CHAPTER 1 - SOME REVISION AND USEFUL DEFINITIONS

§1.1 Some standard distributions revisited

BERNOULLI A basic building block of discrete random variables, the state space is $\{0, 1\}$ and the parameter space is $(0, 1)$:

$$\begin{aligned} X &\sim Bern(p) \\ f_p(x) &= p^x(1-p)^{1-x} \end{aligned}$$

where the parameter p represents the probability of the event $(X = 1)$.

BINOMIAL What happens if we add n iid $Bern(p)$ random variables together (note the common parameter p)?

$$\begin{aligned} X_1, \ldots, X_n &\quad \text{iid} \quad Bern(p) \\ Y = \sum_{i=1}^{n} X_i &\sim Bin(n, p) \\ f_p(y) &= \left(\begin{array}{c} n \\ y \end{array} \right) p^y(1-p)^{n-y} \end{aligned}$$

where the state space is $\{0, 1, \ldots, n\}$ and the parameter space is $p \in (0, 1)$.

GEOMETRIC The Geometric is a discrete time waiting distribution. We run a sequence of iid $Bern(p)$ random variables and wait for the first occurrence of a 1:

$$\begin{aligned} X &\sim Geo(p) \\ f_p(x) &= p(1-p)^{x-1} \end{aligned}$$

where the state space is $\{1, 2, \ldots\}$ and the parameter space is $p \in (0, 1)$.

There is an alternative parameterisation whereby the variable is defined as the number of 0's before the first 1. In that case:

$$\begin{aligned} Z &\sim Geo(p) \\ f_p(z) &= p(1-p)^{z} \end{aligned}$$

where the state space is $\{0, 1, \ldots\}$ and the parameter space is $p \in (0, 1)$.

NEGATIVE BINOMIAL A generalisation of the Geometric distribution, it is the waiting time for the r^{th} 1 in a sequence of iid $Bern(p)$ random variables:

$$\begin{aligned} X &\sim NegBin(r, p) \\ f_p(x) &= \left(\begin{array}{c} x-1 \\ r-1 \end{array} \right) p^r(1-p)^{x-r} \end{aligned}$$

where the state space is $\{r, r+1, \ldots\}$ and the parameter space is $p \in (0, 1)$.

Again, there is an alternative parameterisation whereby the variable is defined as the number of 0's before the r^{th} 1. In that case:

$$\begin{aligned} Z &\sim NegBin(r,p) \\ f_p(z) &= \left(\begin{array}{c} z+r-1 \\ r-1 \end{array} \right) p^r (1-p)^z \end{aligned}$$

where the state space is $\{0, 1, \ldots\}$ and the parameter space is $p \in (0, 1)$.

POISSON The Poisson can be thought of as a limiting case of the Binomial as $n \to \infty$ and $p \to 0$ in such a way that the expectation of the Binomial, np, tends to some finite, non-zero constant λ:

$$\begin{aligned} X &\sim Pois(\lambda) \\ f_\lambda(x) &= \frac{\lambda^x e^{-\lambda}}{x!} \end{aligned}$$

where the state space is $\{0, 1, \ldots\}$ and the parameter space is $\lambda > 0$.

UNIFORM The Uniform is perhaps the simplest continuous distribution:

$$\begin{aligned} X &\sim U(\theta_1, \theta_2) \\ f_{\theta_1,\theta_2}(x) &= \frac{1}{\theta_2 - \theta_1} \end{aligned}$$

where the state space is (θ_1, θ_2) and the parameter space is $-\infty < \theta_1 < \theta_2 < \infty$.

EXPONENTIAL The Exponential is the continuous time equivalent of the Geometric distribution:

$$\begin{aligned} X &\sim \exp(\lambda) \\ f_\lambda(x) &= \lambda e^{-\lambda x} \end{aligned}$$

where the state space is $X > 0$ and the parameter space is $\lambda > 0$. In this parameterisation, the parameter λ is referred to as the rate parameter, and the expectation of X is λ^{-1}. The alternative parameterisation in terms of the mean, μ, is also frequently used:

$$\begin{aligned} X &\sim \exp(\mu) \\ f_\mu(x) &= \frac{1}{\mu} e^{-x/\mu} \end{aligned}$$

where the state space is still $X > 0$ and the parameter space is $\mu > 0$. Clearly $\mu = 1/\lambda$.

