14. Maximum likelihood estimation: MLE (LM 5.2)

14.1 Definition, method, and rationale

(i) The maximum likelihood estimate of parameter θ is the value of θ which maximizes the likelihood $L(\theta)$.

(ii) For data values of an n-sample $x_1, ..., x_n$, outcomes of pdf $f_X(\cdot)$, $L_n(\theta) \;=\; \prod_i f_X(x_i; \theta)$.

(iii) It is therefore often easier to maximize $\ell_n(\theta) \;=\; \log L_n(\theta) \;=\; \sum_i \log f_X(x_i; \theta)$.
This is equivalent to (ii), since log is an increasing function.

(iv) **For the fixed data we have observed**, the MLE value of θ gives higher probability to these data than does any other value of θ. Note we are comparing θ-values as explanations of the observed data $x_1, ..., x_n$. We are not considering other data outcomes we might have got.

(v) But, when we look at the properties of the maximum likelihood estimator (also abbreviated MLE– be careful) (e.g $\overline{X_n}$) then we *are* considering probabilities for other values it might have had.

14.2 Discrete examples

(i) Bernoulli/Binomial: $X_1, ..., X_n$ i.i.d. $Bin(1, \theta)$, $f_X(x) \;=\; \theta^x (1-\theta)^{1-x}$, $x = 0$ or 1. Let $T = \sum_{i=1}^n X_i$.

$$L_n(\theta) \;=\; \prod_i f_X(x_i) \;=\; \theta^{\sum_i x_i} (1-\theta)^{\sum_i (1-x_i)}$$
$$\ell_n(\theta) \;=\; (\sum_{i=1}^n x_i) \log\theta \;+\; (\sum_{i=1}^n (1-x_i)) \log(1-\theta) \;=\; t\log\theta + (n-t)\log(1-\theta)$$

$d\ell/d\theta \;=\; t/\theta - (n-t)/(1-\theta) \;=\; 0$ gives MLE t/n. The maximum likelihood estimator is $T/n = \overline{X_n}$.

(ii) Poisson: $X_1, ..., X_n$ i.i.d. $\mathcal{P}\mathrm{o}(\theta)$, $f_X(x) \;=\; e^{-\theta}\theta^x/x!$, $x = 0, 1, 2, ...$ Let $T = \sum_{i=1}^n X_i$ and $t = \sum_{i=1}^n x_i$.

$$L_n(\theta) \;=\; \prod_i f_X(x_i) \;=\; \exp(-n\theta)\theta^{\sum_i x_i} / \prod_i x_i! \qquad \ell_n(\theta) \;=\; -n\theta \;+\; (\sum_{i=1}^n x_i)\log\theta \;=\; -n\theta + t\log\theta$$

$d\ell/d\theta \;=\; n - t/\theta \;=\; 0$, gives MLE t/n. The maximum likelihood estimator is $T/n = \overline{X_n}$.

14.3 Continuous examples

(i) Exponential: $X_1, ..., X_n$ i.i.d. $\mathcal{E}(\lambda)$, $f_X(x) \;=\; \lambda e^{-\lambda x}$, $x \geq 0$. Let $T = \sum_{i=1}^n X_i$ and $t = \sum_{i=1}^n x_i$.

$$L_n(\lambda) \;=\; \prod_i f_X(x_i) \;=\; \lambda^n \exp(-\lambda \sum_{i=1}^n x_i), \qquad \ell_n(\lambda) \;=\; n\log\lambda \;-\; \lambda(\sum_{i=1}^n x_i) \;=\; n\log\lambda - \lambda t,$$

$d\ell/d\lambda \;=\; n/\lambda - t \;=\; 0$ gives MLE n/t. The maximum likelihood estimator is $n/T = 1/\overline{X_n}$.

(ii) $X_1, ..., X_n$ i.i.d. with $f_X(x; \alpha) = \alpha x^{\alpha-1}$, $0 \leq x \leq 1$. Let $W = \prod_{i=1}^n X_i$ and $w = \prod_{i=1}^n x_i$.

$$L_n(\alpha) \;=\; \prod_i f_X(x_i; \alpha) \;=\; \alpha^n (\prod_{i=1}^n x_i)^{\alpha-1}, \qquad \ell_n(\alpha) \;=\; n\log\alpha \;+\; (\alpha-1)\log w$$

$d\ell/d\alpha \;=\; n/\alpha \;+\; \log w \;=\; 0$ gives MLE $-n/\log w$. The maximum likelihood estimator is $-n/\log W$.

14.4 A non-standard example $X_1, ..., X_n$ uniform $U(0, \theta)$; $f_X(x; \theta) = 1/\theta$, $0 \leq x \leq \theta$.

$L(\theta) = (1/\theta)^n$ provided $0 \leq x_i \leq \theta$ for all i, and 0 otherwise.

