STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 12

- Central limit theorem allows us to directly approximate the sampling distribution of an estimator if it can be written as the sum of i.i.d. random variables.
- This scenario occurs very often. For example, if we are to estimate the mean μ of $N(\mu, \sigma^2)$ using n i.i.d. observations, the MLE for μ is
$$\hat{\mu} = \bar{X}.$$
- Similarly, if we are estimating the mean θ of an Exponential$(1/\theta)$ distribution.

- However, as we have seen, sometimes the MLE is not the sum of random variables.
- For example if we are estimating λ of an Exponential(λ) distribution using n independent observations $X_1, X_2, \ldots, X_n$ from this distribution,
$$\hat{\lambda} = \frac{1}{\bar{X}}.$$
- Can we still find the approximate sampling distribution of $\hat{\lambda}$?
- It turns out that MLEs are quite generally (not always though!) approximately normally distributed, regardless of whether they are sums or not.

Fisher's approximation

(1) If the data are i.i.d. observations $X_1, X_2, \ldots, X_n$ from some distribution $f(x|\theta)$, and

(2) the MLE for $\hat{\theta}$ is *found by solving*

$$\frac{d}{d\theta}L(\theta) = 0 \quad \text{or} \quad \frac{d}{d\theta}\log L(\theta) = 0$$

then for large n, $\hat{\theta}$ is approximately

$$N\left(\theta, \frac{\tau^2(\theta)}{n}\right)$$

where $\tau^2(\theta)$ is such that

$$\frac{1}{\tau^2(\theta)} := I(\theta) = E_\theta\left(\left(\frac{d}{d\theta}\log f(X_1|\theta)\right)^2\right).$$

provided that $0 < \tau^2(\theta) < \infty$.

Fisher's result shows that

- When n is large, the approximate sampling distribution of the MLE $\hat{\theta}$ has mean θ.
- The variance $\frac{\tau^2(\theta)}{n}$ decreases to zero as n increases to infinity.
- So when we have a lot of data, $\hat{\theta}$ "converges" to the actual unknown parameter θ.
- $I(\theta)$ is called *Fisher's information* (from/in a single observation).

One can show that

$$I(\theta) = E_\theta\left(\left(\frac{d}{d\theta}\log f(X_1|\theta)\right)^2\right) = -E_\theta\left(\frac{d^2}{d\theta^2}\log f(X_1|\theta)\right),$$

where

$$E_\theta\left(\frac{d^2}{d\theta^2}\log f(X_1|\theta)\right) = \int_{-\infty}^{\infty}\left(\frac{d^2}{d\theta^2}\log f(x|\theta)\right)f(x|\theta)dx.$$

- This result can become very handy.
- Either way we can compute $I(\theta)$ but sometimes one is easier than the other to compute.

Fisher's information

- It quantifies the amount of "information" one can get from each individual observation for estimating θ.
- Mathematically, it is the *average* "curvature" of the likelihood function at θ, *average* means over repeated experiments with θ being the truth.
- (Draw a figure.)
- Intuitively, the steeper $\log L(\theta)$ is, the more certain we are about our "best" guess for θ. That is why the information $I(\theta)$ is in the *denominator* of the variance.
- In the extreme case, if $\log L(\theta)$ is flat, $I(\theta) = 0$—no information.

A corollary: Consistency of MLEs

- Fisher's approximation implies that as we get more and more observations, the MLE $\hat{\theta}$ "converges" to θ. (Recall the law of large number for $\bar{X}$.)
- The reason is as below. If

$$\hat{\theta} \sim_{approx} N\left(\theta, \frac{\tau^2(\theta)}{n}\right)$$

then

$$\begin{aligned} P(|\hat{\theta} - \theta| < \varepsilon) &\approx \Phi\left(\frac{\sqrt{n}\varepsilon}{\tau(\theta)}\right) - \Phi\left(-\frac{\sqrt{n}\varepsilon}{\tau(\theta)}\right) \\ &= \int_{-\frac{\sqrt{n}\varepsilon}{\tau(\theta)}}^{\frac{\sqrt{n}\varepsilon}{\tau(\theta)}} \phi(x)dx \\ &\approx 1 \quad \text{for large } n. \end{aligned}$$

Consistency of MLEs

- Therefore as n increases, $\hat{\theta}$ will be as close to θ as desired *with high probability*.
- We say that $\hat{\theta}$ converges to θ *in probability*, donoted as

$$\hat{\theta} \to_P \theta.$$

- This property of the MLE is called *consistency*.

