Chapter 6 - Likelihood Inference

Section 6.1. The likelihood function.

Data s and the model $\{P_\theta : \theta \in \Omega\}$.

Start with P_θ that is discrete, given by f_θ; $f_\theta(s) = P_\theta(S = s)$.

Consider $L(\cdot|s) : \Omega \to \mathbb{R}^1$ s.t. $L(\theta|s) = f_\theta(s)$ - the likelihood function. The value $L(\theta|s)$ is the likelihood of θ. Data are fixed, but θ varies. Here, $f_\theta(s)$ is the probability of obtaining s when the true value of the parameter is θ. Belief ordering on Ω: we prefer θ_1 over θ_2 if $f_{\theta_1}(s) > f_{\theta_2}(s)$. Indifferent between θ_1 and θ_2 if $f_{\theta_1}(s) = f_{\theta_2}(s)$.

Interpretation: $L(\theta|s)$ is the probability of s if θ is the true value. It may be very small but it is the ordering that is important, not the absolute value of $L(\theta|s)$.

Ex 6.1.1. Let $S = \{1, 2, \ldots\}$ & $\{P_\theta : \theta \in \{1, 2\}\}$.

P_1 is $\sim$ Unif $\{1, \ldots, 10^3\}$ & P_2 is Unif $\{1, \ldots, 10^6\}$. Observe $s = 10$.

Then $L(1|10) = \frac{1}{10^3}$ and $L(2|10) = \frac{1}{10^6}$. $\theta = 1$ is a 1000 times likelier under Model 1 than under Model 2.

Thus, we're interested only in likelihood ratios: $\frac{L(\theta_1|s)}{L(\theta_2|s)}$ for $\theta_1, \theta_2 \in \Omega$. Thus, for any $c > 0$, the function $L^*(\cdot|s) = cL(\cdot|s)$ can also serve as the likelihood function. $L(\cdot|s)$ and $L^*(\cdot|s)$ are two equivalent likelihoods.

Ex. 6.1.2. A coin is tossed $n = 10$ times, $s = 4$ heads are observed. The appropriate model for the # of heads is $Bin(10, \theta)$, $\theta \in \Omega = [0, 1]$.

$\Rightarrow L(\theta|4) = \binom{10}{4} \theta^4 (1-\theta)^6$. $\theta = 0.4$ is the maximizer with the value 0.2508.

There are different approaches to inference via the likelihood function. At one extreme is the likelihood principle: if two model and data combinations yield equivalent likelihood functions, then inferences about the unknown parameter must be the same. This is a very restrictive approach.

Ex. 6.1.3. Toss a coin (independently) until 4 heads are obtained; the number of tails observed is $s = 6$.
$S \sim \text{NegBin}(4, \theta)$ and $L(\theta/6) = \binom{9}{6}\theta^4(1-\theta)^6$. This is a positive multiple of $L(\theta/4)$ considered in the Ex. (6.1.2). Note that the data were obtained in the entirely different ways!! ... additional model features besides the likelihood function must be taken into account.

An example of the likelihood inference: let $C(s) = \{\theta : L(\theta/s) \geq c\}$ for some $c \geq 0$ – a likelihood region. This may possibly contain the true value θ; if some $\theta^* \notin C(s)$, then $L(\theta^*/s) < L(\theta/s)$ for $\forall \theta \in C(s)$ and so is not well supported by the data compared to a $\theta \in C(s)$. For now, we don't know how to choose θ – leave it for later.

If P_θ is continuous (imagine $f_{\theta_1}(s) > f_{\theta_2}(s)$ and $s \in \mathbb{R}^1$. Then, if $f_\theta(s)$ is continuous at every s, we have

$$P_{\theta_1}(V) = \int f_{\theta_1}(s)\,ds > P_{\theta_2}(V) = \int f_{\theta_2}(s)\,ds \quad \text{for every } V = (a, b)$$

that is small enough and contains 1. ... the data supports θ_1 more than θ_2. If, $s \in \mathbb{R}^n$, $n > 1$ – similar interpretation. Thus, $L(\theta/s) = f_\theta(s)$ again.

