STA261: Lecture 4

Likelihood Inference I

Alex Stringer

July 16th, 2018

Disclaimer

The materials in these slides are intended to be a companion to the course textbook, *Mathematical Statistics and Data Analysis, Third Edition*, by John A Rice. Material in the slides may or may not be taken directly from this source. These slides were organized and typeset by Alex Stringer.

A big thanks to Jerry Brunner as well for providing inspiration for assignment questions.

License

Towards finding estimators

We've talked about how to tell whether an estimator is sufficient, and about how to improve one estimator if we know about a sufficient statistic. Are these tools enough to actually find good estimators?

We have also talked about factoring the joint density of the sample given the parameters, and analyzing how the result depends on the parameters.

Let's take this idea one step further.

Joint Distribution of the Sample

So far we have talked about how we should use all the data we have when making inferences about θ. There is a converse to this statement as well.

One of the main principles behind the "Frequentist" approach to parameter estimation that we adopt in this course is that inferences about a parameter should be based *only* on the observed data.

So we should use all the data, and nothing but.

Of course, this isn't really true, since we make a *ton* of assumptions about the family of distributions from which the data came- but as I said in this course, we are going to take the family as fixed and known, and talk only of parameter estimation.

Joint Distribution of the Sample

How can we achieve this goal?

Consider the joint distribution of the sample, $f(\mathbf{x}|\theta)$. For a fixed, observed dataset $\mathbf{x}$ generated from this distribution, what is the simplest sufficient statistic we can think of, given the above philosophy of using only the observed data to estimate θ?

The Data is Sufficient

Really what we are saying is simply that the data $\mathbf{x}$ itself is sufficient for θ.

Trivially, we can write

$$f(\mathbf{x}|\theta) = g(\mathbf{x}, \theta) \times h(\mathbf{x})$$

and just take $h(\mathbf{x}) = 1$.

What this means is that we are justified in using only $f(\mathbf{x}|\theta)$ to estimate θ.

Likelihood

But I told you before that since we have already *observed* the dataset, there is no randomness left in $\mathbf{x}$. We're conditioning all inference based on the observed data only. So what is the point of looking at its distribution?

Definition: The **likelihood** function is the joint distribution of the data, treated as a function of the parameters:

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

It's exactly the same formula, just treated as a function of θ for fixed $\mathbf{x}$ rather than the other way around.

It's one of the most important definitions in all of Statistics.

Example

Suppose $X_i \sim Bern(p)$ is a random sample of coin flips. Find the likelihood function for p.

Example

Solution: the likelihood function is equal to the probability of observing any particular sequence of 0's and 1's. Since each trial is independent, the probability of observing a sequence is just equal to the product of the probabilities of observing each result:

$$\begin{aligned} L(p|\mathbf{x}) =& p \times p \times p \times \ldots \times (1-p) \times (1-p) \ldots \times (1-p) \\ =& p^{\sum_{i=1}^{n} x_i} \times (1-p)^{\sum_{i=1}^{n} (1-x_i)} \end{aligned}$$

We multiply by p for each 1 in the sequence, and by (1-p) for each 0 in the sequence. There are $\sum_{i=1}^{n} x_i$ 1's, and $\sum_{i=1}^{n}(1-x_i) = n - \sum_{i=1}^{n} x_i$ 0's.

Note any sequence with the same number of 0's and 1's in it gets the same likelihood for p- the order doesn't matter (can you connect this to the concept of sufficiency?).

Example: IID Data

The case where we have IID (Independent, Identically Distributed) data is the most common, and gets special attention.

Suppose $X_i \sim F_\theta$ independently. Denote the corresponding density of each x_i as $f_x(x|\theta)$ Then

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f_{x_i}(x_i|\theta)$$

Likelihood

Treating this as a function of θ, there is a lot we can do with $L(\theta)$ (I'm dropping the dependence on $\mathbf{x}$ in my notation).

