STA261: Week 2

Introduction to Estimation Theory

Alex Stringer

Jan 16th - 20th, 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

Recap

Last week:

- Convergence in probability of sequences of random variables
- Law of Large Numbers
- Convergence in distribution of sequences of random variables
- Central Limit Theorem for a sum of independent random variables

This week

- Introduction to the theory of parameter estimation
- Consistency
- Method of moments

Recall: Probability

Recall: in *probability theory*, we are given all the information about a stochastic (random) process, then ask questions about what realizations (data) of that process might look like.

Example: heights of students are normally distributed with mean $170cm$ and standard deviation $20cm$. What is the probability that a randomly sampled student is taller than me, at $\approx 185cm$?

The inverse problem

What if we didn't know any information about the stochastic process, but we did have a bunch of data generated by it?

Example: I measured a bunch of randomly selected students' heights. How do I know whether the heights of students are normally distributed with mean $170cm$ and standard deviation $10cm$?

The inverse problem

In this course, we will always assume we know the *famliy* of distributions from which the data was generated, and focus on estimating the *parameters* of that family. So, I will tell you "the heights of students are normally distributed" and we will focus on identifying the "mean $= 170cm$" and "standard deviation $= 10cm$" parts of the above question.

The inverse problem

Our intuition is to calculate the sample mean $\bar{X_n}$ and the sample standard deviation $s = \sqrt{1/(n-1) \times \sum_{i=1}^{n} \left(X_i - \bar{X_n}\right)^2}$, and use these to make a conclusion about the mean and standard deviation of the population from which the data came. But how do we *know* this is a good thing to do?

This is an example of *parameter estimation*, the central theme of this course.

Coin toss example

Recall the simple coin toss example of lecture 1. Suppose now that we don't know whether the coin is fair, that is, we don't know whether $P(X = 1) = 1/2$ or not. We flip the coin 10 times and observe 7 heads. What do we do?

Definition: Family of Distributions

From STA257, we are used to seeing things like $X \sim N(\mu, \sigma)$, and saying "X has **the** normal distribution".

Here we have to get more precise: X doesn't have **the** normal distribution, X has **a** normal distribution.

We call a set of distributions $\{F_\theta\}$ a *family* if they have the same functional form, but are specified only up to an unknown parameter

Example: Family of Distributions

$F_\theta = N(\mu, \sigma)$ with $\theta = (\mu, \sigma) \implies$ family of Normal distributions. They share the same pdf

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

up to the specification of the values of μ and σ.

Example: Family of Distributions

$F_\theta = Bin(n, p)$ with $\theta = p$. For a fixed and known n, the pmf is the same for all memebers of this family, up to the constant p.

Formal Statement of the Problem

Let $\{X_i\}_{i=1}^{n}$ be a sequence of random variables generated from a known family of distributions $F_\theta(\cdot)$ indexed by parameter $\theta \in \mathbb{R}^d$. We seek an *estimator* of θ, defined as a function

$$\hat{\theta}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^d$$

that takes in the given sequence and returns an *estimate* of θ. We call such a function an *estimator* of θ.

Statistics, as a subject, is concerned with

- *Finding* such functions
- *Evaluating* the quality of such functions in theory and in practice
- *Using* the output of such functions to make inferences about the family $F_\theta(\cdot)$

Notation

There are a lot of competing and confusing concepts at play here.

θ: **parameter**. a fixed, constant element of the vector space $\mathbb{R}^d$. In general $d > 1$, meaning θ is a vector, but there are many cases where $d = 1$ and it is a single number.

$\hat{\theta}$: **estimator** of θ. This is a *function*; an abstract mathematical object. It is not a number; it is used to *generate* numbers from data.

$\hat{\theta}$: **estimate** of θ. This is an actual number, by plugging a real dataset into the *estimator* $\hat{\theta}$.

What? That must be a typo

It's not a typo. In this course, the same notation will be used for *estimators* and *estimates*. This is consistent with the broader statistics literature. It should always be clear from the context whether we are talking about a function or a particular value of that function.

This isn't weird: in calculus, you write $f(x)$ to indicate a function, like $f(x) = x^2$. You also write things like $f(x) = 4$ when you mean to refer to a specific value of $f(x)$. This is exactly the same idea.

Examples

Here is an example of a **parameter**.

Heights of students are normally distributed with mean $170cm$ and standard deviation $10cm$.

$F_\theta = N(\mu, \sigma)$

$\theta = (\mu, \sigma)$

$d = 2$

Example

Here is an example of an **estimator**.