Ex. 6.1.4. Location normal model.

$(X_1, \ldots, X_n) \sim N(\theta, \sigma_0^2)$ where $\sigma_0^2 > 0$ is known and $\theta \in \Omega = \mathbb{R}^1$.

Then, $L(\theta/x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \left(2\pi\sigma_0^2\right)^{-n/2} e^{-\frac{n}{2\sigma_0^2}(\bar{x}-\theta)^2} e^{-\frac{n-1}{2\sigma_0^2}s^2}$

The simpler version is $L(\theta/x_1, \ldots, x_n) = e^{-\frac{n}{2\sigma_0^2}(\bar{x}-\theta)^2}$

If $n=25$, $\sigma_0^2=1$, observe $\bar{x}=3.3$ ... see Fig. (6.12)
The peak is at $\theta=\bar{x}=3.3$. The likelihood interval

$$C(x)=\{\theta: L(\theta/x_1,\dots,x_n)\geq 0.5\}=\{3.0645, 3.535784\}$$

contains all those θ for which likelihood is at least 0.5.

Ex. 6.1.5. Multinomial models.

Suppose that $k=3$, $(\theta_1,\theta_2,\theta_3)$ are unknown.

$$\Omega=\{(\theta_1,\theta_2,\theta_3): \theta_i\geq 0, i=1,2,3, \sum\theta_i=1\}.$$

Ω is effectively two-dimensional!

We observe $(s_1,\dots,s_n) \Rightarrow L(\theta_1,\theta_2,\theta_3/s_1,\dots,s_n)=$
$=\theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3}$ where x_i's the number of i's in the sample. We can see that the likelihood based on the observed counts (x_1,x_2,x_3) is given by

$$L(\theta_1,\theta_2,\theta_3/x_1,x_2,x_3)=\theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3}$$

Section 6.1.1 Sufficient statistics.

Def A function T defined on the sample space S is a sufficient statistic for the model if, whenever $T(s_1) = T(s_2)$, then

$$L(\cdot|s_1) = c(s_1, s_2) L(\cdot|s_2) \text{ for some } c(s_1, s_2) > 0.$$

Thus, if we observe $T = t$, we can pick any value set $T^{-1}(t) = \{s : T(s) = t\}$ and base the likelihood on it. This is, effectively, a dimension reduction technic.

Ex. Take $S = \{1, 2, 3, 4\}$, $\Omega = \{a, b\}$

	$s=1$	$s=2$	$s=3$	$s=4$
$\theta = a$	$1/2$	$1/6$	$1/6$	$1/6$
$\theta = b$	$1/4$	$1/4$	$1/4$	$1/4$

$L(\cdot|2) = L(\cdot|3) = L(\cdot|4)$, so $\{2, 3, 4\}$ give the same likelihood ratios. Thus, $T: S \to \{0, 1\}$ s.t. $T(1) = 0$, $T(2) = T(3) = T(4) = 1$ is a sufficient statistic. Now, the sample space for T has only two elements.

Factorization Theorem If $f_\theta(s) = h(s) g_\theta(T(s))$, $g_\theta, h \geq 0$, then T is a sufficient statistic

Proof If $T(s_1) = T(s_2)$,

$$L(\cdot|s_1) = h(s_1) g_\theta(T(s_1)) = \frac{h(s_1) g_\theta(T(s_1))}{h(s_2) g_\theta(T(s_2))} h(s_2) g_\theta(T(s_2)) =$$

$$= \frac{h(s_1)}{h(s_2)} h(s_2) g_\theta(T(s_2)) = c(s_1, s_2) L(\cdot|s_2) \text{ because } g_\theta(T(s_1)) = g_\theta(T(s_2)).$$

A sufficient statistic hinges on the choice of parameter - if the parameter is different, $T(\cdot)$ will also change!!

