Asymptotics & optimality
(Module 11)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

School of Mathematics and Statistics
University of Melbourne

Semester 2, 2019

Aims of this module

- Explain some of the theory that we skipped in previous modules
- Show why the MLE is usually a good (or best) estimator
- Explain some related important theoretical concepts

Outline

Previous claims (from modules 2 & 4)

The MLE is asymptotically:

- unbiased
- efficient (has the optimal variance)
- normally distributed

Can use the 2nd derivative of the log-likelihood (the ‘observed information function’) to get a standard error for the MLE.

Motivating example (non-zero binomial)

- Consider a factory producing items in batches. Let θ denote the proportion of defective items. From each batch 3 items are sampled at random and the number of defectives is determined. However, records are only kept if there is at least one defective.
- Let Y be the number of defectives in a batch.
- Then $Y \sim \mathrm{Bi}(3, \theta)$,

$$\Pr(Y = y) = \binom{3}{y} \theta^y (1-\theta)^{3-y}, \quad y = 0, 1, 2, 3$$

- But we only take an observation if $Y > 0$, so the pmf is

$$\Pr(Y = y \mid Y > 0) = \frac{\binom{3}{y} \theta^y (1-\theta)^{3-y}}{1 - (1-\theta)^3}, \quad y = 1, 2, 3$$

- Let X_i be the number of times we observe i defectives and let $n = X_1 + X_2 + X_3$ be the total number of observations.
- The likelihood is,

$$L(\theta) = \frac{n!}{x_1!x_2!x_3!} \left(\frac{3\theta(1-\theta)^2}{1-(1-\theta)^3} \right)^{x_1} \left(\frac{3\theta^2(1-\theta)}{1-(1-\theta)^3} \right)^{x_2} \left(\frac{\theta^3}{1-(1-\theta)^3} \right)^{x_3}$$

- This simplifies to,

$$L(\theta) \propto \frac{\theta^{x_1+2x_2+3x_3}(1-\theta)^{2x_1+x_2}}{(1-(1-\theta)^3)^n}$$

- After taking logarithms and derivatives, the MLE is found to be the smaller root of

$$t\theta^2 - 3t\theta + 3(t-n) = 0$$

where $t = x_1 + 2x_2 + 3x_3$.
- This gives:

$$\hat{\theta} = \frac{3t - \sqrt{-3t^2 + 12tn}}{2t}$$

- We now have the MLE. . .
- . . . but finding its sampling distribution is not straightforward!
- In general, finding the exact distribution of a statistic is often difficult.
- We've used the Central Limit Theorem to approximate the distribution of the sample mean.
- Gave us approximate CIs for a population mean μ of the form,

$$\bar{x} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \times \frac{s}{\sqrt{n}}$$

- Similar results hold more generally for MLEs (and other estimators)

Definitions

- Start with the log-likelihood:

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

- Taking the first derivative gives the score function (also known simply as the score). Let's call it U,

$$U(\theta) = \frac{\partial \ell}{\partial \theta}$$

- Note: we solve $U(\hat{\theta}) = 0$ to get the MLE

- Taking the second derivative, and then it's negative, gives the observed information function (also known simply as the observed information). Let's call it V,

$$V(\theta) = -\frac{\partial U}{\partial \theta} = -\frac{\partial^2 \ell}{\partial \theta^2}$$

- This represents the **curvature** of the log-likelihood. Greater curvature $\Rightarrow$ narrower likelihood around a certain value $\Rightarrow$ the likelihood is more **informative**.

Fisher information

- All of the above are functions of the data (and parameters). Therefore they are random variables and have sampling distributions.
- For example, we can show that $\mathbb{E}(U(\theta)) = 0$.
- An important quantity is $I(\theta) = \mathbb{E}(V(\theta))$, which is the Fisher information function (or just the Fisher information). It is also known as the expected information function (or simply as the expected information).
- Many results are based on the Fisher information.
- For example, we can show that $\mathrm{var}(U(\theta)) = I(\theta)$.
- More importantly, it arises in theory about the distribution of the MLE.

Asymptotic distribution

- The following is a key result:

$$\hat{\theta} \approx \mathrm{N}\left(\theta, \frac{1}{I(\theta)}\right) \quad \text{as } n \to \infty$$

- It requires some conditions for it to hold. The main one being that the parameter should not be defining a boundary of the sample space (e.g. like in the boundary problem examples we've looked at).
- Let's see a proof...