In the above example, we write $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = s$.

Or more compactly, $\hat{\theta} = \left(\bar{X}, s\right)$

You should start thinking about $\bar{X}$ and s as functions from $\mathbb{R}^n \to \mathbb{R}$.

Example

Here is an example of an **estimate**.

We sample data $\mathbf{X} = (145, 189, 172, 166, 159)$, and from this calculate $\bar{X} = 166.2$ and $s = 16.24$.

Our **estimate** of μ is $\hat{\mu} = 166.2cm$ and our *estimate* of σ is $\hat{\sigma} = 16.24cm$.

Our **estimate** of θ is $\hat{\theta} = (166.2cm, 16.24cm)$.

Example

Here is an example of a **parameter**.

We flip a coin n times for some fixed n. The true probability of any individual flip being heads is p. We write $X_i \sim Bin(n, p)$, and our parameter is $\theta = p$.

Example

Here is an example of an **estimator**.

One estimator is $\hat{p} = \bar{X}$, the sample proportion of heads.

In this one-dimensional example, with $\theta = p$, $\hat{\theta} = \hat{p}$ too.

Example

Here is an example of an **estimate**.

We flip the coin $n = 10$ times and observe 7 heads. So $\hat{\theta} = \hat{p} = \bar{X} = 0.7$.

Estimators are Random Variables

Estimators are random variables, because they are functions of random variables.

The probability distribution of an estimator is sometimes referred to as its **sampling distribution**.

E.g. the sampling distribution of $\bar{X}$ is $N(\mu, \sigma/\sqrt{n})$ when $X_i \sim N(\mu, \sigma)$

Often, we don't know this sampling distribution exactly.

Evaluating Estimators

To decide whether a given estimator is “good”, in the sense that it provides reasonable estimates of the population parameters, we study the properties of the estimator and its sampling distribution.

There are four major properites of an estimator that we are typically concerned with. These will form much of the material between now and the midterm.

Important: The first half of the course focusses on evaluating *estimators*, that is, functions. We also evaluate *estimates*, that is, numbers. This will be the focus of the second half of the course.

Evaluating Estimators

Intuitively, here are four big things we might hope our estimator satisfies:

1. As we get more data, we should be able to get as close as we want to the parameter we are estimating, with as high a probability as we want (this should sound familiar. . .).

Evaluating Estimators

Intuitively, here are four big things we might hope our estimator satisfies:

2. We should base our estimator off of all the information in the sample. Our estimator should be a *summary* of the full sample. Whether we know the entire dataset or just our estimate $\hat{\theta}$, we should make the same conclusions regarding θ.

Evaluating Estimators

Intuitively, here are four big things we might hope our estimator satisfies:

3. We should not expect our estimator to vary *systematically* from the true parameter. If we took repeated samples from the same population, and got the corresponding estimates of θ, we might want to know that the mean of *these* would be the true θ.

Evaluating Estimators

Intuitively, here are four big things we might hope our estimator satisfies:

4. We want our estimator to have low variance. Slightly different samples should not yield wildly different estimates.

Consistency (textbook, page 266)

1. As we get more data, we should be able to get as close as we want to the parameter we are estimating, with as high a probability as we want (this should sound familiar. . .).

Definition: Let F_θ be a family of distributions with parameter $\theta \in \mathbb{R}^d$. Let $\hat{\theta}$ be an estimator of θ. We say that $\hat{\theta}$ is **consistent** for θ if $\hat{\theta} \xrightarrow{p} \theta$.

Consistency

Note that we defined convergence in probabilty for scalar random variables only, while consistency is defined for vector-valued random variables. In this course, if asked to prove that $\hat{\theta} = \left(\hat{\theta_1}, \ldots, \hat{\theta_d}\right)$ is consistent for $\theta = (\theta_1, \ldots, \theta_d)$, just show that each element is consistent, $\hat{\theta_k} \xrightarrow{p} \theta_k$ for $k = 1 \ldots d$.

Why do we care about consistency?

Consistency is typically the bare minimum we ask of an estimator. If as we get more and more data, our estimator doesn't get closer and closer (in probability) to the thing it's trying to estimate, we have a problem.

But, while consistency is *necessary* for our estimator to be good, it's certainly not enough on its own.

For example, $\bar{X} = (1/n) \times \sum_{i=1}^{n} X_i$ is consistent for μ. But so is $(1/(n + 1,000,000)) \times \sum_{i=1}^{n} X_i$, and any other silly estimator we can define that still has the same limit in probability.