That is $L(\theta) = (1/\theta)^n$ provided $\max(x_i) \leq \theta$, and 0 otherwise.

So choose θ as small as possible so that $\theta \geq \max(x_i)$. That is the MLE is $\max_i(X_i)$.

15 Conditional pdf and pmf LM 3.11

15.1 Definition: discrete case

(i) For any two discrete random variables X and W, we the conditional probability mass function is $p_{X|W}(x|w) \; = \; P(X = x|W = w) \; = \; p_{X,W}(x, w)/p_W(w)$, for w such that $p_W(w) > 0$.
Note this is a pmf for X.

(ii) For $X_1, ..., X_n$ an n-sample from a discrete distribution p_X, and W some function of $X_1, ..., X_n$ the conditional pmf is $P(x_1, ..., x_n|W = w) = (\prod_{i=1}^n p_X(x_i))/p_W(w)$ over all $(x_1, ..., x_n)$ giving the value $W = w$.

15.2 Examples

(i) $X_1, ..., X_n$ i.i.d $Bin(1, \theta)$, $W = \sum_{i=1}^n X_i \; \sim \; Bin(n, \theta)$ (LM. P.399)

$p_X(x) \; = \; \theta^x(1-\theta)^{1-x}$ and $p_W(w) = \binom{n}{w}\theta^w(1-\theta)^{n-w}$, and $\sum_{i=1}^n x_i \; = \; w$.

$$P(x_1, ..., x_n \mid W = w) \;\; = \;\; (\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i})/(\binom{n}{w}\theta^w(1-\theta)^{n-w}) \;\; = \;\; 1/\binom{n}{w}$$

Given the total number of successes, the probability of any particular sequence is (1 / (number of ways)) of arranging w successes in n trials (and does not depend on θ).

(ii) $X_1, ..., X_n$ i.i.d. $\mathcal{P}\mathrm{o}(\theta)$, $W = \sum_{i=1}^n X_i \; \sim \; \mathcal{P}\mathrm{o}(n\theta)$

$p_X(x) \; = \; e^{-\theta}\theta^x/x!$ and $p_W(w) = e^{-n\theta}(n\theta)^w/w!$, and $\sum_{i=1}^n x_i \; = \; w$.

$$P(x_1, ..., x_n \mid W = w) \;\; = \;\; (\prod_{i=1}^n e^{-\theta}\theta^{x_i}/x!)/(e^{-n\theta}(n\theta)^w/w!) \;\; = \;\; (w!/\prod_{i=1}^n x_i!).(1/n)^w$$

Again we find the conditional probability dows not depend on θ.
Note, if $n = 2$, this conditional pdf is $Bin(w, \frac{1}{2})$.

15.3 The conditional pdf: continuous case (LM 3.11)

(i) We **define** the conditional pdf $f_{X|W}(x|w) \; = \; f_{X,W}(x, w)/f_W(w)$, for w such that $f_W(w) > 0$.
Note, for each w, $f_{X|W}(x|w)$ is a pdf for X.

(ii) This definition is motivated by

$$P(x < X \leq x + \delta x \mid w < W \leq w + \delta w) \;\; = \;\; \frac{P(x < X \leq x + \delta x \cap w < W \leq w + \delta w)}{P(w < W \leq w + \delta w)} \approx \frac{f_{X,W}(x, w) \; \delta x \; \delta w}{f_W(w) \; \delta w}$$

(iii) For $X_1, ..., X_n$ an n-sample from a continuous pdf f_X, and W some function of $X_1, ..., X_n$ the conditional pdf is $f(x_1, ..., x_n \mid W = w) \; = \; \prod_{i=1}^n f_X(x_i)/f_W(w)$ over all $(x_1, ..., x_n)$ giving the value $W = w$.

15.4 Examples

(i) $X_1, ..., X_n$ i.i.d. $\mathcal{E}(\lambda)$, $W \; = \; \sum_{i=1}^n X_i \; \sim \; G(n, \lambda)$.

$f_X(x; \lambda) \; = \; \lambda e^{-\lambda x}$ on $x \geq 0$; $f_W(w; \lambda) \; = \; (\lambda^n/\Gamma(n)).w^{n-1}e^{-\lambda w}$ on $w \geq 0$, and $w = \sum_{i=1}^n x_i$.

The conditional pdf of the sample, given $W = w$, is

$$f(x_1, ..., x_n \mid W = w) \; = \; \frac{\prod_{i=1}^n \lambda \exp(-\lambda x_i)}{(\lambda^n/\Gamma(n)).w^{n-1}e^{-\lambda w}} \; = \; \frac{\Gamma(n)}{w^{n-1}}$$

Again, we have managed to choose a W such that the conditional pdf does not depend on the parameter.