Efficiency of MLEs

- For large n, from Fisher's approximation we see that the MSE of an MLE
$$MSE_{\hat{\theta}}(\theta) \approx \frac{\tau^2(\theta)}{n} = \frac{1}{nI(\theta)}.$$
- Under some further conditions, one can show that *no other* estimator can have an approximating (sampling) distribution with MSE smaller than $\tau^2(\theta)/n$ as n becomes very large.
- This property is sometimes called the (asymptotic) *efficiency* of the MLE.
- So for many (not all!) problems involving a large number of observations, the MLE will do as well as possible in terms of MSE.
- The quantity $\frac{1}{nI(\theta)}$ is sometimes referred to as the *Cramer-Rao lower-bound*.

Example I: Normal data with unknown mean and known variance

- Suppose our data are i.i.d. observations from $N(\mu, \sigma^2)$ where σ^2 is known and we wish to estimate μ.
- From our earlier classes we know that the MLE for μ is

  $$\hat{\mu} = \bar{X}$$

  and in this case the sampling distribution of $\hat{\mu}$ is *exactly* $N(\mu, \sigma^2/n)$.
- What does Fisher's approximation say in this case?

$$\log f(X_1|\mu) = \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_1-\mu)^2}{2\sigma^2}}\right)$$
$$= -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{(X_1-\mu)^2}{2\sigma^2}.$$

To apply Fisher's approximation, we first find

$$\frac{d}{d\mu}\log f(X_1|\mu) = \frac{X_1-\mu}{\sigma^2}.$$

Therefore

$$I(\mu) = E\left[\left(\frac{d}{d\mu}\log f(X_1|\mu)\right)^2\right]$$
$$= E\left[\left(\frac{X_1-\mu}{\sigma^2}\right)^2\right] = \frac{E[(X_1-\mu)^2]}{\sigma^4} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

Alternatively,

Note that

$$\frac{d^2}{d\mu^2}\log f(X_1|\mu) = -\frac{1}{\sigma^2}.$$

Hence

$$I(\mu) = -E\left(\frac{d^2}{d\mu^2}\log f(X_1|\mu)\right) = \frac{1}{\sigma^2}.$$

The computation is quite a bit easier this way!

Either way, we can now apply Fisher's approximation,

$$\frac{\tau^2(\mu)}{n} = \frac{1}{n\,I(\mu)} = \frac{\sigma^2}{n}.$$

So Fisher's approximation says that for large n,

$$\hat{\mu} \sim_{approx} N(\mu, \sigma^2/n).$$

We know the RHS is the exact sampling distribution of $\hat{\mu}$!

Another example

- The data are i.i.d. observations $X_1, X_2, \ldots, X_n$ from an Exponential(λ) distribution. So the p.d.f for each X_i is

$$\begin{aligned} f(x|\lambda) &= \lambda e^{-\lambda x} \quad \text{for } x > 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

- We have found the MLE for λ to be

$$\hat{\lambda} = \frac{1}{\bar{X}}.$$

- This is not the sum of independent random variables so we can't apply CLT.
- But Fisher's approximation still applies.

To apply Fisher's approximation, first we have

$$\log f(X_1|\lambda) = \log(\lambda) - \lambda X_1.$$

So

$$\frac{d}{d\lambda}\log f(X_1|\lambda) = \frac{1}{\lambda} - X_1.$$

Therefore

$$I(\lambda) = E\left[\left(\frac{d}{d\lambda}\log f(X_1|\lambda)\right)^2\right] = E\left[\left(X_1 - \frac{1}{\lambda}\right)^2\right] = \mathrm{Var}(X_1) = \frac{1}{\lambda^2}.$$

Alternatively,

$$\frac{d^2}{d\lambda^2}\log f(X_1|\lambda) = -\frac{1}{\lambda^2}.$$

So

$$I(\lambda) = -E\left(\frac{d^2}{d\lambda^2}\log f(X_1|\lambda)\right) = \frac{1}{\lambda^2}.$$

Therefore,

$$\tau^2(\lambda) = \frac{1}{I(\lambda)} = \lambda^2.$$

By Fisher's approximation,

$$\hat{\lambda} \sim_{approx} N\left(\lambda, \frac{\lambda^2}{n}\right).$$

- To see how good this approximation is, let us compare this approximate sampling distribution to the exact sampling distribution of $\hat{\lambda}$.

