STAT0008 Lecture 6
Likelihood and Bayesian Methods of Point Estimation

Dr. Aidan O'Keeffe

Department of Statistical Science
University College London

12th November 2018

Outline

- Brief re-cap of maximum likelihood estimation.
- Large sample properties of maximum likelihood estimators.
- Bayesian approach to parametric inference.
- Choice of prior distribution.
- Bayesian point estimation.

Maximum Likelihood Estimation

Recall: Suppose we have a sample of data $X_1, \ldots, X_n$ with $X_i \sim \mathcal{D}(\theta)$ and the joint pdf/pmf of $(X_1, \ldots X_n)^\intercal$ is given by $f(\mathbf{x}; \theta)$.

The likelihood function is considered a function of θ and is written

$$\mathcal{L}(\theta \mid \mathbf{x}) = f(\mathbf{x}; \theta).$$

The maximum likelihood estimate of θ is simply the value of θ that maximises the likelihood function $\mathcal{L}(\theta \mid \mathbf{x})$. We call the maximum likelihood estimate $\hat{\theta}$ and write

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, \mathcal{L}(\theta \mid \mathbf{x}).$$

Maximum Likelihood Estimation

Intuitively, the maximum likelihood estimate is the parameter value that is the most compatible with the observed data. Here, the phrase 'most compatible' means the parameter value that makes the observed data most probable.

Two strengths of maximum likelihood estimation are:

1. It is an objective method for constructing estimators.
2. In large samples, at least, the maximum likelihood estimator has desirable properties.

However, for small samples there is no guarantee that the mle will be a desirable estimator.

In general, an mle is not necessarily unbiased and may be inefficient.

Properties of MLEs

We outline some important properties of maximum likelihood estimators:

1. A maximum likelihood estimator is necessarily a function of a minimal sufficient statistic.

   Proof: Follows immediately from the factorisation criterion.

2. If an estimator, $T(\mathbf{X})$, exists such that $T(\mathbf{X})$ is the MVBUE of an unknown parameter, θ, then $T(\mathbf{X})$ is the maximum likelihood estimator of θ.

   Proof: Follows immediately if we write the score function in the linear form

$$\frac{\partial \ell(\theta \mid \mathbf{X})}{\partial \theta} = U(\mathbf{X}; \theta) = \mathcal{I}(\theta) \left(T(\mathbf{X}) - \theta\right).$$

Properties of MLEs

3. If $\hat{\theta}$ is the maximum likelihood estimate of θ, then the maximum likelihood estimate of $g(\theta)$ is $g(\hat{\theta})$, where $g(.)$ is a function of θ.

   More generally, in the multiparameter case (i.e. where our parameter of interest has dimension > 1), if we re-parameterise the likelihood function using functions of the original parameters then the maximum likelihood estimates of our new parameters are the corresponding functions of the maximum likelihood estimates of our original parameters.

   Proof: See next two slides or additional material on the STAT0008 moodle page (proof of the invariance property). Note that this proof is non-examinable.

Proof of Invariance Property

Suppose that $\phi = g(\theta)$ is an invertible function of θ. Then, if $\mathcal{L}(\theta)$ is the likelihood function for θ, we have

$$\begin{aligned}\mathcal{L}(\theta) &= \mathcal{L}(g^{-1}(g(\theta))) \\ &= \mathcal{L}\left(g^{-1}(\phi)\right) \\ &\equiv \tilde{\mathcal{L}}(\phi).\end{aligned}$$

Then

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \theta} &= \frac{\partial \tilde{\mathcal{L}}(\phi)}{\partial \phi}\frac{\partial \phi}{\partial \theta} \\ &= g'(\theta)\frac{\partial \tilde{\mathcal{L}}(\phi)}{\partial \phi} \quad \text{since } \phi = g(\theta).\end{aligned}$$

Proof of Invariance Property

The mle occurs where $\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0$ and this implies that

$$\text{either} \quad g'(\theta) = 0 \quad \text{or} \quad \frac{\partial \tilde{\mathcal{L}}(\phi)}{\partial \phi} = 0.$$

If $g'(\theta) = 0$ then $g(\theta)$ is a constant which is not invertible (and not of interest). Hence, it follows that

$$\frac{\partial \tilde{\mathcal{L}}(\phi)}{\partial \phi} = 0$$

and $g(\hat{\theta}) = \hat{\phi}$ maximises $\tilde{\mathcal{L}}(\phi)$.