(ii) $X_1, ..., X_n$ i.i.d. $U(0, \theta)$, $W \; = \; \max_{i=1,...,n} X_i$

$f_X(x; \theta) \; = \; 1/\theta$ on $0 \leq x \leq \theta$; $f_W(w; \theta) \; = \; nw^{n-1}/\theta^n$ on $0 \leq w \leq \theta$, and $w = \max_i x_i$.

The conditional pdf of the sample, given $W = w$, is

$$f(x_1, ..., x_n \mid W = w) \; = \; \frac{\prod_{i=1}^n (1/\theta)}{nw^{n-1}/\theta^n} \; = \; \frac{1}{nw^{n-1}}$$

Again, we have managed to choose a W such that the conditional pdf does not depend on the parameter.

16. Sufficient statistics and the factorization criterion LM 5.6

16.1 Definition LM P.407.

(i) A statistic $T(X_1, ..., X_n)$ is *sufficient* for inferences about parameter θ is the conditional pmf/pdf of the sample, given the value of T does not depend on θ.

(ii) Examples: 15.2 (i),(ii) and 15.4 (i),(ii): in each case we found a W ($\sum_i X_i$ or $\max_i(X_i)$) for which $f(x_1, ..., x_n|W = w)$ did not depend on the parameter θ. In each case the statistic W is *sufficient* for θ.

(iii) The idea is that the sufficient statistic contains all the information about θ that there is in the entire sample. If you know the value of the sufficient statistic, you will not gain anything more by knowing $(x_1, ..., x_n)$.

(iv) Example: 15.2 (i) is the clearest. If you know the number of sucesses $W = \sum_1^n X_i \sim Bin(n, \theta)$, you know as much about θ as if you know the complete sequence of successes and failures (1 and 0).

16.2 Factorizing the Likelihood LM P.407, Definition 4.6.1.

(i) Note, by definition, $f(x_1, ..., x_n|T = t) = f(x_1, ..., x_n)/f_T(t)$, so in likelihood terms we have

$$\begin{aligned} L_n(\theta) &= f(x_1, ..., x_n; \theta) = f(x_1, ..., x_n|T = t) f_T(t; \theta) \\ \ell_n(\theta) &= \log f(x_1, ..., x_n|T = t) + \log f_T(t; \theta) \end{aligned}$$

Conversely, if the likelihood factorizes in this way, with the first term not depending on θ then T is sufficient.

(ii) Note the MLE will depend only on T; the conditional term is just a "constant" (multiplicative in the likelihood, additive in the log-likelihood). It does not affect the value of θ that maximizes $L(\theta)$ or $\ell_n(\theta)$.

(iii) Recall it is *relative* values of the likelihood that matter; we compare the likelihoods for different θ-values for the same data. Note $L_n(\theta)/L_n(\theta^*) = f_T(t; \theta)/f_T(t; \theta^*)$ also depends only on the value t of the sufficient statistic T, and not otherwise on $x_1, ..., x_n$.

16.3 Fundamental principle: all inferences should be based only on sufficient statistics. Why?

(i) They contain "all the information"

(ii) They determine the (log-)likelihood function, up to a constant factor.

(iii) **Rao-Blackwell Theorem: Approximate Statement only**(LM P.405): If an estimator W is not a function only of the sufficient statistic, and T is sufficient, then there is a function of T which, for every θ,

(a) has the same expectation as W, and (b) has smaller mean square error than W.

(Subject to some conditions, there is one and only one such function of T.)

(iv) Example: $X_1, ..., X_n$ i.i.d. $U(0, \theta)$, $T = \max_i X_i$ is sufficient (see 15.4 (ii)). MoM estimator $2\overline{X_n}$ is unbiased, but not a function of T. Estimator $(n + 1)T/n$ is also unbiased and has smaller variance or mse.

16.4 A (second) factorization criterion (LM P403)

(i) It would be a pain to have to identify T, and then check the conditional pdf $f(x_1, ..., x_n|T = t)$ to make sure it does not depend on θ. Fortunately, we don't have to.

(ii) **Theorem:** $T = h(X_1, ..., X_n)$ is sufficient for θ **if and only if** $L_n(\theta) = g(h(x_1, ..., x_n); \theta).b(x_1, ..., x_n)$.

Proof: If T is sufficient, this holds with $g(h; \theta) \equiv f_T(h; \theta)$ and $b(x_1, ..., x_n) = f(x_1, ..., x_n|T = t)$.

Conversely, if we have the factorization, it can be shown that $g(h; \theta) \propto f_T(h; \theta)$, where the factor does not depend on θ, so then we have previous factorization and T is sufficient (LM P.404).

(iii) **Conclusion:** To find sufficient statistics:

(a) write down the (log)-likelihood; (b) Find what functions of $(x_1, ..., x_n)$ are *inextricably mixed up with* θ;

(c) These are the sufficient statistics!

Note: as n case of $U(0, \theta)$ the "mixed up with" may come through the range on the r.v.s.