STA261: Week 5

Properties of the Sampling Distribution of an Estimator (and Midterm Review)

Alex Stringer

Feb 5th - 9th, 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

Last Week

Last week we talked about the "large sample" properties of the MLE, including its asymptotic sampling distribution.

We said "in large samples, the MLE is approximately Normally distributed with mean θ_0 and variance $I(\theta_0)^{-1}$".

This Week

This week, we will discuss the concept of a sampling distribution of an estimator in more detail.

We will then talk about properties 3 and 4 of an estimator, both of which relate to its sampling distribution

Sampling Distribution

We defined the sampling distribution of an estimator as its probability distribution.

Estimators are random variables, because they are functions of the sample, which is itself random.

So they have probability distributions.

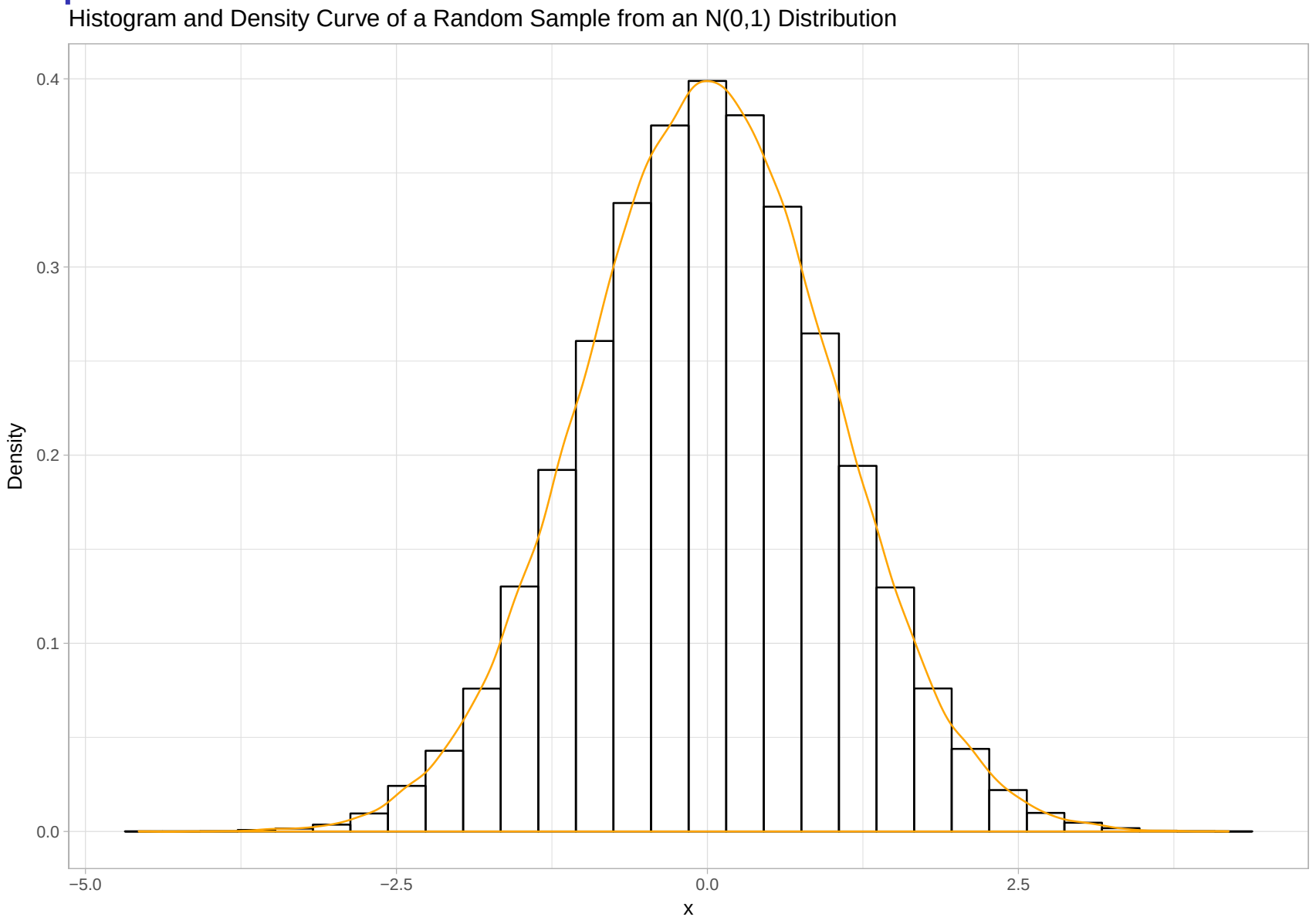
But what does this mean?