Consider this: values of θ that give a higher $L(\theta)$ are more likely to have generated the observed data. (Why?)

Maximum Likelihood

This idea gives us a way to find estimators.

Definition: for an observed sample $\mathbf{x}$ with joint density $f(\mathbf{x}|\theta) = L(\theta)$, the **maximum likelihood estimator** of θ is the value of θ that maximizes $L(\theta)$.

$$\hat{\theta} = \text{argmax}_\theta L(\theta)$$

This gives us *the value of the parameter that was most likely to have generated the data we observed*.

Practically, we just turned our estimation problem into an optimization problem.

Log-likelihood

Of great practical and theoretical interest is the *log-likelihood*

$$\ell(\theta) = \log L(\theta)$$

(base e log, which some may know as ln).

This gives the same MLE (Maximum Likelihood Estimators) as working with $L(\theta)$ because log is a monotone function, so in general $f(x)$ and $\log f(x)$ have the same optima.

Why?

1. The likelihood of an independent sample is a product over the density of each point. The log-likelihood is a sum.
2. Densities themselves are often a product of several factors anyways, and a log of these gives a sum. Sums are way easier to work with than products.
3. The log-likelihood has a bunch of awesome theoretical properties that we will discuss in the coming weeks
4. Numerically more stable. Likelihood is a product of density values, most of which are small, so can get really small really fast. Log likelihood is a sum of logged values, which gets sort of small much less fast.

Example

Simple example: $X_i \sim N(\mu, 1)$. The likelihood is

$$L(\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

The log likelihood, on the other hand, is

$$\ell(\mu) = \log L(\mu) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Example

To optimize this, take two derivatives with respect to μ. Set the first to zero and solve; check that the second one is negative $\Rightarrow$ local maximum.

$$\partial\ell/\partial\mu = \sum_{i=1}^{n}(x_i - \mu) = 0 \implies \hat{\mu} = \bar{x}$$

$$\partial^2\ell/\partial\mu^2 = -n < 0 \forall \mu \implies \hat{\mu} = \bar{x} \text{ maximum of } \ell(\mu)$$

You will work out the details for yourself on Assignment 3, and re-do when the standard deviation is also a parameter to be estimated.

Multivariable Optimization

I'm expecting you're familiar with basic univariate optimization from calculus like we just did. If you're not though, the previous example contains all you need to know.

I'm not expecting you're all familiar with multivariable optimization, so let's briefly discuss the procedure we'll use in this course.

Multivariable Optimization

To optimize a multivariable function $f : \mathbb{R}^d \to \mathbb{R}$:

1. Take derivatives with respect to each variable, holding all others fixed (these are called *partial* derivatives, and is what the ∂ symbols refer to), and set each of these derivatives equal to zero.
2. This gives you a system of equations. Solve the system to get estimators for each parameter.
3. The part we're going to skip is the multivariable generalization of the second derivative test: check whether the negative of the *Hessian* (matrix of second-order partial derivatives) is *positive definite*.

So in this course, just find the estimators using step 1 and 2, because I don't want to spend any more time on this, even though it's super important.

Example

We can revisit the normal example when both parameters are unknown. The log-likelihood function is now

$$\ell(\mu, \sigma) = -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

with derivatives

$$\partial\ell/\partial\mu = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \implies \hat{\mu} = \bar{x}$$

$$\partial\ell/\partial\sigma^2 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

Notes

Notice I did two tricky things:

- Assumed that $\sigma > 0$, which I did twice (where?). You should always first recognize what your *parameter space* - the space of all possible values of θ - is, and only optimize within it.
- Maximized with respect to σ^2, rather than σ. This works because the MLE any one-to-one function of the parameter, $\psi = g(\theta)$, is $\hat{\psi} = g(\hat{\theta})$. This is not a trivial fact, and it is extremely useful.

Example

Find a MLE for the *precision* in the previous example, which is a fancy word sometimes used to refer to the inverse of the variance. That is, find the MLE for $\psi = 1/\sigma^2$.