Example: LLN

The LLN says that the sample mean $\bar{X}$ is consistent for the population mean $E(X)$.

This is a special result because it is true for any *family* of distributions; usually we have to find consistent estimators for each new family we come across.

Example: Population Moments (textbook, page 266)

Actually, the LLN implies that all the sample moments computed based on *independent* samples are consistent for their respective population moments, that is

$$(1/n) \times \sum_{i=1}^{n} X_i^k \xrightarrow{p} E(X^k)$$

for every $k \in \mathbb{N}$. We denote the quantity on the left as $\bar{X}^k$, the quantity on the right as μ^k, and say that $\bar{X}^k$ is consistent for μ^k.

Continuous Functions

Recall assignment 1, textbook question 7: show that if $X_n \xrightarrow{p} x$, a g is a continuous function, $g(X_n) \xrightarrow{p} g(x)$.

So if $\hat{\theta}$ is consistent for θ, then $g(\hat{\theta})$ is consistent for $g(\theta)$.

This gives you most of what you need to do proofs about consistency.

Slutsky

This is sometimes referred to as *Slutsky's Lemma*.

It works for multivariable functions too, which is mostly important in the case of $f(X,Y) = X + Y$ (addition) and $f(X,Y) = XY$ (multiplication):

$$X_n \xrightarrow{p} x, Y_n \xrightarrow{p} y \implies X_n + Y_n \xrightarrow{p} x + y$$

$$X_n \xrightarrow{p} x, Y_n \xrightarrow{p} y \implies X_n Y_n \xrightarrow{p} xy$$

Application: Central Limit Theorem, Again

A key application of these convergence rules is finding consistent estimators- but first, let's look at modifying something we saw last week so we can actually use it.

Recall the CLT: $X_1 \dots X_n$ independent random sample with mean 0 and standard deviation σ, then $\frac{\sum_{i=1}^n X_i}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$.

You may notice that in practice, we don't always know σ. Last week we cheated and used examples where we did know it exactly, but usually we don't.

Application: Central Limit Theorem, Again

Let V_n be an *estimator* of σ^2 having the property that $nV_n \xrightarrow{p} \sigma^2$ (e.g. $V_n = s^2/n$).

Because the function $g(x) = \sigma/\sqrt{nx}$ is continuous, the above rules let us write $\sigma/\sqrt{nV_n} \xrightarrow{p} \sigma/\sigma = 1$.

The rule above about multiplication therefore lets us conclude that

$$\left(\frac{\sqrt{n}\bar{X}_n}{\sigma}\right) \times \left(\frac{\sigma}{\sqrt{s_n^2}}\right) \xrightarrow{p} Z \times 1$$

where $Z \sim N(0,1)$.

Application: Central Limit Theorem, Again

So we can replace σ with a consistent estimate in the CLT definition and the theorem still works.

This is what we would have done anyways without thinking about it.

This is called "Studentizing" after William Gosset, who invented the t-distribution under the pen name "Student". We'll see the connection to the t-distribution when we derive it later in the course.

Example

E.g. let $X_i \sim N(\mu, \sigma)$ independently. Show that the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2$ is consistent for $\sigma^2 = E(X-\mu)^2$.

Proof:

$$
\begin{aligned}
s^2 &= \frac{n}{n-1} \times \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2 \\
&= \left(\frac{n}{n-1}\right) \times \left((1/n) \times \sum_{i=1}^{n} X_i^2 - \left(1/n\sum_{i=1}^{n} X_i\right)^2\right) \\
&= \left(\frac{n}{n-1}\right) \times \left(\bar{X^2} - \left(\bar{X}\right)^2\right) \\
&\to (1) \times \left(E(X^2) - (E(X))^2\right) \\
&= \sigma^2
\end{aligned}
$$

Note

The n vs $n-1$ thing doesn't affect consistency arguments since $\lim_{n\to\infty}\frac{n}{n-1}=1$. It's so not important in this context that I even forgot about it the first time I wrote these slides.

Remember, correcting the bias in the sample variance is an important issue in *small samples*, while here the subject is *limiting arguments*, i.e. what happens when the sample gets infinitely large.

Example

For the previous example, show that the sample standard deviation $s = \sqrt{s^2}$ is consistent for the population standard deviation $\sigma = \sqrt{\sigma^2}$.