Sampling Distribution

When we talk of a random variable having a probability distribution, $X \sim F_\theta$, we usually mean to describe the set of plausible outcomes if we sampled many values of it.

For example if $X \sim N(0, 1)$, we picture a bell curve (the density). If we sampled many values of X and made a histogram, the density curve would touch the tops of the bars.

Example


Histogram and Density Curve of a Random Sample from an N(0,1) Distribution
Density
0.4
0.3
0.2
0.1
0.0
−5.0
−2.5
0.0
2.5
x

Sampling Distribution

For an estimator, we only observe one dataset, and calculate one value.

Key point: the dataset we observed was one of may possible datasets we could have observed.

The data is *random*. If we repeated our experiment, we would get a different dataset, and a different estimate of θ.

$\hat{\theta}$ is a realization of a random variable.

Example

We say that if $X_i \sim N(0,1)$, $\bar{X} \sim N\left(0, \frac{1}{\sqrt{n}}\right)$.

Consider the following procedure:

- Randomly sample $B = 10,000$ datasets of size $n = 5$ from a $N(0,1)$ distribution
- Calculate $\bar{X}$ for each, so we get $B = 10,000 \bar{X}$'s
- Those $B = 10,000$ $\bar{X}$'s are a random sample from the sampling distribution of $\bar{X}$. They should follow a $N\left(0, \frac{1}{\sqrt{5}}\right)$ distribution

Example


Histogram and Density Curve of Xbar
For X sampled from a N(0,1) distribution
Density
0.00
0.25
0.50
0.75
−1
0
1
2
x

Example

```
## Mean of Xbar =  0.007
## SD of Xbar =  0.449

## Theoretical mean of Xbar = 0,
## Theoretical SD of Xbar =  0.447
```

Example

We don't always know the sampling distribution of our estimator exactly.

We saw last class that we can approximate the sampling distribution of the MLE in any problem (assuming the regularity conditions hold) using a normal distribution, and we found the mean and variance.

The mean and variance of the sampling distribution of our estimator are important.

Unbiasedness

Definition: suppose $\hat{\theta}$ is an estimator for θ. The **bias** of $\hat{\theta}$ is defined as

$$bias(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$$

The bias measures the degree by which we expect $\hat{\theta}$ to differ from θ systematically, or on average.

If we repeated our experiment many times and calculated the average of all the resulting estimates of θ, we would expect this to be $bias(\hat{\theta})$ away from θ.

Unbiasedness

Property 3 of an estimator is called **unbiasedness**.

Definition: an estimator $\hat{\theta}$ of θ is called **unbiased** if $E(\hat{\theta}) = \theta$.

This is equivalent to $bias(\hat{\theta}) = 0$.

Why Unbiasedness?

This comes from the principle that we want to pick an estimation procedure that we don't expect to give wrong answers, at least on average.

This is because we are using our estimates to make decisions about the process that generated the data.

Why Unbiasedness?

Example: consider an important scientific experiment, like a medical trial. Something where the outcome, the decision we have to make from data, actually affects people's lives in a tangible way.

Suppose we want to estimate the incidence of a harsh side-effect for some life-saving drug, for example.

Would we use an estimation procedure that we knew, on average, was likely to underestimate this? Would we feel comfortable asserting that the incidence was 0.001%, if we knew across all possible trials we could run, our procedure would give an underestimate of this on average?

Example

Back to math. Let $X_i \sim N(\mu, \sigma)$ and show $\bar{X}$ is unbiased.

We know $E(X) = \mu$, and $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$. Because $E(\bar{X}) = \mu$, $\bar{X}$ is unbiased for μ.