Asymptotic distribution (derivation)

- Assumptions:
  - $X_1, \ldots, X_n$ is a random sample from $f(x, \theta)$
  - Continuous pdf, $f(x, \theta)$
  - θ is not a boundary parameter
- Suppose the MLE satisfies:

$$U(\hat{\theta}) = \frac{\partial \ln L(\hat{\theta})}{\partial \theta} = 0$$

  Note: this requires that θ is not a boundary parameter.
- Taylor series approximation for $U(\hat{\theta})$ about θ:

$$\begin{aligned} 0 = U(\hat{\theta}) = \frac{\partial \ln L(\hat{\theta})}{\partial \theta} &\approx \frac{\partial \ln L(\theta)}{\partial \theta} + (\hat{\theta} - \theta) \frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \\ &= U(\theta) - (\hat{\theta} - \theta) V(\theta) \end{aligned}$$

- We can write this as:

$$V(\theta)\,(\hat{\theta} - \theta) \approx U(\theta)$$

- Remember that we have a random sample (iid rvs), so we have,

$$U(\theta) = \frac{\partial \ln L(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(X_i, \theta)}{\partial \theta}$$

- Since the X_i are iid so are:

$$U_i = \frac{\partial \ln f(X_i, \theta)}{\partial \theta}, \quad i = 1, \ldots, n.$$

- And the same for:

$$V_i = -\frac{\partial^2 \ln f(X_i, \theta)}{\partial \theta^2}, \quad i = 1, \ldots, n.$$

- Determine $\mathbb{E}(U_i)$ by integration by substitution and exchanging the order of integration and differentiation,

$$\begin{aligned}\mathbb{E}(U_i) &= \int_{-\infty}^{\infty} \frac{\partial \ln f(x,\theta)}{\partial \theta} f(x,\theta)\, dx = \int_{-\infty}^{\infty} \frac{\partial f(x,\theta)}{\partial \theta} \frac{f(x,\theta)}{f(x,\theta)}\, dx \\ &= \int_{-\infty}^{\infty} \frac{\partial f(x,\theta)}{\partial \theta}\, dx = \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f(x,\theta)\, dx = \frac{\partial}{\partial \theta} 1 = 0\end{aligned}$$

- To get the variance of U_i, we start with one of the above results,

$$\int_{-\infty}^{\infty} \frac{\partial \ln f(x,\theta)}{\partial \theta} f(x,\theta) dx = 0$$

- Taking another derivative of both sides gives,

$$\int_{-\infty}^{\infty} \left\{ \frac{\partial^2 \ln f(x,\theta)}{\partial \theta^2} f(x,\theta) + \frac{\partial \ln f(x,\theta)}{\partial \theta} \frac{\partial f(x,\theta)}{\partial \theta} \right\} dx = 0$$

- But,

$$\frac{\partial f(x,\theta)}{\partial \theta} = \frac{\partial \ln f(x,\theta)}{\partial \theta} f(x,\theta)$$

- Combining the previous two equations gives,

$$\int_{-\infty}^{\infty} \left\{ \frac{\partial \ln f(x,\theta)}{\partial \theta} \right\}^2 f(x,\theta)\, dx = -\int_{-\infty}^{\infty} \frac{\partial^2 \ln f(x,\theta)}{\partial \theta^2} f(x,\theta)\, dx$$

- In other words,

$$\mathbb{E}(U_i^2) = \mathbb{E}(V_i)$$

- Since $\mathbb{E}(U_i) = 0$ we also have $\mathbb{E}(U_i^2) = \mathrm{var}(U_i)$, so we can conclude,

$$\mathrm{var}(U_i) = \mathbb{E}(V_i)$$

- Thus $U = \sum_i U_i$ is the sum of iid rvs with mean 0 and this variance.
- Thus,

$$\mathrm{var}(U) = n\,\mathbb{E}(V_i)$$

- Also, since $V = \sum_i V_i$, we can conclude that,

$$\mathbb{E}(V) = n\,\mathbb{E}(V_i)$$

- Note that this is just the Fisher information, i.e.

$$\mathbb{E}(V) = \mathrm{var}(U) = I(\theta)$$

- Looking back at,
$$V(\theta)\,(\hat{\theta} - \theta) \approx U(\theta)$$
We want to know what happens to U and V as the sample size gets large.
- U has mean 0 and variance $I(\theta)$.
- Central Limit Theorem $\Rightarrow U \approx \mathrm{N}(0, I(\theta))$.
- V has mean $I(\theta)$.
- Law of Large Numbers $\Rightarrow V \to I(\theta)$
- Putting these together gives, as $n \to \infty$,
$$I(\theta)\,(\hat{\theta} - \theta) \sim \mathrm{N}(0, I(\theta))$$

- Equivalently,

$$\hat{\theta} \sim \mathrm{N}\left(\theta, \frac{1}{I(\theta)}\right)$$

- This is a very powerful result. For large (or even modest) samples we do not need toa find the exact distribution of the MLE but can use this approximation.
- In other words, as a standard error of the MLE we can use:

$$\mathrm{se}(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}}$$

if we know $I(\theta)$, or otherwise replace it with it's realised (observed) version,

$$\mathrm{se}(\hat{\theta}) = \frac{1}{\sqrt{V(\hat{\theta})}}$$

- Furthermore, we use the normal distribution to construct approximate confidence intervals.