Proof: the function $g(x) = \sqrt{x}$ is continuous, and we just proved that s^2 is consistent for σ^2. Therefore,

$$s = \sqrt{s^2} \to \sqrt{\sigma^2} = \sigma$$

Finding Consistent Estimators

It's nice that we can prove that using $\bar{X}$ to estimate μ and s^2 to estimate σ^2 gives us consistent estimators of these quantities. How did we *know* to use these though? Especially the variance one, did we just guess?

Wouldn't it be nice if we could reverse the steps of that consistency proof, and use the population moments to *find* consistent estimators?

We *can* do that

$$\begin{aligned}
\sigma^2 &= E(X^2) - \left(E(X)\right)^2 \\
&\leftarrow \bar{X^2} - \left(\bar{X}\right)^2 \\
&= \frac{1}{n} \times \sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&= \frac{1}{n} \times \sum_{i=1}^{n} (X_i - \bar{x})^2 \\
&= s^2
\end{aligned}$$

I used $1/n$ here- you'll see why shortly.

Method of Moments

The **Method of Moments** gives us a guaranteed way of obtaining consistent estimators.

This amounts to finding the population parameters as functions of the population moments, solving the resulting system of equations, and plugging the sample moments in for the population moments.

The resulting estimators are consistent because all of the functions involved - and their inverses - are continuous.

Method of Moments

Algorithm: Method of Moments let $X_i \sim F_\theta$ independently, $\theta = (\theta_1, \dots, \theta_d)$. The **method of moments** is as follows:

1. Find expressions for the first d population moments in terms of $\theta_1, \dots, \theta_d$,

$$\begin{aligned} E(X) &= g_1(\theta_1, \dots, \theta_d) \\ E(X^2) &= g_2(\theta_1, \dots, \theta_d) \\ &\vdots \\ E(X^d) &= g_d(\theta_1, \dots, \theta_d) \end{aligned}$$

Method of Moments

2. Solve the resulting system of nonlinear equations to get the parameters as continuous functions of the population moments,

$$\begin{aligned}
\theta_1 &= g_1^{-1}(E(X),\ldots,E(X^d)) \\
\theta_2 &= g_2^{-1}(E(X),\ldots,E(X^d)) \\
&\vdots \\
\theta_d &= g_d^{-1}(E(X),\ldots,E(X^d))
\end{aligned}$$

3. Plug in the sample moments $\bar{X}^d$ in place of their respective population moments. The result is a set of consistent estimators for $\theta_1 \ldots \theta_d$, i.e. a consistent estimator for the vector θ.

Example

Let $X_i \sim N(\mu, \sigma)$ independently. Find a MoM estimator for $\theta = (\mu, \sigma^2)$.

Solution:

$$\begin{aligned} E(X) &= \mu \\ E(X^2) &= \sigma^2 + \mu^2 \end{aligned}$$

Solving the system,

$$\begin{aligned} \mu &= E(X) \\ \sigma^2 &= E(X^2) - \mu^2 \\ &= E(X^2) - \big(E(X)\big)^2 \end{aligned}$$

Example

Plugging in the sample moments gives us

$$\hat{\mu} = \bar{X}$$
$$\hat{\sigma^2} = \bar{X^2} - \left(\bar{X}\right)^2$$

as before.

Example

Let $X_i \sim Bern(p)$ be independent draws from the Bernoulli distribution (single coin toss). Find a MoM estimator for p.

Solution

$E(X) = p$, so $\hat{p} = \bar{X}$, the sample proportion.

Example

Let $X_i \sim Unif(0, b)$ (the *continuous* uniform distribution on $(0, b)$, with pdf

$$f_{x_i}(x) = \frac{1}{b} \times I(0 \leq x \leq b)$$

Find a MoM estimator for b.

Example

Evaluate

$$E(X) = \int_0^b x \times \frac{1}{b} dx$$
$$= \frac{b}{2}$$

Note I suppressed the $I(0 \leq x \leq b)$ because this just equals 1 on the interval across which we are integrating.

Example

By the Method of Moments, set

$$E(X) = \bar{X} = \frac{\hat{b}}{2}$$

to get $\hat{b} = 2\bar{X}$.

The Method of Moments provides an intuitive estimator here, since we think of $\bar{X}$ as estimating the centre of the distribution, i.e. half the distance from 0 to b. So $2\bar{X}$ should give us a reasonable estimate of b. At least, we know it's consistent.

We'll see next week we can do even better for the uniform distribution.

Example

Things aren't always this easy. Let $X_i \sim Unif(a,b)$. Find a MoM estimator for $\theta = (a,b)$.

Solution: see assignment 2. *Hint*: $E(X) = \frac{a+b}{2}$, so $b > E(X)$.