Large Sample Properties of MLEs

We assume that we have a regular estimation problem (regularity conditions satisfied - more details on the course moodle page, under 'Additional Material') and we have a sample $X_1, \ldots, X_n$ such that $X_i \sim \mathcal{D}(\theta)$ with a single unknown parameter θ. Let $T_n(\mathbf{X})$ denote an mle based on a sample of size n. Then

1. There exists a unique maximum likelihood estimator of θ.

2. The maximum likelihood estimator is a consistent estimator of θ. That is, for all $\epsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}\left(|T_n(\mathbf{X}) - \theta| > \epsilon\right) = 0.$$

   It follows that the mle is asymptotically unbiased, since

$$\mathbb{E}(T_n(\mathbf{X}) - \theta) \to 0 \quad \text{as} \quad n \to \infty.$$

Large Sample Properties of MLEs

3. The maximum likelihood estimator is asymptotically efficient. That is

$$\lim_{n\to\infty} \text{Var}(T_n(\mathbf{X})) = \frac{1}{\mathcal{I}(\theta)}.$$

In other words, if the sample size is large, the variance of the mle is approximately equal to the Cramér-Rao lower bound.

Since, the mle is also asymptotically unbiased, we infer that the mle is, approximately - for large samples, the MVBUE of θ.

4. The maximum likelihood estimator is asymptotically normally distributed. That is

$$T_n(\mathbf{X}) \sim \mathcal{N}\left(\theta, \frac{1}{\mathcal{I}(\theta)}\right) \quad \text{for large } n.$$

Large Sample Properties: k-dimensional parameter

Asymptotic results also apply in the case of a $k-$dimensional parameter $\boldsymbol{\theta}$. In particular

- The ($k-$dimensional) maximum likelihood estimator $\mathbf{T}_n$ converges in probability to $\boldsymbol{\theta}$ as $n \to \infty$. That is

$$\lim_{n\to\infty} \mathbb{P}\left(\|\mathbf{T}_n - \boldsymbol{\theta}\| > \epsilon\right) = 0 \quad \text{for all } \epsilon > 0.$$

- The ($k-$dimensional) maximum likelihood estimator $\mathbf{T}_n$ has an asymptotic $k-$dimensional multivariate normal distribution such that

$$\mathbf{T}_n \sim \mathcal{N}_k\left(\boldsymbol{\theta}, \mathcal{I}^{-1}(\boldsymbol{\theta})\right) \quad \text{for large } n.$$

Note: For these results to apply, the number of parameters must remain finite as $n \to \infty$.

Asymptotic Distribution of MLE: Example 1

Suppose that $X_1, \ldots, X_n$ are independent and identically distributed random variables where $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Determine the asymptotic distribution of the maximum likelihood estimators $(\hat{\mu}, \hat{\sigma}^2)^\intercal$.

Asymptotic Distribution of MLE: Example 1

Asymptotic Distribution of MLE: Example 1

Asymptotic Distribution of MLE: Example 2

Suppose that $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ for $j = 1, 2$ and $i = 1, \ldots, n$ with the X_{ij} variables independent.

We note that there are two observations per subject with $2n$ observations overall.

In this situation, we note that the number of parameters increases with n, so our standard asymptotic theory does not apply! Let us see what happens in large samples...

Writing $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_n, \sigma^2)^\intercal$, the joint density of $\mathbf{X}$ is

$$f(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{2} (x_{ij} - \mu_i)^2 \right\}.$$

Asymptotic Distribution of MLE: Example 2

The log-likelihood is

$$\ell(\boldsymbol{\theta} \mid \mathbf{x}) = -n \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{2} (x_{ij} - \mu_i)^2.$$

The partial derivatives wrt μ_i are

$$\frac{\partial \ell}{\partial \mu_i} = \frac{1}{\sigma^2} \left[(x_{i1} - \mu_i) + (x_{i2} - \mu_i) \right].$$

Solving the set of score equations for μ_i, the maximum likelihood estimator of μ_i is

$$\hat{\mu}_i = \frac{1}{2} \left(X_{i1} + X_{i2} \right).$$

with $\mathbb{E}(\hat{\mu}_i) = \mu_i$ and $\text{Var}(\hat{\mu}_i) = \frac{\sigma^2}{2}$.