Def. 6.1.2 A sufficient statistic T for a model is a minimal sufficient statistic, whenever the value of $T(s)$ can be calculated once we know the likelihood f-n $\mathcal{L}(\cdot|s)$.

A minimal sufficient statistic gives the greatest reduction of the data in the sense that, if T is a min. suf. and U is suf., then there $\exists$ f-n h such that $T = h(U)$. Sufficiency always depends on the model!

Ex. Location Normal Model.

Factorization Theorem $\rightarrow \bar{x}$ is a sufficient statistic.

Any $\mathcal{L}(\cdot|s)$ is a positive multiple of $\exp\left(-\frac{n}{2\sigma_0^2}(\bar{x}-\theta)^2\right)$

Any such function of θ is completely specified by the maximum point, that is, $\theta = \bar{x} \Rightarrow$ it is a complete sufficient statistic (because $\sigma_0^2 > 0$ is known!).

Ex. Location - Scale Normal model.

$X_1, \dots, X_n \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}^1$, $\sigma^2 > 0$ are unknown.

$\theta = (\mu, \sigma^2) \in \Omega = \mathbb{R}^1 \times (0, \infty)$

$$\Rightarrow \mathcal{L}(\theta | x_1, \dots, x_n) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} =$$

$$= (2\pi\sigma^2)^{-n/2} e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2} e^{-\frac{n-1}{2\sigma^2}s^2} \Rightarrow (\bar{x}, s^2) \text{ is a suf. stat.}$$

If σ^2 is fixed, any positive multiple of $\mathcal{L}(\cdot|x_1, \dots, x_n)$ is maximized at $\mu = \bar{x}$. Fix μ at $\bar{x}$, we have

$$\mathcal{L}((\bar{x}, \sigma^2) | x_1, \dots, x_n) = (2\pi\sigma^2)^{-n/2} e^{-\frac{n-1}{2\sigma^2}s^2} \rightarrow \max$$

doesn't change if taking a log; $\Rightarrow$

$$\frac{\partial \ln \mathcal{L}((\bar{x}, \sigma^2)|x)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{n-1}{2\sigma^4}s^2 \Big| = 0$$

gives $\hat{\sigma}^2 = \frac{n-1}{n}s^2$ - it is a 1-1 function of s^2.

$\Rightarrow (\bar{x}, s^2)$ is a minimal sufficient statistic for this model.

Ex, Multinomial model.

The likelihood function $\mathcal{L}(\theta_1, \theta_2, \theta_3 | s_1, \dots, s_n) =$

$= \theta_1^{x_1} \theta_2^{x_2} \theta_3^{x_3}$, x_i - number of i's in the sample.

It is maximized by taking $(\theta_1, \theta_2, \theta_3) = \left(\frac{x_1}{n}, \frac{x_2}{n}, \frac{x_3}{n}\right) \rightarrow$

→ given the likelihood, we can compute counts

n is assumed known → (x_1, x_2, x_3) is a minimal sufficient statistic.

6.2 MLE – estimation problem.

Need to find $\hat{\theta}(s)$ that maximizes $\mathcal{L}(\theta/s)$, i.e.

$$(*) \mathcal{L}(\hat{\theta}(s)/s) \geq \mathcal{L}(\theta/s) \text{ for } \forall \theta \in \Omega.$$

$\hat{\theta}: S \rightarrow \Omega$ satisfying $(*)$ is an MLE; the value $\hat{\theta}(s)$ is called the maximum likelihood estimate

Ex. Let $S = \{1, 2, 3\}$; $\Omega = \{1, 2\}$,

	$s=1$	$s=2$	$s=3$
$f_1(s)$	0.3	0.4	0.3
$f_2(s)$	0.1	0.7	0.2

Observe $s=1$ (chips labelled 1, 2, 3 in given proportions in two bowls; how to select a bowl if a chip labelled 1 has been selected?),

$\Rightarrow \hat{\theta}(1) = 1$, because $0.3 = \mathcal{L}(1/1) > 0.1 = \mathcal{L}(2/1)$.