GAMMA A generalisation of the Exponential distribution:

$$\begin{aligned} X &\sim Gam(\lambda, t) \\ f_{\lambda,t}(x) &= \frac{\lambda^t}{\Gamma(t)} x^{t-1} e^{-\lambda x} \\ \text{where } \Gamma(t) &= \int_0^\infty y^{t-1} e^{-y} dy \end{aligned}$$

where the state space is $X > 0$ and the parameter space is $\lambda > 0, t > 0$.

One way in which Gamma distributions arise is as the sum of iid $\exp(\lambda)$ random variables:

$$\begin{aligned} X_1, \ldots, X_t \quad & \text{iid} \quad \exp(\lambda) \\ Y = \sum_{i=1}^{t} X_i \quad & \sim \quad Gam(\lambda, t) \end{aligned}$$

NORMAL The Normal arises in a number of ways, for example as the limiting case of a Binomial as $n \to \infty$, or as a result of the Central Limit Theorem:

$$\begin{aligned} X &\sim N(\mu, \sigma^2) \\ f_{\mu,\sigma^2}(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \end{aligned}$$

where the state space is the real line and the parameter space is $\mu \in (-\infty, \infty), \sigma^2 > 0$.

BETA The Beta distribution will be a useful one for those going on to study MA40189:

$$\begin{aligned} X &\sim Beta(a, b) \\ f_{a,b}(x) &= \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \\ \text{where } B(a,b) &= \int_0^1 y^{a-1}(1-y)^{b-1} dy \end{aligned}$$

where the state space is $(0, 1)$ and the parameter space is $a > 0, b > 0$.

§1.2 Exponential families

It will be useful, and elegant, if we can work with a definition which includes many of the distributions in the previous section and which has good theoretical properties.

Definition 1.1

A *k-parameter exponential family* is one whose pmf/pdf can be written in the form

$$f_\theta(x) = \begin{cases} c(\theta)h(x)\exp(\sum_{j=1}^{k} a_j(\theta)b_j(x)) & x \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

and where the state space Ω does not involve the parameter θ.

Examples of testing whether a distribution is a member of an exponential family

1. Exponential

$$f_\lambda(x) = \lambda e^{-\lambda x}$$

There is one parameter, so $k = 1$. We could parameterise $f_\lambda(x)$ using $c(\lambda) = \lambda$, $h(x) = 1$, $a(\lambda) = -\lambda$ and $b(x) = x$. The state space is $X > 0$, which is independent of the parameter λ. Therefore the Exponential distribution is a member of the one parameter exponential family.

2. Normal

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

There are two parameters, so $k = 2$. We could parameterise $f_{\mu,\sigma^2}(x)$ using $c(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\mu^2}{2\sigma^2}}$, $h(x) = 1$, $a_1(\mu, \sigma^2) = -1/2\sigma^2$ with $b_1(x) = x^2$ and $a_2(\mu, \sigma^2) = \mu/\sigma^2$ with $b_2(x) = x$. The state space is the real line, which is independent of the parameters μ and σ^2. Therefore the Normal distribution is a member of the two parameter exponential family.

3. Uniform The Uniform distribution with parameters θ_1 and θ_2 has state space (θ_1, θ_2). Therefore the Uniform distribution is not a member of an exponential family as its state space depends on the parameters.

§1.3 Sufficiency

Another useful concept will be that of *sufficiency*.

Definition 1.2

A *sufficient statistic* is a statistic which exhausts the information in a data set regarding the parameter θ in the sense that the conditional distribution of $X_1, \ldots, X_n$ given the value of the sufficient statistic $T(\mathbf{x}) = t$ does not involve the parameter θ.