Asymptotic Distribution of MLE: Example 2

Unfortunately, $\hat{\mu}_i$ is not consistent for μ_i and the asymptotic arguments fail! (We can show this fairly easily).

But. . . note that $\hat{\mu}_i$ is a linear combination of X_{i1} and X_{i2} (with X_{i1} and X_{i2} normal random variables), so the **marginal distribution** of $\hat{\mu}_i$ is exactly normal.

Now, let us consider the mle of $\hat{\sigma}^2$.

Asymptotic Distribution of MLE: Example 2

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \sum_{j=1}^{2} (x_{ij} - \mu_i)^2.$$

Then, solving the score equation for $\hat{\sigma}^2$ and $\hat{\boldsymbol{\mu}}$, we obtain

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2} (X_{ij} - \hat{\mu}_i)^2 \\
&= \frac{1}{2n} \sum_{i=1}^{n} \left\{ \left(\frac{X_{i1} - X_{i2}}{2} \right)^2 + \left(\frac{X_{i2} - X_{i1}}{2} \right)^2 \right\} \\
&= \frac{1}{4n} \sum_{i=1}^{n} (X_{i1} - X_{i2})^2.
\end{aligned}$$

Asymptotic Distribution of MLE: Example 2

Then

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}^2) &= \frac{1}{4n}\sum_{i=1}^{n}\mathbb{E}[(X_{i1}-X_{i2})^2] \\
&= \frac{1}{4n}\sum_{i=1}^{n}\mathbb{E}[(X_{i1}-\mu_i+\mu_i-X_{i2})^2] \\
&= \frac{1}{4n}\sum_{i=1}^{n}(\mathsf{Var}X_{i1}+\mathsf{Var}X_{i2}) \\
&= \frac{2n\sigma^2}{4n} = \frac{\sigma^2}{2}.
\end{aligned}$$

We see that the mle of σ^2 is biased. Moreover, it is also not asymptotically unbiased (or consistent) for σ^2. The usual asymptotic arguments fail.

Bayesian Approach to Parametric Inference

Recall the Bayesian approach to statistical inference...

As usual, we have a sample of data $X_1, \ldots, X_n$ with $\mathbf{X} \sim \mathcal{D}(\theta)$ for $\theta \in \Theta$.

When using a Bayesian framework for inference, we assume that the parameter(s) of interest, θ, is a random variable (i.e. θ has a probability distribution).

A **prior distribution** for θ is specified, before considering any data $\mathbf{x}$, and the prior pdf/pmf is denoted $\pi(\theta)$ with $\theta \in \Theta$.

The prior distribution summarises our 'degree of belief' about θ before any data are observed.

Bayesian Approach to Parametric Inference

The observed data $\mathbf{x} = (x_1, \ldots, x_n)^\intercal$ have joint density

$$f(\mathbf{x}; \theta) = \mathcal{L}(\theta \mid \mathbf{x}).$$

The prior distribution and information from the observed data (via the likelihood function) are combined to give the **posterior distribution** $\pi(\theta \mid \mathbf{x})$, where

$$\pi(\theta \mid \mathbf{x}) \propto \pi(\theta) \times \mathcal{L}(\theta \mid \mathbf{x}).$$

The posterior distribution represents our updated beliefs about θ, having observed the sample of data $\mathbf{x}$.

When making inference about θ using a Bayesian approach, including point estimation, we use the posterior distribution of θ.

For example, we might calculate the posterior mean or posterior variance.

Prior Distribution

How do we choose a prior distribution?

If we should like our inferential procedure to be objective, then the use of a prior distribution could be problematic and a significant drawback of using a Bayesian approach.