Example (exponential distribution)

- $X_1, \ldots, X_n$ random sample from

$$f(x \mid \theta) = \frac{1}{\theta} e^{-x/\theta}, \quad 0 < x < \infty, \quad 0 < \theta < \infty$$

- MLE is $\bar{X}$.
- $\ln f(x \mid \theta) = -\ln \theta - x/\theta$, so

$$\begin{aligned} U_i(\theta) &= \frac{\partial}{\partial \theta} \ln f(x \mid \theta) = -\frac{1}{\theta} + \frac{x}{\theta^2} \\ V_i(\theta) &= -\frac{\partial^2}{\partial \theta^2} \ln f(x \mid \theta) = -\frac{1}{\theta^2} + \frac{2x}{\theta^3} \end{aligned}$$

- Since $\mathbb{E}(X) = \theta$,

$$I_i(\theta) = \mathbb{E}(V_i(\theta)) = \mathbb{E}\left(-\frac{1}{\theta^2} + \frac{2X}{\theta^3}\right) = -\frac{1}{\theta^2} + \frac{2\theta}{\theta^3} = \frac{1}{\theta^2}$$

- Then $I(\theta) = n/\theta^2$ and $\hat{\theta} \approx \mathrm{N}(\theta, \theta^2/n)$
- Suppose we observe $n = 20$ and $\bar{x} = 3.7$. An approximate 95% CI is,

$$3.7 \pm 1.96\sqrt{\frac{3.7^2}{20}} = (2.1, 5.3)$$

Example (Poisson distribution)

- Same arguments hold for discrete distributions, e.g. $\mathrm{Pn}(\lambda)$.

$$f(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, \ldots, \quad \lambda > 0$$

  We have seen $\hat{\lambda} = \bar{X}$.

- $\ln f(x \mid \lambda) = x \ln \lambda - \lambda - \ln(x!)$, so

$$\frac{\partial \ln f(x \mid \lambda)}{\partial \lambda} = \frac{x}{\lambda} - 1, \quad \text{and} \quad \frac{\partial^2 \ln f(x \mid \lambda)}{\partial \lambda^2} = -\frac{x}{\lambda^2}$$

- Thus

$$-\mathbb{E}\left(-\frac{X}{\lambda^2}\right) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

- Then $\hat{\lambda} \approx \mathrm{N}(\lambda, \lambda/n)$
- Suppose we observe $n = 40$ and $\bar{x} = 2.225$. An approximate 90% CI is,

$$2.225 \pm 1.645\sqrt{\frac{2.225}{40}} = (1.837, 2.612)$$

Cramér–Rao lower bound

- How good can our estimator get?
- Suppose we know that it is unbiased.
- What is the minimum variance we can achieve?
- Under similar assumptions to before (esp. the parameter must not define a boundary), we can find a lower bound on the variance
- This is known as the Cramér–Rao lower bound
- It is equal to the asymptotic variance of the MLE.
- In other words, if we take any unbiased estimator T, then

$$\mathrm{var}(T) \geqslant \frac{1}{I(\theta)}$$

Cramér–Rao lower bound (proof)

- Let T be an unbiased estimator of θ
- Consider its covariance with the score function,

$$\begin{aligned}
\operatorname{cov}(T, U) &= \mathbb{E}(TU) - \mathbb{E}(T)\,\mathbb{E}(U) = \mathbb{E}(TU) \\
&= \int T \frac{\partial \ln L}{\partial \theta} L \, d\mathbf{x} = \int T \frac{\partial L}{\partial \theta} \, d\mathbf{x} \\
&= \frac{\partial}{\partial \theta} \int T L \, d\mathbf{x} = \frac{\partial}{\partial \theta} \, \mathbb{E}(T) = \frac{\partial}{\partial \theta} \theta = 1
\end{aligned}$$

- Using the fact that $\operatorname{cor}(T, U)^2 \leqslant 1$,

$$\begin{aligned}
\operatorname{cov}(T, U)^2 &\leqslant \operatorname{var}(T) \operatorname{var}(U) \\
\operatorname{var}(T) &\geqslant \frac{1}{\operatorname{var}(U)} = \frac{1}{I(\theta)}
\end{aligned}$$