An example of sufficiency

Suppose $X_1, \ldots, X_n$ are iid $Bern(p)$ random variables and that we want to see whether $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ is a sufficient statistic. We will need to find the conditional distribution of $X_1, \ldots, X_n | T = t$ and note whether or not this distribution involves the parameter p.

- What is the distribution of T? As T is the sum of iid $Bern(p)$ random variables, $T \sim Bin(n, p)$

$$\mathsf{P}(T = t) = \left(\begin{array}{c} n \\ t \end{array} \right) p^t (1-p)^{n-t}, \;\; t = 0, 1, \ldots, n$$

- What is the joint distribution of $X_1, \ldots, X_n$? As they are independent, the joint distribution is the product of the marginals

$$\mathsf{P}(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}, \; x_i \in \{0, 1\}$$

- What is the conditional distribution of $\mathbf{X}$ conditional on T taking the value t?

$$\begin{aligned} \mathsf{P}(\mathbf{X} = \mathbf{x} | T = t) &= \frac{\mathsf{P}(\mathbf{X} = \mathbf{x} \cap T = t)}{\mathsf{P}(T = t)} \\ &= \frac{\mathsf{P}(\mathbf{X} = \mathbf{x})}{\mathsf{P}(T = t)} I_{[\sum_{i=1}^n x_i = t]} \\ &= \frac{1}{\left(\begin{array}{c} n \\ t \end{array} \right)} I_{[\sum_{i=1}^n x_i = t]} \end{aligned}$$

As this conditional distribution does not involve the parameter p, we can say that $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for the $Bern(p)$ distribution.

Rather than guessing at a possible sufficient statistic for any particular distribution, and then finding the conditional distribution to check whether the guess is correct, the following theorem enormously simplifies the process in the one parameter case:

Theorem 1.1

The *Factorisation theorem* states that $T(\mathbf{X})$ is sufficient for θ if and only if there exist functions $g(T(\mathbf{x}), \theta)$ and $h(\mathbf{x})$ such that the joint distribution can be factorised

$$f_\theta(\mathbf{x}) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$$

Proof of the Factorisation Theorem in the discrete case:

Suppose $f_\theta(\mathbf{x}) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$

$$\begin{aligned}
\mathsf{P}_\theta(T(\mathbf{X}) = t) &= \sum_{\mathbf{x}:T(\mathbf{x})=t} \mathsf{P}_\theta(\mathbf{X} = \mathbf{x}) \\
&= \sum_{\mathbf{x}:T(\mathbf{x})=t} f_\theta(\mathbf{x}) \\
&= \sum_{\mathbf{x}:T(\mathbf{x})=t} g(T(\mathbf{x}), \theta)h(\mathbf{x}) \\
&= g(t, \theta) \sum_{\mathbf{x}:T(\mathbf{x})=t} h(\mathbf{x})
\end{aligned}$$

Now consider a particular $\mathbf{y}$ such that $T(\mathbf{y}) = t$:

$$\begin{aligned}
\mathsf{P}_\theta(\mathbf{X} = \mathsf{y}|T(\mathbf{X}) = t) &= \frac{\mathsf{P}_\theta(\mathbf{X} = \mathbf{y} \cap T(\mathbf{X}) = t)}{\mathsf{P}_\theta(T(\mathbf{X}) = t)} \\
&= \frac{\mathsf{P}_\theta(\mathbf{X} = \mathbf{y})}{\mathsf{P}_\theta(T(\mathbf{X}) = t)} \text{ since } T(\mathbf{y}) = t \\
&= \frac{g(T(\mathbf{y}), \theta)h(\mathbf{y})}{g(t, \theta) \sum_{\mathbf{x}:T(\mathbf{x})=t} h(\mathbf{x})} \\
&= \frac{h(\mathbf{y})}{\sum_{\mathbf{x}:T(\mathbf{x})=t} h(\mathbf{x})}
\end{aligned}$$

Since this conditional probability does not involve θ, $T(\mathbf{X})$ is sufficient for θ.