Nonetheless, choosing an inferential procedure (or a statistical model within a given inferential procedure) is a subjective choice. As such, subjectivity cannot usually be avoided, even with non-Bayesian approaches to inference.

Personal priors: A truly subjective Bayesian statistician would specify a personal prior. Such priors could differ considerably between statisticians. The motivation for such priors may be expert opinion, deep subject knowledge etc.

Prior Distribution: Personal Prior

A personal prior may be derived that corresponds roughly with the statistician/observer's view of the uncertainty about a parameter value. For example...

A clinician is interested in modelling high blood pressure in a group of patients. He is fairly confident that, for the group of patients he is studying, systolic blood pressure would lie between 130mmHg and 160mmHg, with a likely value of 145mmHg.

Suppose we take 'fairly confident' to imply that the clinician is 95% certain that a patient's systolic blood pressure is in the range 130 to 160. Let μ be the population mean systolic blood pressure.

Taking 145 to be the centre of this range (the mean), we could assume that 2 standard deviations for a normal prior $= 160 - 145 = 15$. Then, the clinician's personal prior could be given as

$$\mu \sim \mathcal{N}(145, 7.5^2).$$

Prior Distribution: Personal Prior

Two observers might have two very different personal priors.

However, if these two observers receive the same data, then the corresponding posterior distributions will be more similar than the priors.

Indeed, in any Bayesian inference, as more and more data are acquired, the posterior distribution depends less on the prior and more on the likelihood function.

The Bayesian approach begins to look similar to a likelihood-based approach, **but**. . . the interpretation of a Bayesian analysis is still very different from a likelihood-based analysis.

Uniform Priors

If we don't want to use a personal prior, we might want to choose a prior that is quite vague.

A **uniform prior** (or flat prior) is simply a prior distribution where each possible value of θ is, a priori, equally likely.

If θ takes one of a set of discrete values then this prior would be a discrete uniform distribution.

If θ is over a continuous range of values, then this prior would be a continuous uniform distribution.

Note: Such a prior might not be applicable at all. For example, if $\theta \in (-\infty, \infty)$ then we cannot specify a continuous uniform prior.

Improper Prior

An **improper prior** is a prior distribution such that the integral of the prior pdf (or sum of the prior pmf) over the parameter space Θ is not equal to 1.

In short, a prior distribution that is, in fact, not a probability distribution at all!

Whilst this sounds counter-intuitive, improper priors are sometimes used. For example, we might specify a prior variance v and choose an improper prior

$$\pi(\theta) = \frac{1}{v} \implies \pi(\theta \mid \mathbf{x}) \propto \mathcal{L}(\theta \mid \mathbf{x}).$$

This would amount to giving the likelihood function a Bayesian interpretation. In some cases, this may be acceptable but we note that many statisticians would make strong arguments against the use of an improper prior.

Uniform Prior - Problem of Invariance

In words, the use of a uniform prior implies that 'all possible values of θ are equally likely' (or, we have no knowledge about θ other than its range of possible values).

Then, logically, if we know 'nothing' about θ, then we should know 'nothing' about any function of θ , say $g(\theta)$.

Unfortunately, the prior distribution of $g(\theta)$ is not necessarily uniform!

We can address the problem of invariance by using a minimally informative prior known as **Jeffreys' prior**.

Jeffreys' Prior

Jeffreys' prior is defined as

$$\pi(\theta) \propto \sqrt{\mathcal{I}(\theta)}.$$

When using Jeffreys' prior, the posterior distribution is invariant under re-parameterisation of θ.

Suppose

1. You choose the prior $\pi(\theta) \propto \sqrt{\mathcal{I}(\theta)}$ and obtain the posterior $\pi(\theta \mid \mathbf{x})$.
2. You take the posterior distribution from 1., $\pi(\theta \mid \mathbf{x})$, and transform this distribution to obtain the distribution $\pi(g(\theta) \mid \mathbf{x})$.

Then the distribution in 2., $\pi(g(\theta) \mid \mathbf{x})$, is the same as the posterior that you would obtain by specifying the prior $\pi(g(\theta)) \propto \sqrt{\mathcal{I}(g(\theta))}$ and using Bayes' theorem directly.