We can find the exact distribution of $\hat{\lambda}$ as follows.

- The X_i's are independent Exponential(λ), or Gamma$(1, \lambda)$ random variables.
- By the method we have learned for finding the distribution of sums of independent random variables, we can show that

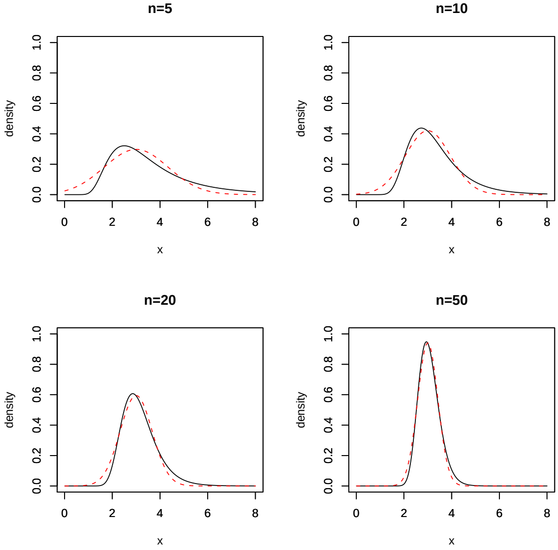
$$\sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \lambda).$$

Accordingly, by a change of variable we get

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim \text{Gamma}(n, n\lambda)$$

- By another change of variable we can find the p.d.f of $\hat{\lambda} = \frac{1}{\bar{X}}$.
- Its sampling distribution is called the *inverse-Gamma$(n, n\lambda)$ distribution*.

Exact sampling distribution for $\hat{\lambda}$ vs Fisher's approximation when $\lambda = 3$



Black solid is the exact pdf. Red dashed is Fisher's approximation.

Condition for Fisher's approximation to apply

- The specific conditions are beyond the scope of this class.
- Roughly speaking the MLE must be a root for the equation

$$\frac{d}{d\theta} \log L(\theta) = 0.$$

- Otherwise the approximation does not apply. Let's look at an example.

Example

- Suppose the data are i.i.d. observations from a Uniform$(0, \theta)$ distribution. So for each X_i the p.d.f is

$$
\begin{aligned}
f(x|\theta) &= \frac{1}{\theta} \quad \text{for } 0 \le x \le \theta \\
&= 0 \quad \text{otherwise.}
\end{aligned}
$$

- What is the MLE?
- The likelihood is

$$
\begin{aligned}
L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) &= \frac{1}{\theta^n} \quad \text{for } \theta \ge \max(x_1, x_2, \ldots, x_n) \\
&= 0 \quad \text{otherwise.}
\end{aligned}
$$

(Draw a figure.)

- So the value of θ that maximizes the likelihood is

$$\hat{\theta} = \max(x_1, x_2, \ldots, x_n).$$

- Thus the MLE is

$$\hat{\theta} = \max(X_1, X_2, \ldots, X_n).$$

- Note that we did not solve for this MLE using

$$\frac{d}{d\theta} L(\theta) \quad \text{or} \quad \frac{d}{d\theta} \log L(\theta) = 0.$$

- In fact,

$$\frac{d}{d\theta} L(\theta) = -\frac{n}{\theta^{n+1}} \neq 0 \quad \text{and} \quad \frac{d}{d\theta} \log L(\theta) = -\frac{n}{\theta} \neq 0$$

for any θ.