Suppose $T(\mathbf{X})$ **is sufficient**

$$\begin{aligned}
f_\theta(\mathbf{x}) &= \mathsf{P}_\theta(\mathbf{X} = \mathbf{x}) \\
&= \sum_{t^*} \mathsf{P}_\theta(\mathbf{X} = \mathbf{x} \cap T(\mathbf{X}) = t^*) \text{ where } t^* \text{ forms a partition} \\
&= \mathsf{P}_\theta(\mathbf{X} = \mathbf{x} \cap T(\mathbf{X}) = t(\mathbf{x})) \text{ since all other joint probabilities are zero} \\
&= \mathsf{P}_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t(\mathbf{x}))\mathsf{P}_\theta(T(\mathbf{X}) = t(\mathbf{x}))
\end{aligned}$$

but this first term is independent of θ by assumption, so denote it by $h(\mathbf{x})$, while the second is a function of $T(\mathbf{x})$ and θ, so denote it by $g(T(\mathbf{x}), \theta)$, i.e.

$$f_\theta(\mathbf{x}) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$$

Examples of finding sufficient statistics using the Factorisation theorem

1. Bernoulli $X_1, \ldots, X_n$iid $Bern(p)$ random variables:

$$\begin{aligned} f_p(\mathbf{x}) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \end{aligned}$$

Take $h(\mathbf{x}) = 1$, $T(\mathbf{x}) = \sum_{i=1}^n x_i$ and $g(T(\mathbf{x}), p) = p^{T(\mathbf{x})}(1-p)^{n-T(\mathbf{x})}$, then by the Factorisation theorem, $\sum_{i=1}^n X_i$ is a sufficient statistic for p.

2. The one parameter exponential family $X_1, \ldots, X_n$iid from the one parameter exponential family

$$\begin{aligned} f_\theta(x) &= c(\theta)h(x)\exp(a(\theta)b(x)),\ x \in \Omega \\ f_\theta(\mathbf{x}) &= c(\theta)^n \prod_{i=1}^{n} h(x_i)\exp(a(\theta)\sum_{i=1}^{n} b(x_i)) \end{aligned}$$

Take $h(\mathbf{x}) = \prod_{i=1}^n h(x_i)$, $T(\mathbf{x}) = \sum_{i=1}^n b(x_i)$ and $g(T(\mathbf{x}), \theta) = c(\theta)^n \exp(a(\theta)T(\mathbf{x}))$, then by the Factorisation theorem, $\sum_{i=1}^n b(X_i)$ is a sufficient statistic for θ.

The concept of sufficiency extends to *joint sufficiency* with a corresponding generalised factorisation theorem (which we will not prove).

Definition 1.3

Suppose $X_1, \ldots, X_n$ have a joint distribution which depends on parameters $\{\theta_1, \ldots, \theta_k\}$. The statistics $\{T_1(\mathbf{X}), \ldots, T_r(\mathbf{X})\}$ are jointly sufficient for $\{\theta_1, \ldots, \theta_k\}$ if the conditional distribution of $X_1, \ldots, X_n$ given the values of $\{T_1(\mathbf{X}), \ldots, T_r(\mathbf{X})\}$ does not depend on any of the $\{\theta_1, \ldots, \theta_k\}$.