Example

The log-likelihood can be re-parametrized in terms of ψ:

$$\ell(\mu, \psi) = -\frac{n}{2}\log 2\pi + \frac{n}{2}\log\psi - \frac{\psi}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

with derivatives

$$\partial\ell/\partial\mu = \psi\sum_{i=1}^{n}(x_i - \mu) = 0 \implies \hat{\mu} = \bar{x}$$

$$\partial\ell/\partial\psi = \frac{n}{2\psi} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \implies \hat{\psi} = \frac{n}{\sum_{i=1}^{n}(x_i - \mu)^2}$$

Example

..or, we could have used the fact that since the function $g(x) = 1/x$ is one-to-one ($x \neq 0$), the MLE for $\psi = 1/\sigma^2$ is

$$\hat{\psi} = 1/\hat{\sigma}^2 = \frac{n}{\sum_{i=1}^{n} (x_i - \mu)^2}$$

In MLE problems, reparametrizing the distribution to make it easier to differentiate is a common useful technique.

You can't always use calculus

You can't maximize the likelihood using calculus techniques in three major common cases:

- The likelihood is not continuous on the whole parameter space, so you can't take derivatives
- The parameter space is not an *open* subset of $\mathbb{R}^d$. This occurs when it includes its boundary. For example, what if we had allowed $\sigma = 0$ in the previous example?
- The support of the distribution depends on the parameters. This one is more subtle, so let's take a look at an example

Example

Let $X_i \sim Unif(0, b)$, the continuous uniform distribution on the open interval $(0, b)$. Find the MLE of b.

The likelihood function is

$$L(b) = \prod_{i=1}^{n} \frac{1}{b}$$

which is a strictly decreasing function of b. In particular, it is unbounded as $b \to 0$.

Or is it?

Example

We made a mistake: we didn't express all of the dependency on b explicitly in $L(b)$. We actually should have written

$$L(b) = \prod_{i=1}^{n} \frac{1}{b} \times I(x_i \leq b)$$

I'll leave it to you to

- Convince yourself that this is true and
- Show that the MLE is $\hat{b} = max(x_i)$

Connection to Sufficiency

The MLE is sufficient. Why?

Exercise (Assignment 3): show that the MLE can depend on the data only through the value of a sufficient statistic.

Because the MLE is a function of a sufficient statistic, it is itself sufficient.

Is it consistent?

Derivatives of the log-likelihood

Definition: the **score statistic** (or score vector, or just score) is the first derivative of the log-likeilhood:

$$S(\theta) = \partial \ell(\theta)/\partial \theta$$

For IID data, the log likelihood is a sum over the dataset, and so as well is the score:

$$S(\theta) = \partial \sum_{i=1}^{n} \ell_i(\theta)/\partial \theta \equiv \sum_{i=1}^{n} s_i(\theta)$$

Derivatives of the log-likelihood

Definition: the **observed information** is the negative second derivative of the log-likelihood, or is the negative derivative of the score:

$$J(\theta) = -\partial^2 \ell(\theta)/\partial\theta^2 = -\partial S(\theta)/\partial\theta$$

Like the score, it adds over the data for IID random samples:

$$J(\theta) = \sum_{i=1}^{n} j_i(\theta)$$

Little j refers to a single datapoint; big J refers to the whole sample.

Derivatives of the log-likelihood

Definition: the *expected* or *Fisher* information is the expected value (over the distribution of $\mathbf{X}$) of the observed information:

$$i(\theta) = E(j(\theta))$$

Expected information also adds across the dataset, but a nice thing happens: because the datapoints are IID, the expected information from each is equal, $E(j_i(\theta)) \equiv E(j(\theta)) \equiv i(\theta)$ for $i = 1 \dots n$. So,

$$I(\theta) = ni(\theta)$$

The information in the whole sample is n times the information in a single point: information increases *linearly* with the sample size.