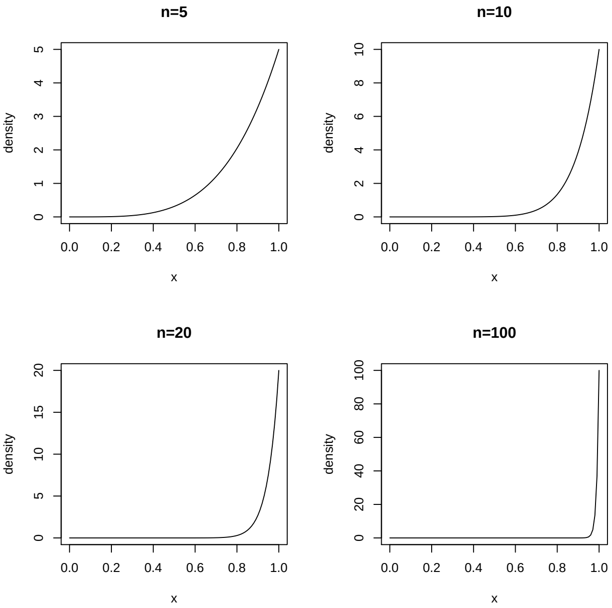
- What is the sampling distribution of $\hat{\theta}$?
- Let us first find the c.d.f of $\hat{\theta}$.

$$
\begin{aligned}
F_{\hat{\theta}}(y) &= P(\hat{\theta} \leq y) \\
&= P(\max(X_1, X_2, \ldots, X_n) \leq y) \\
&= P(X_1 \leq y, X_2 \leq y, \ldots, X_n \leq y) \\
&= \prod_{i=1}^{n} P(X_i \leq y) \\
&= \begin{cases} 0 & \text{if } y < 0 \\ \left(\frac{y}{\theta}\right)^n & \text{if } 0 \leq y \leq \theta \\ 1 & \text{if } y > \theta. \end{cases}
\end{aligned}
$$

- So the p.d.f of $\hat{\theta}$ is

$$
\begin{aligned}
f_{\hat{\theta}}(y|\theta) &= \frac{ny^{n-1}}{\theta^n} \quad \text{for } 0 \leq y \leq \theta \\
&= 0 \quad \text{otherwise.}
\end{aligned}
$$

Sampling distribution of $\hat{\theta}$ when $\theta = 1$



Numerical methods for finding MLEs (optional material)

- In the examples we see in this class, the equation

$$\frac{d}{d\theta} \log L(\theta)$$

is relatively easy to solve.

- In real life problems, the likelihood function can be quite complex and so it is often not easy to find the solution to the above equation.
- Numerical methods such as Newton-Raphson's method can be applied in such cases.

Newton-Raphson's method

- Suppose we want to find a solution $\hat{\theta}$ to an equation

$$g(\theta) = 0.$$

- In our current context, the function

$$g(\theta) = \frac{d}{d\theta} \log L(\theta) \quad \text{or} \quad = \frac{d}{d\theta} L(\theta).$$

- For any θ_0, if θ is close to θ_0, then by the mean value theorem (i.e. the first order Taylor expansion)

$$g(\theta) \approx g(\theta_0) + (\theta - \theta_0) g'(\theta_0).$$

- Now consider $\hat{\theta}$, which is a root, i.e., $g(\hat{\theta}) = 0$, and thus

$$-g(\theta_0) \approx (\hat{\theta} - \theta_0) g'(\theta_0).$$

Therefore,

$$\hat{\theta} - \theta_0 \approx -\frac{g(\theta_0)}{g'(\theta_0)}$$

and so

$$\hat{\theta} \approx \theta_0 - \frac{g(\theta_0)}{g'(\theta_0)}.$$

(Draw a figure.)

- The RHS gives an approximation to the root $\hat{\theta}$ when θ_0 is close to the root.
- But how do we know a particular value of θ_0 is close to $\hat{\theta}$ to apply this approximation?
- We don't. But this suggests an iterative procedure.

Starting from an initial guess at the root, (we denote this initial guess by $\hat{\theta}_0$), for $n = 1, 2, \ldots$, we compute

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \frac{g(\hat{\theta}_n)}{g'(\hat{\theta}_n)},$$

until the estimate changes little.

(Show movie at
`http://www.youtube.com/watch?v=r3KXzyGS2zg`)