Theorem 1.2

The *Generalised Factorisation theorem* states that $\{T_1(\mathbf{X}), \ldots, T_r(\mathbf{X})\}$ are jointly sufficient for $\{\theta_1, \ldots, \theta_k\}$ if and only if there exist functions $g(T_1(\mathbf{X}), \ldots, T_r(\mathbf{X}), \theta_1, \ldots, \theta_k)$ and $h(\mathbf{x})$ such that

$$f_{\theta_1,\ldots,\theta_k}(\mathbf{x}) = g(T_1(\mathbf{X}), \ldots, T_r(\mathbf{X}), \theta_1, \ldots, \theta_k)h(\mathbf{x})$$

Examples of finding joint sufficient statistics using the Generalised Factorisation theorem

1. Normal $X_1, \ldots, X_n$iid $N(\mu, \sigma^2)$ random variables:

$$\begin{aligned} f_{\mu,\sigma^2}(\mathbf{x}) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2) \\ &= (2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2}\left[\frac{\sum_{i=1}^n x_i^2}{\sigma^2} - \frac{2\mu\sum_{i=1}^n x_i}{\sigma^2} + \frac{n\mu^2}{\sigma^2}\right]) \end{aligned}$$

Take $h(\mathbf{x}) = 1$, $T_1(\mathbf{x}) = \sum_{i=1}^n x_i^2$, $T_2(\mathbf{x}) = \sum_{i=1}^n x_i$ and
$g(T_1(\mathbf{x}), T_2(\mathbf{x}), \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2}\left[\frac{T_1(\mathbf{x})}{\sigma^2} - \frac{2\mu T_2(\mathbf{x})}{\sigma^2} + \frac{n\mu^2}{\sigma^2}\right])$, then by the Generalised Factorisation theorem, $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n X_i^2$ are jointly sufficient statistics for μ, σ^2.

2. Uniform $X_1, \ldots, X_n$iid $U(\theta, 2\theta)$ random variables:

$$\begin{aligned} f_\theta(\mathbf{x}) &= \left(\frac{1}{\theta}\right)^n I_{[\theta<x_i<2\theta,\ i=1,\ldots,n]} \\ &= \left(\frac{1}{\theta}\right)^n I_{[\theta<\min x_i<\max x_i<2\theta]} \end{aligned}$$

Take $h(\mathbf{x}) = 1$, $T_1(\mathbf{x}) = \min x_i$, $T_2(\mathbf{x}) = \max x_i$ and $g(T_1(\mathbf{x}), T_2(\mathbf{x}), \theta) = \left(\frac{1}{\theta}\right)^n I_{[\theta<T_1(\mathbf{x})<T_2(\mathbf{x})<2\theta]}$, then by the Generalised Factorisation theorem, $\min X_i$ and $\max X_i$ are jointly sufficient statistics for θ.

§1.4 Revision of properties of maximum likelihood estimation

Recall that the likelihood function is the joint mass/density function evaluated at the observed $x_1, \ldots, x_n$ and regarded as a function of the unknown parameters:

$$L(\theta) = f_\theta(x_1, \ldots, x_n)$$

The maximum likelihood estimate (MLE) is the value of θ (not necessarily unique) in the parameter space Θ which makes $L(\theta)$ as large as possible.

1. Working with the log likelihood In practice, we often work with the loglikelihood, $\ell(\theta) = \ln L(\theta)$.

Example

$X_1, \ldots, X_n$iid $Bern(p)$ random variables, $\Theta = (0, 1)$:

$$\begin{aligned} f_p(\mathbf{x}) &= \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \\ L(p) &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\ \ell(p) &= (\sum_{i=1}^n x_i)\ln p + (n - \sum_{i=1}^n x_i)\ln(1-p) \\ \frac{d\ell(p)}{dp} &= \frac{\sum_{i=1}^n x_i}{p} - \frac{n-\sum_{i=1}^n x_i}{1-p} \\ \frac{d^2\ell(p)}{dp^2} &= -\frac{\sum_{i=1}^n x_i}{p^2} - \frac{n-\sum_{i=1}^n x_i}{(1-p)^2} \end{aligned}$$

Solving the first derivative equal to zero over the parameter space $(0, 1)$, gives as a turning point $\hat{p} = \sum_{i=1}^n x_i/n$, at which point the second derivative is negative, confirming that this $\hat{p}$ is a maximum. Thus the Maximum Likelihood Estimator is $\hat{p} = \sum_{i=1}^n X_i/n$.