If we observed $s=2$, then $\hat{\theta}(2) = 2$; if $s=3$, $\hat{\theta}(3) = 1$.

Remarks (i) MLE is not unique; if $f_2(1) = 0$, $f_2(2) = 0.7$, $f_2(3) = 0.3$, then, in addition to $\hat{\theta}(3) = 1$ we also have $\hat{\theta}(3) = 2$.

(ii) MLE is invariant... Instead of $\theta \in \Omega$ use $\psi \in \Gamma = \{\psi(\theta): \theta \in \Omega\}$. E.g. take $\psi(1) = a$, $\psi(2) = b$, so that $\Gamma = \{a, b\}$. The model is $\{g_\psi: \psi \in \Gamma\}$ where $g_\psi = f_\theta$ for the unique value θ s.t. $\psi(\theta) = \psi$. There is a new parameter $\psi \in \Gamma$ - but these are labels only! Thus...

Th. If $\hat{\theta}(s)$ is an MLE when $\theta \in \Omega$, ψ is a 1-1 function defined on Ω, then $\hat{\psi}(s) = \psi(\hat{\theta}(s))$ is an MLE in the new parameterization.

Proof. If we select $\mathcal{L}^*(\psi/s) = g_\psi(s)$, and the original one is $\mathcal{L}(\theta/s) = f_\theta(s)$, then

$$\mathcal{L}^*(\hat{\psi}(s)/s) = g_{\psi(\hat{\theta}(s))}(s) = f_{\hat{\theta}(s)}(s) = \mathcal{L}(\hat{\theta}(s)/s) \geq$$

$$\geq \mathcal{L}(\theta/s) = \mathcal{L}^*(\psi(\theta)/s) \text{ for } \forall \theta \in \Omega. \Rightarrow$$

$$\Rightarrow \mathcal{L}^*(\hat{\psi}(s)/s) \geq \mathcal{L}^*(\psi/s) \quad \forall \psi \in \Psi \quad \text{q.e.d.}$$

6.2.1 Computation of the MLE

Definition. $\mathcal{L}(\cdot/s) \Rightarrow \ell(\cdot/s)$ is $\ell(\cdot/s) = \ln \mathcal{L}(\cdot/s)$ because ln is an increasing function, this works.

Indeed, $\mathcal{L}(\theta/s_1, \ldots, s_n) = \prod_{i=1}^{n} f_\theta(s_i) \Rightarrow \ell(\theta/s_1, \ldots, s_n) = \sum_{i=1}^{n} \ln f_\theta(s_i)$

Define $S(\theta/s) = \frac{\partial \ell(\theta/s)}{\partial \theta}$ – the score function.

To obtain the MLE have to solve $S(\theta/s) = 0$, to guarantee the solution is at least a ~~global~~ local max., check that $\frac{\partial S(\theta/s)}{\partial \theta}\Big|_{\theta=\hat{\theta}(s)} = \frac{\partial^2 \ell(\theta/s)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}(s)} < 0.$

Ex Location normal model. $\frac{n}{2\sigma_0^2}(\bar{x}-\theta)^2$

Here $\mathcal{L}(\theta/x_1, \ldots, x_n) = e$, check that

$S(\theta/x_1, \ldots, x_n) = \frac{n}{\sigma_0^2}(\bar{x}-\theta_0)\Big| = 0$, $\Rightarrow \hat{\theta}(x_1, \ldots, x_n) = \bar{x}$ – the unique solution. Verify that $\frac{\partial S(\theta/x_1, \ldots, x_n)}{\partial \theta}\Big|_{\theta=\bar{x}} = -\frac{n}{\sigma_0^2} < 0.$

Ex Exponential model.