Implications of the Cramér–Rao lower bound

- If an unbiased estimator attains this bound, then it is best in the sense that it has minimum variance compared with other unbiased estimators.
- Therefore, MLEs are approximately (or exactly) optimal for large sample size because:
  - They are asymptotically unbiased
  - Their variance meets the Cramér–Rao lower bound asymptotically

Efficiency

- We can compare any unbiased estimator against the lower bound
- We define the efficiency of the unbiased estimator T as its variance relative to the lower bound,

$$\mathsf{eff}(T) = \frac{1/I(\theta)}{\mathrm{var}(T)} = \frac{1}{I(\theta)\,\mathrm{var}(T)}$$

- Note that $0 \leqslant \mathsf{eff}(T) \leqslant 1$
- If $\mathsf{eff}(T) \approx 1$ we say that T is an efficient estimator

Example (exponential distribution)

- Sampling from an exponential distribution
- We saw that $I(\theta) = n/\theta^2$
- Therefore, the Cramér–Rao lower bound is θ^2/n.
- Any unbiased estimator must have variance at least as large as this.
- The MLE in this case is the sample mean, $\hat{\theta} = \bar{X}$
- Therefore, $\mathrm{var}(\hat{\theta}) = \mathrm{var}(X)/n = \theta^2/n$
- So the MLE is efficient (for all sample sizes!)

Outline

Sufficiency: a starting example

- We toss a coin 10 times
- Want to estimate the probability of heads, θ
- $X_i \sim \mathrm{Be}(\theta)$
- Suppose we use $\hat{\theta} = \frac{1}{2}(X_1 + X_2)$
- Only uses the first 2 coin tosses
- Clearly, we have not used all of the available information!

Motivation

- Point estimation reduces the whole sample to a few statistics.
- Different methods of estimation can yield different statistics.
- Is there a preferred reduction?
- Toss a coin with probability of heads θ 10 times.
  Observe T H T H T H H T T T.
- Intuitively, knowing we have 4 heads in 10 tosses is all we need.
- But are we missing something? Does the length of the longest run give extra information?

Definition

- Intuition: want to find a statistic so that any **other** statistic provides no additional information about the value of the parameter
- Definition: the statistic $T = g(X_1, \ldots, X_n)$ is sufficient for an underlying parameter θ if the conditional probability distribution of the data $(X_1, \ldots, X_n)$, given the statistic $u(X_1, \ldots, X_n)$, does not depend on the parameter θ.
- Sometimes need more than one statistic, e.g. T_1 and T_2, in which case we say they are jointly sufficient for θ

Example (binomial)

- The pdf is, $f(x \mid p) = p^x(1-p)^{1-x}, \quad x = 0, 1$
- The likelihood is,

$$\prod_{i=1}^{n} f(x_i \mid p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

- Let $Y = \sum X_i$, we have that $Y \sim \mathrm{Bi}(n, p)$ and then,

$$\begin{aligned}
&\Pr(X_1 = x_1, \ldots, X_n = x_n \mid Y = y) \\
&\quad = \frac{\Pr(X_1 = x_1, \ldots, X_n = x_n)}{\Pr(Y = y)} \\
&\quad = \frac{p^{x_1}(1-p)^{1-x_1} \ldots p^{x_n}(1-p)^{1-x_n}}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{1}{\binom{n}{y}}
\end{aligned}$$

- Given $Y = y$, the conditional distribution of $X_1, \cdots X_n$ does not depend on p.
- Therefore, Y is sufficient for p.

Factorisation theorem

- Let $X_1, \dots, X_n$ have joint pdf or pmf $f(x_1, \dots, x_n \mid \theta)$
- $Y = g(x_1, \dots, x_n)$ is sufficient for θ if and only if

$$f(x_1, \dots, x_n \mid \theta) = \phi\{g(x_1, \dots, x_n) \mid \theta\}\, h(x_1, \dots, x_n)$$

- ϕ depends on $x_1, \cdots, x_n$ only through $g(x_1, \cdots, x_n)$ and h doesn't depend on θ.

Example (binomial)

- The pdf is, $f(x \mid p) = p^x(1-p)^{1-x}, \quad x = 0, 1$
- The likelihood is,

$$\prod_{i=1}^{n} f(x_i \mid p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

- So $y = \sum x_i$ is sufficient for p, since we can factorise the likelihood into:

$$\phi(y, p) = p^y(1-p)^{n-y} \quad \text{and} \quad h(x_1, \cdots, x_n) = 1$$

- So in the coin tossing example, the total number of heads is sufficient for θ.