2. Times when calculus does not help If the MLE occurs on the boundary of the parameter space, then calculus may not help to find the maximiser of the likelihood.

Example

$X_1, \ldots, X_n$iid $U(0, \theta)$ random variables

$$\begin{aligned} f_p(\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\theta} I_{[0<x_i<\theta]} \\ L(\theta) &= \theta^{-n}, \theta > x_1, \ldots, x_n \\ \frac{dL(\theta)}{ds\theta} &= -n\theta^{-(n+1)}, \theta > x_1, \ldots, x_n \end{aligned}$$

The first derivative is negative, i.e. the likelihood is decreasing and so to maximise L we should pick the smallest possible value of θ. In this case, the smallest possible value that θ can take is $\max x_i$, that is the Maximum Likelihood Estimator is $\max X_i$.

3. MLEs and sufficient statistics If there is a unique MLE $\hat{\theta}$, then it is a function of the sufficient statistics. To see why this is the case, consider writing the distribution $f_\theta(\mathbf{x})$ in the form of the Generalised Factorisation theorem:

$$L(\theta) = f_\theta(\mathbf{x}) = g(T_1(\mathbf{X}), \ldots, T_r(\mathbf{X}), \theta)h(\mathbf{x})$$

Maximising $L(\theta)$ over θ for fixed $\mathbf{x}$ is equivalent to maximising $g(T_1(\mathbf{X}), \ldots, T_r(\mathbf{X}), \theta)$ over θ, and hence the maximiser, i.e. $\hat{\theta}$, will be a function of the sufficient statistics $T_1(\mathbf{X}), \ldots, T_r(\mathbf{X})$.

Examples

Consider the previous two examples, the Bernoulli and the Uniform.

4. Multiparameter problems In multiparameter problems, we have an optimisation problem in more than one dimension. In some cases, this will be analytically tractable but in other cases numerical methods might be required. Either way, conditions for any turning point to be a maximum should be checked.

Example

$X_1, \ldots, X_n$iid $Gam(\lambda, t)$ random variables

$$\begin{aligned} f_{\lambda,t}(\mathbf{x}) &= \prod_{i=1}^{n} \frac{\lambda^t}{\Gamma(t)} x_i^{t-1} \exp(-\lambda x_i) \\ L(\lambda, t) &= \frac{\lambda^{nt}}{\Gamma(t)^n} \left(\prod_{i=1}^{n} x_i \right)^{t-1} \exp(-\lambda \sum_{i=1}^{n} x_i) \\ \ell(\lambda, t) &= nt \ln \lambda - n \ln \Gamma(t) + (t-1) \ln \prod_{i=1}^{n} x_i - \lambda \sum_{i=1}^{n} x_i \\ \frac{\partial \ell(\lambda, t)}{\partial \lambda} &= \frac{nt}{\lambda} - \sum_{i=1}^{n} x_i \\ \frac{\partial \ell(\lambda, t)}{\partial t} &= n \ln \lambda - \frac{n \partial ln \Gamma(t)}{\partial t} + \ln \prod_{i=1}^{n} x_i \end{aligned}$$

We need to solve simultaneously $\frac{\partial \ell(\lambda,t)}{\partial \lambda} = 0$ and $\frac{\partial \ell(\lambda,t)}{\partial t} = 0$. We can reduce this to a one-dimensional problem by noting from the first equation that we need $\lambda = \frac{nt}{\sum_{i=1}^{n} x_i}$ and substituting this into the second equation, but we are still left with a problem which must be solved numerically.

Once we have a turning point of $\ell(\lambda, t)$, the conditions on the second derivatives to be checked to ensure we have a maximum are:

$$\begin{aligned} \frac{\partial^2 \ell(\lambda,t)}{\partial \lambda^2} &< 0 \\ \frac{\partial^2 \ell(\lambda,t)}{\partial t^2} &< 0 \\ \frac{\partial^2 \ell(\lambda,t)}{\partial t^2}\frac{\partial^2 \ell(\lambda,t)}{\partial \lambda^2} &> \left(\frac{\partial^2 \ell(\lambda,t)}{\partial t \partial \lambda}\right)^2 \end{aligned}$$

5. Maximum likelihood estimates are functionally invariant If $\hat{\theta}$ is the MLE of θ, then when g is any function of θ, the MLE of $g(\theta)$ is $g(\hat{\theta})$, i.e. we can simply plug the MLE into the function.