Example

Let $X_i \sim Exp(\theta)$, with $f(x) = \frac{1}{\theta}e^{-x/\theta}$. Is $\hat{\theta} = \bar{X}$ unbiased for θ?

Compute $E(X) = \theta$, either by integrating or using the MGF (integrating is easier in this example). Then compute

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \theta$$

so $\hat{\theta}$ is unbiased for θ.

Example

Let $X_i \sim Exp(\beta)$, with $f(x) = \beta e^{-\beta x}$. Is $\hat{\beta} = \frac{1}{\bar{X}}$ unbiased for θ?

Compute $E(\hat{\beta}) = E\left(\frac{1}{\bar{X}}\right) = \ldots$

Hmm.

Example

Except for cases when the functional form of the estimator is very simple (e.g. a linear combination of datapoints), there is no reason that the expectation of the estimator would be easy to compute.

Sometimes, we don't even have a formula for the estimator.

This is one reason why the CLT for the MLE is so useful. No matter how messy/intractable it is, we can assert that the MLE is *aymptotically unbiased*, because the CLT implies that $E(\hat{\theta} - \theta_0) \to 0$.

Example

So in this example, compute the MLE for β:

$$\ell(\beta) = n \log \beta - \beta \sum_{i=1}^{n} X_i$$

$$S(\beta) = \frac{n}{\beta} - \sum_{i=1}^{n} X_i$$

$$\implies \hat{\beta} = \frac{1}{\bar{X}}$$

Without doing any more work, we can now state that $E\left(\frac{1}{\bar{X}}\right) \to \beta$.

Variance of an Estimator

We have talked about the mean of the sampling distribution of an estimator, and said that we want it to equal θ.

We can also talk about the variance of this sampling distribution, for fixed n. We'd want it to be low; slightly different samples should not yield wildly different estimates.

How low can we go?

The Cramer-Rao Lower Bound (Textbook, page 300 - 301)

Theorem: Suppose $\hat{\theta}$ is any *unbiased* estimator of θ. Then

$$Var(\hat{\theta}) \geq \frac{1}{nI_0(\theta_0)}$$

where I_0 denotes the Fisher Information for a single datapoint, so nI_0 is the Fisher Information for the whole sample of size n.

In practice, you plug in $\hat{\theta}$ for θ_0.

The conditions required for this to hold are essetially the same as the regularity conditions required for $Var(\hat{\theta}) \to I(\theta_0)$, as discussed last lecture.

Efficiency

This lets us state property 4.

Definition: an estimator $\hat{\theta}$ of θ is **efficient** if it attains the Cramer-Rao Lower bound, that is if

$$Var(\hat{\theta}) = \frac{1}{nI_0(\theta_0)}$$

Corollary: The MLE is *asymptotically efficient*.

Example

Let $X_i \sim N(\mu, 1)$. Show $\bar{X}$ is an efficient estimator of μ.

You have to show it's unbiased, which we did above: $E(\bar{X}) = \mu$.

Then compute the variance of the estimator,

$$\begin{aligned} Var(\bar{X}) &= \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) \\ &= \frac{n(1)}{n^2} \\ &= \frac{1}{n} \end{aligned}$$

Example

Now compute the Fisher Information for a single datapoint,

$$I_0(\mu) = -E\left(\frac{\partial^2 \log f(x|\mu)}{\partial \mu^2}\right)$$
$$= 1$$

so $\frac{1}{nI_0(\mu)} = \frac{1}{n}$. Therefore $\bar{X}$ is an efficient estimator of μ.

Use the CLT for the MLE

Rather than doing all those steps every time, though, we usually just use the fact that the MLE is asymptotically efficient. This is for the same reason as in the case of unbiasedness: except in really simple examples (e.g. linear combinations of independent random variables), computing the variance of an estimator is hard.

Note on Efficiency

Given that we want an unbiased estimator, the CRLB shows us the optimal variance we can get. So we pick the unbiased estimator that achieves this variance.

If we allow a little bias in our estimator, we could maybe get something with much lower variance. This is a commonly used technique in statistical modelling.