: $\mathbf{X} = [x_1, x_2, \ldots, x_n]$

- **Sampling Distribution** - probability distribution function of the sample (represents the probability of every possible sample under given assumptions, e.g., iid): $Pr([X_1, X_2, \ldots, X_n])$

- **Statistic** - function taking any sample outcome to a number: $T(\mathbf{x}) = T([x_1, x_2, \ldots, x_n]) = t$

- **Statistic Sampling Distribution** - probability function of the statistic (represents the probability the statistic can when considering every possible sample): $Pr(T(\mathbf{X}))$
Statistics

$T(x)$  
$Pr(T(X))$

Statistic  
Statistic Sampling Distribution

Sample of size $n$  
Sampling Distribution

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

$X = x$, $Pr(X)$

Random Variable

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

$Pr(\mathcal{F})$

Experiment  
$\Omega$  
$\mathcal{F}$
Event Conditional Probability Mass Function (PMF)

In probability theory, we often consider the behavior of random processes, which are described by random variables. A random variable, denoted by $X$, is a function that maps outcomes from an experiment's sample space, $\mathcal{X}$, to the real numbers. The sample space represents the possible outcomes of the experiment, while the random variable assigns a real number to each outcome.

The probability mass function (PMF) of a random variable $X$, denoted as $\Pr(X)$, is a function that maps each outcome $x_i$ in the sample space $\mathcal{X}$ to its probability $\Pr(X=x_i)$. This function is a measure of the likelihood of each outcome occurring in the experiment.

The PMF encapsulates the basic probabilistic structure of the random variable, providing a fundamental framework for understanding the behavior of random processes. It is a key concept in probability theory, enabling the calculation of various statistical properties and the formulation of probability models.

In summary, the PMF is a cornerstone of probability theory, allowing us to quantify the likelihood of outcomes in random experiments and to make predictions about the behavior of random variables based on this probabilistic framework.
Review of concepts that are the foundation of inference

• Recall that our eventual goal is to use a sample (generated by an experiment) to provide an answer to a question (about a system)

• Inference (informally) is the process of using the output of the experiment = experimental trials (the sample) to answer the question

• For our system and experiment, we are going to assume there is a single “correct” probability function (which in turn defines the probability of our possible random variable outcomes, the probability of possible random vectors that represent samples, and the probability of possible values of a statistic)

• For the purposes of inference, we often assume a parameterized family of probability models determine the possible cases that contain the “true” model that describes the result of the experiment

• This reduces the problem of inference to identifying the “single” value(s) of the parameter that describes this true model
Review of inference

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

• There are two major types of inference we will consider in this course: *estimation* and *hypothesis testing*

• Before we get to these specific forms of inference, let’s review *samples* and *statistics* (and the sampling distributions of samples and statistics)
Review of samples

- **Sample** - repeated observations of a random variable $X$, generated by experimental trials

- We will consider samples that result from $n$ experimental trials (recall the ideal $n$ is infinite!)

- We already have the formalism to represent a sample of size $n$, specifically this is a random vector:

  $$[X = x] = [X_1 = x_1, \ldots, X_n = x_n]$$

- While samples could take a variety of forms, we generally assume that each result of an experiment has the same form, such that they are identically distributed, and that each is independent of the others such that the sample is **independent and identically distributed**, which we abbreviate as i.i.d.
Review: Samples and Inference

• In each of these cases, we would like to use these samples to perform inference (i.e. say something about our parameter of the assumed probability model)

• Using the entire sample is unwieldy, so we do this by defining a statistic

• Note that statistics take the random vector as an input an output a value(s), such that the result of applying a statistic to a specific sample (a “realization of a random vector”) is a single value (or values)

• Note that since statistics take the random vector and output a value(s), the result of considering all values the statistic could take when applied to every possible sample is that the statistic has a probability distribution (the sampling distribution of the statistic!)
Review: Statistics

- **Statistic** - a function on a sample

- Note that a statistic $T$ is a function that takes a vector (a sample) as an input and returns a value (or vector):

$$T(x) = T(x_1, x_2, ..., x_n) = t$$

- For example, one possible statistic is the mean of a sample:

$$T(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- It is critical to realize that, just as a probability model on $X$ induces a probability distribution on a sample, since a statistic is a function on the sample, this induces a probability model on the statistic (the *statistic probability distribution* or the *sampling distribution* of the statistic):

$$\Pr(T(X))$$
Statistics and estimators I

- Recall for the purposes of inference, we would like to use a sample to say something about the specific parameter value (of the assumed) family or probability models that could describe our sample space.
- Said another way, we are interested in using the sample to determine the “true” parameter value that describes the outcomes of our experiment.
- An approach for accomplishing this goal is to define our statistic in a way that it will allow us to say something about the true parameter value.
- In such a case, our statistic is an estimator of the parameter: \( T(x) = \hat{\theta} \).
- There are many ways to define estimators (we will focus on maximum likelihood estimators in this course).
- Each estimator has different properties and there is no perfect estimator.
Statistics and estimators II

- Estimation is a “type” of inference, i.e. where we use a sample to reach a conclusion about a parameter
- Specifically, estimation is the process of saying something about the specific value of the true parameter
- Again, as a reminder, we do this by defining an estimator $\hat{\theta}$, which is a function on our sample
- Intuitively, an estimator is the value for which we have the best evidence for being the true value of the parameter (our “best guess”) based on the sample, given uncertainty and our assumptions
- Note that without an infinite sample, we will never know the true value of the parameter with absolute certainty (!!)
Statistics and estimators III

- **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: $\hat{T}(x) = \theta$.