To see why this works, let $\phi = g(\theta)$. Denote the likelihood function for θ by $L(\theta)$ and the likelihood function for ϕ by $\tilde{L}(\phi)$. Consider the two cases separately, g invertible and not invertible.

g is invertible In this case the likelihood $\tilde{L}(\phi)$ is easy to define since we know exactly which value of θ corresponds to any ϕ:

$$\tilde{L}(\phi) = L(\theta) \text{ where } \theta = g^{-1}(\phi)$$

As a result, since $L(\hat{\theta}) \geq L(\theta)$ by the definition of maximum likelihood

$$\begin{aligned} \tilde{L}(g(\hat{\theta})) = L(\hat{\theta}) &\geq L(\theta) = \tilde{L}(g(\theta)) \\ \text{that is } \tilde{L}(g(\hat{\theta})) &\geq \tilde{L}(\phi)\ \forall \phi \end{aligned}$$

so $g(\hat{\theta})$ maximises the likelihood $\tilde{L}$ and so is the maximum likelihood estimate of ϕ.

g is not invertible Note: this part of the proof is not examinable, it is purely for those who are interested. In this case there is not a unique θ corresponding to each ϕ and to define $\tilde{L}$ we need to make a choice as to which θ to use for each ϕ. Define

$$\tilde{L}(\phi) = \max_{\theta : g(\theta) = \phi} L(\theta)$$

then again we can see that the largest value of $\tilde{L}$ occurs at $g(\hat{\theta})$ since this $\hat{\theta}$ maximises L.

Example of function invariance

Suppose $X_1, \ldots, X_n$ are iid $\exp$ random variables, parameterised either by the rate parameter λ or by the mean parameter μ. What are the MLEs of the two parameters?

The rate parameter λ

$$f_\lambda(\mathbf{x}) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i)$$

$$
\begin{aligned}
L(\lambda) &= \lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i) \\
\ell(\lambda) &= n \ln \lambda - \lambda \sum_{i=1}^{n} x_i \\
\frac{d\ell(\lambda)}{d\lambda} &= \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \\
\frac{d^2\ell(\lambda)}{d\lambda^2} &= -\frac{n}{\lambda^2}
\end{aligned}
$$

Solving $\frac{d\ell(\lambda)}{d\lambda} = 0$ and checking that at this point $\frac{d^2\ell(\lambda)}{d\lambda^2} < 0$, tells us that $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$.

The mean parameter μ

$$
\begin{aligned}
f_\mu(\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\mu} \exp(-x_i/\mu) \\
L(\mu) &= \mu^{-n} \exp(-\sum_{i=1}^{n} x_i/\mu) \\
\ell(\mu) &= -n \ln \mu - \frac{1}{\mu} \sum_{i=1}^{n} x_i \\
\frac{d\ell(\mu)}{d\mu} &= -\frac{n}{\mu} + \frac{\sum_{i=1}^{n} x_i}{\mu^2} \\
\frac{d^2\ell(\mu)}{d\mu^2} &= \frac{n}{\mu^2} - 2\frac{\sum_{i=1}^{n} x_i}{\mu^3}
\end{aligned}
$$

Solving $\frac{d\ell(\mu)}{d\mu} = 0$ and checking that at this point $\frac{d^2\ell(\mu)}{d\mu^2} < 0$, tells us that $\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$.

The point of this example is to notice that $\mu = \lambda^{-1}$ and $\hat{\mu} = \hat{\lambda}^{-1}$ (so, once we had found $\hat{\lambda}$, we could write down $\hat{\mu}$ without resorting to calculus).