Let $X_1, \ldots, X_n \sim Exp\left(\frac{1}{\theta}\right)$ for $\theta > 0$.

$$\Rightarrow L(\theta / x_1, \ldots, x_n) = \frac{1}{\theta^n} e^{-\frac{n\bar{x}}{\theta}} \Rightarrow \ell(\theta / x_1, \ldots, x_n) = -n \ln\theta - \frac{n\bar{x}}{\theta},$$

and $S(\theta / x_1, \ldots, x_n) = -\frac{n}{\theta} + \frac{n\bar{x}}{\theta^2} \Big/ = 0,$

get $\hat{\theta}(x_1, \ldots, x_n) = \bar{x}$; since $\bar{x} > 0$,

$$\frac{\partial S(\theta / x_1, \ldots, x_n)}{\partial \theta} \Big/ \theta = \bar{x} = \frac{n}{\theta^2} - 2\frac{n\bar{x}}{\theta^3} \Big/ \theta = \bar{x} = -\frac{n}{\bar{x}^2} < 0 ;$$

so $\bar{x}$ is the MLE.

Ex. (No simple formulas for MLE):

Three categories - 1, 2, 3. Probabilities: $p_1 = \theta$, $p_2 = \theta^2$, $p_3 = 1 - \theta - \theta^2$ where $\theta \in [0, (\sqrt{5}-1)/2] = [0, 0.61803]$ is unknown. Note that $\theta + \theta^2 \le 1$ imposes the upper bound on θ; obtain this bound using $\theta + \theta^2 - 1 = 0$ for θ. This relationship between p_i's arises in genetics.

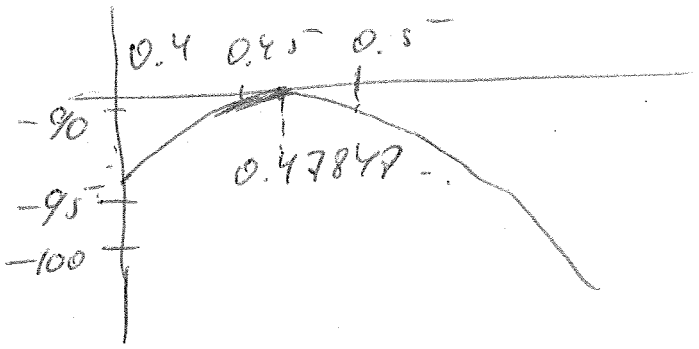
For a sample of n (small relative to the population size) so X_i are iid, $L(\theta / x_1, \ldots, x_n) = \theta^{x_1} \theta^{2x_2} (1 - \theta - \theta^2)^{x_3}$, x_i - sample count in the ith class.

$$\Rightarrow S(\theta / s_1, \ldots, s_n) = \frac{x_1 + 2x_2}{\theta} - \frac{x_3(1+2\theta)}{1 - \theta - \theta^2},$$

$\Rightarrow \hat{\theta}$ is the root of the quadratic equation

$$(x_1 + 2x_2)(1 - \theta - \theta^2) - x_3(\theta + 2\theta^2) = -(x_1 + 2x_2 + 2x_3)\theta^2 - (x_1 + 2x_2 + x_3)\theta + (x_1 + 2x_2)$$

the resulting $\hat{\theta}$ depends on x_1, x_2, x_3 - it is not even guaranteed to be in $[0, 1]$ in general! 4 possible values - 2 roots, & boundary point 0 and 0.61803. Typically, we have to evaluate these numerically by evaluating the likelihood at two points. E.g. if $x_1 = 70$, $x_2 = 5$, $x_3 = 25$, $\theta_1 = -1.28616$, $\theta_2 = 0.47849$,

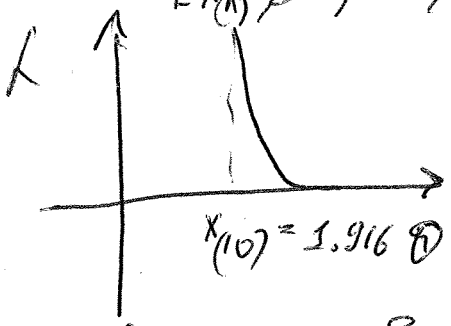


Ex. (6.2.5) Uniform $[0,\theta]$ model – attention to smoothness conditions.