Example (Poisson)

- $X_1, \ldots, X_n$ random sample from a Poisson distribution with mean λ.
- The likelihood is,

$$\prod_{i=1}^{n} f(x_i \mid \lambda) = \frac{\lambda^{\sum x_i} e^{-n\lambda}}{x_1! \ldots x_n!} = (\lambda^{n\bar{x}} e^{-n\lambda}) \left(\frac{1}{x_1! \ldots x_n!} \right)$$

- We see that $\bar{X}$ is sufficient for λ.

Exponential family of distributions

- We often use distributions which have pdfs of the form:

$$f(x \mid \theta) = \exp\{K(x)p(\theta) + S(x) + q(\theta)\}$$

- This is called the exponential family.
- Let $X_1, \cdots, X_n$ be iid from an exponential family. Then $\sum_{i=1}^{n} K(X_i)$ is sufficient for θ.
- To prove this note that the joint pdf is

$$\begin{aligned} &\exp\left\{p(\theta)\sum K(x_i) + \sum S(x_i) + nq(\theta)\right\} \\ &\quad = \left[\exp\left\{p(\theta)\sum K(x_i) + nq(\theta)\right\}\right] \exp\left\{\sum S(x_i)\right\} \end{aligned}$$

- The factorisation theorem then shows sufficiency.

Example (exponential)

- The pdf is,

$$f(x \mid \theta) = \frac{1}{\theta} e^{-x/\theta} = \exp\left[x\left(-\frac{1}{\theta}\right) - \ln\theta\right], \quad 0 < x < \infty$$

- This is of the form

$$f(x \mid \theta) = \exp\{K(x)p(\theta) + S(x) + q(\theta)\}$$

- So $K(x) = x$ and $\sum X_i$ is sufficient for θ (and so is $\bar{X} = \sum X_i/n$).

Sufficiency and MLEs

- If there exist sufficient statistics, the MLE will be a function of them.
- Factorise the likelihood:

$$L(\theta) = f(x_1, \ldots, x_n \mid \theta) = \phi\{g(x_1, \ldots, x_n\} \mid \theta]\, h(x_1, \ldots, x_n)$$

- We find the MLE by maximizing $\phi\{g(x_1, \ldots, x_n) \mid \theta\}$ which is a function of the sufficient statistics and θ
- So the MLE must be a function of the sufficient statistics

Importance of sufficiency

- Why are sufficient statistics important?
- Once the sufficient statistics are known there is no additional information on the parameter in the sample
- Samples that have the same values of the sufficient statistic yield the same estimates
- The optimal estimators/tests are based on sufficient statistics (such as the MLE)
- A lot of statistical theory is based on them
- Easy to find the sufficient statistics in some special cases (e.g. exponential family)

Disclaimer

- But… the concept of sufficiency relies on **knowing the population distribution**
- So, it is mostly important for theoretical work.
- In practice, we want to also look at all aspects of our data
- That is, we should go beyond any putative sufficient statistics, as a sanity check of our assumptions (e.g. QQ plots).

Outline

Previous claims (from module 8)

- The likelihood ratio test (LRT) gives the optimal test
- The likelihood ratio has a known distribution

Neyman–Pearson lemma

- Comparing **simple** hypotheses:

$$H_0\colon \theta = \theta_0 \quad \text{versus} \quad H_1\colon \theta = \theta_1$$

- The Neyman–Pearson lemma states that the most powerful test, for a given significance level, is the LRT
- (Proof of lemma not shown)

Uniformly most powerful tests

- Now consider a **composite alternative** hypothesis,

$$H_1 \colon \theta \in A_1$$

- If the same test (from the LRT) is most powerful for all $\theta_1 \in A_1$, then we say it is uniformly most powerful for $\theta_1 \in A_1$.
- If the form of the LRT **differs** for different values of θ_1, then any given one will only be the best for particular values of θ_1.
- If so, then we do not have a uniformly best test.
- But any given test might still be a reasonably good test for other values of θ_1

Asymptotic distribution of the likelihood ratio*

- Consider the test,

$$H_0\colon \theta = \theta_0 \quad \text{versus} \quad H_1\colon \theta \neq \theta_0$$

- The likelihood ratio is,

$$\lambda = \frac{L_0}{L_1} = \frac{L(\theta_0)}{L(\hat{\theta})}$$

- The function $2\ln(\lambda)$ asymptotically follows a χ^2_1 distribution
- This can be used to set up approximate hypothesis tests
- Is often used to formally compare different models