- 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.
Statistics and estimators IV

- Since our underlying probability model induces a probability distribution on a statistic, and an estimator is just a statistic, there is an underlying probability distribution on an estimator:
  \[ Pr(T(X = x)) = Pr(\hat{\theta}) \]
- Our estimator takes in a vector as input (the sample) and may be defined to output a single value or a vector of estimates:
  \[ T(X = x) = \hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2, ...] \]
- We cannot define a statistic that always outputs the true value of the parameter for every possible sample (hence no perfect estimator!)
- There are different ways to define “good” estimators and lots of ways to define “bad” estimators (examples?)
Method of moments estimator I

- As an example of how to construct estimators, let's construct a method of moments estimator.

- Consider the single coin flip experiment / number of tails random variable / Bernoulli probability model family (parameter p) / fair coin model (assumed and unknown to us!!!) / sample of size $n=10$

- What is the sampling distribution (of the sample) in this case?

- We want to estimate $p$, where a perfectly reasonable estimator is:
  $$T(X = x) = \hat{\theta} = \hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- What is the probability distribution of this statistic in this case?

- e.g. this statistic (=mean of the sample) would equal 0.5 for the following particular sample (will it always?)
  $$x = [1, 1, 0, 1, 0, 0, 0, 1, 1, 0]$$
Method of moments estimator II

- Let’s continue with our example of constructing a method of moments estimator

- Consider the single coin flip experiment / number of tails random variable
  \[ \Omega = \{H, T\} \quad X : X(H) = 0, X(T) = 1 \]

- Bernoulli probability model family (parameter p)
  \[ X \sim p^X (1 - p)^{1-X} \]

- Sample of size n=10
  \[ [X = x] = [X_1 = x_1, X_2 = x_2, ..., X_{10} = x_{10}] \]

- Sampling distribution (pmf of sample) if i.i.d. (!!)
  \[ [X_1 = x_1, X_2 = x_2, ..., X_{10} = x_{10}] \sim p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}...p^{x_{10}}(1-p)^{1-x_{10}} \]
Method of moments estimator III

- Define a statistic $T(X)$
  \[ T(X = x) = T(x) = \bar{X} = \frac{1}{10} \sum_{i=1}^{10} x_i \]
- Note the values the statistic can take (!!), e.g. with $p = 0.5$
  \[ [T_{\min}, \ldots, T_{\max}] = [0, 0.1, \ldots, 1] \leftrightarrow [0, 1, \ldots, 10] \]
- We can therefore write the sampling distribution (pmf) of the statistic as
  \[ Pr(T(x)) \sim \binom{n}{nT(x)} p^{nT(x)} (1 - p)^{n-nT(x)} \]
Method of moments estimator IV

- We are going to use the statistic (mean) of the sample as an estimator of the parameter - and it follows the estimator has the same distribution (!!)

\[ T(x) = \hat{\theta} = \hat{p} \]

\[ \Pr(\hat{p}) \sim \left( \binom{n}{nT(x)} \right) p^{nT(x)} (1 - p)^{n - nT(x)} \]

- Also note that the expected value of this estimator is the true value of the parameter (do we ever know this true value!?)

\[ E\hat{p} = p \]

- In practice, one sample and we estimate a single value for the parameter

\[ x = [1, 1, 0, 1, 0, 0, 0, 1, 1, 0] \]
Method of moments estimator V

- As another example consider the heights experiment / identity random variable / Normal probability model family / with true parameters unknown to us (!!) / sample of size n=10

- A perfectly reasonable estimator $\hat{\mu}$ is:
  \[
  T(X = x) = \bar{X} = \hat{\theta} = \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i
  \]

- Note that the sampling distribution of this statistic is also normal, where this statistic is the expected value of the statistic sampling distribution (why might this be a good thing?):
  \[
  T(X = x) = \bar{X} = \hat{\mu} \sim N(\mu, \sigma^2/n)
  \]
  \[
  E(X) = \int_{-\infty}^{\infty} X f_X(x) dx = \mu
  \]

- e.g. this statistic (=mean of the sample) would equal 0.01 for the following particular sample (will it always?)

  \[
  x = [-2.3, 0.5, 3.7, 1.2, -2.1, 1.5, -0.2, -0.8, -1.3, -0.1]
  \]
Method of moments estimator VI

- For this same example, we could similarly define the following estimator:

\[ T(X = x) = Var(X = x) = \hat{\theta} = \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^2 \]

- What is the sampling distribution of this statistic?

- This is the variance of the sample and is justified because:

\[ Var(X) = \int_{-\infty}^{\infty} (X - \mu)^2 f_X(x) dx = \sigma^2 \]

- In sum, since we are calculating means and variances of samples, and these are “moments” when applied to random variables with a probability distribution, these are method of moments estimators.
Introduction to maximum likelihood estimators (MLE)

- 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.
• 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

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

• However, we could turn this around and fix the sample and let the parameters vary (this is a likelihood!)
• For example, say we have a sample \( n=1 \), where \( x=0.2 \) then the likelihood is:

\[ L(\mu | x = 0.2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(0.2-\mu)^2}{2}} \]
Likelihood II

- **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

\[
L(\theta|x_1, x_2, \ldots, x_n) = L(\theta|x_1)L(\theta|x_2)\ldots L(\theta|x_n)
\]

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

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

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:
Likelihood III

- **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 of our single coin example:

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

• And for our heights example:

\[ MLE(\hat{\mu}) = \frac{1}{n} \sum_{i}^{n} x_i = \bar{x} \quad MLE(\hat{\sigma}^2) = \frac{1}{n} \sum_{i}^{n} (x_i - \bar{x})^2 \]
That’s it for today

• Next lecture, we will continue our discussion of maximum likelihood estimators and begin our discussion of hypothesis testing