Let $X_1, \dots, X_n \sim \text{Unif}[0,\theta] \Rightarrow \theta > 0$ is unknown.

$$\Rightarrow L(\theta / X_1, \dots, X_n) = \begin{cases} \theta^{-n}, & x_i \le \theta,\ i = 1, \dots, n \\ 0, & x_i > \theta \text{ for at least one } i \end{cases}$$

$$= \theta^{-n} \mathbb{I}_{[X_{(n)}, \infty)}(\theta); \quad \text{e.g. plot when } n = 10,\ x_{(n)} = 1.916$$



$\Rightarrow \hat{\theta} = X_{(n)}$ – cannot be obtained by differentiation!

Ex. 6.2.6 Location – Scale Normal Model

$(X_1, \dots, X_n) \sim N(\mu, \sigma^2)$; $\mu \in \mathbb{R}^1$, $\sigma > 0$ are unknown;

$$\theta = (\mu, \sigma^2) \in \Omega = \mathbb{R}^1 \times (0, \infty)$$

$$\Rightarrow L(\mu, \sigma^2 / x_1, \dots, x_n) = (2\pi\sigma^2)^{-n/2} e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2} e^{-\frac{n-1}{2\sigma^2}s^2};$$

$$\Rightarrow \ell(\mu, \sigma^2 / x_1, \dots, x_n) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{n}{2\sigma^2}(\bar{x}-\mu)^2 - \frac{n-1}{2\sigma^2}s^2$$

First, $\mu = \bar{x}$, plugging it in, we get $-\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{n-1}{2\sigma^2}s^2$

$$\Rightarrow \frac{\partial}{\partial\sigma}: -\frac{n}{2\sigma^2} + \frac{n-1}{2(\sigma^2)^2}s^2 = 0$$

The final MLE of (μ, σ^2) is $\left(\bar{x}, \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$.

Often, we need to estimate some $\psi(\theta)$. Can we say that $\psi(\hat{\theta})$ is the MLE of $\psi(\theta)$? If ψ is 1-1, then yes!

If ψ is not 1-1, we can often find a complementary function λ on Ω s.t. (ψ, λ) is a 1-1 function of θ.

Then, $(\hat{\psi}(s), \hat{\lambda}(s)) = (\psi(\hat{\theta}(s)), \lambda(\hat{\theta}(s))$ is the joint MLE but $\hat{\psi}(s)$ is technically not an MLE. Taken separately, it may perform quite poorly.

Ex. 6.29. Let $X_i \sim N(\mu_i, 1)$, $i = 1, \dots, n$ — independent with unknown means. Then, $\theta = (\mu_1, \dots, \mu_n)$, $\Omega = \mathbb{R}^n$. Need to estimate $\psi(\theta) = \mu_1^2 + \dots + \mu_n^2$.

$\Rightarrow \ell(\theta / x_1, \dots, x_n) = -\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_i)^2$, have $\hat{\theta}(x_1, \dots, x_n) = (x_1, \dots, x_n)$. The "plug-in" MLE is $\hat{\psi} = \sum_{i=1}^{n} x_i^2$

However, $E_\theta\left(\sum_{i=1}^{n} X_i^2\right) = \sum_{i=1}^{n}\left[\mu_i^2 + V(X_i)\right] = n + \psi(\theta)$

; here $E_\theta g$ refers to the expectation of $g(s)$ when $s \sim f_\theta$.

When n is large, $\hat{\psi}$ is clearly far from the true value; should be using $\sum_{i=1}^{n} x_i^2 - n$ instead.