Conjugate Priors

A **conjugate prior** is a prior distribution whereby the resulting posterior distribution belongs to the same distributional family as the prior.

We've already seen some examples of conjugate priors (e.g. Tutorial Exercises 2, Qu. 3, Poisson likelihood/sampling distribution with a gamma prior, leads to a gamma posterior).

Often, conjugate priors may be used for mathematical convenience but, in many cases, their use can make a lot of sense if the form of a conjugate prior corresponds to the a priori beliefs about the parameter(s) of interest, θ.

Common Conjugate Priors

Sampling Distribution	**Conjugate Prior**
Normal	Normal
Binomial	Beta
Negative Binomial	Beta
Geometric	Beta
Poisson	Gamma
Exponential	Gamma
Gamma	Gamma

Bayesian Point Estimation

Suppose that $T(\mathbf{X})$ is a point estimator of a single, unknown parameter θ.

Recall, from Lecture 2, we could specify a decision function $\delta(\mathbf{X}) = T(\mathbf{X})$ which is the estimator $T(\mathbf{X})$, given the sample $\mathbf{X}$.

Suppose that the loss incurred in using $T(\mathbf{X})$ to estimate θ is given by the loss function $L(\theta, T(\mathbf{X}))$.

The **Bayes estimate** of θ, denoted θ^*, is the value $t = t(\mathbf{x})$ that minimises the posterior expected loss

$$\mathbb{E}_{\theta|\mathbf{x}}[L(\theta, t)] = \int_{\Theta} L(\theta, t)\pi(\theta \mid \mathbf{x})\mathrm{d}\theta.$$

Bayesian Point Estimation: Quadratic Error Loss

Under quadratic error loss, the loss function is $L(\theta, t) = (\theta - t)^2$ and the Bayes estimate is the posterior mean.

Bayesian Inference: Example

A market research company wishes to know the percentage, p, of the general public who have been aware of a recent advertising campaign. From past records, the company expects this percentage to be about 35% and is confident that the actual population percentage lies between 25% and 45%.

In order to obtain a better estimate of this percentage, a simple random sample of 300 people is taken and, of these 300 people, 123 say that they are aware of the campaign. Use a Bayesian approach to update the company's beliefs about p.

Bayesian Inference: Example

Bayesian Inference: Example

Bayesian Inference: Example

Normal Sampling

Often, we are concerned with normal sampling, where our data $X_1, \ldots, X_n$ are normally distributed ($X_i \sim \mathcal{N}(\mu, \sigma^2)$).

In this case, the prior distribution of μ is usually taken to be normal (the **conjugate prior** for normal sampling).

When using Bayesian methods and a normal distribution, it is often convenient to define

$$\kappa = \frac{1}{\sigma^2}$$

as the precision of the distribution.

This can make some of the algebraic manipulation easier and the precision itself has a meaning (small variance $\implies$ greater precision/accuracy)

Normal Sampling: Example 1

Suppose that the observe a sample $X_1, \ldots, X_n$ of iid $\mathcal{N}(\mu, \sigma^2)$ random variables, where σ is known. The prior distribution of μ is $\mu \sim \mathcal{N}(\theta, \tau^2)$. Derive the posterior distribution of μ.

Normal Sampling: Example 1

Normal Sampling: Example 1

Normal Sampling: Example 1

Normal Sampling: Example 2

Suppose that $X_1, \ldots, X_9$ are independent $\mathcal{N}(\mu, 4)$ and the prior distribution of μ is $\mu \sim \mathcal{N}(25, 10)$. The observed sample mean is $\bar{x} = 20$. Determine the posterior distribution of μ and interpret your answer.

Normal Sampling: Example 2

Learning Outcomes

- Understand the advantages of maximum likelihood estimation.
- Know and be able to use the large sample (asymptotic) properties of MLEs (in one and more dimensions).
- Understand the Bayesian approach to parametric inference.
- Know about and use the various types of prior distributions:
  - Personal priors
  - Uniform and improper priors
  - Jeffreys' prior
  - Conjugate priors
- Understand and apply the Bayesian method of point estimation.